ГРУППА АГРОПРЕДПРИЯТИЙ
РЕСУРС

Документ

СИСТЕМА ИНТЕГРИРОВАННОГО ПЛАНИРОВАНИЯ РАСТЕНИЕВОДСТВА

ООО «Ростовская зерновая компания «Ресурс»





www.gapresurs.ru

Содержание

1. Описание функциональных характеристик системы по модулям

1.1. Система интегрированного планирования Растениеводства (далее по тексту – СИПР) предназначена для автоматизированной поддержки бизнес-процессов долгосрочного и производственного планирования сезона в сегменте растениеводства. Функционал системы включает:

1.1.1. Проведение оптимизационного моделирования в части:
- модели оптимального размещения культур (структуры посевных площадей);
- алгоритмов формирования плана посевных работ;
- проведения моделирования балансировки сельскохозяйственной техники.

1.1.2. Единый интерфейс планирования производства, основанный на ведении технологических карт культур.

1.1.3. Формирование производственных задач.

1.1.4. Отслеживание выполнения задач в реальном времени, с возможностью просмотров отчетов за прошлые периоды.

1.1.5. Интеграцию с производственным оборудованием для получения/передачи информации.

1.1.6. Инструменты оркестрации для запуска и мониторинга работы алгоритмов в среде оптимизации.

1.1.7. Проведение расчетов для формирования показателей отчетности.

1.1.8. Набор механизмов проверки входных данных (сценария) перед запуском моделирования:
- полноты данных;
- корректности данных;
- логических проверок.

1.1.9. Набор пользовательских интерфейсов для выполнения шагов планирования в соответствии с бизнес-ролью пользователя:
- области планирования и формы ввода данных для подготовки входных данных;
- пульт управления интегрированным планированием для запуска процесса моделирования и мониторинга процессов расчета;
- формирование отчетности и анализ результатов;
- факторный анализ и сравнение сценариев.

1.1.10. Механизмы интеграции с внешними системами (системы оперативного контроля, корпоративные хранилища данных).

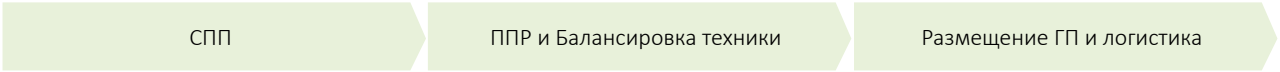
Рисунок № 1. Функциональная архитектура решения СИПР

1.2. Формирование производственной программы – единая платформа интегрированного планирования для поддержки принятия стратегических решений и согласованного с ними оперативного управления. В основе решения лежит создание оптимизационной платформы для автоматизации сквозных расчетов, учитывающих как процессы стратегического планирования, так и процессы производственного планирования сезона, с поддержкой возможности анализа различных сценариев развития Растениеводства.

Подход к автоматизации задач долгосрочного планирования и планирования производственной программы на сезон представляет собой использование единого вычислительного графа, состоящего из набора взаимосвязанных оптимизационных моделей, выполняемых по цепочке. При этом результат работы одной модели является входными данными для работы следующей по цепочке модели.

Общая схема связей и порядка выполнения оптимизационных моделей в вычислительном графе приведена на *Схеме № 1*, где:
СПП – Структура посевных площадей
ППР – План полевых работ
ГП – Готовая продукция

Схема № 1. Связь и порядок выполнения оптимизационных моделей



Автоматизированный подход поддержки процессов планирования включает:
1.2.1. Формирование сценария для моделирования:
- входные данные;
- параметры сценария.

1.2.2. Подготовку и расчет сценария.
1.2.3. Анализ результатов.

Процесс планирования производственной программы выполняется в несколько шагов. Каждый шаг – это сценарий модели конкретного вида (СПП, ППР и пр.). Иерархическая последовательность сценариев формирует цепочку оптимизационных моделей, результатом которой является вариант производственного плана.

Пользователи СИПР формируют несколько вариантов производственного плана, сравнивают их, анализируют финансовый показатели и выбирают план, который утверждается в качестве «производственной программы». Утвержденная производственная программа может быть передана в систему оперативного контроля или сезонного планирования растениеводства для реализации.

1.3. Оптимизационное моделирование – в основе цепочки моделей находится оптимизационная модель СПП, задача которой спланировать посевы на полях. Результатом моделирования СПП являются следующие данные:
1.3.1. структура посевных площадей;
1.3.2. плановая урожайность культур на полях;
1.3.3. оценка потребности в наемной уборочной технике с учетом имеющегося парка сельскохозяйственной техники и оборудования.

Результаты моделирования СПП используются в качестве входных данных для оптимизационной модели ППР. В результате моделирования ППР определяются сроки и ресурсы для работ по выращиванию культур и графику уборки урожая. Реализация вычислительного графа поддерживает возможность пересчета модели ППР без пересчета модели СПП.

Результаты моделирования ППР будут использоваться в качестве входных данных для следующей по цепочке группы оптимизационных моделей по размещению готовой продукции, задача которой на основе уточненного плана полевых работ сформировать план размещения продукции на местах хранения и соответствующую логистику от полей до мест размещения.
Подготовка к запуску процесса моделирования заключается в подготовке входных данных, задании параметров сценария, а также в выполнении проверки качества входных данных.

Вне зависимости от того, запускал ли пользователь проверку качества данных вручную, при запуске расчета модели (процесса оптимизации) предварительным шагом в обязательном порядке автоматически запускается проверка данных. При наличии критических ошибок процесс останавливается, уведомляя пользователя о перечне замечаний к данным. Пользователь обязан устранить критические ошибки. Некритичные замечания к данным не обязательны к устранению, пользователь может запустить расчёт без их устранения.

Если проверка качества данных не выявила критичных замечаний, то осуществляется запуск оптимизационного расчета, при этом сценарий переходит в статус «В процессе расчета». Во время расчета ведется журнал расчета, в котором отражается текущая стадия расчета. По завершении расчета сценарий переходит в один из статусов: «Расчет завершен успешно. Решение найдено», «Расчет завершен успешно. Решение не найдено», «Расчет завершен неуспешно».

1.4. Аналитическая отчетность – в системе реализуется возможность формирования следующих отчетов:
   1.4.1. производственные отчеты (выполняется на базе результатов расчета соответствующих оптимизационных моделей);
   1.4.2. экономические отчеты (выполняется на базе результатов расчета всей цепочки моделей);
   1.4.3. отчеты по сравнению базового сценария и сценария с инвестициями в технику и места хранения (выполняется на базе результатов расчетов всей цепочки моделей двух сценариев).

Производственные отчеты формируются по мере выполнения каждого из блоков моделей. В состав производственных отчетов системы входят следующие:

- Структура посевных площадей;
- Сводная производственная программа;
- План работ сельскохозяйственной техники в формате отчета и гант-диаграммы;
- Отчет по потребности в технике для сезонов долгосрочного планирования (с оценкой дефицита техники);
- План потребности в персонале (с оценкой дефицита);
- План размещения собственной и наемной техники по филиалам;
- План перемещения техники между филиалами;
- Сводная таблица по валовому сбору для очереди уборки;
- Схема размещения готовой продукции (с оценкой дефицита техники);
- Расчет потребности в грузовом автотранспорте для уборочной кампании.

1.5. Уровни доступа к информационным ресурсам – встроенные средства системы позволяют ограничить доступ пользователей СИПР к информационным ресурсам как на уровне доступа к компонентам, так и на уровне объектов полномочий каждого компонента системы. Объектами полномочий в компонентах системы являются:
   1.5.1. Компонент Dashboards:

- аналитические отчеты;
- датасеты.

1.5.2. Компонент формы ввода:
- пользовательские формы;
- сценарии запуска оптимизации;
- датасеты.

1.5.3. Компонент Analytics:
- OLAP-кубы.

1.5.4. Компонент ML:
- оптимизационные модели.

Доступ пользователей к информационным ресурсам СИПР предполагает выполнение процедур аутентификации и авторизации. Процедуры аутентификации и авторизации пользователей выполняются в СИПР с использованием решения для управления идентификацией и доступами Keycloak с возможностью подключения к службе Active Directory.

Предоставляемые на данном уровне полномочия обеспечивают доступ пользователей СИПР непосредственно к информации, обрабатываемой и хранящейся в СИПР, а также к средствам автоматизированной обработки этой информации.

Полномочия на доступ к функциональности и информационным ресурсам Системы определяются ролевой моделью доступа. Ролевая модель обеспечивает разграничение доступа пользователей к информационным ресурсам.

Полномочия пользователей СИПР определяются с учетом:
- типов пользователей системы;
- уровней доступа к информационным ресурсам системы посредствам прикладного программного обеспечения (далее по тексту – ПО);
- разделения функциональных обязанностей.

Управление полномочиями доступа пользователей (предоставление, изменение, аннулирование) обеспечивается средствами прикладного ПО Polymatica в интерфейсе пользователя с правами администратора.

2. Информация об условиях использования программного обеспечения

2.1. Стоимость использования программного обеспечения на условиях неисключительной бессрочной лицензии для одного юридического лица – 15 000 000 (пятнадцать миллионов) рублей 00 копеек.

2.2. Для полнофункциональной работы Системы интегрированного планирования Растениеводства требуется наличие лицензий на ПО Полиматика.

3. Установка необходимого программного обеспечения (Инструкция по инсталляции)

3.1. Для работы СИПР необходима установка СУБД PostgreSQL а также распределенной системы хранения SeaweedFS. Установку рекомендуется выполнять на один или несколько выделенных серверов.
Установка СУБД PostgreSQL выполняется согласно инструкции для соответствующего дистрибутива, расположенной в сети интернет по адресу https://www.postgresql.org/download/.

Последовательность установки SeaweedFS представлена и доступна на ресурсе по адресу https://github.com/chrislusf/seaweedfs/wiki/Getting-Started.

После установки указанного программного обеспечения, для работы системы требуется запустить сервисы ***master***, ***filer***, ***s3***, ***volume***, предварительно сконфигурировав SeaweedFS.

3.2. Конфигурация SeaweedFS – настроить роль для **S3 API** с полными правами, создав файл ***/etc/seaweedfs/s3-config.json*** со следующим содержимым:

```
{
“identities”: [
{
“name”: “anonymous”,
“actions”: [
“Read”,
“List”
]
},
{
“name”: “admin”,
“credentials”: [
{
“accessKey”: “qwerty”,
“secretKey”: “qwerty”
}
],
“actions”: [
“Admin”,
“Read”,
“List”,
“Tagging”,
“Write”
]
}
]
}
```

Данную конфигурацию следует указывать при запуске сервиса S3 через параметр ***config***:-config=/etc/seaweedfs/s3-config.json. Также необходимо создать необходимые бакеты хранения S3, выполнив на сервере SeaweedFS команды:

```
echo “s3.bucket.create-name mm” | weed shell
echo “s3.bucket.create-name md” | weed shell
echo “s3.bucket.create-name cs” | weed shell
echo “s3.bucket.create-name cnf” | weed shell
```

3.3. Установка компонента Polymatica (конфигурация Polymatica ML). Перед установкой следует указать основные параметры развертывания в файле ***charts/polymatica-ml/values.yaml*** в следующих разделах:

3.3.1. Раздел global

- RegistryUrl – адрес для доступа к docker registry в виде <ip-адрес или DNS-имя сервера ***polymatica ml>:5000***
- DomainName – ip-адрес или DNS-имя для доступа к серверу
- GatewayUrl – адрес для доступа к публикуемым сервисам в виде ***http://<ip-адрес или DNS-имя сервера polymatica ml>/***
- MetadataServerUrl – адрес для доступа к сервису метаданных в виде ***http://<ip-адрес или***

DNS-имя сервера polymatica ml>

- Подраздел postgres:
  - host – адрес или имя хоста PostgreSQL
  - port – порт PostgreSQL
  - user – пользователь для подключения к БД PostgreSQL
  - password – пароль для подключения к БД PostgreSQL
  - databaseMeta – имя БД сервиса метаданных в PostgreSQL
  - databaseDD – имя БД сервиса datadiscovery в PostgreSQL
  - databaseMD – имя БД сервиса modeldesigner в PostgreSQL
- Подраздел s3:
  - url – адрес сервиса S3 API в развернутой SeawedFS в виде ***http://<ip-адрес*** или DNS-имя сервера ***sraweedfs>:<порт>***

3.3.2. Раздел modelmanager

- Подраздел postgres
  - user – пользователь для подключения к БД PostgreSQL
  - password – пароль для подключения к БД PostgreSQL
  - databaseMM – имя БД сервиса modelmanager в PostgreSQL
- Подраздел smtp
  - username – пользователь для подключения к SMTP-серверу
  - password – пароль для подключения к SMTP-серверу
  - from – адрес отправителя
  - port – порт SMTP-сервера
  - server – ip-адрес или DNS-имя SMTP-сервера
- Подраздел camunda
  - jdbcUrl – строка подлючения к БД camunda в виде ***jdbc:postgresql://<ip-адрес*** или DNS-имя сервера ***PostgreSQL>:<порт>/cs_dev1?sslmode=disable***
  - jdbcUser – пользователь для подключения к БД PostgreSQL
  - jdbcPassword – пароль для подключения к БД PostgreSQL
  - keycloakUrl – адрес для доступа к сервису keycloak в виде ***http://<ip-адрес или DNS-имя сервера polymatica ml>***
- Подраздел modellauncher
  - devpi_host – адрес для доступа к репозиторию devpi в виде ***http://<ip-адрес*** или DNS-имя сервера ***polymatica ml>:3141***

3.3.3. Раздел keycloak

- Подраздел postgres
  - user – пользователь для подключения к БД PostgreSQL
  - password – пароль для подключения к БД PostgreSQL
  - databaseName – имя БД сервиса keycloak в PostgreSQL

3.4. Установку Polymatica ML следует выполнять под непривилегированным пользователем с правами sudo. Для установки запустить install.sh из каталога с дистрибутивом: DOMAINNAME=<ip-адрес или DNS-имя сервера polymatica-ml> ./install.sh В переменной DOMAINNAME указывается сетевой адрес или DNS-имя сервера, которые будут использоваться для доступа к серверу Polymatica ML. После установки рекомендуется перезапустить службу k3s: sudo systemctl restart k3s

ГРУППА АГРОПРЕДПРИЯТИЙ
РЕСУРС

Документ

СИСТЕМА ИНТЕГРИРОВАННОГО ПЛАНИРОВАНИЯ РАСТЕНИЕВОДСТВА

Руководство пользователя





www.gapresurs.ru

ОГЛАВЛЕНИЕ

1. Общая часть

1.1. Архитектура

Система представляет собой тонкий клиент (thin client) с веб-интерфейсом. В основу серверной части (Backend) Модуля заложена микросервисная архитектура.

На рисунке приведена целевая функциональная архитектура Модуля.

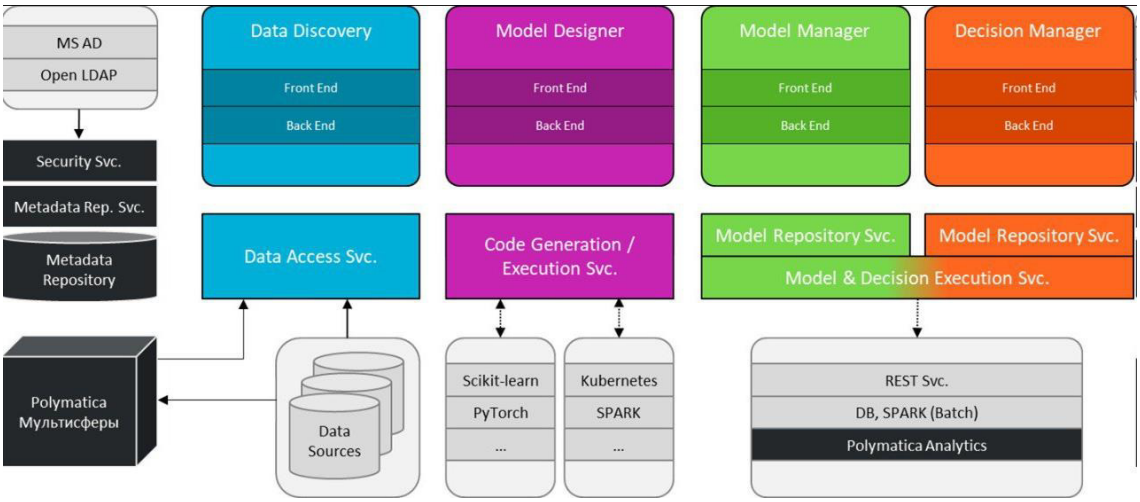


Рисунок 1 Функциональная архитектура Модуля

Основными компонентами архитектуры являются:

- Интерфейс пользователя (Frontend/Backend). Выделяются следующие подкомпоненты пользовательского интерфейса:
    - **Исследование данных (Data Discovery)** — пользовательский интерфейс для выбора исследуемых данных и их последующего интерактивного анализа. В текущей версии Модуля обеспечивается поддержка следующих операций:
        - выбор источника данных и формирования выборки;
        - визуализация выборки или ее части;
        - профилирование данных и оценка выборки по основным статистическим критериям;
        - отбор и преобразование признаков;
        - тестирование статистических гипотез о виде распределения непрерывных и дискретных наблюдений.
    - **Построение моделей машинного обучения (Model Designer)** — пользовательский интерфейс, позволяющий пользователю в графическом режиме в виде последовательности шагов (pipeline) построить модели машинного обучения, провести оценки точности моделей и сохранить модели в репозитории для дальнейшего использования.

В текущей версии Модуля обеспечивается поддержка следующих функций:

- отбор и преобразование признаков;
- разделение выборки на обучающую, валидационную и проверочную;
- построение моделей машинного обучения различными методами;
- автоподбор гиперпараметров модели;
- кросс-валидация;
- тестирование моделей;
- интерпретация моделей;
- скоринг и оценка качества моделей — расчет, табличная и графическая визуализация различных характеристик модели;
- выбор лучшей модели по заданному критерию;
- возможность переобучения моделей при появлении новых данных;
- сохранение и регистрация моделей в репозитории.

- **Управление жизненным циклом моделей (Model Manager)** — интерфейс пользователя для управления репозиторием моделей. Позволяет публиковать модели для применения в пакетном режиме или режиме сервиса, а также настраивать процессы согласования моделей. В текущей версии Модуля обеспечивается поддержка следующих операций:

  - ведение единого репозитория и версионности моделей;
  - публикация моделей для применения в пакетном режиме — формирование скрипта для применения модели на данных в БД;
  - публикация моделей в качестве сервиса;
  - публикация моделей из репозитория в среду Polymatica;
  - сравнение моделей и выбор лучшей;
  - снятие моделей с публикации.

- **Управление решениями (Decision Manager)** – интерфейс пользователя для создания цепочек решений. В текущей версии Модуля обеспечивается поддержка следующих функций:

  - Построение цепочек решений в графическом интерфейсе с возможностью комбинирования моделей ИИ и экспертных правил;
  - Создание экспертных правил в графическом режиме;
  - Тестирование цепочек решений.

- Сервис управления метаданными (Metadata Rep. Svc.) — компонент отвечает за хранение и управление внутренними метаданными модуля, такими как параметры подключения к источникам данных, проекты по построению моделей, параметры для публикации моделей и т.д. Остальные компоненты решения используют сервис управления метаданными для сохранения, чтения, изменения внутренних метаданных.

- Сервис управления безопасностью (Security Svc.) — компонент отвечает за управление правами доступа, интеграцию с каталогами пользователей (Open LDAP, MS ActiveDirectory), аутентификацию и авторизацию пользователей.
- Сервис по доступу к источникам данных (Data Access Svc.) — данный компонент отвечает за доступ к различным источникам и получение из них данных.
- Сервис по генерации и запуску вычислений (Code Generation and Execution Svc.) — данный компонент предназначен для перевода команд в выполняемый код на языке Python. В решении предполагается механизм регистрации различных библиотек Python (scikit-learn и др.), содержащих алгоритмы машинного обучения. При работе с данными в компоненте Data Discovery, а также при создании проекта и построении моделей в компоненте Model Designer — выполняется генерация кода Python в соответствии с заданными настройками.
- Сервис по управлению и публикации моделей (Model Repository Svc.) — компонент содержит процедуры, необходимые для управления жизненным циклом моделей — ведение реестра моделей, процессы согласования, процессы публикации.
- Сервис(ы) исполнения моделей (Model Execution Svc.) — это сервис(ы) исполнения моделей, т.е. каждая опубликованная модель — это исполняемый сервис.

1.2. Интерфейс

1.2.1. Авторизация

При запуске Модуля открывается экран авторизации в системе.

Логин и пароль учетной записи Пользователь должен получить у своего системного администратора.
При корректной авторизации в системе в зависимости от выданных пользователю прав откроется один из компонентов Модуля.

1.2.2. Основное меню

Основным элементом интерфейса является панель меню, расположенная в верхней части экрана (рисунок ниже). Она неизменна для всех компонентов Модуля.

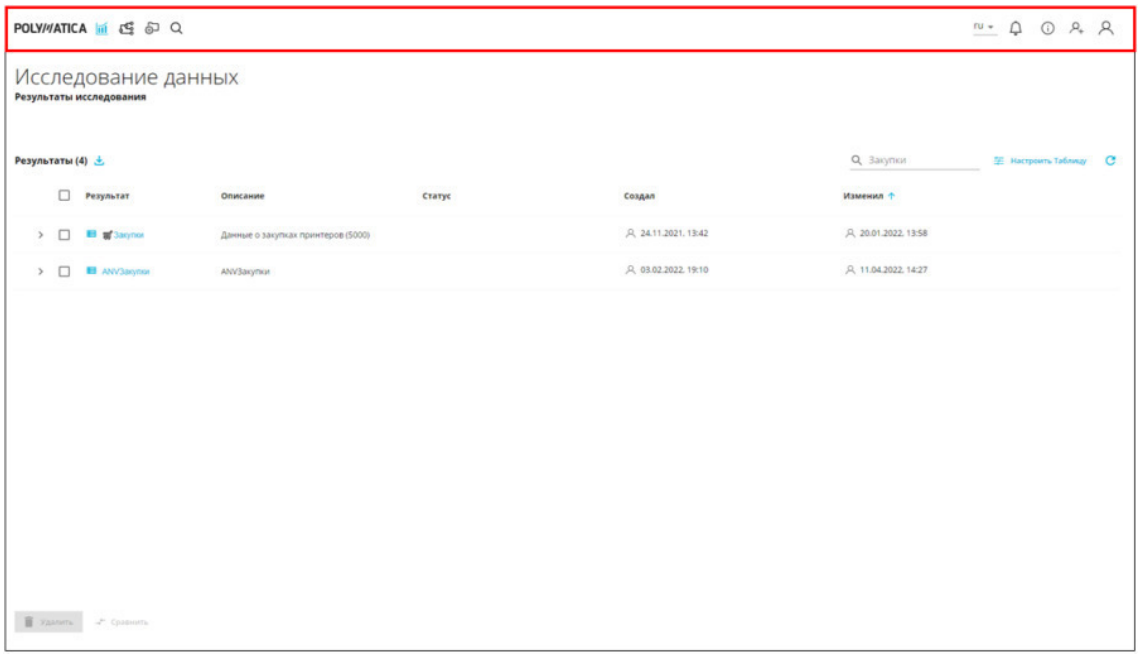


Рисунок 2 Интерфейс модуля. Меню

В левой части панели меню расположены объекты, которые позволяют:

- Выбор иконки  открывает компонент Исследование данных (Data Discovery, DD).
- Выбор иконки  открывает компонент Построение модели (Model Designer, MD).
- Выбор иконки  открывает компонент Разработка решений (Decision Manager, DM).
- Выбор иконки  открывает компонент Управление моделями и решениями (Model Manager, MM).
- Выбор иконки  открывает интерфейс глобального поиска.

В правой части меню расположены объекты, которые позволяют:

- Для выбора языка отображения интерфейса предусмотрен список  с доступными языками.
- Выбор иконки  открывает боковую панель с уведомлениями о завершении расчета результатов Исследований в DD (Рисунок а).
- Выбор иконки  открывает боковую панель с интерактивной справкой (Рисунок б).

- Выбор иконки  открывает раздел Администрирования.

- Выбор иконки  открывает боковую панель с информацией об учетной записи пользователя, о лицензии и кнопкой выхода из Модуля (Рисунок в).

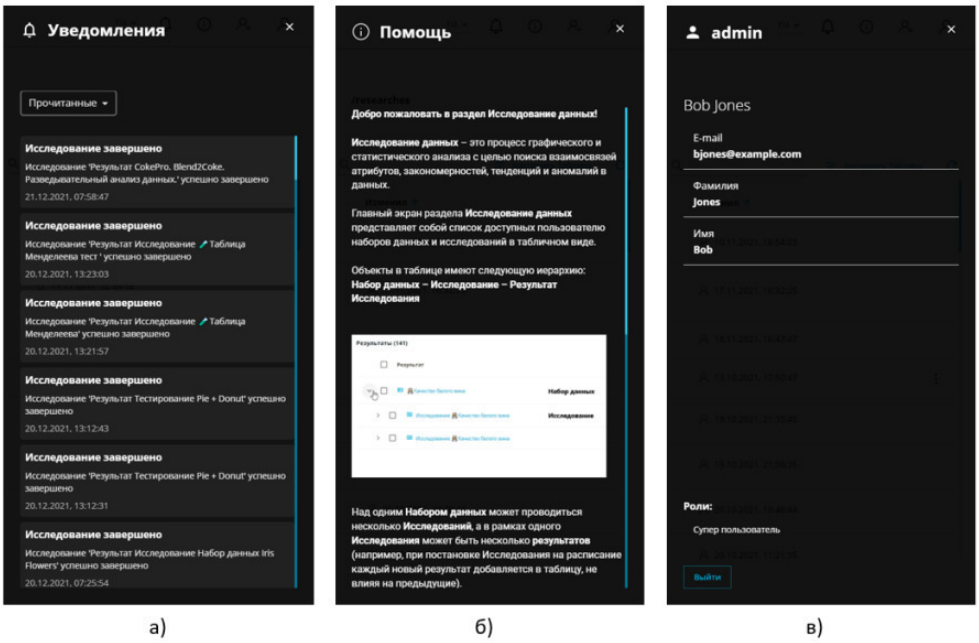


Рисунок 3 Боковая панель Уведомления (а), боковая панель Помощь (б), боковая панель Пользователь (в)

1.2.3. Роли пользователей

В зависимости от выданных пользователю прав в интерфейсе отображаются соответствующие компоненты:

- **Аналитик** имеет доступ к функционалу компонентов Исследование данных (DD) и Построение моделей (MD).
- **Менеджер моделей** имеет доступ ко всему функционалу компонента Управление моделями (MM).
- **Пользователь моделей** имеет доступ к разделу Опубликованные модели MM.
- **Согласующий** имеет доступ к разделу Согласование MM.
- **Суперпользователь** имеет полный доступ к функционалу Модуля.
- **Системный администратор** имеет доступ только к разделу Администрирование.

	DD	MD	MM	Admin
Аналитик	✓	✓		
Менеджер моделей			✓	
Пользователь моделей			✓	
Согласующий			✓	
Суперпользователь	✓	✓	✓	✓
Системный администратор				✓

Таблица 1 Доступ к компонентам по ролям

2. Компонент Исследование данных (Data Discovery, DD)

Этап исследования данных представляет собой процесс графического и статистического анализа, необходимый для проверки общего качества данных и поиска взаимосвязей атрибутов, тенденций и аномалий. Полученная на данном этапе информация позволяет сформулировать гипотезы о том, как данные помогут решить поставленную задачу.

Для этого в компоненте Исследование данных предусмотрены следующие инструменты:

- Графическое представление данных (одномерные и многомерные графики и таблицы).
- Профилирование данных (статистические характеристики).
- Корреляционные матрицы.
- Кластерный анализ.
- Статистические тесты.
- Периодический запуск исследования данных (постановка исследований на расписание).

Компонент включает:

- Главный экран со списком доступных исследований.
- Мастер настройки исследования.
- Экран с результатами исследования.
- Окно постановки исследования на расписание.
- Вспомогательные окна настройки вида, пример данных и т.д.

2.1. Главный экран

2.1.1. Интерфейс главного экрана DD

Главный экран раздела **Исследование данных** открывается при выборе иконки  в левом верхнем меню и представляет собой список доступных пользователю наборов данных и исследований в табличном виде (рисунок ниже).

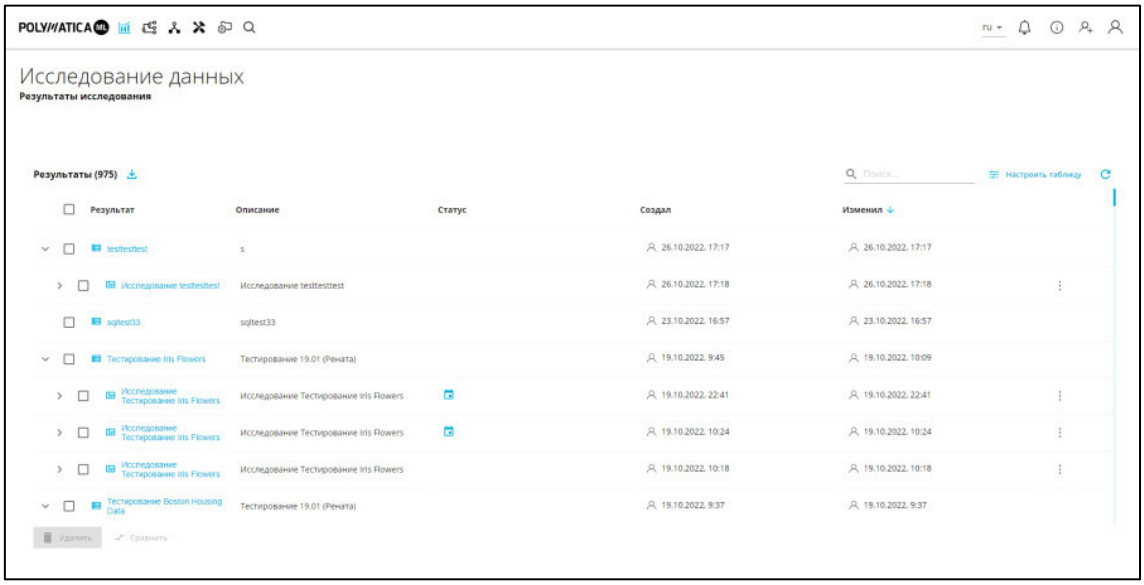


Рисунок 4 Главный экран компонента Исследование данных (Data Discovery)

Объекты в таблице имеют Иерархию в соответствии с рисунком ниже.

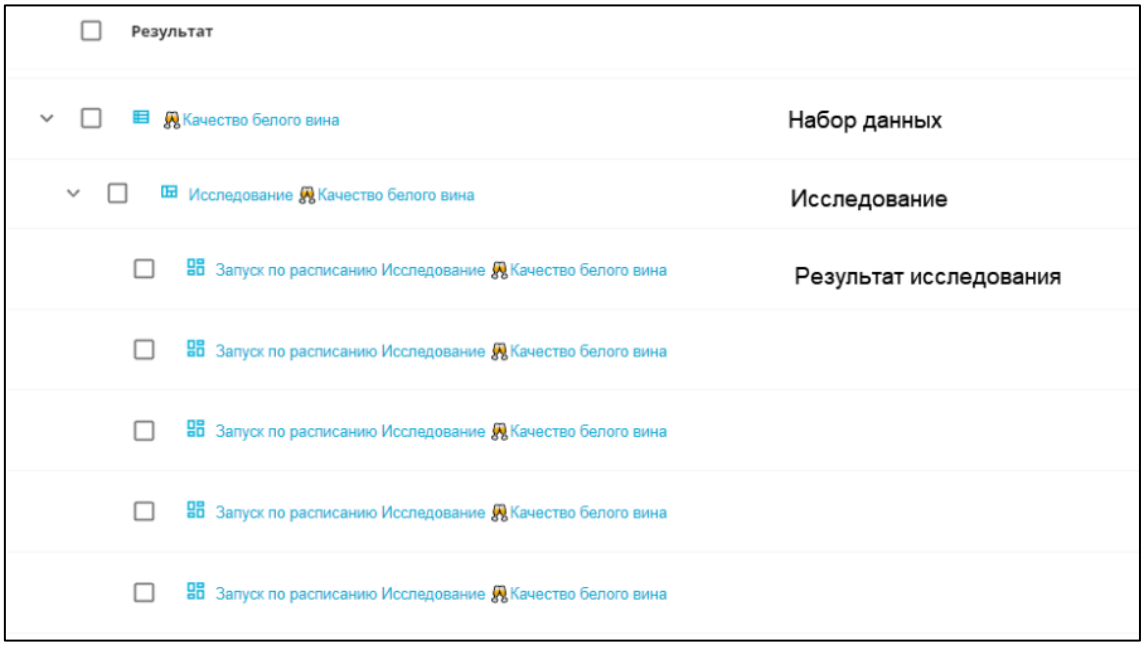


Рисунок 5 Пример иерархии объектов в таблице Главного экрана компонента Исследование данных

Для одного **Набора данных** (имеет символьное обозначение) может проводиться несколько **Исследований** (обозначается), например, с разным наборов визуализаций и объектов. В рамках одного Исследования может быть несколько **Результатов** (обозначается) — например, при постановке Исследования на расписание.

Таблица с доступными исследованиями имеет гибкие настройки отображения. Так пользователь может:

- изменить ширину любого столбца (для этого необходимо перетащить границу его заголовка до нужной ширины);
- сортировать таблицу (для этого необходимо выбрать иконку рядом с заголовком сортируемого столбца);
- скрывать/отображать столбцы и изменять их порядок в окне **Вид таблицы** (для открытия окна необходимо выбрать Настроить Таблицу в правом верхнем углу таблицы; при выборе иконки столбец скроется, при наведении на иконку активируется возможность перемещения столбца);
- сбросить внесенные изменения также в окне **Вид таблицы** (для этого выбрать кнопку **«Сбросить»**).

Для быстрого поиска объекта в таблице предусмотрено поле Поиск... в правой верхней части таблицы.
Объекты таблицы можно выгрузить в формате Excel. Для этого нужно выбрать иконку **Экспорта в excel** .

Другие подсказки по интерфейсу Главного экрана:

- Описание набора данных подтягивается при его регистрации (подробнее в Руководстве Администратора), для Исследования задается в окне **Мастера настройки**.
- Исследование может иметь **Статус**, связанный с постановкой на расписание (иконка — Запланировано), результат исследования может иметь **Статус**, связанный с ходом выполнения исследования (иконка — Обрабатывается, иконка — Готово).
- Иконка рядом с датой создания и изменения позволяет узнать, кто является автором исследования/изменения исследования.
- Иконка позволяет открыть выпадающее меню, в котором Пользователь может **Запустить исследование.**

2.1.2. Создание исследования

Для создания нового Исследования необходимо выполнить следующие шаги:

- Выбрать необходимый набор данных из списка доступных.
- В открывшемся **Мастере настройки** выбрать и задать настройки:
  - визуализаций (подробнее в разделе Визуализации),
  - корреляции (подробнее в разделе Корреляция),
  - кластеризации (подробнее в разделе Кластеризация),
  - статистических тестов (подробнее в разделе Статистические тесты).

- Запустить расчет Исследования, выбрав пункт **Запустить** в выпадающем меню рядом с необходимым Исследованием.
- После успешного выполнения Исследования (статус «Готово» ) открыть результаты Исследования, выбрав из списка необходимый.

2.1.3. Удаление исследования

Для удаления Исследования (результата Исследования) необходимо выполнить следующие шаги:

- Выбрать чекбокс рядом с удаляемым объектом (при выборе чекбокса рядом с Набором данных выбираются все объекты, нижестоящие в иерархии — это все Исследования данного набора и их результаты).
- Выбрать кнопку **«Удалить»**.

2.2. Мастер настройки

2.2.1. Интерфейс Мастера настройки

При создании нового исследования открывается окно **Мастер настройки**, который включает в себя следующие вкладки:

- Параметры.
- Роли атрибутов.
- Визуализации.
- Корреляция.
- Кластеризация.
- Статистические тесты.

2.2.2. Параметры

Во вкладке **Параметры** пользователь может задать Название исследования и Описание (рисунок ниже), которые отображаются в таблице на Главном экране компонента.

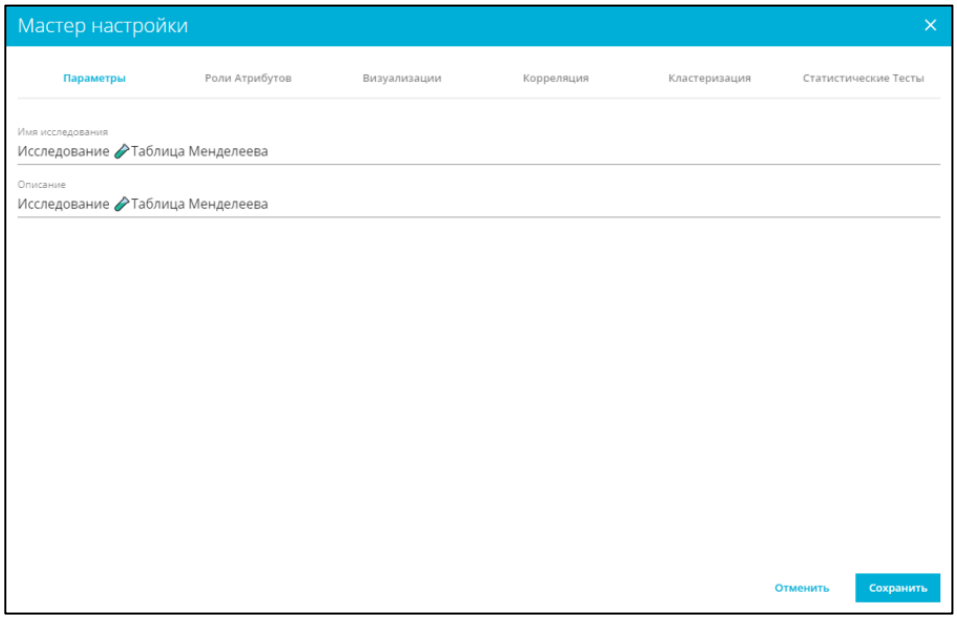


Рисунок 6 Вкладка «Параметры» Мастера настройки

2.2.3. Роли атрибутов

Во вкладке **Роли атрибутов** Пользователь может ознакомиться с параметрами атрибутов набора данных и изменить тип переменной.

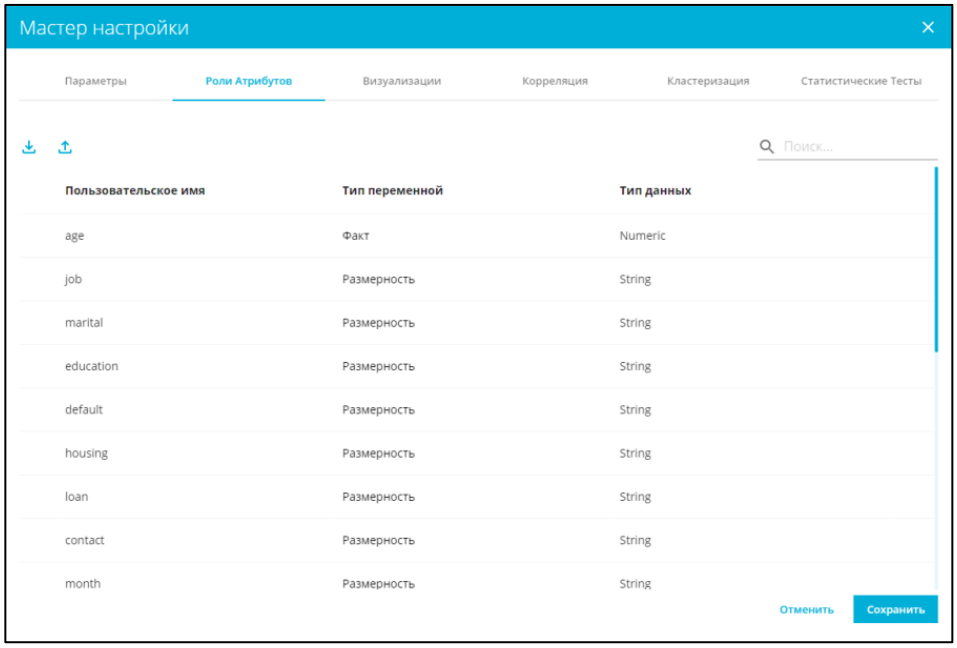


Рисунок 7 Вкладка «Роли атрибутов» Мастера настройки

Пользовательское имя задается при добавлении Набора данных в разделе Администрирования (подробнее **Руководство администратора**).

В компоненте предусмотрены следующие типы переменных и типы данных (задаются по умолчанию):

- **Факт**:
  - Numeric (включает и целые числа, и числа с плавающей точкой).
  - Date (дата).
  - DateTime (дата и время).

- **Размерность**:
  - String (строка).

Т.к. визуализации могут строится лишь с определенными типами переменных (подробнее Визуализации), предусмотрено изменение Типа переменной. Для этого необходимо щелкнуть левой кнопкой мыши по выбранному атрибуту и в открывшемся меню выбрать необходимый тип. Тип данных изменить нельзя.

Изменить Тип переменных также можно выгрузив Excel из системы, внести изменения, и загрузить обратно. Для этого нужно выбрать иконки **Экспорта в Excel** и **Импорта из Excel** .

2.2.4. Визуализации

Данная вкладка представляет собой список визуализаций, которые будут рассчитаны в ходе Исследования.

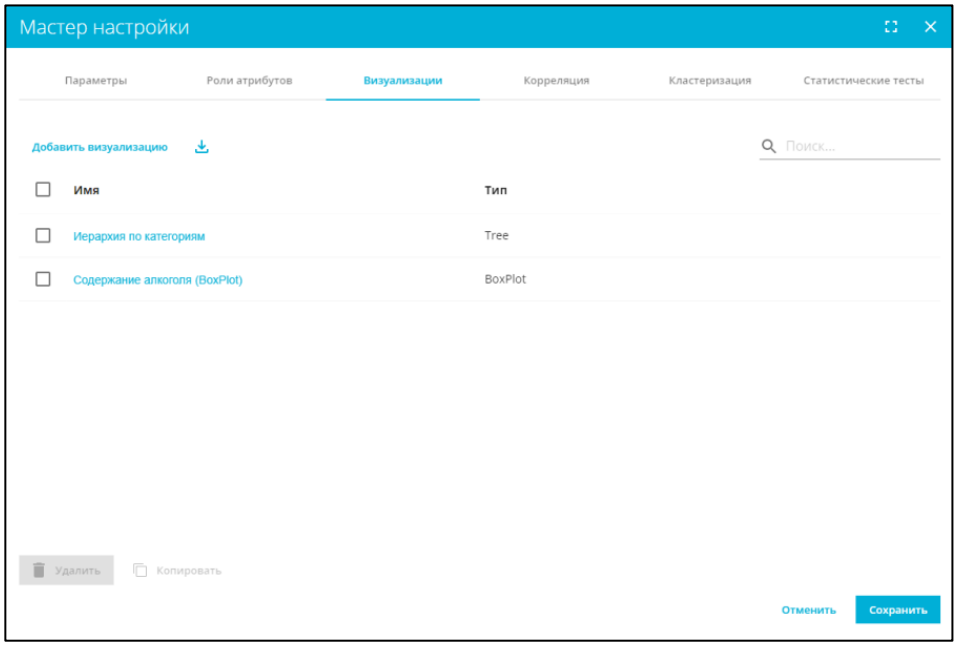


Рисунок 8 Вкладка «Визуализации» Мастера настройки

Для создания новой Визуализации необходимо выбрать **«Добавить визуализацию»**. В результате откроется окно **Настройки визуализации**.

В компоненте предусмотрены следующие типы визуализаций:

- Столбчатая диаграмма (Bar Chart).
- Диаграмма с накоплением (Stacked Bar Chart).
- Круговая диаграмма (Pie Chart).
- Кольцевая диаграмма (Donut Chart).
- Линейный график (Line Chart).
- Точечный график (Scatter Plot).
- Точечный 3D график (Scatter Plot 3D).
- Диаграмма размаха (BoxPlot).
- Древовидная диаграмма (Tree).
- Тепловая карта (HeatMap).
- Матрица сопряженности.

2.2.4.1. Столбчатая диаграмма (Bar Chart)

Столбчатая диаграмма отображает числовое сравнение между разными категориями. На одной оси (параметр **Ось X**) представлены конкретные сравниваемые категории, а на другой (параметр **Ось Y**) — шкала числовых значений. Диаграмма также позволяет отражать группировку по другой категории при помощи цвета (параметр **Группировка**).

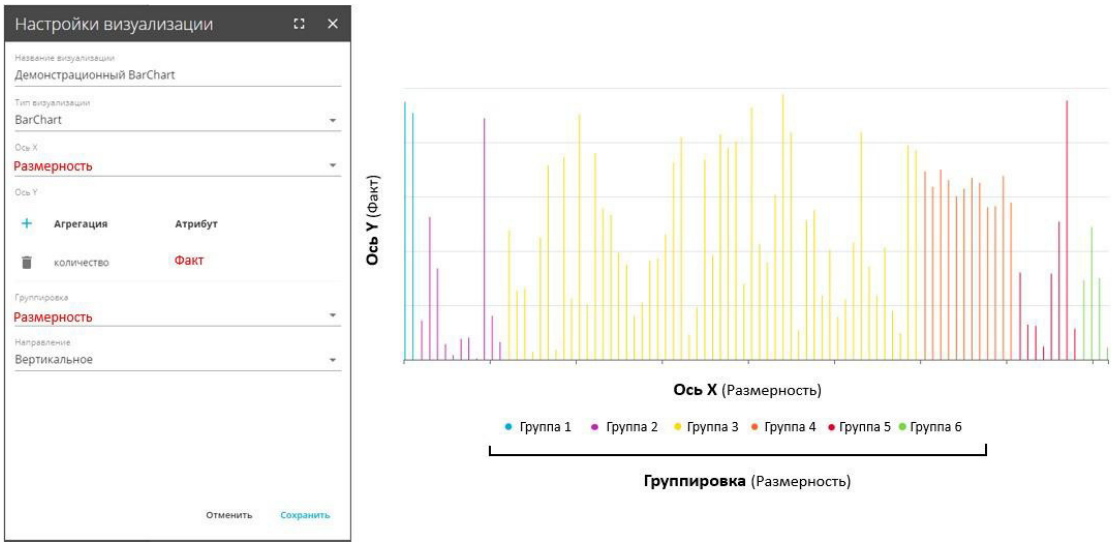


Рисунок 9 Столбчатая диаграмма с вертикальным направлением столбцов и группировкой

Столбцы могут быть расположены вертикально или горизонтально (параметр **Направление**).

Можно выбрать несколько атрибутов (параметр **Ось Y**), но тогда будет недоступна Группировка.

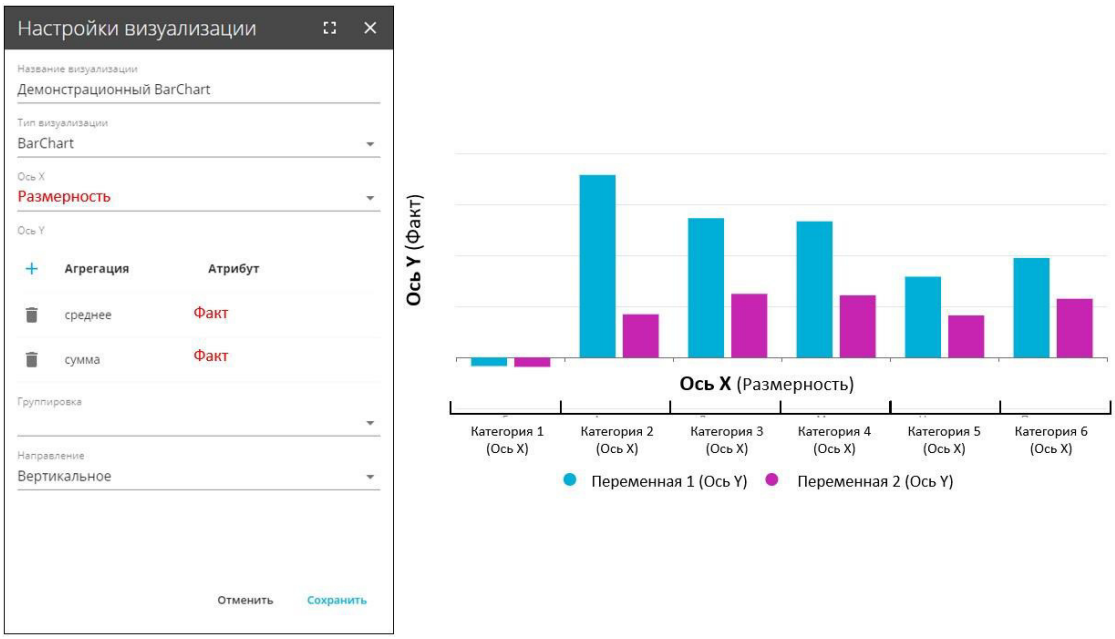


Рисунок 10 Столбчатая диаграмма с несколькими атрибутами

2.2.4.2. Диаграмма с накоплением (Stacked Bar Chart)

Диаграмма с накоплением показывает вклад нескольких элементов данных (параметр **Ось Y**) в суммирующий результат в виде столбцов, расположенных друг над другом. Высота каждого столбца пропорциональна значению соответствующего элемента данных. Сравнение происходит между разными категориями (параметр **Ось X**). Можно выбрать несколько атрибутов (параметр **Ось Y**), но тогда будет недоступна Группировка (параметр **Группировка**).

Столбцы могут быть расположены вертикально или горизонтально (параметр **Направление**).

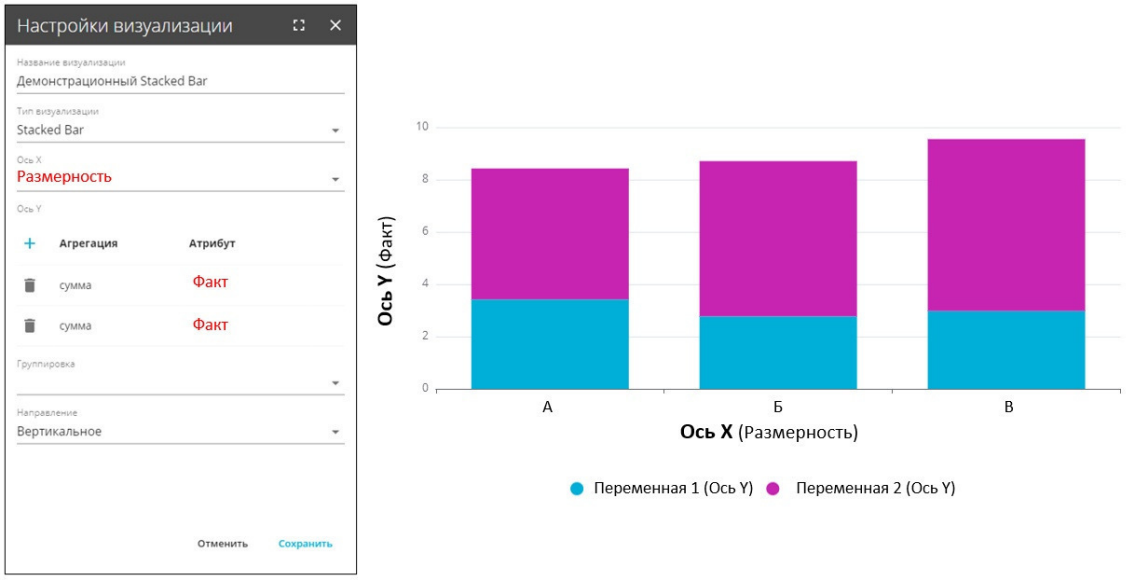


Рисунок 11 Диаграмма с накоплением и несколькими атрибутами Ось Y

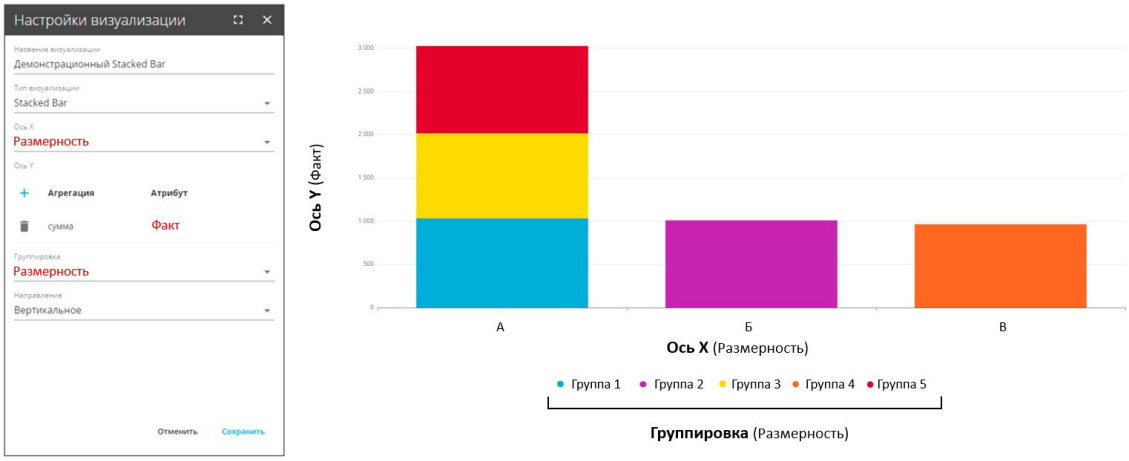


Рисунок 12 Диаграмма с накоплением и параметром Группировка

2.2.4.3. Гистограмма с группировкой (Grouped Histogram)

Гистограмма агрегирует числовые данные (**параметр Ось X**) по группам с равными интервалами, которые называют **бинами**, и отображает частоту встречаемости значений в каждом из бинов **(параметр Количество бинов)**.

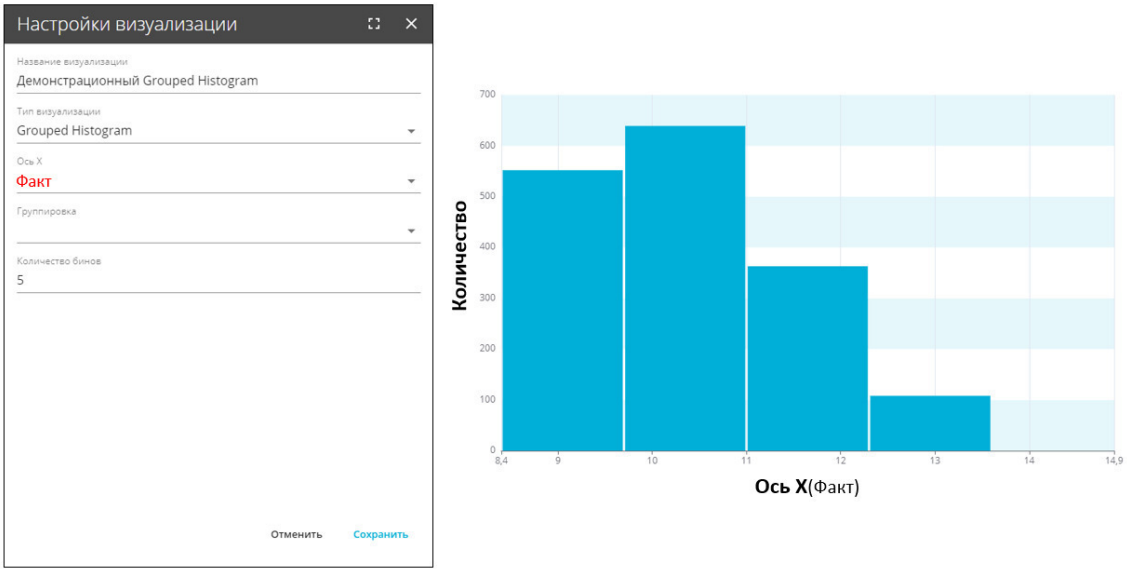


Рисунок 13 Гистограмма

Для построения гистограммы с группировкой необходимо задать атрибут для Группировки (**параметр Группировка**).

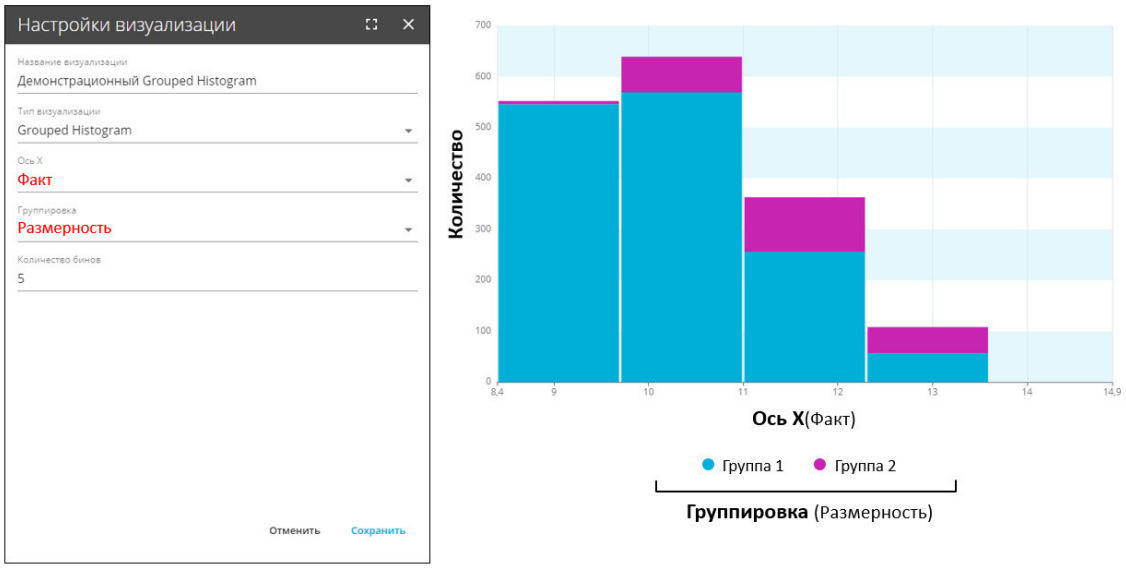


Рисунок 14 Сгруппированная гистограмма

2.2.4.4. Круговая диаграмма (Pie Chart)

Круговая диаграмма показывает процентное соотношение между категориями (параметр **Категория**). Длина каждой дуги представляет собой пропорциональную длину категории (параметр **Значение**) от длины всей окружности, равной 100%.

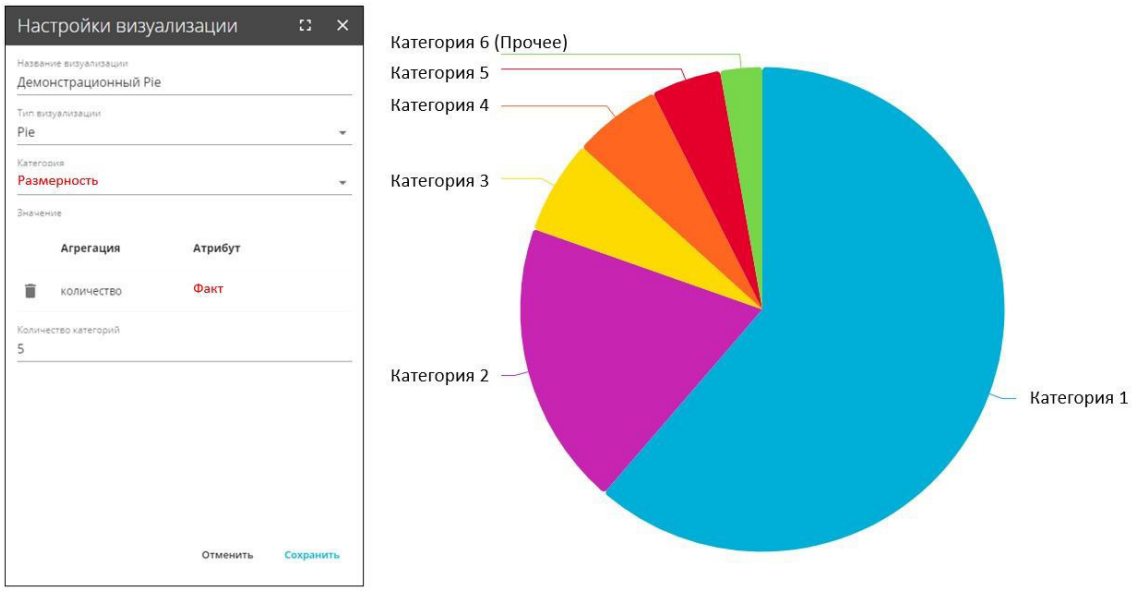


Рисунок 15 Круговая диаграмма

Если данные содержат множество категорий, пользователь может сократить количество отображаемых на диаграмме категорий для удобства интерпретации (параметр **Количество категорий**). Не вошедшие категории (согласно сортировки по убыванию) объединяются в единую категорию.

2.2.4.5. Кольцевая диаграмма (Donut Chart)

Кольцевая диаграмма представляет собой круговую диаграмму, но с вырезанной центральной частью. Это позволяет пользователю сосредоточить внимание на длине дуг, а не на сопоставлении пропорций сегментов.

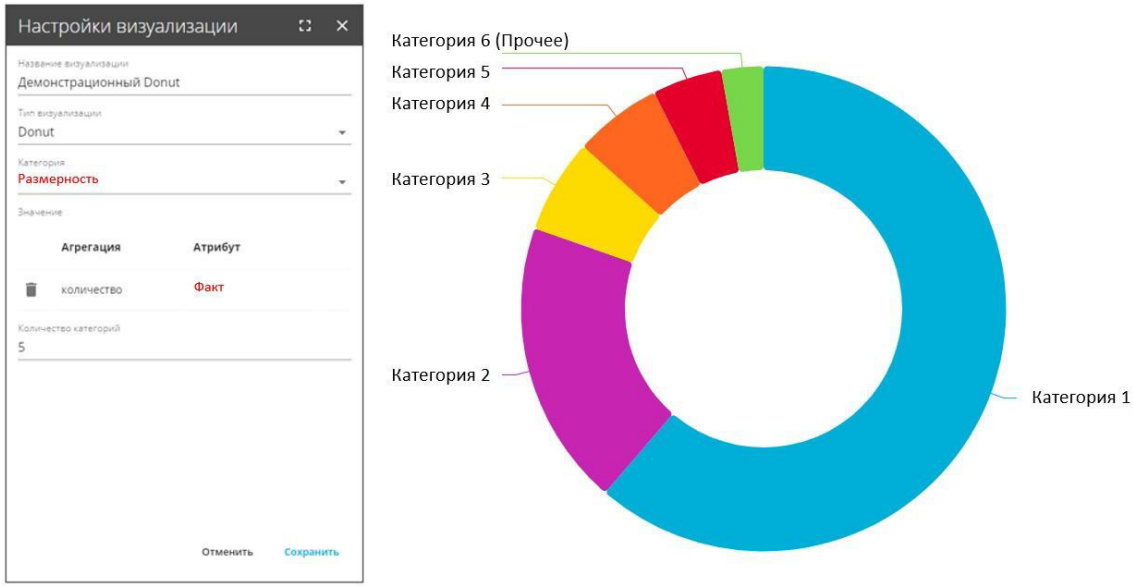


Рисунок 16 Кольцевая диаграмма

Аналогично круговой диаграмме пользователь может сократить количество отображаемых на диаграмме категорий (параметр **Количество категорий**). Не вошедшие категории (согласно сортировки по убыванию) объединяются в единую категорию.

2.2.4.6. Линейный график (Line Chart)

Линейный график отображает изменение показателей (параметр **Ось Y**) за некоторый интервал (параметр **Ось X**). Как правило, на оси Y отмечаются количественные значения, а на оси X шкала последовательностей.

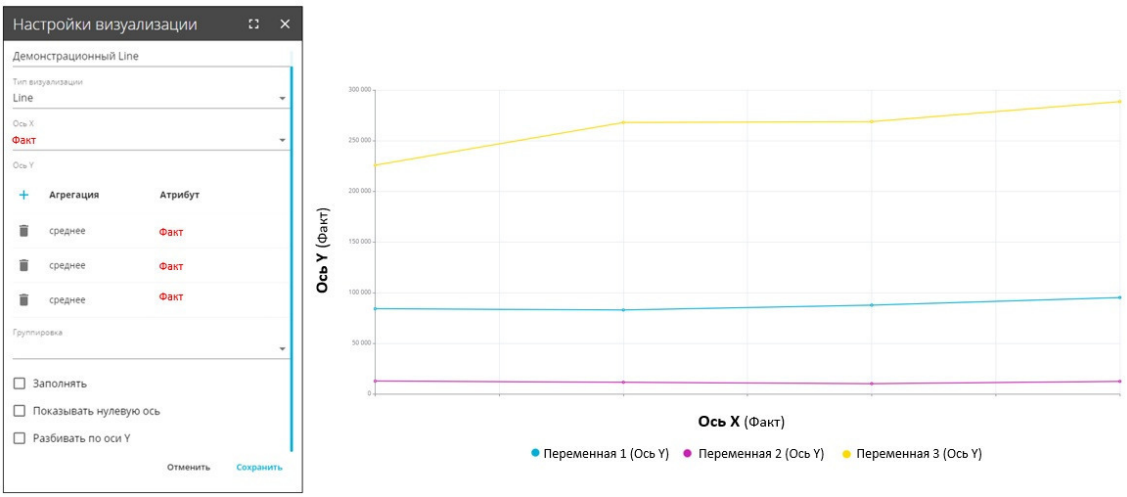


Рисунок 17 Линейный график

Также предусмотрена возможность группировки наблюдений по цвету (параметр **Группировка**).

Выбрав чекбокс **Заполнять**, пользователь может построить накопительную диаграмму, в которой область ниже линии заполнена определённым цветом.

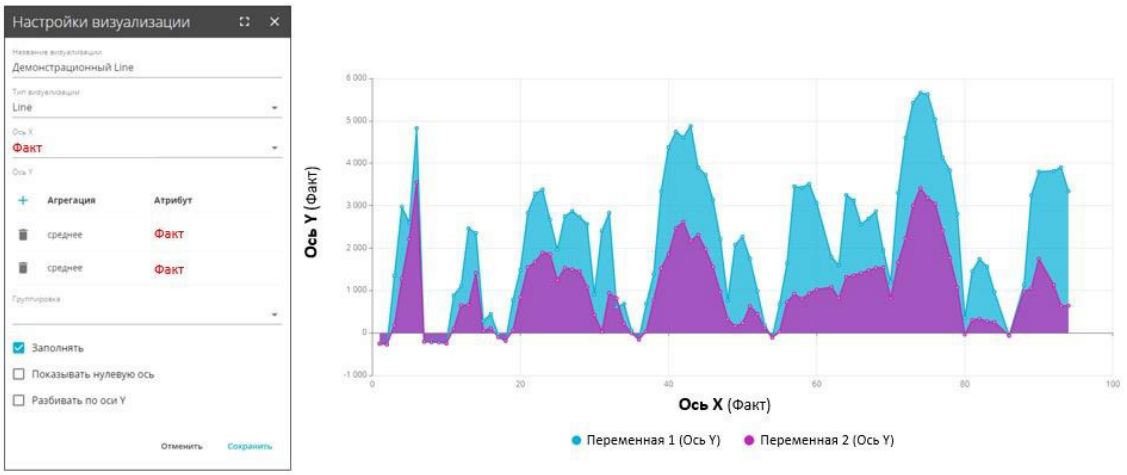


Рисунок 18 Линейный график с накоплением

Многоосевая диаграмма позволяет сравнить несколько показателей с различными диапазонами значений и использует для этого две или более оси Y и одну общую ось X (чекбокс **Разбивать по оси Y**).

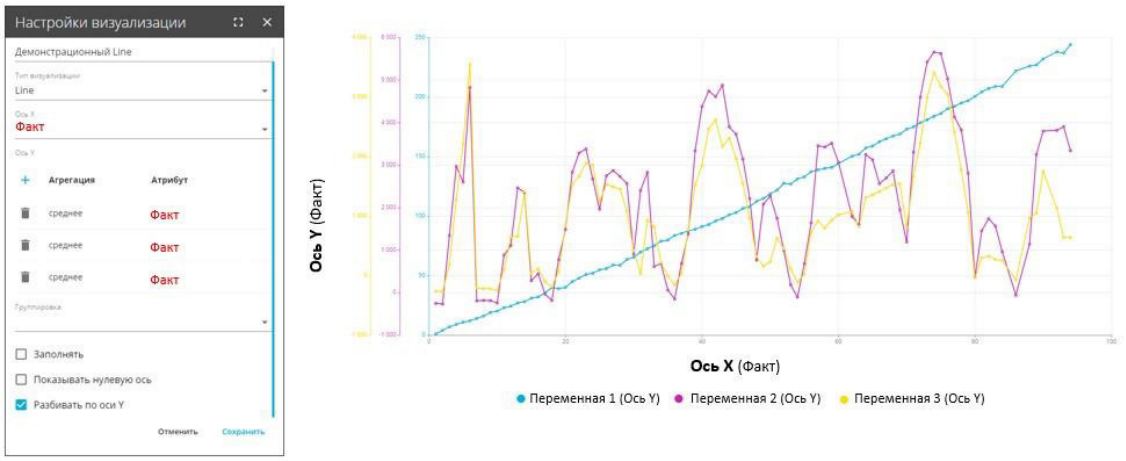


Рисунок 19 Многоосевой линейный график

2.2.4.7. Точечный график (Scatter Plot)

Точечный график активно используется для:

- Оценки корреляции переменных;
- Обнаружения выбросов и ошибок в данных;
- Оценки скученности наблюдений.

Для построения графика пользователю необходимо задать две количественные переменные (параметры **Ось X** и **Ось Y**).

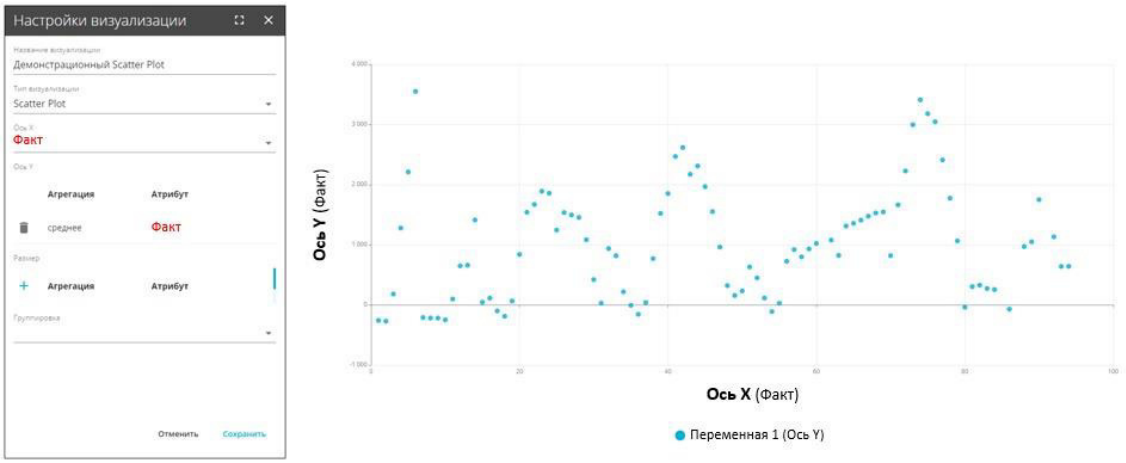


Рисунок 20 Точечный график

При необходимости можно масштабировать наблюдения согласно значению третьей количественной переменной (параметр **Размер**).

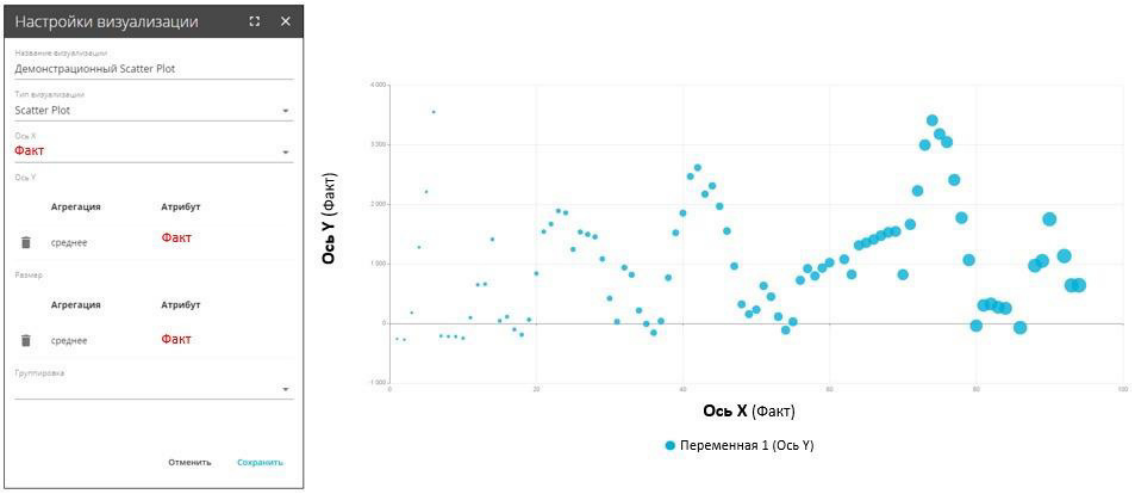


Рисунок 21 Точечный график с заданным параметром Размер

Пользователь также может присвоить цвет наблюдениям согласно значению некоторой категориальной переменной (параметр **Группировка**).

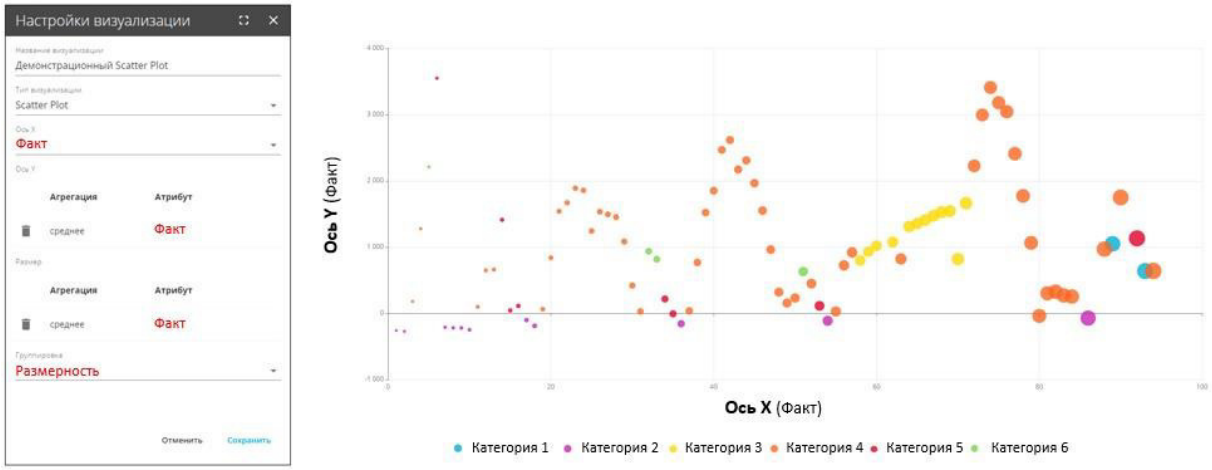


Рисунок 22 Точечный график с заданными параметрами Размер и Группировка

2.2.4.8. Точечный 3D график (Scatter Plot 3D)

Точечный 3D график отображает отношение трех количественных переменных (параметры **Ось X**, **Ось Y** и **Ось Z**) в трехмерном пространстве, позволяет масштабировать наблюдения согласно четвертой количественной переменной (параметр **Размер**) и цветовым градиентом выделить различия по пятой количественной переменной (параметр **Цвет**).

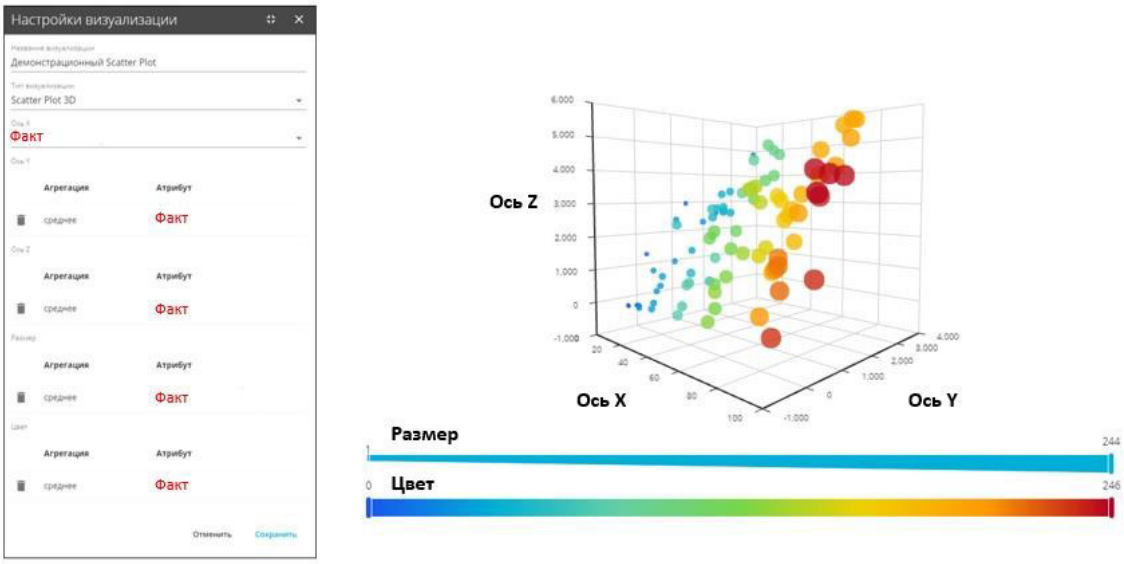


Рисунок 23 Точечный 3D график с заданными параметрами Размер и Цвет

2.2.4.9. Диаграмма размаха (Boxplot)

Диаграмма размаха («ящик с усами») — удобный способ визуального представления непрерывных переменных (параметр **Переменная**), который позволяет:

- Отобразить значение ключевых статистик: медиана, верхний/нижний квартили, максимальное и минимальное значение выборки.
- Выявить выбросы.
- Определить степень разброса (дисперсии) и асимметрии данных.

Несколько таких диаграмм можно нарисовать бок о бок, чтобы визуально сравнивать распределение переменной в разных группах (параметр **Группировка**). Диаграммы можно располагать как горизонтально, так и вертикально (параметр **Направление**).

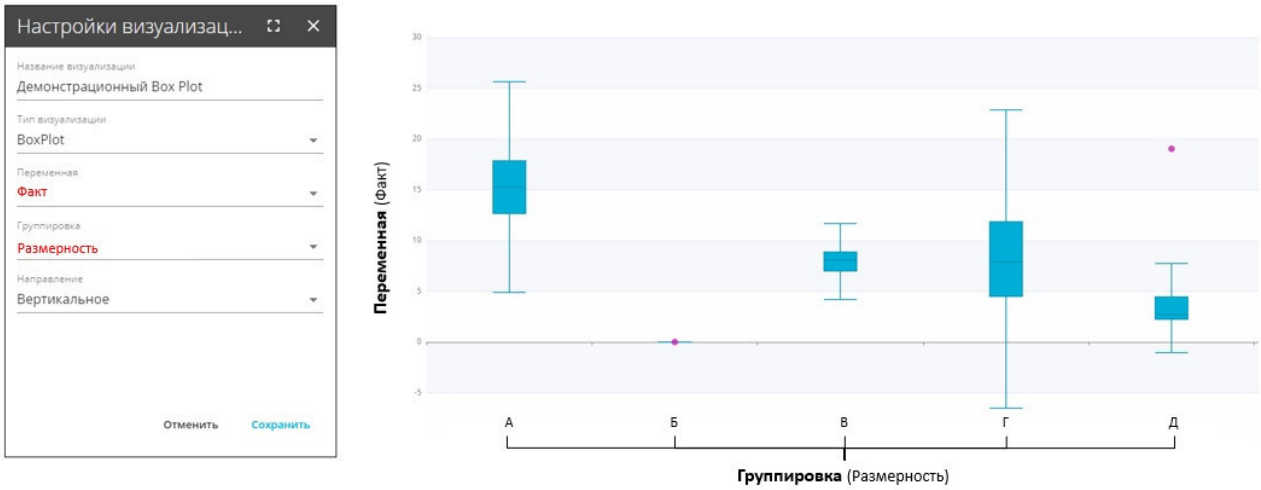


Рисунок 24 Диаграмма размаха

При наведении на ящик появляется панель со следующими статистиками:

- Min — минимальное значение;
- Q1 — нижний квартиль;
- Median — медианное значение;
- Q3 — верхний квартиль;
- Max — максимальное значение.

Помимо самих ящиков на диаграмме размаха отображаются выпадающие значения в виде фиолетовых точек (за пределами полутора межквартильных интервалов).

2.2.4.10. Древовидная диаграмма (Tree)

Древовидная диаграмма — это метод визуального представления иерархии в древовидной структуре.

Структура древовидной диаграммы состоит из следующих элементов:

- Корневой узел — элемент, не имеющий вышестоящего элемента,
- Узлы и соединяющие их линии — ветви, которые обозначают взаимосвязи и отношения между элементами,
- Листовые узлы — элементы, у которых нет дочерних элементов.

Для создания данной диаграммы необходимо последовательно указать категориальные переменные (параметр **Категории**) и количественную переменную, которая будет отражаться в каждом узле (параметр **Значение**).

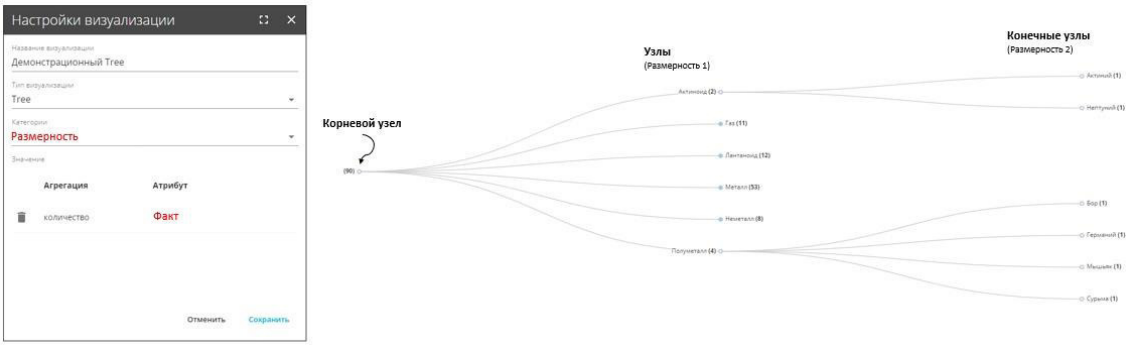
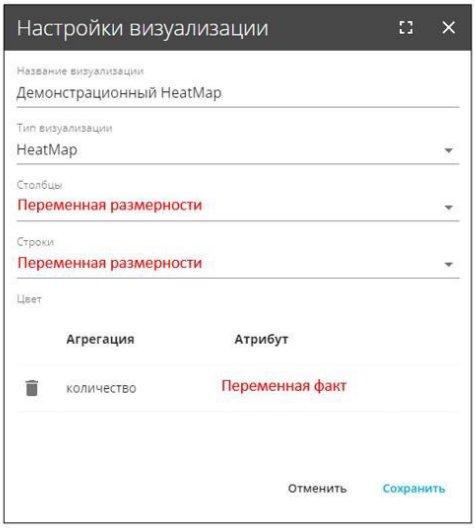


Рисунок 25 Древовидная диаграмма

2.2.4.11. Тепловая карта (Heat Map)

Тепловые карты позволяют анализировать многомерные данные за счет распределения переменных по рядам и столбцам и закрашивания цветом ячеек таблицы.

Ряды отображают одну категорию (параметр **Строки**), а все столбцы — другую (параметры **Столбцы**). Отдельные ряды и столбцы делятся на подкатегории, пересекающиеся друг с другом в рамках матрицы. В ячейках таблицы содержатся отображаемые в цвете количественные данные (параметр **Цвет**). Данные каждой ячейки основаны на взаимосвязи двух переменных: в строке и в столбце.



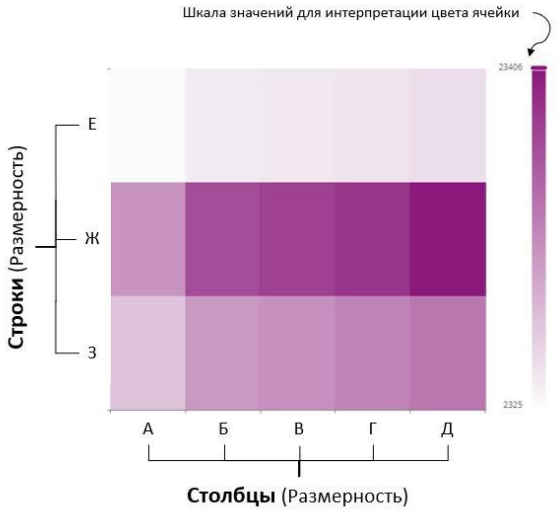
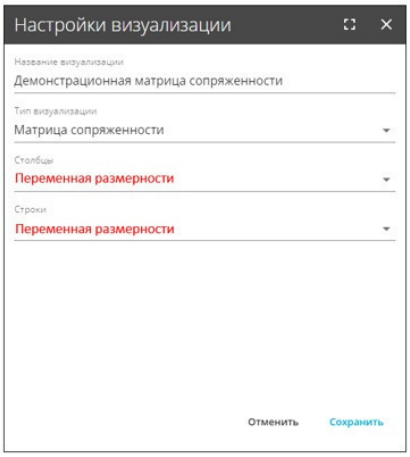
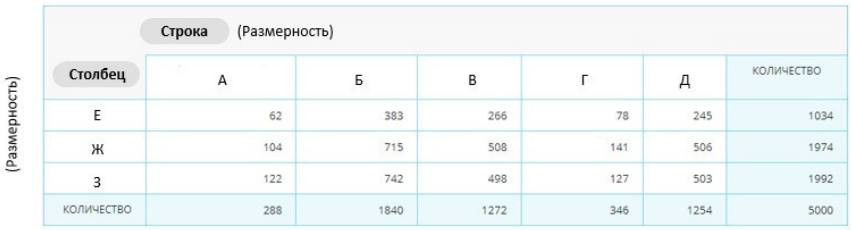


Рисунок 26 Тепловая карта

2.2.4.12. Матрица сопряженности (Pivot)

Таблица сопряженности служит для описания связи двух номинальных переменных, на пересечение столбов (параметр **Столбцы**) и строк (параметр **Строки**) которой указывается частота совместного появления соответствующих значений двух признаков.





(Размерность)

Строка (Размерность) / Столбец	А	Б	В	Г	Д	КОЛИЧЕСТВО
Е	62	383	266	78	245	1034
Ж	104	715	508	141	506	1974
З	122	742	498	127	503	1992
КОЛИЧЕСТВО	288	1840	1272	346	1254	5000

Рисунок 27 Матрица сопряженности

2.2.4.13. Диаграмма-радар (Radar)

Диаграмма-радар представляет собой линейные графики, в которых значения **Оси X** обернуты на 360 градусов и для каждого значения x имеются значения **Оси Y**. Для корректного отображения графика необходимо задать не менее 3 агрегаций по **Оси Y**.

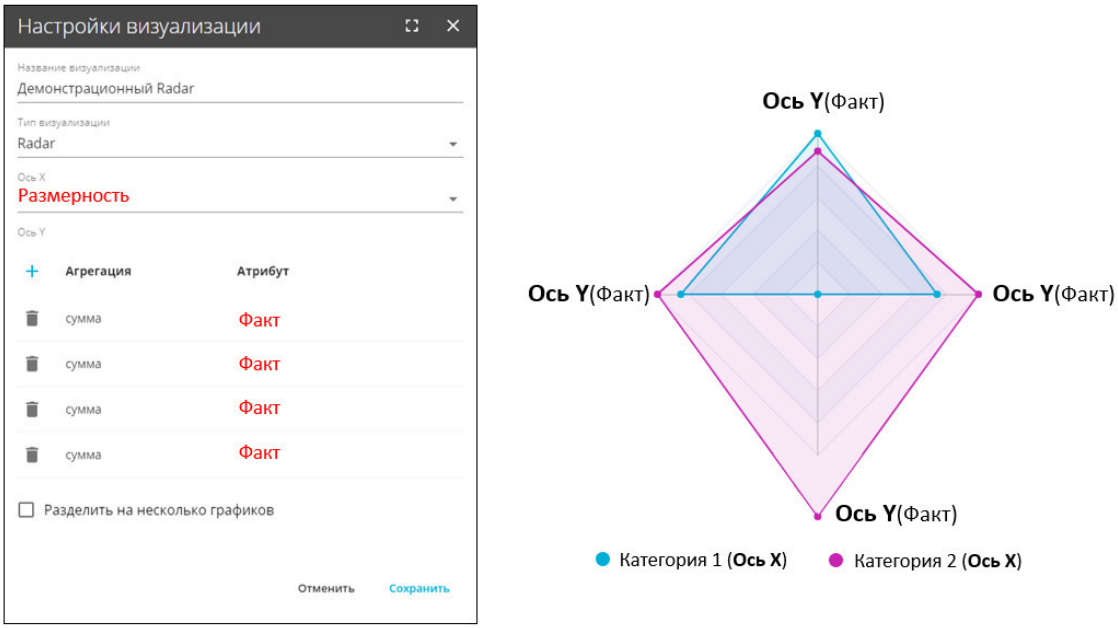


Рисунок 28 Диаграмма-радар

При выборе чекбокса **Разделить на несколько графиков** будут построены графики, количеством совпадающие с числом уникальных значений переменной **Оси X**.

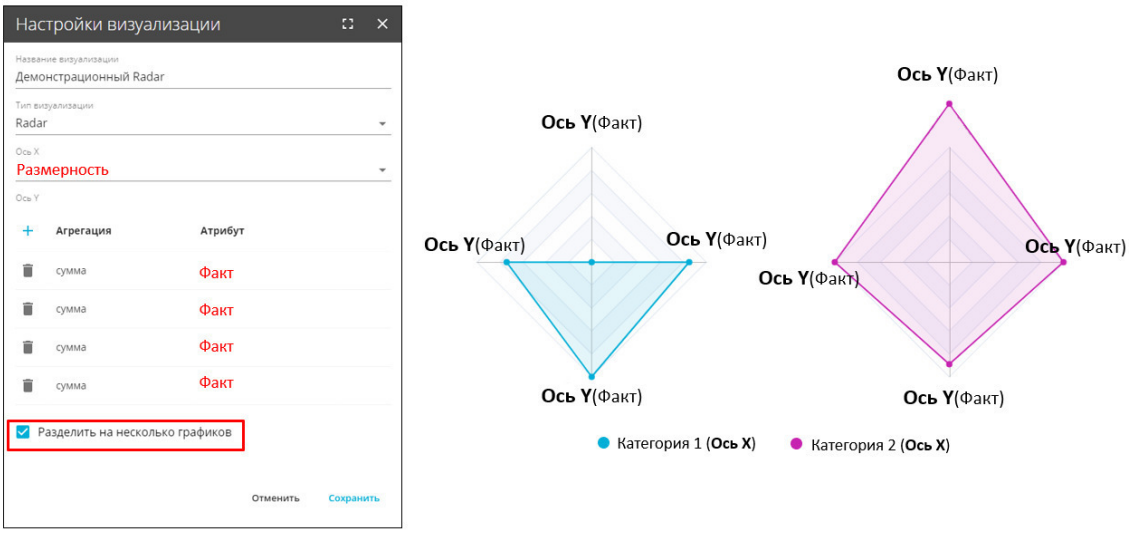


Рисунок 29 Диаграмма-радар с разделением на 2 графика

2.2.5. Корреляция

Данная вкладка представляет собой список корреляционных матриц, которые будут рассчитаны в ходе Исследования.

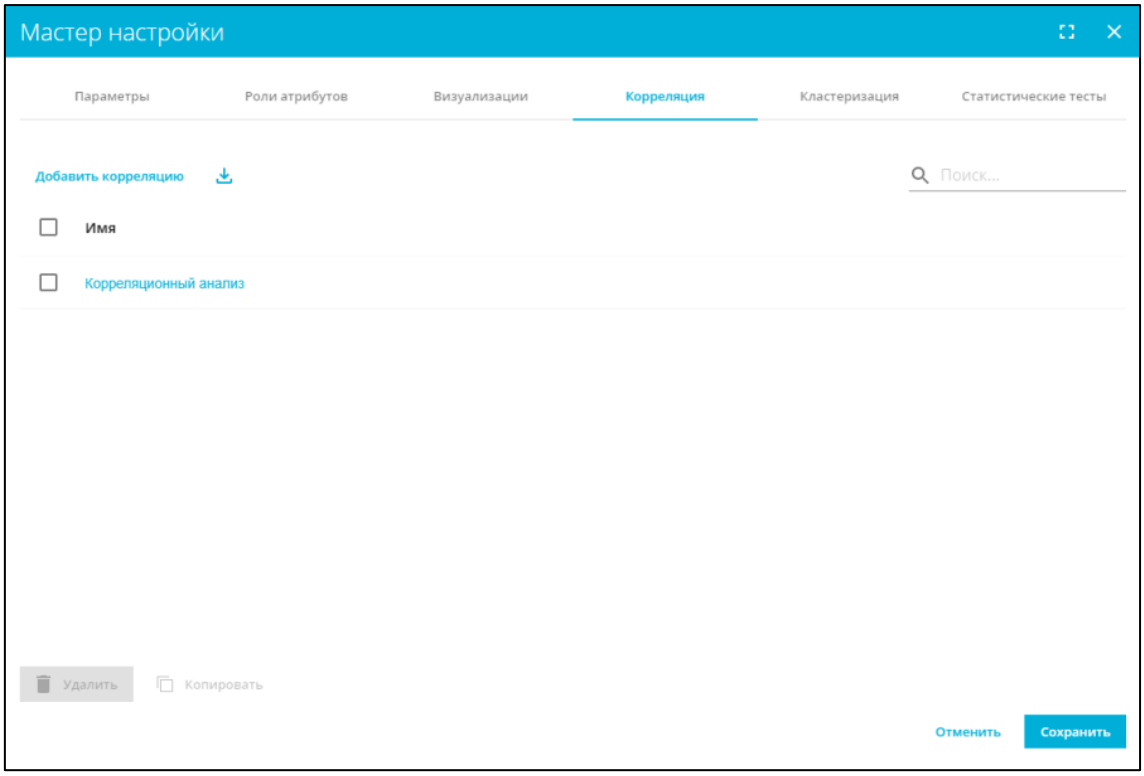


Рисунок 30 Вкладка «Корреляция» Мастера настройки

Корреляционная матрица — это квадратная таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции для соответствующей пары признаков. **Коэффициент корреляции** — статистическая мера, которая отражает силу связи между признаками.

Для создания новой корреляционной матрицы необходимо:

- Выбрать **«Добавить корреляцию».**
- В открывшемся окне **Настройки корреляции** задать Название корреляции.
- Добавить необходимые для анализа признаки, выбрав соответствующие чекбоксы рядом, и нажать на стрелку переноса в правое поле (для добавления всех имеющихся атрибутов выбрать чекбокс рядом с **Добавить в корреляцию** в левом поле.
- Сохранить параметры.

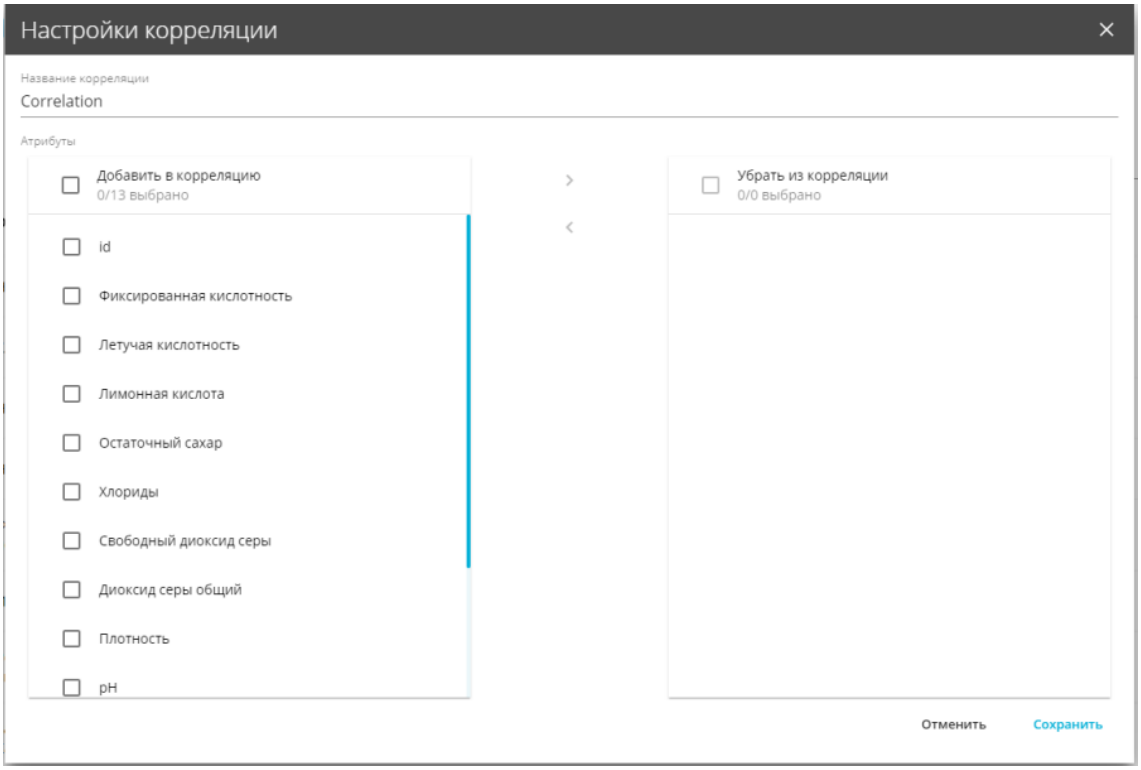


Рисунок 31 Окно Настройки корреляции

2.2.6. Кластеризация

Данная вкладка представляет собой список с процедурами кластерного анализа, которые будут рассчитаны в ходе Исследования.

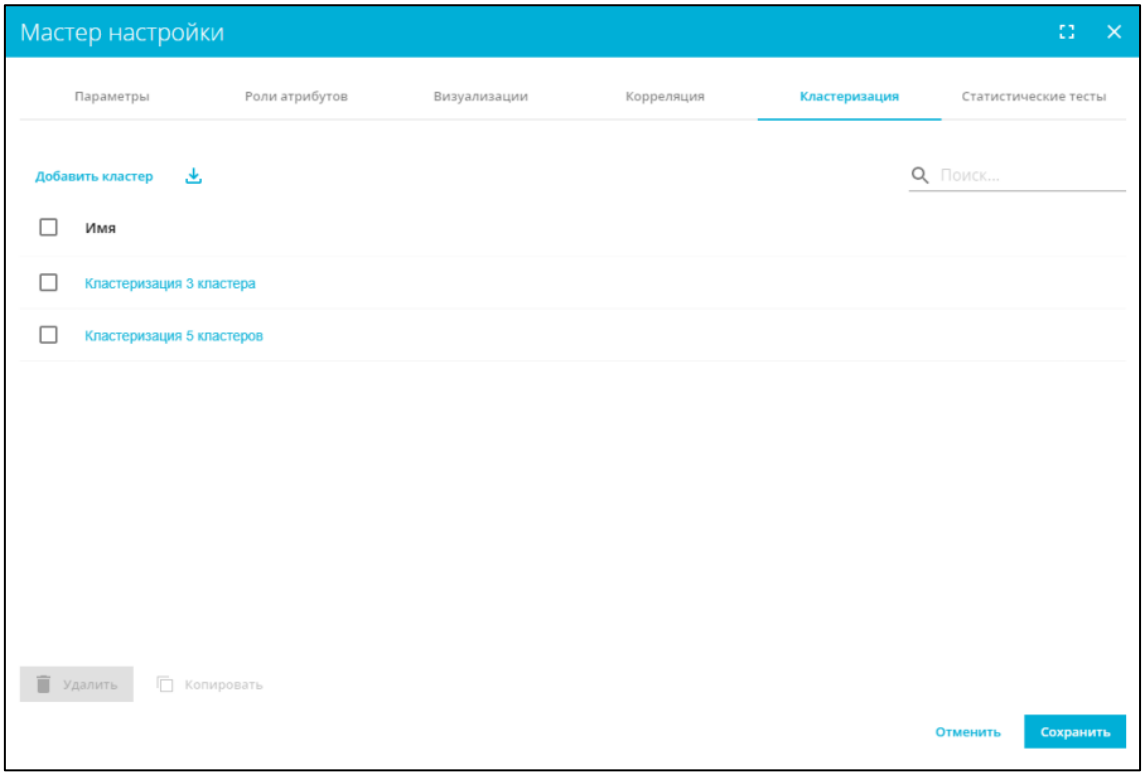


Рисунок 32 Вкладка «Кластеризация» Мастера настройки

Кластеризация — это разбиение множества объектов на подмножества (кластеры) по заданному критерию.

На этапе исследования полезно получить априорную информацию об исходных данных, чтобы на этапе построения модели k-средних указать параметр, который будет увеличивать ее предсказательную способность.

Для создания новой процедуры кластерного анализа необходимо:

- Выбрать **«Добавить кластер».**
- В открывшемся окне **Настройки кластеризации** задать параметры (подробнее в таблице ниже).
- Сохранить параметры.

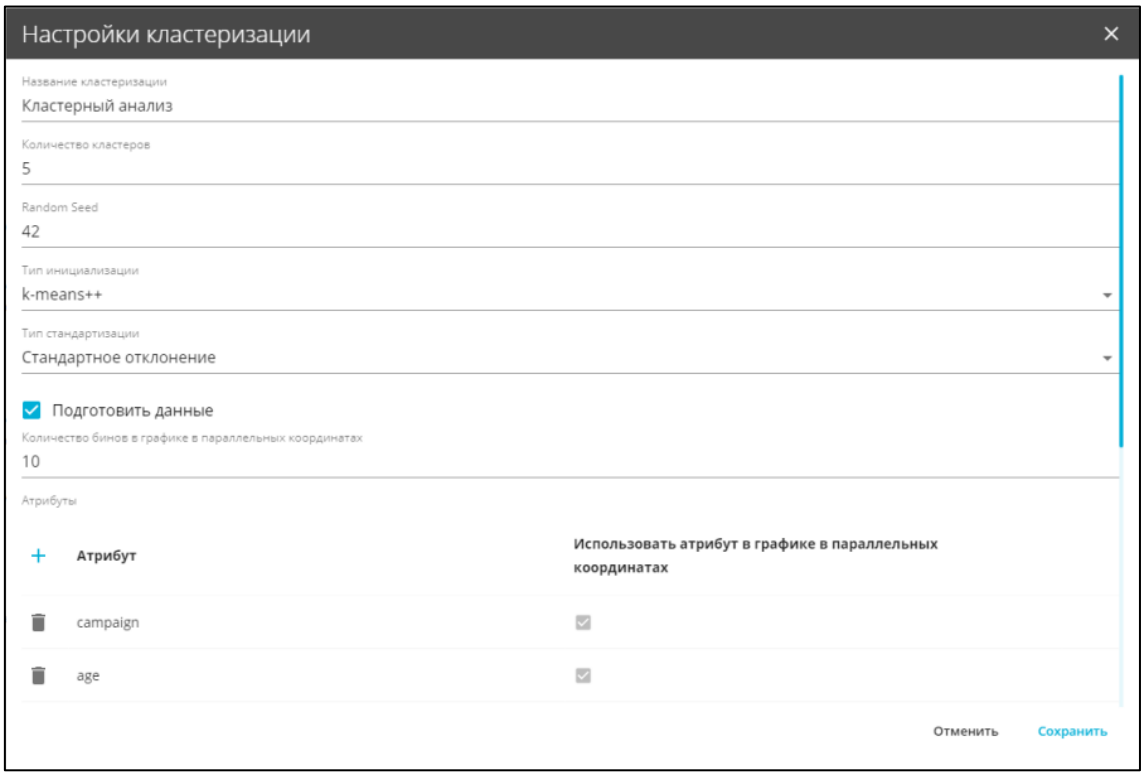


Рисунок 33 Окно Настройки кластеризации

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название, которое будет отображаться в результатах исследования
Количество кластеров	Ручной ввод целочисленного значения больше 0 По умолчанию — 5	Задание числа кластеров, на которые будет делиться векторное пространство.
Random Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел

Параметр	Возможные значения и ограничения	Описание
Тип инициализации	Раскрывающийся список со следующими значениями: • k-means++ (По умолчанию) • Forgy • Random	Данный параметр отвечает за выбор метода инициализации начальных точек кластеров. Предусмотрены: • **k-means++** Идея метода k-means++ состоит в том, чтобы выбрать начальные точки, которые находятся как можно дальше друг от друга. • **Forgy** Метод Forgy случайным образом выбирает k наблюдений (по числу заданных кластеров) из набора данных и использует их в качестве начальных значений. • **Random** Метод Random сначала случайным образом назначает кластер каждому наблюдению, а затем переходит к этапу обновления, таким образом вычисляя начальное среднее значение как центроид случайно назначенных точек кластера.
Тип стандартизации	Раскрывающийся список со следующими значениями: • Не выбрано • Стандартное отклонение (По умолчанию) • Нормализация	Данный параметр задает тип стандартизации данных. Предусмотрены: • **Не выбрано** Не стандартизировать • **Стандартное отклонение** Из каждой записи вычитается среднее значение и результат делится на стандартное отклонение • **Нормализация** Из каждой записи вычитается минимальное значение и результат делится на разницу между максимальным и минимальным значением
Подготовить данные	Чекбокс	Выбор данного чекбокса указывает на необходимость: • Заменить пропущенные значения количественной переменной на mean • Заменить пропущенные значения категориальной переменной на текстовый None
Количество бинов в графике в параллельных координатах	Ручной ввод целочисленного значения Больше 0 По умолчанию — 10	Данный параметр задает количество бинов, на которое делятся наблюдения для отображения на графике в параллельных координатах
Атрибуты	Список атрибутов, доступных в наборе данных	Выбор атрибутов набора данных для проведения кластерного анализа

Таблица 2 Параметры кластеризации

2.2.7. Статистические тесты

Данная вкладка представляет собой список статистических тестов, которые будут рассчитаны в ходе Исследования.

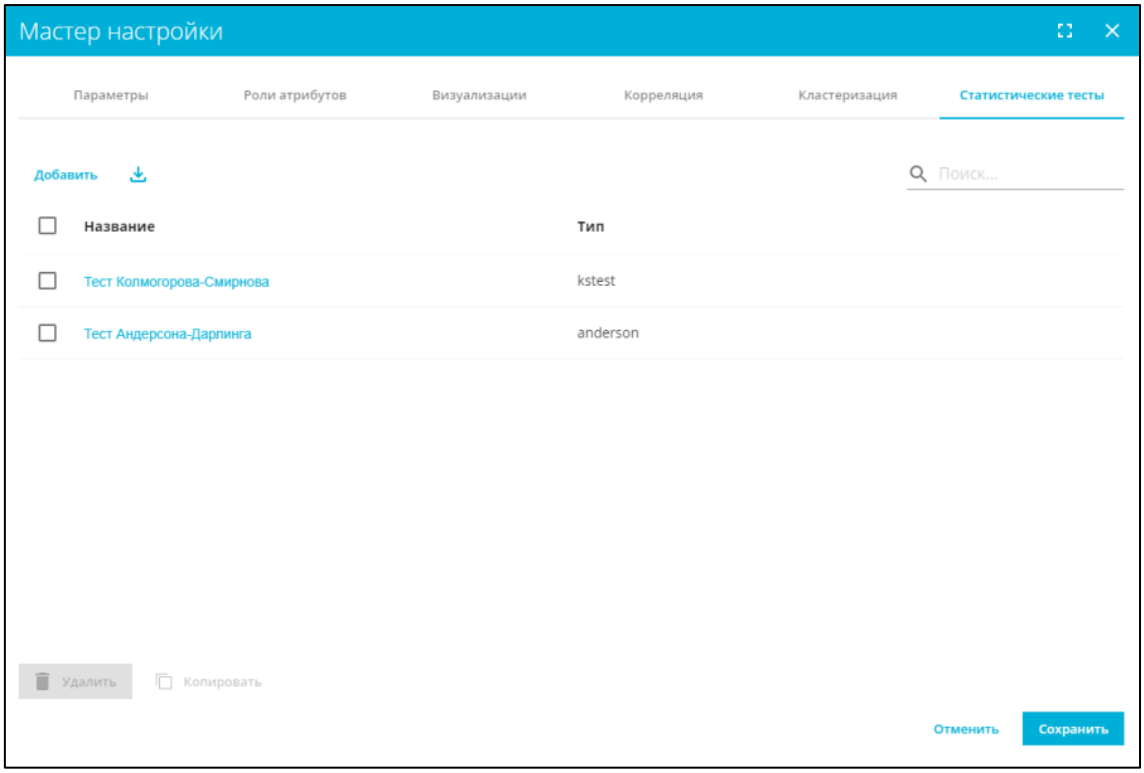


Рисунок 34 Вкладка «Статистические тесты» Мастера настройки

Статистические тесты используются для тестирования статистических гипотез о виде распределения наблюдений:

- Нулевая гипотеза: выборка принадлежит некоторому закону распределения;
- Альтернативная гипотеза: нулевая гипотеза не верна.

В компоненте **Исследование данных** предусмотрены следующие статистические тесты.

Для непрерывных наблюдений:

- Тест Колмогорова-Смирнова.
- Тест Крамера-фон-Мизеса.
- Тест Андерсона-Дарлинга.

Для дискретных наблюдений:

- Критерий согласия хи-квадрат Пирсона.

Для расчета статистического теста необходимо:

- Выбрать **«Добавить».**
- В открывшемся окне **Новый статистический тест** задать название и тип теста (подробнее в таблице ниже), а также параметры, специфичные для каждого из статистических тестов.

- Сохранить параметры.

Рисунок 35 Окно Новый статистический тест

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название, которое будет отображаться в результатах исследования
Тип	Раскрывающийся список со следующими значениями: • Непрерывные наблюдения. Тест Колмогорова-Смирнова • Непрерывные наблюдения. Тест Крамера-фон-Мизеса • Непрерывные наблюдения. Тест Андерсона-Дарлинга • Дискретные наблюдения. Критерий согласия хи-квадрат Пирсона	Данный параметр задает тип рассчитываемого статистического теста

Таблица 3 Параметры статистического теста

В окне **Настройки статистического теста** отображается различный набор параметров, зависящий от типа статистического теста.

Для расчета **Теста Колмогорова-Смирнова** нужно указать следующие параметры:

- Переменная, функция распределения F(x) которой будет проверяться на принадлежность некоторому закону распределения.
- cdf — теоретическое распределение G(x).
- Alternative — варианты нулевой и соответствующей альтернативной гипотезы:

  - two-sided (двусторонняя) — нулевая гипотеза состоит в том, что два распределения идентичны, F(x) = G(x) для всех x; альтернативная — они не идентичны;
  - less (левосторонняя) — Нулевая гипотеза состоит в том, что F(x)> = G(x) для всех x; альтернативная — F(x) <G(x) хотя бы для одного x;
  - greater (правосторонняя) — Нулевая гипотеза состоит в том, что F(x) <= G(x) для всех x; альтернативная — F(x)> G(x) хотя бы для одного x.

- Режим — определяет распределение, используемое для расчета p-значения:
  - auto — автоматический подбор;
  - exact — использует распределение тестовой статистики;
  - approx — приближает двустороннюю вероятность с удвоенной односторонней вероятностью;
  - asymp — использует асимптотическое распределение тестовой статистики.

Для расчета **Теста Крамера-фон-Мизеса** необходимо указать следующие параметры:

- Переменная, функция распределения которой будет проверяться на принадлежность некоторому закону распределения.
- cdf — теоретическое распределение.

Для расчета **Теста Андерсона-Дарлинга** нужно указать следующие параметры:

- Переменная, функция распределения которой будет проверяться на принадлежность некоторому закону распределения.
- Dist — теоретическое распределение.

Для расчета **Критерия согласия хи-квадрат Пирсона** нужно указать следующие параметры:

- Категориальная переменная, ожидаемые частоты которой предполагаются однородными и определяются как среднее значение наблюдаемых частот.
- ddof — поправка на число степеней свободы.

2.3. Расписание

Данная функция позволяет сформировать список расписаний, согласно которым будет запускаться расчет Исследований.

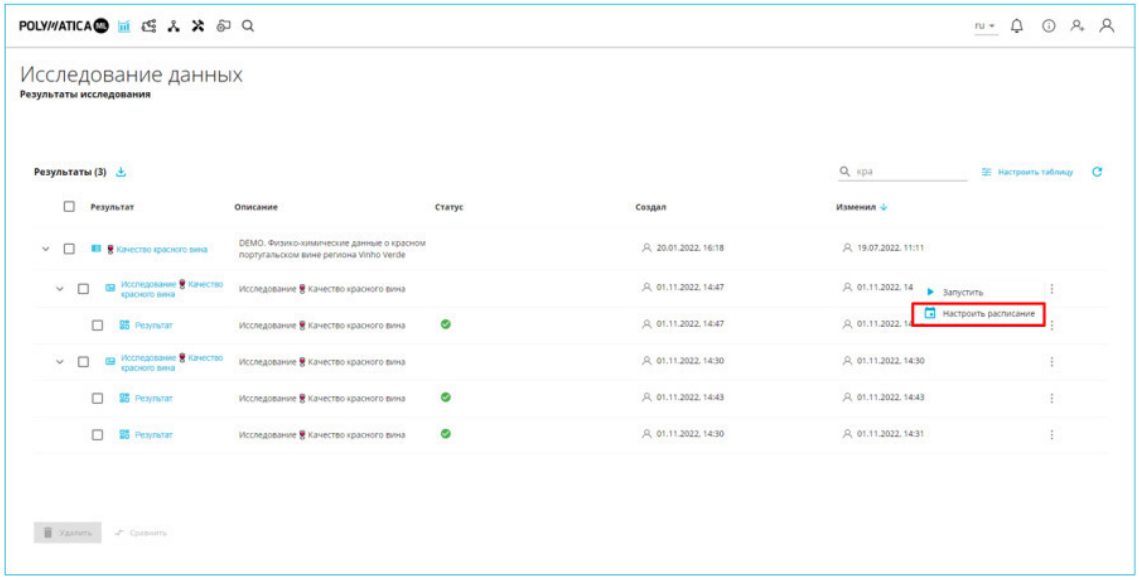


Рисунок 36 Функция настройки расписания запуска Исследований

Для постановки Исследования на расписание необходимо:

- Нажать кнопку **«Настроить расписание»**.
- Выбрать **Добавить Правило**.
- В открывшемся окне **Настройка расписания** задать:
    - Название, которое далее будет отображаться в системе.
    - Выбрать дату и время первого запуска.
    - Настроить повторение (период — ежедневно, еженедельно, ежемесячно и время повторного запуска).
    - Окончание расписания (Всегда запускать/Число выполнений (указать число запусков)/Дата окончания (указать последнюю дату запуска)).
    - Сохранить настройки.

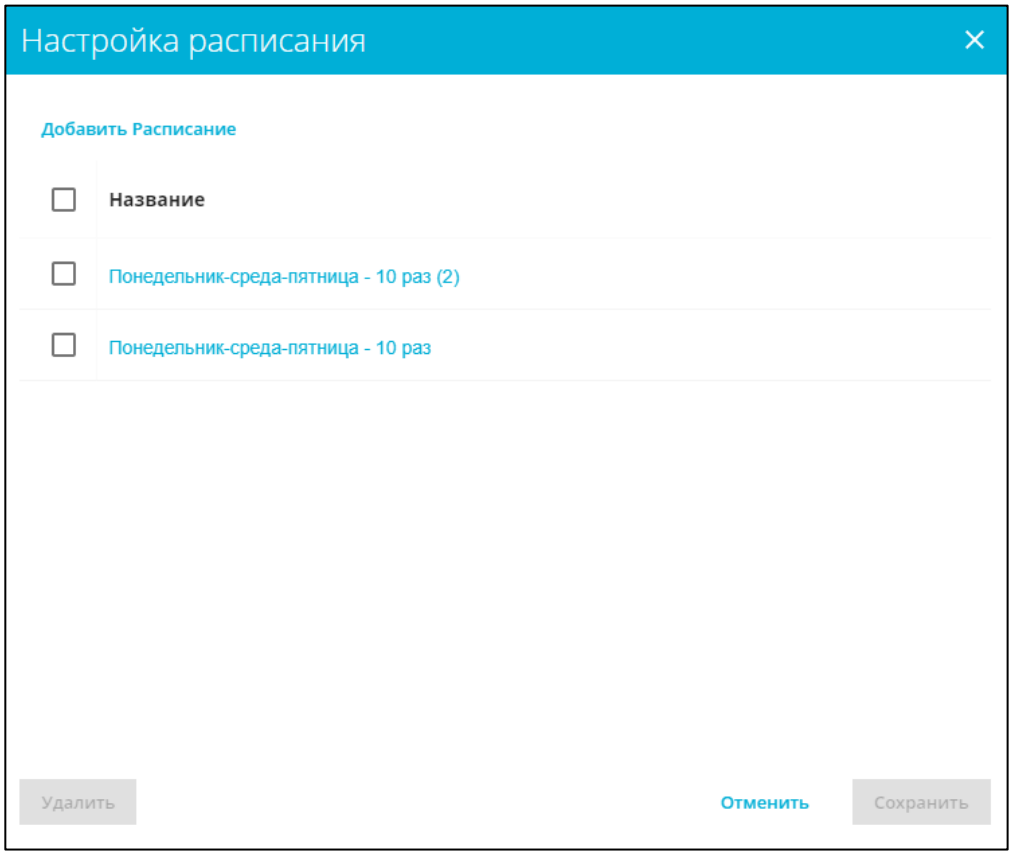


Рисунок 37 Окно Настройка расписания

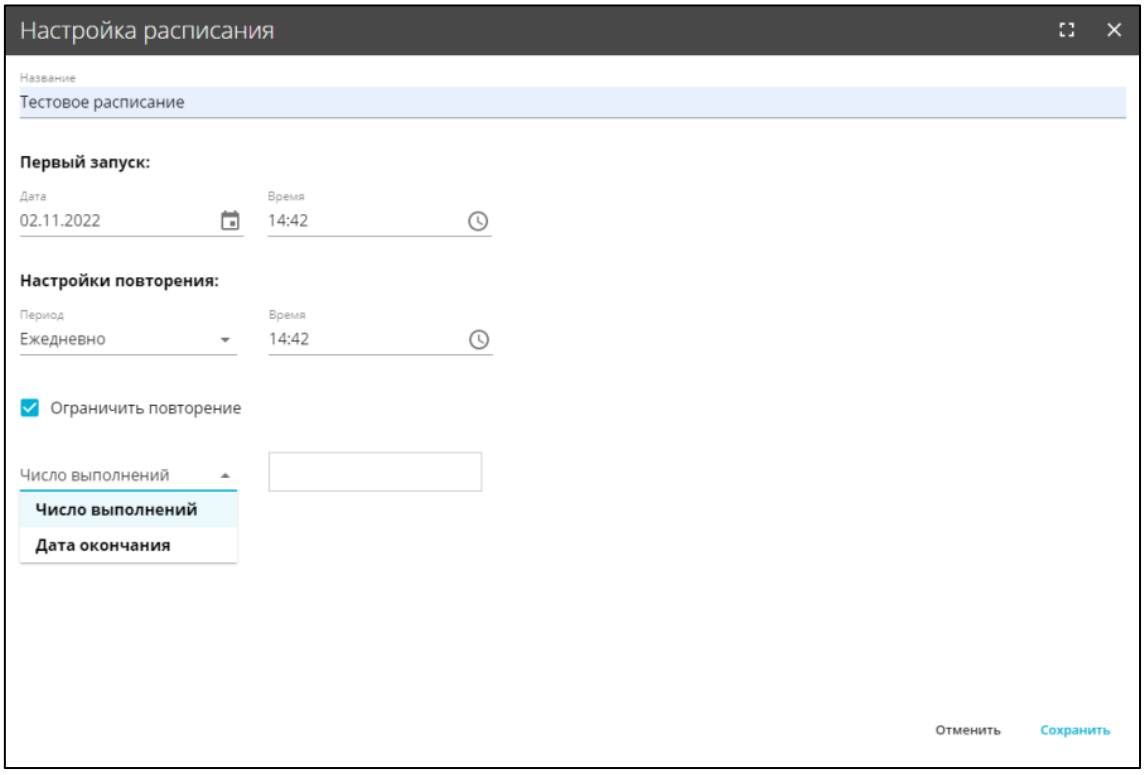


Рисунок 38 Окно Настройка расписания

После постановки Исследования на расписание в колонке **Статус** появится статус **Запланировано** и после каждого расчета результаты будут отображаться ниже по иерархии, никак не влияя на предыдущие и последующие расчеты.

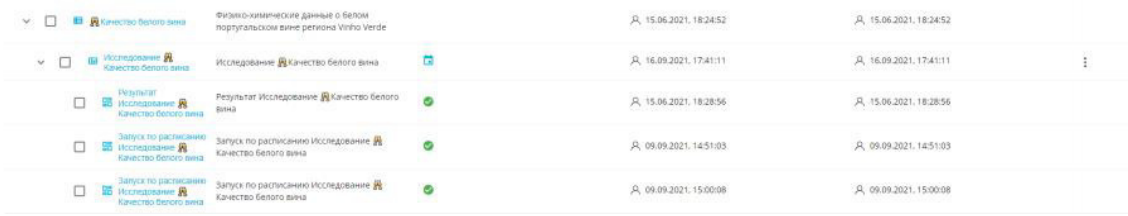


Рисунок 39 Пример Исследования, поставленного на расписание

2.4. Результаты исследования

2.4.1. Интерфейс экрана Результаты Исследования

На экране Результаты Исследования представлены построенные при помощи Мастера настройки визуализации, корреляционные матрицы и другие объекты.

Для перехода на экран Результаты Исследования необходимо на главном экране **компонента Исследование данных** выбрать необходимый результат исследования.

Рисунок 40 Пример экрана с результатами исследования и выделенными основными элементами

На экране с результатами исследования каждый объект расположен в отдельном контейнере. Для каждого из них предусмотрены изменение размера и положения — в верхней части контейнера иконки и , соответственно. Отрегулировать размер контейнера также можно ухватив левой кнопкой мыши правый нижний угол.

Также предусмотрена возможность сохранить визуализацию в формате .png. Для этого необходимо выбрать иконку в верхней части контейнера, либо при нажатии на контейнер правой кнопкой мыши в выпадающем меню выбрать **Сохранить картинку как..**. Копировать визуализацию в буфер обмена можно в том же выпадающем меню по пункту **Копировать картинку**.

Посмотреть на данные, которые лежат в основе визуализации, можно выбрав иконку в верхней части контейнера.

Дополнительно настроить контейнер можно в выпадающем меню в правом нижнем углу (элемент Настройка вида). Там же можно ознакомиться с примером исходного набора данных (элемент Пример данных) и выгрузить его локально (для этого в верхней части таблицы нужно выбрать иконку экспорта в Excel).

Визуальные объекты результатов исследования динамически настраиваемы. Они изменяют свой размер в соответствии с размером контейнера. Для многих типов визуализаций предусмотрены скрытие категорий/элементов. При наведении на элементы графиков выводятся всплывающие подсказки с подробной информацией о содержимом.

Для перехода между результатами исследований одного набора данных предусмотрена навигационная цепочка в верхней части экрана. Первая категория возвращает Пользователя на главную страницу компонента Исследование данных, вторая – позволяет выбрать интересующее исследование в рамках одного набора данных в выпадающем меню, тут же можно выбрать и интересующий результат, третья категория позволяет выбрать результат в рамках интересующего исследования.

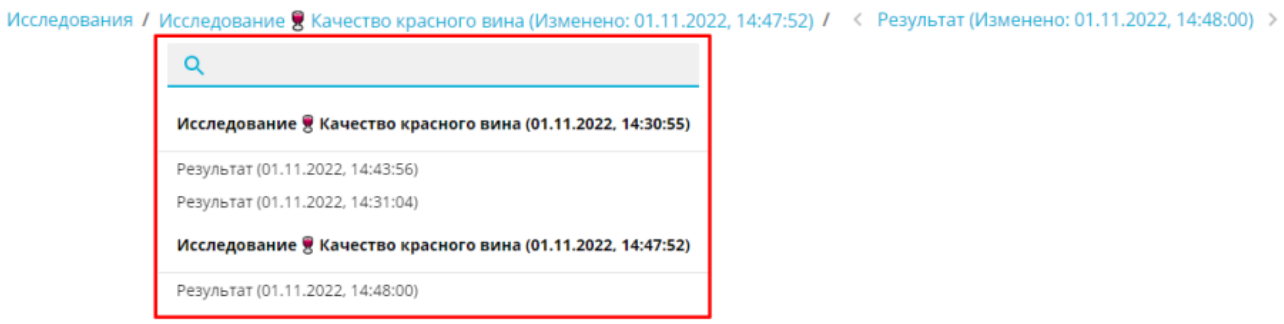


Рисунок 41 Экран с результатами нескольких исследований

На экране с результатами Исследования можно отобразить несколько результатов.

Для этого на Главном экране компонента необходимо выбрать чекбоксы рядом с двумя интересующими расчетами и выбрать **«Сравнить»** в нижней части таблицы (Рисунок 39). Таким образом, откроется экран, содержащий два результата исследования.

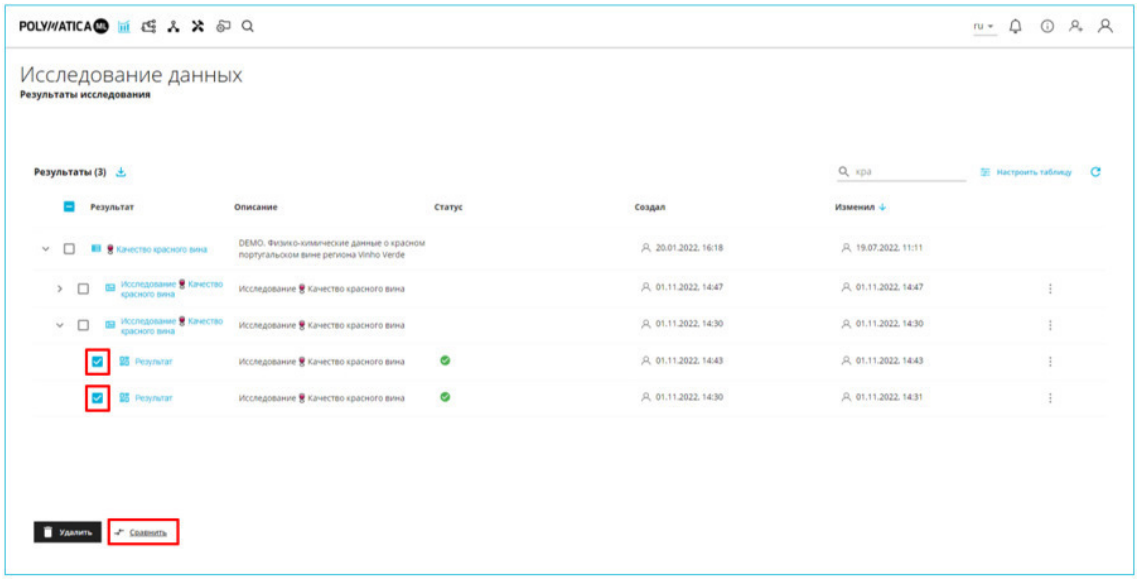


Рисунок 42 Пример выбора двух результатов для сравнения

Рисунок 43 Экран с результатами нескольких исследований

2.4.2. Профилирование

На экране с результатами исследования также представлен посчитанный в автоматическом режиме **Профиль данных**.

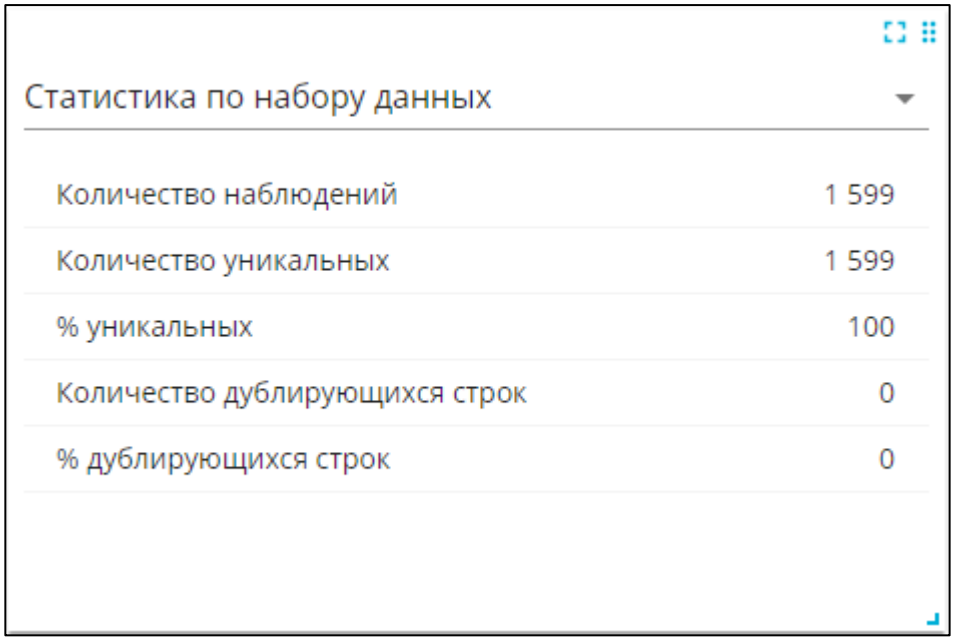


Рисунок 44 Пример Профиля данных

Профилирование представляет собой анализ данных с целью выяснения статистических характеристик переменных, таких как характер распределения, наличие выбросов, параметры выборки, пропущенные значения и другие.

Статистики считаются по всему набору данных и по каждой из переменных и зависят от ее типа.

2.4.2.1. Набор данных

Результат профилирования для всего набора данных зависит от размера анализируемого файла.

Так для набора данных размером менее 1 Гб считается следующий набор статистик:

- Количество наблюдений.
- Количество уникальных наблюдений.
- Процент уникальных наблюдений (в процентах).
- Количество дублирующих строк.
- Процент дублирующих строк (в процентах).
- Количество количественных переменных.
- Количество категориальных переменных.
- Количество переменных дат.

Для набора данных размером более 1 Гб:

- Количество наблюдений.
- Количество количественных переменных.
- Количество категориальных переменных.
- Количество переменных дат.

2.4.2.2. Категориальные переменные (Тип данных String)

Для категориальных переменных считается набор статистик в соответствии с таблицей ниже.

Статистика	Описание
Количество уникальных значений	Количество неповторяющихся значений
Процент уникальных значений	Процент уникальных значений от общего количества значений
Количество пропущенных значений	Количество пропущенных значений
Процент пропущенных значений	Процент пропущенных значений от общего количества значений
Максимальная длина	Количество знаков максимального по длине значения
Минимальная длина	Количество знаков минимального по длине значения
Top	3 значения с максимальной частотой
Bottom	3 значения с минимальной частотой

Таблица 4 Набор статистик, рассчитываемых для категориальных переменных

Для удобства интерпретации и анализа также предусмотрено построение **Облака слов**.

Облако слов (или облако тегов) демонстрирует частотность появления значения переменной. Размер облака отражает частоту появления значения. Цветовая гамма не несет в себе дополнительного смысла и выполняет исключительно эстетическую функцию.

Посмотреть **Облако слов** можно в том же контейнере с профилированием, выбрав в правом верхнем углу иконку  .

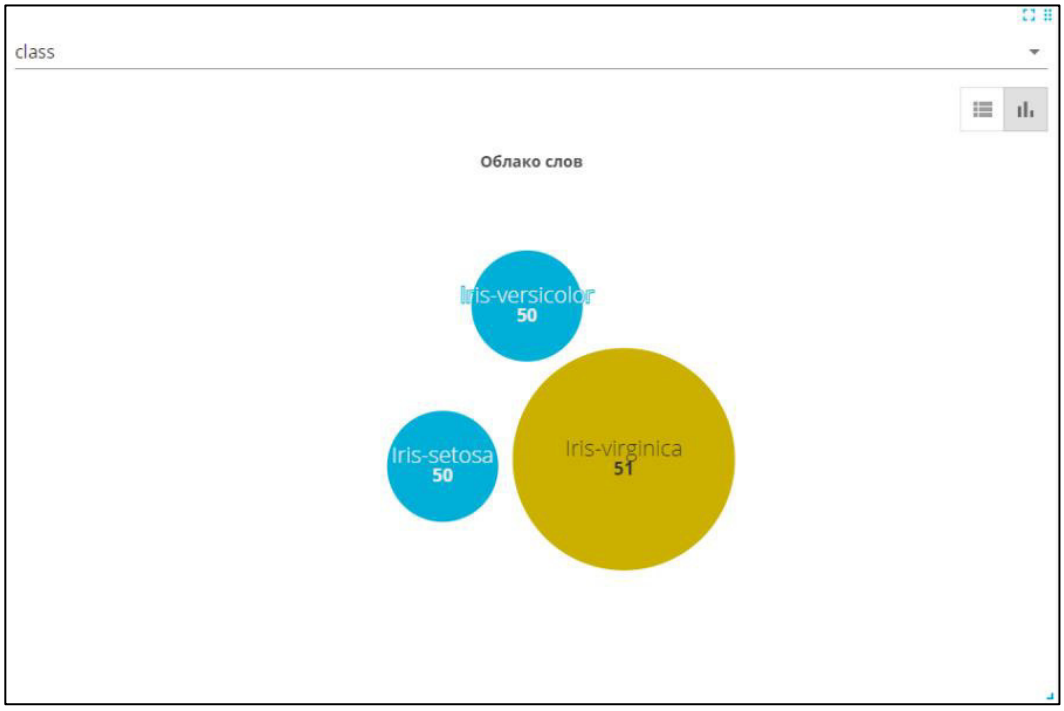


Рисунок 45 Пример Облака слов

2.4.2.3. Количественные переменные (Тип данных Numeric)

Для количественных переменных считается набор статистик в соответствии с таблицей ниже.

Статистика	Описание
Количество уникальных значений	количество неповторяющихся значений переменной
Процент уникальных значений	процент неповторяющихся значений переменной
Количество пропущенных значений	количество пропущенных значений переменной
Процент пропущенных значений	процент пропущенных значений
Минимальное значение	наименьшее значение переменной
Максимальное значение	наибольшее значение переменной
Среднее значение	сумма всех значений переменной, разделенная на число этих значений
5-я персентиль	это некоторое значение X из данного ряда, которое делит все имеющиеся в нем значения на две группы: 5% значений, которые меньше X, и оставшиеся значения (то есть 95%), которые превышают X.
95-я персентиль	это некоторое значение X из данного ряда, которое делит все имеющиеся в нем значения на две группы: 95% значений, которые меньше X, и оставшиеся значения (то есть 5%), которые превышают X.
1-я квартиль	25-я персентиль
3-я квартиль	75-я персентиль
Медиана	значение, которое делит распределение пополам (его площадь в т.ч.): половина значений больше медианы, половина — не больше.
Межквартильный размах	разница между 3-м и 1-м квартилями
Количество выбросов	Выбросами считаются наблюдения, которые отклоняются от своего математического ожидания более чем на три среднеквадратических отклонения (правило трех сигм).
Коэффициент вариации	величина, равная отношению стандартного (среднеквадратичного) отклонения случайной величины к ее математическому ожиданию. Он применяется для сравнения вариативности одного и того же признака в нескольких совокупностях с различным средним арифметическим. Если значение коэффициента вариации не превышает 33%, то совокупность считается однородной, а если больше 33%, то — неоднородной.
Коэффициент эксцесса	характеризует меру высоты графика. Если коэффициент больше нуля, то распределение является более высоким («островершинным») относительно «эталонного» нормального распределения. Если коэффициент ниже нуля, то более низким и пологим.

Статистика	Описание
Медианное абсолютное отклонение	вычисляется как медиана абсолютного значения для каждого значения минус медианное значение группы. Является статистикой, более устойчивой к выбросам в наборе данных, чем стандартное отклонение
Асимметрия	характеризует меру скошенности графика влево/вправо. Если коэффициент асимметрии отрицателен, то скос левосторонний. Если коэффициент положителен, то скос правосторонний. И чем коэффициент больше по модулю, тем сильнее скос распределения.
Стандартное отклонение	статистическая характеристика распределения случайной величины, показывающая среднюю степень разброса значений величины относительно математического ожидания. Большее значение среднеквадратического отклонения показывает больший разброс наблюдаемых значений признака относительно среднего; меньшее значение, соответственно, показывает, что величины в множестве сгруппированы вокруг среднего.
Дисперсия	величина, которая характеризует меру разброса значений случайной величины относительно ее математического ожидания.
Т-статистика	Т-статистика — это разница между средним по выборке и гипотетическим средним (предполагаемым равным нулю), деленная на расчетную стандартную ошибку среднего.
Пи-значение	Уровень значимости — вероятность получить Т-значение, равное или превышающее то значение, которое мы в действительности рассчитали по имеющимся выборочным данным (при условии, что нулевая гипотеза верна)

Таблица 5 Набор статистик, рассчитываемых для количественных переменных

Для удобства интерпретации и анализа предусмотрено построение **Гистограммы**.

Гистограмма визуализирует распределение данных в рамках непрерывного интервала. Каждая полоса представляет в табличной форме частотность за определенный бин. Количество бинов не изменяемо и по умолчанию равно 50. Посмотреть гистограмму можно в том же контейнере, выбрав в правом верхнем углу иконку  .

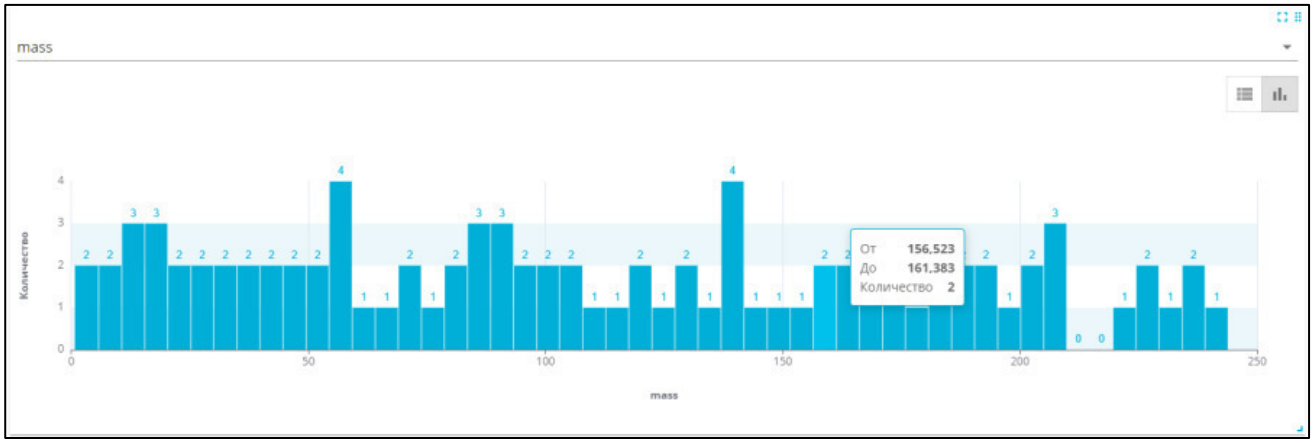


Рисунок 46 Пример Гистограммы

2.4.3. Результат расчета корреляции

Результатом расчета является корреляционная матрица, представляющая собой квадратную тепловую карту, в которой на пересечении выбранных признаков (одинаковый набор для столбцов и строк) находится значение коэффициента корреляции, закодированного при помощи градиентной шкалы. Синий цвет отражает максимальное значение коэффициента корреляции (равный 1), красный — минимальное (равный -1).

Матрица является симметричной, с единичной диагональю. При наведении указателя мыши на ячейку можно получить детализированную информацию:

- Коэффициент корреляции.
- Коэффициент ковариации.
- p-value — уровень значимости.

Коэффициент ковариации — это мера линейной зависимости двух случайных величин. Она является ненормированной версией коэффициента корреляции.

Коэффициент корреляции — статистическая мера, которая отражает силу связи между двумя порядковыми признаками. Коэффициент корреляции может принимать значения от −1 до +1. Если значение по модулю находится ближе к 1, то это означает наличие сильной связи, а если ближе к 0 — связь слабая или вообще отсутствует.

Вычисление коэффициента корреляции производится тремя методами:

- Корреляция Пирсона.
- Корреляция Спирмена.
- Корреляция Кендалла.

Для каждого метода предусмотрено построение своей корреляционной матрицы. Выбрать требуемую можно в выпадающем меню в верхней части панели. Ползунки боковой градиентной шкалы позволяют фильтровать ячейки корреляционной матрицы.

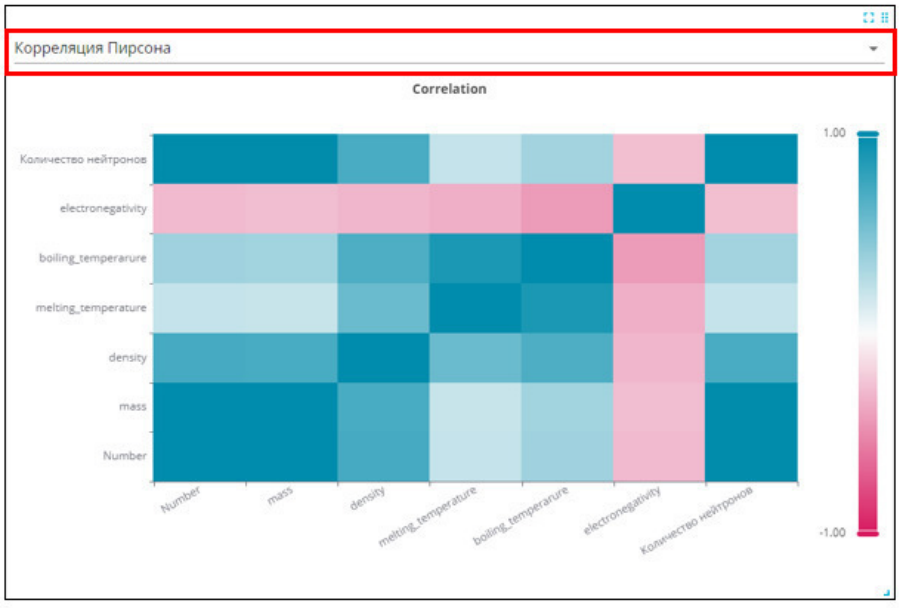


Рисунок 47 Результат расчета корреляционной матрицы

2.4.4. Результат расчета кластеризации

Результатом расчета является разбиение совокупности наблюдений на однородные группы, или кластеры.

Для удобства интерпретации результаты кластерного анализа представлены в отдельном контейнере со следующими объектами:

- **Круговая диаграмма** с количеством наблюдений в каждом кластере.

При наведении курсора мыши на сектор кластера можно узнать количество наблюдений в нем.

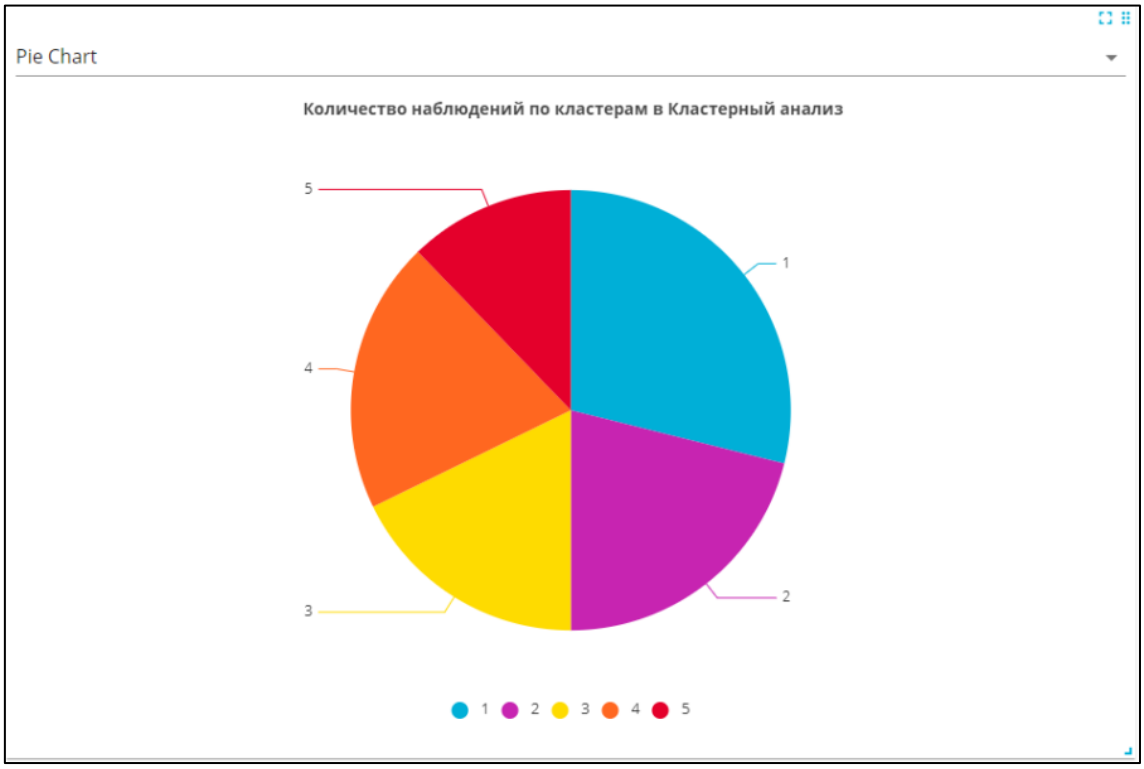


Рисунок 48 Пример Круговой диаграммы с результатами кластеризации

- **График в параллельных координатах** — Parallel Coordinates Plot.

Диаграмма с параллельными координатами позволяет интерпретировать построенные кластеры.

На диаграмме с параллельными координатами каждой переменной присваивается собственная ось. Оси располагаются параллельно друг другу, и каждая имеют свою собственную шкалу. Начальная ось отражает кластер, к которому модель отнесла наблюдение. Каждое наблюдение наносится на график в виде линии, пересекающейся с каждой из осей. Таким образом, пользователь может выявить паттерны и корреляции между разными переменными.

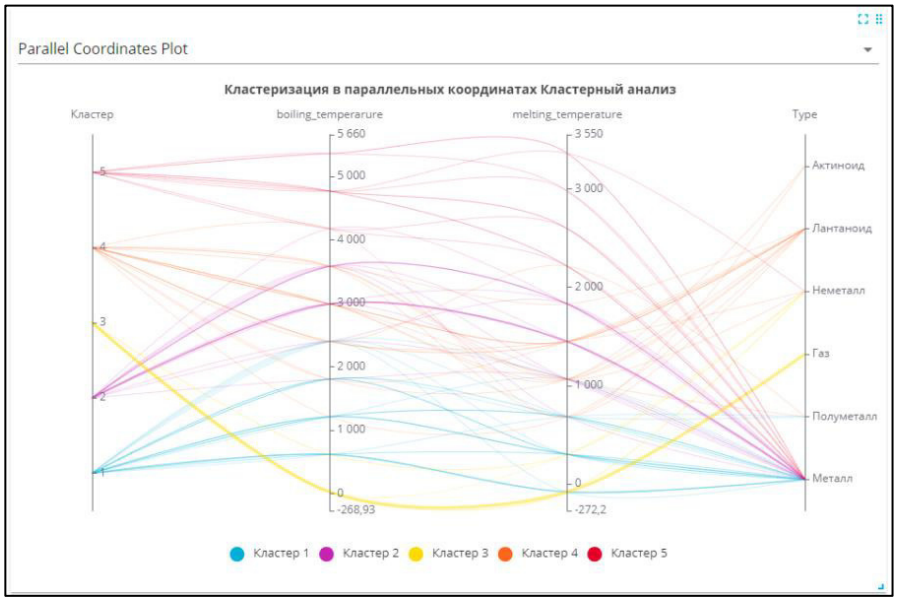


Рисунок 49 Пример Графика в параллельных координатах

- **Силуэт** — Silhouette Plot.

Значение Silhouette для каждого наблюдения является мерой того, насколько это наблюдение похоже на наблюдения в собственном кластере по сравнению с наблюдениями в других кластерах.

Значение Silhouette находится в диапазоне от -1 до 1. Высокое значение указывает на то, что наблюдение хорошо соответствует собственному кластеру и плохо соответствует другим кластерам.
Если большинство наблюдений имеют низкое или отрицательное значение Silhouette, тогда пользователь должен перестроить кластеризацию с большим или меньшим количеством кластеров.

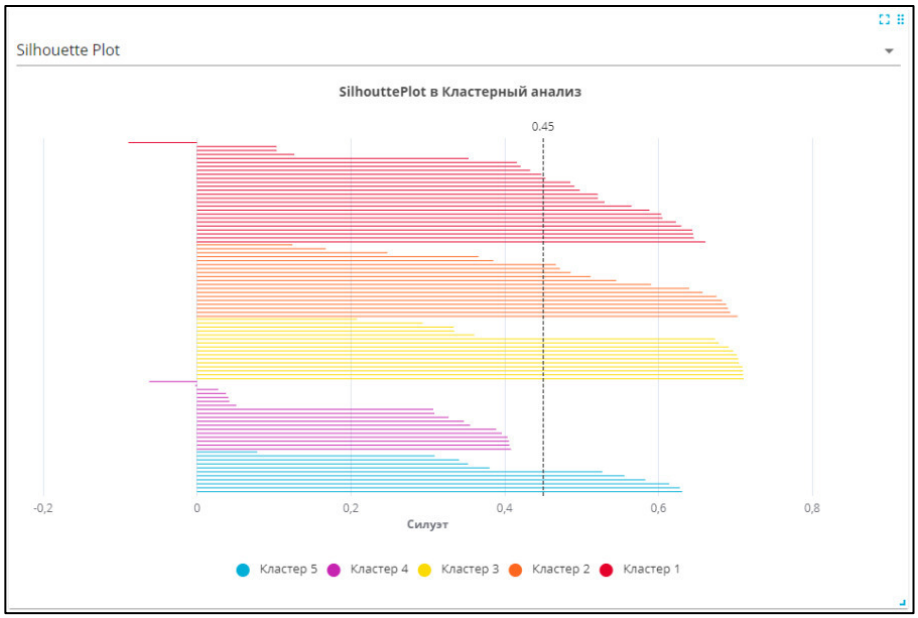
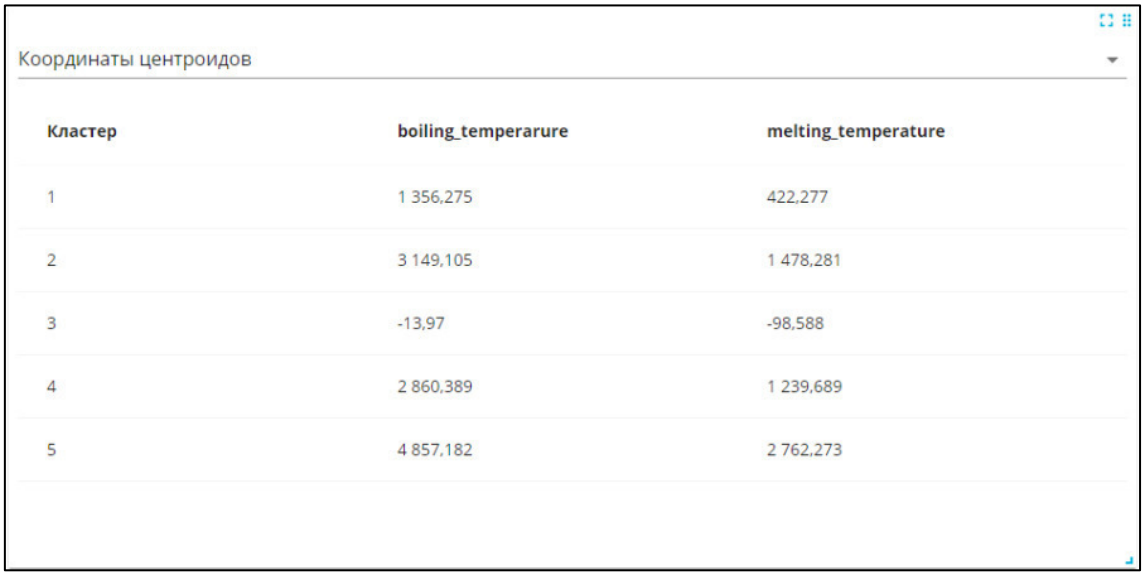


Рисунок 50 Пример Silhouette Plot

- **Таблица с координатами центроидов**, где в качестве строк выступают номера кластеров и значения переменных, в которых находятся центроиды этих кластеров.

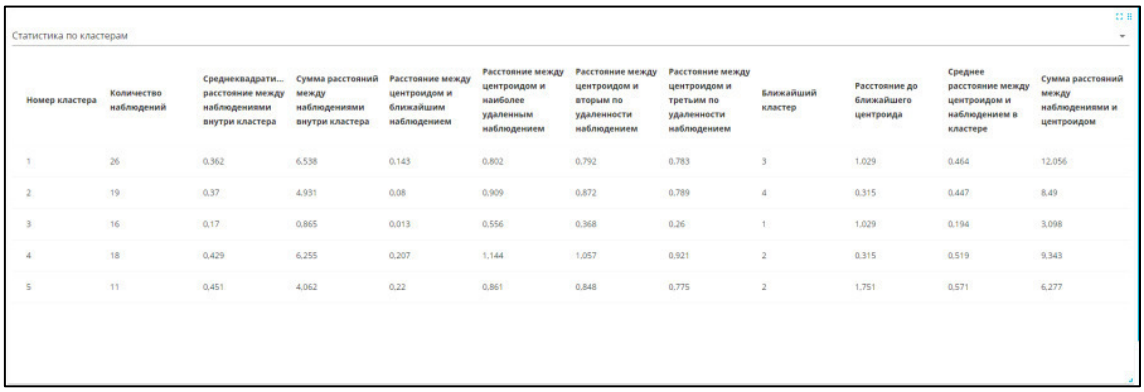


Координаты центроидов

Кластер	boiling_temperarure	melting_temperature
1	1 356,275	422,277
2	3 149,105	1 478,281
3	-13,97	-98,588
4	2 860,389	1 239,689
5	4 857,182	2 762,273

Рисунок 51 Пример таблицы с координатами центроидов

- **Таблица со статистиками по кластерам**. Содержит следующие статистики:
    - Номер кластера.
    - Количество наблюдений.
    - Среднеквадратичное расстояние между наблюдениями внутри кластера.
    - Сумма расстояний между наблюдениями внутри кластера.
    - Расстояние между центроидом и ближайшим наблюдением.
    - Расстояние между центроидом и наиболее удаленным наблюдением.
    - Расстояние между центроидом и вторым по удаленности наблюдением.
    - Расстояние между центроидом и третьим по удаленности наблюдением.
    - Ближайший кластер.
    - Расстояние до ближайшего центроида.
    - Среднее расстояние между центроидом и наблюдениями в кластере.
    - Сумма расстояний между наблюдениями и центроидом.



Статистика по кластерам

Номер кластера	Количество наблюдений	Среднеквадрати... расстояние между наблюдениями внутри кластера	Сумма расстояний между наблюдениями внутри кластера	Расстояние между центроидом и ближайшим наблюдением	Расстояние между центроидом и наиболее удаленным наблюдением	Расстояние между центроидом и вторым по удаленности наблюдением	Расстояние между центроидом и третьим по удаленности наблюдением	Ближайший кластер	Расстояние до ближайшего центроида	Среднее расстояние между центроидом и наблюдением в кластере	Сумма расстояний между наблюдениями и центроидом
1	26	0,362	6,538	0,143	0,802	0,792	0,783	3	1,029	0,464	12,056
2	19	0,37	4,931	0,08	0,909	0,872	0,789	4	0,315	0,447	8,49
3	16	0,17	0,865	0,013	0,556	0,368	0,26	1	1,029	0,194	3,098
4	18	0,429	6,255	0,207	1,144	1,057	0,921	2	0,315	0,519	9,343
5	11	0,451	4,062	0,22	0,861	0,848	0,775	2	1,751	0,571	6,277

Рисунок 52 Пример таблицы со статистиками кластеров

- **Информация по параметрам запуска**. Соответствует интерфейсу окна **Параметры кластеризации** в **Мастере настройки**.

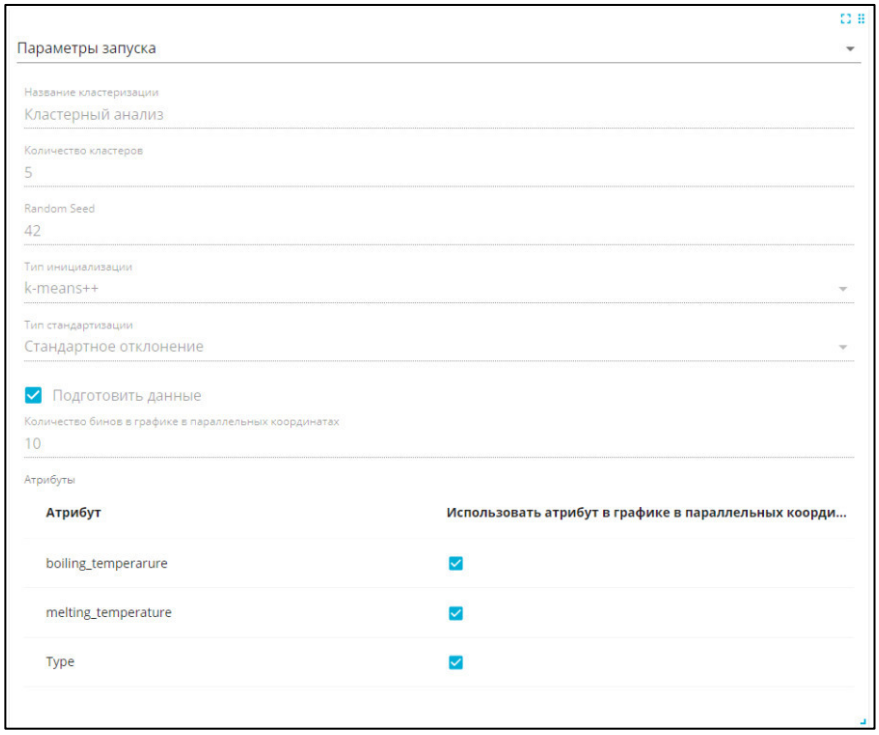


Рисунок 53 Пример вкладки с параметрами запуска кластерного анализа

- **Таблица со статистиками по переменным кластера**. По каждому кластеру отражены среднее и стандартное отклонение для каждой переменной.

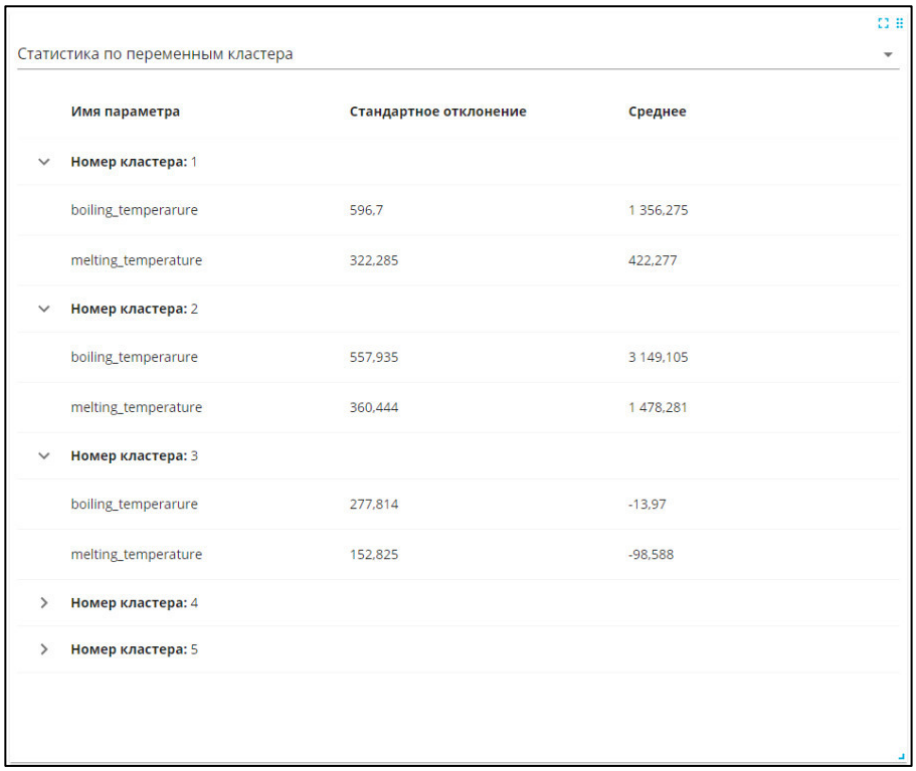


Рисунок 54 Пример таблицы со статистиками переменных по кластерам

2.4.5. Результат расчета статистических тестов

Результаты всех рассчитанных в исследовании статистических тестов отображаются в едином контейнере.

Каждый из статистических тестов имеет свои результаты. Выбрать результаты интересующего можно в выпадающем меню в верхней части контейнера.

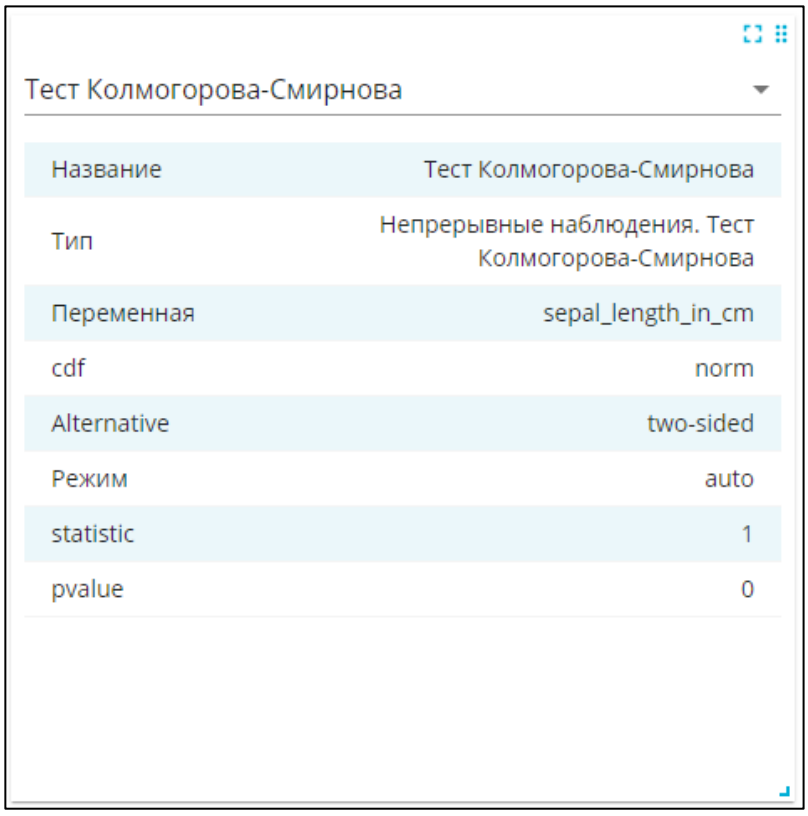


Рисунок 55 Контейнер с результатами статистических тестов

Помимо расчетных параметров, заданных при создании статистического теста, в данном контейнере отражаются следующие результаты:

- **Тест Колмогорова-Смирнова**.
  - statistic — статистика Колмогорова-Смирнова.
  - p-value — уровень значимости.

- **Тест Крамера-фон-Мизеса**.
  - statistic — статистика теста Крамера-фон-Мизеса.
  - p-value — уровень значимости.

- **Тест Андерсона-Дарлинга**.
  - statistic — статистика теста Андерсона-Дарлинга.
  - critical_values — критические значения для этого распределения.
  - significance_level — уровни значимости для соответствующих критических значений в процентах. Функция возвращает критические значения для различного набора уровней значимости в зависимости от тестируемого распределения. (Если возвращаемая статистика превышает эти критические значения, то для соответствующего уровня

значимости нулевая гипотеза о том, что данные принадлежат выбранному распределению, может быть отклонена.).
    - normal/exponential — 15%, 10%, 5%, 2.5%, 1%.
    - logistic — 25%, 10%, 5%, 2.5%, 1%, 0.5%.
    - Gumbel — 25%, 10%, 5%, 2.5%, 1%.

- **Критерий согласия хи-квадрат Пирсона**.
    - statistic — статистика критерий согласия хи-квадрат Пирсона.
    - p-value — уровень значимости.

2.4.6. Автоматическое построение модели в MD

В рамках модуля DD существует возможность быстрого построения пайплайна моделирования в MD.

Для этого на главном экране необходимо выбрать результат интересующего набора данных и на открывшемся экране выбрать кнопку **Построить модель**.

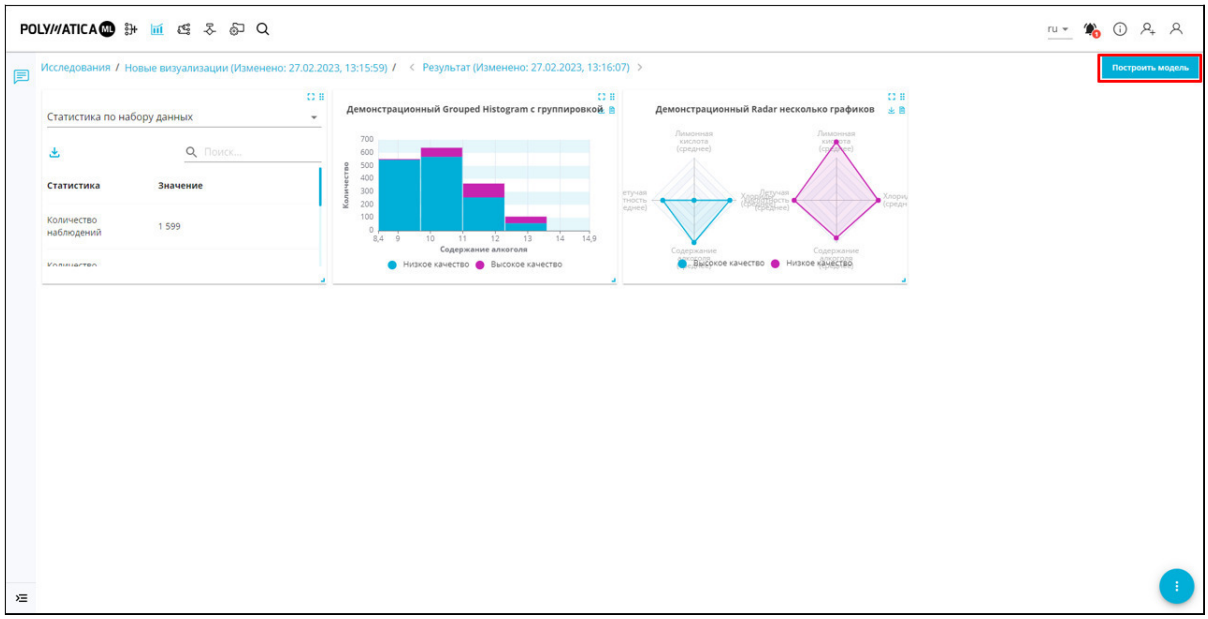


Рисунок 56 Расположение кнопки Построить модель

В окне **Построить модель** необходимо задать Название, Описание будущего проекта MD, а также выбрать Целевую переменную и отобрать необходимые для моделирования атрибуты из исходного набора данных. Нажать кнопку **Сохранить**.

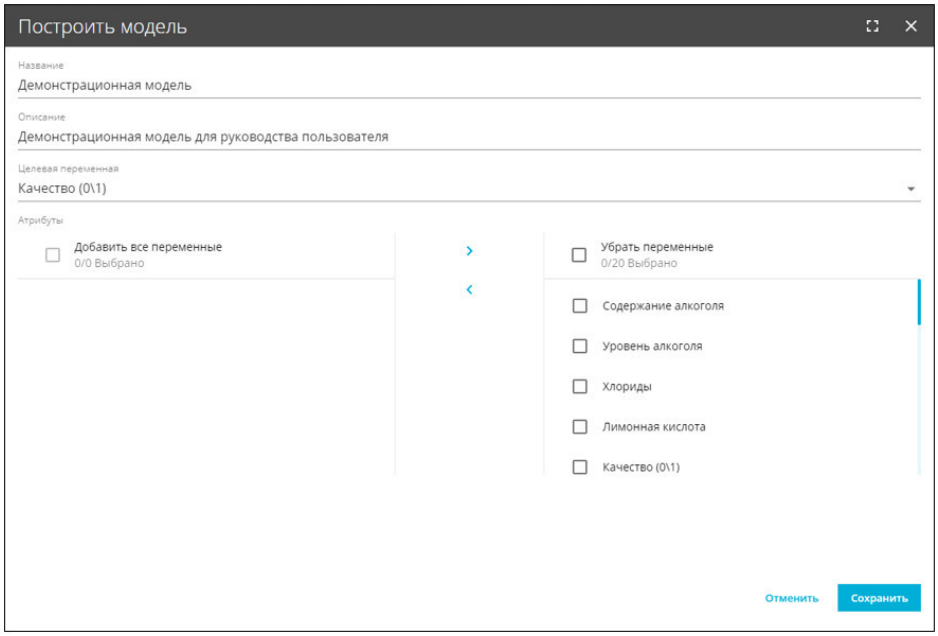


Рисунок 57 Окно Построить модель

После выполнения необходимых расчетов будет доступна кнопка **Перейти по ссылке**, которая переведет Пользователя к созданному проекту MD.

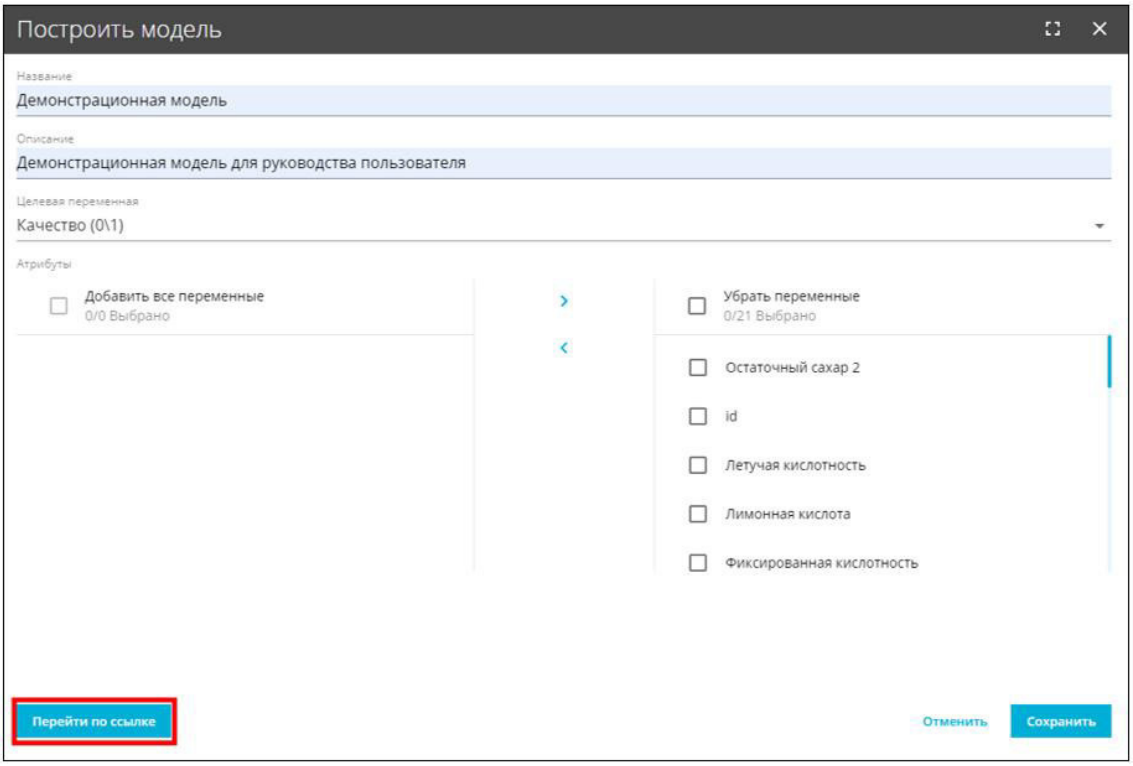


Рисунок 58 Кнопка Перейти по ссылке

В автоматически созданном проекте предусмотрены следующие узлы моделирования:

- Узел "Набор данных" (подробнее Узел «Набор данных»).
- Узел "Метаданные" (подробнее Узел «Метаданные»).
- Узел "Разделение выборки" (подробнее Узел «Разделение выборки»).
- Узел "AutoML" (подробнее Узел «AutoML»).

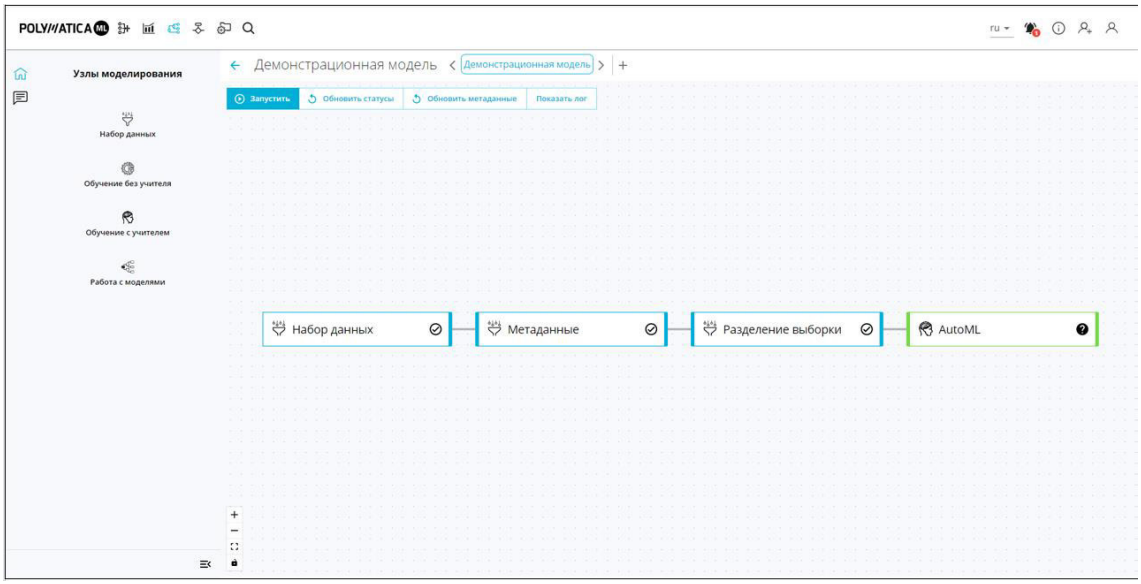


Рисунок 59 Пример созданного проекта MD

3. Компонент Построение моделей (Model Designer, MD)

Этап построения моделей включает в себя обширный пласт работ. Первоначально необходимо нормализовать и доработать данные согласно полученной на этапе их исследования информации и решаемой задачи. Сам процесс обучения происходит итерационно — пробуются разные модели, перебираются гиперпараметры, сравниваются значения выбранной метрики качества и выбирается лучшая комбинация. Также перед переходом к внедрению нужно убедиться, что результат моделирования понятен и логичен.

С учетом этих шагов в компоненте Построение моделей предусмотрены инструменты:

- Подготовки данных (обработка пропусков, расчет новых показателей, фильтрация и т.д.).
- Алгоритмы машинного обучения (включающие в себя широкий список методов обучения с учителем и обучения без учителя).
- Разделение набора данных на выборки для обучения, валидации и тестирования.
- Кросс-валидация.
- Автоподбор параметров.
- Оценки результатов (метрики и диаграммы оценки с учетом типа задачи).
- Сравнения моделей.
- Интерпретации моделей.
- Регистрация моделей в Репозиторий.

Компонент включает в себя:

- Главный экран со списком доступных проектов моделирования (проектов MD).
- Конструктор сценариев.
- Вспомогательные окна настройки вида, создания проекта и т.д.

3.1. Главный экран MD

3.1.1. Интерфейс главного экрана MD

Главный экран компонента **Построение моделей** открывается при выборе иконки  в левом верхнем меню и представляет собой список доступных пользователю **Проектов** в табличном виде.

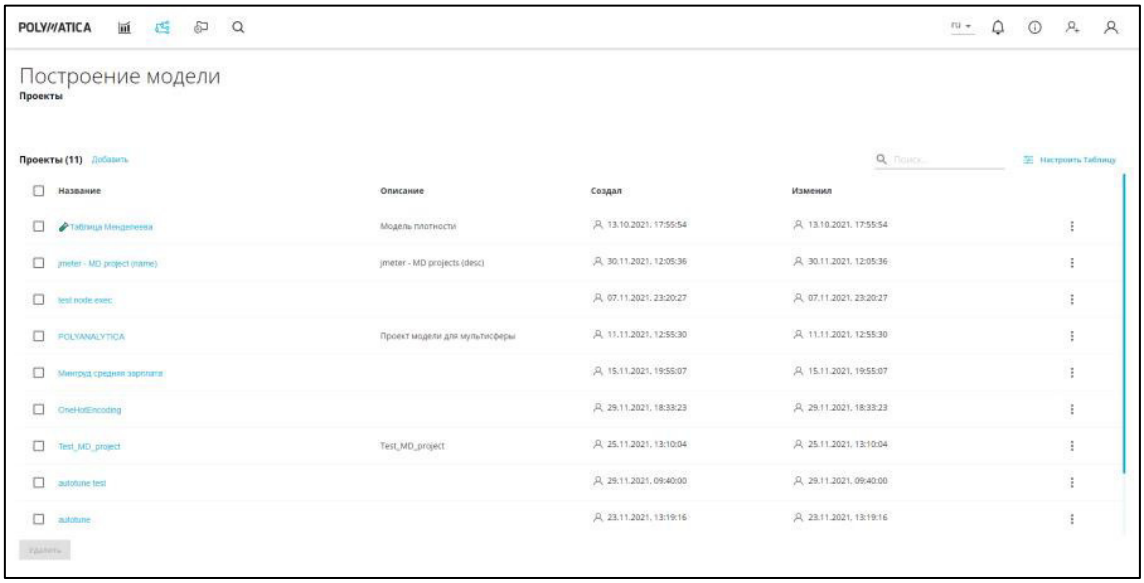


Рисунок 60 Главный экран компонента Построение моделей (Model Designer)

Проект MD — это сценарий моделирования, настраиваемый пользователем для решения конкретной задачи.
Таблица с доступными проектами имеет гибкие настройки отображения. Так, пользователь может.

- Изменить ширину любого столбца (для этого необходимо перетащить границу его заголовка до нужной ширины).
- Сортировать таблицу (для этого необходимо выбрать иконку рядом с заголовком сортируемого столбца).
- Скрывать/отображать столбцы и изменять их порядок в окне **Вид таблицы** (для открытия окна необходимо выбрать Настроить Таблицу в правом верхнем углу таблицы; при выборе иконки столбец скроется, при наведении на иконку активируется возможность перемещения столбца).
- Сбросить внесенные изменения также в окне **Вид таблицы** (для этого выбрать кнопку **«Сбросить»**).

Для быстрого поиска объекта в таблице предусмотрено поле Поиск... в правой верхней части таблицы.
Объекты таблицы можно выгрузить в формате Excel. Для этого нужно выбрать иконку **Экспорта в excel** .

3.1.2. Создание проекта MD

Для создания нового проекта необходимо выполнить следующие шаги:

- Выбрать кнопку **«Добавить»** в верхней части таблицы с Проектами.
- В открывшемся окне **Создание проекта** задать название и описание.
- Сохранить изменения.

3.1.3. Удаление проекта MD

Для удаления проекта необходимо выполнить следующие шаги:

- Выбрать чекбокс рядом с удаляемым проектом.
- Выбрать кнопку **«Удалить»**.

3.1.4. Копирование проекта MD

Для копирования проекта необходимо выполнить следующие шаги:

- С правой стороны от проекта выбрать иконку в виде трех вертикальных точек и в выпадающем меню выбрать **Копировать.**
- В открывшемся окне **Копирование проекта** задать название и описание нового проекта.
- Сохранить изменения.

3.1.5. Редактирование проекта MD

Для переименования и изменения описания проекта необходимо выполнить следующие шаги:

- С правой стороны от проекта выбрать иконку в виде трех вертикальных точек и в выпадающем меню выбрать **Редактировать.**
- В открывшемся окне **Редактирование проекта** внести необходимые изменения в названии и описании проекта.
- Сохранить изменения.

3.1.6. Работа с проектом MD

Для начала работы с проектом необходимо выбрать его из списка доступных.

3.2. Конструктор сценариев

При выборе проекта на главной странице компонента открывается **Конструктор сценариев**.

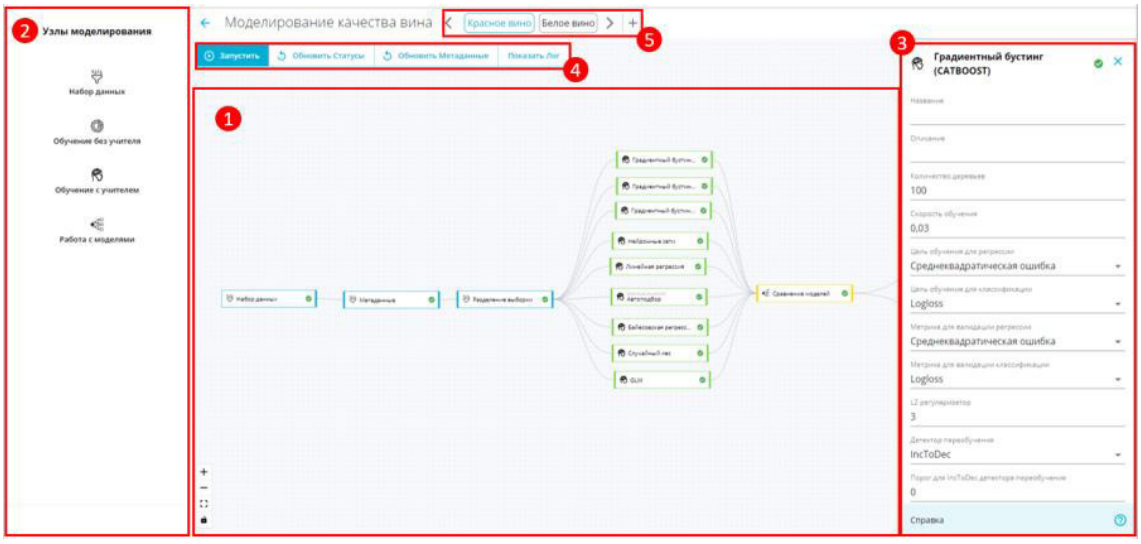


Рисунок 61 Элементы Конструктора сценариев

Сценарий — это комбинация узлов моделирования, настраиваемая пользователем в зависимости от решаемой задачи.

3.2.1. Интерфейс конструктора сценариев

Интерфейс **Конструктора сценариев** состоит из следующих элементов:

1. Рабочей области, в которую помещаются узлы моделирования.
2. Левой боковой панели с узлами моделирования.
3. Правой боковой панели с настройками узла (открывается при выборе узла, размещенного в рабочем поле).
4. Верхней панели с кнопками запуска узлов, обновления статусов и метаданных и раскрытием журнала логирования.
5. Панель создания нескольких сценариев в рамках одного проекта

Для удобства пользования рабочей областью в левом нижнем углу предусмотрены кнопки масштабирования (кнопки + и − для приближения и отдаления, соответственно, кнопка ⛶ для масштабирования на объектах, расположенных в рабочем поле) и кнопка, блокирующая перемещение объектов в рабочем пространстве (🔒).

Левую боковую панель можно скрыть, выбрав кнопку ⇤ в нижней части панели. Раскрыть скрытую панель можно выбрав ⇥ также в нижней части панели. Правая боковая панель открывается при выборе узла из рабочего поля. Скрыть эту панель можно выбрав иконку × в правом верхнем углу.

Панель с кнопками включает в себя:

- Кнопку **«Запустить»**, которая запускает процесс расчета сценария.
- Кнопку **«Обновить Статусы»,** которая обновляет статусы узлов.
- Кнопку **«Обновить Метаданные»**, которая обновляет метаданные узлов.
- Кнопку **«Показать Лог»,** которая раскрывает панель с логами сценария.

3.2.2. Логирование

Для удобства работы в компоненте предусмотрен просмотр лога событий, связанных с расчетом сценария моделирования.
При выборе кнопки **«Показать Лог»** в нижней части страницы откроется панель.

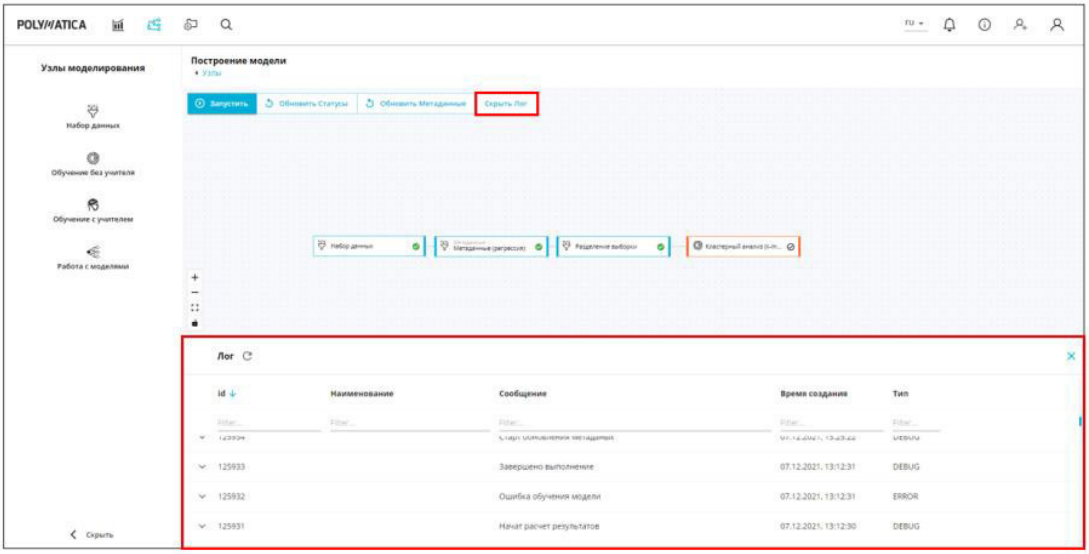


Рисунок 62 Журнал логирования

В этой панели отображена информация о событиях сценария, а именно:

- id события.
- Сообщение — описание события.
- Наименование – наименование узла MD.
- Дата и время события.
- Тип события. В лог записываются события типов INFO, DEBUG и ERROR.

В случае событий типа ERROR в описание также включен traceback. Для отображения необходимо выбрать иконку ⌄ рядом с событием.

Рисунок 63 Пример отображения traceback

Для удобства поиска записей в логе можно использовать фильтры. Фильтры предусмотрены для каждой из колонок таблицы.

3.2.3. Создание сценария

Для создания сценария необходимо:

- Поместить в рабочую область **узел Набор данных**. Он является первоначальным узлом любого сценария, т.к. подгружает набор данных, на основе которого будет строиться модель.
- Выбрать необходимый набор данных, задать остальные параметры узла и запустить расчет, выбрав кнопку **«Запустить».**
- Поместить следующий необходимый узел, создать связь с предыдущим, задать параметры и запустить.
- Последовательно проделать аналогичные шаги со всеми необходимыми узлами (о последовательности узлов подробнее в разделе Пример базового сценария).

3.2.4. Создание нескольких сценариев

В рамках моделирования может возникнуть необходимость построить несколько сценариев с разными настройками. Для создания нового сценария в рамках одного проекта необходимо:

- В верхней части конструктора рядом с названием проекта выбрать иконку + .
- В открывшемся окне **Создание сценария** задать название и описание нового сценарий и сохранить изменения.
- В панели рядом с названием проекта появится название созданного сценария.

Переключение между сценариями производится в панели рядом с названием проекта моделирования.

3.2.5. Узлы

Каждый **узел** отвечает за определенный этап машинного обучения.
Для работы с узлом необходимо выбрать его левой кнопкой мыши в панели с узлами и поместить в рабочее поле.

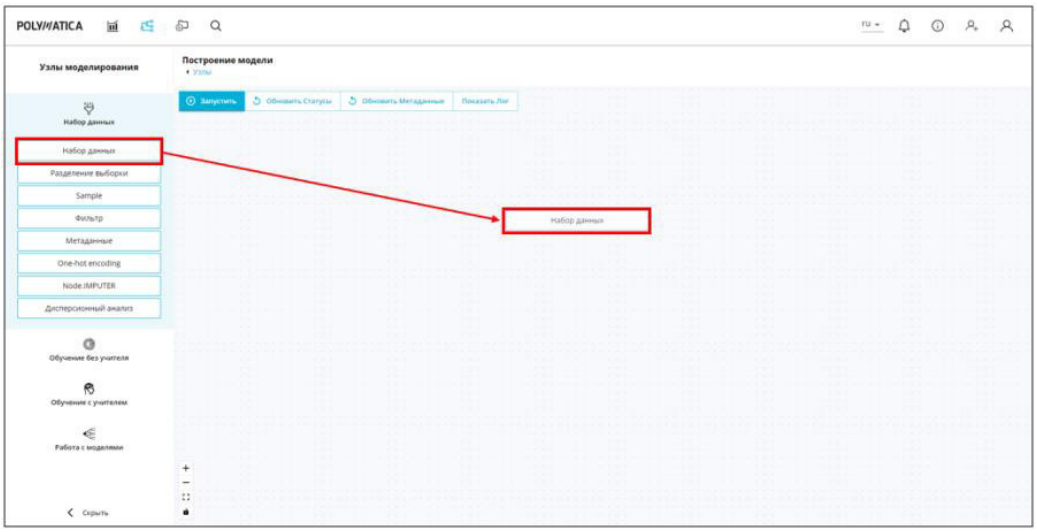


Рисунок 64 Размещение узла в рабочее поле

3.2.5.1. Статус Узла

Каждый узел имеет статус. Отображается он с правой стороны узла, а также в верхней части боковой панели с настройками узла.

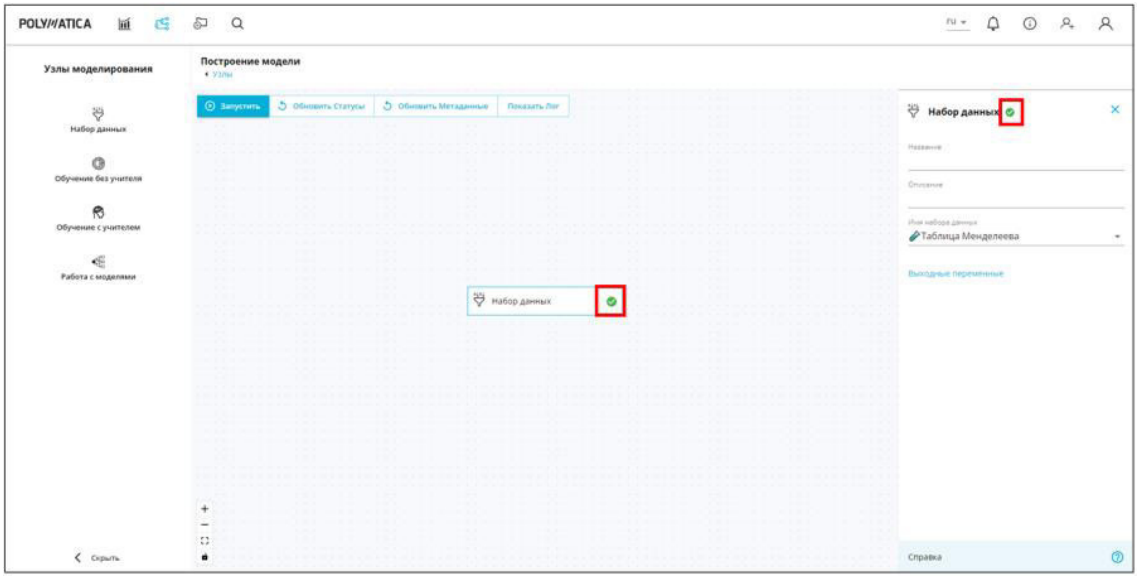


Рисунок 65 Пример статуса узла Набор данных

Узлы могут иметь следующие статусы:

- — необходимо обратить внимание на узел (при наведении на иконку появится всплывающая подсказка).
- — Узел готов к запуску.
- — Ошибка выполнения узла (понять причину ошибки можно в журнале логирования).
- — Узел выполнен.
- — Узел в процессе выполнения.

3.2.5.2. Действия над узлом

После размещения узла в рабочем поле над ним можно совершить действия. Для этого нужно нажать правой кнопкой мыши на узел и в выпадающем меню выбрать необходимое. Узел можно:

- Запустить.
- Копировать.
- Посмотреть результат выполнения (В результате выбора откроется окно с результатами расчета. Для каждого узла предусмотрены свои результаты).
- Удалить.
- Обновить статус.
- Загрузить pickle (Локальная выгрузка pickle файла).
- Загрузить CSV (Локальная выгрузка результатов узла в виде csv-файла).

Рисунок 66 Пример выпадающего меню

3.2.5.3. Связь между узлами

Для создания связи между узлами необходимо правой кнопкой мыши выбрать область с правой стороны узла и протянуть появившуюся линию связи до левой стороны следующего узла.

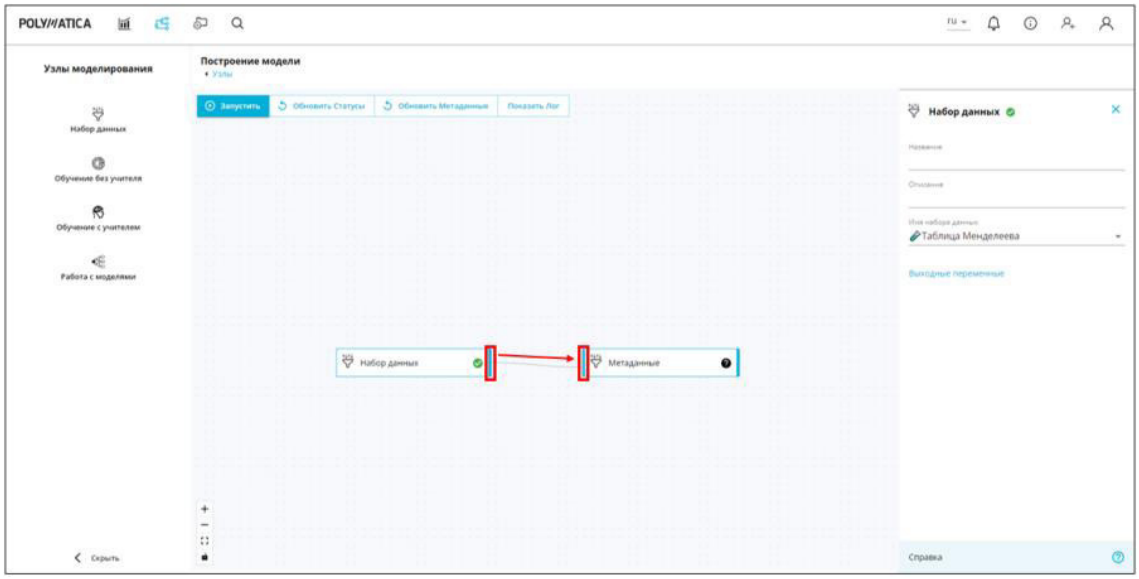


Рисунок 67 Создание связи между узлами

Для удаления связи необходимо левой кнопкой мыши выбрать линию связи и нажать на клавиатуре клавишу **Delete**.
В таблице ниже представлен список узлов, между которыми можно создавать связь.

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
Набор данных	Разделение выборки Sample Фильтр Метаданные Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов	–
Разделение выборки	Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM	Набор данных Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Sample	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM	Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Интерпретация Подбор отсечки (Cut off)
Фильтр	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Интерпретация	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Метаданные	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Кластерный анализ (k-means)

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM	Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
One-hot encoding	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей	Разделение выборки Sample Фильтр Метаданные One-hot encoding Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Интерпретация
Заполнение пропусков	Разделение выборки Sample Фильтр Метаданные Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means)	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Трансформация Бининг/энкодинг Стандартизация

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей	Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Интерпретация Регистрация модели
Трансформация	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Бининг/энкодинг Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Сравнение моделей	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Бининг/энкодинг Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Сравнение моделей Подбор отсечки (Cut off)
Бининг/энкодинг	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Линейные модели Нейронная сеть LDA Сравнение моделей	Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Сравнение моделей Подбор отсечки (Cut off)
Дисперсионный анализ	Бининг/энкодинг Фильтр	Разделение выборки Sample Фильтр Метаданные Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Сравнение моделей
Стандартизация	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Веса классов	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Кластерный анализ (k-means) Иерархическая	Набор данных Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Кластерный анализ (k-means)

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)	Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Кластерный анализ (k-means)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Регистрация модели	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Подбор отсечки (Cut off)
Иерархическая кластеризация	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM	Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Подбор отсечки (Cut off)
Дерево решений	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Подбор отсечки (Cut off)
Случайный лес	Разделение выборки Sample Фильтр Метаданные	Разделение выборки Sample Фильтр Метаданные

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Подбор отсечки (Cut off)
Байесовская регрессия	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost)

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	GLM Сравнение моделей Интерпретация Подбор отсечки (Cut off)
Линейная регрессия	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Логистическая регрессия	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Линейные модели	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)
Нейронная сеть	Sample Фильтр Метаданные Трансформация Бининг/энкодинг	Разделение выборки Sample Фильтр Метаданные Трансформация

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Бининг/энкодинг Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть Сравнение моделей
LDA	Разделение выборки Sample Фильтр Метаданные Трансформация Бининг/энкодинг Дисперсионный анализ Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные Трансформация Бининг/энкодинг Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели LDA Сравнение моделей Регистрация моделей Интерпретация
Градиентный бустинг (XGBoost)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Интерпретация Подбор отсечки (Cut off)
Градиентный бустинг (LightGBM)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (CatBoost) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)
Градиентный бустинг (CatBoost)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) GLM Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) GLM Сравнение моделей Подбор отсечки (Cut off)

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
GLM	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) Сравнение моделей Регистрация моделей Интерпретация Подбор отсечки (Cut off)	Разделение выборки Sample Фильтр Метаданные One-hot encoding Заполнение пропусков Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) Сравнение моделей Подбор отсечки (Cut off)
Сравнение моделей	Разделение выборки Sample Фильтр Метаданные One-hot encoding Трансформация Заполнение пропусков Бининг/энкодинг Дисперсионный анализ Стандартизация Веса классов Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Регистрация моделей Интерпретация	One-hot encoding Трансформация Заполнение пропусков Бининг/энкодинг Стандартизация Веса классов Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Подбор отсечки (Cut off)
Регистрация модели	-	Кластерный анализ (k-means) Дерево решений

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
		Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей
Интерпретация	Sample Фильтр Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей Подбор отсечки (Cut off)	Фильтр Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Нейронная сеть LDA Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM Сравнение моделей
Подбор отсечки (Cut off)	Разделение выборки Sample Метаданные Фильтр Трансформация Бининг/энкодинг Стандартизация Веса классов Кластерный анализ (k-means) Иерархическая кластеризация Дерево решений Случайный лес Байесовская регрессия Линейная регрессия Логистическая регрессия Линейные модели Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг	Стандартизация Веса классов Дерево решений Случайный лес Байесовская регрессия Логистическая регрессия Линейные модели Нейронная сеть Градиентный бустинг (XGBoost) Градиентный бустинг (LightGBM) Градиентный бустинг (CatBoost) GLM

Исходный узел	Исходящий узел (в который входит связь из исходного узла)	Входящий узел (из которого выходит связь в исходный узел)
	(CatBoost) GLM Сравнение моделей	

Таблица 6 Список возможной связи узлов

3.2.5.4. Входные и выходные данные узла

Каждый узел получает данные на вход, преобразует их и подает на выход. Входом текущего узла является выход предыдущего. Ознакомиться с входными и выходными параметрами можно в правой панели с настройками узла, щелкнув по ссылкам **«Входные переменные»** и **«Выходные переменные»**.

Для **узла «Набор данных»** не предусмотрены **ссылки** (и соответственно окна) **«Входные переменные»** и **«Выходные переменные»**. Ознакомиться с входными переменными **узла «Набор данных»** можно в **окне «Конфигурация переменных»** (подробнее в разделе Узел «Набор данных»)

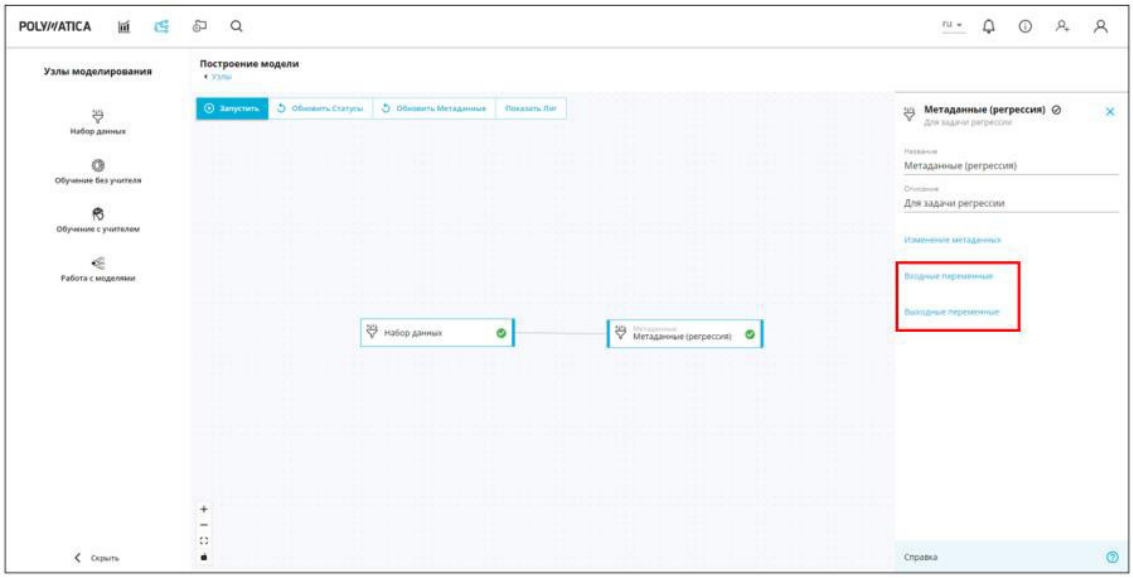


Рисунок 68 Расположение ссылок для просмотра Входных и Выходных переменных узла

При выборе ссылки **«Входные переменные»** откроется одноименное окно, в котором будут отражены подающиеся на вход узла переменные и их метаданные. При выборе же ссылки **«Выходные переменные»** откроется аналогичное окно, но с измененными и созданными переменными на выходе узла.

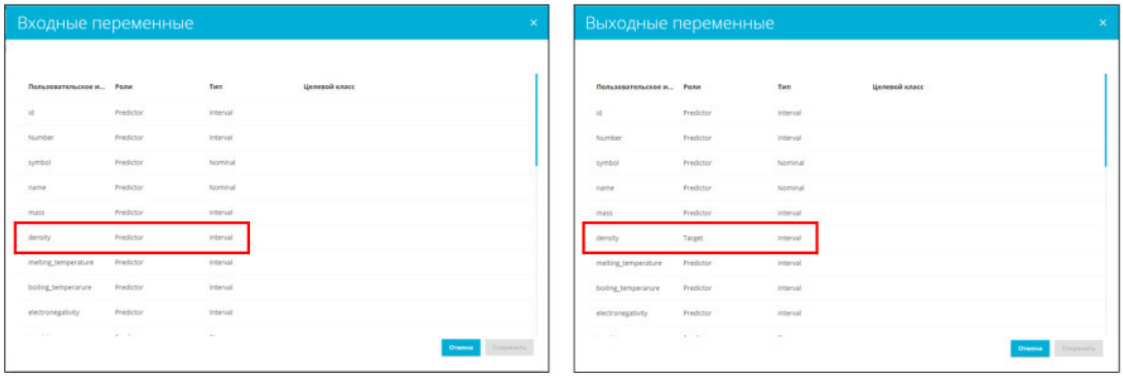


Рисунок 69 Пример Входных и выходных переменных узла «Метаданные»

К метаданным переменной относятся **Роль**, **Тип** и **Целевой класс**.

В Модуле предусмотрены следующие **Роли** переменной:

- **Target** — **Целевая переменная** — зависимая переменная, используемая в качестве цели моделирования.
- **Predictor** — **Предиктор** — независимая переменная, используемая в качестве прогнозирующей целевую переменную.
- **Excluded** — **Исключен** — данная роль позволяет исключить переменную из процесса моделирования.
- **Weight** — **Вес**— данная роль задается переменной, которая указывает модели на важность наблюдений. Пример: для прогноза могут быть важны тенденции за последний период (например, сезон).
- **Cluster_ID** — **Сегмент** — переменная с данной ролью рассчитывается при решении задачи кластеризации.
- **Prediction** — **Прогноз** — переменная с данной ролью рассчитывается при решении задачи регрессии.
- **Classification** — **Прогноз класс** — переменная с данной ролью рассчитывается при решении задачи классификации.
- **Partition** — **Разделение** — переменная с данной ролью задает разделение набора данных на обучающую, валидационную и тестовую выборки (рассчитывается в узле «Разделение выборки»).
- **ClassWeight** — **Вес класса** — переменная с данной ролью корректирует несбалансированность классов (задается в узле «Веса классов»).

Первоначально при выборе набора данных в узле «Набор данных» все атрибуты имеют роль Предиктор.

Предусмотрены следующие **Типы** переменной:

- **Номинальный (Nominal)** — дискретные значения без числовой связи между категориями.

- **Порядковый (Ordinal)** — дискретные значения, которые можно ранжировать или сортировать (Пример — хорошо, великолепно, превосходно).
- **Двоичный (Binary)** — это дискретные данные, которые могут относиться только к одной из двух категорий: 0 и 1.
- **Интервальный (Interval)** — числовые значения.

Целевой класс используется в процессе моделирования и может быть указан только для Целевой переменной.

Вручную изменить метаданные атрибутов можно при помощи узла «Метаданные».

3.2.5.5. Описание узлов

Каждый **Узел** выполняет отдельную операцию. Для удобства все узлы разделены на группы в зависимости от выполняемых функций:

- Группа «**Набор данных**» включает в себя узлы для подготовки и преобразования данных перед построением моделей.
- Группы «**Обучение с учителем**» и «**Обучение без учителя**» представлены узлами-алгоритмами машинного обучения.
- Узлы группы «**Работа с моделями**» используются на завершающих этапах моделирования для интерпретации, отсечки, сравнения моделей и регистрации в Репозиторий.

Сведения о конкретном узле можно получить в его **Справке**, которая находится в нижней части боковой панели с настройками узла. Для получения подробной информации об узле необходимо щелкнуть ссылку **«Подробнее».** В результате откроется новая вкладка в браузере с описанием узла из настоящего руководства пользователя.

Рисунок 70 Справка узла «Набор данных»

В текущей версии Модуля предусмотрены 37 узлов. Ознакомиться с кратким описанием каждого узла можно в таблице ниже.

Краткое описание узлов

Группа узлов	Название узла	Краткое описание
Набор данных	**Узел «Набор данных»**	Данный узел позволяет выбрать необходимый для моделирования набор данных из списка зарегистрированных
	Узел «Разделение выборки»	Данный узел разбивает набор данных на обучающую, валидационную и тестовую выборки
	Узел «Sample»	Данный узел корректирует неравномерное распределение классов в исходном наборе данных
	Узел «Фильтр»	Данный узел позволяет по заданным условиям удалить наблюдения из процесса моделирования
	Узел «Метаданные»	Данный узел позволяет изменить метаданные переменных
	Узел «One-hot encoding»	Данный узел преобразует категориальные переменные в числовые данные
	Узел «Заполнение пропусков»	Данный узел обрабатывает пропущенные значения
	Узел «Трансформация»	Данный узел позволяет рассчитать новые переменные
	Узел «Биннинг/энкодинг»	Данный узел включает в себя методы бинаризации интервальных переменных и кодирования категориальных переменных.
	Узел «Дисперсионный анализ»	Данный узел позволяет исследовать значимость различия между средними значениями зависимой количественной переменной по группам фактора (независимой переменной).
	Узел «Стандартизация»	Данный узел приводит признаки в разных единицах измерения и диапазонах значений к единому виду, который позволит сравнивать их между собой или использовать для расчета схожести объектов.
	Узел «Веса классов»	Данный узел корректирует несбалансированность классов (в обучающей выборке доли объектов разных классов существенно различаются)
	Узел «Автоэнкодер (PyTorch)»	
	Узел «PCA»	
	Узел «Профилирование»	
Обучение без учителя	**Узел «Кластерный анализ (k-means)»**	Данный узел группирует наблюдения в подмножества (кластеры) таким образом, чтобы наблюдения внутри одного кластера были похожи друг на друга, но различались с наблюдениями из других кластеров.
	Узел «Иерархическая кластеризация»	Данный узел создает иерархии вложенных подмножеств (кластеров).

Группа узлов	Название узла	Краткое описание
	Узел «Детекция аномалий (PyTorch)»	
	Узел «Ассоциативные правила»	
Обучение с учителем	**Узел «Дерево решений»**	Данный узел обобщает наблюдения правилами вида «Если…, то…» в иерархическую, последовательную структуру в виде дерева. Используется для решения задач классификации и регрессии
	Узел «Случайный лес»	В основе данного узла лежит алгоритм, который представляет собой ансамбль деревьев решений. Используется для решения задач классификации и регрессии
	Узел «Байесовская регрессия»	Данный узел представляет собой линейную регрессию с применением распределения вероятностей параметров, а не точечных оценок Используется для решения задач регрессии
	Узел «Линейная регрессия»	В результате данного узла строится модель зависимости между входными и выходными переменными с линейной функцией связи Используется для решения задач регрессии
	Узел «Логистическая регрессия»	В основе данного узла лежит метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам. Используется для решения задач классификации
	Узел «Линейные модели»	Данный узел объединяет в себе линейные классификаторы и регрессоры
	Узел «Нейронная сеть»	В основе данного узла лежит упрощенная модель биологической нейронной сети. Используется для решения задач классификации и регрессии
	Узел «LDA» (Линейный дискриминантный анализ)	Данный узел применяется для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий.
	Узел «Градиентный бустинг (XGBOOST)»	В основе данного узла лежит алгоритм градиентного бустинга на деревьях поиска решений. Используется для решения задач классификации и регрессии.
	Узел «Градиентный бустинг (XGBOOST)»	В основе узла лежит реализация алгоритма градиентного бустинга на деревьях поиска решений, который включает в себя две ключевые идеи: Градиентная односторонняя выборка (GOSS) и Объединение взаимоисключающих признаков (EFB). Используется для решения задач классификации и регрессии.

Группа узлов	Название узла	Краткое описание
	Узел «Градиентный бустинг (XGBOOST)»	В основе узла лежит реализация алгоритма градиентного бустинга, которая оптимизирована под работу с категориальными признаками и хорошо работает с параметрами по умолчанию. Используется для решения задач классификации и регрессии.
	Узел «GLM»	Данный узел обобщает линейную регрессию и допускает наличие у зависимой переменной распределения, отличающегося от нормального. GLM связывает зависимую переменную с факторами посредством задаваемой функции связи.
	Узел «Нейронная сеть (PyTorch)»	
	Узел «AutoML»	
Работа с моделями	**Узел «Сравнение моделей»**	Данный узел оценивает построенные модели и выбирает лучшую.
	Узел «Регистрация модели»	Данный узел сохраняет построенную модель в выбранном проекте репозитория Model Manager.
	Узел «Интерпретация»	Данный узел, включает в себя методы, которые позволяют объяснить принципы и закономерности, которые использует модель в ходе прогнозирования.
	Узел «Подбор отсечки (Cut off)»	Данный узел позволяет определить оптимальный порог отсечения для высокого соотношения истинных и ложных срабатываний модели

3.2.5.6. Группа узлов «Набор данных»

3.2.5.6.1. Узел «Набор данных»

Узел «Набор данных» является начальным узлом любого сценария моделирования, позволяет выбрать необходимый для моделирования набор данных из списка зарегистрированных и задать конфигурацию переменных.

При конфигурировании переменных Пользователь может указать необходима ли переменная далее в процессе моделирования, а также указать необходимо ли регистрировать переменную при регистрации модели.

Помимо этого, Пользователь может отфильтровать, а также случайным образом отобрать наблюдения, которые далее будут использоваться в сценарии моделирования.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе

Параметр	Возможные значения и ограничения	Описание
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Имя набора данных	Раскрывающийся список с названиями наборов данных	Список доступных пользователю наборов данных
Конфигурация переменных	Кнопка	При выборе кнопки откроется окно **Конфигурация переменных** (подробное описание ниже).
Фильтр	Поле ввода условия фильтрации	При выборе поля ввода откроется окно **Формула** (подробное описание ниже).
Доля наблюдений для выборки	Ручной ввод Значение больше 0 и меньше или равно 1	Данный параметр задает долю наблюдений исходного набора данных, которые будут включены в итоговый набор данных
Максимальное количество загружаемых строк	Ручной ввод целочисленного значения По умолчанию – 100000	Данный параметр ограничивает количество строк, которые будут использоваться в ходе моделирования

Таблица 7 Параметры узла "Набор данных"

Окно Конфигурация переменных

В окне **Конфигурация переменных** Пользователь может ознакомиться с входными переменными узла и указать на необходимость регистрировать переменную при сохранении модели в репозиторий (столбец **Регистрация**) и необходимость данной переменной при построении дальнейшего сценария (столбец **Дроп**).

При выборе **Регистрация** (значение Да) данная переменная будет зарегистрирована при сохранении модели в репозиторий ММ.

При выборе **Дроп** (значение Да) далее в сценарии данная переменная использоваться не будет и вернуть ее в процесс будет невозможно. Для временного исключения атрибута необходимо назначить ему роль Исключен (подробнее узел «**Метаданные**»).

> **ВАЖНО:** Целевой переменной на данном шаге необходимо задать значение Нет в столбце **Регистрация**. Это необходимо, чтобы далее данная переменная не фигурировала при регистрации и при публикации в ММ.

Роль и **Тип переменной** можно изменить в узле "**Метаданные**". **Тип источника** задается автоматически: если переменная из исходного набора данных, то указано значение **Base**, если рассчитана в системе - значение **Computed**.

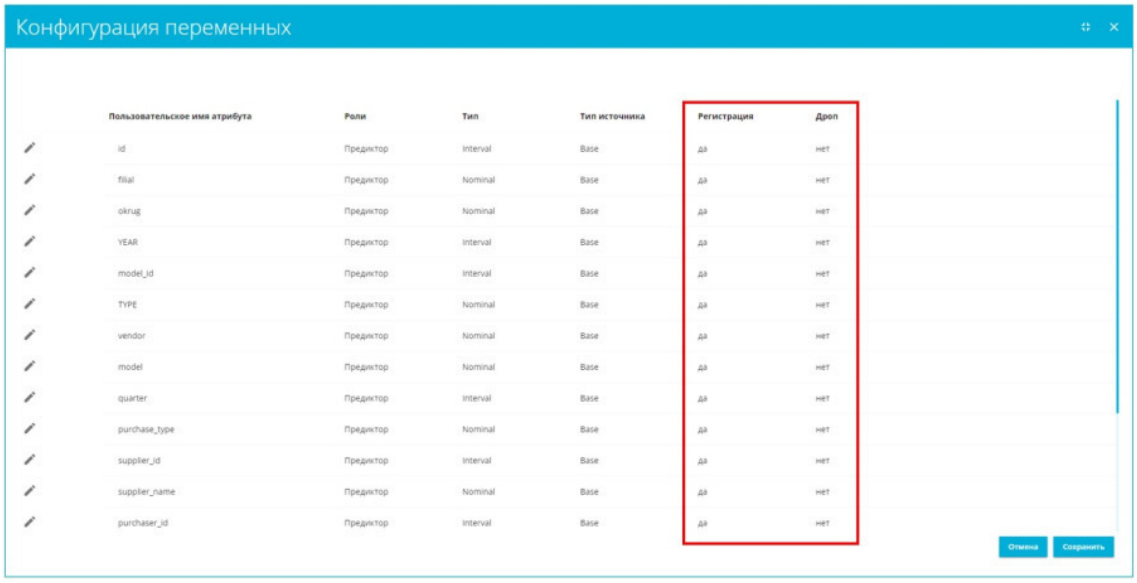


Рисунок 71 Окно Конфигурация переменных

Окно Формула

В окне **Формула** задается условие отбора наблюдений из исходного набора данных.

Основными элементами окна Формула являются:

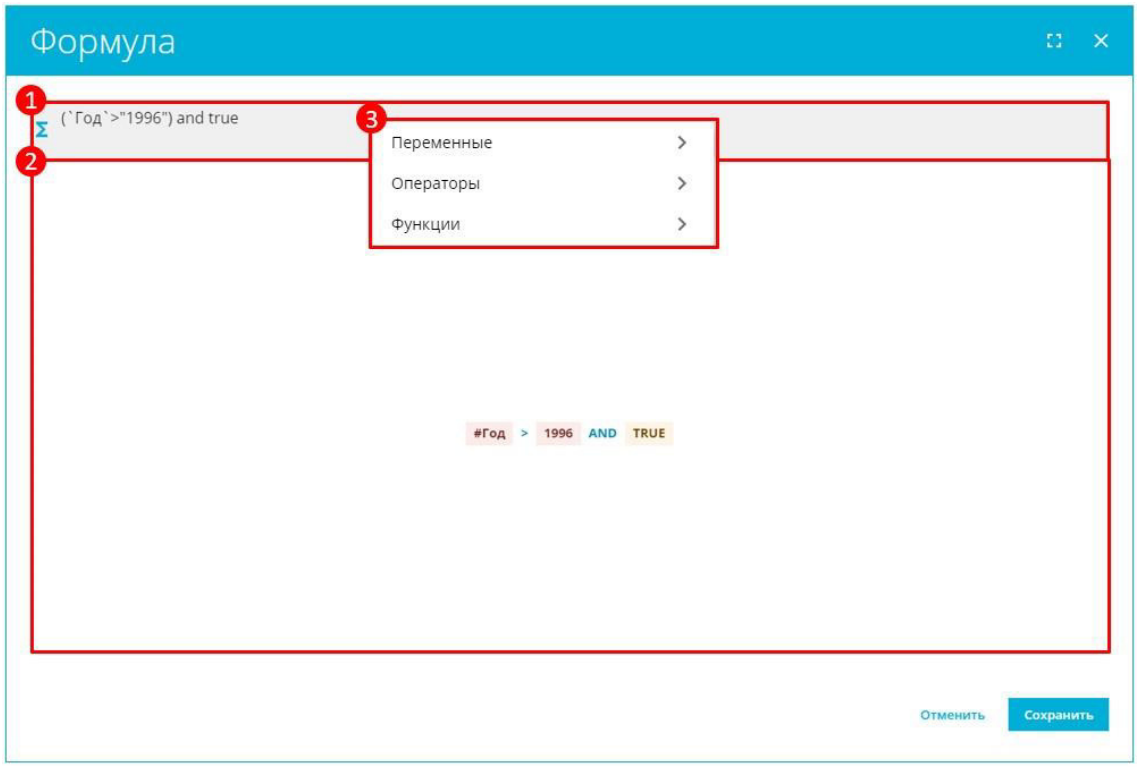


Рисунок 72 Окно Формула

1. Строка ввода текстового представления условия.
2. Рабочее поле, в котором строится графическое представление условия.
3. Вкладка, которая открывается при щелчке правой кнопкой мыши в строке ввода, и включает в себя:
    - Переменные из набора данных.
    - Операторы и функции, представленные в таблице ниже.

Позволяет просматривать переменные набора данных и строить конструкции условий.
Для **задания условия** необходимо:

- В строку текстового представления ввести условие, используя операторы, функции и переменные набора данных.
    - Название переменной должно быть указано в обратных одинарных кавычках (` `). Пример: **`Год`** > 1996.
    - В качестве десятичного разделителя используется точка.
    - Строка должна указываться в двойных кавычках. Пример: `Пол` = **"мужской"**.
- Посмотреть список операторов, функций и переменных, а также добавить их в условие можно и в панели, которая открывается при щелчке правой кнопкой мыши по строке ввода.

Оператор/Функция	Описание	Текстовое представление (формула)
And	Логическая операция «И»	true **and** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 and `Пол`= "Женский"
Or	Логическая операция «ИЛИ»	true **or** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 or `Пол`= "Женский"
+	Операция сложения	0 **+** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`+ ` Цена продукта М`
–	Операция вычитания	0 **-** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`- 1000
×	Операция умножения	0 ***** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М` * `Скидка`
/	Операция деления	0 **/** 0 Вместо 0 нужно вставить числовое значение/числовую переменную

Оператор/Функция	Описание	Текстовое представление (формула)
		Пример: `Выручка за первый квартал` / `Выручка за год`
<	Меньше	"" **<** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` < 32
≤	Меньше или равно	"" **<=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` <= 120
>	Больше	"" **>** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` > "32"
≥	Больше или равно	"" **>=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` >= 120
=	Равно	"" **=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` = 120
≠	Не равно	"" **!=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` != 120
like	Проверяет, удовлетворяет ли символьная строка заданному образцу, который может содержать поисковые символы. Учитывает регистр	"" **like** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` like "Москва"
ilike	Проверяет, удовлетворяет ли символьная строка заданному образцу, который может содержать поисковые символы.	"" **ilike** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` ilike "Москва"
startswith	Проверяет, есть ли в начале одной текстовой строки другая текстовая строка	"" **startswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` startswith "Сан"
endswith	Проверяет, есть ли в конце одной текстовой строки другая текстовая строка	"" **endswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` endswith "бург"
contains	Проверяет, встречается ли указанная строка внутри другой строки	"" **contains** "" Вместо "" нужно вставить переменную/значение **Пример**: `Название` contains "consulting"

Оператор/Функция	Описание	Текстовое представление (формула)
between	Проверяет, входит ли значение в заданный диапазон	0 **between**(0,0) Вместо 0 нужно вставить числовую переменную/значение **Пример**: `ID` between (1, 100500)
in	Проверяет наличие элемента в последовательности	"" **in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` in ("Москва", "Московская область")
not in	Проверяет отсутствие элемента в последовательности	"" **not_in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` not_in ("Москва", "Московская область")
coalesce	Возвращает данные из первого столбца, содержащего значение, отличное от NULL	**coalesce**("") Вместо "" нужно вставить переменную/значение **Пример**: coalesce(`ProductNumber`, `ProductName`)
not	Задает противоположное условие	not(true) Вместо true нужно вставить условие **Пример**: not(`Год` > 1996)
create_date	Создает переменную даты из последовательно введенных года, месяца и дня	create_date(0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_date(2016,10,25)
current_date	Возвращает текущую дату	current_date() Вводится без дополнительных параметров
current_timestamp	Возвращает текущие дату и время	current_timestamp() Вводится без дополнительных параметров
now	Возвращает текущие дату и время	now() Вводится без дополнительных параметров
create_datetime	Создает переменную даты времени из последовательно введенных года месяца дня часа минут и секунд	create_datetime(0,0,0,0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_datetime(2016,10,25,12,18,0)
char_length	Возвращает длину строки	char_length("") Вместо "" нужно вставить строку/строковую переменную **Пример**: char_length(`Код_продукта`)
random	Возвращает случайное число	random() Вводится без дополнительных параметров
ln	Натуральный логарифм	ln(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: ln(`Числовая переменная`)
exp	Экспонента	exp(0)

Оператор/Функция	Описание	Текстовое представление (формула)
		Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: exp (`Числовая переменная`)
power	Возведение в степень	power(0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: power(`Числовая переменная`,2)
sqrt	Квадратный корень	sqrt(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: sqrt (`Числовая переменная`)
abs	Абсолютное значение	abs(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: abs (`Числовая переменная`)
ceil	Возвращает наименьшее целое число, которое больше или равно текущему значению	ceil(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: ceil(25.1) Вернет значение 26
floor	Возвращает наибольшее целое число, которое меньше или равно текущему значению	floor(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: floor(25.1) Вернет значение 25
extract_from_datetime	Получает указанную часть (день, месяц, год, час, минута, секунда) из значения даты и времени	extract_from_datetime("","") Вместо первых кавычек нужно указать необходимую часть для извлечения: · SECOND · MINUTE · HOUR · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа datetime **Пример:** extract_from_datetime("DAY",`datetime`)
extract_from_date	Получает указанную часть (день, месяц, год) из значения даты	extract_from_date("","") В первые кавычки нужно вписать необходимую часть для извлечения: · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа date **Пример:** extract_from_date("DAY",`date`)

Оператор/Функция	Описание	Текстовое представление (формула)
concat	Объединяет в единую строку указанные значения	concat("") Вместо "" нужно вставить строки/строковые переменные **Пример:** concat (`Фамилия`, " ", `Имя`)

Таблица 8 Операторы и функции

Результаты выполнения узла:

- **Таблица с примером данных**. Отображаются первые 100 наблюдений.

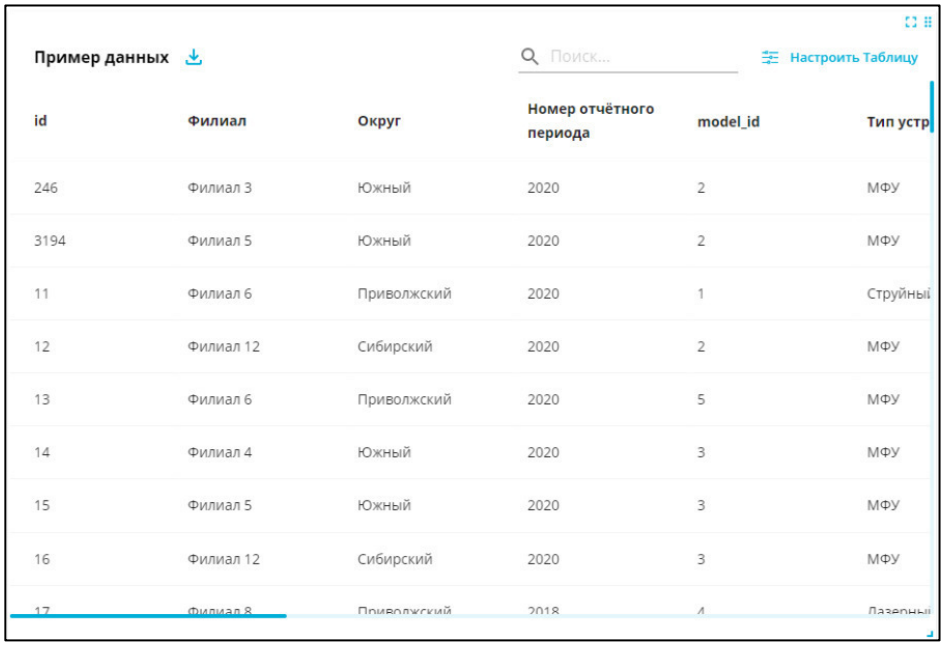


Рисунок 73 Таблица с примером данных

- **Таблица с результатами сэмплирования**.

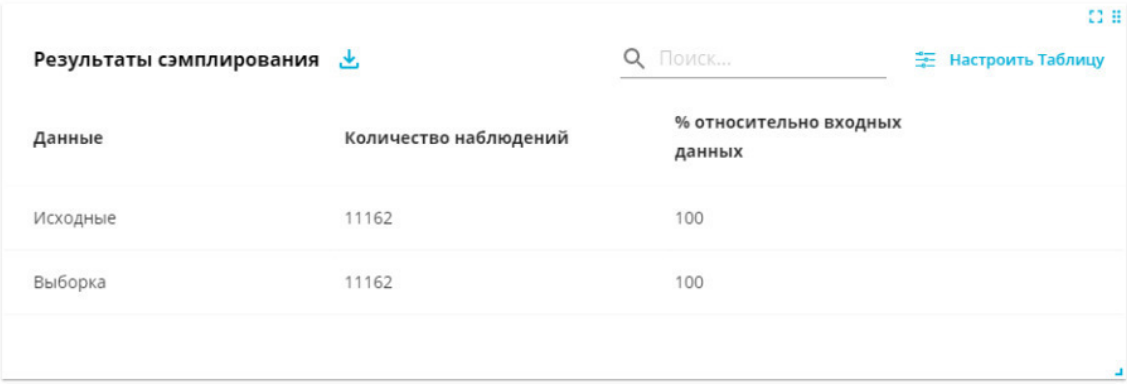


Рисунок 74 Таблица с результатами сэмплирования

3.2.5.6.2. Узел «Разделение выборки»

Узел «Разделение выборки» разбивает набор данных на части: **обучающую** (используемую в процессе обучения модели), **валидационную** (используемую для подбора оптимального набора гиперпараметров модели) и **тестовую** согласно заданным Пользователем пропорциям.

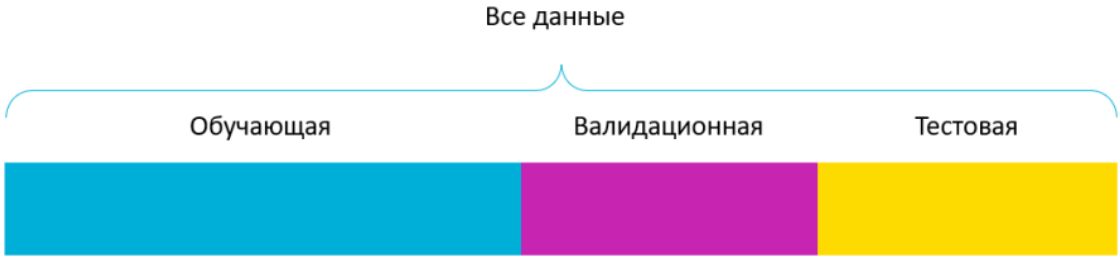


Рисунок 75 Принцип работы узла «Разделение выборки»

Разбиение можно произвести двумя способами:

- **Простая случайная выборка** — все наблюдения имеют одинаковую вероятность быть отобранными.
- **Выборка со стратификацией** — случайный отбор наблюдений производится в пределах каждого класса (т.е. с помощью стратификации можно избежать "непредставительной" выборки (когда в выборку попадают наблюдения только одной страты/класса) что не гарантируется простой случайной выборкой)

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Разделение выборки на части	Ручной ввод доли (в %) для каждой части Сумма долей должна быть равна 100%	Доли обучающей, валидационной и тестовой выборок в исходном наборе данных
Метод разбиения	Раскрывающийся список со следующими значениями: • Simple Random (по умолчанию) • Stratified	Метод разбиения исходного набора данных. Предусмотрены: • **Simple Random** — все наблюдения имеют одинаковый шанс быть отобранными • **Stratified** — случайный отбор наблюдений выполняется в пределах каждого класса (при выборе данного метода появится поле «**Список входных переменных**» для указания переменной, по которой будет проводиться стратификация)

Параметр	Возможные значения и ограничения	Описание
Seed	Ручной ввод числового значения По умолчанию — 12345	Начальное числовое значение для генератора случайных чисел Используется для воспроизведения результатов при повторном запуске

Таблица 9 Параметры узла «Разделение выборки»

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.

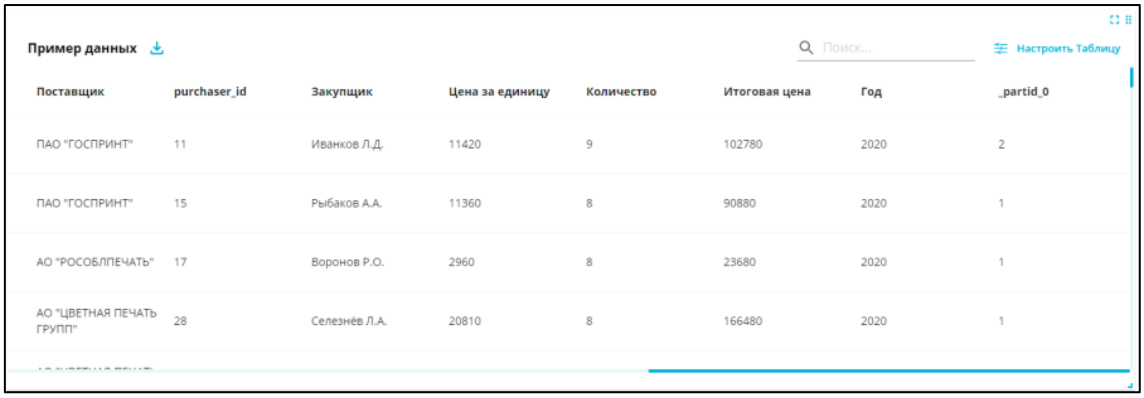


Рисунок 76 Таблица с примером данных

В результате выполнения узла будет рассчитана новая переменная, по которой далее будет производиться разделение набора данных на выборки (переменная **_partid_0**).

- Таблица с указанием долей и количества наблюдений, попавших в соответствующую выборку.

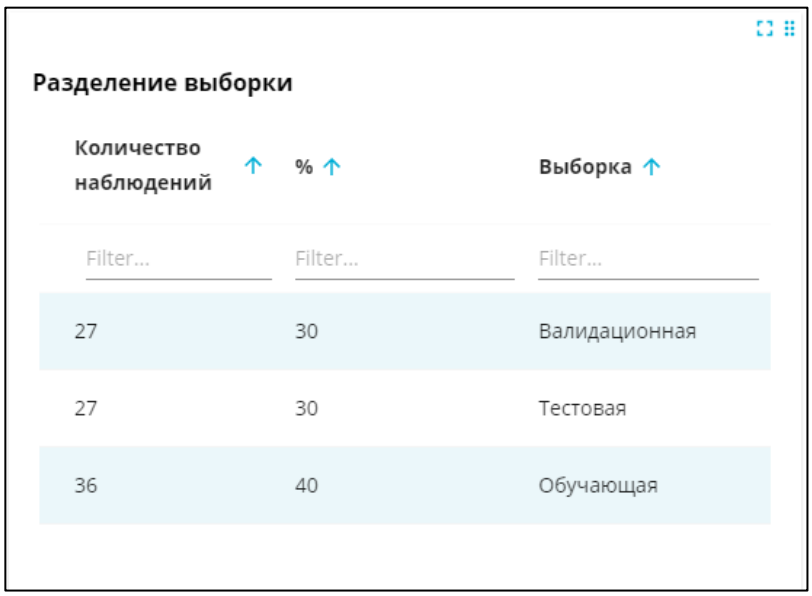


Рисунок 77 Пример таблицы с указанием выборок и количества наблюдений

3.2.5.6.3. Узел «Sample»

Узел «Sample» позволяет сформировать репрезентативную выборку из исходного набора данных, а также скорректировать неравномерное распределение классов. Предусмотрены два метода построения выборки:

- **Простая случайная выборка** — все наблюдения имеют одинаковую вероятность быть отобранными.
- **Выборка со стратификацией** — случайный отбор наблюдений производится в пределах каждого класса (т.е. с помощью стратификации можно избежать "непредставительной" выборки (когда в выборку попадают наблюдения только одной страты/класса) что не гарантируется простой случайной выборкой).

В **задаче классификации** данные называются **несбалансированными**, когда в обучающей выборке доли объектов разных классов существенно различаются.

Для корректировки неравномерного распределения классов предусмотрена **субдискретизация (undersampling)**, которая заменяет больший класс подвыборкой по мощности равной малому классу.

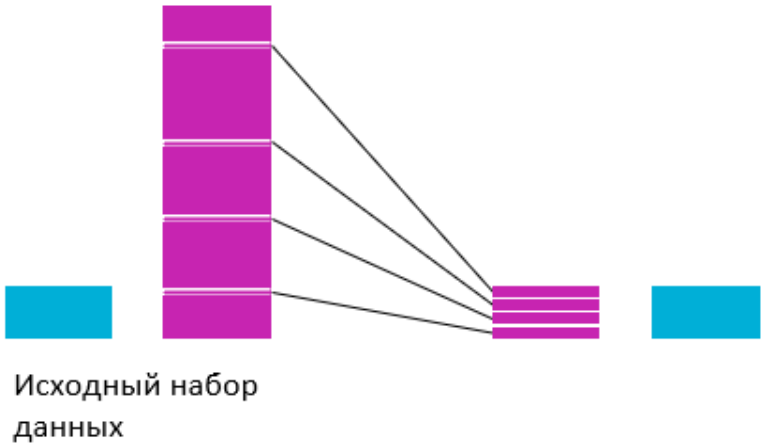


Рисунок 78 Принцип работы метода Undersampling

При выборе чекбокса **Использовать undersampling** параметр **Доля наблюдений в выборке** игнорируется.

Если выбраны **Выборка со стратификацией** и чекбокс **Использовать undersampling**, то стратификация делается отдельно для выборки с целевым классом и отдельно для выборки с прочими классами. При этом может получиться в целом нестратифицированная выборка, т.к. распределения переменных в выборке по целевому классу может отличаться от распределений во всей выборке.

Целевой класс задается **в узле «Метаданные»**.

Если выбрать чекбокс **Использовать Undersampling** и **Долю целевого класса в выборке** задать меньше доли целевого класса во входных данных, то возникнет ошибка. Это связано с тем, что возможна ситуация, когда в выборке должно быть больше наблюдений с прочими классами, чем их есть во всем исходном наборе.

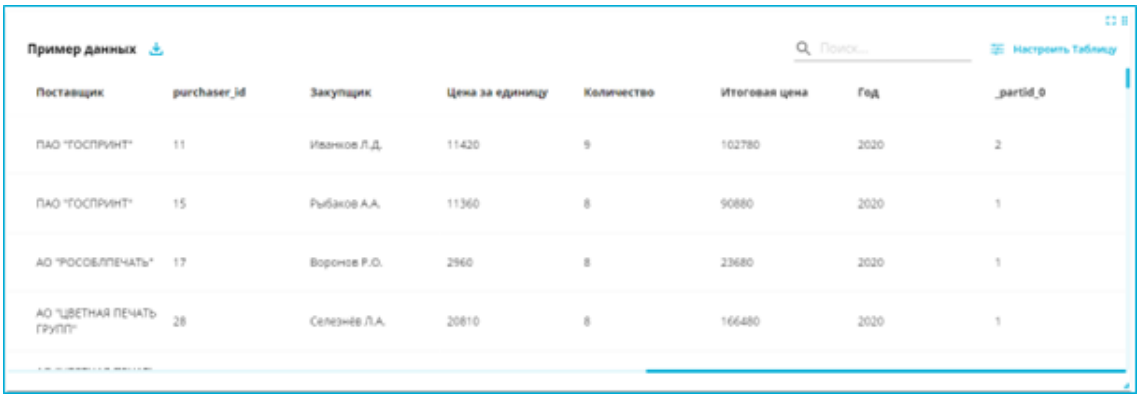
Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значения нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значения нет	Описание узла
Метод построения выборки	Раскрывающийся список со следующими значениями: • Простая случайная выборка • Выборка со стратификацией	Метод разбиения исходного набора данных. Предусмотрены: • **Простая случайная выборка** — все наблюдения имеют одинаковую вероятность быть отобранными • **Выборка со стратификацией** — случайный отбор наблюдений производится в пределах каждого класса
Доля наблюдений в выборке	Ручной ввод числового значения Значение не должно быть больше 1 и меньше или равно 0 По умолчанию — 0,5	Данный параметр задает долю наблюдений, которые попадут в выборку из входных данных
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске
Переменные для стратификации	Поле выбора со списком переменных	В случае выбора Выборки со стратификацией нужно указать переменную, по которой будет проводиться стратификация
Использовать Undersampling	Чекбокс	Выбор данного чекбокса указывает на стратегию сэмплирования — **субдискретизация**, которая удаляет наблюдения из большего класса.
Доля целевого класса относительно входных данных	Ручной ввод числового значения Значение не должно быть больше 1 и меньше или равно 0 По умолчанию — 1	Данный параметр задает долю целевого класса относительно входных данных
Доля целевого класса в выборке	Ручной ввод числового значения Значение не должно быть больше 1 и меньше или равно 0 По умолчанию — 0,2	Данный параметр задает долю целевого класса в выборке

Таблица 10 Параметры узла «Sample»

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.



- Таблица с результатами сэмплирования.

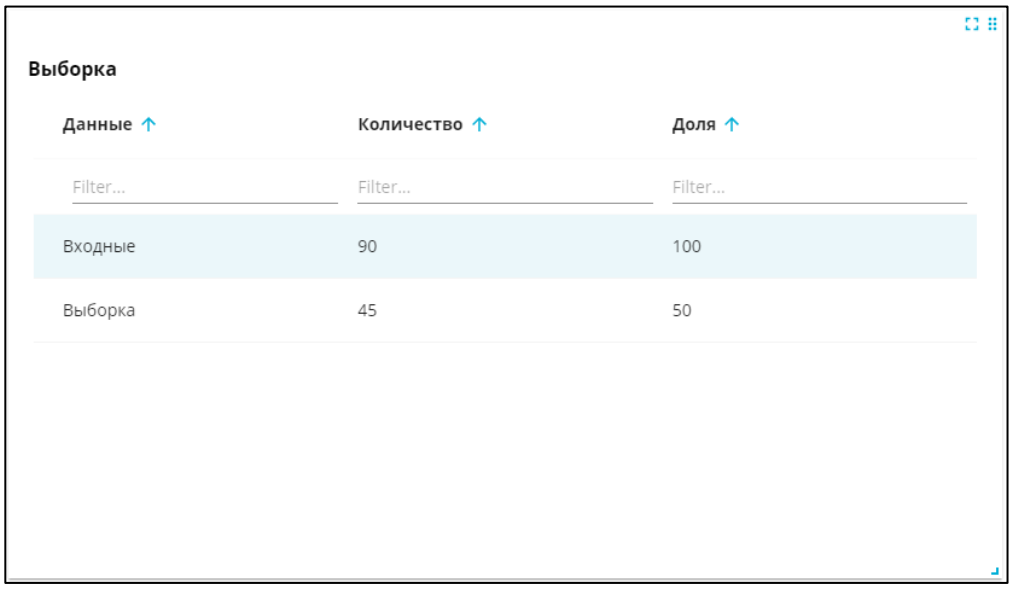


Рисунок 79 Пример таблицы с результатами сэмплирования

Рисунок 80 Пример таблицы с результатами сэмплирования (с undersampling)

- Таблица с результатами стратификации (отображается при выборе метода построения выборки **Выборка со стратификацией**).

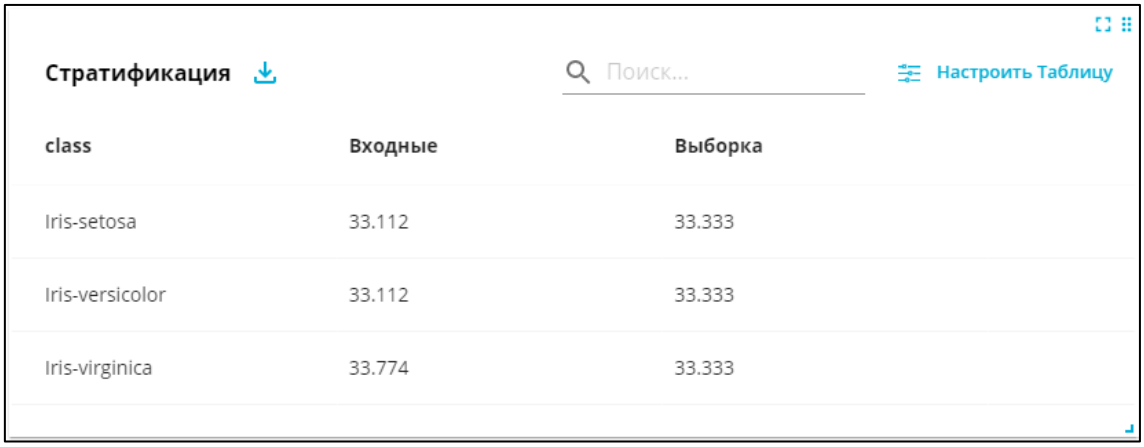


Рисунок 81 Пример таблицы с результатами стратификации

- Таблица с результатами стратификации для целевого класса (отображается при выборе чекбокса **Использовать Undersampling** и **Выборке со стратификацией**).

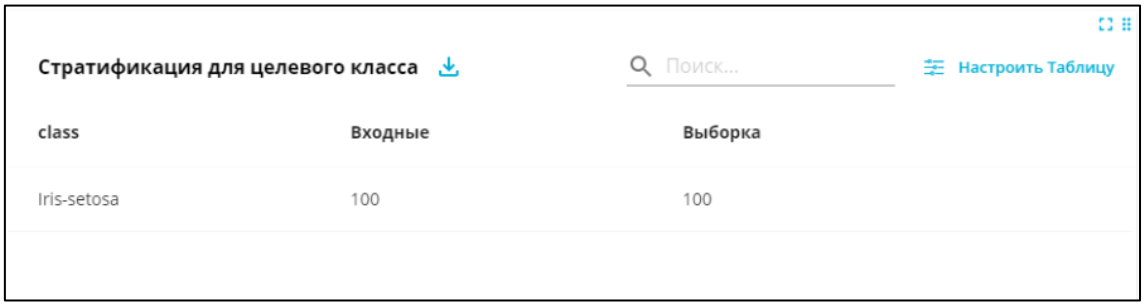


Рисунок 82 Пример таблицы с результатами стратификации для целевого класса

- Таблица с результатами стратификации для прочих классов (отображается при выборе чекбокса **Использовать Undersampling** и **Выборке со стратификацией**).

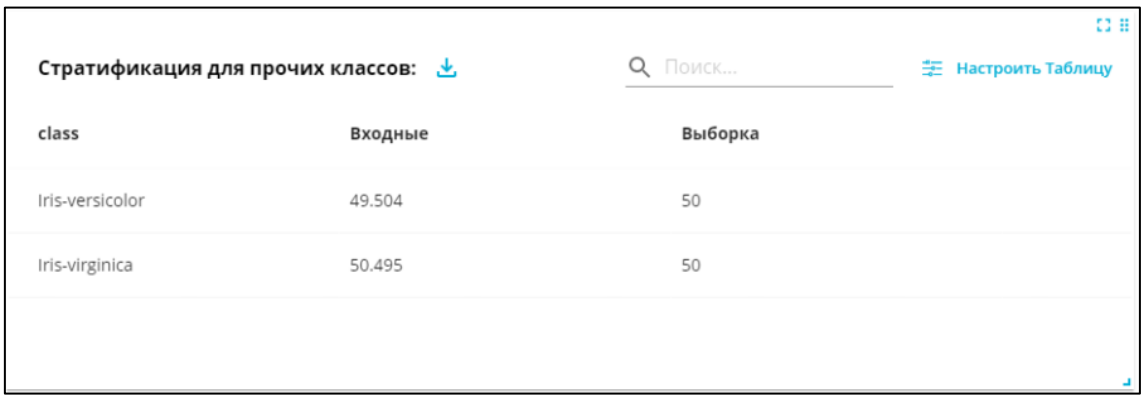


Рисунок 83 Пример таблицы с результатами стратификации для прочих классов

3.2.5.6.4. Узел «Фильтр»

Узел «Фильтр» позволяет по заданным условиям удалить часть наблюдений из процесса моделирования. Отфильтрованные (не соответствующие критериям фильтрации) данные не попадут на вход последующих узлов.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Поле для ввода формулы	Поле ввода условия фильтрации	При выборе поля ввода откроется окно **Формула** (подробное описание ниже).

Таблица 11 Параметры узла «Фильтр»

Окно Формула

В окне **Формула** задается условие отбора наблюдений из исходного набора данных.

Основными элементами окна Формула являются:

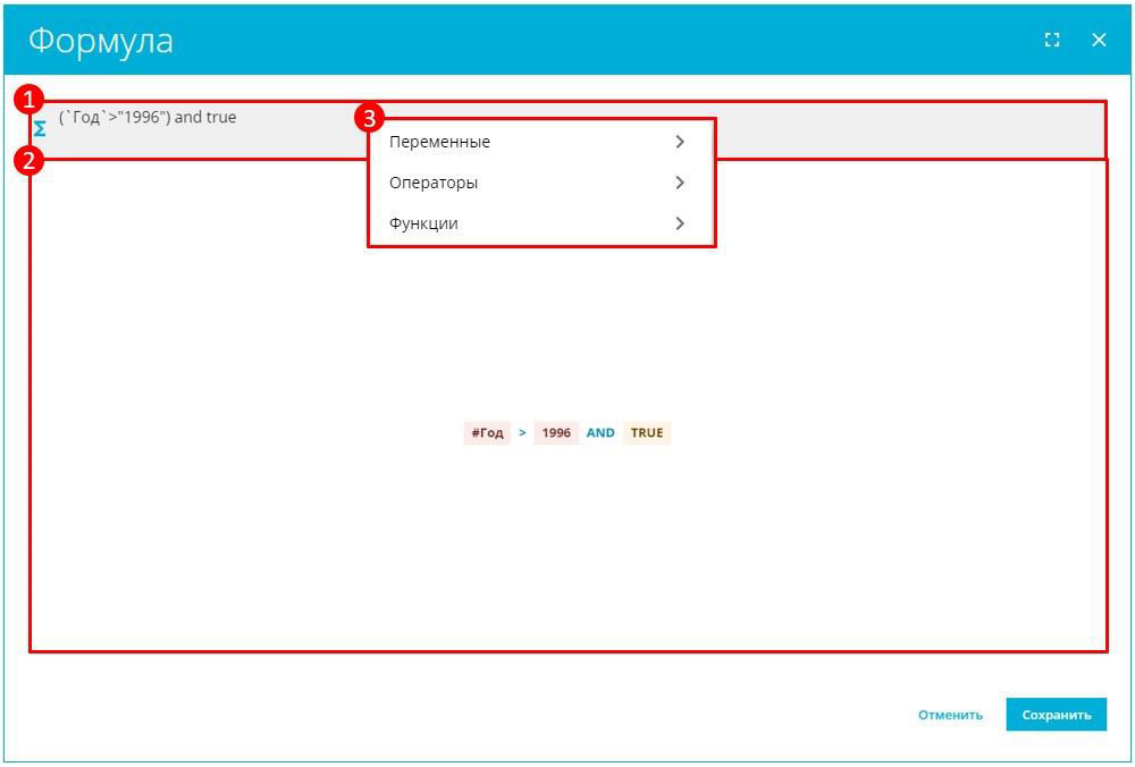


Рисунок 84 Окно Формула

1. Строка ввода текстового представления условия.
2. Рабочее поле, в котором строится графическое представление условия.
3. Вкладка, которая открывается при щелчке правой кнопкой мыши в строке ввода, и включает в себя:
    - Переменные из набора данных.
    - Операторы и функции, представленные в таблице ниже.

Позволяет просматривать переменные набора данных и строить конструкции условий.

Для **задания условия** необходимо:

- В строку текстового представления ввести условие, используя операторы, функции (Таблица 8) и переменные набора данных.
    - Название переменной должно быть указано в обратных одинарных кавычках (` `). Пример: **`Год`** > 1996.
    - В качестве десятичного разделителя используется точка.
    - Строка должна указываться в двойных кавычках. Пример: `Пол` = **"мужской"**.
- Посмотреть список операторов, функций и переменных, а также добавить их в условие можно и в панели, которая открывается при щелчке правой кнопкой мыши по строке ввода.

Список функций и операторов

Оператор/Функция	Описание	Текстовое представление (формула)
And	Логическая операция «И»	true **and** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 and `Пол` = "Женский"
Or	Логическая операция «ИЛИ»	true **or** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 or `Пол` = "Женский"
+	Операция сложения	0 **+** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`+ ` Цена продукта М`
–	Операция вычитания	0 **-** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`- 1000
×	Операция умножения	0 ***** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М` * `Скидка`
/	Операция деления	0 **/** 0

Оператор/Функция	Описание	Текстовое представление (формула)
		Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Выручка за первый квартал` / `Выручка за год`
<	Меньше	"" **<** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` < 32
≤	Меньше или равно	"" **<=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` <= 120
>	Больше	"" **>** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` > "32"
≥	Больше или равно	"" **>=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` >= 120
=	Равно	"" **=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` = 120
≠	Не равно	"" **!=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` != 120
like	Проверяет, удовлетворяет ли символьная строка заданному образцу, который может содержать поисковые символы. Учитывает регистр	"" **like** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` like "Москва"
ilike	Проверяет, удовлетворяет ли символьная строка заданному образцу, который может содержать поисковые символы.	"" **ilike** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` ilike "Москва"
startswith	Проверяет, есть ли в начале одной текстовой строки другая текстовая строка	"" **startswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` startswith "Сан"
endswith	Проверяет, есть ли в конце одной текстовой строки другая текстовая строка	"" **endswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` endswith "бург"
contains	Проверяет, встречается ли указанная строка внутри другой строки	"" **contains** "" Вместо "" нужно вставить переменную/значение

Оператор/Функция	Описание	Текстовое представление (формула)
		Пример: `Название` contains "consulting"
between	Проверяет, входит ли значение в заданный диапазон	0 **between**(0,0) Вместо 0 нужно вставить числовую переменную/значение **Пример**: `ID` between (1, 100500)
in	Проверяет наличие элемента в последовательности	"" **in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` in ("Москва", "Московская область")
not in	Проверяет отсутствие элемента в последовательности	"" **not_in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` not_in ("Москва", "Московская область")
coalesce	Возвращает данные из первого столбца, содержащего значение, отличное от NULL	**coalesce**("") Вместо "" нужно вставить переменную/значение **Пример**: coalesce(`ProductNumber`, `ProductName`)
not	Задает противоположное условие	not(true) Вместо true нужно вставить условие **Пример**: not(`Год` > 1996)
create_date	Создает переменную даты из последовательно введенных года, месяца и дня	create_date(0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_date(2016,10,25)
current_date	Возвращает текущую дату	current_date() Вводится без дополнительных параметров
current_timestamp	Возвращает текущие дату и время	current_timestamp() Вводится без дополнительных параметров
now	Возвращает текущие дату и время	now() Вводится без дополнительных параметров
create_datetime	Создает переменную даты времени из последовательно введенных года месяца дня часа минут и секунд	create_datetime(0,0,0,0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_datetime(2016,10,25,12,18,0)
char_length	Возвращает длину строки	char_length("") Вместо "" нужно вставить строку/строковую переменную **Пример**: char_length(`Код_продукта`)
random	Возвращает случайное число	random() Вводится без дополнительных параметров
In	Натуральный логарифм	ln(0) Вместо 0 нужно вставить числовое значение/числовую переменную

Оператор/Функция	Описание	Текстовое представление (формула)
		Пример: ln(`Числовая переменная`)
exp	Экспонента	exp(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: exp (`Числовая переменная`)
power	Возведение в степень	power(0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: power(`Числовая переменная`,2)
sqrt	Квадратный корень	sqrt(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: sqrt (`Числовая переменная`)
abs	Абсолютное значение	abs(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: abs (`Числовая переменная`)
ceil	Возвращает наименьшее целое число, которое больше или равно текущему значению	ceil(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: ceil(25.1) Вернет значение 26
floor	Возвращает наибольшее целое число, которое меньше или равно текущему значению	floor(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: floor(25.1) Вернет значение 25
extract_from_datetime	Получает указанную часть (день, месяц, год, час, минута, секунда) из значения даты и времени	extract_from_datetime("","") Вместо первых кавычек нужно указать необходимую часть для извлечения: · SECOND · MINUTE · HOUR · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа datetime **Пример:** extract_from_datetime("DAY",`datetime`)
extract_from_date	Получает указанную часть (день, месяц, год) из значения даты	extract_from_date("","") В первые кавычки нужно вписать необходимую часть для извлечения: · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа date **Пример:** extract_from_date("DAY",`date`)

Оператор/Функция	Описание	Текстовое представление (формула)
concat	Объединяет в единую строку указанные значения	concat("") Вместо "" нужно вставить строки/строковые переменные **Пример:** concat (`Фамилия`, " ", `Имя`)

Таблица 12 Операторы и функции

Результаты выполнения узла:

- Таблица с примером отфильтрованного набора данных. Отображаются первые 100 наблюдений.

Рисунок 85 Пример таблицы с отфильтрованным набором данных (условие id <10)

3.2.5.6.5. Узел «Метаданные»

Узел «Метаданные» позволяет изменить метаданные переменных.

К метаданным переменной относятся **Роль**, **Тип** и **Целевой класс**.

В Модуле предусмотрены следующие **Роли** переменной:

- **Target** — **Целевая переменная** — зависимая переменная, используемая в качестве цели моделирования.
- **Predictor** — **Предиктор** — независимая переменная, используемая в качестве прогнозирующей целевую переменную.
- **Excluded** — **Исключен** — данная роль позволяет исключить переменную из процесса моделирования. Вернуть ее обратно в процесс можно также при помощи узла «Метаданные», установленного далее в сценарии.
- **Weight** — **Вес**— данная роль задается переменной, которая указывает модели на важность наблюдений. Пример: для прогноза могут быть важны тенденции за последний период (например, сезон).
- **Cluster_ID** — **Сегмент** — переменная с данной ролью рассчитывается при решении задачи кластеризации.

- **Prediction** — **Прогноз** — переменная с данной ролью рассчитывается при решении задачи регрессии.
- **Classification** — **Прогноз класс** — переменная с данной ролью рассчитывается при решении задачи классификации.
- **Partition** — **Разделение** — переменная с данной ролью задает разделение набора данных на обучающую, валидационную и тестовую выборки (рассчитывается в узле «Разделение выборки»).
- **ClassWeight** — **Вес класса** — переменная с данной ролью корректирует несбалансированность классов (задается в узле «Веса классов»).

Первоначально при выборе набора данных **в узле «Набор данных»** все атрибуты имеют роль **Предиктор**.

Предусмотрены следующие **Типы** переменной:

- **Номинальный (Nominal)** — дискретные значения без числовой связи между категориями.
- **Порядковый (Ordinal)** — дискретные значения, которые можно ранжировать или сортировать (Пример — хорошо, великолепно, превосходно).
- **Двоичный (Binary)** — это дискретные данные, которые могут относиться только к одной из двух категорий: 0 и 1.
- **Интервальный (Interval)** — числовые значения.

Целевой класс используется в процессе моделирования (например, при стратификации) и может быть указан только для **Целевой переменной**.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Изменение метаданных	Кнопка	При выборе кнопки откроется окно **Изменение метаданных.**

Таблица 13 Параметры узла «Метаданные»

Окно Изменение метаданных

В окне **Изменение метаданных** Пользователь имеет возможность изменить **Роль**, **Тип** переменной и указать **Целевой класс**. Для этого необходимо:

- Рядом с интересующей переменной выбрать иконку  .
- В выпадающих меню выбрать **Роль** и/или **Тип**, указать **Целевой класс** (доступно только для переменной типа **Целевая переменная** –**Target**).
- Сохранить изменения, выбрав иконку  .

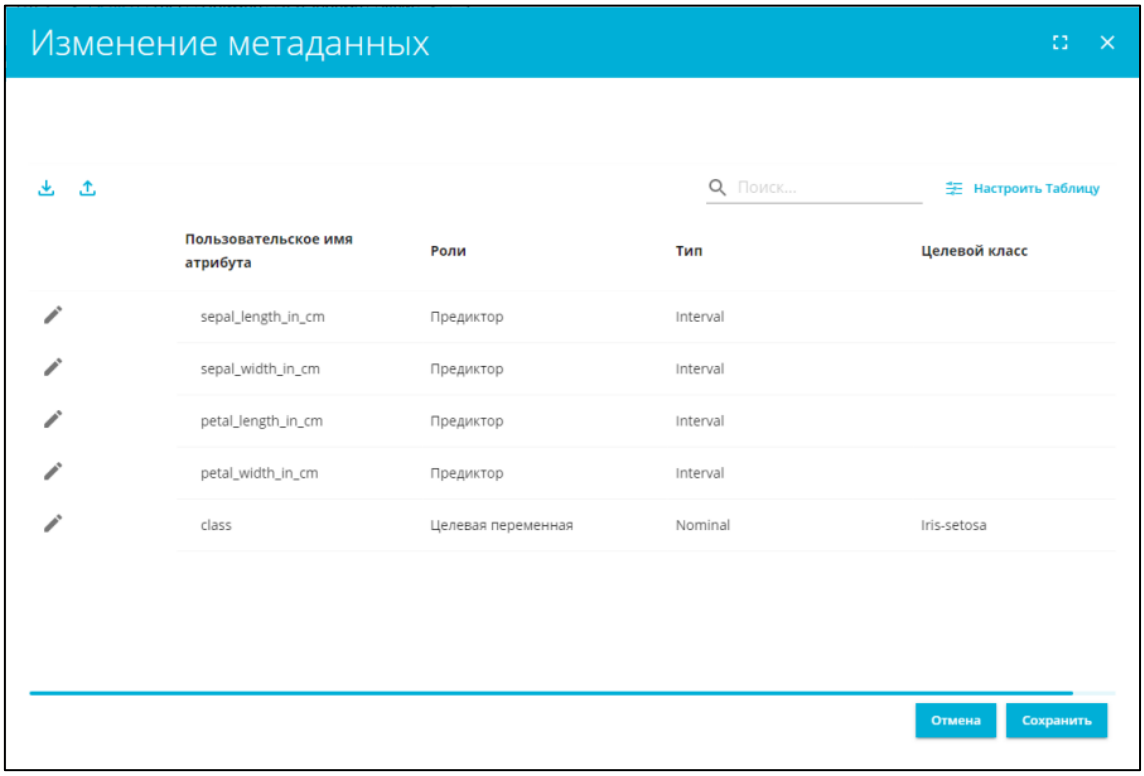


Рисунок 86 Окно Изменение метаданных

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.

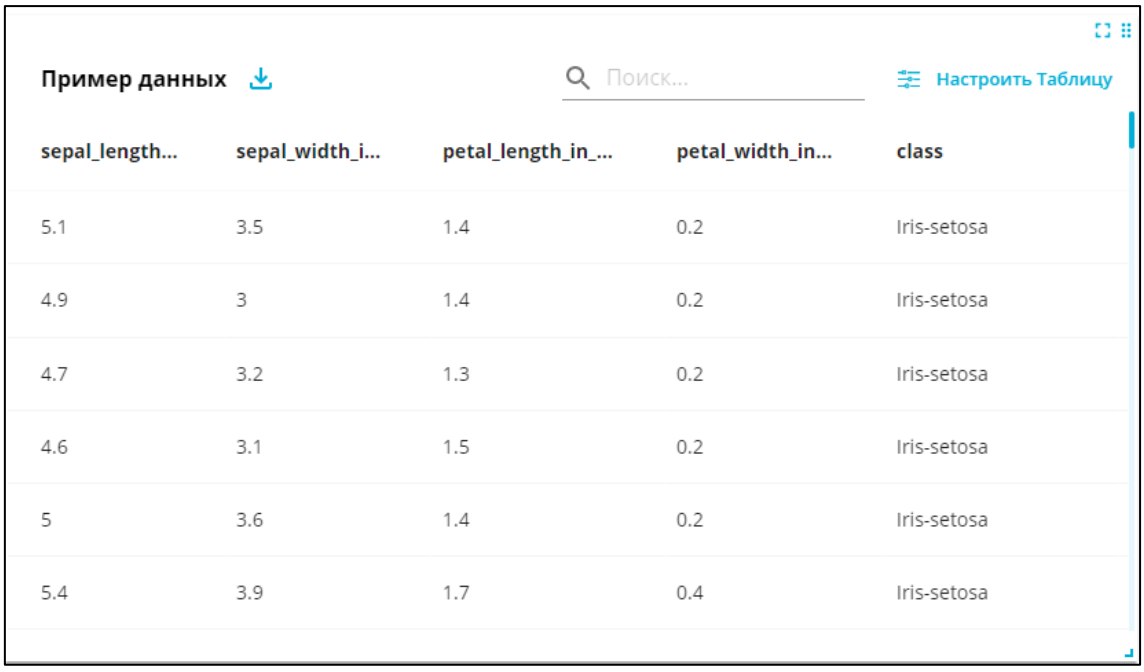


Рисунок 87 Таблица с примером данных

3.2.5.6.6. Узел «One-hot encoding»

Узел «One-hot encoding» преобразует категориальные данные в числовую форму.

Многие алгоритмы не могут напрямую работать с категориальными переменными. Для этого предусмотрен метод **One-hot encoding,** который преобразует категориальные данные в числовую форму. Для этого создаются дополнительные столбцы-индикаторы наличия/отсутствия категории с помощью значений 1 или 0 соответственно. Таким образом, если категориальная переменная имеет k возможных значений, то на выходе получится k столбцов для ее представления. Алгоритмы машинного обучения могут принимать эти столбцы в качестве входных данных.

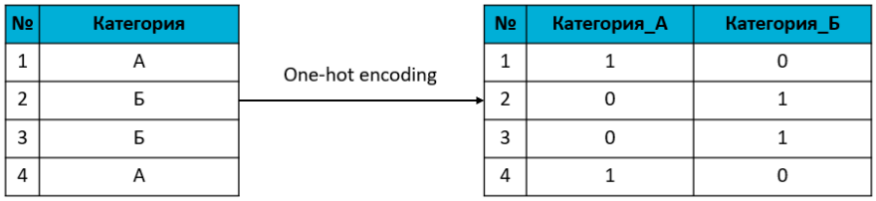


Рисунок 88 Принцип работы узла «One-hot encoding»

В кодировании One-Hot есть некоторая избыточность. Например, переменная «Пол» может принимать два значения - мужчина или женщина. При кодировании достаточно использовать в качестве предиктора лишь одну из этих двух фиктивных переменных. Для этого

Пользователю нужно выбрать чекбокс **Исключить первую категорию** или вручную указать исключаемую переменную в окне **Выбор категории**.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Исключить первую категорию	Чекбокс	Выбор данного чекбокса указывает методу на необходимость исключить первую категорию
Исключить оригинальную переменную	Чекбокс	Выбор данного чекбокса указывает методу на необходимость исключить оригинальную переменную, на основе которой были вычислены фиктивные, из дальнейшего процесса моделирования. В окне **Выходные переменные** Роль данного атрибута изменится на Исключен (Excluded) (про Роли переменной подробнее Узел «Метаданные»).
Выбор категории	Кнопка	При выборе кнопки откроется окно **Выбор категории.**

Таблица 14 Параметры узла «One-hot encoding»

Окно Выбор категории

В окне **Выбор категории** Пользователь имеет возможность выбрать категориальную переменную для кодирования (Рисунок 83). Для этого необходимо:

- Рядом с интересующей переменной выбрать иконку .
- В столбце **Выбрать** выбрать чекбокс и при необходимости указать исключаемые из кодирования значения (перечислить через запятую).
- Сохранить изменения, выбрав иконку .

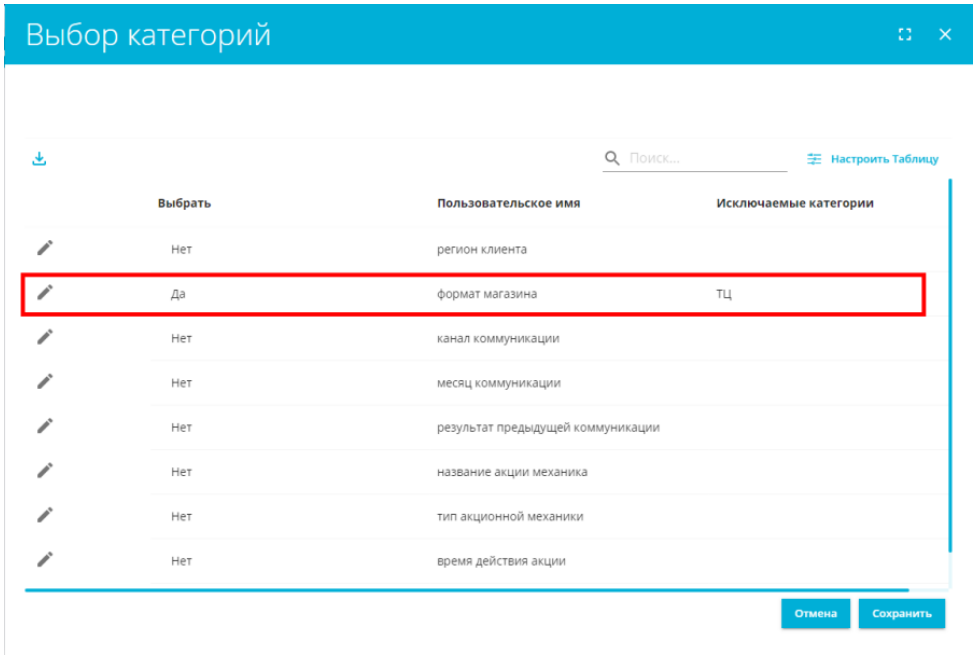


Рисунок 89 Окно Выбор категории

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.

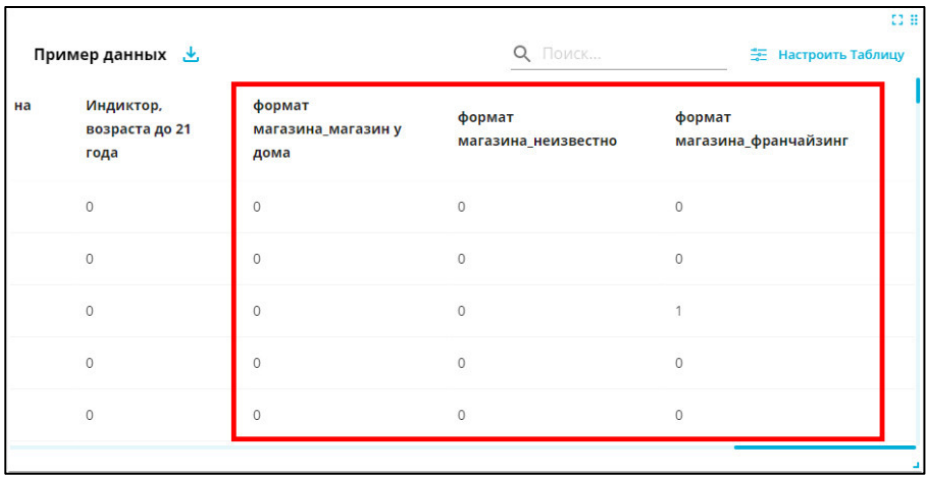


Рисунок 90 Таблица с примером посчитанных трех фиктивных переменных

- Таблица с результатами кодирования.

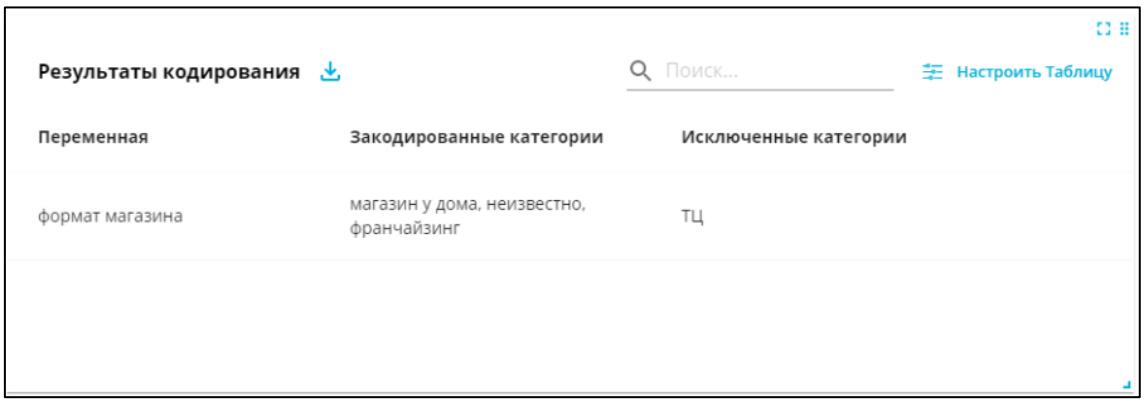


Рисунок 91 Пример таблицы с результатами кодирования

В результате выполнения в выходных параметрах узла появятся столбцы-индикаторы с ролью первоначального столбца.

3.2.5.6.7. Узел «Заполнение пропусков»

Узел «Заполнение пропусков» обрабатывает пропущенные значения.

В зависимости от задачи Пользователь может использовать тот или иной метод заполнения отсутствующих элементов. Заменить пропуски можно на:

- **Моду** – наиболее часто встречающееся значение (подходит для категориальных переменных).
- **Константу** – выбранное Пользователем конкретное значение.
- **Среднее** – находится суммированием всех чисел в выборке и делением полученной суммы на количество чисел.
- **Медиану** – если взять все элементы множества и отсортировать, то это число делит множество пополам. Одна половина множества равна или больше этого числа, а другая меньше или равна этому числу.
- **Минимум**.
- **Максимум**.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Создать общий индикатор	Чекбокс	Выбор данного чекбокса указывает на необходимость расчета общего для всех переменных набора данных индикатора

Параметр	Возможные значения и ограничения	Описание
пропущенных значений		пропущенных значений (столбца, в котором будет указано наличие или отсутствие пропущенных значений в наблюдении). После успешного выполнения узла созданный индикатор можно найти по ссылке **«Выходные параметры».** Ему будет назначена роль **Предиктора** с типом ***Binary***.
Создать индивидуальные индикаторы пропущенных значений	Чекбокс	Выбор данного чекбокса указывает на необходимость расчета индивидуальных индикаторов пропущенных значений для выбранной далее переменной (для каждой из выбранных переменных будет посчитан свой столбец). После успешного выполнения узла созданный индикатор можно найти по ссылке **«Выходные параметры».** Ему будет назначена роль **Предиктора** с типом ***Binary***.
Максимальная доля пропусков	Ручной ввод	Данный параметр указывает максимальную долю пропусков
Выбор переменных для заполнения пропусков	Кнопка	При выборе кнопки откроется окно **Выбор переменных для заполнения пропусков.**

Таблица 15 Параметры узла «Заполнение пропусков»

Окно Выбор переменных для заполнения пропусков

В окне **Выбор переменных для заполнения пропусков** Пользователь может задать метод замены пропущенных данных (Рисунок 86). Для этого необходимо:

- Рядом с интересующей переменной выбрать иконку .
- В столбце **Выбрать** нажать на чекбокс и при необходимости изменить **Стратегию заполнения**.
    - При выборе **Стратегии заполнения** константой нужно указать значение в соответствующем столбце **Константа**.
- Сохранить изменения, выбрав иконку .

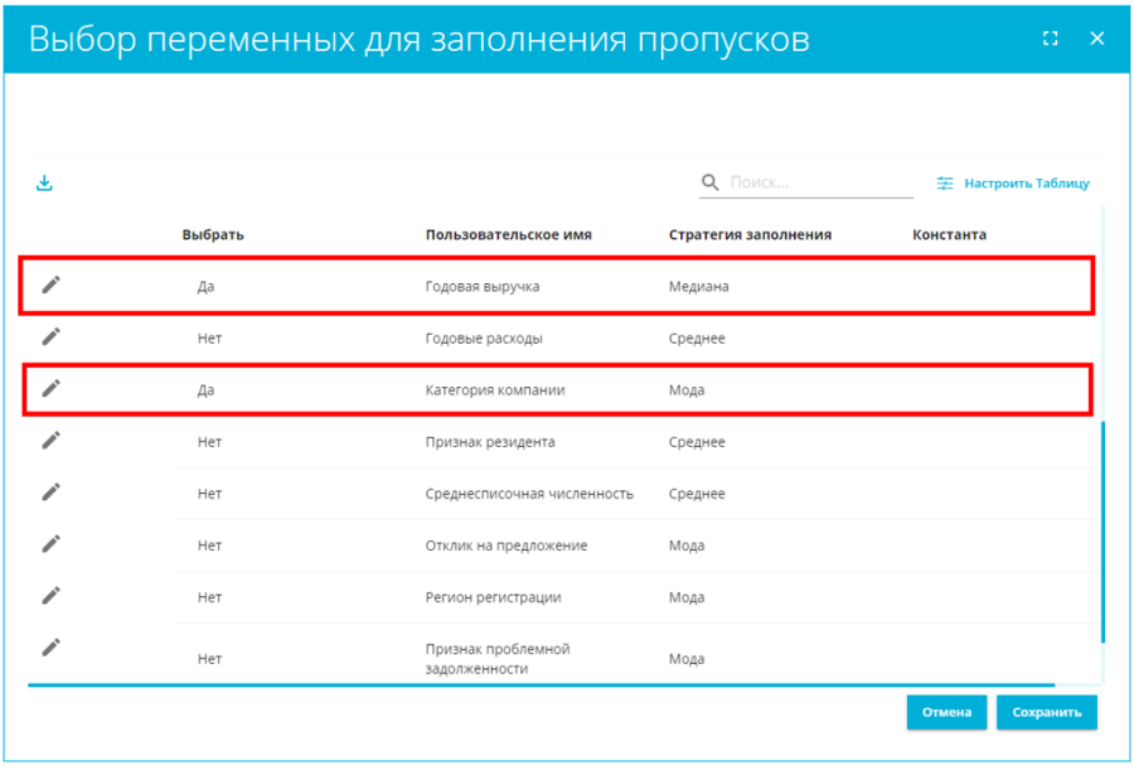


Рисунок 92 Окно Выбор переменных для заполнения пропусков

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.

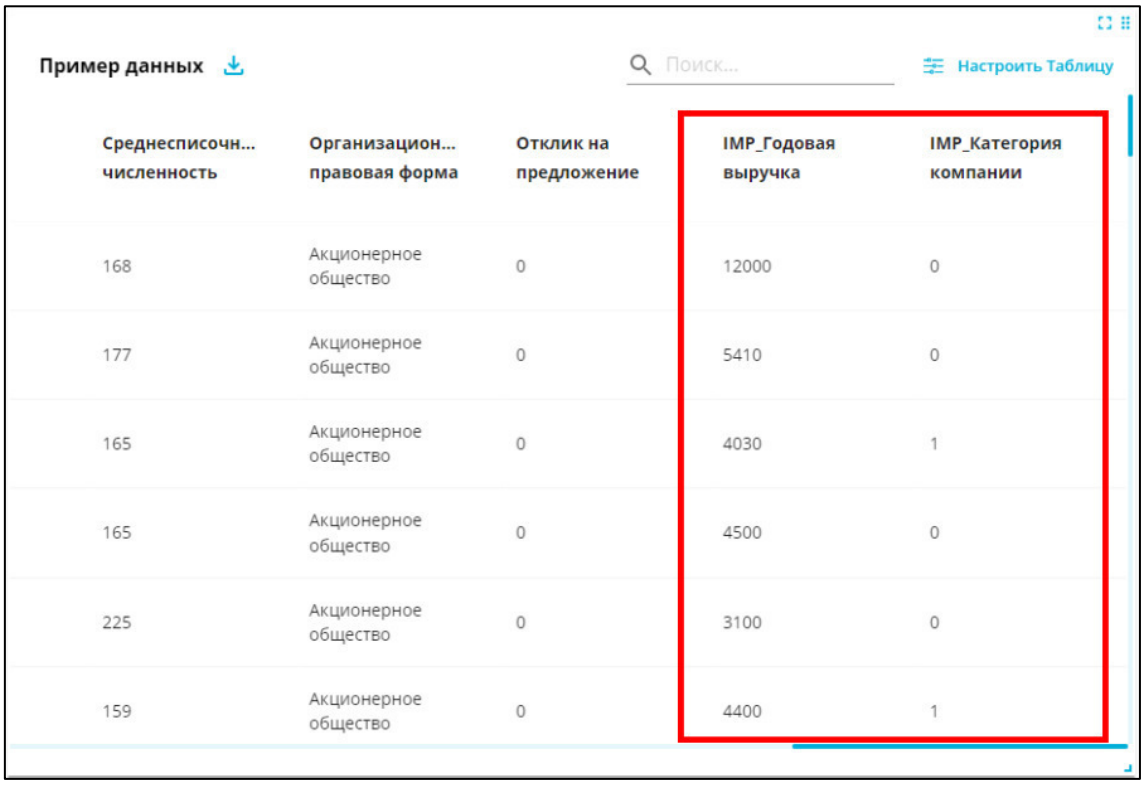


Рисунок 93 Окно Выбор переменных для заполнения пропусков

В результате выполнения узла в наборе данных будут рассчитаны новые переменные с заполненными пропущенными значениями (с префиксом **IMP_**).

- Таблица с количеством пропусков по каждой переменной набора данных.

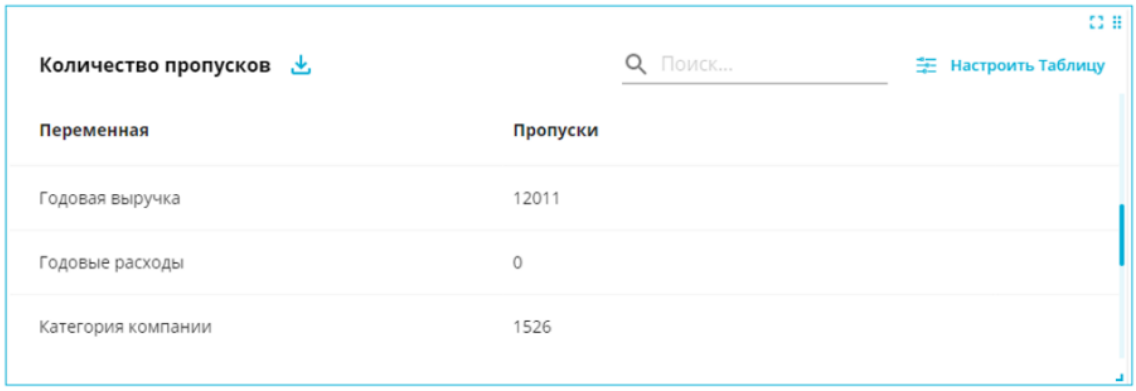


Рисунок 94 Пример таблицы с количеством пропущенных значений по каждой переменной

- Таблица замененных значений.

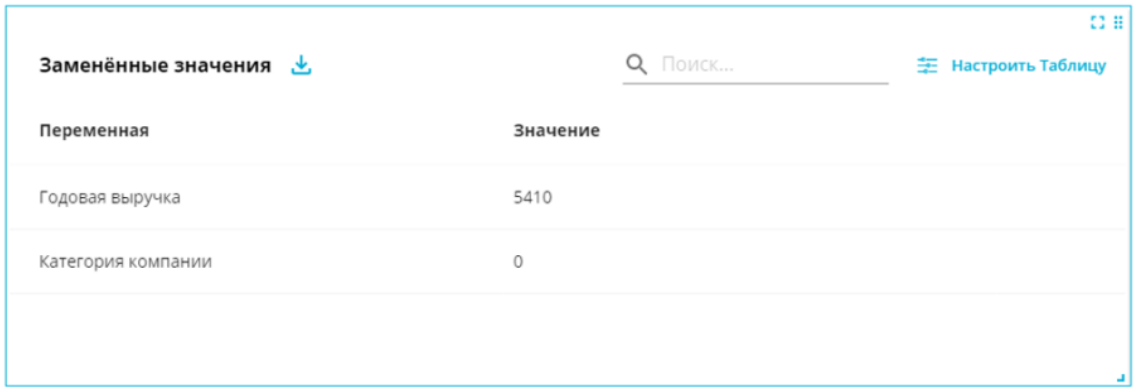


Рисунок 95 Пример таблицы с заменными значениями

3.2.5.6.8. Узел «Трансформация»

Узел «Трансформация» позволяет рассчитать новые переменные.

Пользователь может получить дополнительные сведения для моделирования, выполнив вычисления на основе имеющихся переменных. Для этого в узле «Трансформация» предусмотрено создание вычисляемых показателей при помощи расчетных формул.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Расчетные атрибуты	Кнопка	При выборе кнопки откроется окно **Расчетные атрибуты.**

Таблица 16 Параметры узла «Трансформация»

Окно Расчетные атрибуты

В окне **Расчетные атрибуты** Пользователь имеет возможность создать новые переменные. Для этого необходимо:

- Выбрать иконку + .
- Задать следующие параметры:
  - Имя атрибута.
  - Формулу. Для этого нужно нажать на поле ввода, откроется окно **Формула** в котором и задается формула расчета.
  - Тип (подробнее в разделе Входные и выходные данные узла). Задается автоматически после задания расчетной формулы.
  - Роль (подробнее в разделе Входные и выходные данные узла).
- Сохранить изменения, выбрав иконку .

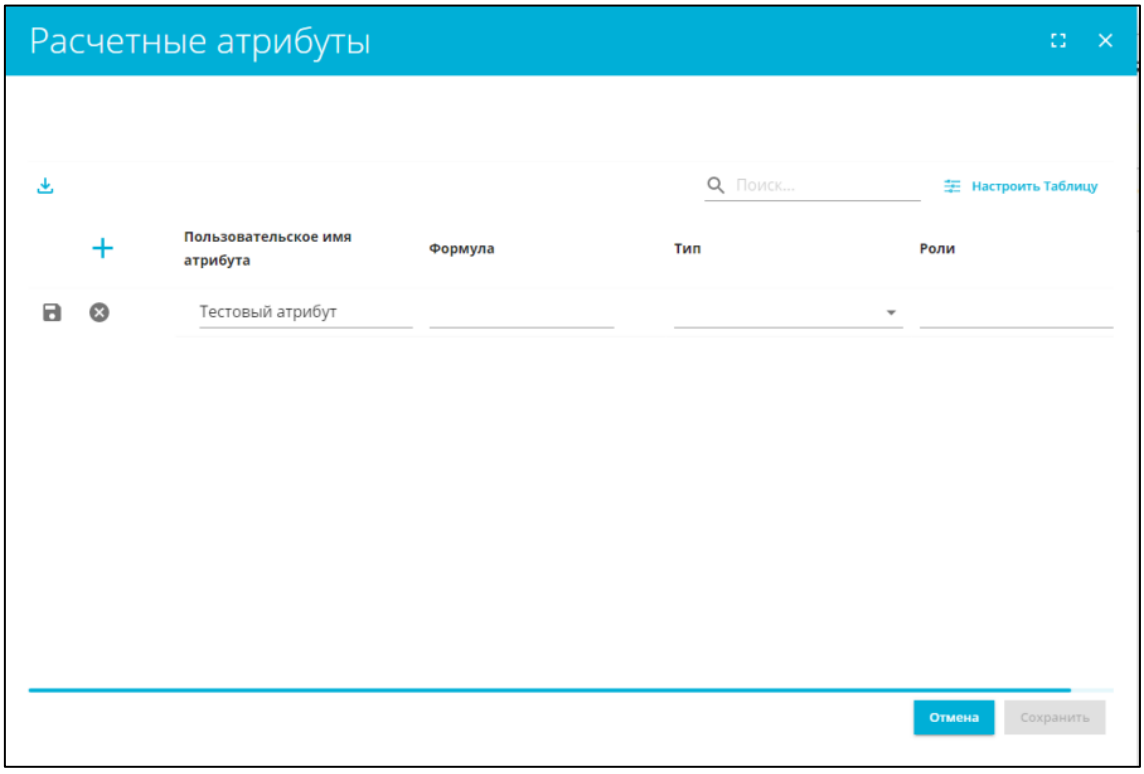


Рисунок 96 Окно Расчетные атрибуты

Окно Формула

В окне **Формула** задается расчетная формула.

Основными элементами окна Формула являются:

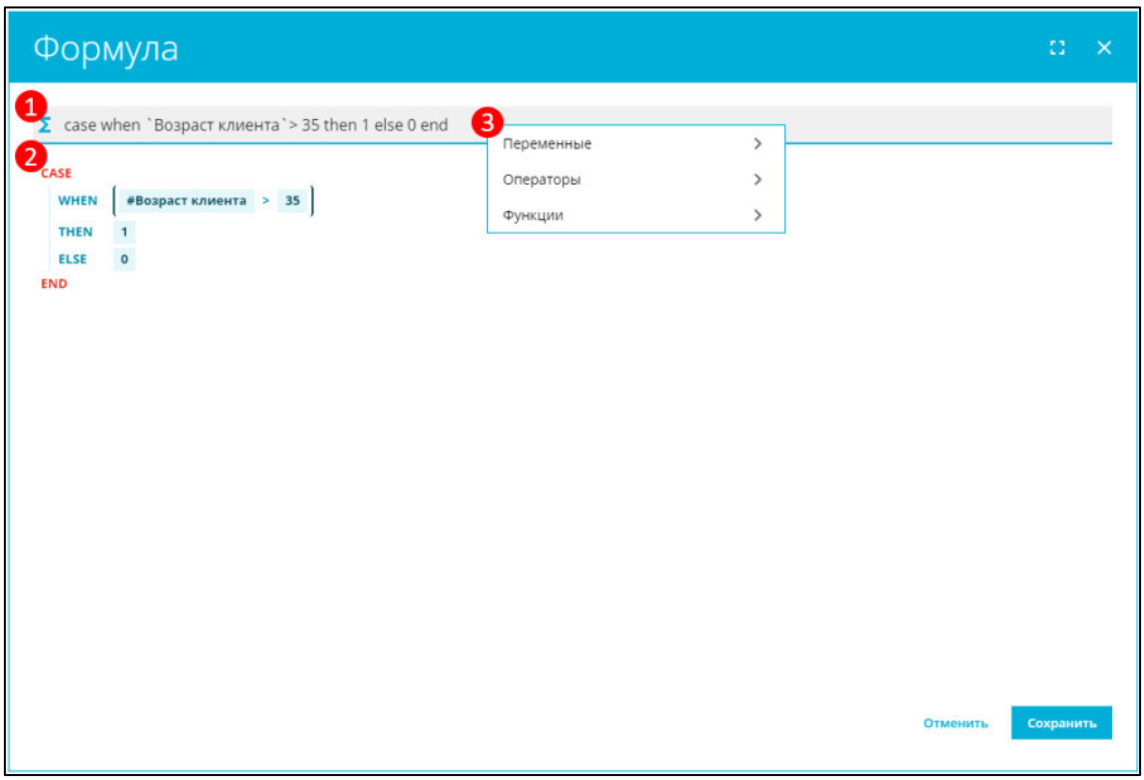


Рисунок 97 Окно Формула

1. Строка ввода текстового представления условия.
2. Рабочее поле, в котором строится графическое представление условия.
3. Вкладка, которая открывается при щелчке правой кнопкой мыши в строке ввода, и включает в себя:
    - Переменные из набора данных.
    - Операторы и функции, представленные в таблице ниже.

Позволяет просматривать переменные набора данных и строить расчетные формулы.

Для **задания условия** необходимо:

- В строку текстового представления ввести формулу, используя операторы, функции и переменные набора данных.
    - Название переменной должно быть указано в обратных одинарных кавычках (` `). Пример: **`Год`** > 1996.
    - В качестве десятичного разделитель используется точка.
    - Строка должна указываться в двойных кавычках. Пример: `Пол` = **"мужской"**.
- Посмотреть список операторов, функций и переменных, а также добавить их в условие можно и в панели, которая открывается при щелчке правой кнопкой мыши по строке ввода.

Оператор/Функция	Описание	Текстовое представление (формула)
And	Логическая операция «И»	true **and** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 and `Пол`= "Женский"
Or	Логическая операция «ИЛИ»	true **or** true Вместо true нужно вставить условие **Пример**: `Год` > 1996 or `Пол`= "Женский"
+	Операция сложения	0 **+** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`+ ` Цена продукта М`
–	Операция вычитания	0 **-** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М`- 1000
×	Операция умножения	0 ***** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Цена продукта М` * `Скидка`
/	Операция деления	0 **/** 0 Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: `Выручка за первый квартал` / `Выручка за год`
<	Меньше	"" **<** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` < 32
≤	Меньше или равно	"" **<=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` <= 120
>	Больше	"" **>** "" Вместо "" нужно вставить переменную/значение **Пример**: `Возраст` > “32”
≥	Больше или равно	"" **>=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` >= 120
=	Равно	"" **=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` = 120
≠	Не равно	"" **!=** "" Вместо "" нужно вставить переменную/значение **Пример**: `Температура` != 120
like	Проверяет, удовлетворяет ли символьная строка	"" **like** "" Вместо "" нужно вставить переменную/значение

Оператор/Функция	Описание	Текстовое представление (формула)
	заданному образцу, который может содержать поисковые символы. Учитывает регистр	**Пример**: `Город` like "Москва"
ilike	Проверяет, удовлетворяет ли символьная строка заданному образцу, который может содержать поисковые символы.	"" **ilike** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` ilike "Москва"
startswith	Проверяет, есть ли в начале одной текстовой строки другая текстовая строка	"" **startswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` startswith "Сан"
endswith	Проверяет, есть ли в конце одной текстовой строки другая текстовая строка	"" **endswith** "" Вместо "" нужно вставить переменную/значение **Пример**: `Город` endswith "бург"
contains	Проверяет, встречается ли указанная строка внутри другой строки	"" **contains** "" Вместо "" нужно вставить переменную/значение **Пример**: `Название` contains "consulting"
between	Проверяет, входит ли значение в заданный диапазон	0 **between**(0,0) Вместо 0 нужно вставить числовую переменную/значение **Пример**: `ID` between (1, 100500)
in	Проверяет наличие элемента в последовательности	"" **in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` in ("Москва", "Московская область")
not in	Проверяет отсутствие элемента в последовательности	"" **not_in**("") Вместо "" нужно вставить переменную/значение **Пример**: `Регион` not_in ("Москва", "Московская область")
CASE	Позволяет проверить несколько условий и выполнить разные операции на основе этих условий	**case when** true **then** "" **else** "" **end** Вместо true нужно вставить условие, вместо "" – значение Для проверки дополнительных условий необходимо добавить дополнительные «**when** true **then** ""» **Пример**: case when `Возраст` > 32 then 0 else 1 end
coalesce	Возвращает данные из первого столбца, содержащего значение, отличное от NULL	**coalesce**("") Вместо "" нужно вставить переменную/значение **Пример**: coalesce(`ProductNumber`, `ProductName`)

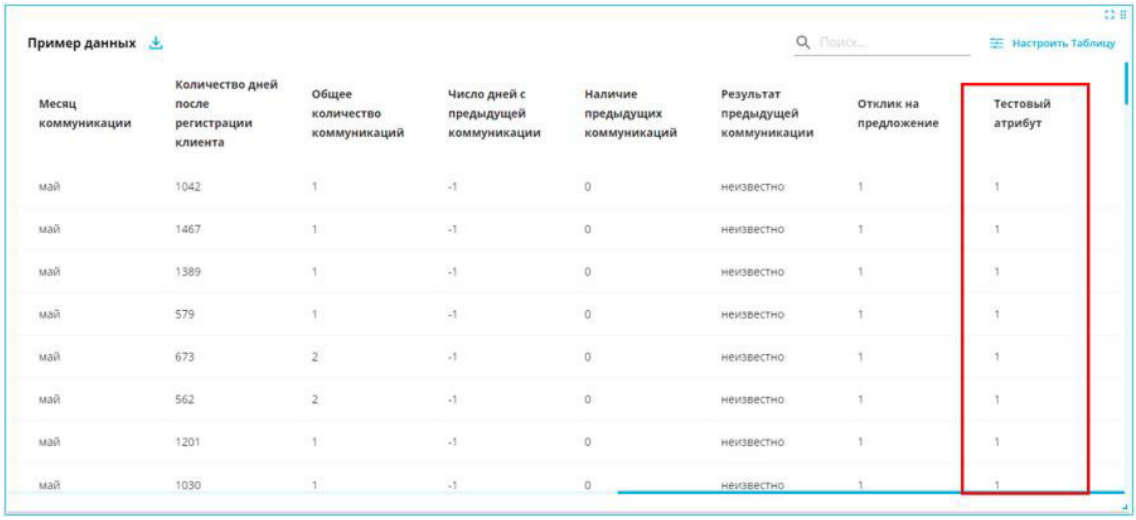
Оператор/Функция	Описание	Текстовое представление (формула)
not	Задает противоположное условие	not(true) Вместо true нужно вставить условие **Пример**: not(`Год` > 1996)
create_date	Создает переменную даты из последовательно введенных года, месяца и дня	create_date(0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_date(2016,10,25)
current_date	Возвращает текущую дату	current_date() Вводится без дополнительных параметров
current_timestamp	Возвращает текущие дату и время	current_timestamp() Вводится без дополнительных параметров
now	Возвращает текущие дату и время	now() Вводится без дополнительных параметров
create_datetime	Создает переменную даты времени из последовательно введенных года месяца дня часа минут и секунд	create_datetime(0,0,0,0,0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: create_datetime(2016,10,25,12,18,0)
char_length	Возвращает длину строки	char_length("") Вместо "" нужно вставить строку/строковую переменную **Пример**: char_length(`Код_продукта`)
random	Возвращает случайное число	random() Вводится без дополнительных параметров
In	Натуральный логарифм	ln(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: ln(`Числовая переменная`)
exp	Экспонента	exp(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: exp (`Числовая переменная`)
power	Возведение в степень	power(0,0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: power(`Числовая переменная`,2)
sqrt	Квадратный корень	sqrt(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: sqrt (`Числовая переменная`)
abs	Абсолютное значение	abs(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: abs (`Числовая переменная`)

Оператор/Функция	Описание	Текстовое представление (формула)
ceil	Возвращает наименьшее целое число, которое больше или равно текущему значению	ceil(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: ceil(25.1) Вернет значение 26
floor	Возвращает наибольшее целое число, которое меньше или равно текущему значению	floor(0) Вместо 0 нужно вставить числовое значение/числовую переменную **Пример**: floor(25.1) Вернет значение 25
extract_from_datetime	Получает указанную часть (день, месяц, год, час, минута, секунда) из значения даты и времени	extract_from_datetime("","") Вместо первых кавычек нужно указать необходимую часть для извлечения: · SECOND · MINUTE · HOUR · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа datetime **Пример:** extract_from_datetime("DAY",`datetime`)
extract_from_date	Получает указанную часть (день, месяц, год) из значения даты	extract_from_date("","") В первые кавычки нужно вписать необходимую часть для извлечения: · DAY · MONTH · YEAR Вместо вторых кавычек нужно указать переменную типа date **Пример:** extract_from_date("DAY",`date`)
concat	Объединяет в единую строку указанные значения	concat("") Вместо "" нужно вставить строки/строковые переменные **Пример:** concat (`Фамилия`, " ", `Имя`)

Таблица 17 Операторы и функции

Результаты выполнения узла:

- Таблица с примером набора данных с посчитанными атрибутами. Отображаются первые 100 наблюдений.



Пример данных

Поиск...

Настроить Таблицу

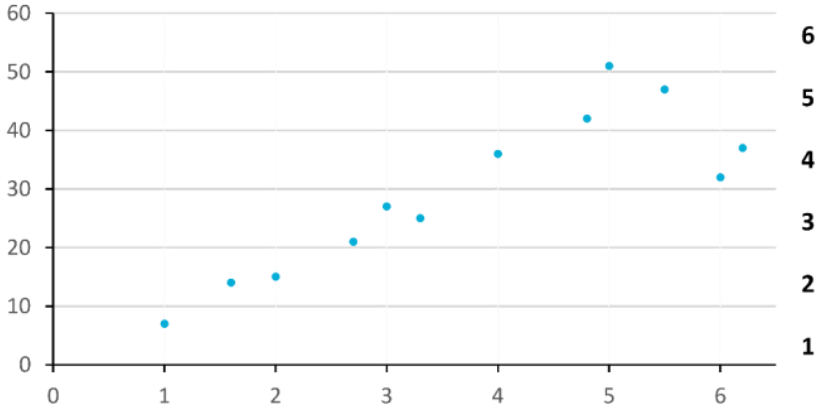
Месяц коммуникации	Количество дней после регистрации клиента	Общее количество коммуникаций	Число дней с предыдущей коммуникации	Наличие предыдущих коммуникаций	Результат предыдущей коммуникации	Отклик на предложение	Тестовый атрибут
май	1042	1	-1	0	неизвестно	1	1
май	1467	1	-1	0	неизвестно	1	1
май	1389	1	-1	0	неизвестно	1	1
май	579	1	-1	0	неизвестно	1	1
май	673	2	-1	0	неизвестно	1	1
май	562	2	-1	0	неизвестно	1	1
май	1201	1	-1	0	неизвестно	1	1
май	1030	1	-1	0	неизвестно	1	1

Рисунок 98 Пример таблицы с набором данных с рассчитанным атрибутом Тестовый атрибут

3.2.5.6.9. Узел «Биннинг/энкодинг»

Узел «Биннинг/энкодинг» включает в себя методы биннинга интервальных переменных и кодирования категориальных переменных.

Биннинг – процесс разделения диапазона непрерывной переменной на k интервалов. Может использоваться для сокращения размерности данных, что часто повышает точность модели.



X	Y	Бин
1	7	1
2	15	2
3	27	3
4	36	4
5	51	5
6	32	4
1,6	14	2
2,7	21	3
3,3	25	3
4,8	42	6
5,5	47	5
6,2	37	4

Рисунок 99 Пример биннинга по значениям переменной Y

Предусмотрены следующие методы бинаризации переменной:

- **Одинаковая ширина**. Данная стратегия разделяет диапазон значений переменной на указанное количество равных интервалов.
- **Квантильный**. Данная стратегия разделяет диапазон значений переменной таким образом, чтобы в каждый из них попало примерно одинаковое количество значений.
- **Дерево.** Данная стратегия разделяет диапазон значений переменной с помощью дерева решений, что предполагает связь с целевой переменной.

Разбиение начинается со всех наблюдений, которые представлены корневым узлом дерева. Алгоритм разбивает этот родительский узел на дочерние узлы (и листья) таким образом, чтобы значения (уровни) целевой переменной в пределах каждого дочернего региона были максимально похожи (критерий разбиения задается в параметрах **Критерий разбиения для регрессии** и **Критерий разбиения для классификации**).

Соответственно, параметр **Количество бинов** равен **Максимальной глубине дерева** в исходном алгоритме дерева решений, а параметр **Минимальное количество наблюдений** – **Минимальному количество наблюдений в листе** дерева (подробнее про работу алгоритма в справке **узла «Дерево решений»**).

Если хотя бы для одной переменной используется бинаризация на основе дерева, то должна быть задана **целевая переменная** (задается **в узле «Метаданные»**).

Предусмотрены отдельные бины для пропущенных значений и значений вне диапазона (параметры **Обработка пропущенных значений** и **Обработка значений вне диапазона).**

Большинство алгоритмов машинного обучения не могут обрабатывать категориальные переменные. **Энкодинг** – процесс преобразования текстовых атрибутов в числовые значения.

Закодировать категориальные переменные в числовые можно четырьмя методами:

- **Количество** (Count Encoding) – для каждой категории ставится количество наблюдений с этой категорией.
- **Частота** (Freq Encoding) – для каждой категории ставится частота наблюдений с этой категорией.
- **Целевая** (Target Encoding) – для каждой категории ставится средневзвешенное среднего целевой на подвыборке соответствующей заданной категории и среднего целевой на всей обучающей выборке (train). Параметр веса рассчитывается через количество наблюдений в подвыборке, соответствующей заданной категории, и параметры **Сдвига** и **Масштаба.**

Для корректной работы необходимо указать целевую переменную (задается в узле «**Метаданные**»).

- **WOE** (Weight of Evidence) – для каждой категории ставится WOE, рассчитанный для подвыборки, соответствующей выбранной категории. Математически **WOE** определяется как логарифм отношения доли «хороших» наблюдений к доле «плохих» наблюдений.

Для корректной работы необходимо указать целевую переменную (задается в узле «**Метаданные**»).

Предусмотрены **Обработка пропущенных значений** и **Обработка неизвестных значений**, которые зависят от выбранного метода энкодинга (подробнее в описании соответствующих параметров).

Если целевая переменная категориальная и в ней больше 2-х категорий, то кодировщики, основанные на целевой переменной (**Target Encoding** и **WOE**),

исключают одну из категорий целевой переменной и для каждой из оставшихся рассматривают соответствующую задачу бинарной классификации (1 – выбранная категория, 0 – все остальные) и строятся по ней. Таким образом, если у целевой переменной K категорий, K > 2, то каждый кодировщик, основанный на целевой переменной, генерирует K - 1 столбцов в результатах.

Список параметров узла представлен в таблице.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Префикс выходных переменных	Ручной ввод По умолчанию — BIN_	Данный параметр задает префикс для выходных переменных узла Используется для биннинга интервальных переменных
Обработка значений вне диапазона	Раскрывающийся список со следующими значениями: • Включить в крайние бины (по умолчанию) • Отдельные бины • Игнорировать	Данный параметр указывает что делать со значениями вне диапазона. Используется для биннинга интервальных переменных
Бин для пропущенных значений	Чекбокс	Выбор данного чекбокса указывает, что необходимо посчитать отдельный бин для пропущенных значений Используется для биннинга интервальных переменных
Переменные	Кнопка	Используется для биннинга интервальных переменных При выборе кнопки **«Переменные»** открывается окно **Переменные**, в котором необходимо выбрать переменные для биннинга и указать необходимые параметры. Параметр **Метод** задает метод, который будет использоваться для бинаризации переменной. Предусмотрены следующие методы: • **Одинаковая ширина** (по умолчанию) • **Квантильный** • **Дерево** Параметр **Количество бинов** имеет значение по умолчанию, равное 10. Для **метода Дерево** значение должно быть больше или равно 2. Для остальных методов – больше или равно 1.

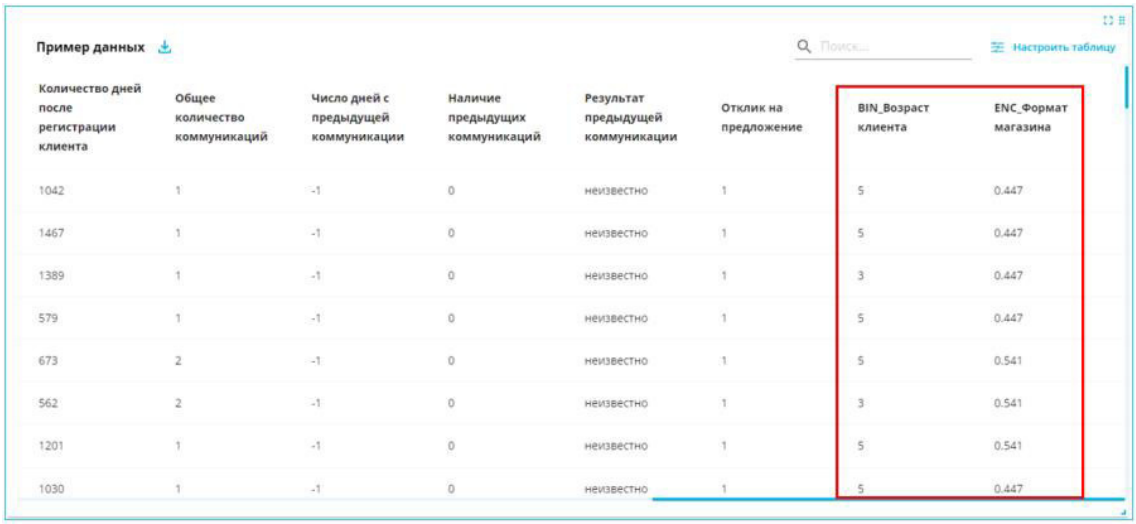
Параметр	Возможные значения и ограничения	Описание
Минимальное количество наблюдений в бине	Ручной ввод целочисленного значения больше ли равно 1 По умолчанию — 1	Является дополнительным параметром для **биннинга на основе дерева** и задает минимальное количество наблюдений в бине
Критерий разбиения для регрессии	Раскрывающийся список со следующими значениями: • squared error (по умолчанию) • friedman mse • absolute error • poisson	Является дополнительным параметром для **биннинга на основе дерева** и задает критерий разбиения для регрессии. Предусмотрены следующие критерии: • **squared error** (среднеквадратичная ошибка) • **friedman mse** (среднеквадратичная ошибка с оценкой улучшения Фридмана) • **absolute error** (средняя абсолютная ошибка) • **poisson** (отклонение Пуассона)
Критерий разбиения для классификации	Раскрывающийся список со следующими значениями: • gini (по умолчанию) • entropy	Является дополнительным параметром для **биннинга на основе дерева** и задает критерий разбиения для классификации. Предусмотрены следующие критерии: • **gini** (коэффициент Джини) • **entropy** (критерий прироста информации, энтропия)
Префикс выходных переменных	Ручной ввод По умолчанию — ENC_	Данный параметр используется для кодирования категориальных переменных и задает префикс для выходных переменных узла
Обработка пропущенных значений	Раскрывающийся список со следующими значениями: • Значение • Пропуск	Данный параметр используется для кодирования категориальных переменных и задает метод обработки пропущенных значений. Предусмотрены следующие методы: • **Пропуск** – будет поставлено NaN. • **Значение** – зависит от выбранного метода энкодинга. Для Count и Freq Encoding – пропуск как отдельная категория, Target Encoding – среднее значение целевой переменной на обучающей выборке (train), WOE Encoding – значение 0.
Обработка неизвестных значений	Раскрывающийся список со следующими значениями: • Значение • Пропуск	Данный параметр используется для кодирования категориальных переменных и задает метод обработки неизвестных значений. Предусмотрены следующие методы: • **Пропуск** – будет поставлено NaN. • **Значение** – зависит от выбранного метода энкодинга. Для Count, Freq и WOE Encoding – значение 0, Target Encoding – среднее значение целевой переменной на обучающей выборке (train).
Минимальная частота значений	Ручной ввод Число больше или	Данный параметр используется для кодирования категориальных переменных и

Параметр	Возможные значения и ограничения	Описание
	равно 0 и меньше или равно 1 По умолчанию — 0	задаёт минимальный % значений с категорией для того, чтобы не считать её редкой. Все редкие категории собираются в одну категорию и при скоринге для всех редких категорий будет проставлено одно и то же значение. Пропущенные значения не входят в редкие категории.
Количество итераций кросс-валидации	Ручной ввод Число больше или равно 0 По умолчанию — 0	Данный параметр используется для кодирования категориальных переменных и задает количество итераций кросс-валидации. Кросс-валидация используется для борьбы с переобучением при использовании кодировщиков, основанных на целевой переменной (методы **Target Encoding** и **WOE**).
Переменные	Кнопка	Используется для кодирования категориальных переменных При выборе кнопки **«Переменные»** открывается окно **Переменные**, в котором необходимо выбрать переменные для энкодинга и указать необходимые параметры. Параметр **Метод** задает метод, который будет использоваться для энкодинга. Предусмотрены следующие методы: • **Количество** (Count Encoding) • **Частота** (Freq Encoding) • **Целевая** (Target Encoding) • **WOE** (Weight of Evidence)
Сдвиг	Ручной ввод По умолчанию — 1	Параметр для **Target Encoding** Данный параметр используется для кодирования категориальных переменных и задает сдвиг
Масштаб	Ручной ввод Число больше 0 По умолчанию — 0,001	Параметр для **Target Encoding** Данный параметр используется для кодирования категориальных переменных и задает масштаб
Рандомизация	Чекбокс	Параметр для **WOE Encoding** Выбор чекбокса указывает на необходимость рандомизации
Стандартное отклонение	Ручной ввод Число больше или равно 0 По умолчанию — 0,05	Параметр для **WOE Encoding** Данный параметр используется для кодирования категориальных переменных и задает стандартное отклонение
Регуляризация	Ручной ввод Число больше или равно 0 По умолчанию — 1	Параметр для **WOE encoding** Данный параметр используется для кодирования категориальных переменных и задает значение регуляризации

Таблица 18 Параметры узла «Биннинг»

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.



Пример данных

Поиск...

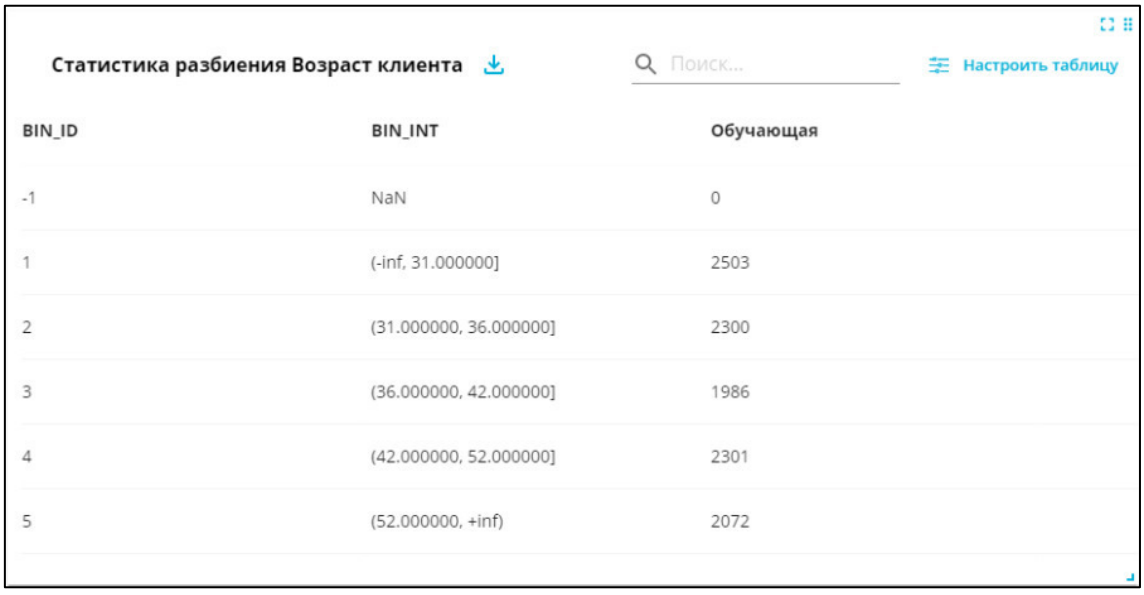
Настроить таблицу

Количество дней после регистрации клиента	Общее количество коммуникаций	Число дней с предыдущей коммуникации	Наличие предыдущих коммуникаций	Результат предыдущей коммуникации	Отклик на предложение	BIN_Возраст клиента	ENC_Формат магазина
1042	1	-1	0	неизвестно	1	5	0.447
1467	1	-1	0	неизвестно	1	5	0.447
1389	1	-1	0	неизвестно	1	3	0.447
579	1	-1	0	неизвестно	1	5	0.447
673	2	-1	0	неизвестно	1	5	0.541
562	2	-1	0	неизвестно	1	3	0.541
1201	1	-1	0	неизвестно	1	5	0.541
1030	1	-1	0	неизвестно	1	5	0.447

Рисунок 100 Таблица с примером данных

В результате выполнения узла в наборе данных будут рассчитаны новые переменные с указанными в параметрах префиксами и ролью Предиктор. Переменная, полученная в ходе биннинга, будет иметь тип Nominal, в ходе кодирования – Interval.

- Таблица со статистиками биннинга.



Статистика разбиения Возраст клиента

Поиск...

Настроить таблицу

BIN_ID	BIN_INT	Обучающая
-1	NaN	0
1	(-inf, 31.000000]	2503
2	(31.000000, 36.000000]	2300
3	(36.000000, 42.000000]	1986
4	(42.000000, 52.000000]	2301
5	(52.000000, +inf)	2072

Рисунок 101 Пример таблицы со статистиками бинаризации

- Таблица со статистиками кодирования.

Статистика кодирования Формат магазина

Поиск...

Настроить таблицу

Формат магазина	Freq	Rare level	ENC_Формат магазина
ТЦ	0.49	0	0.447
магазин у дома	0.134	0	0.394
неизвестно	0.044	0	0.507
франчайзинг	0.33	0	0.541
NaN	0	0	0.473
Unknown levels	0	0	0.473

Рисунок 102 Пример таблицы со статистиками кодирования

3.2.5.6.10. Узел «Дисперсионный анализ»

Узел «Дисперсионный анализ» применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную.

В ходе дисперсионного анализа тестируется гипотеза о равенстве средних (целевой непрерывной переменной) в группах (которые задаются одной или несколькими категориальными переменными) с помощью сравнения (анализа) дисперсий.

Дисперсия случайной величины — мера разброса значений случайной величины относительно её математического ожидания. Разделение общей дисперсии на несколько источников, позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью.

ВАЖНО! Для работы узла необходимо, чтобы целевая переменная была **интервального типа (Interval)**.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла

Таблица 19 Параметры узла «Дисперсионный анализ»

Результаты выполнения узла:

- Табличное представление результатов дисперсионного анализа. Основные понятия таблицы с результатами дисперсионного анализа:
  - **Intercept** — коэффициент β0.
  - **Grouping** — межгрупповая дисперсия.
  - **Residual** — внутригрупповая дисперсия.
  - **Сумма квадратов** — суммы квадратов отклонений.
  - **Степени свободы** – количество значений, которые могут свободно изменяться.
  - **F-статистика** — значение критерия Фишера
  - **p-value** — вероятность получить F-значение, равное или превышающее то значение, которое рассчитано по имеющимся выборочным данным (при условии, что нулевая гипотеза верна).

Результаты дисперсионного анализа

	Сумма квадратов	Степени свободы	F-статистика	p-value
Intercept	106.183	1	24.401	0
Grouping	92.922	3	7.117	0.009
Residual	39.164	9	-	-

Рисунок 103 Пример таблицы с результатами дисперсионного анализа

- Табличное представление средних значений зависимой переменной для каждой группы.

Средние по группам

Type	density
Газ	0.002
Лантаноид	8.305
Металл	8.944

Рисунок 104 Пример таблицы со средними значениями целевой переменной (density) по группам (Type)

3.2.5.6.11. Узел «Стандартизация»

Узел «Стандартизация» приводит признаки в разных единицах измерения и диапазонах значений к общей шкале.

Стандартизация – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью.
Предусмотрены следующие методы:

- **Стандартное отклонение (std)** – преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1.
- **Диапазон (range)** – линейно преобразует значения переменных в диапазон [0, 1].

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Префикс выходных переменных	Ручной ввод	Данный параметр задает префикс стандартизированным выходным переменным
Выбор переменных	Кнопка	При выборе кнопки **«Выбор переменных»** открывается окно **Выбор переменных**, в котором необходимо выбрать переменные для стандартизации.

Таблица 20 Параметры узла «Стандартизация»

В окне **Выбор переменных** Пользователь имеет возможность выбрать атрибут для стандартизации и указать метод. Для этого необходимо:

- Рядом с интересующей переменной выбрать иконку 
- В выпадающих меню выбрать **Метод** (std или range) и ограничить диапазон при стандартизации в [0, 1] на скоринге (столбец **Clip**). Если в столбце **Clip** стоит нет, то при стандартизации в [0, 1] на обучающей выборке всё точно будет в [0, 1], а на других выборках не обязательно.
- Сохранить изменения, выбрав иконку  .

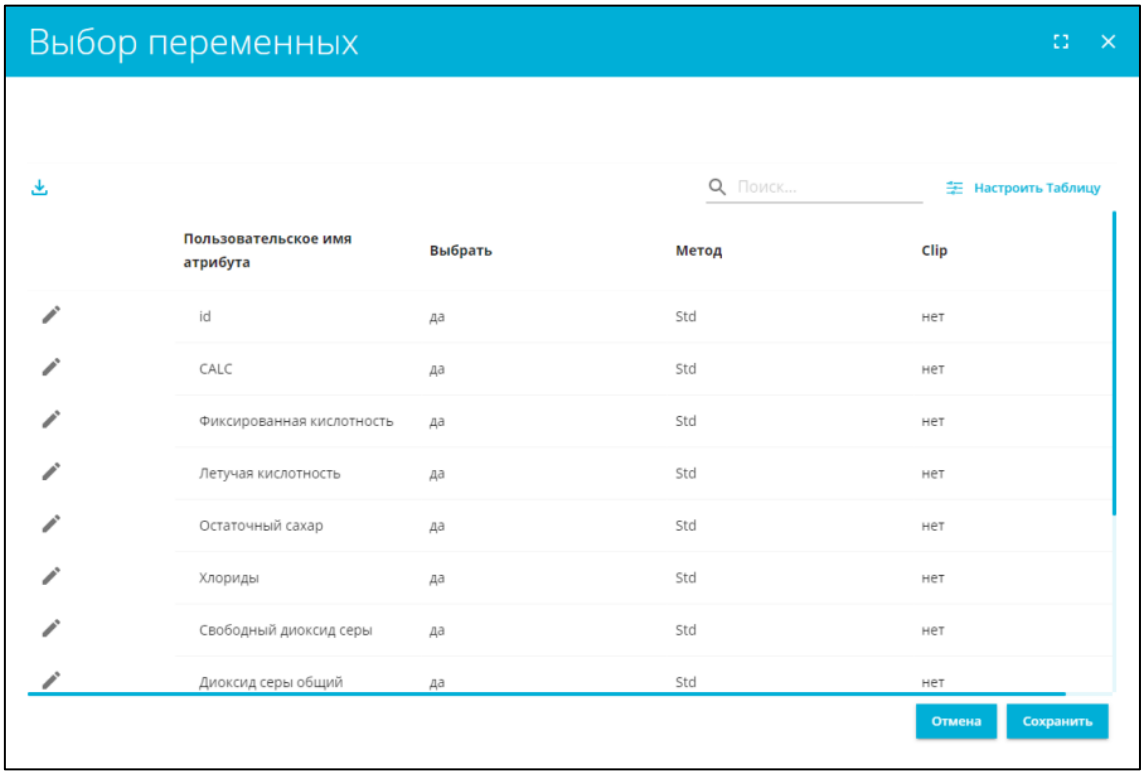


Рисунок 105 Окно Выбор переменных

Результаты выполнения узла:

- Таблица с примером полученных данных. Отображаются первые 100 наблюдений.

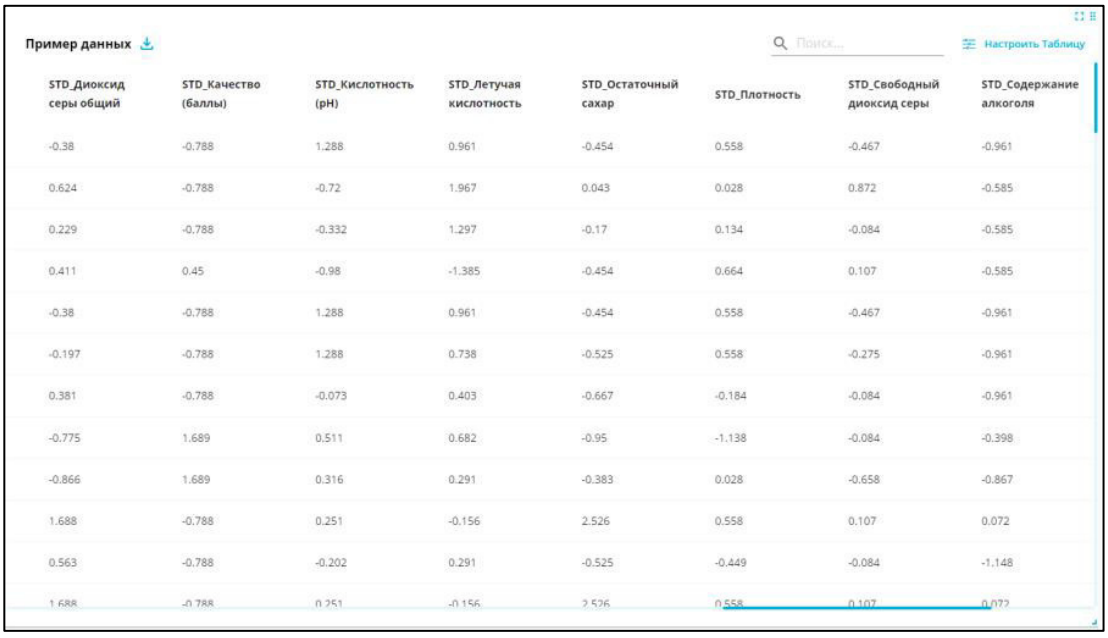


Рисунок 106 Таблица с примером полученных данных

В результате выполнения узла в наборе данных будут рассчитаны новые стандартизированные переменные (с префиксом **STD_**).

3.2.5.6.12. Узел «Веса классов»

Узел «Веса классов» позволяет скорректировать дисбаланс классов при помощи задания весов.

В **задаче классификации** данные называются **несбалансированными**, когда в обучающей выборке доли объектов разных классов существенно различаются. Для такой ситуации существуют разные стратегии перебалансировки данных (например, замена большого класса подвыборкой по мощности равной малому классу, или **undersampling**, который предусмотрен в **узле «Sample»**).

В **узле «Веса классов»** Пользователь может задать веса для объектов каждого из классов вручную, либо сбалансировать автоматически.

Список параметров узла представлен в таблице.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Сбалансировать веса	Чекбокс	Выбор данного чекбокса указывает, что необходимо использовать автоматическую корректировку весов, которая считается обратно пропорционально частотам классов во входных данных.
Задать веса	Кнопка	При выборе кнопки открывается окно **Задать веса**, в котором необходимо задать вес каждому из классов.

Таблица 21 Параметры узла «Веса классов»

В окне **Задать веса** Пользователь может вручную задать веса каждому из классов. Для этого необходимо:

- Рядом с интересующей переменной выбрать иконку  .
- В столбце **Вес** задать необходимое значение.
- Сохранить изменения, выбрав иконку  .

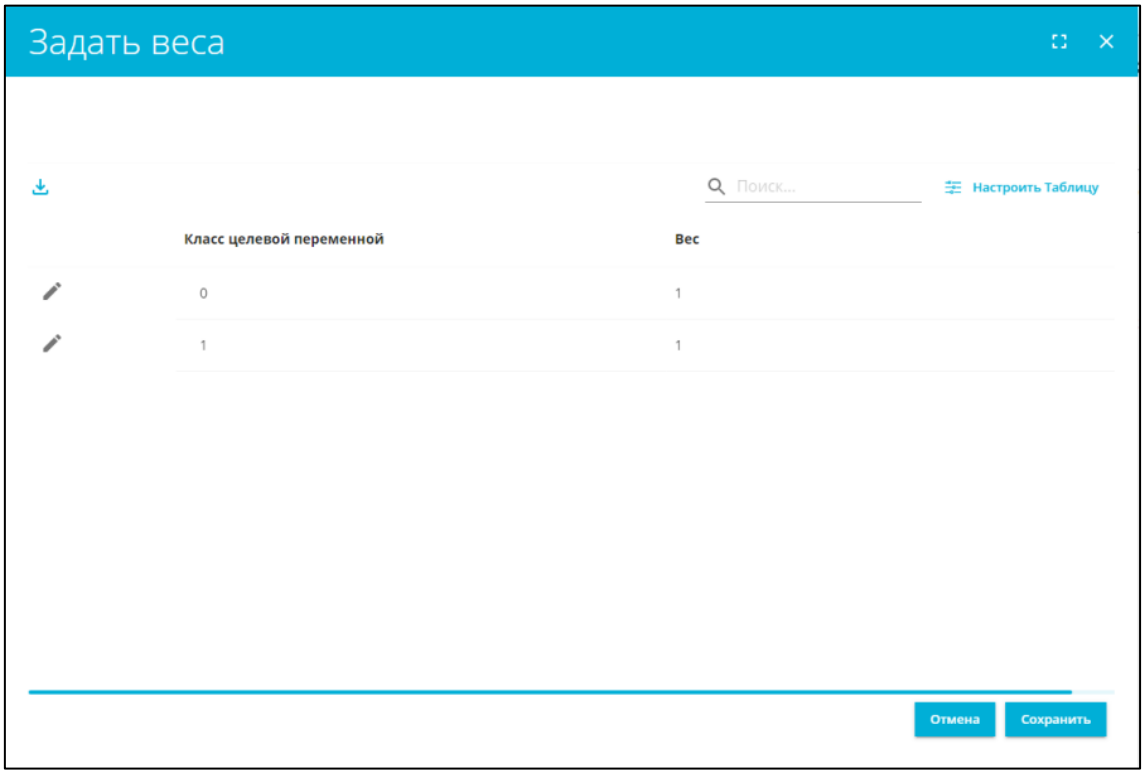


Рисунок 107 Окно Выбор переменных

Для работы узла достаточно выбрать чекбокс **Сбалансировать веса** и запустить расчет узла. Весы классов посчитаются автоматически (без необходимости указывать их вручную в окне **Задать вес**).

Результаты выполнения узла:

- Таблица с примером полученных данных. Отображаются первые 100 наблюдений.

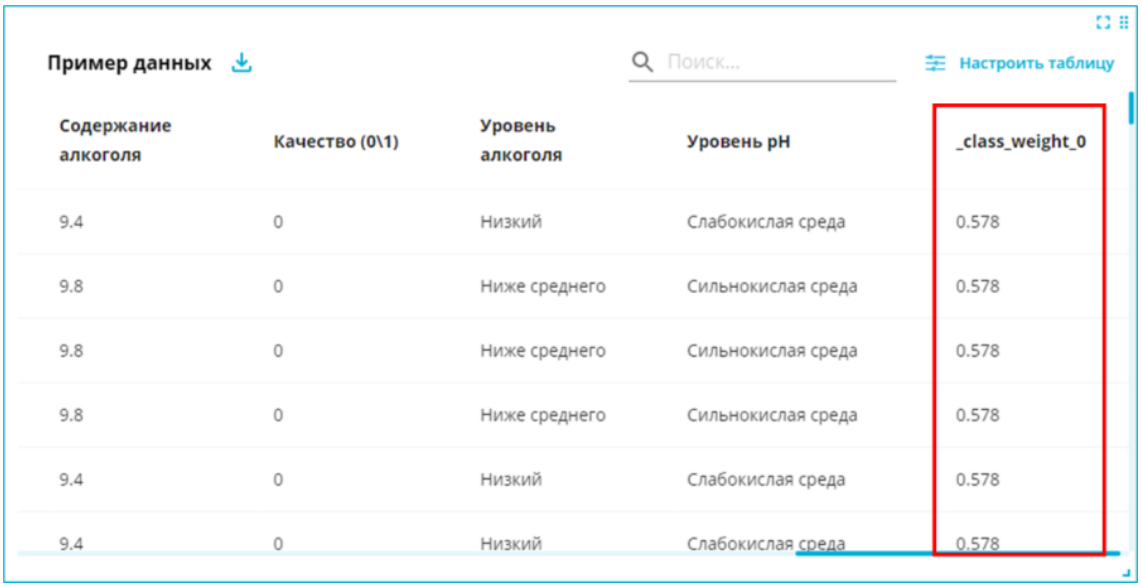


Рисунок 108 Таблица с примером полученных данных

В результате выполнения узла в наборе данных будет рассчитана новая переменная с весами и ролью **ClassWeight** (имя переменной **_class_weight_0**).

- Таблица с весами классов.

Рисунок 109 Таблица с весами классов

3.2.5.6.13. Узел «Автоэнкодер (PyTorch)»

Узел "Автоэнкодер (PyTorch)" позволяет создать сжатое представление данных (т.е. сократить размерность) в скрытом слое за счет специальной архитектуры нейронной сети.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе	Общие параметры
Описание	Ручной ввод Ограничений на значение нет	Описание узла	Общие параметры
Конфигурация слоев автоэнкодера	Кнопка	При выборе данной кнопки откроется окно **Конфигурация слоев нейросети**, где можно задавать слои, количество нейронов в слоях и функции активации для каждого узла. Для этого необходимо выбрать кнопку **Добавить** и в появившемся списке выбрать необходимую функцию активации и настроить параметры. Предусмотрены: • **Функция активации ReLU (Rectified Linear Unit)** – $f(x) = \max(0, x)$ • **Функция активации CELU** – $f(x) = \max(0, x) + \min(0, \alpha * (\exp(x/\alpha) - 1))$ Дополнительно необходимо задать α • **Функция активации ELU** – $f(x) = \begin{cases} \alpha * (exp(x) - 1) & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases}$	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		Дополнительно необходимо задать α • **Функция активации Sigmoid** – $\sigma(x) = \frac{1}{1+e^{-x}}$ • **Функция активации Softmax** – $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad \text{for } i = 1,2,\dots,K$ • **Линейный слой** – линейная функция, результат пропорционален переданному аргументу $f(x) = x$ Дополнительно необходимо задать **Количество выходных переменных** и при необходимости выбрать чекбокс **Добавить константу** • **Функция активации Logsigmoid** – $\text{LogSigmoid}(x) = \log(\frac{1}{1+e^{-x}})$ • **Исключение (Dropout)** – Во время обучения случайным образом обнуляет некоторые элементы входного тензора с вероятностью p, используя выборки из распределения Бернулли. Дополнительно необходимо задать **Вероятность исключения** • **Tanh** – функция гиперболического тангенса $f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}}$ После настройки необходимой конфигурации слоев нейросети выбрать кнопку **Сохранить**.	
Функция потерь	Раскрывающийся список со следующими значениями: • MAE • MSE • Poisson loss • Negative log likelihood • Cross entropy	Данный параметр задает функцию потерь. Предусмотрены: • **MAE** (средняя абсолютная ошибка) • **MSE** (среднеквадратическая ошибка) • **Poisson loss** • **Negative log likelihood** (отрицательное логарифмическое правдоподобие) • **Cross entropy** (перекрестная энтропия)	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Алгоритм оптимизации	Раскрывающийся список со следующими значениями: • SGD • Adam (по умолчанию) • Adadelta • Adamax • LBFGS • AdamW • ASGD • NAdam • RAdam • Rprop • RMSprop	Данный параметр задает метод оптимизации, который будет использоваться для обновления весов нейронов скрытых слоев нейронной сети. Предусмотрены следующие методы: • **SGD (stochastic gradient descent)** – стохастический градиентный спуск. Данный метод делает шаг постоянной величины в направлении, указанном градиентом в текущей точке • **Adam (adaptive moment estimation)** – адаптивная оценка момента. Данный метод сочетает в себе идею метода Momentum о накоплении градиента и идею методов Adadelta и RMSProp об экспоненциальном сглаживании информации о предыдущих значениях квадратов градиентов. • **Adadelta (adaptive learning rate)** – метод адаптивной скорости обучения • **Adamax** – модификация метода Adam, основанная на бесконечной норме (max) • **LBFGS (limited-memory BFGS)** – BFGS с ограниченной памятью. • **AdamW.** Данный метод основан на адаптивной оценке моментов первого и второго порядка с добавленным методом уменьшения весов • **ASGD (average stochastic gradient descent)** – усредненный стохастический градиентный спуск. Данный метод усредняет веса, вычисляемые на каждой итерации. • **Nadam (Nesterov-accelerated adaptive momentum).** Данный метод представляет собой модификацию оптимизатора Adam с добавлением	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		момента Нестерова при вычислении градиентов. • **RAdam (Rectified Adam).** Данный метод является модификацией Adam, более устойчивой к изменению значений скорости обучения. • **Rprop (resilient backpropagation)** – устойчивый алгоритм обратного распространения. Данный метод использует только знаки частных производных для подстройки весовых коэффициентов. Также Rprop поддерживает отдельные дельты для каждого веса и смещения и адаптирует эти дельты во время обучения. • **RMSPROP (root mean square propagation)** – среднеквадратичное распространение корня. Данный метод использует усредненный по истории квадрат градиента.	
Скорость обучения	Ручной ввод По умолчанию - 0,001	Данный параметр задает скорость обучения, которая управляет размером шага при обновлении весов.	Общие параметры
Количество эпох	Ручной ввод По умолчанию - 10	Данный параметр задает сколько раз алгоритм обучения будет обрабатывать весь набор обучающих данных.	Общие параметры
Размер пакета	Ручной ввод По умолчанию - 128	Данный параметр определяет количество выборок, которые необходимо обработать перед обновлением параметров модели.	Общие параметры
Seed	Ручной ввод По умолчанию - 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске узла.	Общие параметры
L2 регуляризация	Ручной ввод По умолчанию - 0,1	Данный параметр задает значение L2-регуляризации	Общие параметры
Доля валидационной выборки	Ручной ввод По умолчанию - 0,1	Данный параметр задает долю валидационной выборки, которая будет отобрана из исходной тестовой	Общие параметры
Количество итераций без	Ручной ввод По умолчанию - 5	Данный параметр задает количество эпох без улучшения,	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
существенного улучшения		после которых скорость обучения будет снижена. Пример: если значение параметра = 2, то первые 2 эпохи без улучшений loss будут проигнорированы, и только после 3-й эпохи скорость обучения уменьшится.	
Режим динамического расчета порога	Раскрывающийся список со следующими значениями: • rel • abs	Данный параметр задет режим расчета порога Предусмотрены: • **rel** – dynamic_threshold = best * (1 - threshold) • **abs** – dynamic_threshold = best - threshold	Общие параметры
Порог	Ручной ввод По умолчанию - 0,0001	Данный параметр задает порог расчета нового оптимума, чтобы сосредоточиться только на значительных изменениях	Общие параметры
Стандартизация	Раскрывающийся список со следующими значениями: • no • std (по умолчанию) • range	Данный параметр отвечает за выбор метода стандартизации числовых переменных. **Стандартизация** – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью. Предусмотрены следующие методы: • **no** — стандартизация не нужна • **std** — стандартное отклонение - преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • **range** — диапазон - линейно преобразует значения переменных в диапазон [0, 1].	Общие параметры
Beta1	Ручной ввод числа с плавающей точкой По умолчанию - 0,9	Данный параметр задает коэффициент, используемый для управления скоростью затухания скользящих средних значений градиента (первого момента)	Параметры алгоритма оптимизации **Adam, Adamax, AdamW,**

Параметр	Возможные значения и ограничения	Описание	Группа параметров
			RAdam, Nadam
Beta2	Ручной ввод числа с плавающей точкой По умолчанию - 0,999	Данный параметр задает коэффициент, используемый для управления скоростью затухания средних значений вторых моментов градиентов (нецентрированной дисперсии)	Параметры алгоритма оптимизации **Adam, Adamax, AdamW, RAdam, Nadam**
Epsilon	Ручной ввод числа с плавающей точкой По умолчанию - 1e-8	Данный параметр задает значение, добавляемое к знаменателю для улучшения числовой стабильности Minimal decay applied to lr. If the difference between new and old lr is smaller than eps, the update is ignored. Default: 1e-8.	Параметры алгоритма оптимизации **Adam, Adamax, AdamW, RAdam, Nadam**
Использовать алгоритм AMSGrad	Чекбокс	Выбор данного чекбокса указывает, что необходимо использовать вариант AMSGrad этого алгоритма. Разница между AMSgrad и Adam заключается в рассчитанном векторе второго момента, который используется для обновления параметров.	Параметры алгоритма оптимизации **Adam**, **AdamW**
Импульс (Momentum)	Ручной ввод числа с плавающей точкой По умолчанию - 0	Данный параметр задает коэффициент импульса (запоминает скорость на предыдущем шаге и добавляет в указанное число раз меньшую величину на следующем шаге)	Параметры алгоритма оптимизации **SGD, RMSProp**
Dampening	Ручной ввод числа с плавающей точкой По умолчанию - 0	Данный параметр задает демпфирование импульса. Демпфирование гарантирует, что оптимизатор не сделает слишком больших шагов, что может произойти, если использовать только импульс. Чем выше градиент, тем больше демпфирование уменьшает размер шага.	Параметры алгоритма оптимизации **SGD**
Момент Нестерова	Чекбокс	Выбор данного чекбокса включает импульс Нестерова (использует производную не в текущей точке, а в следующей, если бы мы продолжали двигаться в этом же направлении без изменений)	Параметры алгоритма оптимизации **SGD**
Rho	Ручной ввод По умолчанию - 0,9	Данный параметр задает коэффициент, используемый для вычисления скользящего среднего квадратов градиентов	Параметры алгоритма оптимизации **Adadelta**

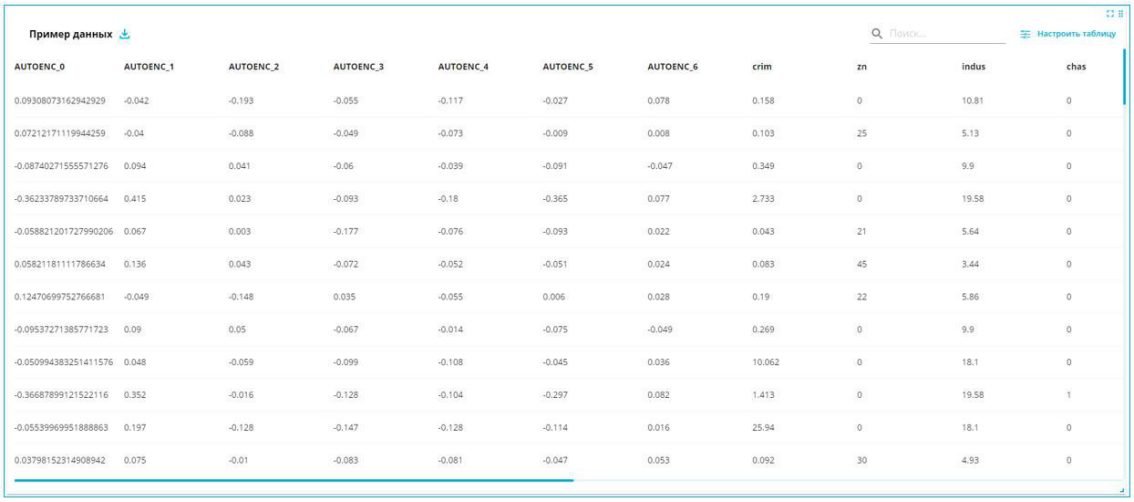
Параметр	Возможные значения и ограничения	Описание	Группа параметров
Epsilon	По умолчанию - 1e-8	Данный сглаживающий параметр задает значение, предотвращающее деление на 0.	Параметры алгоритма оптимизации **Adadelta, RMSProp**
Максимум итераций за шаг оптимизации	Ручной ввод целочисленного значения По умолчанию - 20	Данный параметр задает максимальное число итераций за шаг оптимизации	Параметры алгоритма оптимизации **LBFGS**
Максимум вычислений за шаг оптимизации	Ручной ввод целочисленного значения По умолчанию - 1	Данный параметр задает максимальное число вычислений функции за шаг оптимизации	Параметры алгоритма оптимизации **LBFGS**
Tolerance grad	Ручной ввод числа с плавающей точкой По умолчанию - 0,00001	Данный параметр задает допуск завершения при оптимальности первого порядка	Параметры алгоритма оптимизации **LBFGS**
Tolerance change	Ручной ввод числа с плавающей точкой По умолчанию - 1e-9	Данный параметр задает допуск завершения при изменении значения/параметра функции	Параметры алгоритма оптимизации **LBFGS**
Количество запоминаемых шагов оптимизации	Ручной ввод целочисленного значения По умолчанию - 100	Данный параметр задает количество запоминаемых шагов оптимизации	Параметры алгоритма оптимизации **LBFGS**
Line search	Список: • no (по умолчанию) • strong wolfe	Данный параметр задает метод линейного поиска	Параметры алгоритма оптимизации **LBFGS**
Lambda	Ручной ввод числа с плавающей точкой По умолчанию - 0,0001	Данный параметр задает затухание	Параметры алгоритма оптимизации **ASGD**
Alpha	Ручной ввод числа с плавающей точкой По умолчанию - 0,75	Данный параметр задает мощность для обновления скорости обучения	Параметры алгоритма оптимизации **ASGD**
t0	Ручной ввод числа с плавающей точкой	Данный параметр задает точку, с которой начинается усреднение. Если требуемое количество итераций меньше данного	Параметры алгоритма оптимизации **ASGD**

Параметр	Возможные значения и ограничения	Описание	Группа параметров
	По умолчанию - 100000	значения, то усреднение не произойдет.	
Сокращение импульса	Ручной ввод По умолчанию - 0,004	Данный параметр задает значение сокращения импульса	Параметры алгоритма оптимизации **Nadam**
Коэффициент уменьшения (eta minus)	Ручной ввод Число больше 0 и меньше 1 По умолчанию - 0,5	Данный параметр задает мультипликативный коэффициент уменьшения. Если на текущем шаге частная производная по соответствующему весу поменяла свой знак, значит последнее изменение было большим, и алгоритм проскочил локальный минимум. Следовательно, величину коррекции необходимо уменьшить на значение данного параметра и вернуть предыдущее значение весового коэффициента.	Параметры алгоритма оптимизации **Rprop**
Коэффициент увеличения (eta plus)	Ручной ввод Число больше 1 По умолчанию - 1,2	Данный параметр задает мультипликативный коэффициент увеличения. Если на текущем шаге частная производная по соответствующему весу не поменяла свой знак, значит нужно увеличить величину коррекции на значение данного параметра для достижения более быстрой сходимости.	Параметры алгоритма оптимизации **Rprop**
Минимальный размер шага	Ручной ввод По умолчанию - 0,000001	Данный параметр задает минимальный размер шага. Он необходим, чтобы не допустить слишком маленьких значений весов, ограничивает величину коррекции снизу.	Параметры алгоритма оптимизации **Rprop**
Максимальный размер шага	Ручной ввод По умолчанию - 50	Данный параметр задает максимальный размер шага. Он необходим, чтобы не допустить слишком больших значений весов, ограничивает величину коррекции сверху.	Параметры алгоритма оптимизации **Rprop**
Alpha	Ручной ввод По умолчанию - 0,99	Данный параметр задает константу сглаживания	Параметры алгоритма оптимизации **RMSProp**
Центрировать	Чекбокс	Выбор данного чекбокса указывает, что необходимо вычислить центрированный RMSProp, градиент которого нормализуется по оценке его дисперсии	Параметры алгоритма оптимизации **RMSProp**

Таблица 22 Параметры узла «Автоэнкодер (PyTorch)»

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.



Пример данных

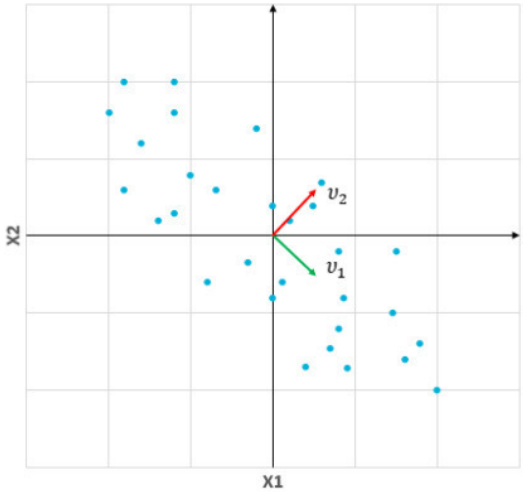
AUTOENC_0	AUTOENC_1	AUTOENC_2	AUTOENC_3	AUTOENC_4	AUTOENC_5	AUTOENC_6	crim	zn	indus	chas
0.09308073162942929	-0.042	-0.193	-0.055	-0.117	-0.027	0.078	0.158	0	10.81	0
0.072121711199442 59	-0.04	-0.088	-0.049	-0.073	-0.009	0.008	0.103	25	5.13	0
-0.087402715555571276	0.094	0.041	-0.06	-0.039	-0.091	-0.047	0.349	0	9.9	0
-0.36233789733710664	0.415	0.023	-0.093	-0.18	-0.365	0.077	2.733	0	19.58	0
-0.0588212017279902 06	0.067	0.003	-0.177	-0.076	-0.093	0.022	0.043	21	5.64	0
0.058211811111786634	0.136	0.043	-0.072	-0.052	-0.051	0.024	0.083	45	3.44	0
0.12470699752766681	-0.049	-0.148	0.035	-0.055	0.006	0.028	0.19	22	5.86	0
-0.095372713857717 23	0.09	0.05	-0.067	-0.014	-0.075	-0.049	0.269	0	9.9	0
-0.0509943832514115 76	0.048	-0.059	-0.099	-0.108	-0.045	0.036	10.062	0	18.1	0
-0.366878991215221 16	0.352	-0.016	-0.128	-0.104	-0.297	0.082	1.413	0	19.58	1
-0.055399699518888 63	0.197	-0.128	-0.147	-0.128	-0.114	0.016	25.94	0	18.1	0
0.037981523149089 42	0.075	-0.01	-0.083	-0.081	-0.047	0.053	0.092	30	4.93	0

Рисунок 110 Таблица с примером данных

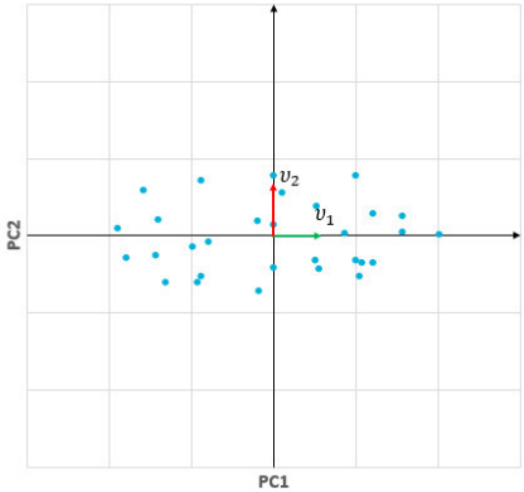
3.2.5.6.14. Узел «PCA»

Анализ главных компонент (PCA – Principal component analysis) – это метод, который преобразует большой набор переменных в меньший с минимальными потерями информативности. Использование PCA позволяет ускорить расчет модели за счет уменьшения количества входных переменных.

PCA используется для разложения многомерного набора данных на набор последовательных ортогональных компонентов, которые объясняют максимальную величину дисперсии (дисперсия – степень разброса данных).
Снижение размерности достигается с помощью **SVD** (Singular Value Decomposition, сингулярное разложение).



Данные в исходных координатах



Данные в новой системе координат, задаваемой главными компонентами

В узле предусмотрены следующие **Алгоритмы главных компонент**:

- **auto** – алгоритм подбирается автоматически исходя из числа переменных и количества наблюдений в наборе данных: если входные данные больше 500x500, а количество извлекаемых компонентов меньше 80% наименьшего размера данных, то будет выбран метод **randomized**. В остальных случаях будет выбран **full**.
- **full** – точный полный SVD (стандартный решатель LAPACK) и отбор компонент с помощью постобработки (**параметр Методы определения количества компонент**)
- **arpack** – SVD, усеченный количеством компонент в пределе строго больше 0 и строго меньше ***min(количество наблюдений, количество переменных)*** (решатель ARPACK).
- **randomized** – алгоритм находит приближенное усеченное SVD, используя рандомизацию для ускорения вычислений (по методу Halko).

Рекомендуется нормализовать данные перед использованием PCA, иначе можно получить вводящие в заблуждение компоненты (**чекбокс Нормализация компонент**).

PCA работает с числовыми данными. Поэтому если в наборе данных присутствует категориальная переменная, то она кодируется методом One-hot encoding и количество рассчитанных главных компонент будет равно (**исходное кол-во некатегориальных атрибутов) + (столбцы One-hot encoding)**.

Пример: изначально было 14 атрибутов один из которых категориальный (с 3 категориями), то в результатах получится 15 (подробнее про One-hot encoding в описании одноименного узла).

Для отбора компонент предусмотрены следующие **Методы определения количества компонент**:

- **full** – отбирается количество компонент в соответствии с формулой ***min(количество наблюдений, количество переменных) – 1.***
- **number** – отбираемое количество компонент задается вручную.
- **variance** – будут отобраны компоненты, которые равны или превышают заданный процент объясненной дисперсии.
- **mle** –для отбора компонентов используется MLE Минка.

Список параметров узла представлен в таблице ниже.

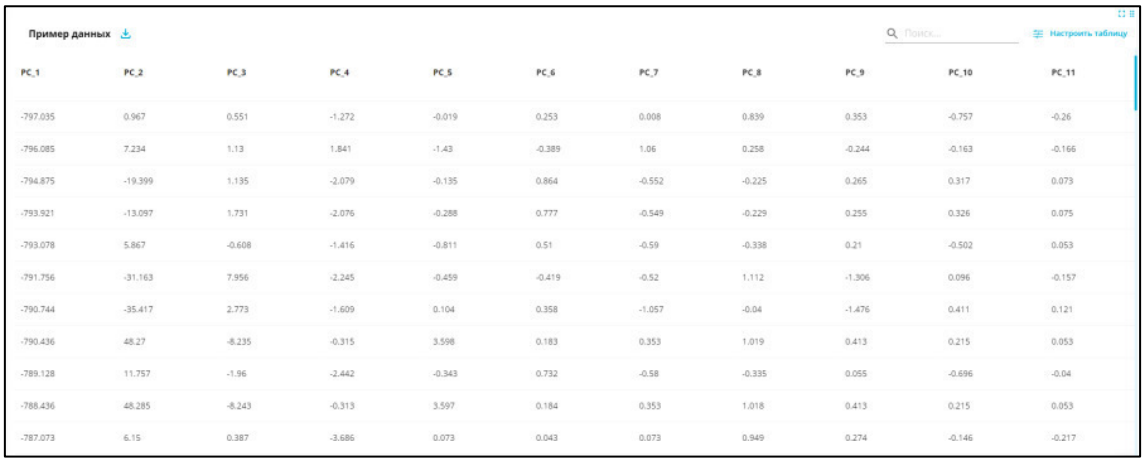
Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Метод определения количества компонент	Раскрывающийся список со следующими значениями: • full (по умолчанию) • number • variance • mle	Данный параметр задает метод определения количества компонент. Предусмотрены: • full • number • variance • mle
Количество компонент	Ручной ввод целочисленного значения По умолчанию - 1	Актуален при выборе **Метода определения количества компонент = number.** Задает ограничение на количество компонент, которые в конечном итоге будут отобраны алгоритмом.
Процент объясненной дисперсии	Ручной ввод Значение больше 0 и меньше или равно 1 По умолчанию – 0,95	Актуален при выборе **Метода определения количества компонент = variance.** Задает ограничение на процент объясненной дисперсии компонентами, которые в конечном итоге будут отобраны алгоритмом.
Нормализация компонент	Чекбокс	Данный чекбокс отвечает за преобразование числовых наблюдений с целью приведения их к общей шкале. Необходимость нормализации вызвана тем, что разные признаки из исходного набора могут быть представлены в разных масштабах и изменяться в разных диапазонах
Алгоритм главных компонент	Раскрывающийся список со следующими значениями: • auto (по умолчанию) • full • arpack • randomized	Данный параметр задает алгоритм главных компонент. Предусмотрены следующие варианты: • auto • full • arpack • randomized
Погрешность	Ручной ввод числового значения Значение должно быть в диапазоне [0.0, inf) По умолчанию – 0	Актуален при выборе **Алгоритма главных компонент = arpack** Задает допустимую погрешность для сингулярных значений.
Задать количество итераций степенного метода	Чекбокс	Актуален при выборе **Алгоритма главных компонент = randomized** Выбор данного чекбокса указывает, что

Параметр	Возможные значения и ограничения	Описание
		необходимо задать количество итераций степенного метода
Количество итераций степенного метода	Ручной ввод числового значения Значение должно быть в диапазоне [0, inf) По умолчанию – 10	Актуален при выборе **Алгоритма главных компонент = randomized** Перед увеличением **Количества итераций степенного метода** следует увеличивать **Количество дополнительных случайных векторов**, поскольку принцип рандомизированного метода заключается в том, чтобы избежать использования этих более дорогостоящих шагов итерации.
Количество дополнительных случайных векторов	По умолчанию – 10	Актуален при выборе **Алгоритма главных компонент = randomized** Задает дополнительное количество случайных векторов, что обеспечивает лучшую аппроксимацию сингулярных векторов и сингулярных значений.
Нормализация итераций	Раскрывающийся список со следующими значениями: • auto (по умолчанию) • QR • LU • randomized	Актуален при выборе **Алгоритма главных компонент = randomized** Предусмотрены следующие методы: • auto – не применяет нормализацию, если **Количество итераций степенного метода** <=2, и переключается на LU разложение в противном случае • QR – пошаговое разложение матрицы. Самое медленное, но наиболее точное • LU – LU разложение матрицы. Численно стабильное, но может терять в точности • randomized – наиболее быстрый метод, но нестабильный, если **Количество итераций степенного метода** большое (5 и больше)
Seed	Ручной ввод числового значения Значение больше 0 По умолчанию — 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске
Исключить оригинальные предикторы	Чекбокс	Выбор данного чекбокса указывает, что необходимо удалить оригинальные предикторы

Таблица 23 Параметры узла «PCA»

Результаты выполнения узла:

- Таблица с примером данных. Отображаются первые 100 наблюдений.



PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_10	PC_11
-797.035	0.967	0.551	-1.272	-0.019	0.253	0.008	0.839	0.353	-0.757	-0.26
-796.085	7.234	1.13	1.841	-1.43	-0.389	1.06	0.258	-0.244	-0.163	-0.166
-794.875	-19.399	1.135	-2.079	-0.135	0.864	-0.552	-0.225	0.265	0.317	0.073
-793.921	-13.097	1.731	-2.076	-0.288	0.777	-0.549	-0.229	0.255	0.326	0.075
-793.078	5.867	-0.608	-1.416	-0.811	0.51	-0.59	-0.338	0.21	-0.502	0.053
-791.756	-31.163	7.956	-2.245	-0.459	-0.419	-0.52	1.112	-1.306	0.096	-0.157
-790.744	-35.417	2.773	-1.609	0.104	0.358	-1.057	-0.04	-1.476	0.411	0.121
-790.436	48.27	-8.235	-0.315	3.598	0.183	0.353	1.019	0.413	0.215	0.053
-789.128	11.757	-1.96	-2.442	-0.343	0.732	-0.58	-0.335	0.055	-0.696	-0.04
-788.436	48.285	-8.243	-0.313	3.597	0.184	0.353	1.018	0.413	0.215	0.053
-787.073	6.15	0.387	-3.686	0.073	0.043	0.073	0.949	0.274	-0.146	-0.217

Рисунок 111 Таблица с примером данных

В данной таблице будут отображены вычисленные компоненты (переменные **PC_1**, **PC_2** и т.д.).

- Столбчатая диаграмма, на которой отображены компоненты (их номера) и значения объясненной ими дисперсии (какая доля общего разброса точек приходится на каждую из новых координат).

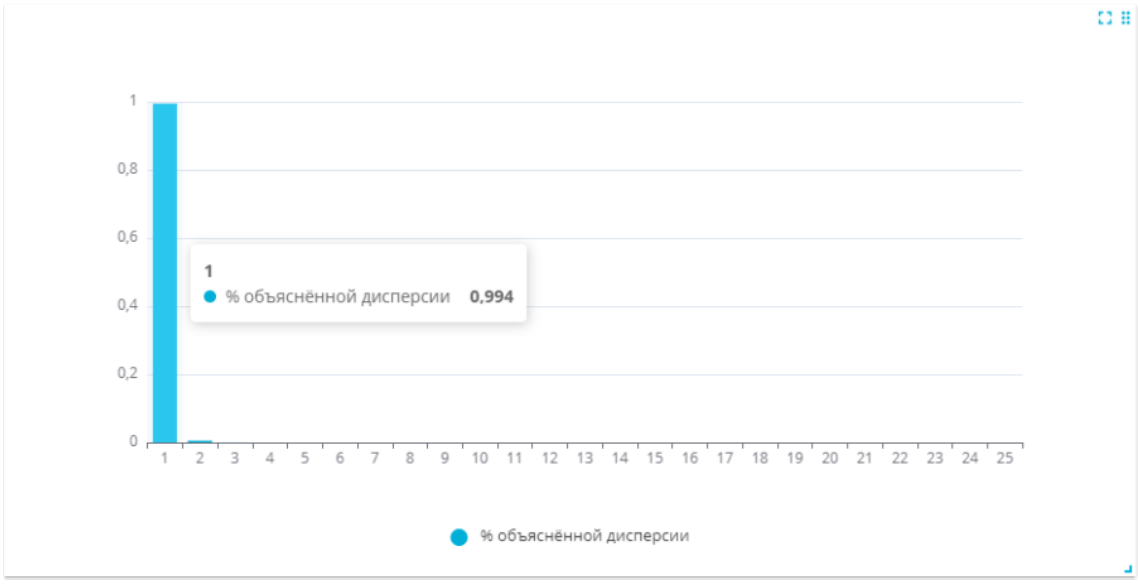


Рисунок 112 Пример столбчатой диаграммы

- Столбчатая диаграмма аналогичная первой, но с кумулятивной суммой.

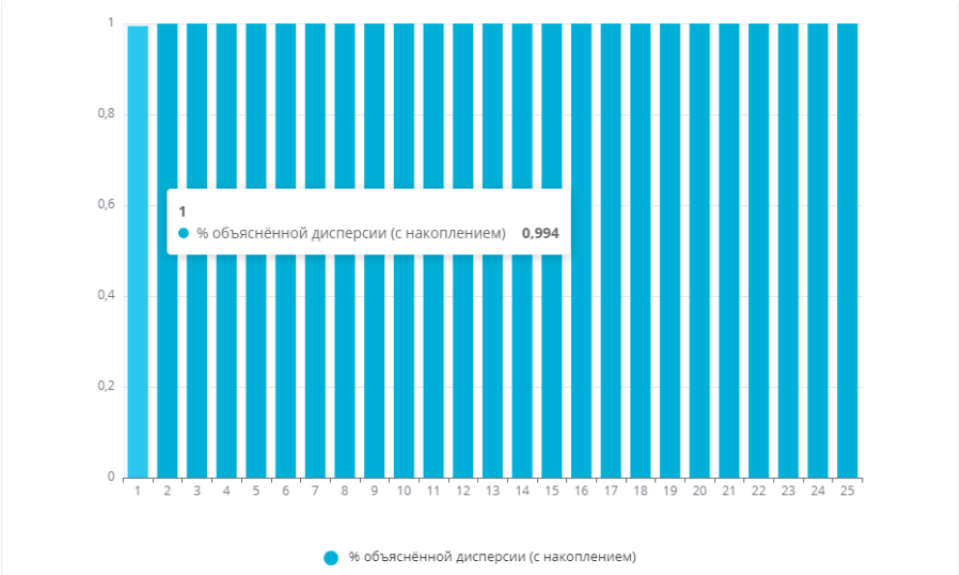


Рисунок 113 Пример столбчатой диаграммы с кумулятивной суммой

3.2.5.6.15. Узел «Профилирование»

Узел «Профилирование» позволяет исследовать данные с целью выяснения статистических характеристик переменных.

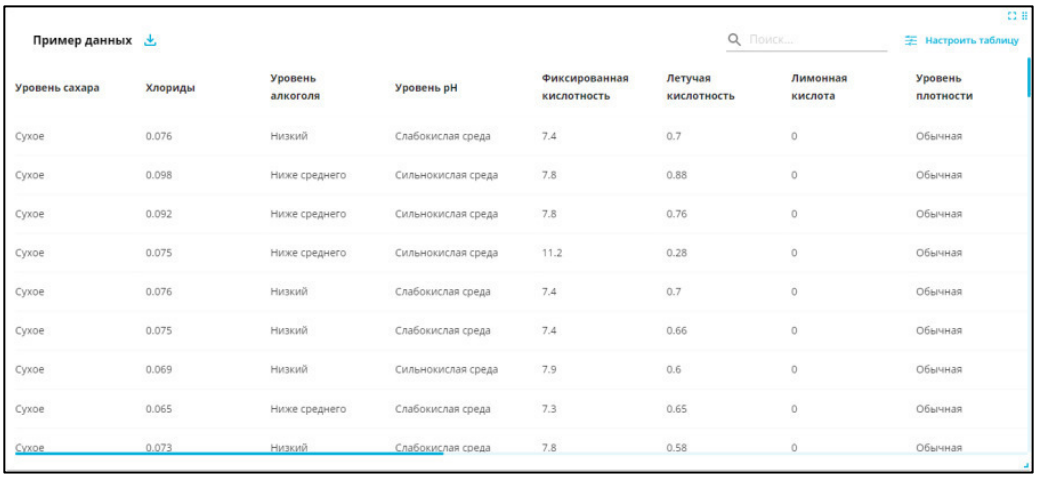
Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла

Таблица 24 Параметры узла «Веса классов»

Результаты выполнения узла:

- **Таблица с примером данных.** Отображаются первые 100 наблюдений.



Уровень сахара	Хлориды	Уровень алкоголя	Уровень pH	Фиксированная кислотность	Летучая кислотность	Лимонная кислота	Уровень плотности
Сухое	0.076	Низкий	Слабокислая среда	7.4	0.7	0	Обычная
Сухое	0.098	Ниже среднего	Сильнокислая среда	7.8	0.88	0	Обычная
Сухое	0.092	Ниже среднего	Сильнокислая среда	7.8	0.76	0	Обычная
Сухое	0.075	Ниже среднего	Сильнокислая среда	11.2	0.28	0	Обычная
Сухое	0.076	Низкий	Слабокислая среда	7.4	0.7	0	Обычная
Сухое	0.075	Низкий	Слабокислая среда	7.4	0.66	0	Обычная
Сухое	0.069	Низкий	Сильнокислая среда	7.9	0.6	0	Обычная
Сухое	0.065	Ниже среднего	Слабокислая среда	7.3	0.65	0	Обычная
Сухое	0.073	Низкий	Слабокислая среда	7.8	0.58	0	Обычная

Рисунок 114 Таблица с примером данных

- **Профиль данных.**

Профиль данных рассчитывается по всему набору данных и по каждой из переменных и зависит от ее типа (для категориальных и количественных переменных разный набор статистик).

Для того, чтобы отобразить статистики по нужной переменной, необходимо выбрать ее в списке.

Рисунок 115 Пример профиля данных

Общие статистики

Для всех наблюдений набора данных считается следующий набор статистик (**пункт Статистика по набору данных**):

- Количество значений.
- Количество уникальных значений.
- Процент уникальных значений (в процентах).
- Количество дублирующих строк.
- Процент дублирующих строк (в процентах).

Категориальные переменные

Для категориальных переменных рассчитывается набор статистик в соответствии с таблицей ниже.

Статистика	Описание
Количество уникальных значений	Количество неповторяющихся значений
Процент уникальных значений	Процент уникальных значений от общего количества значений
Количество пропущенных значений	Количество пропущенных значений
Процент пропущенных значений	Процент пропущенных значений от общего количества значений
Максимальная длина	Количество знаков максимального по длине значения
Минимальная длина	Количество знаков минимального по длине значения
Top	3 значения с максимальной частотой
Bottom	3 значения с минимальной частотой

Таблица 25 Набор статистик, рассчитываемых для категориальных переменных

Для удобства интерпретации и анализа также предусмотрено построение **Облака слов**.

Облако слов (или облако тегов) визуализирует частоту появления значения переменной. Размер облака отражает частоту появления значения. Цветовая гамма не несет в себе смысла и выполняет исключительно эстетическую функцию.

Посмотреть **Облако слов** можно в том же контейнере с профилированием, выбрав в правом верхнем углу иконку  .

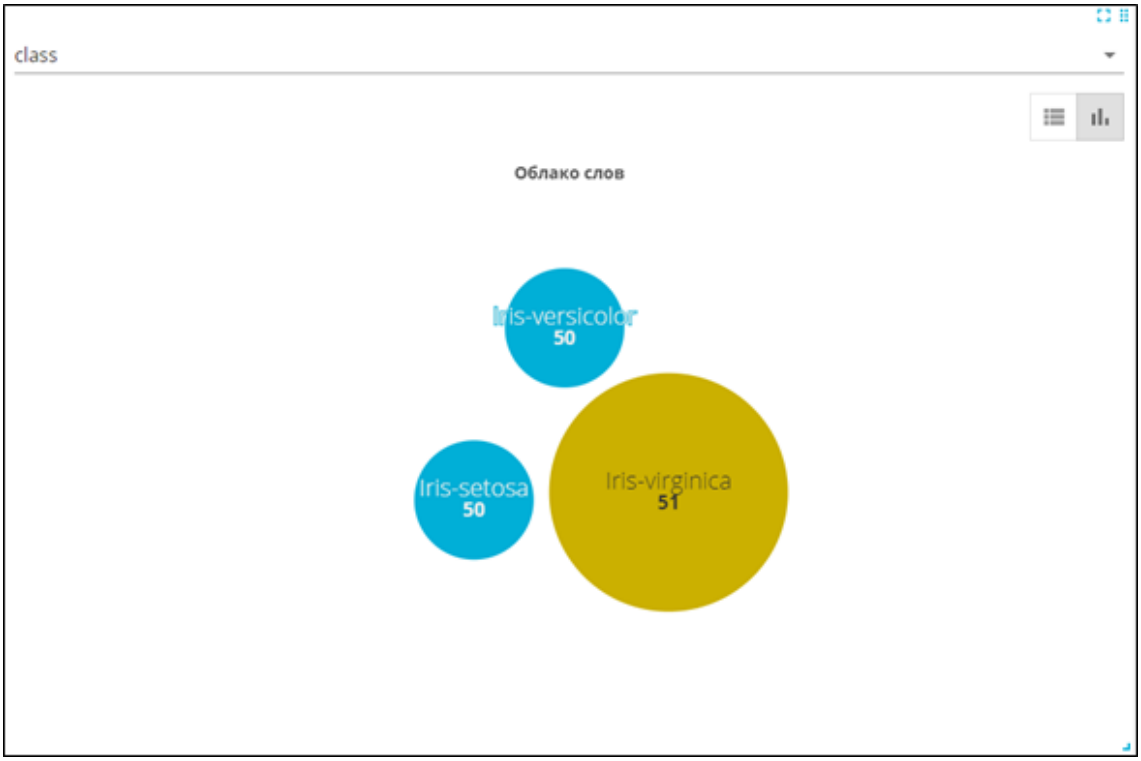


Рисунок 116 Пример Облака слов

Количественные переменные

Для количественных переменных посчитывается набор статистик в соответствии с таблицей ниже.

Статистика	Описание
Количество уникальных значений	количество неповторяющихся значений переменной
Процент уникальных значений	процент неповторяющихся значений переменной
Количество пропущенных значений	количество пропущенных значений переменной
Процент пропущенных значений	процент пропущенных значений
Минимальное значение	наименьшее значение переменной

Статистика	Описание
Максимальное значение	наибольшее значение переменной
Среднее значение	сумма всех значений переменной, разделенная на число этих значений
5-я персентиль	это некоторое значение X из данного ряда, которое делит все имеющиеся в нем значения на две группы: 5% значений, которые меньше X, и оставшиеся значения (то есть 95%), которые превышают X.
95-я персентиль	это некоторое значение X из данного ряда, которое делит все имеющиеся в нем значения на две группы: 95% значений, которые меньше X, и оставшиеся значения (то есть 5%), которые превышают X.
1-я квартиль	25-я персентиль
3-я квартиль	75-я персентиль
Медиана	значение, которое делит распределение пополам (его площадь в т.ч.): половина значений больше медианы, половина — не больше.
Межквартильный размах	разница между 3-м и 1-м квартилями
Количество выбросов	Выбросами считаются наблюдения, которые отклоняются от своего математического ожидания более чем на три среднеквадратических отклонения (правило трех сигм).
Коэффициент вариации	величина, равная отношению стандартного (среднеквадратичного) отклонения случайной величины к ее математическому ожиданию. Он применяется для сравнения вариативности одного и того же признака в нескольких совокупностях с различным средним арифметическим. Если значение коэффициента вариации не превышает 33%, то совокупность считается однородной, а если больше 33%, то — неоднородной.
Коэффициент эксцесса	характеризует меру высоты графика. Если коэффициент больше нуля, то распределение является более высоким («островершинным») относительно «эталонного» нормального распределения. Если коэффициент ниже нуля, то более низким и пологим.
Медианное абсолютное отклонение	вычисляется как медиана абсолютного значения для каждого значения минус медианное значение группы. Является статистикой, более устойчивой к выбросам в наборе данных, чем стандартное отклонение
Асимметрия	характеризует меру скошенности графика влево/вправо. Если коэффициент асимметрии отрицателен, то скос левосторонний. Если коэффициент положителен, то скос правосторонний. И чем коэффициент больше по модулю, тем сильнее скос распределения.
Стандартное отклонение	статистическая характеристика распределения случайной величины, показывающая среднюю степень разброса значений величины относительно математического ожидания. Большее значение среднеквадратического отклонения показывает больший разброс наблюдаемых значений признака относительно среднего; меньшее значение, соответственно, показывает, что величины в множестве сгруппированы вокруг среднего.
Дисперсия	величина, которая характеризует меру разброса значений случайной величины относительно ее математического ожидания.

Статистика	Описание
Т-статистика	Т-статистика — это разница между средним по выборке и гипотетическим средним (предполагаемым равным нулю), деленная на расчетную стандартную ошибку среднего.
Пи-значение	Уровень значимости — вероятность получить Т-значение, равное или превышающее то значение, которое мы в действительности рассчитали по имеющимся выборочным данным (при условии, что нулевая гипотеза верна)

Таблица 26 Набор статистик, рассчитываемых для количественных переменных

Для удобства интерпретации и анализа предусмотрено построение **Гистограммы**.

Гистограмма визуализирует распределение данных в рамках непрерывного интервала. На горизонтальной оси отмечаются интервалы (бины), а на вертикальной оси отмечается частота попаданий наблюдений в каждый интервал. Количество бинов не изменяемо и по умолчанию равно 50.

Посмотреть гистограмму можно в том же контейнере, выбрав в правом верхнем углу иконку  .

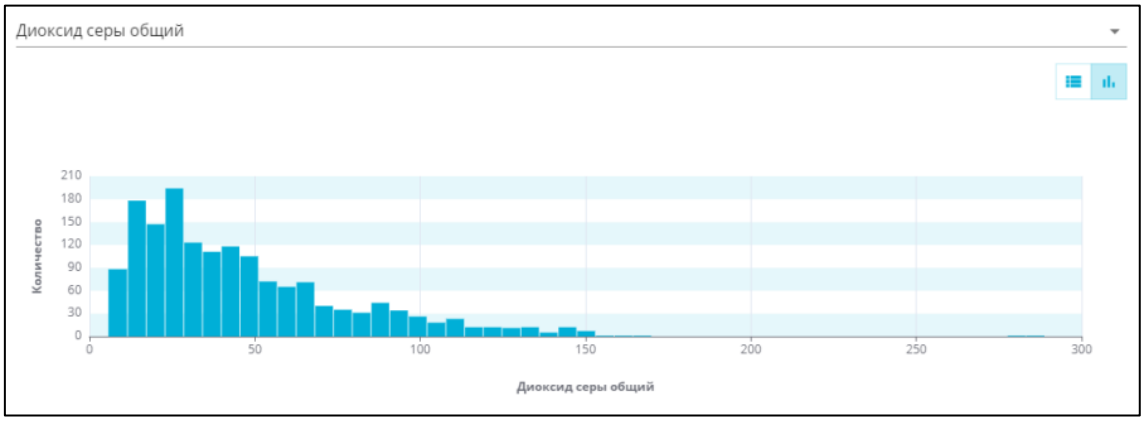


Рисунок 117 Пример Гистограммы

3.2.5.7. Группа узлов «Обучение без учителя»

3.2.5.7.1. Узел «Кластерный анализ (k-means)»

Узел «Кластерный анализ (k-means)» используется для кластеризации набора данных в отдельные группы (кластеры) исходя из выявленных шаблонов во входном наборе данных. Наблюдения группируются таким образом, чтобы они были похожи друг на друга внутри кластера, но различались с наблюдениями из других кластеров.

Алгоритм работы: Модель k-средних определяет начальный набор центроидов для кластеров (исходя из параметров **Количество кластеров** и **Метод инициализации кластеров**). Затем каждое наблюдение определяется в кластер с наиболее близким центроидом. Центроиды кластеров обновляются в соответствии с набором наблюдений, назначенным в каждый кластер. Далее итерационно проверяется необходимо ли переназначить наблюдение в другой кластер. Данный процесс продолжается до момента достижения максимального числа итераций (параметр **Максимальное количество итераций**).

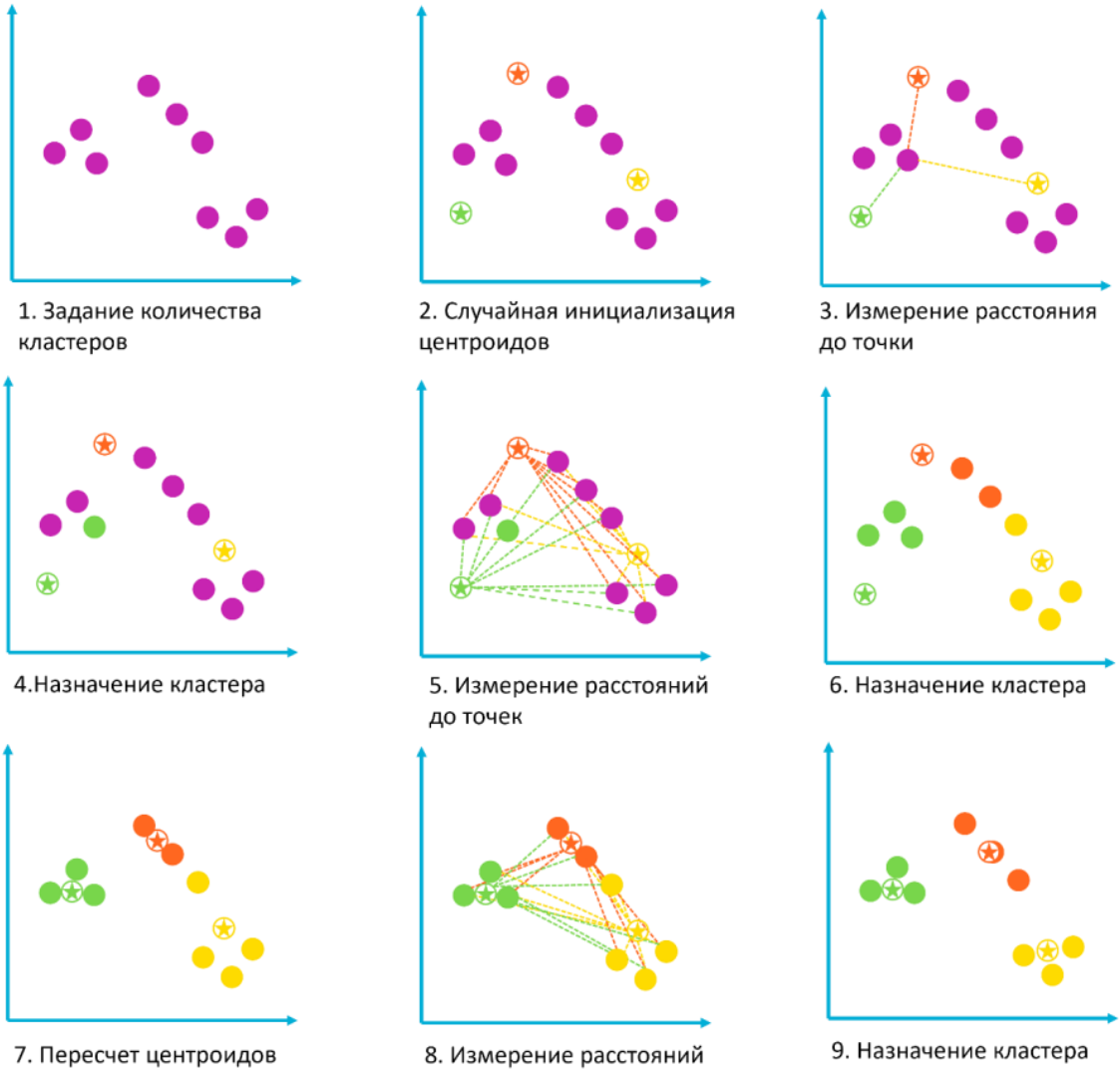


Рисунок 118 Принцип работы узла «Кластерный анализ (k-means)»

Запуск алгоритма с определением начального набора центроидов происходит ограниченное количество раз (в соответствии с заданным параметром **Количество запусков**). После выполнения всех запусков выбирается запуск с минимальным критерием **инерции** (суммой квадратов расстояний между точками и центроидом внутри кластеров).

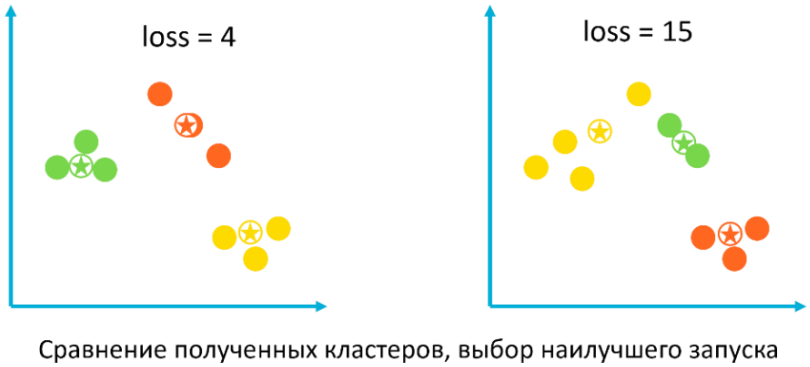


Рисунок 119 Принцип работы узла «Кластерный анализ (k-means)»

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Количество кластеров	Ручной ввод целочисленного значения По умолчанию — 5	Задание числа кластеров, на которые будет делиться пространство признаков. Для определения количества кластеров можно воспользоваться априорной информацией об исходных данных в разделе Исследования данных при Кластеризации исходного набора данных.
Seed	Ручной ввод целочисленного значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске узла
Метод инициализации кластеров	Раскрывающийся список со следующими значениями: • k-means++ (по умолчанию) • Forgy	Данный параметр отвечает за выбор метода инициализации начальных точек кластеров. Предусмотрены следующие методы: • **k-means++** Идея метода k-means++ состоит в том, чтобы случайным образом выбрать начальные точки, которые находятся как можно дальше друг от друга. • **Forgy** Метод Forgy случайным образом выбирает k наблюдений (по числу заданных кластеров) из набора данных и использует их в качестве начальных значений.
Стандартизация	Раскрывающийся список со следующими значениями: • Нет • Стандартное отклонение (по умолчанию) • Диапазон	Данный параметр отвечает за выбор метода стандартизации числовых переменных. **Стандартизация** – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью. Предусмотрены следующие методы: • **Нет**. • **Стандартное отклонение** – преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • **Диапазон** – линейно преобразует значения переменных в диапазон [0, 1].

Параметр	Возможные значения и ограничения	Описание
Количество запусков	Ручной ввод целочисленного значения По умолчанию — 10	Данный параметр задает число запусков алгоритма с разными начальными центроидами
Максимальное количество итераций	Ручной ввод целочисленного значения По умолчанию — 300	Данный параметр задает максимальное количество итераций в рамках одного запуска
Алгоритм K-средних	Раскрывающийся список со следующими значениями: • elkan • full	Данный параметр отвечает за выбор алгоритма k-средних. • **elkan** – может быть более эффективным для некоторых наборов данных с четко определенными кластерами за счет использования неравенства треугольника. Однако требует больше памяти • **full** – классический алгоритм
Размер выборки	Ручной ввод целочисленного значения По умолчанию — 1000	Данный параметр задает размер выборки, которая будет отображена на графике **Силуэт** в результатах узла
Расстояние между кластерами	Ручной ввод целочисленного значения По умолчанию — 25	Данный параметр задает расстояние между кластерами на графике **Силуэт** в результатах узла
Количество бинов	Ручной ввод целочисленного значения По умолчанию — 10	Данный параметр задает количество бинов, на которое будет делиться количественная переменная на **графике с параллельными осями** в результатах узла
Переменные, по которым делать оси	Раскрывающийся список с выбором нескольких переменных	Данный параметр задает переменные, которые будут отражены на **графике с параллельными осями** в результатах узла
Максимальное количество линий	Ручной ввод По умолчанию — 50	Данный параметр задает максимальное количество линий, которые будут отображаться на **графике с параллельными осями** в результатах узла

Таблица 27 Параметры узла «Кластерный анализ (k-means)»

Результаты выполнения узла:

- Круговая диаграмма с количеством наблюдений по кластерам.

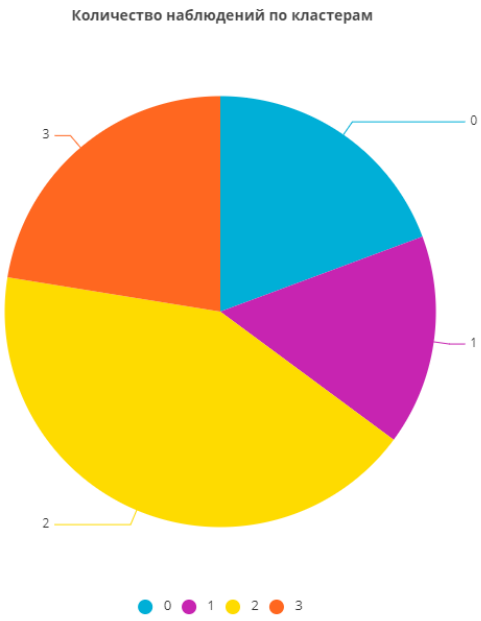


Рисунок 120 Пример Круговой диаграммы с результатами кластеризации

При наведении курсора мыши на сектор кластера можно узнать количество наблюдений в нем.

- Силуэт – Silhouette Plot

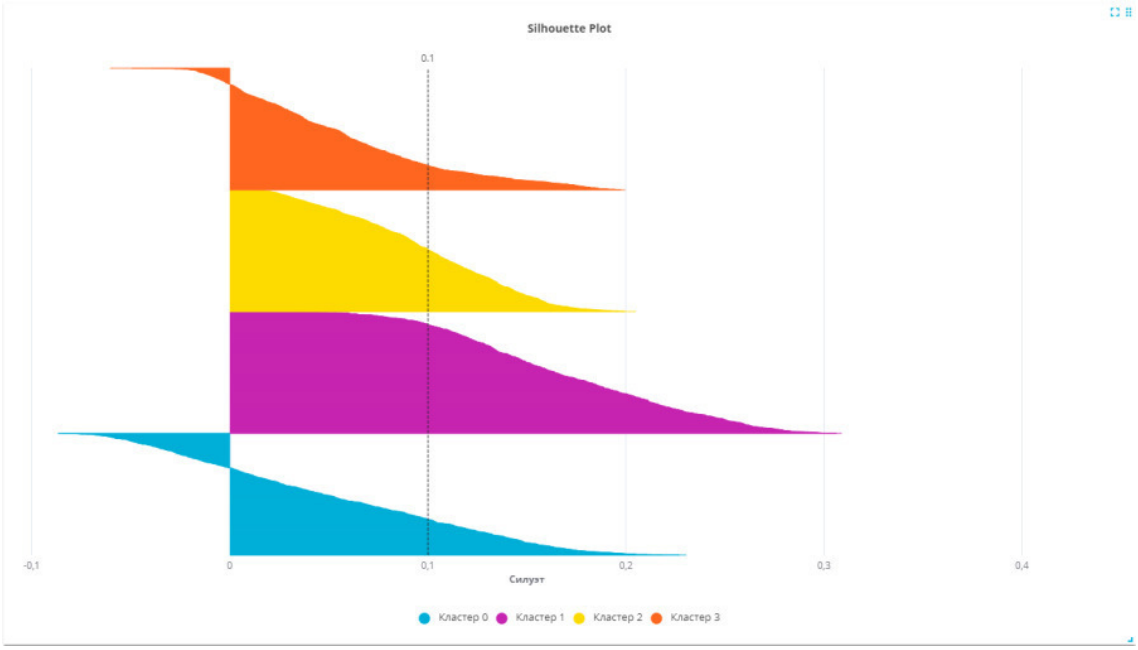


Рисунок 121 Пример Silhouette Plot

Значение Silhouette для каждого наблюдения (на графике отображается указанное в параметре **Размер выборки число наблюдений**) является мерой того, насколько это наблюдение похоже на наблюдения в собственном кластере по сравнению с наблюдениями в других кластерах.

Значение Silhouette находится в диапазоне от -1 до 1. Высокое значение указывает на то, что наблюдение хорошо соответствует собственному кластеру и плохо соответствует другим кластерам.

Если большинство наблюдений имеют низкое или отрицательное значение Silhouette, тогда пользователь должен перестроить кластеризацию с большим или меньшим количеством кластеров.

- График в параллельных координатах.

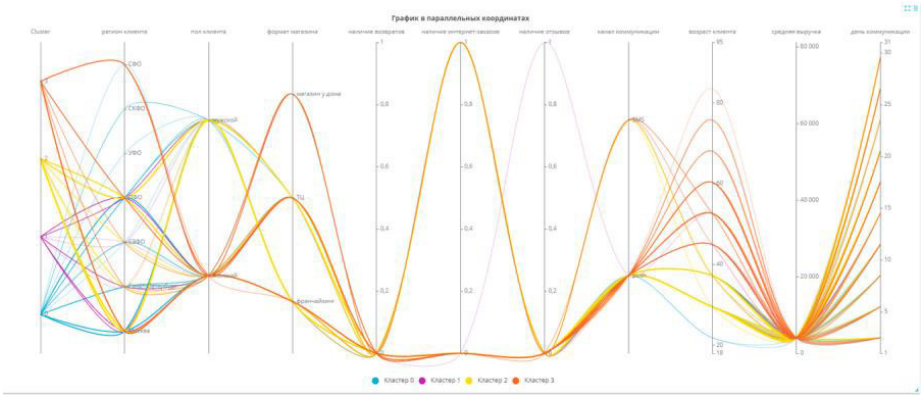
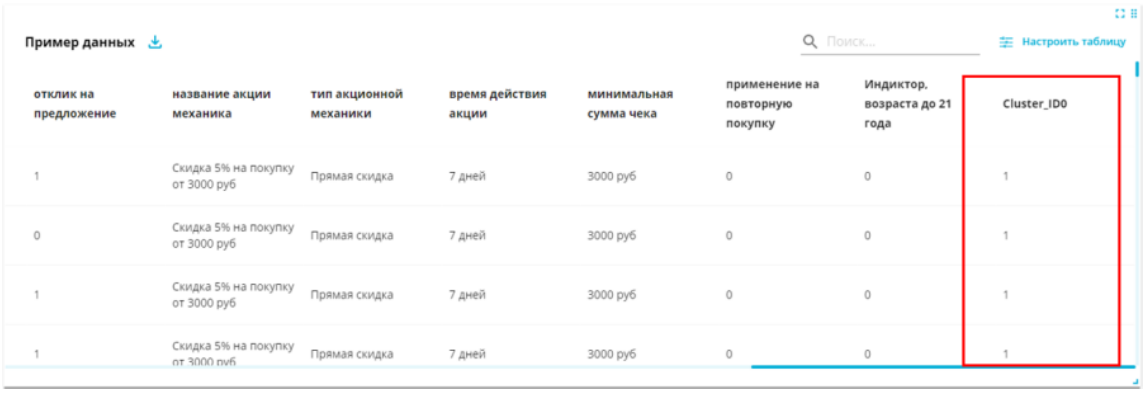


Рисунок 122 Пример графика в параллельных координатах

График в параллельных координатах позволяет интерпретировать построенные кластеры.

На данном графике каждой переменной присваивается собственная ось (согласно параметру **Переменные, по которым делать оси**). Оси располагаются параллельно друг другу, и каждая имеет свою собственную шкалу. Начальная ось отражает кластер, к которому модель отнесла наблюдение. Каждое наблюдение наносится на график в виде линии (параметр **Максимальное количество линий**), пересекающейся с каждой из осей. Таким образом, пользователь может выявить паттерны и корреляции между разными переменными.

- Таблица с примером данных. Отображаются первые 100 наблюдений.

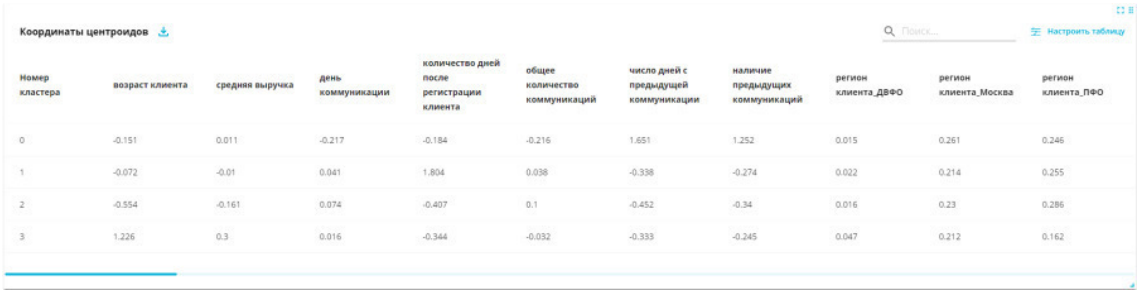


отклик на предложение	название акции механика	тип акционной механики	время действия акции	минимальная сумма чека	применение на повторную покупку	Индиктор, возраста до 21 года	Cluster_ID0
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	1
0	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	1
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	1
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	1

Рисунок 123 Таблица с примером данных

В результате выполнения узла будет рассчитана новая переменная с результатами кластеризации (переменная **Cluster_ID0**).

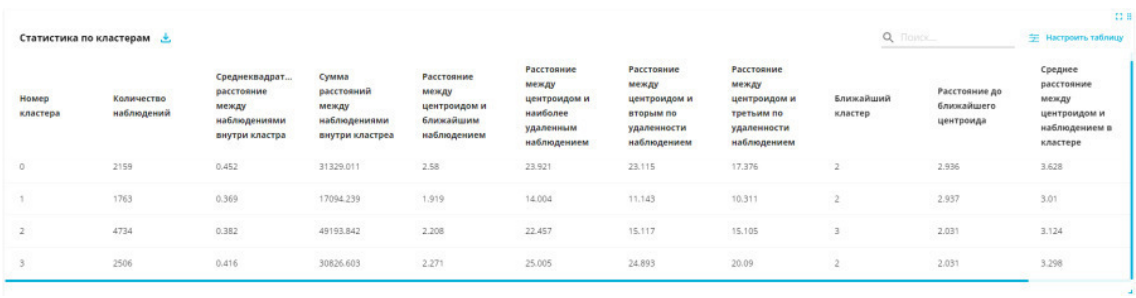
- Таблица с координатами центроидов, где в качестве строк выступают номера кластеров и значения переменных, в которых находятся центроиды этих кластеров.


Координаты центроидов

Номер кластера	возраст клиента	средняя выручка	день коммуникации	количество дней после регистрации клиента	общее количество коммуникаций	число дней с предыдущей коммуникации	наличие предыдущих коммуникаций	регион клиента_ДВФО	регион клиента_Москва	регион клиента_ПФО
0	-0.151	0.011	-0.217	-0.184	-0.216	1.651	1.252	0.015	0.261	0.246
1	-0.072	-0.01	0.041	1.804	0.038	-0.338	-0.274	0.022	0.214	0.255
2	-0.554	-0.161	0.074	-0.407	0.1	-0.452	-0.34	0.016	0.23	0.286
3	1.226	0.3	0.016	-0.344	-0.032	-0.333	-0.245	0.047	0.212	0.162

Рисунок 124 Пример таблицы с координатами центроидов кластеров

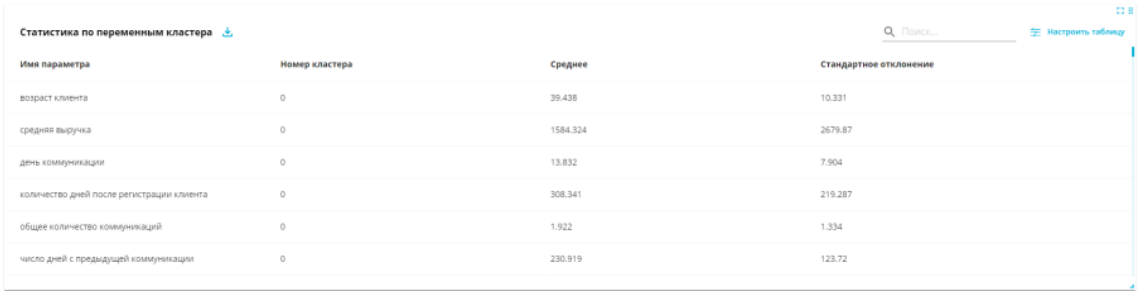
- Таблица со статистиками по кластерам
  - Номер кластера.
  - Количество наблюдений.
  - Среднеквадратичное расстояние между наблюдениями внутри кластера.
  - Сумма расстояний между наблюдениями внутри кластера.
  - Расстояние между центроидом и ближайшим наблюдением.
  - Расстояние между центроидом и наиболее удаленным наблюдением.
  - Расстояние между центроидом и вторым по удаленности наблюдением.
  - Расстояние между центроидом и третьим по удаленности наблюдением.
  - Ближайший кластер.
  - Расстояние до ближайшего центроида.
  - Среднее расстояние между центроидом и наблюдениями в кластере.
  - Сумма расстояний между наблюдениями и центроидом.


Статистика по кластерам

Номер кластера	Количество наблюдений	Среднеквадрат... расстояние между наблюдениями внутри кластра	Сумма расстояний между наблюдениями внутри кластреа	Расстояние между центроидом и ближайшим наблюдением	Расстояние между центроидом и наиболее удаленным наблюдением	Расстояние между центроидом и вторым по удаленности наблюдением	Расстояние между центроидом и третьим по удаленности наблюдением	Ближайший кластер	Расстояние до ближайшего центроида	Среднее расстояние между центроидом и наблюдением в кластере
0	2159	0.452	31329.011	2.58	23.921	23.115	17.376	2	2.936	3.628
1	1763	0.369	17094.239	1.919	14.004	11.143	10.311	2	2.937	3.01
2	4734	0.382	49193.842	2.208	22.457	15.117	15.105	3	2.031	3.124
3	2506	0.416	30826.603	2.271	25.005	24.893	20.09	2	2.031	3.298

Рисунок 125 Пример со статистиками по кластерам

- Таблица со статистиками по переменным кластера. По каждому кластеру отражены среднее и стандартное отклонение для каждой переменной.



Статистика по переменным кластера

Поиск... Настроить таблицу

Имя параметра	Номер кластера	Среднее	Стандартное отклонение
возраст клиента	0	39.438	10.331
средняя выручка	0	1584.324	2679.87
день коммуникации	0	13.832	7.904
количество дней после регистрации клиента	0	308.341	219.287
общее количество коммуникаций	0	1.922	1.334
число дней с предыдущей коммуникации	0	230.919	123.72

Рисунок 126 Пример таблицы со статистиками по переменным кластера

3.2.5.7.2. Узел «Иерархическая кластеризация»

В основе **узла «Иерархическая кластеризация»** лежит алгоритм кластеризации, направленный на создание иерархии вложенных кластеров.

Алгоритм работы: Каждое наблюдение начинается в своем собственном кластере (Шаг 0), далее кластеры последовательно объединяются. Так, первоначально рассчитываются расстояния (расчет расстояния задает параметр **Метрика**) между наблюдениями, ближайшие из них объединяются в один кластер. Параметр **Критерий объединения** определяет стратегию слияния кластеров. Затем вычисляется расстояние между кластерами и ближайшие объединяются в один большой кластер. Слияние кластеров происходит до тех пор, пока не будет синтезирован один единый кластер (Шаг N).

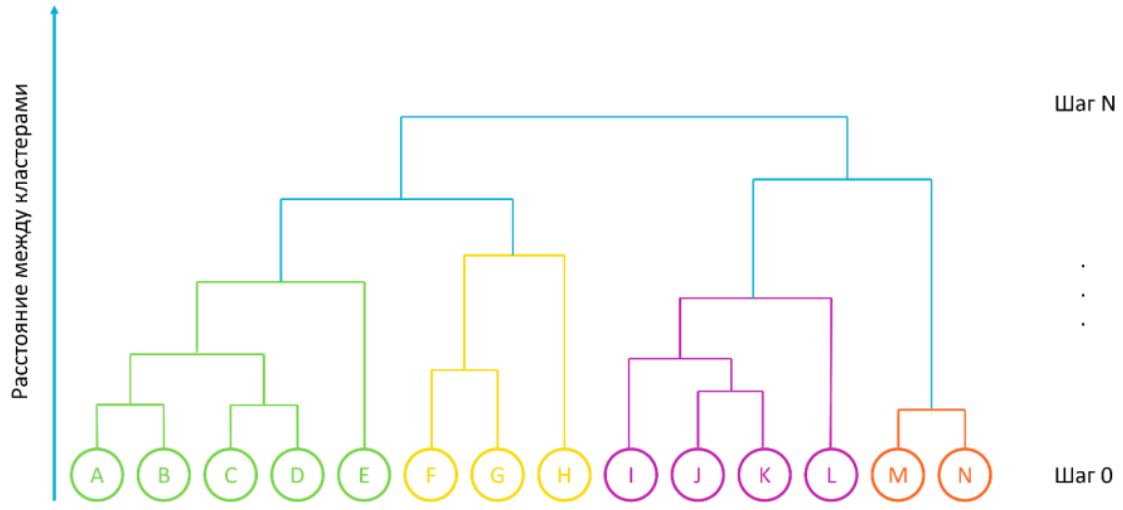


Рисунок 127 Схема работы алгоритма Иерархической кластеризации

Для остановки алгоритма необходимо указать в параметре **Критерий остановки** требуемый вариант – по достижению заданного количества кластеров, либо по минимальному расстоянию между кластерами.

Для определения значения количества кластеров или расстояния между кластерами рекомендуется воспользоваться дендрограммой в результатах узла.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Критерий остановки	Раскрывающийся список со следующими значениями: • Количество кластеров (по умолчанию) • Расстояние	Данный параметр отвечает за выбор критерия остановки алгоритма. Предусмотрены следующие методы: • **Количество кластеров** Достигнуто указанное в параметре **Количество кластеров** значение • **Расстояние** Достигнуто указанное в параметре **Минимальное расстояние** значение.
Количество кластеров	Ручной ввод целочисленного значения Число больше или равно 1 По умолчанию — 5	Данный параметр задает число кластеров, на которые будет делиться пространство признаков. Действителен при выбранном **Критерии остановки Количество кластеров**
Минимальное расстояние	Ручной ввод Число больше или равно 0 По умолчанию — 0	Данный параметр задает минимальное расстояние между кластерами для остановки алгоритма. Действителен при выбранном **Критерии остановки Расстояние** Для определения значения минимального расстояния можно воспользоваться **Дендрограммой** в результатах узла.
Стандартизация	Раскрывающийся список со следующими значениями: • Нет • Стандартное отклонение • Диапазон	Данный параметр отвечает за выбор метода стандартизации числовых переменных. **Стандартизация** – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью. Предусмотрены следующие методы: • Нет. • Стандартное отклонение – преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • Диапазон – линейно преобразует значения переменных в диапазон [0, 1].

Параметр	Возможные значения и ограничения	Описание
Метрика	Раскрывающийся список со следующими значениями: • Евклидова метрика • Манхэттенская метрика • Косинус	Данный параметр отвечает за выбор метрики, которая задает расчет расстояния между наблюдениями. Выбор метрики влияет на форму кластеров, поскольку некоторые элементы могут быть относительно ближе друг к другу по одной метрике, чем по другой. Предусмотрены следующие метрики: • Евклидова метрика • Манхэттенская метрика • Косинус
Критерий объединения	Раскрывающийся список со следующими значениями: • Ward (можно использовать только если в качестве метрики близости наблюдений выбрана Евклидова метрика) • Average • Maximum • Minimum	Метрика, используемая для объединения кластеров. Предусмотрены следующие метрики близости кластеров: • Ward минимизирует сумму квадратов разностей во всех кластерах • Average минимизирует среднее расстояние между всеми наблюдениями пар кластеров. • Maximum сводит к минимуму максимальное расстояние между наблюдениями пар кластеров • Minimum минимизирует расстояние между ближайшими наблюдениями пар кластеров
Seed	Ручной ввод целочисленного значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске узла
Размер выборки	Ручной ввод целочисленного значения Значение больше или равно 2 По умолчанию — 1000	Данный параметр задает размер выборки, которая будет отображена на графике **Силуэт** в результатах узла
Расстояние между кластерами	Ручной ввод Значение больше или равно 0 По умолчанию — 25	Данный параметр задает расстояние между кластерами на графике **Силуэт** в результатах узла
Количество бинов	Ручной ввод целочисленного значения Значение больше или равно 1 По умолчанию — 10	Данный параметр задает количество бинов, на которое будет делиться количественная переменная на **графике с параллельными осями** в результатах узла
Переменные, по которым делать оси	Раскрывающийся список с выбором нескольких переменных	Данный параметр задает переменные, которые будут отражены на **графике с параллельными осями** в результатах узла

Параметр	Возможные значения и ограничения	Описание
Максимальное количество линий	Ручной ввод целочисленного значения Значение больше или равно 1 По умолчанию — 50	Данный параметр задает максимальное количество линий, которые будут отражены на **графике с параллельными осями** в результатах узла

Таблица 28 Список параметров узла "Иерархическая кластеризация"

Параметры узла «Иерархическая кластеризация»

Результаты выполнения узла:

- Круговая диаграмма с количеством наблюдений по кластерам.

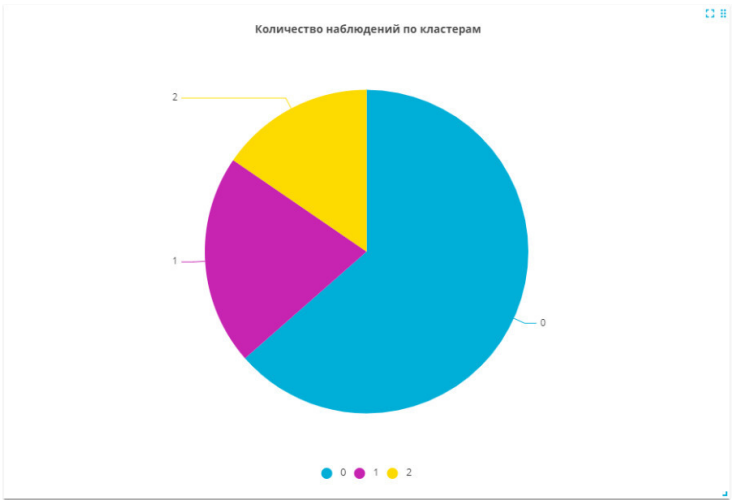


Рисунок 128 Пример круговой диаграммы

При наведении курсора мыши на сектор кластера можно узнать количество наблюдений в нем.

- Силуэт – Silhouette Plot.

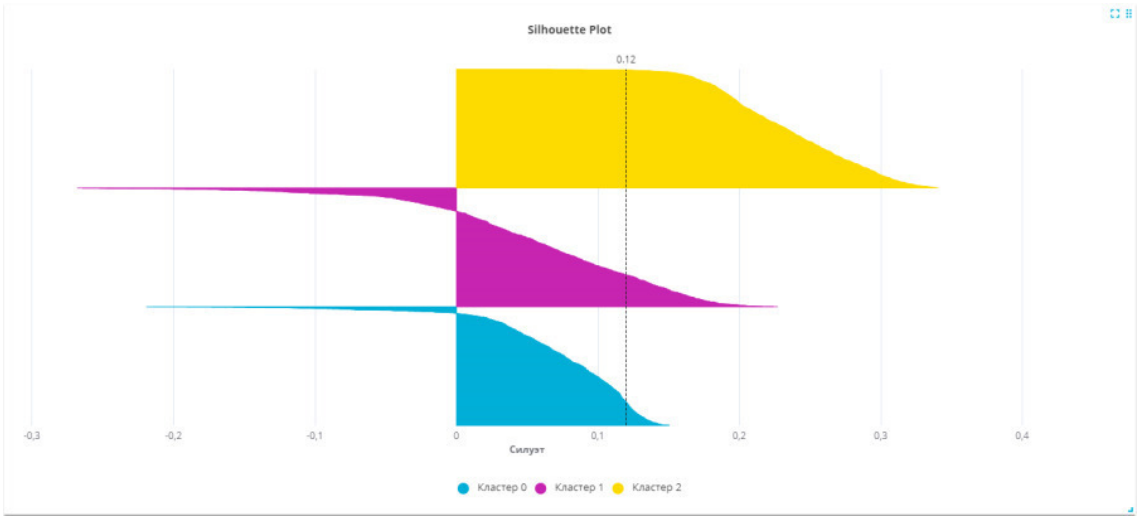


Рисунок 129 Пример Silhouette Plot

Значение Silhouette для каждого наблюдения (на графике отображается указанное в параметре **Размер выборки число наблюдений**) является мерой того, насколько это наблюдение похоже на наблюдения в собственном кластере по сравнению с наблюдениями в других кластерах.

Значение Silhouette находится в диапазоне от -1 до 1. Высокое значение указывает на то, что наблюдение хорошо соответствует собственному кластеру и плохо соответствует другим кластерам.

Если большинство наблюдений имеют низкое или отрицательное значение Silhouette, тогда пользователь должен перестроить кластеризацию с большим или меньшим количеством кластеров.

- График в параллельных координатах.

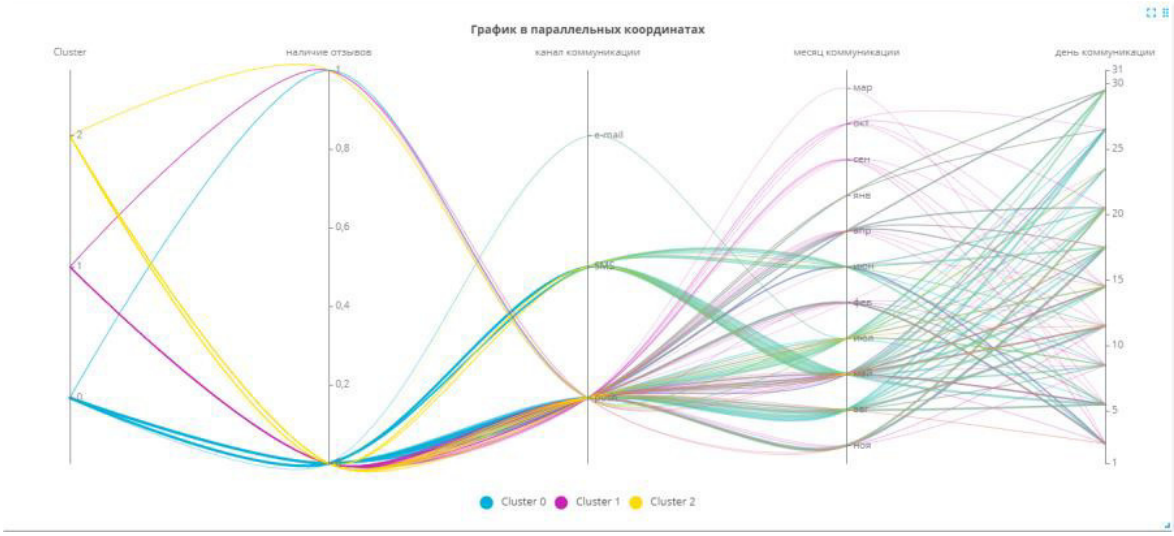


Рисунок 130 Пример графика в параллельных координатах

График в параллельных координатах позволяет интерпретировать построенные кластеры.

На данном графике каждой переменной присваивается собственная ось (согласно параметру **Переменные, по которым делать оси**). Оси располагаются параллельно друг другу, и каждая имеет свою собственную шкалу. Начальная ось отражает кластер, к которому модель отнесла наблюдение. Каждое наблюдение наносится на график в виде линии (параметр **Максимальное количество линий**), пересекающейся с каждой из осей. Таким образом, пользователь может выявить паттерны и корреляции между разными переменными.

- Дендрограмма кластеризации.

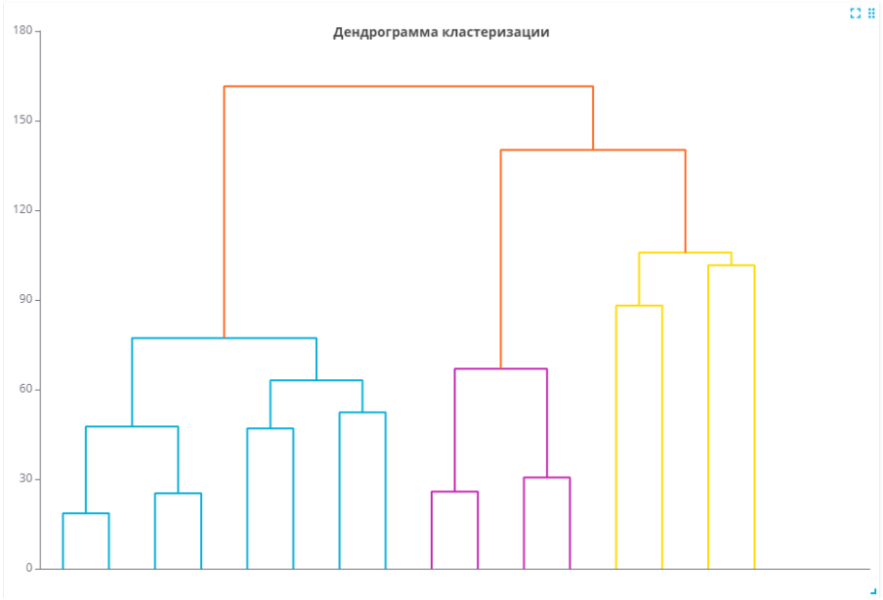
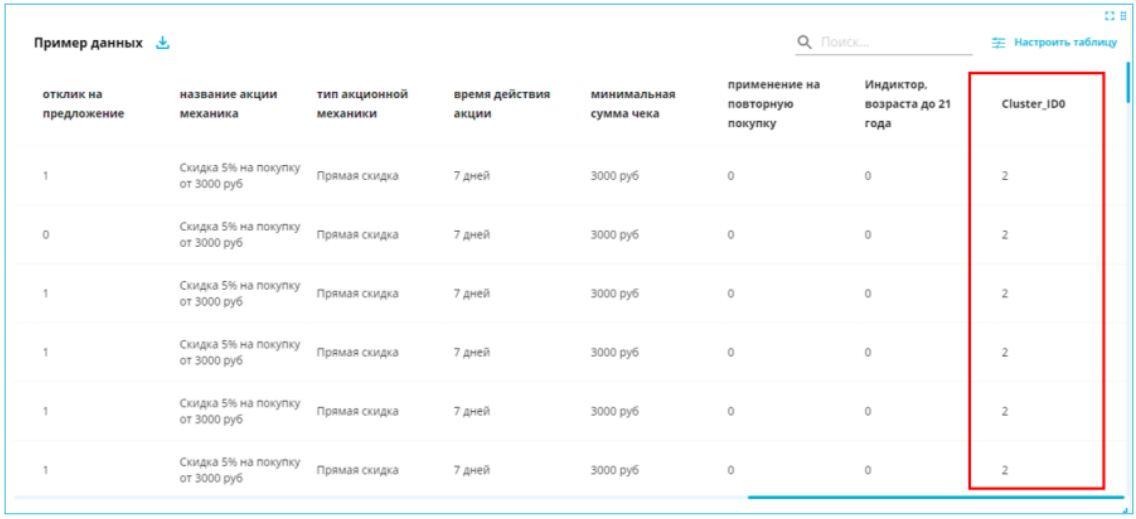


Рисунок 131 Пример дендограммы

Дендрограмма показывает близость отдельных наблюдений и кластеров, а также последовательность их объединения. Количество уровней соответствует количеству слияний кластеров. По оси Y расположена шкала, на которой откладывается расстояние между объектами в пространстве признаков.

- Таблица с примером данных. Отображаются первые 100 наблюдений.

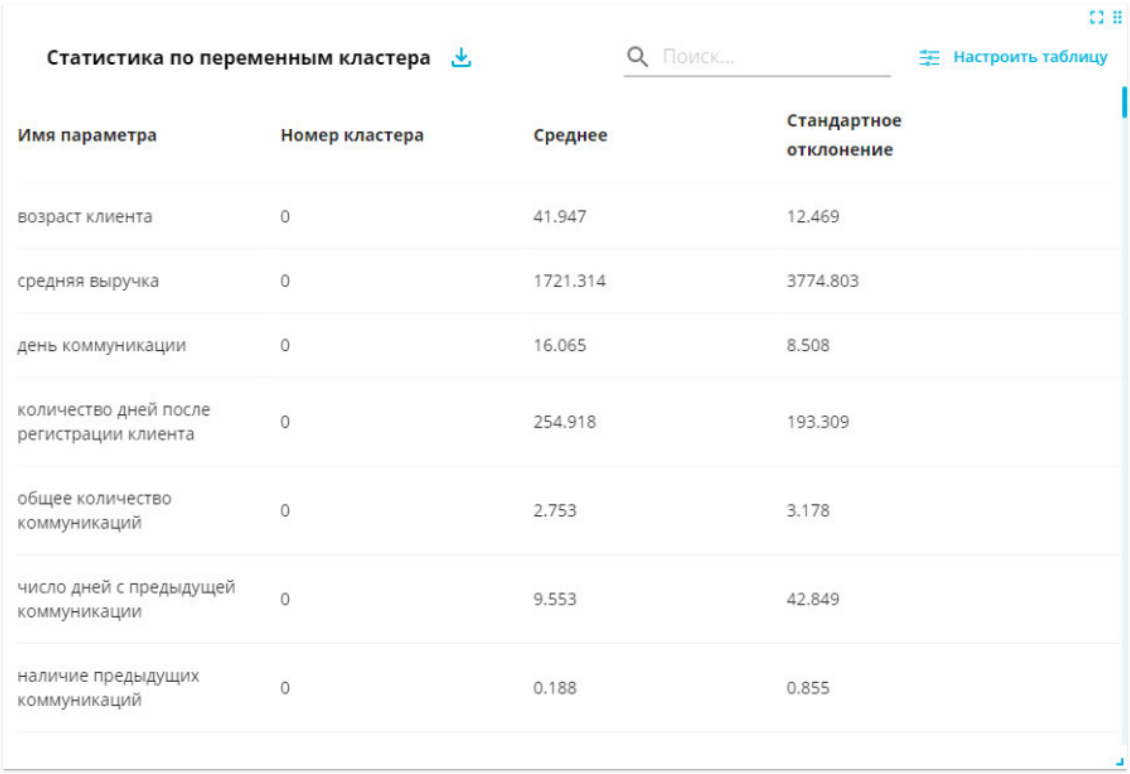


отклик на предложение	название акции механика	тип акционной механики	время действия акции	минимальная сумма чека	применение на повторную покупку	Индиктор, возраста до 21 года	Cluster_ID0
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2
0	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2
1	Скидка 5% на покупку от 3000 руб	Прямая скидка	7 дней	3000 руб	0	0	2

Рисунок 132 Пример таблицы

В результате выполнения узла будет рассчитана новая переменная с результатами кластеризации (переменная **Cluster_ID0**).

- Таблица со статистиками по переменным кластера. По каждому кластеру отражены среднее и стандартное отклонение для каждой переменной.



Статистика по переменным кластера

Поиск...

Настроить таблицу

Имя параметра	Номер кластера	Среднее	Стандартное отклонение
возраст клиента	0	41.947	12.469
средняя выручка	0	1721.314	3774.803
день коммуникации	0	16.065	8.508
количество дней после регистрации клиента	0	254.918	193.309
общее количество коммуникаций	0	2.753	3.178
число дней с предыдущей коммуникации	0	9.553	42.849
наличие предыдущих коммуникаций	0	0.188	0.855

Рисунок 133 Пример таблицы со статистиками по переменным кластера

3.2.5.8. Группа узлов «Обучение с учителем»

3.2.5.8.1. Узел «Дерево решений»

В основе **узла «Дерево решений»** лежит алгоритм, обобщающий наблюдения правилами вида «Если…, то…» в иерархическую, последовательную структуру в виде дерева. Правила генерируются в процессе обучения.

Алгоритм работы: Процесс построения деревьев решений представляет собой последовательное, рекурсивное разбиение множества наблюдений на подмножества с применением решающих правил в **узлах**. Разбиение продолжается до момента, пока не будет достигнуто условие остановки алгоритма. Последний узел, который не осуществляет проверку и разбиение, становится **листом**.

В основе алгоритма построения дерева решений лежит принцип жадной максимизации прироста информации – на каждом шаге выбирается тот признак, при разделении по которому прирост информации оказывается наибольшим. Дальше процедура повторяется рекурсивно, пока энтропия не окажется равной нулю или какой-то малой величине.

Борьба с переобучением:

- Ограничить максимальную глубину дерева (параметр **Максимальная глубина**)

- Ограничить минимальное число объектов листе (параметр **Минимальное количество наблюдений в листе**)
- Ограничить максимальное количества листьев в дереве (параметр **Максимальное количество листов)**

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Критерий разбиения для классификации	Раскрывающийся список со следующими значениями: • gini (по умолчанию) • entropy	Данный параметр задает критерий разбиения на узлы для классификации. Предусмотрены следующие критерии: • gini (неопределенность Джини) – направлен на максимизацию количества пар объектов, одного класса, оказавшихся в одном поддереве. • entropy (критерий прироста информации, энтропия) – направлен на максимизацию прироста информации
Критерий разбиения для регрессии	Раскрывающийся список со следующими значениями: • squared error (по умолчанию) • friedman mse • absolute error • poisson	Данный параметр задает критерий разбиения для регрессионной задачи. Предусмотрены следующие критерии: • squared error (среднеквадратичная ошибка) • friedman mse (среднеквадратичная ошибка с оценкой улучшения Фридмана) • absolute error (средняя абсолютная ошибка) • poisson (отклонение Пуассона)
Стратегия разбиения	Раскрывающийся список со следующими значениями: • best (по умолчанию) • random	Данный параметр задает стратегию разделения на каждом узле. Предусмотрены следующие стратегии: • best – выбор наилучшей функции сегментации и точки сегментации • random – случайное разделение
Максимальная глубина	Ручной ввод Неотрицательное число По умолчанию — 5	Данный параметр задает максимальную глубину дерева, после достижения которой алгоритм останавливает работу
Минимальное количество наблюдений для разбиения	Ручной ввод Неотрицательное число По умолчанию — 2	Данный параметр задает минимальное количество наблюдений, которое должно быть в разбиении
Минимальное количество	Ручной ввод Неотрицательное	Данный параметр задает минимальное количество наблюдений, которое может быть в листе

Параметр	Возможные значения и ограничения	Описание
наблюдений в листе	число По умолчанию — 5	
Минимальная доля веса наблюдений в листе	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр определяет минимальный весовой коэффициент выборки в листовом узле. По умолчанию наблюдения имеют одинаковый вес
Максимальное количество признаков	Раскрывающийся список со следующими значениями: • all (по умолчанию) • sqrt • log2 • number • frac	Данный параметр определяет максимальное количество признаков, которое будет учитываться при поиске лучшего разделения. Предусмотрены следующие варианты: • all – учитывать все доступные признаки • sqrt – учитывать sqrt(число всех доступных признаков) • log2 – учитывать log2(число всех доступных признаков) • number – учитывать указанное число признаков • frac – учитывать int(указанное число * число всех доступных признаков) При выборе number или frac появится дополнительный параметр Число (вводится int) и Frac (вводится float) соответственно.
Seed	Ручной ввод По умолчанию — 12345	Начальное числовое значение для генератора случайных чисел
Максимальное количество листов	Ручной ввод Неотрицательное число	Данный параметр определяет максимальное количество листов в дереве
Минимальное снижение неоднородности	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр определяет минимальное снижение неоднородности Узел будет разделен, если это разделение вызовет уменьшение неоднородности большее или равное указанному значению
ccp_alpha (отсечение с минимизацией стоимости-сложности) Метод отсечения дерева	Ручной ввод По умолчанию — 0	Данный параметр регулирует количество отсекаемых узлов. Чем больше значение ccp_alpha, тем большее количество узлов удаляется из дерева

Таблица 29 Параметры узла «Дерево решений»

Результаты выполнения узла:

Узел «Дерево решений» имеет разные результаты в зависимости от решаемой задачи.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC .

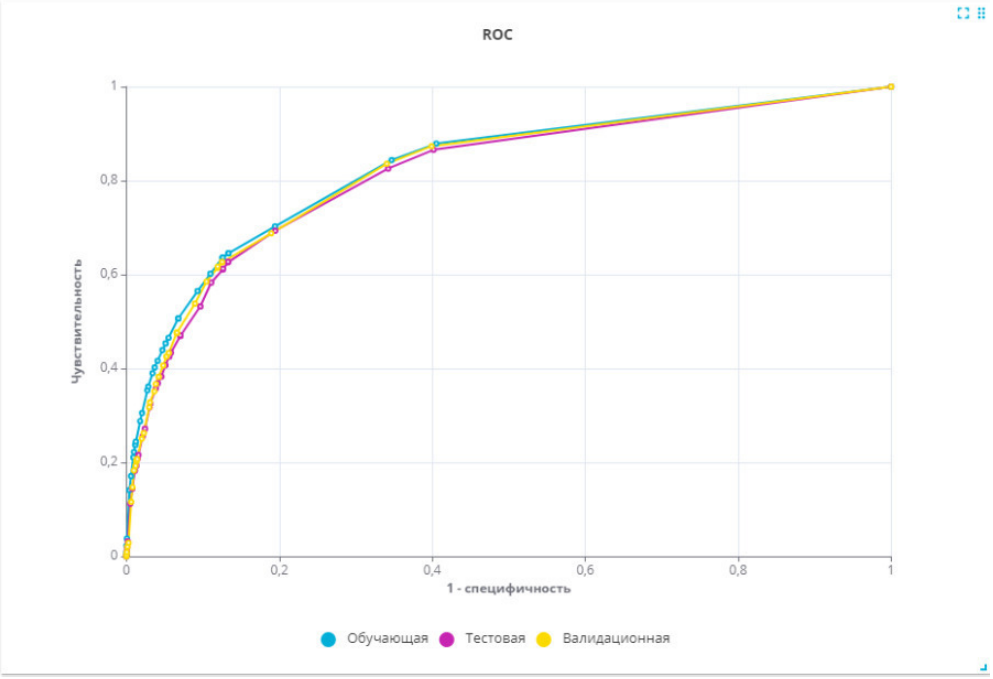


Рисунок 134 Пример графика ROC

- График Lift.

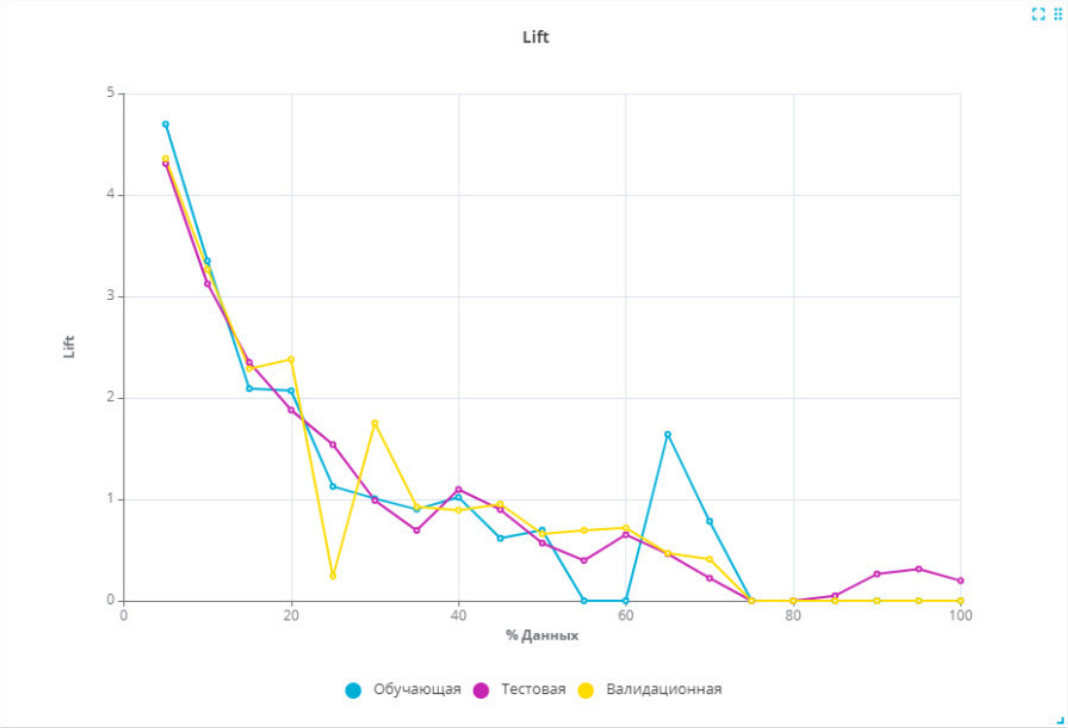


Рисунок 135 Пример графика Lift

- График Cumulative Lift.

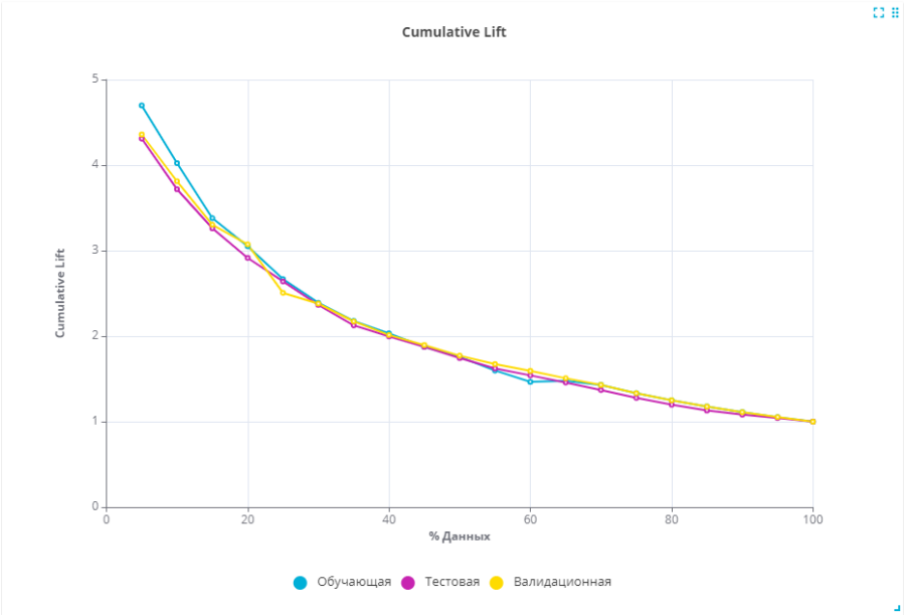


Рисунок 136 Пример графика Cumulative Lift

- График Gain.

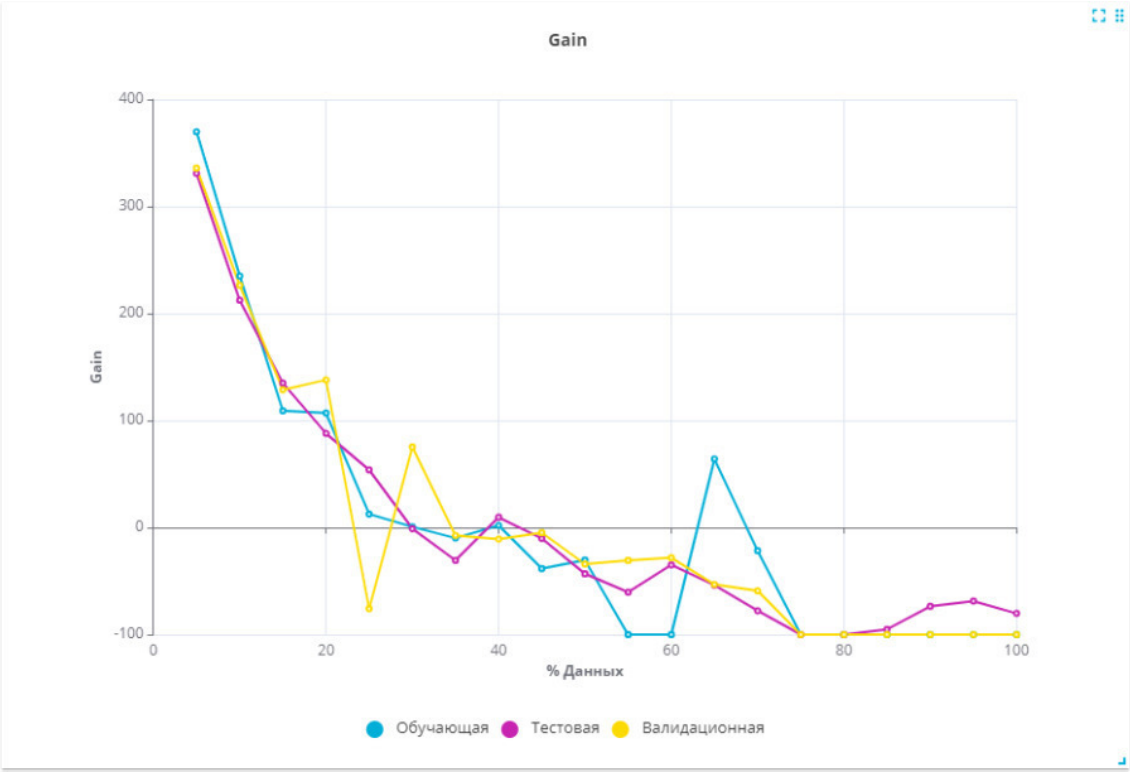


Рисунок 137 Пример графика Gain

- График Cumulative Gain.

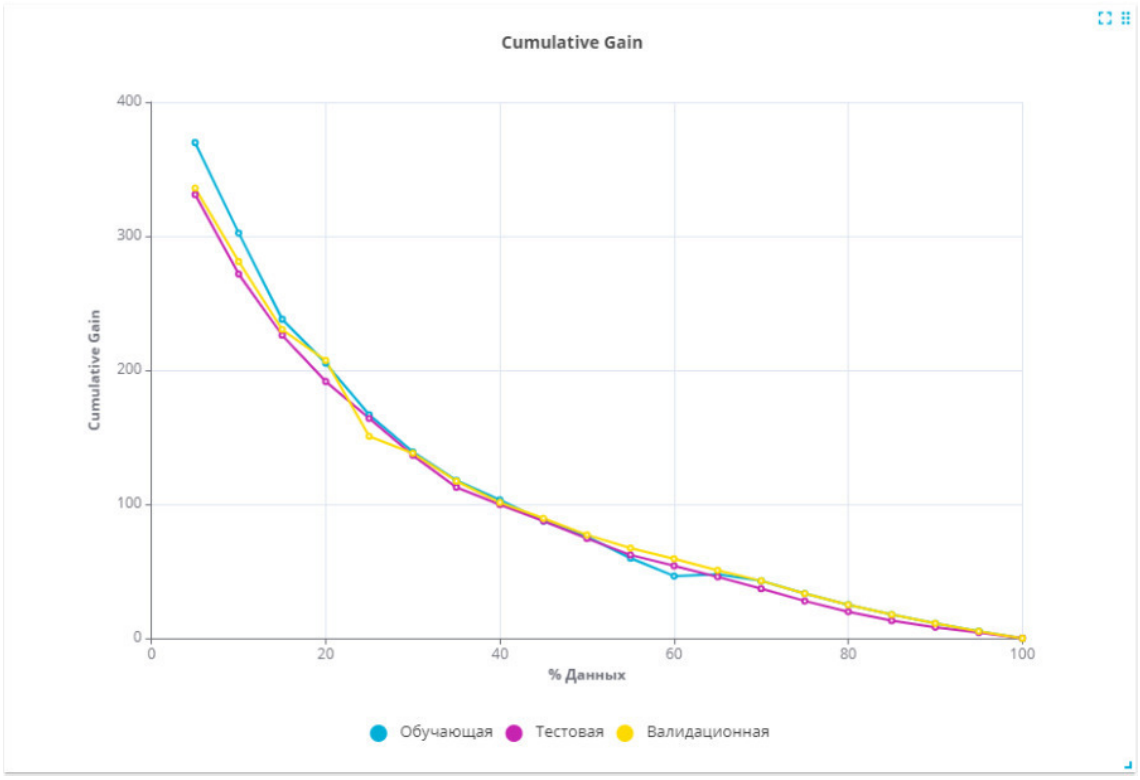


Рисунок 138 Пример графика Cumulative Gain

- Диаграмма дерева решений.

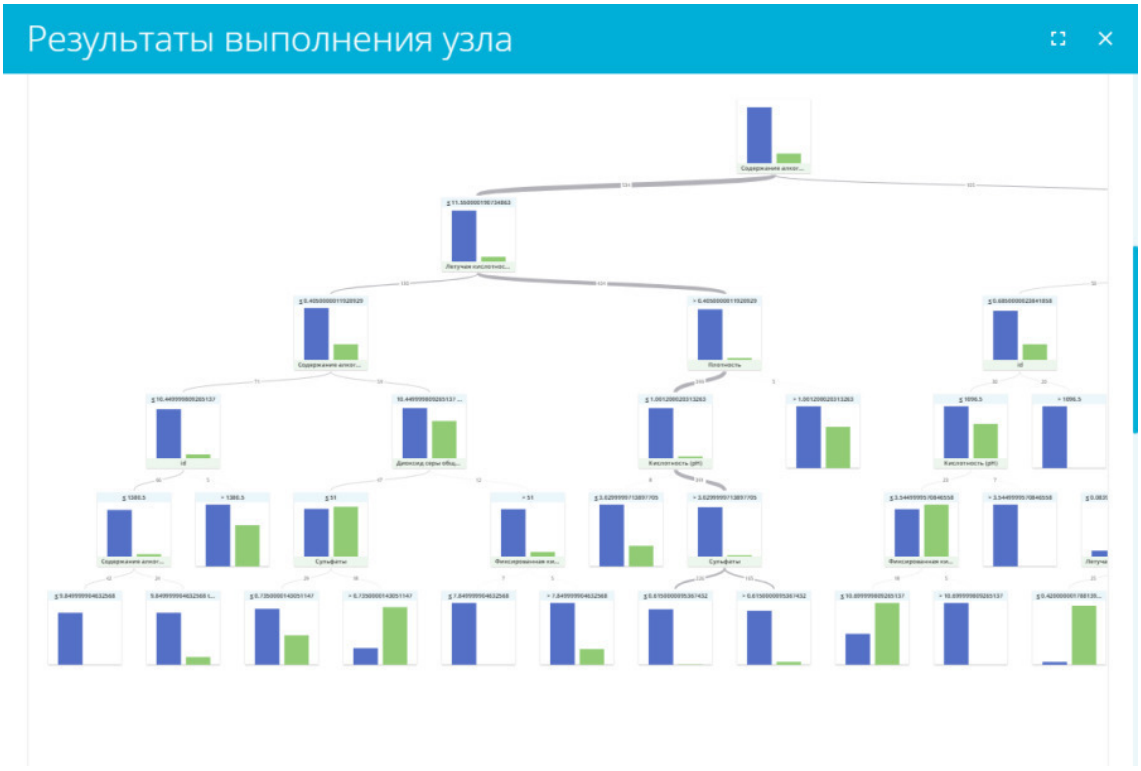
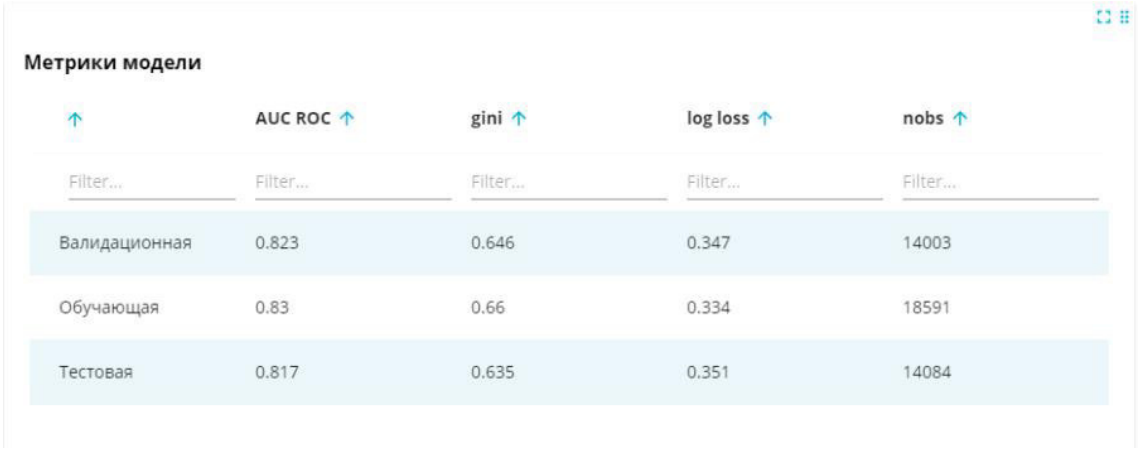


Рисунок 139 Пример графика дерева решений

- Таблица с метриками качества модели.

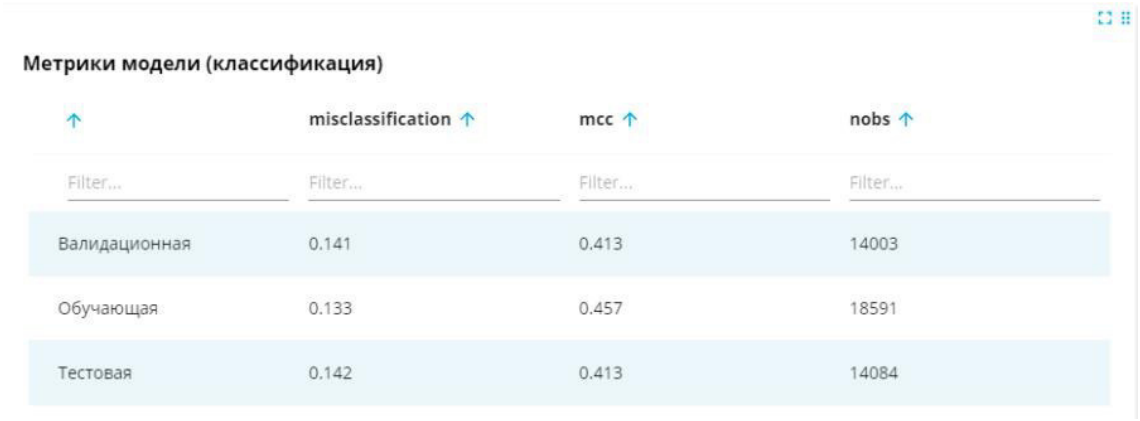


Метрики модели

	AUC ROC	gini	log loss	nobs
Filter...	Filter...	Filter...	Filter...	Filter...
Валидационная	0.823	0.646	0.347	14003
Обучающая	0.83	0.66	0.334	18591
Тестовая	0.817	0.635	0.351	14084

Рисунок 140 Пример таблицы с метриками качества модели

- Таблица с метриками качества модели для задачи классификации.

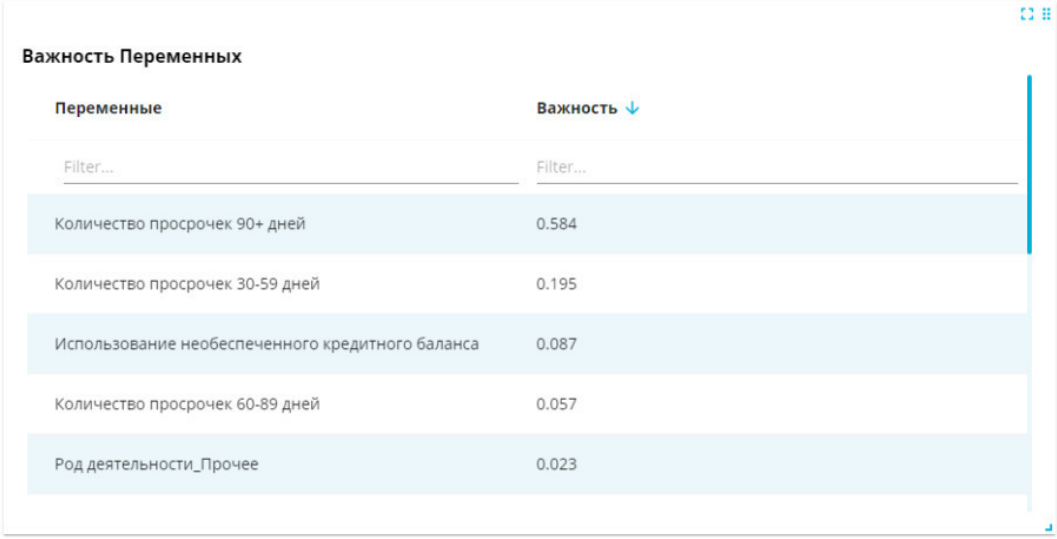


Метрики модели (классификация)

	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.141	0.413	14003
Обучающая	0.133	0.457	18591
Тестовая	0.142	0.413	14084

Рисунок 141 Пример таблицы с метриками качества модели для задачи классификации

- Таблица со списком переменных, сортированных по важности.



Важность Переменных

Переменные	Важность
Filter...	Filter...
Количество просрочек 90+ дней	0.584
Количество просрочек 30-59 дней	0.195
Использование необеспеченного кредитного баланса	0.087
Количество просрочек 60-89 дней	0.057
Род деятельности_Прочее	0.023

Рисунок 142 Пример таблицы со списком переменных, отсортированных по важности

Результаты **многоклассовой классификации** представлены следующими объектами:

- Диаграмма дерева решений.
- Таблица с метриками качества модели.

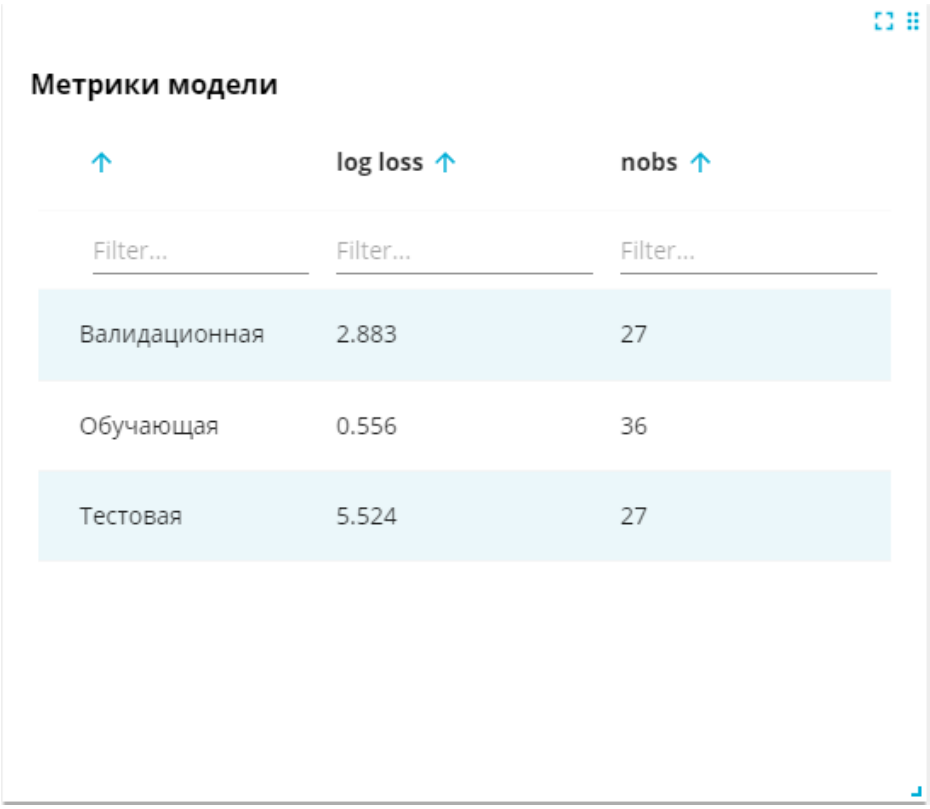


Рисунок 143 Пример таблицы с метриками качества модели

- Таблица с метриками качества модели для задачи классификации.

Рисунок 144 Пример таблицы с метриками качества модели для задачи классификации

- Таблица со списком переменных, сортированных по важности.

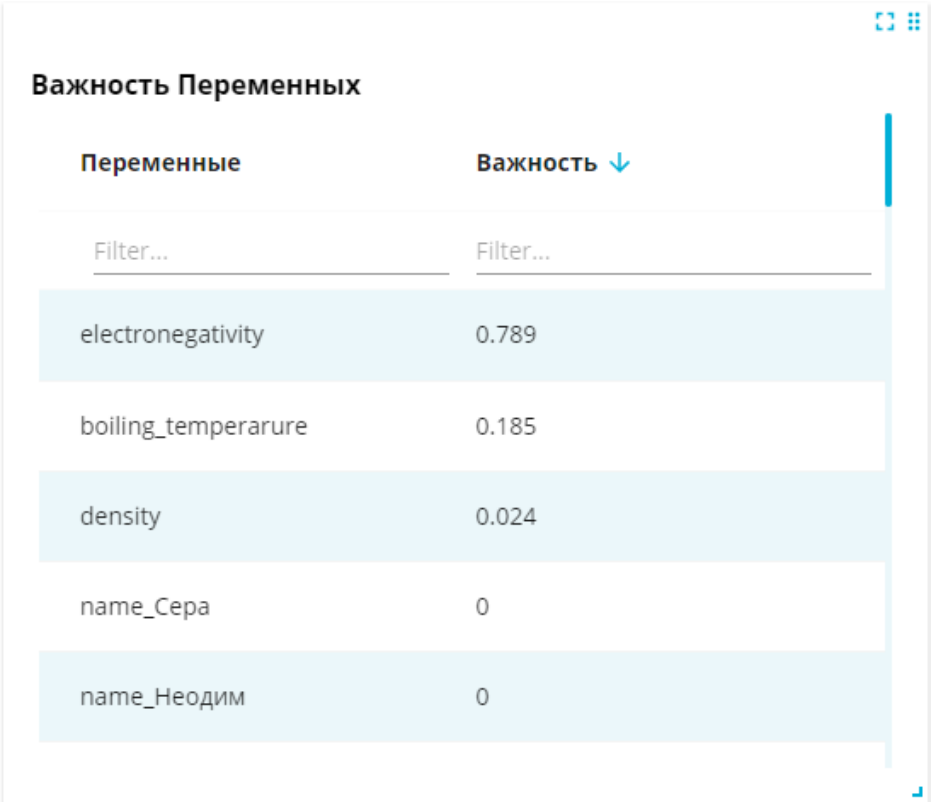


Рисунок 145 Пример таблицы со списком переменных, сортированных по важности

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты обучающей, тестовой и валидационной выборок.

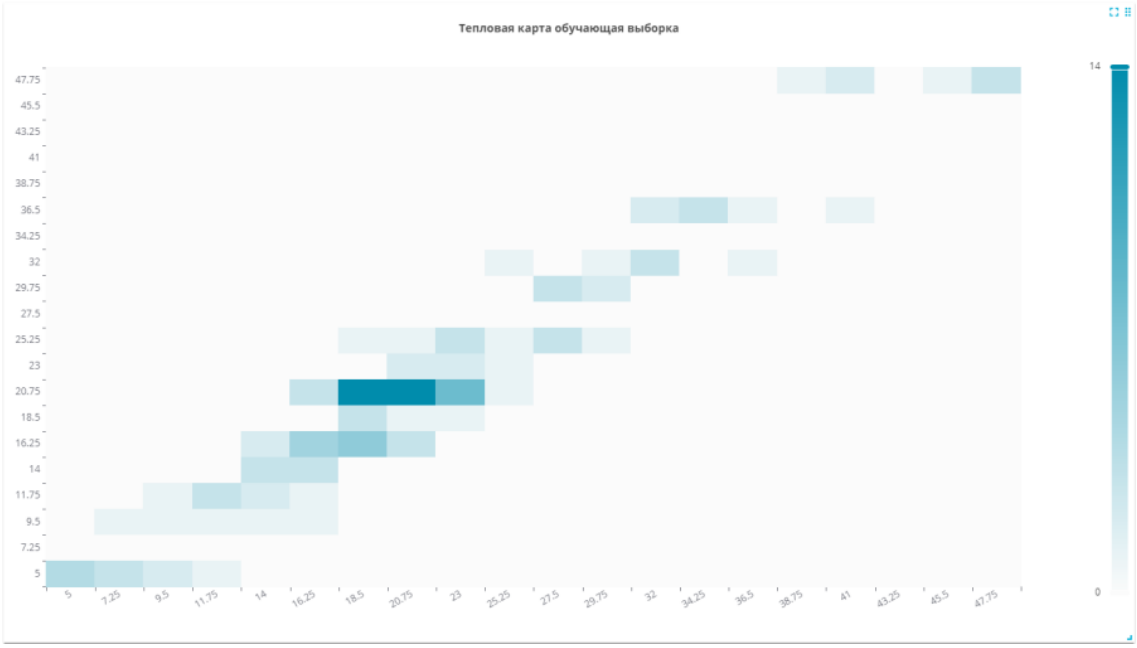


Рисунок 146 Пример тепловой карты на данных обучающей выборки

«Тепловые» карты, отражающие корреляцию или ассоциацию типа хи-квадрат первоначальных значений с целевыми признаками по сегментам.

- Диаграмма дерева решений.

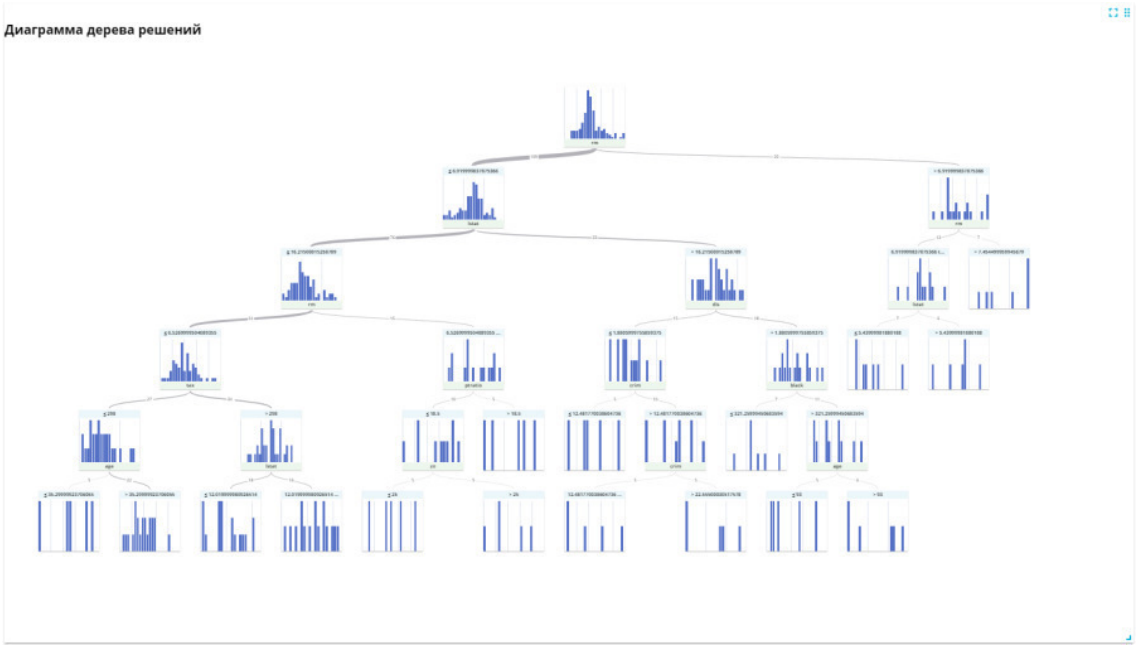
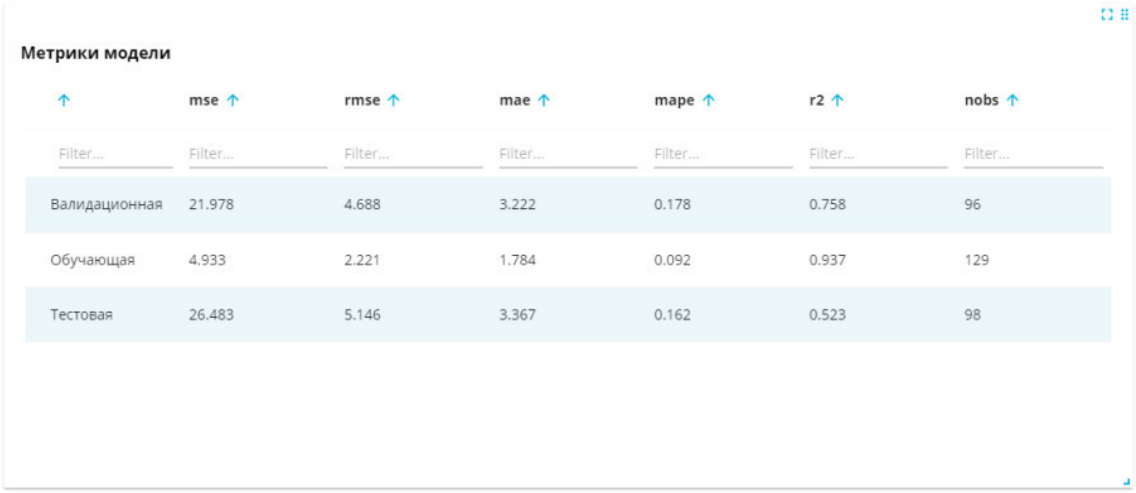


Рисунок 147 Пример диаграммы дерева решений для задачи регрессии

- Таблица с метриками качества модели.



Метрики модели

	mse	rmse	mae	mape	r2	nobs
Валидационная	21.978	4.688	3.222	0.178	0.758	96
Обучающая	4.933	2.221	1.784	0.092	0.937	129
Тестовая	26.483	5.146	3.367	0.162	0.523	98

Рисунок 148 Пример таблицы с метриками качества модели

- Таблица со списком переменных, сортированных по важности (рисунок выше).

3.2.5.8.2. Узел «Случайный лес»

В основе **узла «Случайный лес»** лежит алгоритм машинного обучения, который представляет собой ансамбль деревьев решений.

Алгоритм работы: Ансамблирование — это тип обучения, при котором объединяются различные типы алгоритмов или тот же алгоритм несколько раз, что позволяет сформировать более мощную прогнозную модель. Так, в алгоритме **Случайного леса** каждое дерево предсказывает класс (в случае задачи классификации) или значение (в случае задачи регрессии) на основании своего разбиения, и выбирается то предсказание, которое получило наибольшее количество голосов (в случае задачи классификации) или среднее значение всех предсказанных значений (в случае задачи регрессии).

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Количество деревьев	Ручной ввод Неотрицательное число	Данный параметр определяет количество деревьев в случайном лесу
Критерий разбиения для классификации	Раскрывающийся список со следующими значениями: • Gini • entropy	Данный параметр задает критерий разбиения на узлы для классификации. Предусмотрены следующие критерии: • gini (коэффициент Джини) • entropy (критерий прироста информации, энтропия)
Критерий разбиения для регрессии	Раскрывающийся список со следующими значениями: • squared error • friedman mse • absolute error • poisson	Данный параметр задает критерий разбиения для регрессионной задачи. Предусмотрены следующие критерии: • squared error (среднеквадратичная ошибка) • friedman mse (среднеквадратичная ошибка с оценкой улучшения Фридмана) • absolute error (средняя абсолютная ошибка) • poisson (отклонение Пуассона)
Максимальная глубина	Ручной ввод Неотрицательное число По умолчанию — 5	Данный параметр задает максимальную глубину дерева, после достижения которой алгоритм останавливает работу.
Минимальное количество наблюдений для разбиения	Ручной ввод Неотрицательное число По умолчанию — 2	Данный параметр задает минимальное количество наблюдений, которое должно быть в разбиении
Минимальное количество наблюдений в листе	Ручной ввод Неотрицательное число По умолчанию — 5	Данный параметр задает минимальное количество наблюдений, которое может быть в листе

Параметр	Возможные значения и ограничения	Описание
Максимальное количество признаков	Раскрывающийся список со следующими значениями: • all • sqrt • log2 • number • frac	Данный параметр определяет максимальное количество признаков, которое будет учитываться при поиске лучшего разделения. Предусмотрены следующие варианты: • all – учитывать все доступные признаки • sqrt – учитывать sqrt(число всех доступных признаков) • log2 – учитывать log2(число всех доступных признаков) • number – учитывать указанное число признаков • frac – учитывать int(указанное число * число всех доступных признаков) При выборе number или frac появится дополнительный параметр Число (вводится int) и Frac (вводится float) соответственно.
Seed	Ручной ввод числового значения По умолчанию — 12345	Начальное числовое значение для генератора случайных чисел
Максимальное количество листов	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр определяет максимальное количество листов в дереве
Минимальное снижение неоднородности	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр определяет минимальное снижение неоднородности Узел будет разделен, если это разделение вызовет уменьшение неоднородности большее или равное указанному значению
Использовать бутстреп	Чекбокс	Данный чекбокс указывает на необходимость использования метода повторной выборки наблюдений
Размер бутстреп-выборок	Ручной ввод Неотрицательное число По умолчанию — 1	Данный параметр задает размер бутстреп-выборок
Минимальная доля веса наблюдений в листе	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр определяет минимальный весовой коэффициент выборки в листовом узле. По умолчанию наблюдения имеют одинаковый вес
Количество параллельных сессий	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр задает количество параллельных сессий
ccp_alpha	Ручной ввод По умолчанию — 0	Данный параметр регулирует количество отсекаемых узлов. Чем больше значение ccp_alpha, тем большее количество узлов удаляется из дерева

Таблица 30 Параметры узла «Случайный лес»

Результаты выполнения узла:

Узел «Случайный лес» имеет разные результаты в зависимости от решаемой задачи.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC (аналогично узлу «Дерево решений»).
- График Lift (аналогично узлу «Дерево решений»).
- График Cumulative Lift (аналогично узлу «Дерево решений»).
- График Gain (аналогично узлу «Дерево решений»).
- График Cumulative Gain (аналогично узлу «Дерево решений»).
- Таблица с метриками качества модели (аналогично узлу «Дерево решений»).
- Таблица с метриками качества модели для задачи классификации (аналогично узлу «Дерево решений»).
- Таблица со списком переменных, сортированных по важности (аналогично узлу «Дерево решений»).

Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели (аналогично узлу «Дерево решений»).
- Таблица с метриками качества модели для задачи классификации (аналогично узлу «Дерево решений»).
- Таблица со списком переменных, сортированных по важности (аналогично узлу «Дерево решений»).

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты обучающей, тестовой и валидационной выборок (аналогично узлу «Дерево решений»).
- Таблица с метриками качества модели (аналогично узлу «Дерево решений»).
- Таблица со списком переменных, сортированных по важности (аналогично узлу «Дерево решений»).

3.2.5.8.3. Узел «Байесовская регрессия»

Узел «Байесовская регрессия» представляет собой линейную регрессию с применением распределения вероятностей параметров, а не точечных оценок.

Алгоритм работы: В основе узла лежит Байесовская гребневая регрессия, где оптимизируются параметры регуляризации lambda (точность весов) и alpha (точность шума).

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе

Параметр	Возможные значения и ограничения	Описание
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Количество итераций	Ручной ввод Неотрицательное число, больше или равно 1 По умолчанию — 300	Данный параметр задает количество итераций, после достижения которого алгоритм останавливается
Добавить константу в модель	Чекбокс	Выбор данного чекбокса добавит константу в модель
Допустимая погрешность	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает допустимую погрешность, после достижения которой алгоритм останавливается
Стандартизация	Раскрывающийся список со следующими значениями: • no (по умолчанию) • std • range	Данный параметр отвечает за выбор метода стандартизации данных. Предусмотрены следующие методы: • no — нет • std — стандартное отклонение - преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • range — диапазон - линейно преобразует значения переменных в диапазон [0, 1].
Alpha 1	Ручной ввод Неотрицательное число По умолчанию — 0,000001	Данный параметр представляет собой параметр формы для предварительного распределения Гамма над параметром альфа
Alpha 2	Ручной ввод Неотрицательное число По умолчанию — 0,000001	Данный параметр представляет собой обратный параметр масштаба (параметр скорости) для предварительного распределения Гамма над параметром альфа
Alpha init	Ручной ввод Неотрицательное число По умолчанию — 0	Данный параметр представляет собой начальное значение для alpha (точность шума)
Lambda 1	Ручной ввод Неотрицательное число По умолчанию — 0,000001	Данный параметр представляет собой параметр формы для предварительного распределения Гамма над параметром лямбда
Lambda 2	Ручной ввод Неотрицательное число По умолчанию — 0,000001	Данный параметр представляет собой обратный параметр масштаба (параметр скорости) для предварительного распределения Гамма над параметром лямбда
Lambda init	Ручной ввод Неотрицательное число По умолчанию — 1	Данный параметр представляет собой начальное значение для лямбды (точность весов)

Таблица 31 Параметры узла «Байесовская регрессия»

Результаты выполнения узла:

- Тепловые карты для обучающей, валидационной и тестовой выборок.

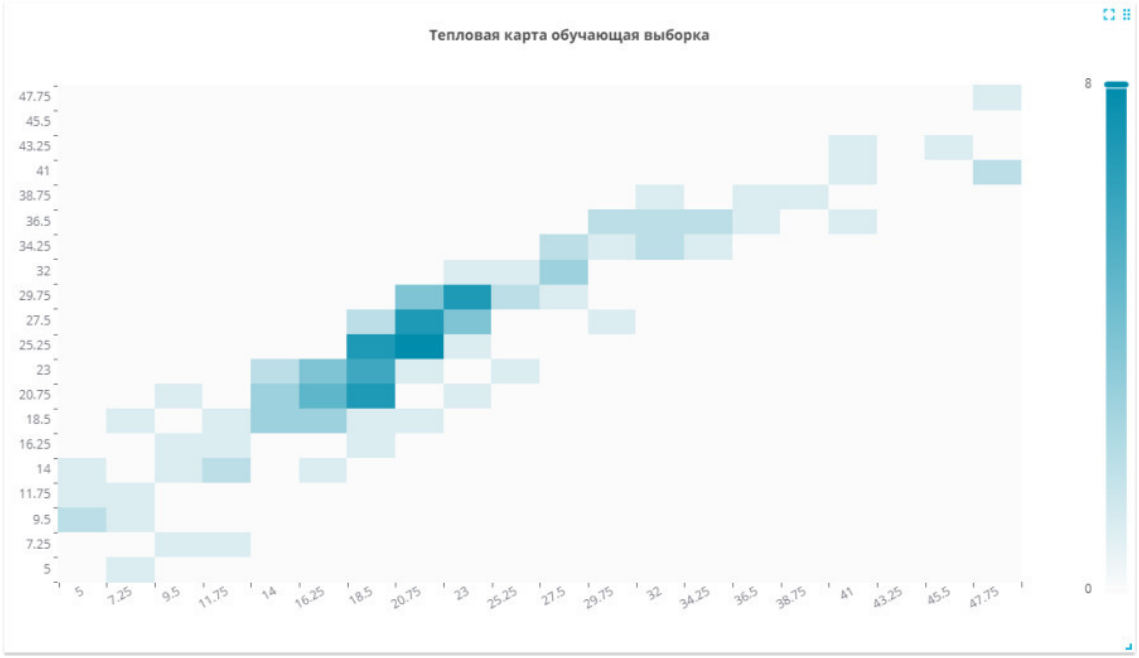
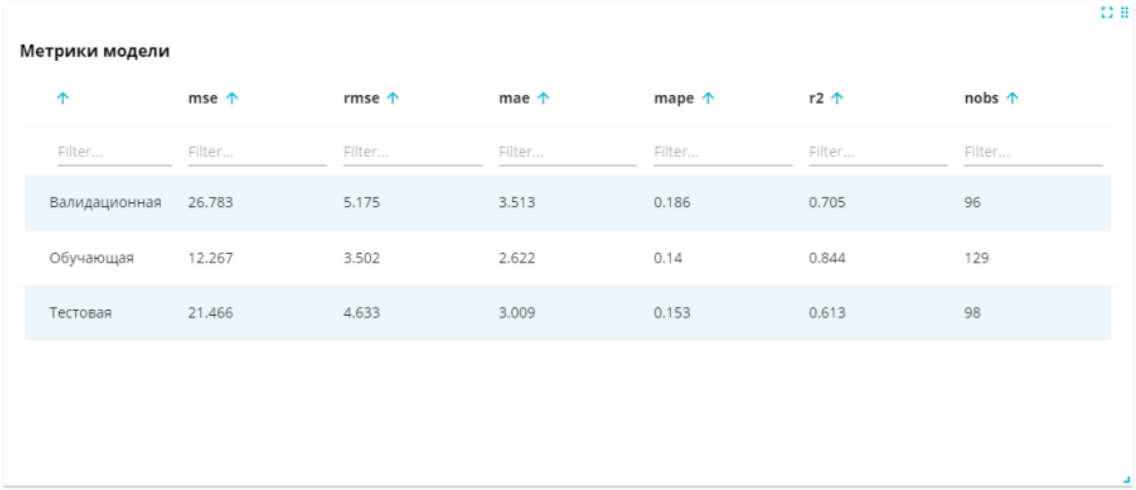


Рисунок 149 Пример тепловой карты на данных обучающей выборки

- Таблица с метриками качества модели.



Метрики модели

	mse	rmse	mae	mape	r2	nobs
Валидационная	26.783	5.175	3.513	0.186	0.705	96
Обучающая	12.267	3.502	2.622	0.14	0.844	129
Тестовая	21.466	4.633	3.009	0.153	0.613	98

Рисунок 150 Пример таблицы с метриками качества модели

- Таблица с коэффициентами переменных.

Рисунок 151 Пример таблицы с коэффициентами переменных

3.2.5.8.4. Узел «Линейная регрессия»

В основе **узла «Линейная регрессия»** лежит модель зависимости между входными и выходными переменными с линейной функцией связи.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Добавить константу в модель	Чекбокс	Выбор данного чекбокса добавит константу в модель
L1	Ручной ввод Неотрицательное число По умолчанию — 0,5	Данный параметр задает значение L1-регуляризации
L2	Ручной ввод Неотрицательное число По умолчанию — 0,5	Данный параметр задает значение L2-регуляризации
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел

Параметр	Возможные значения и ограничения	Описание
Количество итераций	Ручной ввод Неотрицательное число По умолчанию — 100	Данный параметр задает количество итераций, после достижения алгоритм останавливается
Метод оптимизации для гребневой регрессии	Раскрывающийся список со следующими значениями: • svd • cholesky • lsqr (по умолчанию) • sparse_cg • sag • saga	Данный параметр задает метод оптимизации для гребневой регрессии. Предусмотрены следующие варианты: • svd - использует сингулярное разложение • cholesky • lsqr - использует специальную процедуру упорядоченных наименьших квадратов • sparse_cg - использует решатель сопряженных градиентов • sag - использует stochastic average gradient descent • saga - использует улучшенную версию stochastic average gradient descent
Допустимая погрешность	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает допустимую погрешность, после достижения которой алгоритм останавливается
Правило обновления коэффициентов модели	Раскрывающийся список со следующими значениями: • cyclic (по умолчанию) • random	Данный параметр задает правило обновления коэффициентов модели. Предусмотрены следующие варианты: • cyclic - последовательный перебор • random - случайные коэффициенты обновляются каждую итерацию

Таблица 32 Параметры узла «Линейная регрессия»

Результаты выполнения узла:

- Тепловые карты для обучающей, валидационной и тестовой выборок (Аналогично узлу «Байесовская регрессия»).
- Таблица с метриками качества модели (Аналогично узлу «Байесовская регрессия»).
- Таблица с коэффициентами переменных (Аналогично узлу «Байесовская регрессия»).

3.2.5.8.5. Узел «Логистическая регрессия»

В основе **узла «Логистическая регрессия»** лежит метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Тип многоклассовой классификации	Раскрывающийся список со следующими значениями: • Бинарная классификация (по умолчанию) • Многоклассовая классификация	Данный параметр задает тип многоклассовой классификации. Предусмотрены следующие методы: • Бинарная классификация – метод «один против остальных» • Многоклассовая классификация – метод «многие против многих» Если решается бинарная логистическая регрессия, разницы между рассматриваемыми методами нет.
Добавить константу в модель	Чекбокс	Выбор данного чекбокса добавит константу в модель
L1	Ручной ввод Число больше или равное 0 По умолчанию — 0,5	Данный параметр задает значение L1-регуляризации
L2	Ручной ввод Число больше или равное 0 По умолчанию — 0,5	Данный параметр задает значение L2-регуляризации
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел
Количество итераций	Ручной ввод Неотрицательное число По умолчанию — 100	Данный параметр задает количество итераций, после достижения алгоритм останавливается
Метод оптимизации	Раскрывающийся список со следующими значениями: • lbfgs (по умолчанию) • newton-cg • liblinear • sag • saga	Данный параметр задает метод оптимизации. Предусмотрены следующие методы: • lbfgs – использует вторую производную матрицу функции потерь для итерационной оптимизации функции потерь • newton-cg – использует алгоритм Newton-Conjugate-Gradient • liblinear – использует метод спуска оси координат для итеративной оптимизации функции потерь • sag – стохастический средний градиентный спуск • saga – использует улучшенную версию стохастического среднего градиентного спуска Метод liblinear доступен только при выборе бинарной классификации.

Параметр	Возможные значения и ограничения	Описание
		Методы newton-cg, lbfgs, sag и saga обрабатывают L2 или без штрафа. Методы liblinear и saga также обрабатывают штраф L1. Метод saga также поддерживает штраф elasticnet (L1 + L2). Для небольших наборов данных метод liblinear - хороший выбор, тогда как sag и saga быстрее для больших. В случае неправильно выставленных пользователем методов оптимизации, автоматически будет произведена замена.
Допустимая погрешность	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает допустимую погрешность для оптимизации, после достижения которой алгоритм останавливается

Таблица 33 Параметры узла «Логистическая регрессия»

Результаты выполнения узла:

- Таблица с метриками качества модели.

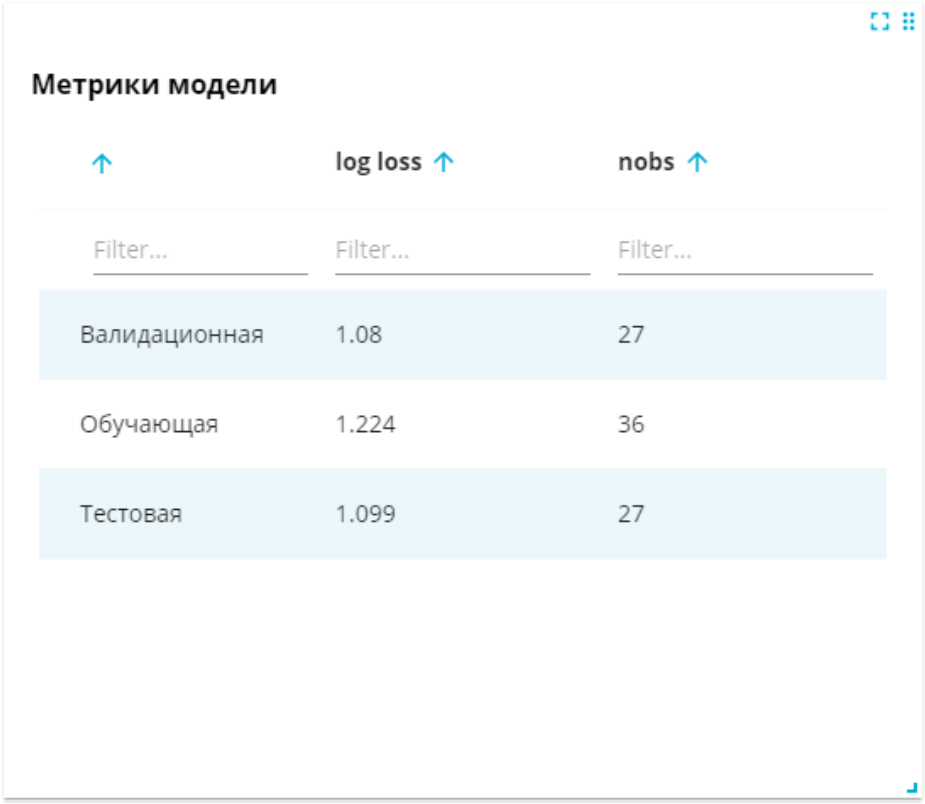
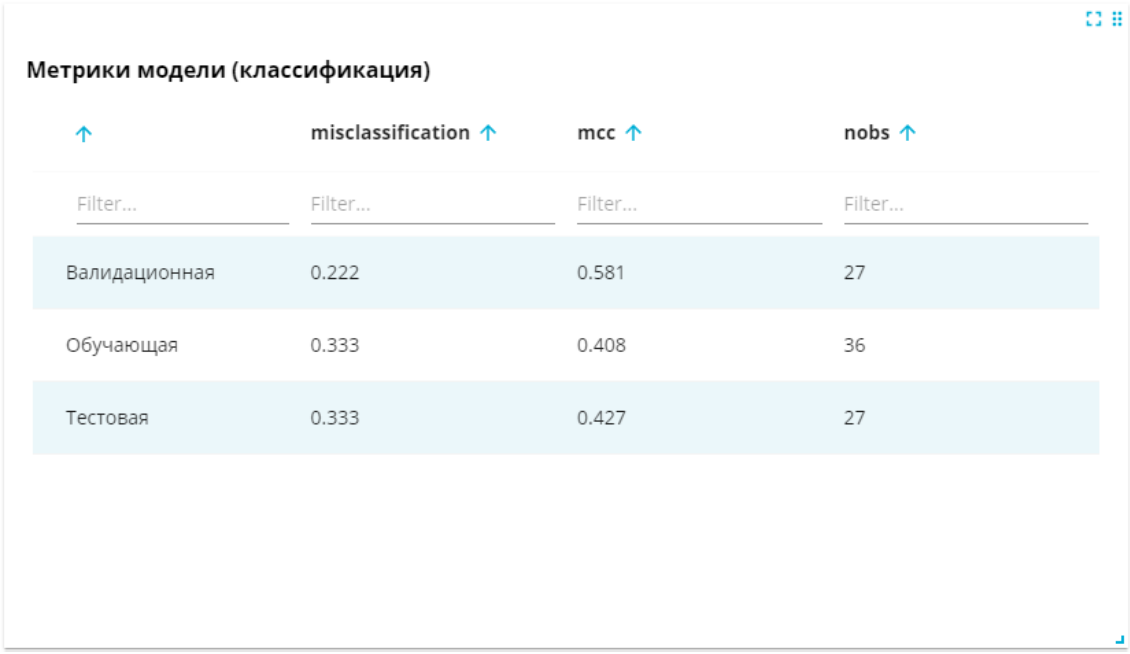


Рисунок 152 Пример таблицы с метриками качества модели

- Таблица с метриками качества модели задачи классификации.

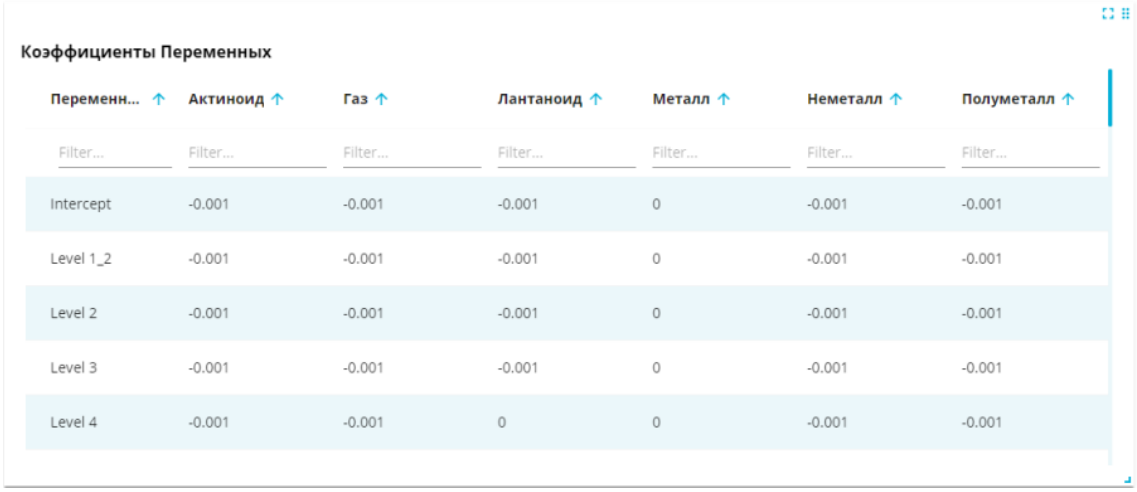


Метрики модели (классификация)

	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.222	0.581	27
Обучающая	0.333	0.408	36
Тестовая	0.333	0.427	27

Рисунок 153 Пример таблицы с метриками качества модели задачи классификации

- Таблица с коэффициентами переменных.



Коэффициенты Переменных

Переменн...	Актиноид	Газ	Лантаноид	Металл	Неметалл	Полуметалл
Filter...	Filter...	Filter...	Filter...	Filter...	Filter...	Filter...
Intercept	-0.001	-0.001	-0.001	0	-0.001	-0.001
Level 1_2	-0.001	-0.001	-0.001	0	-0.001	-0.001
Level 2	-0.001	-0.001	-0.001	0	-0.001	-0.001
Level 3	-0.001	-0.001	-0.001	0	-0.001	-0.001
Level 4	-0.001	-0.001	0	0	-0.001	-0.001

Рисунок 154 Пример таблицы с коэффициентами переменных

3.2.5.8.6. Узел «Линейные модели»

Узел «Линейные модели» объединяет в себе линейные классификаторы и регрессоры с обучением методом стохастического градиентного спуска и поддерживает различные функции потерь и штрафы.

Список параметров узла представлен в таблице ниже

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Функция потерь для классификации	Раскрывающийся список со следующими значениями: • hinge (по умолчанию) • log • modified_huber • squared_hinge • perceptron • squared_loss • huber • epsilon_insensitive • squared_epsilon_insensitive	Данный параметр задает функцию потерь для классификационной задачи. Предусмотрены следующие функции: • hinge – средняя потеря петель • log – логистическая регрессия • modified_huber – сглаженная потеря петли • squared_hinge – похож на hinge, но его штраф возведен в квадрат • perceptron – перцептрон • squared_loss – метод наименьших квадратов • huber – потеря Хубера • epsilon_insensitive • squared_epsilon_insensitive
Функция потерь для регрессии	Раскрывающийся список со следующими значениями: • squared_error (по умолчанию) • huber • epsilon_insensitive • squared_epsilon_insensitive	Данный параметр задает функцию потерь для регрессионной задачи. Предусмотрены следующие функции: • squared_error – метод наименьших квадратов • huber – потеря Хубера • epsilon_insensitive • squared_epsilon_insensitive
Epsilon	Ручной ввод Неотрицательное число По умолчанию — 0,1	Данный параметр задает Epsilon в функцию потерь
L1	Ручной ввод Число больше или равно 0 По умолчанию — 0,5	Данный параметр задает значение L1-регуляризации
L2	Ручной ввод Число больше или равно 0 По умолчанию — 0,5	Данный параметр задает значение L2-регуляризации
Правило определения скорости обучения	Раскрывающийся список со следующими значениями: • constant • optimal (по умолчанию) • invscaling • adaptive	Данный параметр задает правило определения скорости обучения. Предусмотрены следующие: • constant – постоянная скорость обучения • optimal – определяется на основе эвристики, предложенной Леоном Ботту • invscaling – обратное масштабирование • adaptive – адаптивное уменьшение скорости обучения

Параметр	Возможные значения и ограничения	Описание
Начальная скорость обучения	Ручной ввод Неотрицательное число По умолчанию — 0,001	Данный параметр задает начальную скорость обучения
Стандартизация	Раскрывающийся список со следующими значениями: • no (по умолчанию) • std • range	Данный параметр отвечает за выбор метода стандартизации данных. Предусмотрены следующие методы: • no — нет • std — стандартное отклонение - преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • range — диапазон - линейно преобразует значения переменных в диапазон [0, 1].
Метод построения вероятностей	Раскрывающийся список со следующими значениями: • linear (по умолчанию) • logistic	Данный параметр задает метод построения вероятностей
Показатель степени для изменения скорости обучения	Ручной ввод Неотрицательное число По умолчанию — 0,5	Данный параметр задает показатель степени для изменения скорости обучения
Добавить константу в модель	Чекбокс	Выбор данного чекбокса добавит константу в модель
Перемешать наблюдения	Чекбокс	Выбор данного чекбокса указывает на необходимость перемешать наблюдения
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел
Усреднение коэффициентов	Ручной ввод Число больше или равно 0 По умолчанию — 0	Данный параметр задает вычисление усредненных коэффициентов в результирующей линейной модели
Количество итераций	Ручной ввод Неотрицательное число По умолчанию — 1000	Данный параметр задает количество итераций, после достижения алгоритм останавливается
Ранняя остановка	Чекбокс	Выбор данного чекбокса указывает на необходимость ранней остановки алгоритма, если валидационная оценка не улучшается
Размер % валидационной выборки для ранней остановки	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,1	Данный параметр задает долю обучающих данных, которые нужно отложить в качестве валидационного набора для ранней остановки

Параметр	Возможные значения и ограничения	Описание
Допустимая погрешность	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает допустимую погрешность для оптимизации

Таблица 34 Параметры узла «Линейные модели»

Результаты выполнения узла:

Узел «Линейные модели» имеет разные результаты в зависимости от решаемой задачи.

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты для обучающей, валидационной и тестовой выборок (Аналогично узлу «Байесовская регрессия»).
- Таблица с метриками качества модели (Аналогично узлу «Байесовская регрессия»).
- Таблица с коэффициентами переменных (Аналогично узлу «Байесовская регрессия»).

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC.
- График Lift.
- График Cumulative Lift.
- График Gain.
- График Cumulative Gain.
- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица с коэффициентами переменных.

Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели (Аналогично узлу «Логистическая регрессия»).
- Таблица с метриками качества модели задачи классификации (Аналогично узлу «Логистическая регрессия»).
- Таблица с коэффициентами переменных (Аналогично узлу «Логистическая регрессия»).

3.2.5.8.7. Узел «Нейронная сеть»

Узел «Нейронная сеть» позволяет строить нейросеть типа **MLP (multilayer perceptron, многослойный перцептрон)**, в которой входной сигнал преобразуется в выходной, проходя последовательно через скрытые слои.

Искусственная нейронная сеть представляет собой сложную дифференцируемую функцию, которая в процессе обучения способна выявлять сложные зависимости между входными и выходными данными и выполнять обобщение.

3.2.5.8.7.1. Архитектура нейронной сети

Нейрон – базовый элемент для построения искусственной нейронной сети (далее просто нейронной сети). Представляет собой вычислительный элемент, который имеет несколько входов и один выход. Каждый вход имеет некоторый **вес**, на который умножается поступившее на данный вход значение. Далее в нейроне происходит суммирование взвешенных входных данных, и полученная сумма преобразуется с помощью **функции активации** в значение, которое и будет являться выходом нейрона.

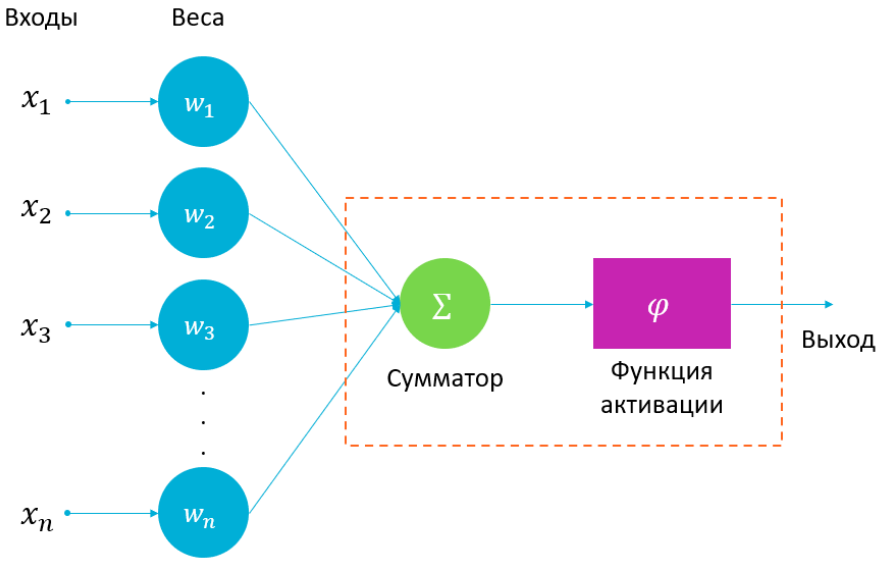


Рисунок 155 Строение искусственного нейрона

В многослойных сетях нейроны сгруппированы в **слои**, причем каждый нейрон предыдущего слоя связан со всеми нейронами следующего слоя, а внутри слоев связь между нейронами отсутствует.

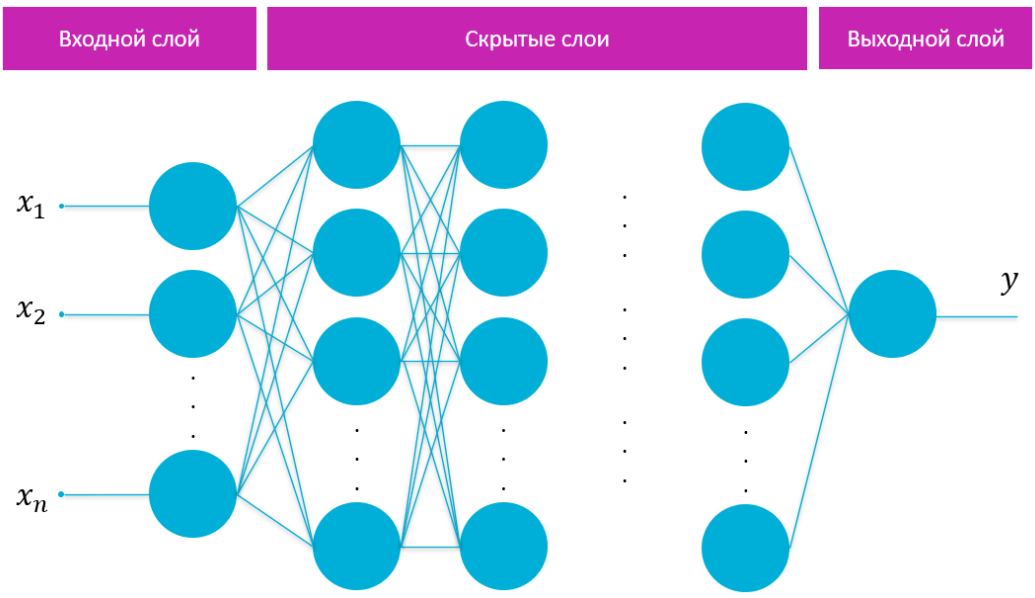


Рисунок 156 Слои нейронной сети

Первый слой сети называется **входным**: его нейроны принимают элементы вектора признаков и передают их нейронам следующего слоя без какой-либо обработки.

Далее располагаются один или несколько **скрытых слоев** (число скрытых слоев и количество нейронов в них задается **параметром Количество нейронов в скрытом слое**, подробнее в таблице с описанием параметров ниже). Каждый нейрон в скрытом слое преобразует значения из предыдущего слоя с взвешенным линейным суммированием функцией активации, которая порождает более информативные признаковые описания, преобразую данные нелинейным образом (функция активации для скрытых слоев задается **параметром Функция активации**).

Последний слой – **выходной**, получает значения из последнего скрытого слоя и преобразует их в выходные значения сети исходя из решаемой задачи.

3.2.5.8.7.2. Обучение нейронной сети

Обучение нейронной сети является итерационным процессом (задается параметром **Количество итераций**), которое заключается в нахождении параметров модели (весов и смещений). На каждой итерации происходит два прохода сети:

1. Настройка нейронной сети начинается с **Прямого распространения.** Данные (задается параметром **Количество наблюдений для одной итерации**) подаются на вход сети, через все слои распространяются в направлении выходов, в ходе чего случайным образом генерируются значения весовых коэффициентов и смещений модели. Таким образом, формируются выходные значения. Далее *фактические выходные значения* оцениваются относительно *ожидаемых выходных значений (целевых значений, target)* с помощью **функции потерь** (задается параметрами **Метрика для ранней остановки (регрессия)** или **Метрика для ранней остановки (классификация)** в зависимости от решаемой задачи).
2. **Обратное распространение ошибки (Backward propagation)** направлено на минимизацию функции потерь путем корректировки параметров нейронной сети (весов и смещений нейронов). Скорректировать эти параметры позволяют методы оптимизации на основе градиентного спуска (задается параметром **Метод оптимизации**).

Процесс оптимизации будет выполняться до тех пор, пока не будет достигнут критерий остановки (задается параметрами **Ранняя остановка**, **Погрешность оптимизации** и **Количество итераций**).

3.2.5.8.7.3. Список параметров узла

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Количество нейронов в скрытом слое	Для изменения количества нейронов и слоев необходимо: • Выбрать кнопку **Изменить** • Добавить необходимое количество слоев (Иконка + позволяет добавить скрытый слой) • Указать необходимое число нейронов в каждом слое • Выбрать кнопку **«Сохранить»** По умолчанию — 1 слой, в котором 100 нейронов	Данный параметр задает количество скрытых слоев и нейронов в скрытых слоях
Функция активации	Раскрывающийся список со следующими значениями: • Identity • logistic • tanh • relu (по умолчанию)	Данный параметр задает функцию активации для скрытых слоев. Предусмотрены следующие функции: • **Identity** – линейная функция, результат пропорционален переданному аргументу $f(x) = x$ • **logistic** – сигмоидная (логистическая) функция $\sigma(x) = \frac{1}{1+e^{-x}}$ • **tanh** – функция гиперболического тангенса • $f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}}$ • **relu** – функция активации ReLu (Rectified Linear Unit) $f(x) = \max(0, x)$
Метод оптимизации	Раскрывающийся список со следующими значениями: • lbfgs • sgd	Данный параметр задает метод оптимизации, который будет использоваться для обновления весов нейронов скрытых слоев нейронной сети. Предусмотрены следующие методы:

Параметр	Возможные значения и ограничения	Описание
	• adam (по умолчанию)	• **sgd** – стохастический градиентный спуск. • **lbfgs** – алгоритм оптимизации из семейства квазиньютоновских методов, который аппроксимирует алгоритм Бройдена–Флетчера–Гольдфарба–Шанно (BFGS) с использованием ограниченного объема памяти. • **adam** – метод адаптивной оценки моментов **Примечание** Метод **SGD** стоит использовать на небольших сбалансированных наборах данных, в которых достаточно равномерно представлены элементы каждого класса. В случае несбалансированности исходной выборки, стохастический градиент не производит качественного распознавания редких значений признаков, также низкая скорость сходимости проявляется на большом объеме данных. Метод **adam** довольно хорошо работает с большими наборами данных с точки зрения как времени обучения, так и валидационной оценки. Однако для небольших наборов данных метод **lbfgs** может сходиться быстрее и работать лучше.
L2 регуляризация	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает значение L2-регуляризации
Правило определения скорости обучения	Раскрывающийся список со следующими значениями: • constant (по умолчанию)	Данный параметр задает правило определения скорости обучения для обновления веса. Предусмотрены следующие варианты:

Параметр	Возможные значения и ограничения	Описание
	• invscalling • adaptive	• **constant** – постоянная скорость обучения, которая задается в **параметре Начальная скорость обучения** • **invscalling** – постепенно снижает скорость обучения на каждом шаге, используя показатель обратный указанному в **параметре Показатель степени для изменения скорости обучения** • **adaptive** – поддерживает скорость обучения постоянной на уровне значения, указанного в параметре **Начальная скорость обучения**, пока потери при обучении продолжают уменьшаться. Каждый раз, когда две последовательные эпохи не могут уменьшить потери при обучении по крайней мере на значение, указанное в параметре **Погрешность оптимизации**, или не могут увеличить оценку проверки по крайней мере на это же значение, если включена **Ранняя остановка**, текущая скорость обучения делится на 5. **Примечание** Данный параметр можно задать только в случае выбранного **метода оптимизации sgd**.
Начальная скорость обучения	Ручной ввод Число больше 0 По умолчанию — 0,001	Данный параметр задает начальную скорость обучения, которая управляет размером шага при обновлении весов. **Примечание** Данный параметр можно задать только в случае выбранного **метода оптимизации sgd или adam**.
Показатель степени для изменения скорости обучения	Ручной ввод Число больше 0 По умолчанию — 0,5	Данный параметр задает показатель степени для изменения скорости обучения, который используется для обновления скорости обучения, когда для параметра **Правило определения скорости обучения** указано значение **invscalling**.

Параметр	Возможные значения и ограничения	Описание
		Примечание Данный параметр можно задать только в случае выбранного **метода оптимизации sgd**.
Перемешивание наблюдений	Чекбокс	Выбор данного чекбокса указывает на необходимость перемешать наблюдения в каждой итерации. **Примечание** Данный параметр можно задать только в случае выбранного **метода оптимизации sgd или adam**.
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске узла
Максимальное количество итераций (эпох)	Ручной ввод Число больше 0 По умолчанию — 200	Данный параметр задает максимальное количество итераций. Оптимизатор выполняет итерации до сходимости (определяемой значением **параметра Погрешность оптимизации**) или до указанного количества итераций.
Количество наблюдений для одной итерации	Ручной ввод Число больше 0 По умолчанию — 200	Данный параметр задает количество наблюдений для одной итерации
Ранняя остановка	Чекбокс	Выбор данного чекбокса указывает на необходимость ранней остановки алгоритма, если валидационная оценка не улучшается на значение, указанное в параметре **Погрешность оптимизации**, в течение указанного в параметре **Количество итераций** без существенного улучшения значения. **Примечание** Ранняя остановка эффективна только в случае выбранного **метода оптимизации sgd или adam**.
Метрика для ранней остановки (регрессия)	Раскрывающийся список со следующими значениями: • MSE (по умолчанию)	Данный параметр задает метрику для ранней остановки (задача регрессии). Предусмотрены следующие метрики: • **MSE** – среднеквадратичная ошибка

Параметр	Возможные значения и ограничения	Описание
	• MAE • MAPE • R2	• **MAE** – средняя абсолютная ошибка • **MAPE** – средняя абсолютная ошибка в процентах • **R2** – коэффициент детерминации
Метрика для ранней остановки (классификация)	Раскрывающийся список со следующими значениями: • Logloss (по умолчанию) • Accuracy • MCC • AUC ROC	Данный параметр задает метрику для ранней остановки (задача классификации). Предусмотрены следующие метрики: • **Logloss** – логистическая функция потерь • **Accuracy** – доля правильных ответов алгоритма • **MCC** – коэффициент корреляции Мэтьюза • **AUC ROC** – площадь под ROC-кривой
Размер валидационной выборки для ранней остановки (%)	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,1	Данный параметр задает долю обучающих данных, которые нужно отложить в качестве валидационного набора для ранней остановки. Используется только в случае выбора **чекбокса Ранняя остановка**.
Количество итераций без существенного улучшения	Ручной ввод Число больше 0 По умолчанию — 10	Данный параметр задает количество итераций без существенного улучшения **Примечание** Данный параметр можно задать только в случае выбранного **метода оптимизации sgd или adam**.
Погрешность оптимизации	Ручной ввод Число больше 0 По умолчанию — 0,0001	Данный параметр задает допустимую погрешность для оптимизации. В случае, если потери не улучшаются на указанное значение в течение указанного в параметре Количество итераций без существенного улучшения числа, то считается, что сходимость достигнута, и обучение останавливается.
Инерция	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию — 0,9	Данный параметр задает инерцию для обновления градиентного спуска **Примечание** Данный параметр можно задать только в случае выбранного **метода оптимизации sgd**
Использовать инерцию Нестерова	Чекбокс	Выбор данного чекбокса указывает на необходимость использовать инерцию Нестерова

Параметр	Возможные значения и ограничения	Описание
		Примечание Данный параметр можно задать только в случае выбранного **метода оптимизации sgd** и значения **Инерции > 0**
Beta 1	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,9	Данный параметр задает скорость экспоненциального убывания для оценок вектора первого момента при **методе оптимизации adam**
Beta 2	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,999	Данный параметр задает скорость экспоненциального убывания для оценок вектора второго момента при **методе оптимизации adam**
Epsilon	Ручной ввод Число больше 0 По умолчанию — 1e-8	Данный параметр задает значение числовой устойчивости при **методе оптимизации adam**
Стандартизация	Раскрывающийся список со следующими значениями: • No (по умолчанию) • std • range	Данный параметр отвечает за выбор метода стандартизации числовых переменных. **Стандартизация** – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью. Предусмотрены следующие методы: • **no** — стандартизация не нужна • **std** — стандартное отклонение - преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • **range** — диапазон - линейно преобразует значения переменных в диапазон [0, 1].
Максимальное количество вызовов целевой функции	Ручной ввод целочисленного значения больше 0 По молчанию - 15000	Данный параметр задает максимальное количество вызовов функции потерь. Т.е. алгоритм выполняет итерации до сходимости (**параметр Погрешность оптимизации**), если количество итераций достигает значения параметра Максимальное количество итераций или заданного значения параметра вызова функции потерь.

Параметр	Возможные значения и ограничения	Описание
		Примечание Данный параметр можно задать только в случае выбранного **метода оптимизации lbfgs**

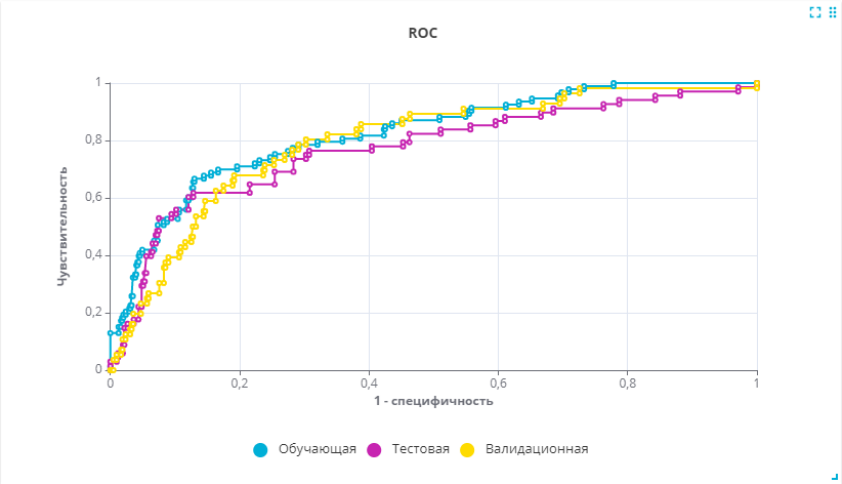
Таблица 35 Параметры узла «Нейронная сеть»

3.2.5.8.7.4. Результаты выполнения узла

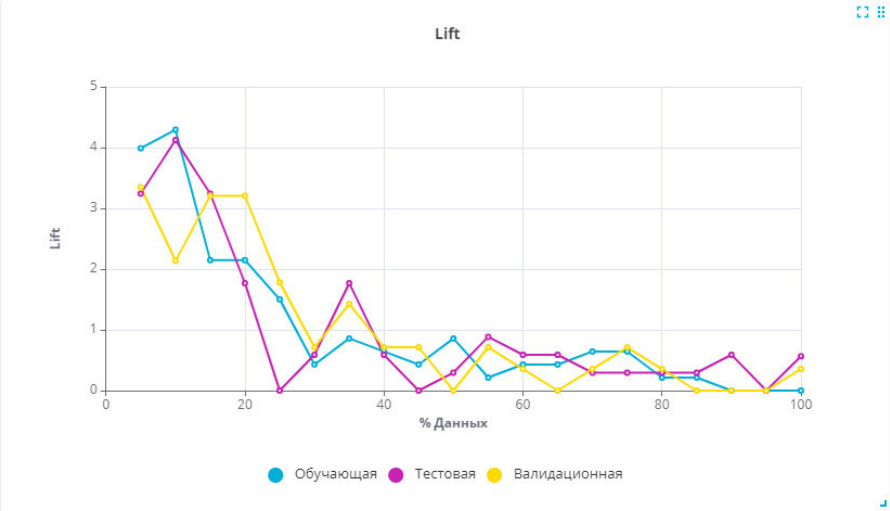
Узел «Нейронная сеть» имеет разные результаты в зависимости от решаемой задачи.

Результаты бинарной классификации представлены следующими объектами:

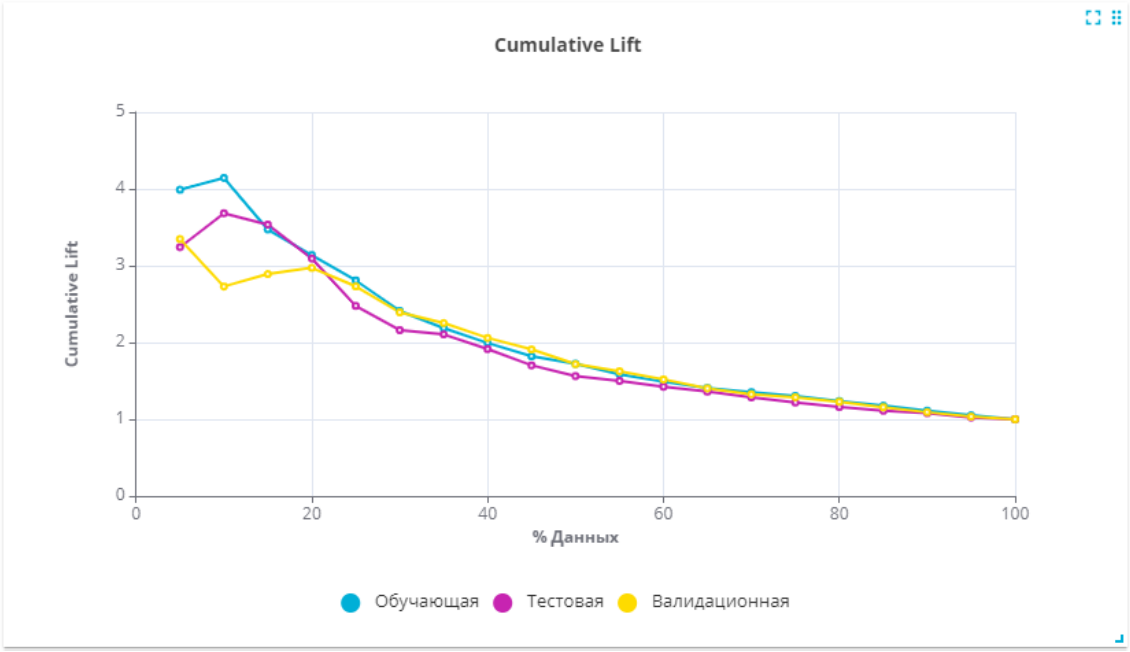
- График ROC.



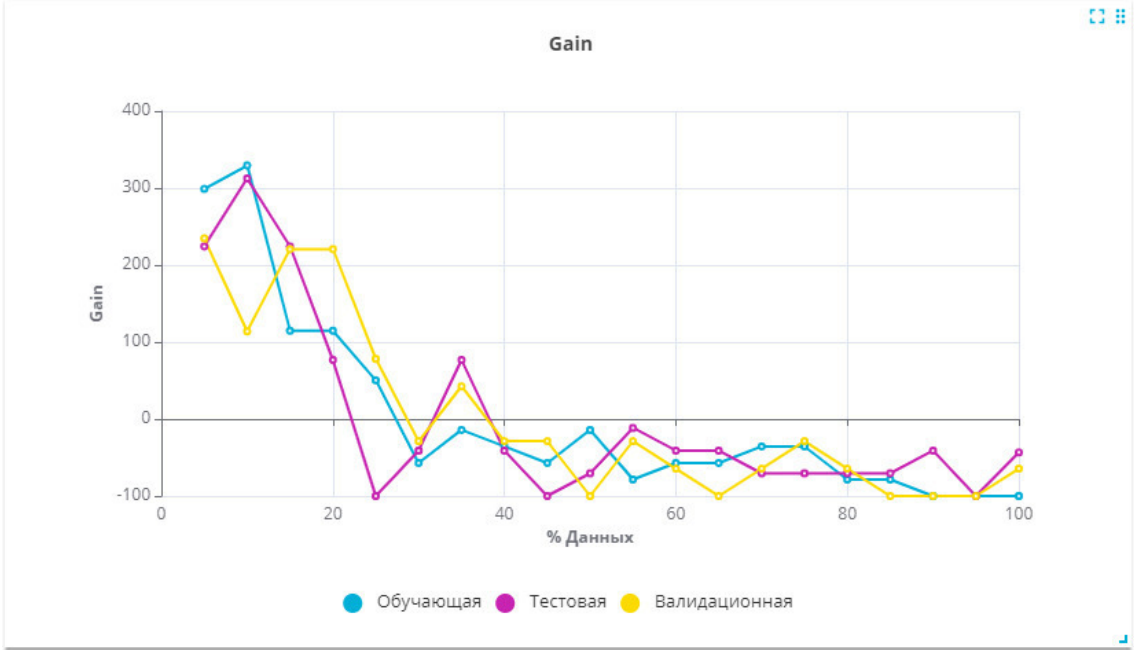
- График Lift.



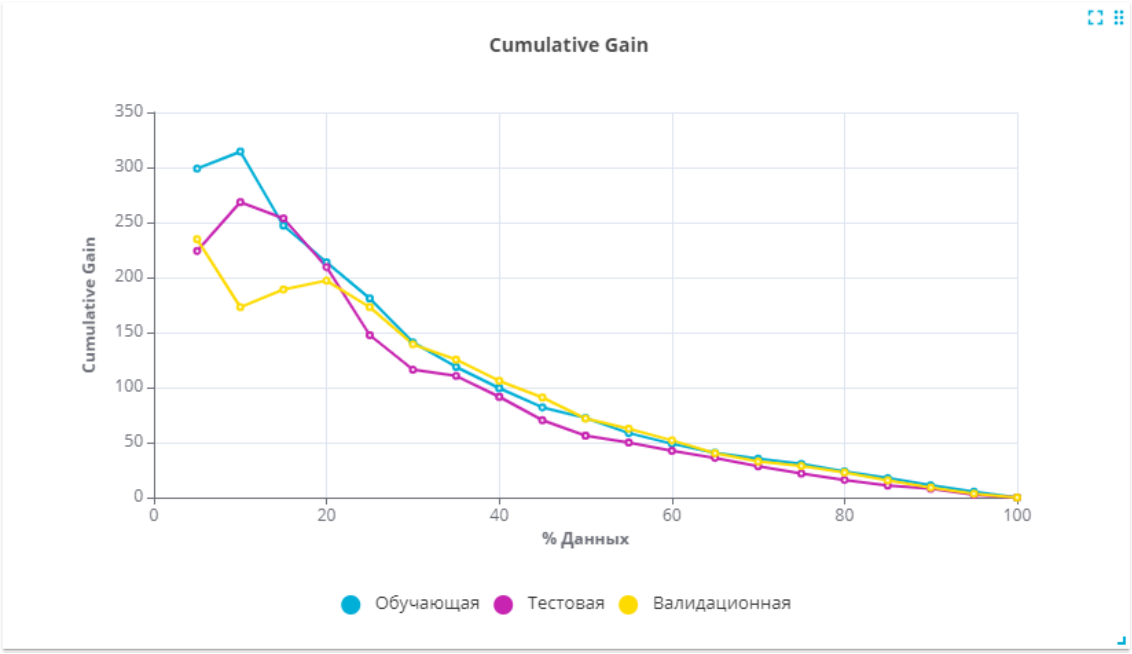
- График Cumulative Lift.



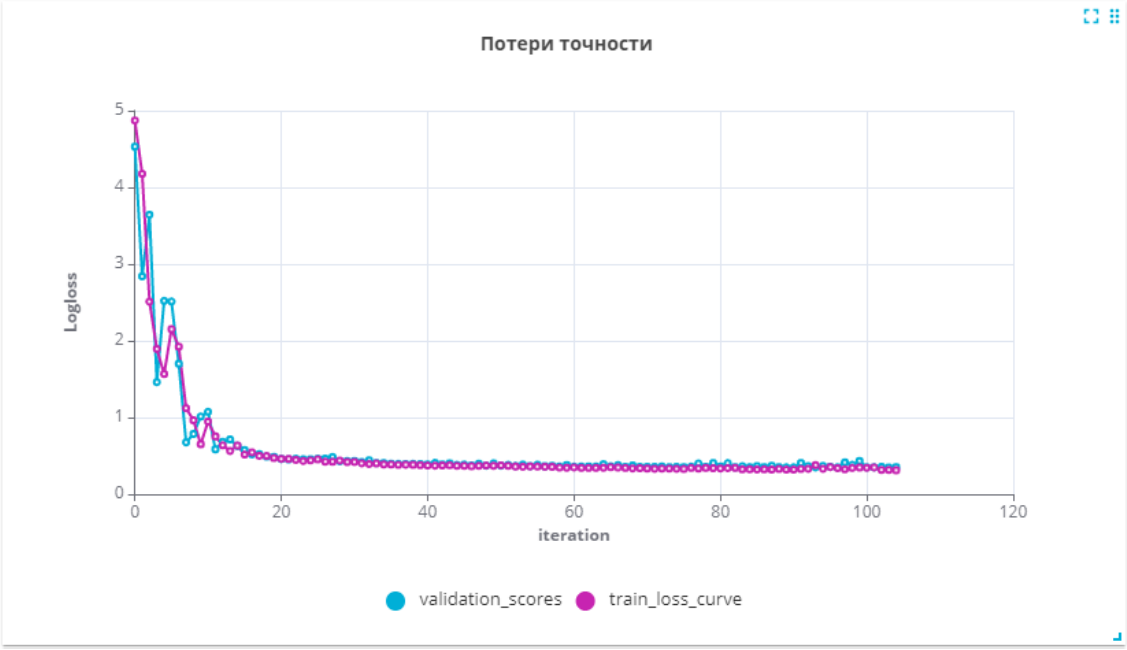
- График Gain.



- График Cumulative Gain.



- График потери точности.

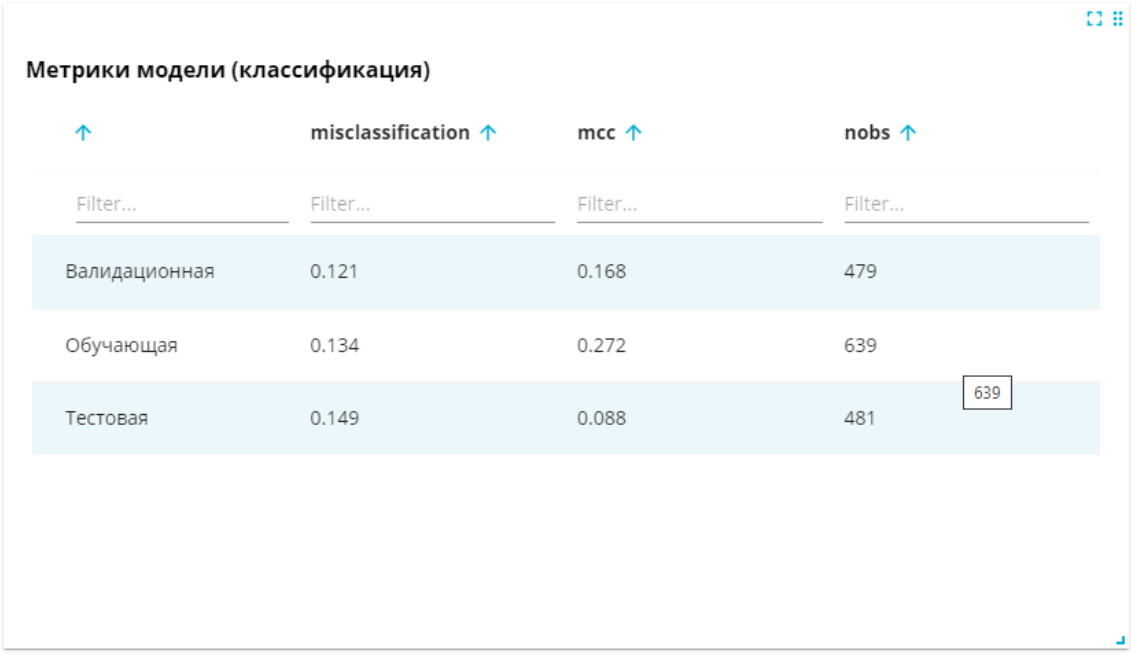


- Таблица с метриками качества модели.

Метрики модели

	AUC ROC	gini	log loss	nobs
Filter...	Filter...	Filter...	Filter...	Filter...
Валидационная	0.793	0.587	0.336	479
Обучающая	0.822	0.644	0.318	639
Тестовая	0.77	0.541	0.362	481

- Таблица с метриками качества модели задачи классификации.



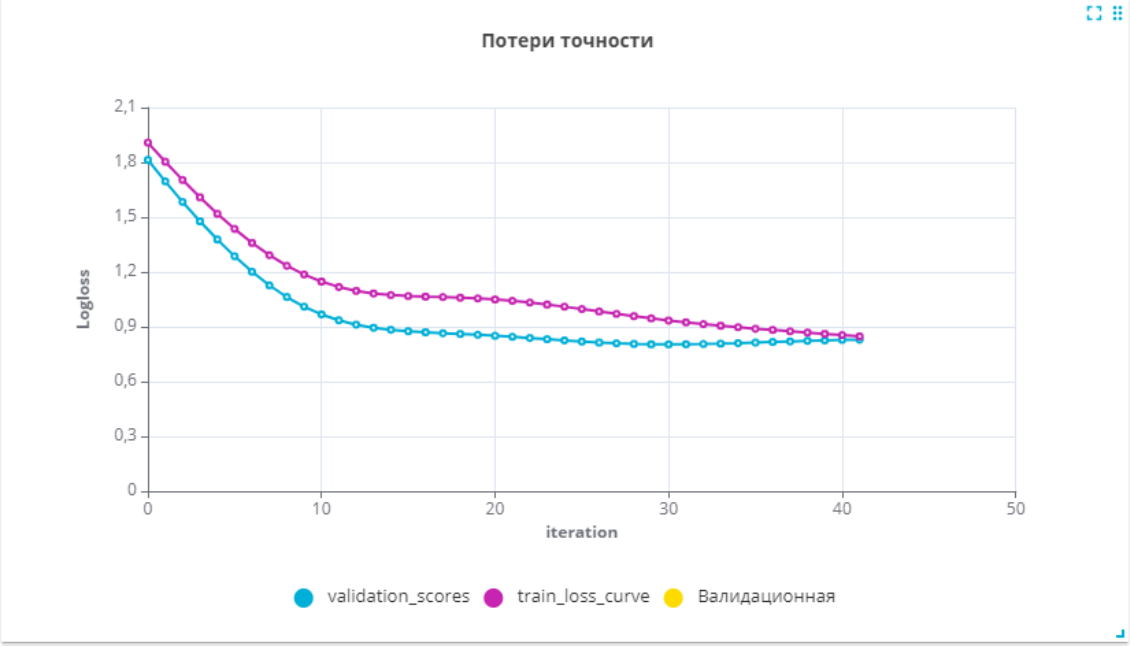
Метрики модели (классификация)

	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.121	0.168	479
Обучающая	0.134	0.272	639
Тестовая	0.149	0.088	481

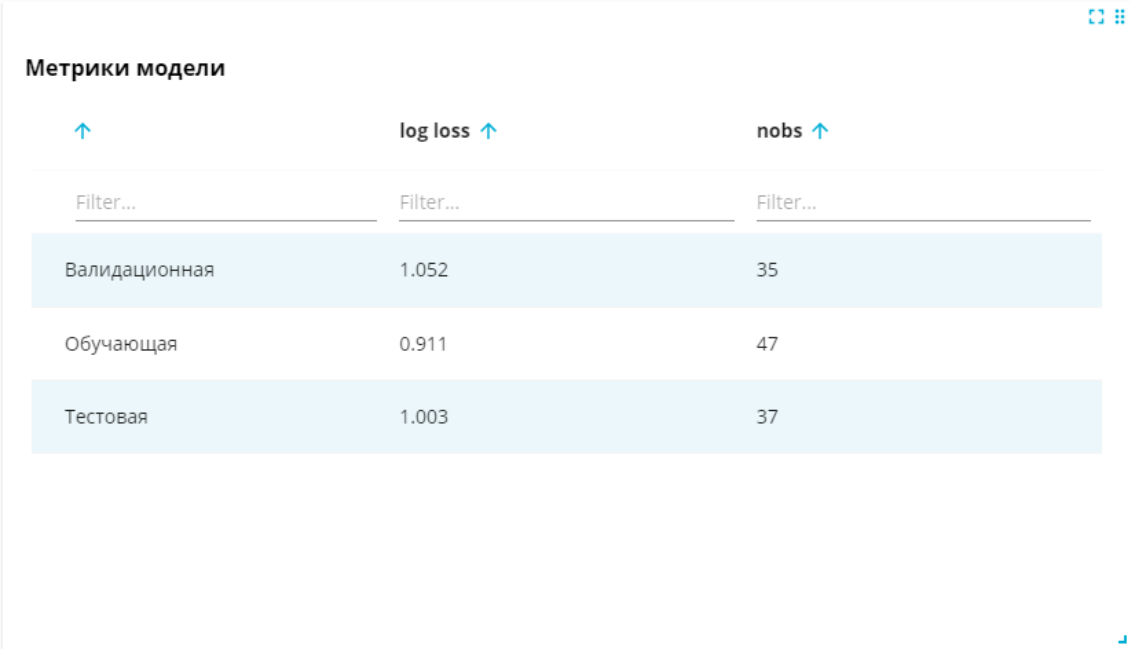
639

Результаты многоклассовой классификации представлены следующими объектами:

- График потери точности.



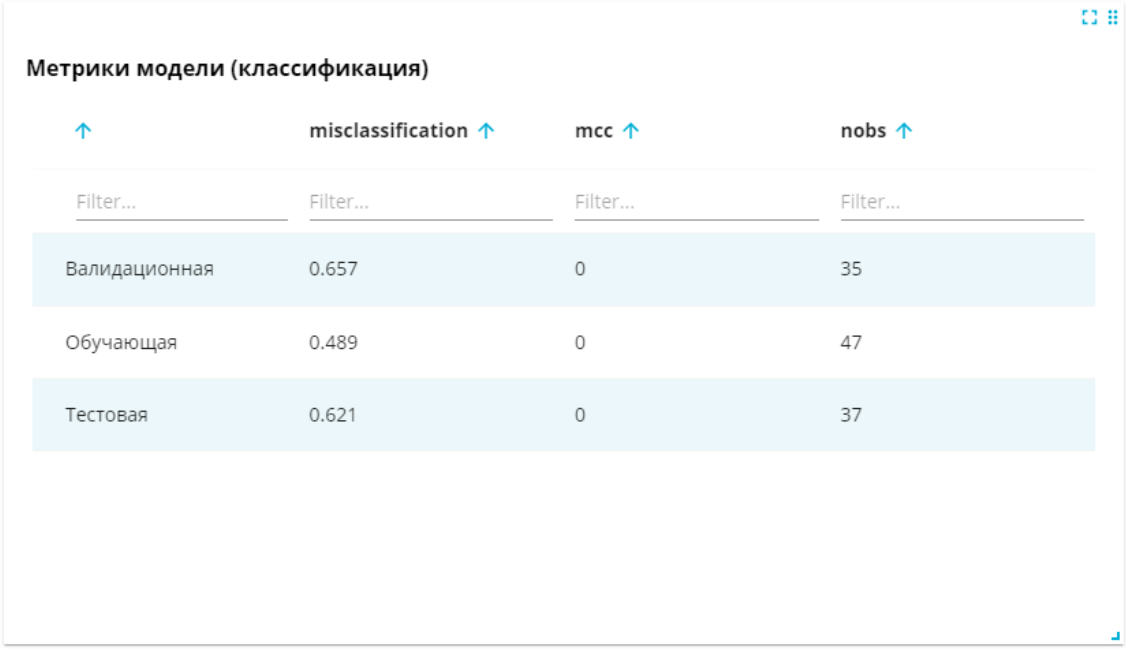
- Таблица с метриками качества модели.



Метрики модели

	log loss	nobs
Filter...	Filter...	Filter...
Валидационная	1.052	35
Обучающая	0.911	47
Тестовая	1.003	37

- Таблица с метриками качества модели задачи классификации.

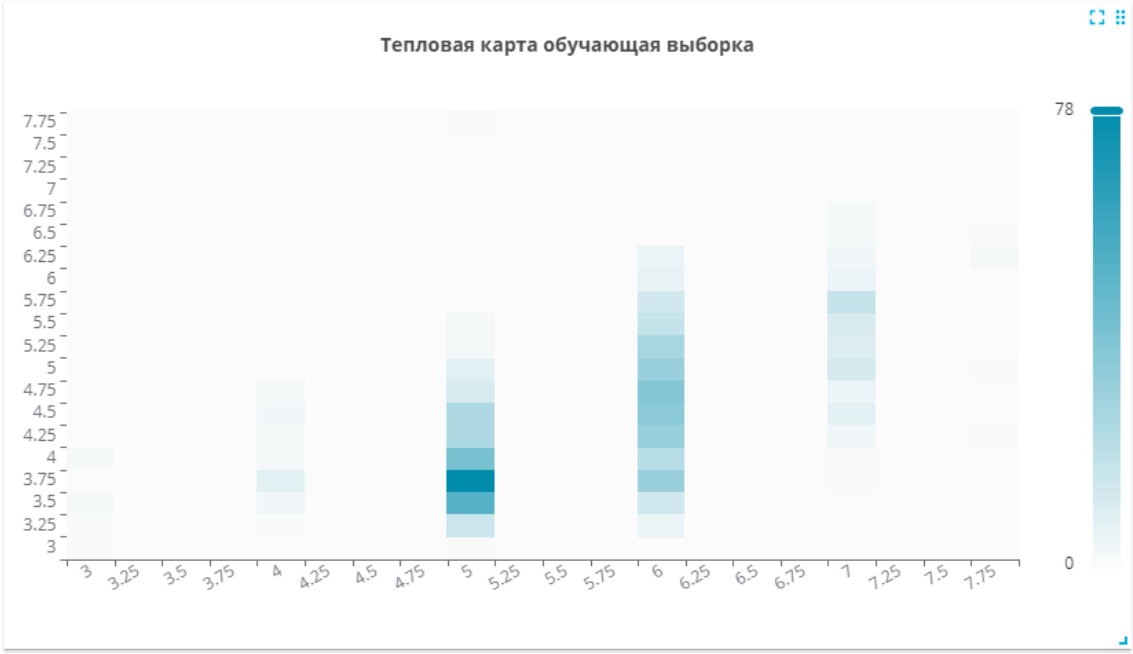


Метрики модели (классификация)

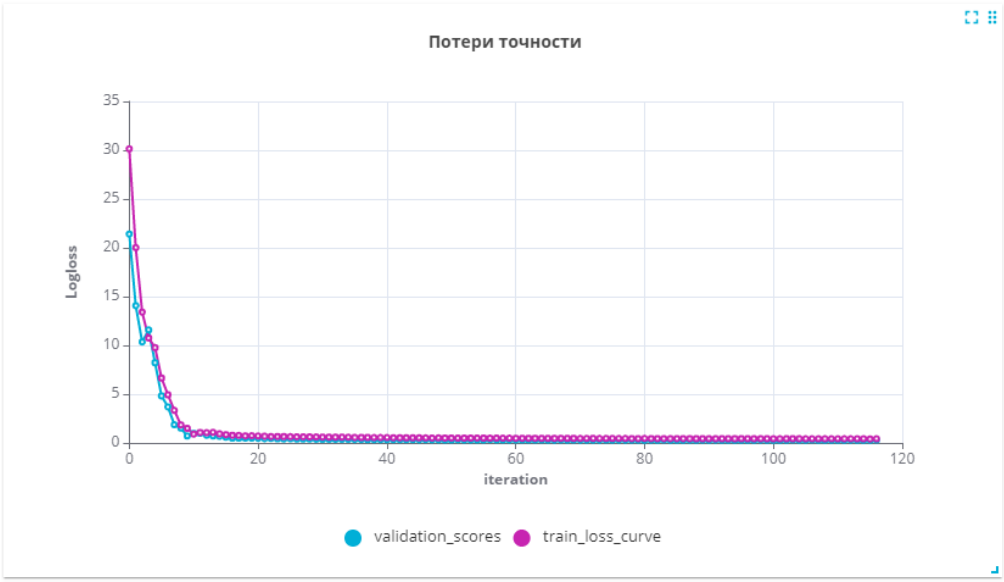
	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.657	0	35
Обучающая	0.489	0	47
Тестовая	0.621	0	37

Результаты регрессии представлены следующими объектами:

- Тепловые карты для обучающей, валидационной и тестовой выборок.



- График потери точности.



- Таблица с метриками качества модели.

Метрики модели

	mse	rmse	mae	mape	r2	nobs
CV (валидационна	0.562	0.745	0.566	0.104	0.084	127.8
CV (обучающая)	0.509	0.705	0.533	0.098	0.218	511.2
Валидационная	0.527	0.726	0.567	0.103	0.178	479
Обучающая	0.412	0.642	0.487	0.09	0.361	639
Тестовая	0.471	0.686	0.529	0.096	0.293	481

- Таблицы с метриками качества в разбиениях обучающей и валидационной выборках.

Метрики в разбиениях обучающей выборки

Номер разбие...	fit_time	mse	rmse	mae	mape	r2	nobs
0	0.381	0.465	0.682	0.511	0.096	0.315	511
1	0.397	0.409	0.64	0.489	0.092	0.34	511
2	0.348	0.413	0.642	0.481	0.087	0.332	511
3	0.096	0.837	0.915	0.71	0.129	-0.224	511
4	0.342	0.42	0.648	0.474	0.086	0.328	512

3.2.5.8.8. Узел «LDA»

Узел «LDA» представляет линейный дискриминантный анализ, который применяется для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Метод оптимизации	Раскрывающийся список со следующими значениями: • svd (по умолчанию) • lsqr • eigen	Данный параметр задает метод, используемый для поиска матрицы объектов гиперплоскости LDA. Предусмотрены следующие методы: • svd – разложение по сингулярным значениям • lsqr – метод наименьших квадратов • eigen – разложение по собственным значениям
Задать величину сжатия автоматически	Чекбокс	Выбор данного чекбокса указывает на необходимость задать величину сжатия автоматически
Сжатие	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0	Данный параметр задает значение сжатия
Допустимая погрешность	Ручной ввод Неотрицательное число По умолчанию — 0,0001	Данный параметр задает допустимую погрешность для оптимизации Данный параметр можно задать только в случае выбранного метода оптимизации svd

Таблица 36 Параметры узла «LDA»

Результаты выполнения узла:
Узел «LDA» имеет разные результаты в зависимости от решаемой задачи.
Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица с коэффициентами переменных.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC.
- График Lift.

- График Cumulative Lift.
- График Gain.
- График Cumulative Gain.
- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица с коэффициентами переменных.

3.2.5.8.9. Узел «Градиентный бустинг (XGBOOST)»

В основе **узла "XGBoost"** лежит алгоритм градиентного бустинга на деревьях поиска решений или линейных моделях. Используется для решения задач классификации и регрессии.

Алгоритм работы градиентного бустинга:

Градиентный бустинг – алгоритм машинного обучения, который строит модель предсказания в виде ансамбля слабых предсказывающих моделей (в основном Дерево решений, но также есть вариант с линейными моделями). На каждой итерации вычисляется отклонение предсказаний уже обученного ансамбля (всех предыдущих построенных моделей) на обучающей выборке. Следующая добавляемая в ансамбль модель будет сводить среднее отклонение предыдущей к минимуму.
Новые деревья добавляются в ансамбль до тех пор, пока уменьшается ошибка, либо пока не выполнится одно из правил «ранней остановки».

Особенности реализации XGBoost (Extreme Gradient Boosting) в сравнении со стандартным алгоритмом градиентного бустинга:

XGBoost поддерживает как алгоритм предварительной сортировки, так и алгоритм на основе гистограммы.

1. В данной реализации каждый лист дерева не проверяется на предмет того, что можно ли его разделить дальше. Здесь заранее задается максимальная глубина дерева, глубже которой узлы и листья не строятся. Далее после построения всех листьев до заданной глубины проверяется дает ли сплит (деление) в конкретном узле прирост информации (Information Gain). Для этого Information Gain сравнивается с заданным порогом, который настраивается экспериментальным путем. Если значение меньше порога, то листья этого узла отсекаются. Таким образом избегается переобучение
2. Используется регуляризация для избежания малого количества наблюдений в листе, что также позволяет бороться с переобучением.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе	Общий параметр
Описание	Ручной ввод Ограничений на значение нет	Описание узла	Общий параметр
Бустер	Раскрывающийся список со следующими значениями: • gbtree (по умолчанию) • gblinear • DART	Данный параметр задает тип базового алгоритма для бустинга. Предусмотрены следующие типы: • gbtree – бустинг на основе деревьев • gblinear – бустинг на основе линейных моделей • DART – модификация gbtree (отбрасывает деревья, для предотвращения переобучения)	Общий параметр
Количество оценочных функций	Ручной ввод Число больше 0 По умолчанию — 100	Данный параметр задает число итераций градиентного бустинга	Общий параметр
Скорость обучения	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,3	Данный параметр задает скорость обучения модели и контролирует, с каким весом предсказания каждой следующей модели суммируются с предсказаниями ансамбля. Значение по умолчанию (0,3) является слишком большим, обычно хорошо работают значения меньше 0.1	Общий параметр
Цель обучения для регрессии	Раскрывающийся список со следующими значениями: • Квадрат ошибки (по умолчанию) • Функция потерь Хьюбера • Пуассоновская регрессия • Регрессия Твиди	Данный параметр задает используемую при обучении функцию потерь. Предусмотрены следующие: • Квадрат ошибки – квадратичная функция потерь • Функция потерь Хьюбера – функция квадратична для малых значений остатка (разница между	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
	• Гамма регрессия	наблюдаемым и предсказанным значением), и линейна для больших значений остатка • Пуассоновская регрессия – предназначена для прогнозирования счетчиков (неотрицательных целых чисел) (например, количество дождевых явлений в год или количество событий прерывания производства в год). Поэтому использовать данную функцию следует в случае соответствия следующим условиям: • переменная ответа имеет распределение Пуассона, • метки не должны быть отрицательными • бессмысленно использовать для дробных чисел • Регрессия Твиди – предназначена для прогнозирования целевой переменной, имеющей распределение Твиди (например, общее количество осадков в год или общее время прерывания в год) • Гамма регрессия – предназначена для прогнозирования целевой переменной,	

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		имеющей гамма-распределение (например, количество осадков на одно событие или продолжительность прерывания)	
Цель обучения для классификации	Раскрывающийся список со следующими значениями: • Бинарная логистическая регрессия (по умолчанию) • Бинарная с hinge loss • Мультиклассовая с softprob • Мультиклассовая с softmax	Данный параметр задает используемую при обучении функцию потерь. Предусмотрены следующие: • Бинарная логистическая регрессия – возвращает прогнозируемую вероятность (не класс) • Бинарная с hinge loss – кусочно-линейная функция потерь для бинарной классификации. Возвращает принадлежность одному из двух классов – 0 или 1 • Мультиклассовая с softprob – функция softprob для мультиклассовой классификации, возвращает матрицу со значениями вероятности каждого класса • Мультиклассовая с softmax – функция softmax для мультиклассовой классификации, возвращает класс с максимальной вероятностью принадлежности	Общий параметр
Дисперсия распределения Твиди	Ручной ввод Число больше 1 и меньше 2 По умолчанию — 1,5	Данный параметр используется для управления дисперсией распределения Твиди. Значение ближе к 2 переходит в гамма-	Параметр актуален при заданных Цели обучения для регрессии = Регрессия Твиди и Метриках для

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		распределение, значение ближе к 1 – распределение Пуассона.	валидации регрессии, связанными с распределением Твиди.
Размер (%) валидационной выборки для ранней остановки	Ручной ввод Число больше 0 и меньше 1 По умолчанию – 0,1	Данный параметр задает размер (%) валидационной выборки для ранней остановки	Общий параметр
Количество итераций до ранней остановки	Ручной ввод По умолчанию – 0	Данный параметр задает количество итераций до ранней остановки	Общий параметр
Метрика для валидации регрессии	Раскрывающийся список со следующими значениями: • RMSE (по умолчанию) • MAE • MAPE • poisson nloglik • gamma nloglik • gamma deviance • tweedie nloglik	Данный параметр задает метрику качества на валидационной выборке. Предусмотрены следующие метрики: • RMSE – Среднеквадратическая ошибка • MAE – Средняя абсолютная ошибка • MAPE – Средняя абсолютная ошибка в процентах • poisson nloglik – отрицательная логарифмическая функция правдоподобия для регрессии Пуассона • gamma nloglik – отрицательная логарифмическая функция правдоподобия для гамма-регрессии • gamma deviance – остаточное отклонение для гамма-регрессии • tweedie nloglik – отрицательная логарифмическая функция правдоподобия для регрессии Твиди	Общие параметры
Метрика для валидации	Раскрывающийся список со	Данный параметр задает метрику качества на валидационной выборке.	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
классификации	следующими значениями: • logloss (по умолчанию) • error • merror • mlogloss • auc • aucpr	Предусмотрены следующие метрики: • logloss – логистическая функция ошибки • error – частота ошибок бинарной классификации, рассчитывается как неправильно классифицированные объекты/все объекты. При прогнозировании положительными экземплярами будут считаться наблюдения со значением прогноза больше 0,5, остальные – как отрицательные • merror – частота ошибок мультиклассовой классификации, рассчитывается как неправильно классифицированные объекты/все объекты • mlogloss – мультиклассовая логистическая функция ошибки • auc – количественная интерпретация кривой ошибок, площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций • aucpr – площадь под PR-кривой (Precision-Recall curve)	
Cutoff для метрики error	Ручной ввод По умолчанию – 0,5	Данный параметр задает порог отсечения, чтобы относить новые примеры к одному из двух классов (задача бинарной классификации).	Общие параметры Актуален при выбранной метрике error для валидации классификации.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Способ определения важности переменных	Раскрывающийся список со следующими значениями: • gain (по умолчанию) • weight • cover • total gain • total cover	Данный параметр определяет метод оценки важности переменных. Предусмотрены следующие методы: • gain – средний выигрыш по всем разделениям (splits), в которых используется переменная • weight – количество раз, когда переменная используется для разделения данных по всем деревьям • cover – среднее количество наблюдений для каждой фичи по всем разделениям, в которых используется переменная • total gain – общий выигрыш по всем разделениям, в которых используется переменная • total cover – общее количество наблюдений для каждой фичи всех разделений, в которых используется переменная	Параметры для Tree Booster (Бустер = gbtree). Для Бустер = gblinear значение параметра = 'weight'
Количество параллельных потоков	Ручной ввод Число больше или равно 0 По умолчанию – 0	Данный параметр задает количество параллельных потоков, используемых для запуска xgboost 0 означает использование всех доступных потоков (CPU)	Общие параметры
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел	Общие параметры
Метод построения дерева	Раскрывающийся список со следующими значениями: • auto (по умолчанию)	Данный параметр задает метод построения дерева. Предусмотрены следующие методы: • auto – для небольшого набора	Параметры для Tree Booster и DART (Бустер = gbtree или dart)

Параметр	Возможные значения и ограничения	Описание	Группа параметров
	• exact • approx • hist	данных будет использован exact, для большого набора данных – approx • exact – жадный алгоритм, который перебирает все наблюдения входного набора в ходе процедуры поиска разделения. Данный метод более точен среди других жадных методов, но медленнее в вычислительной производительности • approx – приближенный жадный алгоритм, который использует quantile sketch (квантильные наброски) и gradient histogram (приближенные гистограммы статистики градиента) • hist – более быстрый приближенный жадный алгоритм, оптимизированный для histogram.	
Метод добавления узлов к дереву	Раскрывающийся список со следующими значениями: • depthwise (по умолчанию) • lossguide	Данный параметр задает способ добавления новых узлов к дереву. Предусмотрены следующие методы: • depthwise – разделение в узлах, ближайших к корню • lossguide – разделение на узлы с наибольшим изменением значения функции потерь	Параметры для Tree Booster и DART (Бустер = gbtree или dart) Поддерживается только если метод построения дерева выбран как hist

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Количество параллельных деревьев	Ручной ввод Число больше или равно 0 По умолчанию – 1	Данный параметр задает количество параллельных деревьев, построенных на каждой итерации	Параметры для Tree Booster и DART (Бустер = gbtree или dart)
Максимальная глубина дерева	Ручной ввод Число больше 0 По умолчанию – 6	Данный параметр задает максимальную глубину дерева	Параметры для Tree Booster и DART (Бустер = gbtree или dart)
Минимальное снижение потери для разбиения	Ручной ввод Число больше или равно 0, float По умолчанию – 0	Данный параметр задает значение минимального уменьшения функции потерь для разбиения	Параметры для Tree Booster и DART (Бустер = gbtree или dart)
Максимальное количество листов	Ручной ввод Число больше или равно 0 По умолчанию – 0	Данный параметр задает максимальное количество листов.	Параметры для Tree Booster и DART (Бустер = gbtree или dart) Актуален, если выбран Метод добавления узлов к дереву = lossguide
Максимальное количество бинов для интервальных переменных	Ручной ввод Число больше 0 По умолчанию – 256	Данный параметр задает максимальное количество бинов для интервальных переменных. Увеличение этого числа повышает оптимальное разделение за счет увеличения времени вычислений.	Параметры для Tree Booster и DART (Бустер = gbtree или dart) Актуален, если выбран метод построения дерева = hist
Относительное количество бинов (sketch_eps)	Ручной ввод Число больше 0 и меньше 1 По умолчанию – 0,03	Данный параметр задает относительное количество бинов, примерно соответствующее 1 / sketch_eps. По сравнению с прямым выбором количества бинов дает теоретическую гарантию точности набросков (sketch).	Параметры для Tree Booster и DART (Бустер = gbtree или dart) Актуален, если выбран метод построения дерева = approx
Соотношение колонок для каждого дерева	Ручной ввод Число больше 0 и меньше или равно 1 По умолчанию – 1	Данный параметр задает долю переменных, используемых на каждой итерации (при построении каждого дерева). Подвыборка происходит один раз для каждого построенного дерева	Параметры для Tree Booster и DART (Бустер = gbtree или dart) Данные параметры работают кумулятивно. Например, комбинация {'colsample_bytree':0,5, 'colsample_bylevel':0,5, 'colsample_bynode':0,5} с 64 функциями оставит 8 функций

Параметр	Возможные значения и ограничения	Описание	Группа параметров
			на выбор при каждом разбиении.
Соотношение колонок для каждого уровня	Ручной ввод Число больше 0 и меньше или равно 1 По умолчанию – 1	Данный параметр задает долю подвыборки признаков, которые будут использованы для обучения каждого уровня. Подвыборка происходит один раз для каждого нового уровня глубины, достигнутого в дереве. Колонки выбираются из набора колонок, выбранных для текущего дерева.	
Соотношение колонок для каждого разбиения	Ручной ввод Число больше 0 и меньше или равно 1 По умолчанию – 1	Данный параметр задает долю подвыборки признаков, которые будут использованы для каждого разделения (узла). Подвыборка происходит каждый раз, когда оценивается новое разбиение. Колонки (столбцы) выбираются из набора колонок, выбранного для текущего уровня	
Минимальный вес для потомка	Ручной ввод Число больше или равно 0 По умолчанию – 0	Данный параметр задает минимальный вес потомка, необходимый для разделения. Если шаг разделения дерева приводит к листу с суммой весов меньше, чем заданное данным параметром значение, то процесс построения откажется от дальнейшего разделения. В задаче линейной регрессии это соответствует минимальному количеству наблюдений, которые должны быть в каждом узле. Чем больше значение, тем более консервативен алгоритм.	Параметры для Tree Booster и DART (Бустер = gbtree или dart)
Максимальный шаг в листе	Ручной ввод Число больше или равно 0 По умолчанию – 0	Данный параметр задает максимальный шаг обновления для каждого выходного листа.	Параметры для Tree Booster и DART (Бустер = gbtree или dart)

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		Если значение равно 0, это означает, что ограничения отсутствуют. Данный параметр может помочь при логистической регрессии, когда классы несбалансированы.	
Соотношение случайной подвыборки в обучающей выборке	Ручной ввод Число больше 0 или меньше или равно 1 По умолчанию – 1	Данный параметр задает долю объектов обучающей выборки, используемых на каждой итерации. При значении равном 0,5 XGBoost будет случайным образом отбирать половину обучающих данных перед обучением (growing) дерева, что предотвращает переобучение. Подвыборка будет происходить один раз в каждой итерации бустинга	Бэггинг
Метод сэмплинга	Раскрывающийся список со следующими значениями: • uniform (по умолчанию) • gradient based	Данный параметр задает метод сэмплинга. Предусмотрены следующие методы: • uniform – каждое наблюдение имеет равную вероятность быть выбранным. Для хороших результатов необходимо выбрать соотношение случайной подвыборки в обучающей выборке = 0,5 • gradient based – вероятность выбора наблюдения пропорциональна регуляризованному абсолютному значению градиентов	Бэггинг
L1 регуляризатор	Ручной ввод По умолчанию – 0	Регуляризация добавляет ограничения алгоритма относительно аспектов модели, которые не	Регуляризация Параметры для Linear Booster и Tree Booster (Бустер =

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		зависят от данных для обучения. Регуляризация обычно используется, чтобы избежать переобучения. Регуляризация L1 применяется для получения максимально разреженной модели	gblinear или Бустер = gbtree)
L2 регуляризатор	Ручной ввод По умолчанию – 1	Регуляризация L2 ограничивает чрезмерный рост какой-либо отдельной координаты весового вектора. Регуляризация L2 полезна в том случае, если целью является создание модели, имеющей в целом малые значения веса.	Регуляризация Параметры для Linear Booster и Tree Booster (Бустер = gblinear или Бустер = gbtree)
Тип сэмплинга	Раскрывающийся список со следующими значениями: • uniform (по умолчанию) • weighted	Данный параметр задает тип алгоритма сэмплинга. Предусмотрены следующие: • uniform –деревья имеют одинаковую вероятность быть отброшенными • weighted – отбрасываемые деревья выбираются пропорционально весу	Параметры для DART (Бустер = dart)
Тип алгоритма нормализации	Раскрывающийся список со следующими значениями: • Дерево (по умолчанию) • Лес	Данный параметр задает тип алгоритма нормализации. Предусмотрены следующие варианты: • Дерево – новые деревья имеют такой же вес, как и каждое из отбрасываемых деревьев. Вес новых деревьев равен 1/(k + скорость обучения). Отбрасываемые деревья масштабируются с коэффициентом k /(k + скорость обучения)	Параметры для DART (Бустер = dart)

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		• Лес – новые деревья имеют тот же вес, что и сумма отбрасываемых деревьев (леса). Вес новых деревьев равен 1/(k + скорость обучения). Отбрасываемые деревья масштабируются с коэффициентом k /(k + скорость обучения)	
Доля отбрасываемых деревьев	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0	Данный параметр задает долю отбрасываемых деревьев	Параметры для DART (Бустер = dart)
Отбрасывать хотя бы одно дерево	Чекбокс	Выбор данного чекбокса указывает на то, что по крайней мере одно дерево будет отбрасываться	Параметры для DART (Бустер = dart)
Вероятность пропуска отбрасывания деревьев	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0	Данный параметр задает вероятность пропуска процедуры отбрасывания деревьев. Если процедура отбрасывания деревьев пропущена, новые деревья добавляются так же, как и gbtree.	Параметры для DART (Бустер = dart)
Алгоритм	Раскрывающийся список со следующими значениями: • shortgun (по умолчанию) • coord descent	Данный параметр задает алгоритм для обучения линейной модели. Предусмотрены следующие алгоритмы: • shortgun – алгоритм спуска с параллельными координатами • coord descent – алгоритм спуска по обычным координатам	Параметры для Linear Booster (Бустер = gblinear)
Метод выбора переменных	Раскрывающийся список со следующими значениями: • cyclic (по умолчанию) • shuffle	Данный параметр задает метод выбора переменных. Предусмотрены следующие методы: • cyclic – детерминированны	Параметры для Linear Booster (Бустер = gblinear)

Параметр	Возможные значения и ограничения	Описание	Группа параметров
	• random • greedy • thrifty	й выбор путем циклического перебора признаков по одному • shuffle – похоже на cyclic, но со случайным перемешиванием функций перед каждым обновлением • random – случайный (с заменой) селектор координат • greedy – выбирается координата с наибольшей величиной градиента. • thrifty – перед циклическими обновлениями функции переупорядочиваются по убыванию величины их одномерных изменений веса	
Количество отбираемых переменных для жадных алгоритмов	Числовое значение По умолчанию – 0	Данный параметр задает количество отбираемых переменных для жадных алгоритмов. Значение, равное 0, означает использование всех функций.	Параметры для Linear Booster (Бустер = gblinear) Актуален, если выбраны метод выбора переменных = greedy или thrifty

Таблица 37 Параметры узла «Градиентный бустинг (XGBOOST)»

Результаты выполнения узла:

Узел «Градиентный бустинг (XGBOOST)» имеет разные результаты в зависимости от решаемой задачи.

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты обучающей, валидационной и тестовой выборок.
- Таблица с метриками качества модели.
- Таблица со списком переменных, сортированных по важности.

Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица со списком переменных, сортированных по важности.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC.
- График Lift.
- График Cumulative Lift.
- График Gain.
- График Cumulative Gain.
- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица со списком переменных, сортированных по важности.

3.2.5.8.10. Узел «Градиентный бустинг (LightGBM)»

В основе **узла «LightGBM»** лежит реализация алгоритма градиентного бустинга на деревьях поиска решений, который включает в себя две ключевые идеи: **GOSS** и **EFB**.

Используется для решения задач классификации и регрессии.

Алгоритм работы градиентного бустинга:

Градиентный бустинг – алгоритм машинного обучения, который строит модель предсказания в виде ансамбля слабых предсказывающих моделей (в основном Дерево решений). На каждой итерации вычисляется отклонение предсказаний уже обученного ансамбля на обучающей выборке. Следующая добавляемая в ансамбль модель будет сводить среднее отклонение предыдущей к минимуму.
Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил «ранней остановки».

Особенности реализации LightGBM:

- Используется **алгоритм роста дерева по листьям** (в каждом узле рассчитывается дает ли разделение листа прирост информации – inf gain), поэтому также необходимо ограничивать построение дерева максимальной глубиной для предотвращения переобучения.
- LightGBM использует **алгоритмы на основе гистограмм**, которые разбивают значения непрерывных признаков (атрибутов) на дискретные ячейки.
- **Градиентная односторонняя выборка (Gradient-based One-Side Sampling - GOSS)**

Для расчета берутся не все наблюдения, а лишь с большой ошибкой и небольшая часть наблюдений с маленькой ошибкой. Данный подход позволяет уменьшить расчетную часть алгоритма, за счет чего растет скорость расчета.
При расчете прироста информации GOSS вводит постоянный множитель для экземпляров данных с небольшими градиентами. Т. е. GOSS сначала сортирует наблюдения в соответствии с абсолютными значениями их градиентов и выбирает верхние экземпляры a*100% (a - **коэффициент сохранения большого градиента**). Затем он случайным образом отбирает наблюдения b*100% (b - **коэффициент сохранения большого градиента**) из остальных данных. После этого GOSS усиливает выборочные данные с небольшими градиентами на константу (1-a)/b при расчете прироста информации. Это позволяет учитывать исходное распределение данных.

- **Встроенная кодировка категориальных переменных:**

Считается сумма градиентов по каждой категории и категории объединяются в группы.

- **Объединение взаимоисключающих признаков (**Exclusive Feature Bundling - **EFB)**

Данный подход подразумевает объединение разреженных (в основном нулевых взаимоисключающих признаков, таких как категориальные переменные, закодированные унитарным кодированием.
Переменные, которые не принимают одновременно ненулевые значения, можно «связать». Уменьшаем количество переменных, объединяя взаимоисключающие в 1 переменную, увеличивается скорость расчета.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе	Общий параметр
Описание	Ручной ввод Ограничений на значение нет	Описание узла	Общий параметр
Бустер	Раскрывающийся список со следующими значениями: • gbtree (по умолчанию) • DART • GOSS • random forest	Данный параметр задает тип базового алгоритма для бустинга. Предусмотрены следующие типы: • gbtree – бустинг на основе деревьев • DART – модификация gbtree (отбрасывает деревья, для предотвращения переобучения)	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		• GOSS – используются только те экземпляры, которые приводят к большому градиенту ошибки, для обновления модели и удаления остальных экземпляров. • random forest – Случайный лес	
Линейные модели в листах	Чекбокс	Выбор данного чекбокса указывает на необходимость обучения линейных моделей в листах каждого дерева	Общий параметр
Количество оценочных функций	Ручной ввод Целочисленное значение больше или равно 0 По умолчанию - 100	Данный параметр задает число итераций градиентного бустинга	Общий параметр
Скорость обучения	Ручной ввод Число больше 0 и меньше 1 По умолчанию — 0,3	Данный параметр задает скорость обучения модели и контролирует, с каким весом предсказания каждой следующей модели суммируются с предсказаниями ансамбля.	Общий параметр
Цель обучения для регрессии	Раскрывающийся список со следующими значениями: • Среднеквадратическая ошибка (по умолчанию) • Средняя абсолютная ошибка • Функция потерь Хьюбера • Регрессия Твиди • Гамма регрессия • Пуассоновская регрессия • Квантильная регрессия • MAPE	Данный параметр задает используемую при обучении функцию потерь. Предусмотрены следующие: • Среднеквадратическая ошибка • Средняя абсолютная ошибка • Функция потерь Хьюбера – функция квадратична для малых значений остатка (разница между наблюдаемым и предсказанным значением), и линейна для	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		больших значений остатка • Регрессия Твиди – предназначена для прогнозирования целевой переменной, имеющей распределение Твиди (например, общее количество осадков в год или общее время прерывания в год). • Гамма регрессия – предназначена для прогнозирования целевой переменной, имеющей гамма-распределение (например, количество осадков на одно событие или продолжительность прерывания) • Пуассоновская регрессия – предназначена для прогнозирования счетчиков (неотрицательных целых чисел) (например, количество дождевых явлений в год или количество событий прерывания производства в год). • Квантильная регрессия – предназначена для моделирования взаимосвязи между набором предикторов и	

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		определенными процентилями (квантилями) целевой переменной • MAPE - средняя абсолютная ошибка в процентах	
Цель обучения для классификации	Раскрывающийся список со следующими значениями: • Бинарная логистическая регрессия (по умолчанию) • Мультиклассовая с softmax	Данный параметр задает используемую при обучении функцию потерь. Предусмотрены следующие: • Бинарная логистическая регрессия – возвращает прогнозируемую вероятность (не класс) • Мультиклассовая с softmax – функция softmax для мультиклассовой классификации, возвращает класс с максимальной вероятностью принадлежности	Общий параметр
Метрика для валидации регрессии	Раскрывающийся список со следующими значениями: Среднеквадратическая ошибка (по умолчанию) Средняя абсолютная ошибка RMSE MAPE Пуассоновская регрессия Функция потерь Хьюбера Квантильная регрессия Гамма регрессия Гамма-отклонение Регрессия Твиди	Данный параметр задает метрику качества на валидационной выборке. Предусмотрены следующие метрики: • Среднеквадратическая ошибка (по умолчанию) • Средняя абсолютная ошибка • RMSE • MAPE • Пуассоновская регрессия • Функция потерь Хьюбера • Квантильная регрессия • Гамма регрессия • Гамма-отклонение • Регрессия Твиди	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Метрика для валидации классификации	Раскрывающийся список со следующими значениями: • Бинарная логистическая регрессия (по умолчанию) • Бинарная ошибка • Мультиклассовая логистическая регрессия • Мультиклассовая ошибка • AUC	Данный параметр задает метрику качества на валидационной выборке. Предусмотрены следующие метрики: • Бинарная логистическая регрессия – логистическая функция ошибки • Бинарная ошибка – частота ошибок бинарной классификации, рассчитывается как неправильно классифицированные объекты/все объекты. При прогнозировании положительными экземплярами будут считаться наблюдения со значением прогноза больше 0,5, остальные – как отрицательные • Мультиклассовая логистическая регрессия – мультиклассовая логистическая функция ошибки • Мультиклассовая ошибка – частота ошибок мультиклассовой классификации, рассчитывается как неправильно классифицированные объекты/все объекты • AUC – количественная интерпретация кривой ошибок, площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Количество итераций до ранней остановки	Ручной ввод По умолчанию - 0	Данный параметр задает количество итераций до ранней остановки	Общий параметр
Размер (%) валидационной выборки для ранней остановки	Ручной ввод Число больше 0 и меньше 1 По умолчанию – 0,1	Данный параметр задает размер (%) валидационной выборки для ранней остановки	Общий параметр
Seed	Ручной ввод числового значения По умолчанию – 42	Начальное числовое значение для генератора случайных чисел	Общий параметр
Количество параллельных потоков	Ручной ввод По умолчанию – -1	Данный параметр задает количество параллельных потоков, используемых для запуска	Общий параметр
Максимальная глубина дерева	Ручной ввод По умолчанию – -1	Данный параметр задает максимальную глубину дерева. Значение меньше или равно 0 означает отсутствие ограничений	Построение дерева
Максимальное количество листов	Ручной ввод Целочисленное значение больше 0 По умолчанию - 31	Данный параметр задает максимальное количество листов в дереве	Построение дерева
Минимальное количество наблюдений для создания бина	Ручной ввод Целочисленное значение больше 0 По умолчанию – 200000	Данный параметр задает количество наблюдений для создания бина	Построение дерева
Максимальное количество бинов	Ручной ввод Целочисленное значение больше 1 По умолчанию – 255	Данный параметр задает максимальное количество бинов, в которые будут группироваться значения признаков	Построение дерева
Минимальное количество наблюдений в бине	По умолчанию – 3	Данный параметр задает минимальное число наблюдений в бине	Построение дерева
Использовать EFB	Чекбокс	Выбор данного чекбокса указывает на необходимость использовать EFB	Построение дерева
Минимальный gain для разбиения	Ручной ввод По умолчанию – 0	Данный параметр задает минимальное снижение потерь, необходимое для создания дальнейшего раздела на листовом узле дерева	Построение дерева
Минимальный вес для разбиения	По умолчанию – 0,0001	Данный параметр задает минимальную сумму весов экземпляров	Построение дерева

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		(Гессиан), необходимую для дочернего элемента (листа)	
Минимальное количество наблюдений в листе	По умолчанию – 20	Данный параметр задает минимальное количество наблюдений в листе	Построение дерева
Доля выборки признаков в каждой итерации	По умолчанию – 1	Данный параметр задает долю переменных, используемых на каждой итерации (при построении каждого дерева).	Построение дерева
Доля выборки признаков в каждом узле дерева	По умолчанию – 1	Данный параметр задает долю подвыборки признаков, которые будут использованы для каждого разделения (узла).	Построение дерева
Использовать экстремально рандомизированные деревья	Чекбокс	Выбор данного чекбокса указывает на необходимость использования экстремально рандомизированных деревьев. В экстремально рандомизированных деревьях используется случайное подмножество объектов-кандидатов, но вместо поиска наиболее отличительных пороговых значений пороги выбираются случайным образом для каждого объекта-кандидата, и лучший из этих случайно сгенерированных пороговых значений выбирается в качестве правила разделения. Обычно это позволяет еще немного уменьшить дисперсию модели за счет чуть большего увеличения смещения	Построение дерева
Доля отбрасываемых деревьев	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0,1	Данный параметр задает долю отбрасываемых деревьев	Параметры для DART (Бустер = dart)

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Максимальное количество отбрасываемых деревьев	По умолчанию – 50	Данный параметр задает максимальное число отбрасываемых деревьев	Параметры для DART (Бустер = dart)
Вероятность отсутствия отбрасывания	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0,5	Данный параметр задает вероятность пропуска процедуры отбрасывания деревьев. Если процедура отбрасывания деревьев пропущена, новые деревья добавляются так же, как и gbtree	Параметры для DART (Бустер = dart)
Использовать XGBoost DART	Чекбокс	Выбор данного чекбокса указывает на необходимость использования XGBoost DART	Параметры для DART (Бустер = dart)
Использовать uniform drop	Чекбокс	Выбор данного чекбокса указывает на то, что отбрасываемые деревья выбираются равномерно.	Параметры для DART (Бустер = dart)
Коэффициент сохранения большого градиента	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0,2	Данный параметр задает коэффициент сохранения наблюдений с большим градиентом	Параметры для GOSS (Бустер = GOSS)
Коэффициент сохранения маленького градиента	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 0,1	Данный параметр задает коэффициент сохранения наблюдений с маленьким градиентом	Параметры для GOSS (Бустер = GOSS)
Соотношение случайной подвыборки в обучающей выборке	Ручной ввод Число больше или равно 0 и меньше или равно 1 По умолчанию – 1	Данный параметр задает соотношение случайной подвыборки в обучающей выборке. Для того, чтобы процесс бэггинга был запущен необходимо указать значение меньше 1.	Параметры бэггинга
Частота случайной подвыборки	Ручной ввод целочисленного значения По умолчанию – 0	Данный параметр задает выполнение бэггинга на каждой **k** (заданной значением) итерации. Каждая **k**-я итерация LightGBM будет случайным образом отбирать '**Соотношение случайной подвыборки в обучающей выборке**'*100% данных для использования в следующих k итерациях. Значение параметра	Параметры бэггинга

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		равное 0 означает отключение бэггинга	
L1 регуляризатор	По умолчанию – 0	Регуляризация добавляет ограничения алгоритма относительно аспектов модели, которые не зависят от данных для обучения. Регуляризация обычно используется, чтобы избежать переобучения. Регуляризация L1 применяется для получения максимально разреженной модели	Параметры регуляризации
L2 регуляризатор	По умолчанию – 1	Регуляризация L2 ограничивает чрезмерный рост какой-либо отдельной координаты весового вектора. Регуляризация L2 полезна в том случае, если целью является создание модели, имеющей в целом малые значения веса	Параметры регуляризации
L2 регуляризатор для линейной регрессии	По умолчанию – 0	Данный параметр задает линейную регуляризацию дерева	Параметры регуляризации
L2 регуляризатор в категориальном разбиении	Ручной ввод Число больше или равно 0 По умолчанию – 10	Данный параметр используется для задания регуляризации категориального разделения (сплита)	Параметры регуляризации
Минимальное количество наблюдений в категориальной переменной	Ручной ввод Число больше 0 По умолчанию – 100	Данный параметр задает минимальное количество наблюдений в категориальной переменной	Параметры для обработки категориальных признаков
Лимит количества разбиений для категориальной переменной	По умолчанию – 32	Данный параметр задает ограничение на количество разбиений (сплитов) для категориальной переменной	Параметры для обработки категориальных признаков
Псевдосчет в сглаживании Лапласа	Ручной ввод Число больше или равно 0 По умолчанию – 10	Данный параметр позволяет уменьшить влияние шумов на категориальные переменные, особенно для категорий с небольшим количеством данных.	Параметры для обработки категориальных признаков

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Максимальное количество категории для One-Hot Encoding	Ручной ввод Число больше 0 По умолчанию – 4	Когда количество категорий одной переменной меньше или равно данному значению, будет использоваться алгоритм разделения one-vs-other («один против другого»)	Параметры для обработки категориальных признаков

Таблица 38 Параметры узла «LightGBM»

Результаты выполнения узла:

Узел «LightGBM» имеет разные результаты в зависимости от решаемой задачи.

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты обучающей, валидационной и тестовой выборок.
- Таблица с метриками качества модели.
- Таблица со списком переменных, сортированных по важности.

Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица со списком переменных, сортированных по важности.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC.
- График Lift.
- График Cumulative Lift.
- График Gain.
- График Cumulative Gain.
- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица со списком переменных, сортированных по важности.

3.2.5.8.11. Узел «Градиентный бустинг (CatBoost)»

В основе **узла «CatBoost»** лежит реализация алгоритма градиентного бустинга, которая оптимизирована под работу с категориальными признаками и хорошо работает с параметрами по умолчанию.

Используется для решения задач классификации и регрессии.

Алгоритм работы градиентного бустинга:

Градиентный бустинг – алгоритм машинного обучения, который строит модель предсказания в виде ансамбля слабых предсказывающих моделей (в основном Дерево решений). На каждой итерации вычисляется отклонение предсказаний уже обученного ансамбля на обучающей выборке. Следующая добавляемая в ансамбль модель будет сводить среднее отклонение предыдущей к минимуму.
Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил «ранней остановки».
Особенности реализации CatBoost:

- CatBoost строит симметричные (сбалансированные) деревья, т.е. на каждом шаге листья предыдущего дерева разделяются по одному и тому же условию.
- Алгоритм сам выбирает лучший способ обработки категориальных признаков в зависимости от решаемой задачи
- Хорошо работает с гиперпараметрами по умолчанию

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе	Общий параметр
Описание	Ручной ввод Ограничений на значение нет	Описание узла	Общий параметр
Количество деревьев	Ручной ввод По умолчанию - 100	Данный параметр задает максимальное количество деревьев	Общий параметр
Скорость обучения	Ручной ввод По умолчанию – 0,03	Данный параметр задает скорость обучения, которая определяет насколько быстро или медленно модель будет учиться	Общий параметр
Цель обучения для регрессии	Раскрывающийся список со следующими значениями: • Среднеквадратическая ошибка • Средняя абсолютная ошибка • Функция потерь Хьюбера • Регрессия Твиди • Пуассоновская регрессия • Квантильная регрессия • MAPE	Данный параметр задает регрессионный показатель, используемый для обучения. Предусмотрены следующие показатели: • Среднеквадратическая ошибка • Средняя абсолютная ошибка • Функция потерь Хьюбера • Регрессия Твиди • Пуассоновская регрессия	Общий параметр

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		• Квантильная регрессия • MAPE	
Цель обучения для классификации	Раскрывающийся список со следующими значениями: • Logloss (по умолчанию) • CrossEntropy • MultiClass	Данный параметр задает классификационный показатель, используемый для обучения. Предусмотрены следующие показатели: • Logloss • CrossEntropy • MultiClass	Общий параметр
Метрика для валидации регрессии	Раскрывающийся список со следующими значениями: • Среднеквадратическая ошибка (по умолчанию) • Средняя абсолютная ошибка • MAPE • SMAPE • R2 • Пуассоновская регрессия • Функция потерь Хьюбера • Квантильная регрессия • Регрессия Твиди	Данный параметр задает метрику, используемую для обнаружения переобучения. Предусмотрены следующие: • Среднеквадратическая ошибка • Средняя абсолютная ошибка • MAPE • SMAPE • R2 • Пуассоновская регрессия • Функция потерь Хьюбера • Квантильная регрессия • Регрессия Твиди	Общий параметр
Метрика для валидации классификации	Раскрывающийся список со следующими значениями: • Logloss (по умолчанию) • CrossEntropy • MultiClass • Precision • F1 • Accuracy • AUC	Данный параметр задает метрику, используемую для обнаружения переобучения	Общий параметр
L2 регуляризатор	Ручной ввод По умолчанию – 3	Данный параметр задает коэффициент при члене регуляризации L2 функции потерь	Общий параметр
Детектор переобучения	Раскрывающийся список со следующими значениями: • IncToDec (по умолчанию) • Iter	Данный параметр задает тип детектора переобучения. Предусмотрены: • IncToDec – обучение	Параметр ранней остановки

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		останавливается при достижении порога • Iter –обучение останавливается прсле указанного количества итераций, начиная с итерации с оптимальным значением метрики	
Порог для IncToDec детектора переобучения	Ручной ввод По умолчанию - 0	Данный параметр задает порог для IncToDec детектора переобучения.	Параметр ранней остановки Актуален при заданном Детекторе переобучения = IncToDec
Количество итераций до ранней остановки	Ручной ввод По умолчанию - 0	Данный параметр задает количество итераций до ранней остановки	Параметр ранней остановки
Количество потоков для обучения	Ручной ввод По умолчанию - 0	Данный параметр задает количество потоков для обучения	Параметр производительности
Коэффициент регуляризации размера модели	Ручной ввод Целое число больше или равно 0 По умолчанию – 0,5	Данный параметр задает коэффициент регуляризации модели. Чем больше значение, тем меньше размер модели	Параметр производительности
Seed	Ручной ввод Целочисленное значение По умолчанию - 42	Начальное числовое значение для генератора случайных чисел	Общий параметр
Максимальная глубина дерева	Ручной ввод Целочисленное значение По умолчанию - 6	Данный параметр задает максимальную глубину дерева	Параметры построения дерева
Стратегия построения дерева	Раскрывающийся список со следующими значениями: • SymmetricTree (по умолчанию) • Depthwise • Lossguide	Данный параметр определяет, как будет применяться жадный алгоритм поиска. Предусмотрены: • SymmetricTree – дерево строится уровень за уровнем, пока не достигнет необходимой глубины. На каждом шаге листья с предыдущего дерева	Параметры построения дерева

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		разделяются с тем же условием. • Depthwise – дерево строится шаг за шагом, пока не достигнет необходимой глубины. Листья разделяются с использованием условия, которое приводит к лучшему уменьшению потерь. • Lossguide – дерево строится по листьям до тех пор, пока не будет достигнуто заданное количество листьев. На каждом шаге разделяется нетерминальный лист с лучшим уменьшением потерь.	
Минимальное количество наблюдений в листе	Ручной ввод По умолчанию – 1	Данный параметр задает минимальное количество обучающих наблюдений в листе	Параметры построения дерева
Скоринговая функция	Раскрывающийся список со следующими значениями: • Cosine (по умолчанию) • L2	Данный параметр задает тип оценки, используемый для выбора следующего разбиения при построении дерева. Предусмотрены следующие: • Cosine • L2	Параметры построения дерева
Доля выборки признаков в каждом разбиении	Ручной ввод По умолчанию – 1	Данный параметр задает долю переменных, используемых в каждом разбиении	Параметры построения дерева
Максимальное количество бинов	Ручной ввод По умолчанию – 254	Данный параметр задает максимальное количество бинов, в которые будут группироваться значения признаков	Параметры построения дерева

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Тип квантилизации интервальных переменных	Раскрывающийся список со следующими значениями: • GreedyLogSum (по умолчанию) • Median • Uniform • UniformAndQuantiles • MaxLogSum • MinEntropy	Данный параметр задает тип квантования интервальных переменных. Предусмотрены следующие: • GreedyLogSum • Median – включение примерно одинакового количества объектов в bucket • Uniform – разбиение происходит разделением сегмента (минимальное значение переменной – максимальное значение переменной) на подсегменты одинаковой длины. В этом случае используются абсолютные значения признака • UniformAndQuantiles – комбинируются сплиты, полученные в типах Median и Uniform, предварительно уменьшив размер квантования двое • MaxLogSum • MinEntropy	Параметры построения дерева
Максимальное количество листов	Ручной ввод По умолчанию – 31	Данный параметр задает максимальное количество листьев в результирующем дереве. Актуален, если Стратегия построения дерева = Lossguide	Параметры построения дерева
Метод обработки пропусков	Раскрывающийся список со следующими значениями:	Данный параметр задает метод работы с пропущенными значениями.	Параметры построения дерева

Параметр	Возможные значения и ограничения	Описание	Группа параметров
	• Forbidden (по умолчанию) • Min • Max	Предусмотрены следующие: • Forbidden – наличие пропущенных значений вызовет ошибку • Min – пропущенные значения будут приняты за максимальные значения для данного признака • Max – пропущенные значения будут приняты как минимальные значения для данного признака.	
Метод вычисления значений в листах	Раскрывающийся список со следующими значениями: • Auto (по умолчанию) • Newton • Gradient • Exact	Данный параметр задает метод вычисления значений в листах. Данный параметр зависит от режима и выбранной функции потерь: • Регрессия с функциями потерь Quantile или MAE — одна точная итерация (Exact). • Регрессия с любой функцией потерь, кроме Quantile или MAE — одна итерация градиента (Gradient). • Режим классификации – Десять ньютоновских итераций (Newton). • Режим мультиклассификации – одна итерация по Ньютону (Newton).	Параметры построения дерева

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Количество итераций в листах	Ручной ввод По умолчанию – 0	Данный параметр регулирует количество шагов, выполняемых в каждом дереве при вычислении значений листьев	Параметры построения дерева
Тип отступа при вычислении значений в листах	Раскрывающийся список со следующими значениями: • AnyImprovement (по умолчанию) • No • Armijo	Данный параметр задает тип бэктрекинга, использующийся при градиентном спуске. Предусмотрены следующие: • AnyImprovement – уменьшает шаг спуска до того, как значение функции потерь будет меньшим, чем оно было на последней итерации. • No • Armijo – уменьшает шаг спуска до тех пор, пока не будет выполнено услови е Вольфе.	Параметры построения дерева
Тип Bootstrap	Раскрывающийся список со следующими значениями: • Bernoulli (по умолчанию) • Bayesian • MVS • No	Данный параметр определяет метод семплинга весов объектов	Бэггинг
Соотношение случайной подвыборки в обучающей выборке	Ручной ввод По умолчанию – 0,8	Данный параметр задает соотношение случайной подвыборки в обучающей выборке	Бэггинг
Частота сэмлирования весов при построении деревьев	Раскрывающийся список со следующими значениями: • PerTreeLevel (по умолчанию) • PerTree	Данный параметр задает частоту выборки весов и объектов при построении деревьев. Предусмотрены: • PerTreeLevel – перед построением каждого нового дерева • PerTree – перед выбором каждого нового	Бэггинг

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		разделения дерева	
Распределение весов в Байесовском Bootstrap	Ручной ввод Целое число больше или равно 0 По умолчанию – 1	Данный параметр определяет распределение из которого выбираются веса для байесовского типа Bootstrap. Веса выбираются из экспоненциального распределения, если значение этого параметра установлено на 1. Все веса равны 1, если значение этого параметра установлено на 0.	Бэггинг
Вес знаменателя	Ручной ввод Число больше или равно 0 По умолчанию – 1	Данный параметр влияет на вес знаменателя и может использоваться для балансировки между сэмплинга на основе важности и Бернулли сэмплинг. Значение приближенное к 0 подразумевает использование сэмплинга на основе важности.	Бэггинг
Тип бустинга	Раскрывающийся список со следующими значениями: • Ordered (по умолчанию) • Plain	Данный параметр задает схему бустинга. Предусмотрены следующие: • Ordered – упорядоченная схема, обеспечивает лучшее качество на небольших наборах данных • Plain – простая для классической схемы градиентного бустинга	Параметры бустинга
Тип сжатия модели	Раскрывающийся список со следующими значениями: • Constant (по умолчанию) • Decreasing	Данный параметр определяет как фактический коэффициент сжатия модели рассчитывается на каждой итерации.	Параметры бустинга
Коэффициент сжатия модели	Ручной ввод Число в диапазоне (0, 1] По умолчанию – 1	Данный параметр задает константу, используемую для расчета коэффициента	Параметры бустинга

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		умножения модели на каждой итерации.	
Максимальное количество категорий для One-hot encoding	Ручной ввод По умолчанию – 2	Данный параметр задает использование one-hot encoding для всех категориальных признаков с количеством уникальных значений меньшим или равным заданному значению.	Параметры для обработки категориальных переменных
Максимальное количество объединяемых категорий	Ручной ввод Целое число больше или равное 1 и меньше или равно 16 По умолчанию – 4	Данный параметр задает максимальное количество объединяемых категорий	Параметры для обработки категориальных переменных
Максимальное количество листов с категориальными переменными	Ручной ввод Целое число больше или равно 0 По умолчанию – 0	Данный параметр задает максимальное количество листьев с категориальными признаками. Если количество превышает указанное значение, часть листьев отбраковывается	Параметры для обработки категориальных переменных

Таблица 39 Параметры узла «LightGBM»

Результаты выполнения узла:

Узел «CatBoost» имеет разные результаты в зависимости от решаемой задачи.

Результаты **регрессии** представлены следующими объектами:

- Тепловые карты обучающей, валидационной и тестовой выборок.
- Таблица с метриками качества модели.
- Таблица со списком переменных, сортированных по важности.

Результаты **многоклассовой классификации** представлены следующими объектами:

- Таблица с метриками качества модели.
- Таблица с метриками качества модели задачи классификации.
- Таблица со списком переменных, сортированных по важности.

Результаты **бинарной классификации** представлены следующими объектами:

- График ROC.
- График Lift.
- График Cumulative Lift.
- График Gain.

- График Cumulative Gain.
- Таблица с метриками качества модели.

3.2.5.8.12. Узел «GLM»

Обобщенная линейная модель (GLM), которая лежит в основе данного узла, обобщает линейную регрессию и допускает наличие у зависимой переменной распределения, отличающегося от нормального. GLM связывает зависимую переменную с факторами посредством задаваемой функции связи.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Тип распределения	Раскрывающийся список со следующими значениями: • Распределение Пуассона • Гамма распределение • Распределение Твиди	Данный параметр задает тип распределения обобщенной линейной модели. Предусмотрены следующие: • Распределение Пуассона • Гамма распределение • Распределение Твиди
Функция связи	Раскрывающийся список со следующими значениями: • auto • identity • log	Данный параметр задает функцию связи. Предусмотрены следующие: • auto • identity – для нормального распределения • log – для типа распределения Пуассона и гамма распределения
Распределение целевой переменной	Ручной ввод Число больше или равное 0 и меньше или равное 3 По умолчанию - 0	Данный параметр задает распределение целевой переменной, где: • 0 задает нормальное распределение • 1 задает распределение Пуассона • (1,2) задает составное распределение Пуассона - гамма • 2 задает гамма распределение • 3 задает обратное гауссово распределение Данный параметр доступен при выборе Типа распределения = Распределение Твиди
Добавить константу в модель	Чекбокс	Выбор данного чекбокса добавит константу в модель

Параметр	Возможные значения и ограничения	Описание
Стандартизация	Раскрывающийся список со следующими значениями: • Нет • Стандартное отклонение (по умолчанию) • Диапазон	Данный параметр отвечает за выбор метода стандартизации данных. Предусмотрены следующие методы: • Нет. • Стандартное отклонение – преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • Диапазон – линейно преобразует значения переменных в диапазон [0, 1].
L2	Ручной ввод Число больше или равное 0 По умолчанию – 0	Данный параметр задает константу, которая определяет силу регуляризации. L2 = 0 эквивалентна GLM без штрафов
Количество итераций	Ручной ввод Целое число больше или равно 1 По умолчанию - 100	Данный параметр задает максимальное количество итераций после достижения которого алгоритм останавливается
Допустимая погрешность	Ручной ввод Число больше 0 По умолчанию – 0,0001	Данный параметр задает допустимую погрешность, после достижения которой алгоритм останавливается

Таблица 40 Параметры узла «LightGBM»

Результаты выполнения узла:

- Тепловые карты обучающей, валидационной и тестовой выборок.
- Таблица с метриками качества модели.
- Таблица с коэффициентами переменных

3.2.5.8.13. Узел «Нейронная сеть (PyTorch)»

Узел "Нейронная сеть (PyTorch)" позволяет строить нейросеть типа **MLP (multilayer perceptron, многослойный перцептрон)**, в которой входной сигнал преобразуется в выходной, проходя последовательно через скрытые слои.

Данный узел отличается от узла "Нейронная сеть" (подробнее Узел «Нейронная сеть») возможностью более глубокой настройки конфигурации слоев нейросети и широким выбором оптимизаторов.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Описание	Ручной ввод Ограничений на значение нет	Описание узла	Общие параметры
Конфигурация слоев нейросети	Кнопка	При выборе данной кнопки откроется окно **Конфигурация слоев нейросети**, где можно задавать слои, количество нейронов в слоях и функции активации для каждого узла. Для этого необходимо выбрать кнопку **Добавить** и в появившемся списке выбрать необходимую функцию активации и настроить параметры. Предусмотрены: • **Функция активации ReLU (Rectified Linear Unit)** – $f(x) = \max(0, x)$ • **Функция активации CELU** – $f(x) = \max(0, x) + \min(0, \alpha * (\exp(x/\alpha) - 1))$ Дополнительно необходимо задать α • **Функция активации ELU** – $f(x) = \begin{cases} \alpha * (exp(x) - 1) & \text{for} \quad x \leq 0 \\ x & \text{for} \quad x > 0 \end{cases}$ Дополнительно необходимо задать α • **Функция активации Sigmoid** – $\sigma(x) = \frac{1}{1 + e^{-x}}$ • **Функция активации Softmax** – $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad \text{for } i = 1,2,\dots,K$ • **Линейный слой** – линейная функция, результат пропорционален переданному аргументу $f(x) = x$ Дополнительно необходимо задать **Количество выходных переменных** и при необходимости выбрать чекбокс **Добавить константу** • **Функция активации Logsigmoid** – $\text{LogSigmoid}(x) = \log(\frac{1}{1 + e^{-x}})$ • **Исключение (Dropout)** – Во время обучения случайным образом обнуляет некоторые	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		элементы входного тензора с вероятностью p, используя выборки из распределения Бернулли. Дополнительно необходимо задать **Вероятность исключения** • **Tanh** – функция гиперболического тангенса $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$ После настройки необходимой конфигурации слоев нейросети выбрать кнопку **Сохранить**.	
Функция потерь	Раскрывающийся список со следующими значениями: • MAE • MSE • Poisson loss • Negative log likelihood • Cross entropy	Данный параметр задает функцию потерь. Предусмотрены: • **MAE** (средняя абсолютная ошибка) • **MSE** (среднеквадратическая ошибка) • **Poisson loss** • **Negative log likelihood** (отрицательное логарифмическое правдоподобие) • **Cross entropy** (перекрестная энтропия)	Общие параметры
Алгоритм оптимизации	Раскрывающийся список со следующими значениями: • SGD • Adam (по умолчанию) • Adadelta • Adamax • LBFGS • AdamW • ASGD • NAdam • RAdam • Rprop • RMSprop	Данный параметр задает метод оптимизации, который будет использоваться для обновления весов нейронов скрытых слоев нейронной сети. Предусмотрены следующие методы: • **SGD (stochastic gradient descent)** – стохастический градиентный спуск. Данный метод делает шаг постоянной величины в направлении, указанном градиентом в текущей точке • **Adam (adaptive moment estimation)** – адаптивная оценка момента. Данный метод сочетает в себе идею метода Momentum о накоплении градиента и идею методов Adadelta и RMSProp об экспоненциальном сглаживании информации о предыдущих значениях квадратов градиентов. • **Adadelta (adaptive learning rate)** – метод адаптивной скорости обучения	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		• **Adamax** – модификация метода Adam, основанная на бесконечной норме (max) • **LBFGS (limited-memory BFGS)** – BFGS с ограниченной памятью. • **AdamW.** Данный метод основан на адаптивной оценке моментов первого и второго порядка с добавленным методом уменьшения весов • **ASGD (average stochastic gradient descent)** – усредненный стохастический градиентный спуск. Данный метод усредняет веса, вычисляемые на каждой итерации. • **Nadam (Nesterov-accelerated adaptive momentum).** Данный метод представляет собой модификацию оптимизатора Adam с добавлением момента Нестерова при вычислении градиентов. • **RAdam (Rectified Adam).** Данный метод является модификацией Adam, более устойчивой к изменению значений скорости обучения. • **Rprop (resilient backpropagation)** – устойчивый алгоритм обратного распространения. Данный метод использует только знаки частных производных для подстройки весовых коэффициентов. Также Rprop поддерживает отдельные дельты для каждого веса и смещения и адаптирует эти дельты во время обучения. • **RMSPROP (root mean square propagation)** – среднеквадратичное распространение корня. Данный метод использует усредненный по истории квадрат градиента.	

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Скорость обучения	Ручной ввод По умолчанию - 0,001	Данный параметр задает скорость обучения, которая управляет размером шага при обновлении весов.	Общие параметры
Количество эпох	Ручной ввод По умолчанию - 10	Данный параметр задает сколько раз алгоритм обучения будет обрабатывать весь набор обучающих данных.	Общие параметры
Размер пакета	Ручной ввод По умолчанию - 128	Данный параметр определяет количество выборок, которые необходимо обработать перед обновлением параметров модели.	Общие параметры
Seed	Ручной ввод По умолчанию - 42	Начальное числовое значение для генератора случайных чисел. Используется для воспроизведения результатов при повторном запуске узла.	Общие параметры
L2 регуляризация	Ручной ввод По умолчанию - 0,1	Данный параметр задает значение L2-регуляризации	Общие параметры
Доля валидационной выборки	Ручной ввод По умолчанию - 0,1	Данный параметр задает долю валидационной выборки, которая будет отобрана из исходной тестовой	Общие параметры
Количество итераций без существенного улучшения	Ручной ввод По умолчанию - 5	Данный параметр задает количество эпох без улучшения, после которых скорость обучения будет снижена. Пример: если значение параметра = 2, то первые 2 эпохи без улучшений loss будут проигнорированы, и только после 3-й эпохи скорость обучения уменьшится.	Общие параметры
Режим динамического расчета порога	Раскрывающийся список со следующими значениями: • rel • abs	Данный параметр задет режим расчета порога Предусмотрены: • **rel** – dynamic_threshold = best * (1 - threshold) • **abs** – dynamic_threshold = best - threshold	Общие параметры
Порог	Ручной ввод По умолчанию - 0,0001	Данный параметр задает порог расчета нового оптимума, чтобы сосредоточиться только на значительных изменениях	Общие параметры
Стандартизация	Раскрывающийся список со следующими значениями: • no • std (по умолчанию) • range	Данный параметр отвечает за выбор метода стандартизации числовых переменных. **Стандартизация** – преобразование числовых наблюдений с целью приведения их к некоторой общей шкале. Необходимость стандартизации вызвана тем, что разные признаки	Общие параметры

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		из обучающего набора могут быть представлены в разных масштабах и изменяться в разных диапазонах, что влияет на выявление некорректных зависимостей моделью. Предусмотрены следующие методы: • **no** — стандартизация не нужна • **std** — стандартное отклонение - преобразует наблюдения таким образом, чтобы их среднее значение равнялось нулю, а стандартное отклонение равнялось 1. • **range** — диапазон - линейно преобразует значения переменных в диапазон [0, 1].	
Beta1	Ручной ввод числа с плавающей точкой По умолчанию - 0,9	Данный параметр задает коэффициент, используемый для управления скоростью затухания скользящих средних значений градиента (первого момента)	Параметры алгоритма оптимизации **Adam, Adamax, AdamW, RAdam, Nadam**
Beta2	Ручной ввод числа с плавающей точкой По умолчанию - 0,999	Данный параметр задает коэффициент, используемый для управления скоростью затухания средних значений вторых моментов градиентов (нецентрированной дисперсии)	Параметры алгоритма оптимизации **Adam, Adamax, AdamW, RAdam, Nadam**
Epsilon	Ручной ввод числа с плавающей точкой По умолчанию - 1e-8	Данный параметр задает значение, добавляемое к знаменателю для улучшения числовой стабильности Minimal decay applied to lr. If the difference between new and old lr is smaller than eps, the update is ignored. Default: 1e-8.	Параметры алгоритма оптимизации **Adam, Adamax, AdamW, RAdam, Nadam**
Использовать алгоритм AMSGrad	Чекбокс	Выбор данного чекбокса указывает, что необходимо использовать вариант AMSGrad этого алгоритма. Разница между AMSgrad и Adam заключается в рассчитанном векторе второго момента, который используется для обновления параметров.	Параметры алгоритма оптимизации **Adam**, **AdamW**

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Импульс (Momentum)	Ручной ввод числа с плавающей точкой По умолчанию - 0	Данный параметр задает коэффициент импульса (запоминает скорость на предыдущем шаге и добавляет в указанное число раз меньшую величину на следующем шаге)	Параметры алгоритма оптимизации **SGD, RMSProp**
Dampening	Ручной ввод числа с плавающей точкой По умолчанию - 0	Данный параметр задает демпфирование импульса. Демпфирование гарантирует, что оптимизатор не сделает слишком больших шагов, что может произойти, если использовать только импульс. Чем выше градиент, тем больше демпфирование уменьшает размер шага.	Параметры алгоритма оптимизации **SGD**
Момент Нестерова	Чекбокс	Выбор данного чекбокса включает импульс Нестерова (использует производную не в текущей точке, а в следующей, если бы мы продолжали двигаться в этом же направлении без изменений)	Параметры алгоритма оптимизации **SGD**
Rho	Ручной ввод По умолчанию - 0,9	Данный параметр задает коэффициент, используемый для вычисления скользящего среднего квадратов градиентов	Параметры алгоритма оптимизации **Adadelta**
Epsilon	По умолчанию - 1e-8	Данный сглаживающий параметр задает значение, предотвращающее деление на 0.	Параметры алгоритма оптимизации **Adadelta, RMSProp**
Максимум итераций за шаг оптимизации	Ручной ввод целочисленного значения По умолчанию - 20	Данный параметр задает максимальное число итераций за шаг оптимизации	Параметры алгоритма оптимизации **LBFGS**
Максимум вычислений за шаг оптимизации	Ручной ввод целочисленного значения По умолчанию - 1	Данный параметр задает максимальное число вычислений функции за шаг оптимизации	Параметры алгоритма оптимизации **LBFGS**
Tolerance grad	Ручной ввод числа с плавающей точкой По умолчанию - 0,00001	Данный параметр задает допуск завершения при оптимальности первого порядка	Параметры алгоритма оптимизации **LBFGS**
Tolerance change	Ручной ввод числа с плавающей точкой По умолчанию - 1e-9	Данный параметр задает допуск завершения при изменении значения/параметра функции	Параметры алгоритма оптимизации **LBFGS**

Параметр	Возможные значения и ограничения	Описание	Группа параметров
Количество запоминаемых шагов оптимизации	Ручной ввод целочисленного значения По умолчанию - 100	Данный параметр задает количество запоминаемых шагов оптимизации	Параметры алгоритма оптимизации **LBFGS**
Line search	Список: • no (по умолчанию) • strong wolfe	Данный параметр задает метод линейного поиска	Параметры алгоритма оптимизации **LBFGS**
Lambda	Ручной ввод числа с плавающей точкой По умолчанию - 0,0001	Данный параметр задает затухание	Параметры алгоритма оптимизации **ASGD**
Alpha	Ручной ввод числа с плавающей точкой По умолчанию - 0,75	Данный параметр задает мощность для обновления скорости обучения	Параметры алгоритма оптимизации **ASGD**
t0	Ручной ввод По умолчанию - 100000	Данный параметр задает точку, с которой начинается усреднение. Если требуемое количество итераций меньше данного значения, то усреднение не произойдет.	Параметры алгоритма оптимизации **ASGD**
Сокращение импульса	Ручной ввод По умолчанию - 0,004	Данный параметр задает значение сокращения импульса	Параметры алгоритма оптимизации **Nadam**
Коэффициент уменьшения (eta minus)	Ручной ввод Число больше 0 и меньше 1 По умолчанию - 0,5	Данный параметр задает мультипликативный коэффициент уменьшения. Если на текущем шаге частная производная по соответствующему весу поменяла свой знак, значит последнее изменение было большим, и алгоритм проскочил локальный минимум. Следовательно, величину коррекции необходимо уменьшить на значение данного параметра и вернуть предыдущее значение весового коэффициента.	Параметры алгоритма оптимизации **Rprop**
Коэффициент увеличения (eta plus)	Ручной ввод Число больше 1 По умолчанию - 1,2	Данный параметр задает мультипликативный коэффициент увеличения. Если на текущем шаге частная производная по соответствующему весу не поменяла свой знак, значит нужно увеличить величину коррекции на значение данного	Параметры алгоритма оптимизации **Rprop**

Параметр	Возможные значения и ограничения	Описание	Группа параметров
		параметра для достижения более быстрой сходимости.	
Минимальный размер шага	Ручной ввод По умолчанию - 0,000001	Данный параметр задает минимальный размер шага. Он необходим, чтобы не допустить слишком маленьких значений весов, ограничивает величину коррекции снизу.	Параметры алгоритма оптимизации **Rprop**
Максимальный размер шага	Ручной ввод По умолчанию - 50	Данный параметр задает максимальный размер шага. Он необходим, чтобы не допустить слишком больших значений весов, ограничивает величину коррекции сверху.	Параметры алгоритма оптимизации **Rprop**
Alpha	Ручной ввод По умолчанию - 0,99	Данный параметр задает константу сглаживания	Параметры алгоритма оптимизации **RMSProp**
Центрировать	Чекбокс	Выбор данного чекбокса указывает, что необходимо вычислить центрированный RMSProp, градиент которого нормализуется по оценке его дисперсии	Параметры алгоритма оптимизации **RMSProp**

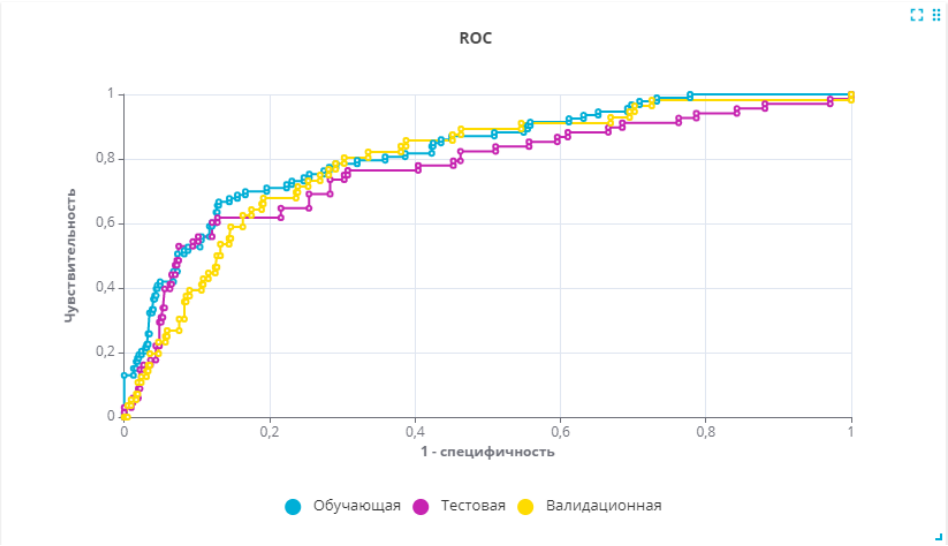
Таблица 41 Параметры узла «Нейронная сеть (PyTorch)»

3.2.5.8.13.1. Результаты выполнения узла

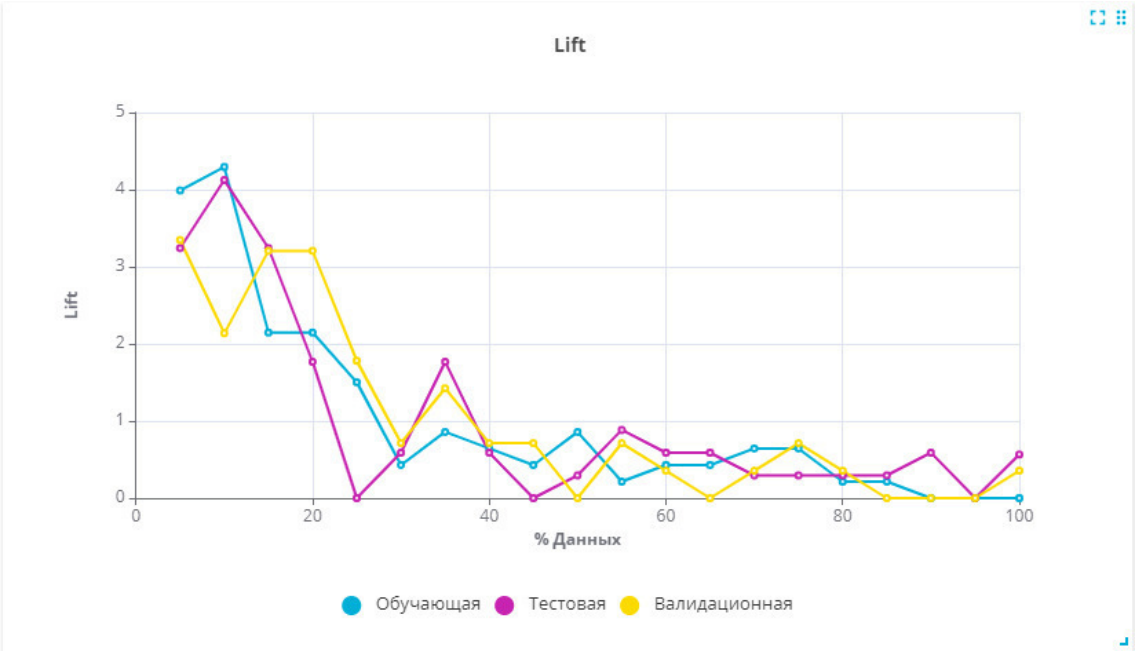
Узел «Нейронная сеть» имеет разные результаты в зависимости от решаемой задачи.

Результаты бинарной классификации представлены следующими объектами:

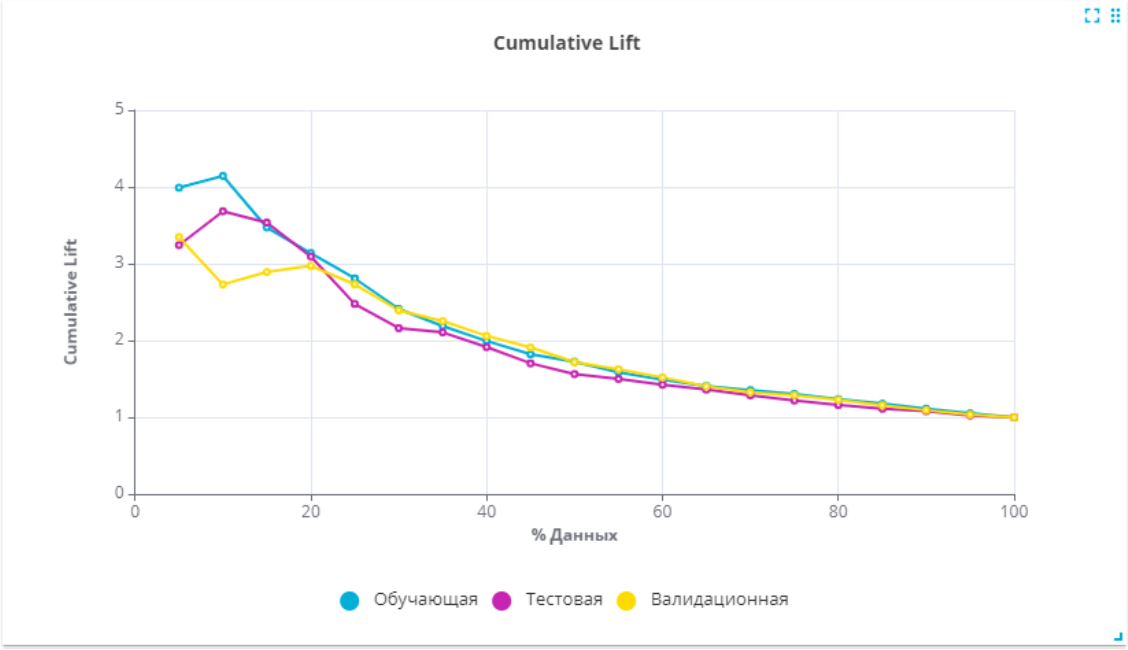
- График ROC.



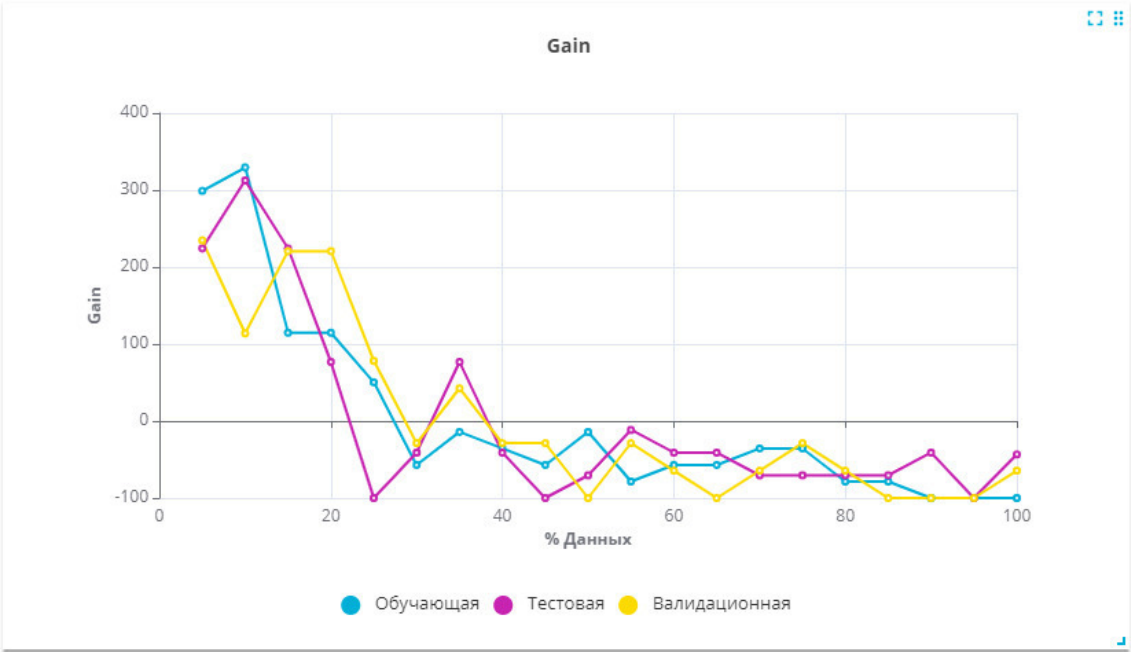
- График Lift.



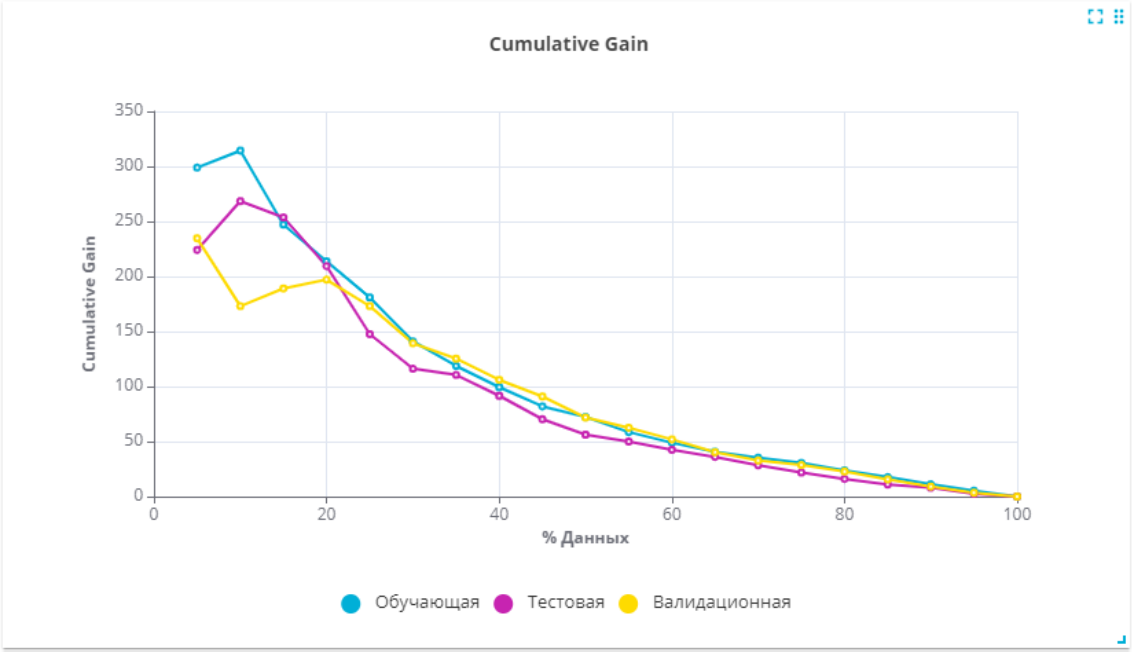
- График Cumulative Lift.



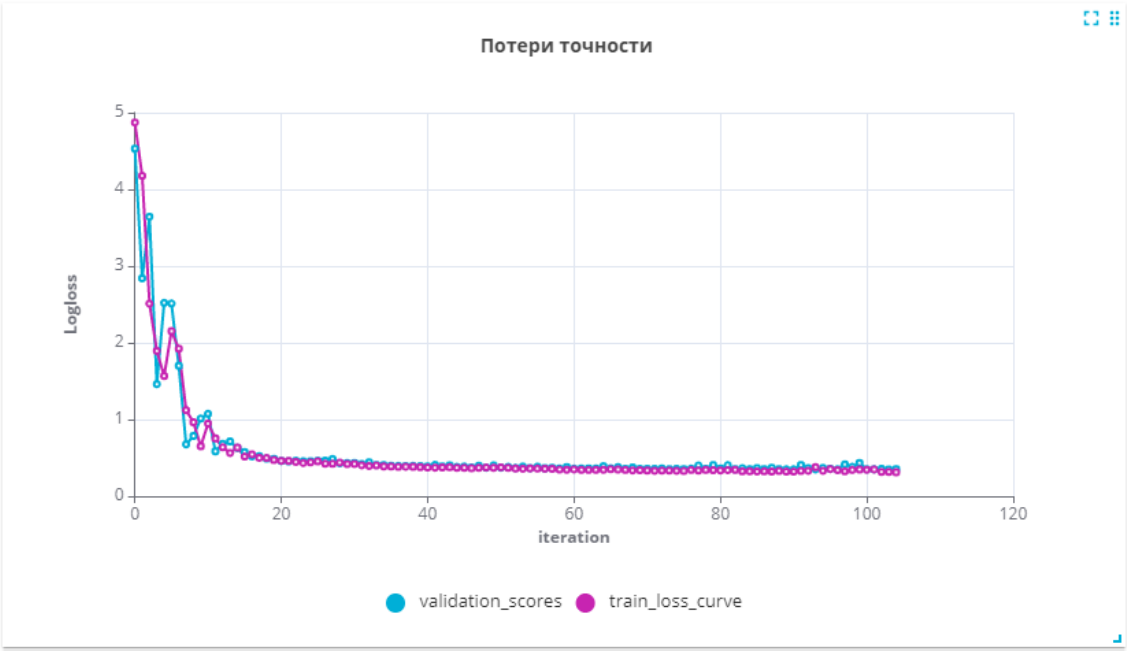
- График Gain.



- График Cumulative Gain.



- График потери точности.

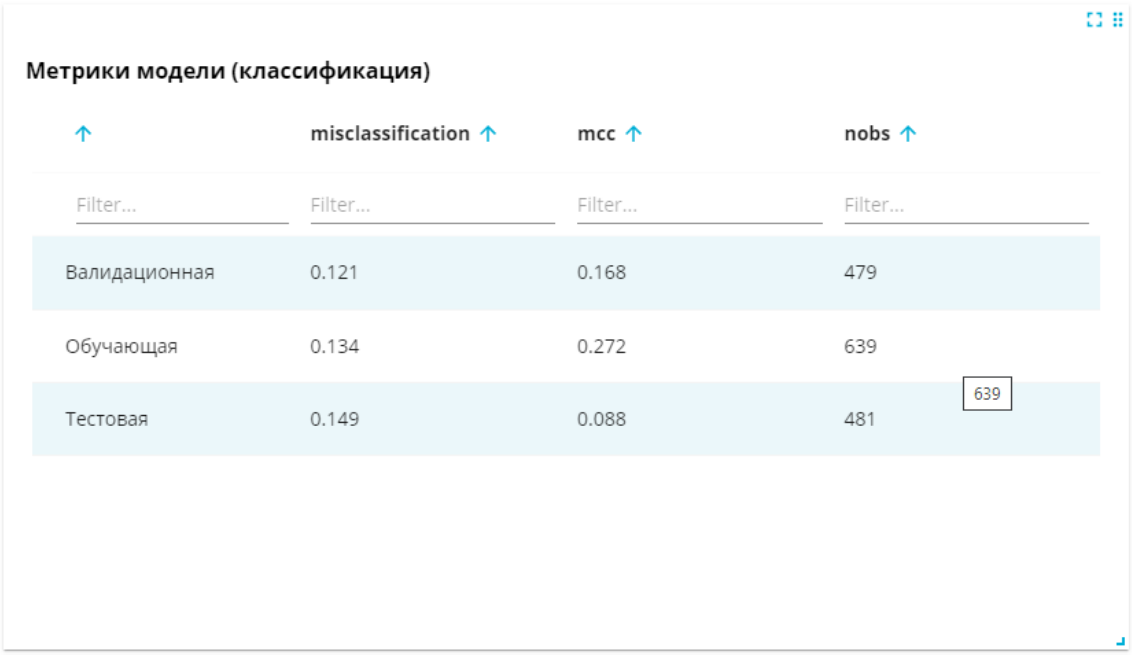


- Таблица с метриками качества модели.

Метрики модели

	AUC ROC	gini	log loss	nobs
Валидационная	0.793	0.587	0.336	479
Обучающая	0.822	0.644	0.318	639
Тестовая	0.77	0.541	0.362	481

- Таблица с метриками качества модели задачи классификации.



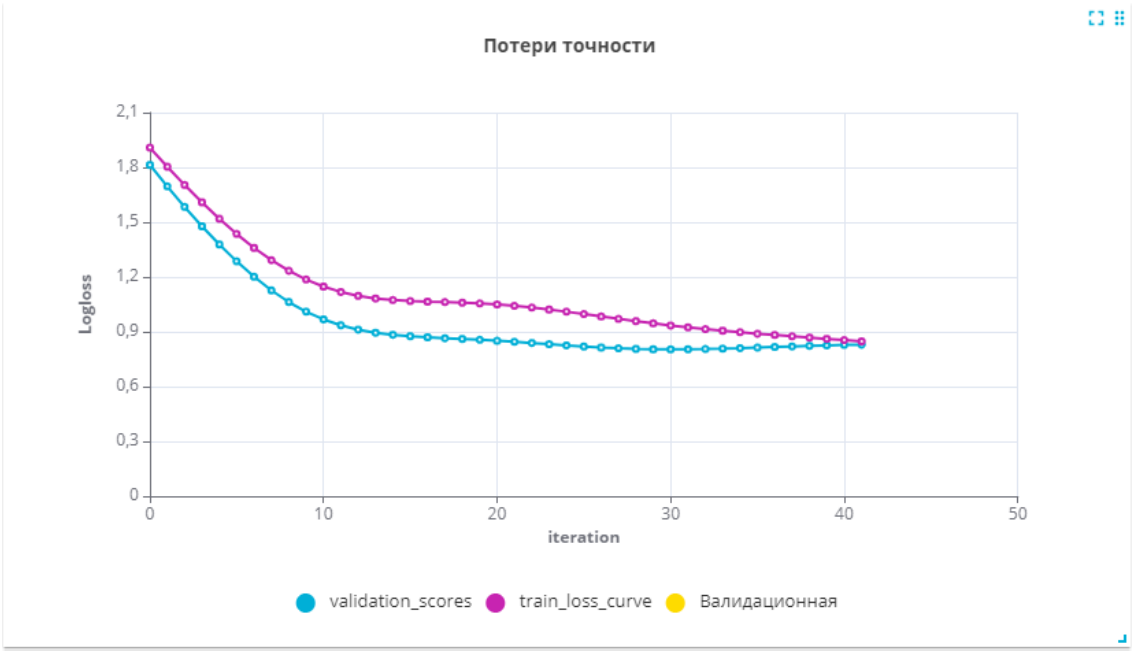
Метрики модели (классификация)

	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.121	0.168	479
Обучающая	0.134	0.272	639
Тестовая	0.149	0.088	481

639

Результаты многоклассовой классификации представлены следующими объектами:

- График потери точности.



- Таблица с метриками качества модели.

Метрики модели

	log loss	nobs
Filter...	Filter...	Filter...
Валидационная	1.052	35
Обучающая	0.911	47
Тестовая	1.003	37

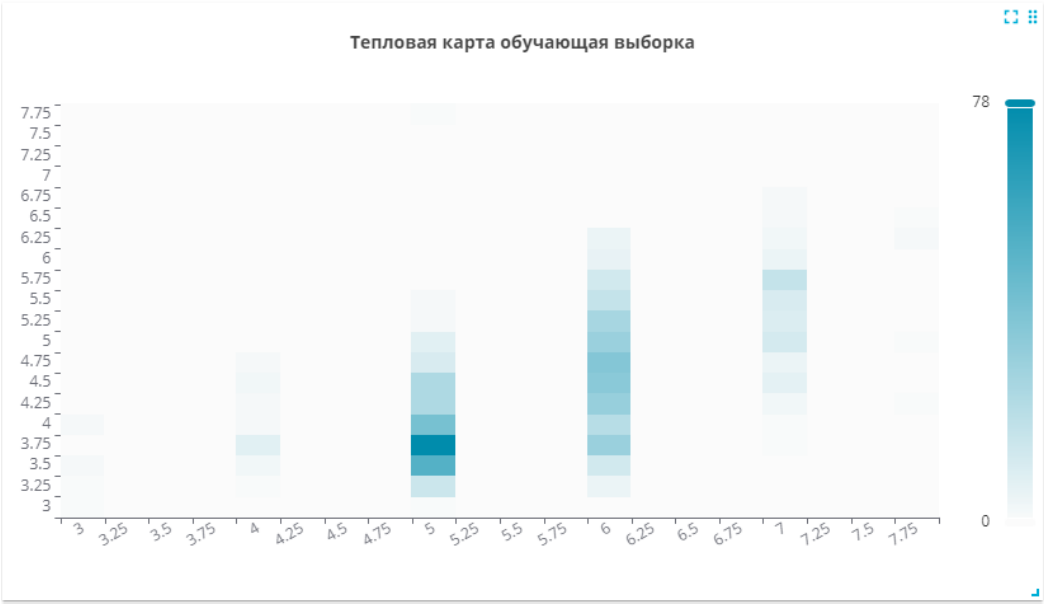
- Таблица с метриками качества модели задачи классификации.

Метрики модели (классификация)

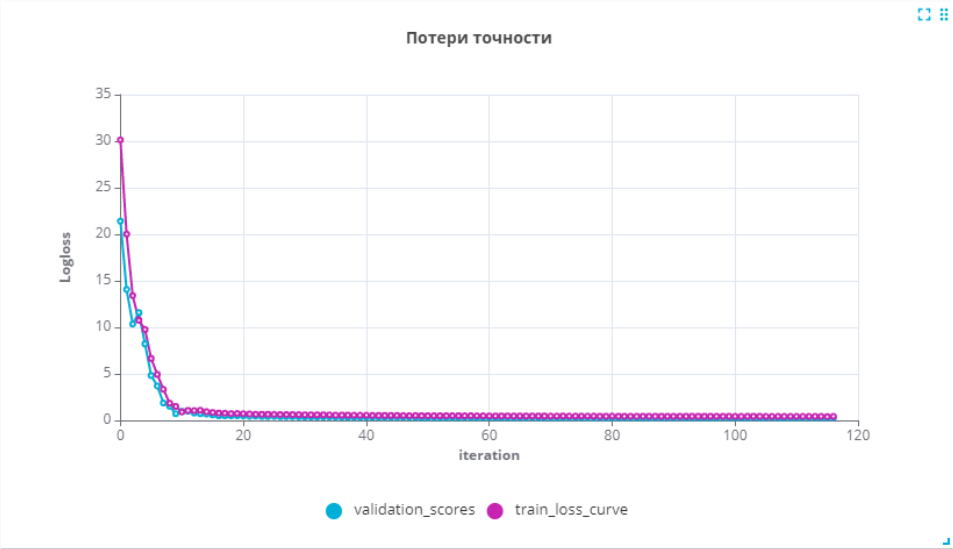
	misclassification	mcc	nobs
Filter...	Filter...	Filter...	Filter...
Валидационная	0.657	0	35
Обучающая	0.489	0	47
Тестовая	0.621	0	37

Результаты регрессии представлены следующими объектами:

- Тепловые карты для обучающей, валидационной и тестовой выборок.



- График потери точности.



- Таблица с метриками качества модели.

Метрики модели

	mse	rmse	mae	mape	r2	nobs
CV (валидационна	0.562	0.745	0.566	0.104	0.084	127.8
CV (обучающая)	0.509	0.705	0.533	0.098	0.218	511.2
Валидационная	0.527	0.726	0.567	0.103	0.178	479
Обучающая	0.412	0.642	0.487	0.09	0.361	639
Тестовая	0.471	0.686	0.529	0.096	0.293	481

- Таблицы с метриками качества в разбиениях обучающей и валидационной выборках.

Метрики в разбиениях обучающей выборки

Номер разбие...	fit_time	mse	rmse	mae	mape	r2	nobs
0	0.381	0.465	0.682	0.511	0.096	0.315	511
1	0.397	0.409	0.64	0.489	0.092	0.34	511
2	0.348	0.413	0.642	0.481	0.087	0.332	511
3	0.096	0.837	0.915	0.71	0.129	-0.224	511
4	0.342	0.42	0.648	0.474	0.086	0.328	512

3.2.5.8.14. Узел «AutoML»

Узел "AutoML" позволяет в автоматизированном режиме построить алгоритмы машинного обучения с различными параметрами и выбрать из них лучшую, исходя из заданной метрики качества.

Автоматическое машинное обучение (AutoML) позволяет автоматизировать процесс дизайна ML-пайплайнов. Это позволяет сократить требуемые ресурсы на процесс отбора оптимальных гиперпараметров модели и на тестирование.

AutoML состоит из трех шагов:

- предобработка данных;
- предобработка признаков;
- обучение и тестирование модели

Рисунок 157 Принцип работы узла "AutoML"

В результате запуска узла AutoML Система автоматически перебирает различные методы балансировки, предобработки и моделирования и объединяет модели в ансамбли.

При необходимости пользователь может посмотреть, какие модели выбрала система и какие модели тестировались.

На рисунке ниже представлен пример результата AutoML.

Рисунок 158 Пример результата узла AutoML

В представленном примере итоговая модель состоит из 5 моделей с различными предобработками и балансировками. Результаты каждой модели имеют вес 0,2.

$$Model = 0{,}2 \cdot f_1 + 0{,}2 \cdot f_2 + 0{,}2 \cdot f_3 + 0{,}2 \cdot f_4 + 0{,}2 \cdot f_5$$

Рисунок 159 Пример итоговой модели

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Время на построение всех моделей, с	По умолчанию - 300	Количество времени в секундах, которым ограничено выполнение данного узла
Время на построение одной модели, с	По умолчанию - 30	Количество времени в секундах, которым ограничено время построения одной модели
Количество конфигураций подбора гиперпараметров	По умолчанию - 25	Количество конфигураций подбора гиперпараметров для ускорения подбора с помощью алгоритма SMAC

Параметр	Возможные значения и ограничения	Описание
Размер ансамблевых моделей	По умолчанию - 5	Максимальное количество моделей в ансамбле
Количество лучших моделей в ансамбле	По умолчанию - 50	Количество лучших моделей, участвующих в отборе в ансамбль
Максимальное количество моделей на диске	По умолчанию - 50	Максимальное количество моделей, которое будет построено в рамках временных ограничений
Seed	По умолчанию - 42	Начальное числовое значение для генератора случайных чисел
Лимит памяти модели	По умолчанию - 3072	Лимит памяти для обучения модели
Типы предобработки переменных	Раскрывающийся список со следующими значениями: • densifier • extra_trees_preproc • fast_ica • feature_agglomeration • kernel_pca • kitchen_sinks • no_preprocessing • nystroem_sampler • pca • polynomial • random_trees_embedding • select_percentile • select_rates • truncatedSVD	Набор методов предобработки переменных, который будет использован при переборе вариантов моделей. Предусмотрены: • densifier (Перевод данных из разреженного представления в плотное) • extra_trees_preproc (Отбор признаков с помощью ансамбля сильно рандомизированных деревьев) • fast_ica (Быстрый метод независимых компонент) • feature_agglomeration (Кластеризация признаков) • kernel_pca (Ядерный метод главных компонент) • kitchen_sinks (Аппроксимация ядерной функции методом случайных признаков Фурье (метод кухонных раковин)) • no_preprocessing (Без предобработки) • nystroem_sampler (Аппроксимация ядерной функции методом Nyström’a) • pca (Метод главных компонент) • polynomial (Полиноминальные предикторы) • random_trees_embedding (Бинарное кодирование сильно рандомизированными деревьями)

Параметр	Возможные значения и ограничения	Описание
		• select_percentile (Отбор признаков по перцентилю) • select_rates (Отбор признаков по частотам ошибок) • truncatedSVD (Усечённое сингулярное разложение)
Типы используемых моделей	Раскрывающийся список со следующими значениями: • adaboost • ard_regression • bernoulli_nb • decision_tree • extra_trees • gaussian_nb • gausssian_process • gradient_boosting • k_nearest_neigbors • Ida • liblinear_svm • libsvm_svc • mlp • multinomial_nb • passive_aggressive • qda • random_forest • sgd	Набор алгоритмов, который будет использован при построении вариантов моделей. Предусмотрены: • adaboost (Адаптивный бустинг (AdaBoost)) • ard_regression (Байесовская линейная регрессия с автоматическим определением актуальности (ARD Regression)) • bernoulli_nb (Наивный байесовский классификатор с распределением Бернулли) • decision_tree (Дерево решений) • extra_trees (Ансамбль сильно рандомизированных деревьев решений) • gaussian_nb (Наивный байесовский классификатор с нормальным распределением) • gausssian_process (Гауссовский процесс) • gradient_boosting (Градиентный бустинг) • k_nearest_neigbors (Метод K-ближайших соседей) • Ida (Линейный дискриминантный анализ) • liblinear_svm (Линейная машина опорных векторов) • libsvm_svc (Машина опорных векторов) • mlp (Многослойный перцептрон) • multinomial_nb (Наивный байесовский классификатор с полиномиальным распределением)

Параметр	Возможные значения и ограничения	Описание
		• passive_aggressive (Пассивно-агрессивный алгоритм) • qda (Квадратичный дискриминантный анализ) • random_forest (Случайный лес) • sgd (Линейная модель, обучаемая методом стохастического градиентного спуска)
Тип ресемплинга	Раскрывающийся список со следующими значениями: • holdout • cv	Тип ресемплинга выборки. Предусмотрены: • holdout (Отложенная выборка) • cv (Кросс-валидация)
% обучающей выборки для AutoML	По умолчанию - 0,67	% Обучающей выборки, который будет использован для AutoML
Перемешать наблюдения	Чекбокс	При выборе этой опции наблюдения будут перемешаны заново
Количество параллельных потоков	По умолчанию - 0	Параллельное исполнение
Метрика	Раскрывающийся список со следующими значениями: • AUC ROC • Logloss • Accuracy • Balanced accuracy • F1 • Logloss • MSE • MAE • R2	Метрика, которую оптимизирует AutoML. Прдусмотрены: Для задачи бинарной классификации: • AUC ROC • Logloss • Accuracy • Balanced accuracy • F1 Для задачи многоклассовой классификации: • Logloss Для задачи регрессии: • MSE • MAE • R2
Коэффициент сжатия датасета	По умолчанию - 1	Сжатие датасета для оптимизации памяти. Сжатие выполняется за счёт сокращения точности чисел с плавающей точкой.

Таблица 42 Параметры узла «AutoML»

3.2.5.9. Группа узлов «Работа с моделями»

3.2.5.9.1. Узел «Сравнение моделей»

Узел «Сравнение моделей» оценивает построенные модели по метрикам качества и выбирает лучшую.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Метрика для сравнения. Регрессия	Раскрывающийся список со следующими значениями: • MSE • RMSE • MAE • MAPE • R2	Данный параметр задает метрику, по которой будет производиться сравнение регрессионных моделей
Метрика для сравнения. Классификация	Раскрывающийся список со следующими значениями: • log loss • Gini • AUC ROC	Данный параметр задает метрику, по которой будет производиться сравнение моделей классификации
Выборка для сравнения	Раскрывающийся список со следующими значениями: • Обучающая • Валидационная • Тестовая	Данный параметр задает выборку для сравнения

Таблица 43 Параметры узла «Сравнение моделей»

Результаты выполнения узла:

- Таблица отранжированных по выбранной метрике моделей с указанием модели-победителя.

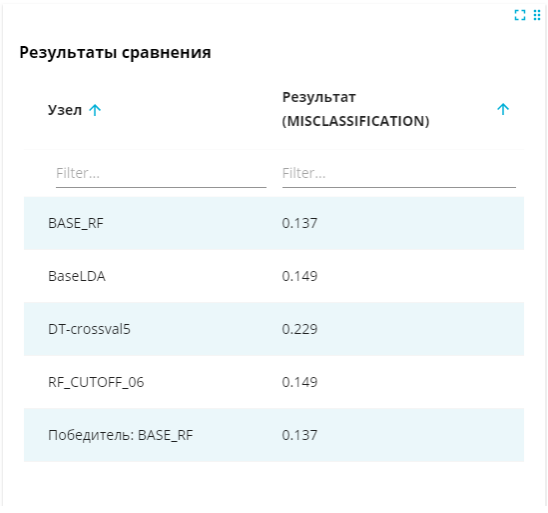


Рисунок 160 Пример таблицы с результатами сравнения моделей

3.2.5.9.2. Узел «Регистрация модели»

Узел «Регистрация модели» позволяет сохранить построенную модель в выбранном проекте репозитория Model Manager.

> **ВАЖНО!** Перед тем как регистрировать модель из MD в Репозиторий (при помощи узла «Регистрация модели») необходимо создать **Проект ММ**.

Является конечным узлом сценария моделирования.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Имя модели	Ручной ввод Ограничений на значение нет	Наименование модели, которое будет отображаться в репозитории
Проект ММ	Раскрывающийся список с доступными проектами	Выбор проекта ММ, в котором будет сохранена модель

Таблица 44 Параметры узла «Регистрация модели»

Результаты выполнения узла:

- Графические и табличные представления не предусмотрены.

3.2.5.9.3. Узел «Интерпретация»

Методы **узла «Интерпретация»** позволяют объяснить, как отдельные признаки и элементы модели влияют на целевую переменную. Рекомендуется устанавливать узел **«Фильтр»** перед узлом **«Интерпретация»** для отбора наблюдений, которые необходимо интерпретировать. Большое количество наблюдений значительно увеличивает время расчета.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Тип интерпретации	Раскрывающийся список со следующими значениями: • LIME • PD • SHAP • ICE	Данный параметр задет тип интерпретации. Предусмотрены следующие методы: • LIME – Локальные интерпретируемые объяснения модели • PD – Частичная зависимость • SHAP – Аддитивные объяснения Шэпли • ICE – Индивидуальные условные ожидания

Таблица 45 Параметры узла «Интерпретация»

Каждый из указанных методов интерпретации имеет свои параметры и результаты.

- **Метод LIME** (локально интерпретируемое объяснение модели).

Данный метод строит модель линейной регрессии, чтобы аппроксимировать предсказания исходной неинтерпретируемой модели локально, а не глобально.

Для этого создается новый набор данных из наблюдений, которые находятся вокруг выбранного для интерпретации наблюдения. Затем этот новый набор данных используется для обучения интерпретируемой линейной модели.

Коэффициенты этой линейной регрессии позволяют оценить важность и направление влияния каждого из предикторов при построении прогноза для выбранного наблюдения.

Список параметров метода LIME представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Количество признаков	Ручной ввод целочисленного значения По умолчанию — 5	Данный параметр задает количество признаков
Размер выборки	Ручной ввод целочисленного значения По умолчанию — 5000	Данный параметр задает размер выборки
Количество объясняемых классов	Ручной ввод целочисленного значения По умолчанию — 0	Данный параметр задает количество объясняемых классов
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел

Таблица 46 Параметры метода LIME

Результаты выполнения метода:
В окне с результатами во вкладке **Выбор наблюдений** необходимо указать наблюдение для интерпретации. Во вкладках **Графические результаты** и **Табличные результаты** будут отображены список объяснений, отражающих вклад каждой функции в прогноз выборки данных в виде графиков и таблицы соответственно (рисунок ниже). Это обеспечивает локальную интерпретируемость, а также позволяет определить, какие изменения характеристик окажут наибольшее влияние на прогноз.

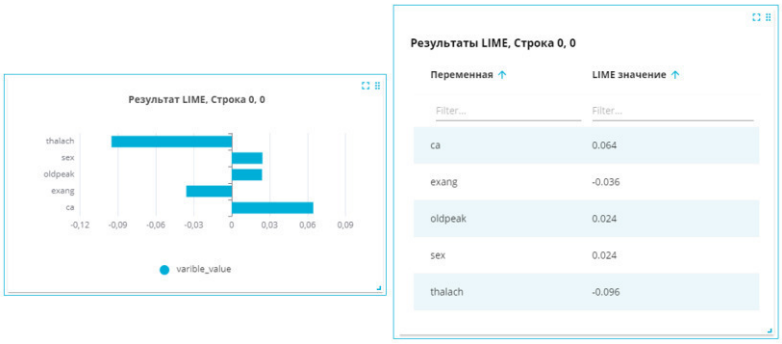


Рисунок 161 Пример графических и табличных результатов метода LIME

- **Метод PD.**

График частичной зависимости (Partial Dependence Plots – PDP) показывает, как меняется средний прогноз при изменении одного из предикторов.

Алгоритм работы метода PD: выбирается переменная и непрерывно изменяется ее значение. На график наносятся эти значения переменной и соответствующее среднее значение прогноза по выборке. Таким образом получается график зависимости прогнозируемых результатов от значений переменной.

Данный график частичной зависимости может показать, является ли отношение между целью и признаком линейным, монотонным или более сложным. Например, при применении к модели линейной регрессии графики частичной зависимости всегда показывают линейную зависимость.

Список параметров метода PD представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Размер выборки	Ручной ввод целочисленного значения По умолчанию — 5000	Данный параметр задает размер выборки
Количество бинов	Ручной ввод целочисленного значения По умолчанию — 5	Данный параметр задает количество бинов
Квантили	Таблица со значениями левого и правого квантилей	Для редактирования значений квантилей необходимо выбрать ссылку **Редактировать**. В открывшемся окне задать новые значения квантилей.
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел
Полный биннинг	Чекбокс	Выбор данного чекбокса указывает на необходимость полного биннинга

Таблица 47 Параметры метода PD

Результаты выполнения метода:
В окне с результатами во вкладке **Графические результаты** выбрать предиктор и класс, чтобы график отразил связь между предиктором и прогнозом.

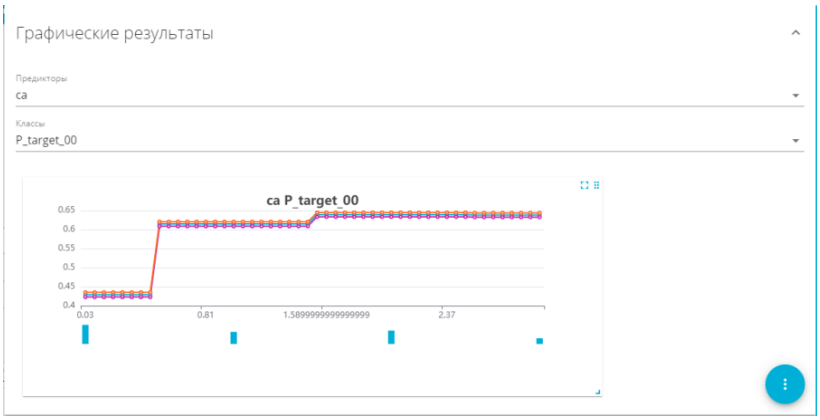


Рисунок 162 Пример графических результатов метода PD

- **Метод SHAP (Shapley Additive Explanation).**

SHAP value показывает средний вклад каждого предиктора в формирование прогноза для наблюдения, выбранного для интерпретации. Усреднение делается по всевозможным комбинациям всех остальных предикторов.

Алгоритм расчета SHAP values: Для каждого возможного упорядочивания признаков берутся все признаки, стоящие перед i-м признаком, и считается величина (прирост эффективности от добавления признака i в комбинацию признаков), равная разности между фактическим прогнозом и прогнозом с учетом текущего набора значений признака. После чего полученные значения усредняются по всем упорядочиваниям. Это означает, что SHAP values описывают ожидаемый прирост выходного значения модели при добавлении i-го признака в текущем примере.

Список параметров метода SHAP представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Размер выборки	Ручной ввод целочисленного значения По умолчанию — 100	Данный параметр задает размер выборки
Количество перестановок	Ручной ввод целочисленного значения По умолчанию — 10	Данный параметр задает количество перестановок

Таблица 48 Параметры метода SHAP

Результаты выполнения метода:
В окне с результатами во вкладке **Выбор наблюдений** необходимо указать наблюдение для интерпретации. Во вкладках **Графические результаты** и **Табличные результаты** будут отображены список объяснений, отражающих вклад каждой функции в прогноз выборки данных в виде графиков и таблицы соответственно.

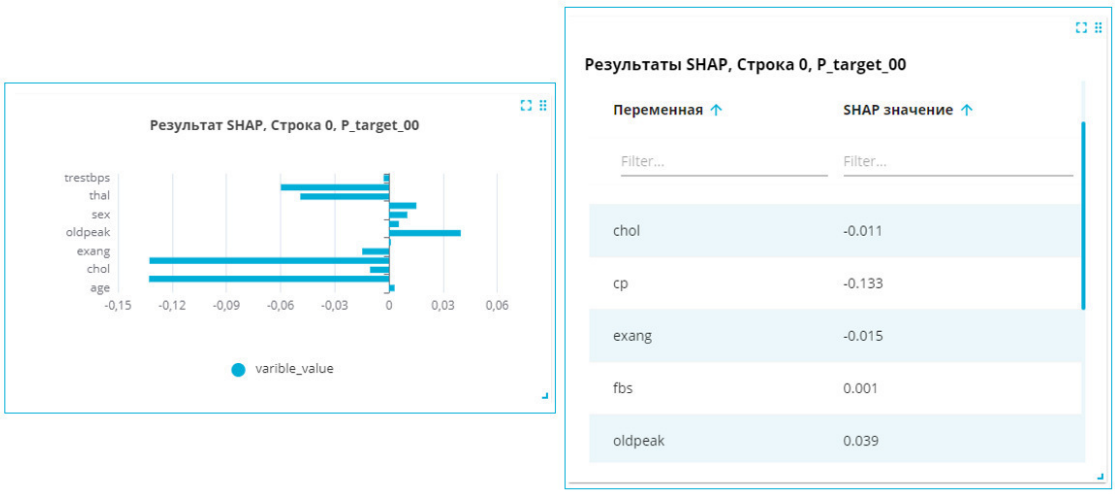


Рисунок 163 Пример графических и табличных результатов метода SHAP

- **Метод ICE — Индивидуальные условные ожидания.**

График индивидуального условного ожидания (ICE) показывает, как меняется прогноз модели на одном или нескольких наблюдениях при изменении одного из предикторов. Однако, в отличие от PDP, который показывает средний эффект входной характеристики, график ICE визуализирует зависимость прогноза от характеристики для каждой выборки отдельно с одной строкой на выборку.

Список параметров метода ICE представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Размер выборки	Ручной ввод целочисленного значения По умолчанию — 5000	Данный параметр задает размер выборки
Количество бинов	Ручной ввод целочисленного значения По умолчанию — 5	Данный параметр задает количество бинов
Квантили	Таблица со значениями левого и правого квантилей	Для редактирования значений квантилей необходимо выбрать ссылку **Редактировать**. В открывшемся окне задать новые значения квантилей.
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел
Полный биннинг	Чекбокс	Выбор данного чекбокса указывает на необходимость полного биннинга

Таблица 49 Параметры метода ICE

Результаты выполнения метода:
В окне с результатами во вкладке **Графические результаты** выбрать предиктор и класс, чтобы график отразил связь между предиктором и значением.

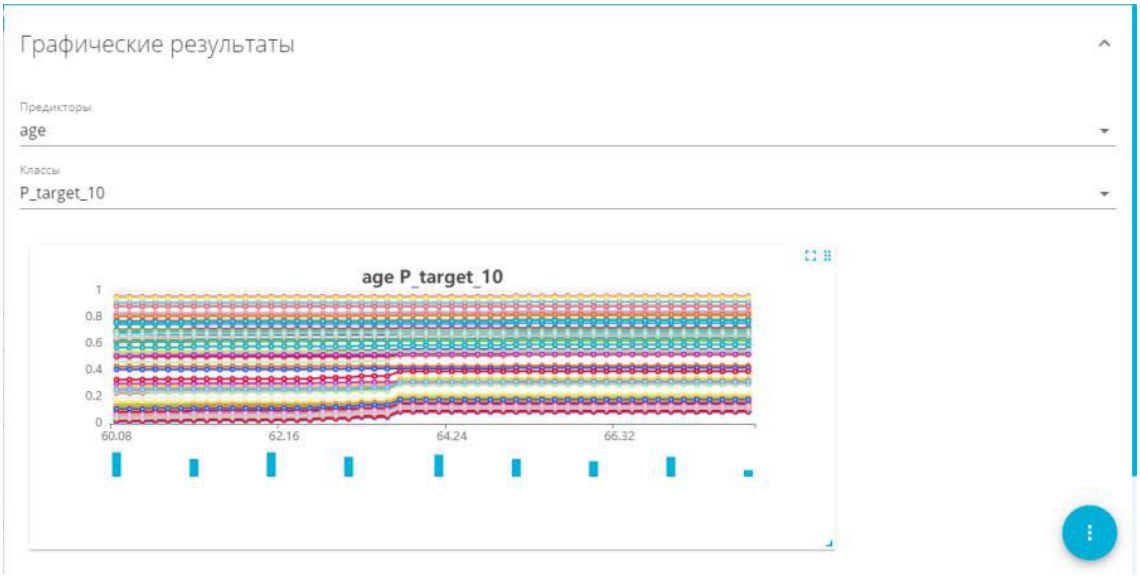


Рисунок 164 Пример графических результатов метода ICE

3.2.5.9.4. Узел «Подбор отсечки (Cut off)»

Узел «Подбор отсечки» позволяет определить оптимальный порог отсечения. Порог отсечения нужен для того, чтобы относить новые примеры к одному из двух классов (задача бинарной классификации).
Список параметров узла представлен в таблице ниже.

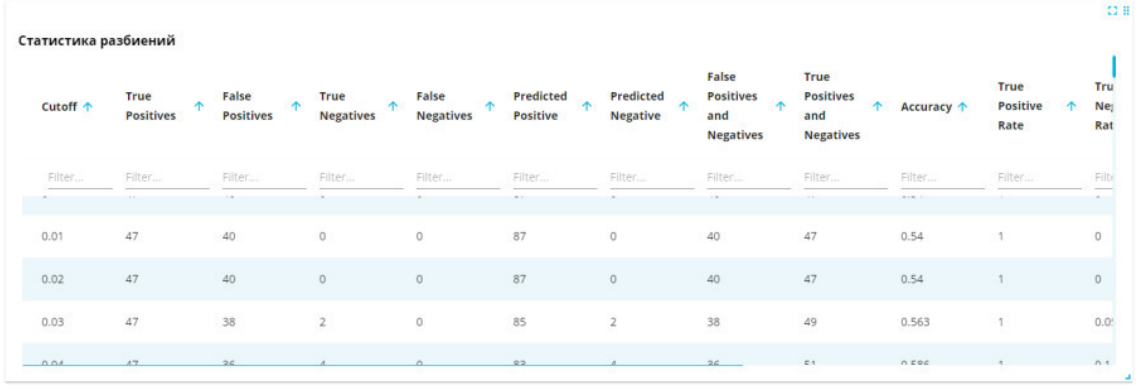
Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Количество разбиений	Ручной ввод целочисленного значения По умолчанию — 100	Данный параметр задает количество разбиений
Размер бина для Lift	Ручной ввод целочисленного значения По умолчанию — 20	Данный параметр задает размер бина для расчета Lift
Критерий отсечки	Раскрывающийся список со следующими значениями: • Число (по умолчанию) • Колмогорова-Смирнова	Данный параметр задает критерий выбора порога отсечения
Отсечка	Ручной ввод Число больше 0 меньше 1 По умолчанию — 0,5	Данный параметр задает порог отсечки

Таблица 50 Параметры узла «Подбор отсечки (Cut off)»

Результаты выполнения узла:

- Таблица со статистикой разбиения, в которой отражены следующие показатели:
  - Cutoff value – значение порога отсечения (считается в соответствии с заданным параметром Количество разбиений).
  - True Positives – количество верно классифицированных положительных примеров.
  - False Positives – количество неверно классифицированных положительных примеров.
  - True Negatives – количество верно классифицированных отрицательных примеров.
  - False Negatives – количество неверно классифицированных отрицательных примеров.
  - Predicted Positive – сумма True Positives и False Positives.
  - Predicted Negative – сумма True Negatives и False Negatives.
  - False Positives and Negatives – сумма False Positives и False Negatives
  - True Positives and Negatives – сумма True Positives и True Negatives

- Accuracy – результат деления суммы True Positive и True Negative на сумму всех значений (доля правильных ответов).
- True Positive Rate (Recall, полнота) – доля найденных объектов класса к общему числу объектов класса. Показывает, насколько хорошо классификатор находит объекты из класса.
- True Negative Rate – доля верно классифицированных отрицательных примеров от общего количества отрицательных примеров.
- False Positive Rate – доля неверных срабатываний классификатора к общему числу объектов за пределами класса. Показывает, насколько часто классификатор ошибается при отнесении того или иного объекта к классу.
- Precision (точность) – показывает долю объектов класса среди объектов, выделенных классификатором.
- F-score – гармоническое среднее между точностью и полнотой.
- KS (Kolmogorov-Smirnov statistic).
- partition_id (если был в выборке).


Статистика разбиений

Cutoff	True Positives	False Positives	True Negatives	False Negatives	Predicted Positive	Predicted Negative	False Positives and Negatives	True Positives and Negatives	Accuracy	True Positive Rate	Tru Ne Rat
0.01	47	40	0	0	87	0	40	47	0.54	1	0
0.02	47	40	0	0	87	0	40	47	0.54	1	0
0.03	47	38	2	0	85	2	38	49	0.563	1	0.0

Рисунок 165 Пример таблицы со статистикой разбиения

- Таблица с метриками.


Метрикии классификации

	misclassification	mcc	nobs
0	0.12	0.757	116
1	0.16	0.707	87
2	0.149	0.704	87

Рисунок 166 Пример таблицы с метриками

3.2.6. Кросс-валидация

Для повышения надежности моделей предусмотрена кросс-валидация.

Кросс-валидация (перекрестная проверка) — метод обучения и оценки модели, при котором исходное множество данных разделяется на несколько блоков и модель обучается на этих блоках.

Для использования кросс-валидации необходимо найти в боковой панели раздел Кросс-валидация и выбрать чекбокс **Использовать кросс-валидацию**. В результате должны появиться следующие параметры в соответствии с таблицей ниже.

Параметр	Возможные значения и ограничения	Описание
Количество разбиений	Ручной ввод Неотрицательное целое число	Задает количество блоков, на которые будет разделено исходное множество данных
Использовать стратификацию	Чекбокс	Выбор данного чекбокса указывает на необходимость использовать стратификацию
Переменные для стратификации	Список переменных набора данных	Данный параметр позволяет выбрать переменные для стратификации.

Таблица 51 Параметры кросс-валидации

3.2.7. Автоподбор гиперпараметров

Подбор параметров — одна из важных задач для построения модели машинного обучения. Изменение параметров модели может принципиально повлиять на ее качество. Перебор этих параметров вручную может занять колоссальное количество времени. Для этого предусмотрен **автоподбор параметров** модели.

Для использования автоподбора параметров необходимо найти в боковой панели раздел **Автоподбор гиперпараметров (общее)** и выбрать чекбокс **Автоподбор параметров**. В результате должны появиться следующие параметры в соответствии с таблицей ниже.

Параметр	Возможные значения и ограничения	Описание
Количество разбиений кросс-валидации	Ручной ввод целочисленного значения По умолчанию — 5	Данный параметр задает количество разбиений кросс-валидации
Использовать стратификацию	Чекбокс	Выбор данного чекбокса указывает на необходимость использовать стратификацию
Переменные для стратификации	Открывается в случае выбора чекбокса **Использовать стратификацию** Список переменных набора данных	Данный параметр позволяет выбрать переменные для стратификации
Количество итераций подбора	Ручной ввод целочисленного значения По умолчанию — 10	Данный параметр задает количество итераций подбора

Параметр	Возможные значения и ограничения	Описание
Метрика для оптимизации классификации	Раскрывающийся список со следующими значениями: • LoglossNominal • LoglossBinary	Данный параметр задает метрику для оптимизации классификации
Метрика для оптимизации регрессии	Раскрывающийся список со следующими значениями: • MSE • MAPE	Данный параметр задает метрику для оптимизации регрессии
Оценка точности автоподбора	Раскрывающийся список со следующими значениями: • Заданная валидационная выборка • Новая валдационная выборка • Кросс-валидация	Данный параметр задает оценку точности автоподбора
Seed	Ручной ввод числового значения По умолчанию — 42	Начальное числовое значение для генератора случайных чисел

Таблица 52 Параметры автоподбора параметров модели

Параметры последующей секции **Автоподбор гиперпараметров** зависит от типа модели. Для них задается диапазон параметров, в котором и производиться автоподбор.
С подробным описанием параметров моделей можно ознакомиться в описании конкретного узла.

3.2.8. Результаты моделирования

Результаты предусмотрены для следующих узлов:

- Разделение выборки.
- Sample.
- Фильтр.
- One-hot encoding.
- Заполнение пропусков.
- Трансформация.
- Биннинг/Энкодинг.
- Дисперсионный анализ.
- Стандартизация.
- Веса классов.
- Кластерный анализ (k-means).
- Иерархическая кластеризация.
- Дерево решений.
- Случайный лес.

- Байесовская регрессия.
- Линейная регрессия.
- Логистическая регрессия.
- Линейные модели.
- Нейронная сеть.
- LDA.
- Градиентный бустинг (XGBoost).
- Градиентный бустинг (LightGBM).
- Градиентный бустинг (CatBoost).
- GLM
- Сравнение моделей.
- Интерпретация.
- Подбор отсечки (Cut off).

Для просмотра результатов выполнения узла необходимо:

- Правой кнопкой мыши нажать по интересующему узлу.
- В открывшемся списке выбрать **«Посмотреть результат».**
- Откроется окно **Результаты выполнения узла** с табличным и/или графическим представлением результатов (наполнение окна зависит от узла, подробнее о результатах можно узнать в разделе с описанием каждого узла).

3.3. Пример базового сценария

В таблице ниже представлен пример базового сценария с разбивкой по шагам.

№	Шаг	Узлы
1	**Подключение набора данных**	В любом из сценариев первоначальным узлом всегда должен быть «**Набор данных**». Он задает набор данных, с которым далее будет производиться моделирование.
2	**Задание метаданных**	Узел **«Метаданные»** позволяет изменить метаданные (например, задать предикторы и целевую переменную). Целевая переменная определяет решаемую задачу: • Классификация (Binary – бинарная классификация, Nominal – многоклассовая классификация) • Регрессия (Interval) • Кластеризация - нет необходимости указывать целевую переменную • Ассоциативные правила (необходимо задать переменные с ролями **Идентификатор** и **Предмет**) Также узел **«Метаданные»** может использоваться на последующих этапах моделирования (например, при подготовке данных), если необходимо править метаданные (например, убрать атрибут из сценария или изменить тип атрибута).

№	Шаг	Узлы
3	**Подготовка данных**	Данный этап включает в себя процессы подготовки данных: • Заполнение пропущенных значений (узел **«Заполнение пропусков»**) • Фильтрация данных (узел **«Фильтр»**) • Создание новых расчетных атрибутов (узел **«Трансформация»**) • Преобразование категориальных переменных в числовые (узел **«One-hot encoding»**) • Бинаризация интервальных переменных и кодирование категориальных (узел **«Биннинг/энкодинг»**) • Дисперсионный анализ • Корректировка неравномерного распределения классов в исходном наборе данных (узел **«Sample»**) • Разделение набора данных на обучающую, валидационную и тестовую выборки (узел **«Разделение выборки»**) • Исследование данных для выяснения статистических характеристик переменных (узел «**Профилирование**») • Преобразование числовых наблюдений с целью приведения их к общей шкале (узел «**Стандартизация**») • Корректировка дисбаланса классов при помощи задания весов (узел «**Веса классов**») • Уменьшение размерности - преобразование большого набора переменных в меньший (узлы «**PCA**» и «**Автоэнкодер (PyTorch)**»)
4	**Построение ML моделей**	Узлы моделирования подразделяются в зависимости от решаемой задачи: • Классификация Узел «**Дерево решений**» Узел «**Случайный лес**» Узел «**Логистическая регрессия**» Узел «**Линейные модели**» Узел «**Нейронная сеть**» Узел «**LDA**» Узел «**Градиентный бустинг (XGBoost)**» Узел «**Градиентный бустинг (LightGBM)**» Узел «**Градиентный бустинг (CatBoost)**» Узел «**Нейронная сеть (PyTorch)**» Узел «**AutoML**» В результате выполнения данных узлов рассчитываются новые переменные, одна из которых – класс наблюдения, остальные – вероятность принадлежности к одному из классов (для каждого класса рассчитывается своя вероятность). • Регрессия Узел «**Дерево решений**» Узел «**Случайный лес**» Узел «**Байесовская регрессия**» Узел «**Линейная регрессия**» Узел «**Линейные модели**» Узел «**Нейронная сеть**» Узел «**Градиентный бустинг (XGBoost)**» Узел «**Градиентный бустинг (LightGBM)**» Узел «**Градиентный бустинг (CatBoost)**» Узел «**GLM**» Узел «**Нейронная сеть (PyTorch)**» Узел «**AutoML**»

№	Шаг	Узлы
		В результате решения задачи регрессии в данных узлах рассчитывается переменная с результирующим значением. • Кластеризация Узел «**Кластерный анализ (k-means)**» Узел «**Иерархическая кластеризация**» В результате решения задачи кластеризации рассчитывается переменная, в которой указывается кластер, к которому относится данное наблюдение. • Ассоциативные правила Узел «**Ассоциативные правила»** • Обнаружение аномалий Узел «**Детекция аномалий**» Каждый узел из группы «Обучение с учителем» имеет также параметры для кросс-валидации (метод оценки модели) и автоподбору гиперпараметров.
5	**Работа с моделями**	После процесса построения моделей и перебора гиперпараметров, идет этап интерпретации и сравнения построенных моделей. Для этого предусмотрены следующие узлы: • Узел «**Сравнение моделей**» – сравнение полученных моделей и выбор лучшей • Узел «**Интерпретация**» включает в себя методы для интерпретации предсказаний модели – PD, LIME, ICE, SHAP • Узел «**Подбор отсечки (Cut off)**» позволяет подобрать отсечку для разделения на классы при бинарной классификации
6	**Регистрация модели**	На конечном этапе модель-победитель можно зарегистрировать в Репозитории ММ (узел «**Регистрация модели**»).

Таблица 53 Пример базового сценария

4. Компонент Разработка решений (Decision Manager, DM)

Разработка решений включает в себя обширный пласт работ. Первоначально необходимо создать переменные решения, указать доступные для цепочки Подключения, а затем настроить последовательность шагов решения, используя различные типы узлов. Перед регистрацией решения необходимо выполнить тестирования и убедиться, что результат соответствует ожиданиям.

Компонент включает в себя:

- Главный экран со списком доступных проектов решений (проектов DM).
- Конструктор цепочек решений.
- Вспомогательные окна настройки вида, создания проекта и т.д.

4.1. Главный экран DM

4.1.1. Интерфейс главного экрана DM

Главный экран компонента **Разработка решений** открывается при выборе иконки  в левом верхнем меню и представляет собой список доступных пользователю **Проектов** в табличном виде.

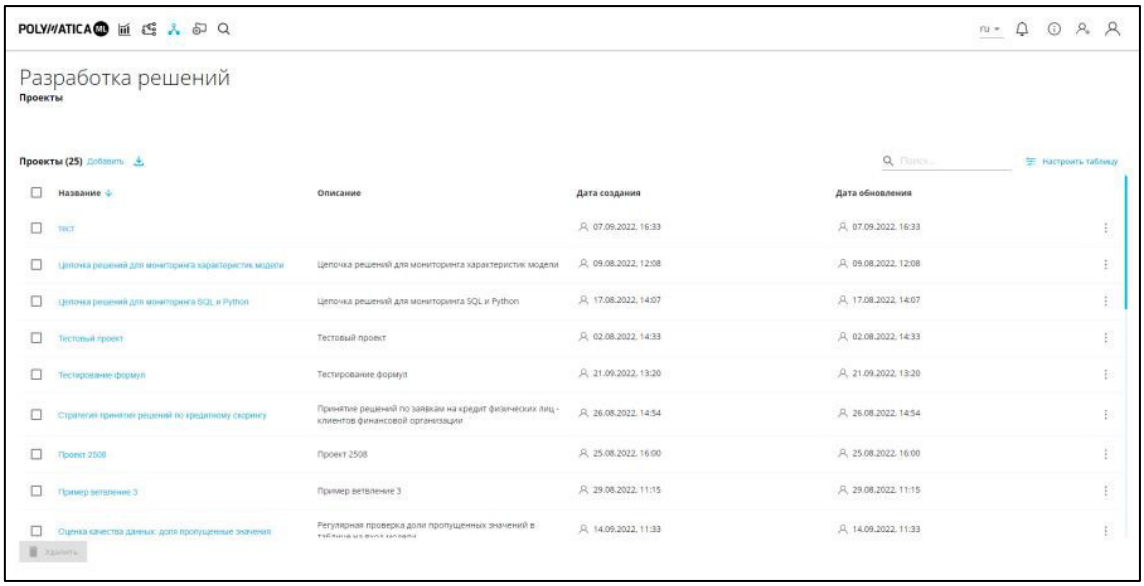


Рисунок 167 Главный экран компонента Разработка решений (Decision Manager)

Проект DM — это набор цепочек решений, настраиваемых пользователем для решения конкретной задачи.
Таблица с доступными проектами имеет гибкие настройки отображения. Так, пользователь может:

- Изменить ширину любого столбца (для этого необходимо перетащить границу его заголовка до нужной ширины).

- Сортировать таблицу (для этого необходимо выбрать иконку рядом с заголовком сортируемого столбца).
- Скрывать/отображать столбцы и изменять их порядок в окне **Вид таблицы** (для открытия окна необходимо выбрать Настроить Таблицу в правом верхнем углу таблицы; при выборе иконки столбец скроется, при наведении на иконку активируется возможность перемещения столбца).
- Сбросить внесенные изменения также в окне **Вид таблицы** (для этого выбрать кнопку **«Сбросить»**).

Для быстрого поиска объекта в таблице предусмотрено поле Поиск... в правой верхней части таблицы.
Объекты таблицы можно выгрузить в формате Excel. Для этого нужно выбрать иконку **Экспорта в excel** .

4.1.2. Создание проекта DM

Для создания нового проекта необходимо выполнить следующие шаги:

- Выбрать кнопку **«Добавить»** в верхней части таблицы с Проектами.
- В открывшемся окне **Создание проекта** задать название и описание.
- Сохранить изменения.

4.1.3. Удаление проекта DM

Для удаления проекта необходимо выполнить следующие шаги:

- Выбрать чекбокс рядом с удаляемым проектом.
- Выбрать кнопку **«Удалить»**.

4.1.4. Копирование проекта DM

Для копирования проекта необходимо выполнить следующие шаги:

- С правой стороны от проекта выбрать иконку в виде трех вертикальных точек и в выпадающем меню выбрать **Копировать.**
- В открывшемся окне **Копирование проекта** задать название и описание нового проекта.
- Сохранить изменения.

4.1.5. Редактирование проекта DM

Для переименования и изменения описания проекта необходимо выполнить следующие шаги:

- С правой стороны от проекта выбрать иконку в виде трех вертикальных точек и в выпадающем меню выбрать **Редактировать.**

- В открывшемся окне **Редактирование проекта** внести необходимые изменения в названии и описании проекта.
- Сохранить изменения.

4.1.6. Работа с проектом DM

Для начала работы с проектом необходимо выбрать его из списка доступных.

4.2. Конструктор цепочек решений

4.2.1. Интерфейс Конструктора цепочек решений

Данный экран открывается при выборе любого из доступных проектов и представляет собой **Конструктор цепочек решений**.

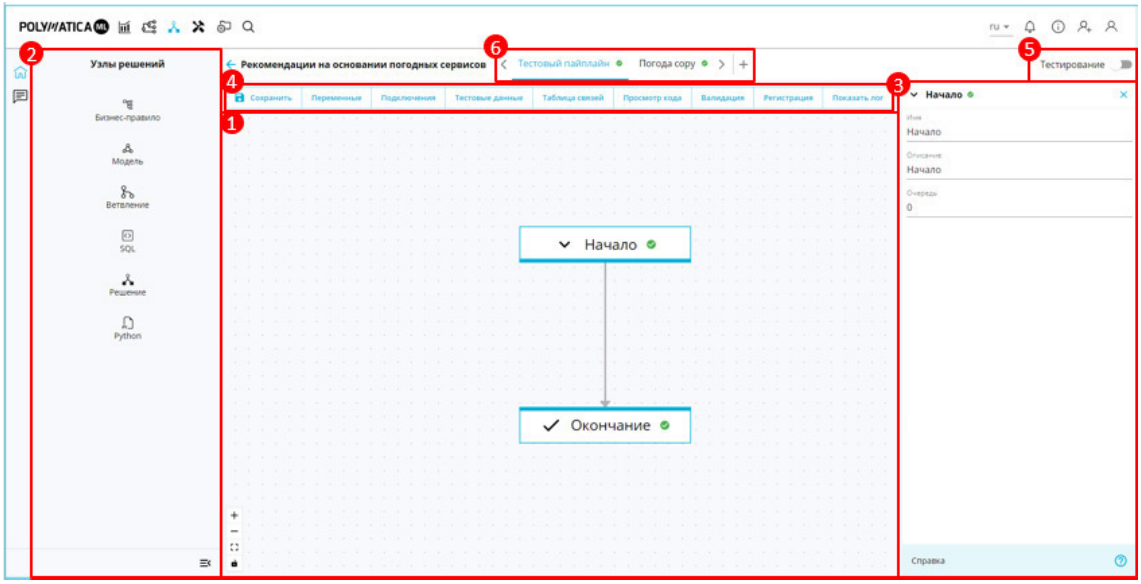


Рисунок 168 Элементы Конструктора цепочек решений

Интерфейс **Конструктора цепочек решений** состоит из следующих элементов:

1. Рабочей области, в которую помещаются узлы решений между исходными узлами **Начало** и **Окончание**.
2. Левой боковой панели с узлами решений.
3. Правой боковой панели с настройками узла (открывается при выборе узла, размещенного в рабочем поле).
4. Верхней панели с кнопками для настройки цепочки решения.
5. Кнопка-переключатель режима Тестирования.
6. Панель создания нескольких сценариев в рамках одного проекта DM.

Для удобства пользования рабочей областью в левом нижнем углу предусмотрены кнопки масштабирования (кнопки

и

для приближения и отдаления, соответственно, кнопка

для масштабирования на объектах, расположенных в рабочем поле) и кнопка, блокирующая перемещение объектов в рабочем пространстве (

).
Левую боковую панель можно скрыть, выбрав кнопку

в нижней части панели. Раскрыть скрытую панель можно выбрав

также в нижней части панели.
Правая боковая панель открывается при выборе узла из рабочего поля. Скрыть эту панель можно выбрав иконку

в правом верхнем углу.
Панель с кнопками включает в себя:

- Кнопку **«Сохранить»**, для сохранения изменений в построенной цепочке решения.
- Кнопку **«Переменные»**, для добавления переменных, которые будут фигурировать в цепочке решений.
- Кнопку **«Подключения»**, для задания и отображения используемых в цепочке решений подключений к БД.
- Кнопку **«Тестовые данные»**, для задания тестовых данных на основе которых будет производиться тестирование цепочки (сценария).
- Кнопку **«Таблица связей»**, для корректирования связей между узлами решений.
- Кнопку **«Просмотр кода»**, для просмотра кода, лежащего за цепочкой решения.
- Кнопку **«Регистрация»**, для регистрации решения в Репозиторий ММ.

4.2.2. Узлы решений

4.2.2.1. Описание узлов решений

4.2.2.1.1. Узел "Начало"

Узел «Начало» является базовым узлом цепочки решения.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения

Таблица 54 Параметры узла "Начало"

4.2.2.1.2. Узел "Бизнес-правило"

Узел "Бинес-правила" используется для реализации бизнес-логики в процессе принятия решения.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения
Правила	Кнопка	При выборе кнопки открывается окно **Правила** для задания условий.

Таблица 55 Параметры узла "Бизнес-правило"

4.2.2.1.3. Узел "Модель"

Узел "Модель" позволяет интегрировать публикацию версии модели из компонента **Управление моделями и решениями (ММ)**.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе

Параметр	Возможные значения и ограничения	Описание
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения
Настройка модели	Кнопка	При выборе кнопки открывается окно **Настройка модели** для выбора необходимой публикации и мэппинга переменных

Таблица 56 Параметры узла "Модель"

4.2.2.1.4. Узел "Ветвление"

Узел "Ветвление" используется для обработки различных сценариев принятия решения.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения
Правила ветвления	Кнопка	При выборе кнопки открывается окно **Правила ветвления** для создания нескольких веток сценария.

Таблица 57 Параметры узла "Ветвление"

4.2.2.1.5. Узел "SQL"

Узел "SQL" используется для написания sql-запросов к базе данных для обогащения данных в процессе принятия решений.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод	Порядок отображения при тестировании решения

Параметр	Возможные значения и ограничения	Описание
	Числовое неотрицательное значение	
Сопоставление	Кнопка	При выборе кнопки открывается окно **Сопоставление** для написания sql-запросов (подробное описание ниже)

Таблица 58 Параметры узла "SQL"

Окно Сопоставление

В окне **Сопоставление** Пользователь может написать sql-запрос.

Основными элементами окна **Сопоставление** являются:

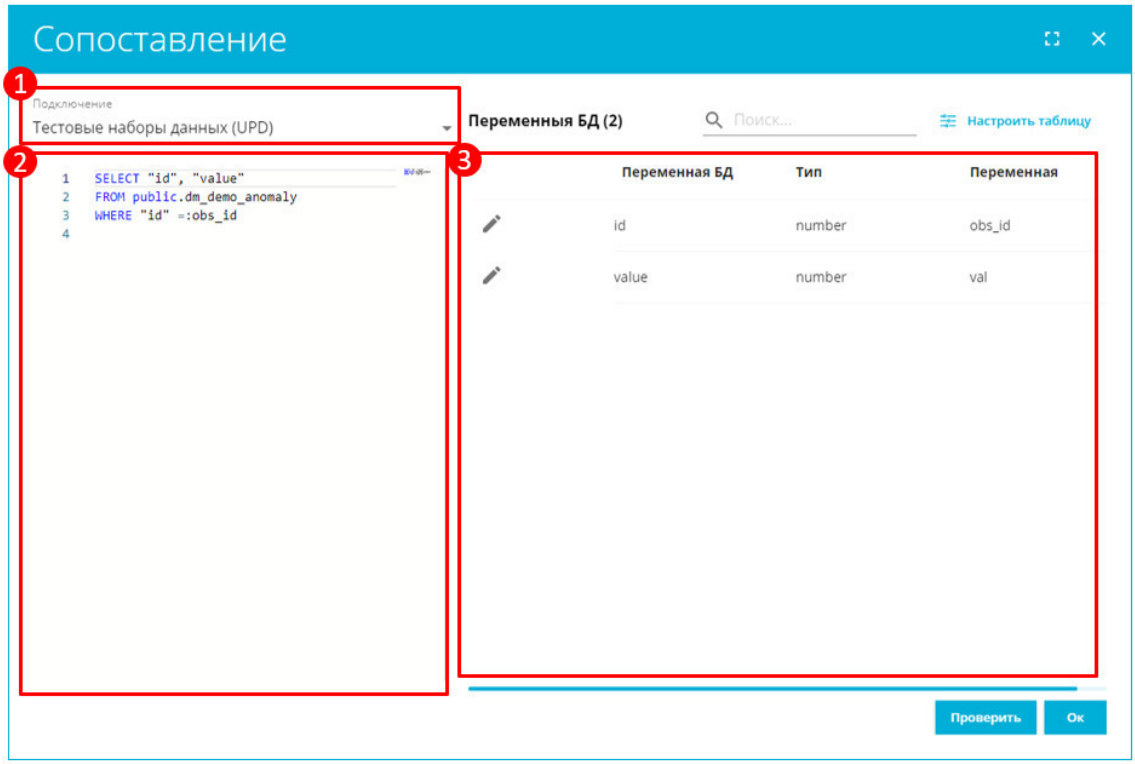


Рисунок 169 Окно Сопоставление

1. Список для выбора необходимого Подключения.
2. Поле ввода sql-запроса.
3. Мэппинг переменных БД и переменных сценария.

Для задания sql-запроса необходимо:

1. Выбрать Подключение из списка доступных для сценария (задается на экране Конструктора сценария по кнопке Подключения).
2. Ввести необходимый запрос в поле ввода и нажать Проверить.
3. Если запрос корректный, то отобразится таблица, в которой необходимо задать мэппинг переменных БД и переменных сценария.
4. Сохранить.

4.2.2.1.6. Узел "Решение"

Узел "Решение" используется для вызова другого решения и интеграции его в эту цепочку принятия решений.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения
Сопоставление	Кнопка	При выборе кнопки открывается окно **Сопоставление** для выбора встраиваемого решения из списка существующих и мэппинга переменных.

Таблица 59 Параметры узла "Решение"

4.2.2.1.7. Узел "Python"

Узел "Python" позволяет интегрировать произвольный python-код для реализации самых сложных нетиповых задач.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения
Код	Кнопка	При выборе кнопки открывается окно **Код** с редактором, куда необходимо добавить произвольный python-код.

Таблица 60 Параметры узла "Python"

4.2.2.1.8. Узел "Окончание"

Узел «Окончание» является базовым узлом цепочки решения.

Список параметров узла представлен в таблице ниже.

Параметр	Возможные значения и ограничения	Описание
Название	Ручной ввод Ограничений на значение нет	Название узла, которое будет отображаться в интерфейсе
Описание	Ручной ввод Ограничений на значение нет	Описание узла
Очередь	Ручной ввод Числовое неотрицательное значение	Порядок отображения при тестировании решения

Таблица 61 Параметры узла "Окончание"

5. Компонент Управление моделями и решениями (Model Manager, MM)

Основная задача, решаемая **компонентом Управление моделями и решениями** — ввод моделей в эксплуатацию и управление их жизненным циклом. В компоненте предусмотрен следующий функционал:

- Репозиторий моделей.
- Импорт сторонних python моделей.
- Публикация моделей.
- Согласование моделей.
- Мониторинг качества работы моделей.

5.1. Интерфейс MM

Раздел **Управление моделями и решениями** открывается при выборе иконки  в левом верхнем меню.

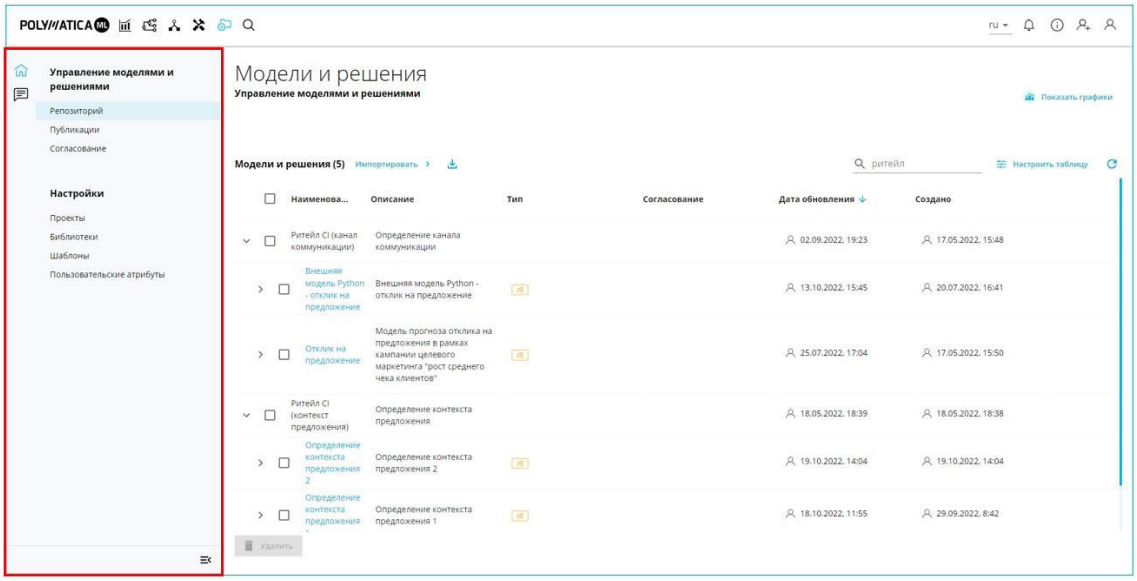


Рисунок 170 Интерфейс компонента Управление моделями

Основным элементом навигации в разделе **Управление моделями и решениями** является боковая панель, которая содержит следующие подразделы:

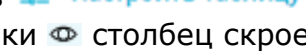
- **Управление моделями и решениями**.
  - **Репозиторий** — представляет собой централизованное хранилище моделей.
  - **Опубликованные модели** — представляет собой список моделей, которые были опубликованы.
  - **Согласование** — включает в себя функционал, необходимый для согласования моделей.

- **Настройки**.
    - **Проекты** — представляет список проектов.
    - **Библиотеки** — представляет собой список Python библиотек, используемых в моделях.
    - **Шаблоны** — представляет собой список доступных для выполнения процессов согласования моделей.
    - **Пользовательские атрибуты** — позволяют пользователю задать любую дополнительную информацию о версии модели.

Для удобства пользования левую боковую панель можно скрыть, выбрав кнопку

в нижней части панели. Раскрыть скрытую панель можно выбрав

также в нижней части панели.
Таблицы в разделе **Управление моделями и решениями** имеют гибкие настройки отображения. Так, пользователь может:

- изменить ширину любого столбца (для этого необходимо перетащить границу его заголовка до нужной ширины).
- сортировать таблицу (для этого необходимо выбрать иконку  рядом с заголовком сортируемого столбца).
- скрывать/отображать столбцы и изменять их порядок в окне **Вид таблицы** (для открытия окна необходимо выбрать  в правом верхнем углу таблицы; при выборе иконки столбец скроется, при наведении на иконку  активируется возможность перемещения столбца).
- сбросить внесенные изменения также в окне **Вид таблицы** (для этого выбрать кнопку **«Сбросить»**).

Для быстрого поиска объекта в таблице предусмотрено поле  в правой верхней части таблицы.
Объекты таблицы можно выгрузить в формате Excel. Для этого нужно выбрать иконку **Экспорта в excel**  .

5.2. Репозиторий

Репозиторий представляет собой централизованное хранилище моделей и их версий и позволяет управлять их жизненным циклом.
Перейти в **Репозиторий** можно, выбрав одноименный раздел боковой панели.

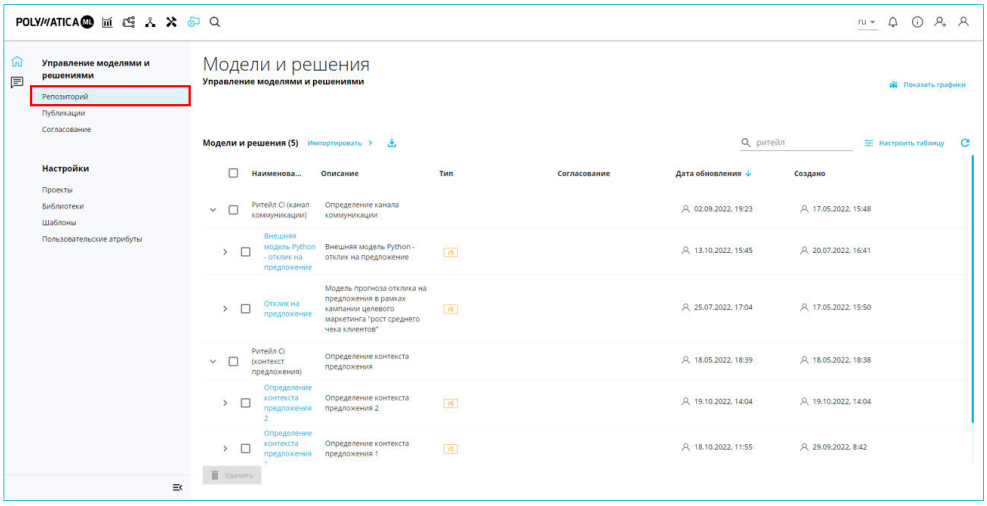


Рисунок 171 Выбор раздела Репозиторий боковой панели

Репозиторий состоит из двух основных элементов:

1. Кольцевые диаграммы, отражающие характеристики моделей в Репозитории (открывается при выборе кнопки Показать графики в верхней части раздела):
    a. Статусы согласования версии модели (возможные значения — Draft (черновик), Approved (согласовано), Rejected (отказано в согласовании), Awaiting (в процессе согласования)).
    b. Используемые алгоритмы.
    c. Библиотеки.
    d. Статусы сборок для публикации (возможные значения — Ready (готово), Error (ошибка), InProgress (в процессе)).
2. Список доступных пользователю моделей в табличном виде.

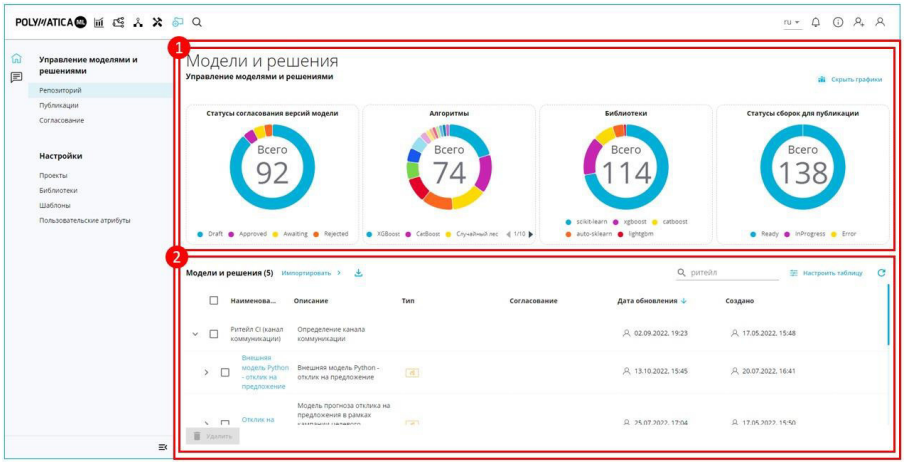


Рисунок 172 Раздел Репозиторий

В таблице отображается иерархическая структура Репозитория в соответствии с рисунком ниже. При нажатии на кнопку › список раскрывается и сворачивается.

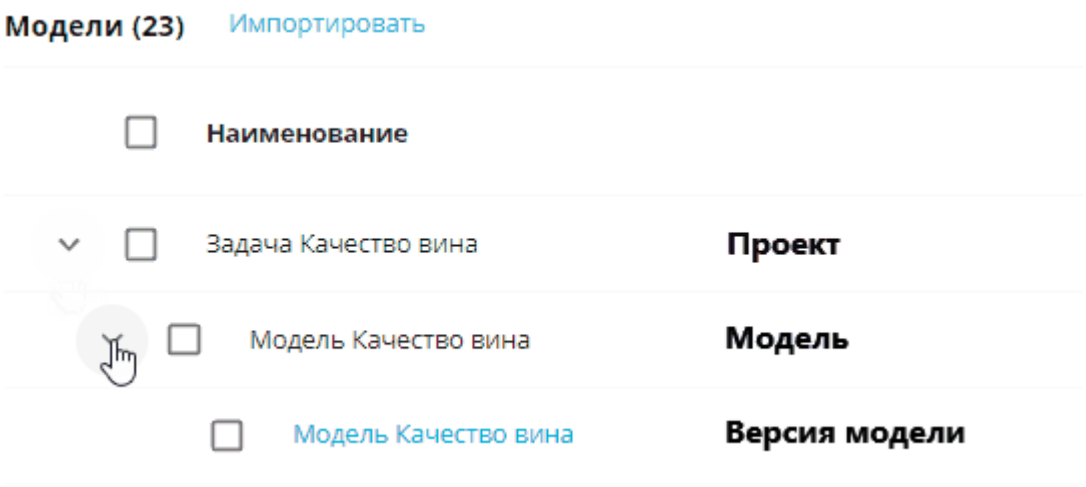


Рисунок 173 Пример иерархии объектов в Репозитории

Проект ММ представляет собой логическое пространство, объединяющее несколько моделей по конкретной бизнес-задаче.
Модель — метаинформация, описывающая программное решение, позволяющее произвести трансформацию или анализ переменных, которые она ожидает на вход (Входные Переменные) и вернуть результаты (Выходные Переменные).
Версия модели — итеративная реализация модели.
В рамках **Проекта** может быть несколько **Моделей** и **Версий модели** (например, первоначальная версия модели и версия модели, полученная после дообучения на новой порции данных).
Для того, чтобы расширить таблицу с моделями и проектами, необходимо убрать панель с графиками, выбрав кнопку 

.

> В Репозитории можно хранить как модели построенные и зарегистрированные из Model Designer и Decision Manager, так и сторонние python модели (и их версии).

5.3. Проект

Для импорта модели в **Репозиторий** необходимо первоначально создать **Проект**.

Проект представляет собой логическое пространство, объединяющее несколько моделей по конкретной бизнес-задаче. Параметры **Проекта** задают минимальные вычислительные ресурсы, с которыми запустится несогласованная модель (количество памяти (в Мб) и ядер), и шаблоны согласования версий модели и их публикаций.

При выборе пункта **Проекты** боковой панели открывается одноименный раздел со списком созданных Проектов ММ.

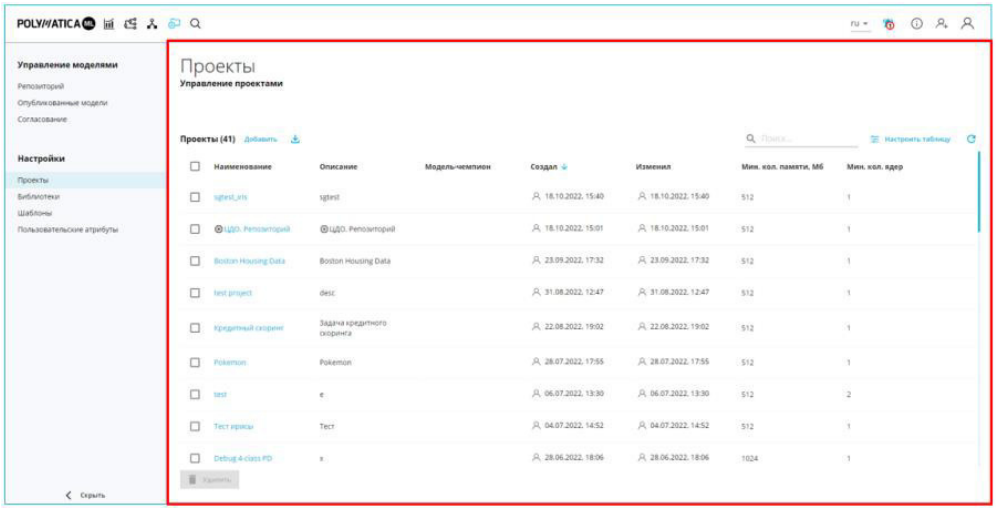


Рисунок 174 Раздел проекты

5.3.1. Создание проекта ММ

Для создания нового Проекта необходимо:

- Выбрать **«Добавить»** в верхней части таблицы раздела **Проекты**.
- В открывшемся окне **Создание проекта** задать:
  - параметры – **Имя проекта**, **Описание**, **Код** (Идентификатор), **Минимальное количество памяти**, **Минимальное количество ядер**;
  - шаблоны **согласования версии модели** и **публикаций** из списка
- Сохранить изменения.

Новый проект отобразится в списке созданных проектов раздела **Проекты**.

> Созданный проект отобразится в **Репозитории** только после импорта в него Модели.

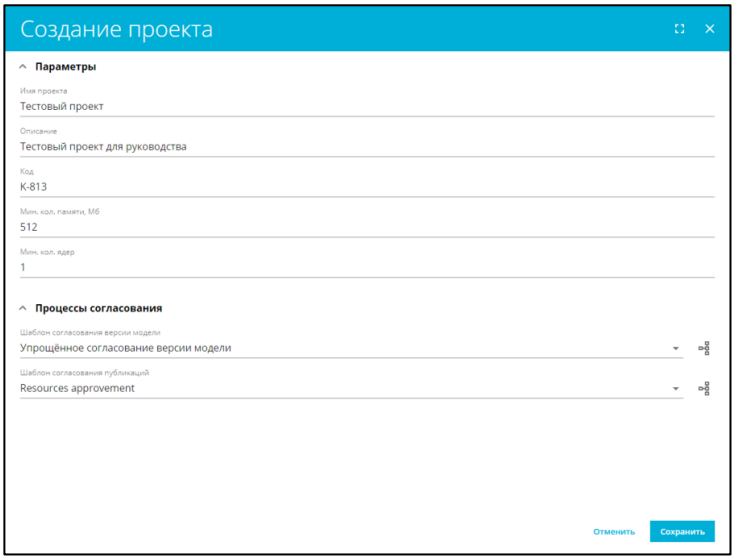


Рисунок 175 Окно Создание проекта

Ознакомиться с выбранным шаблоном согласования можно, выбрав кнопку рядом с полем выбора шаблона, а также в разделе **Шаблоны** боковой панели (подробнее в разделе Шаблон согласования).

5.3.2. Удаление проекта ММ

Для удаления Проекта необходимо:

- Выбрать чекбокс рядом с удаляемым проектом.
- Выбрать кнопку **«Удалить»**.

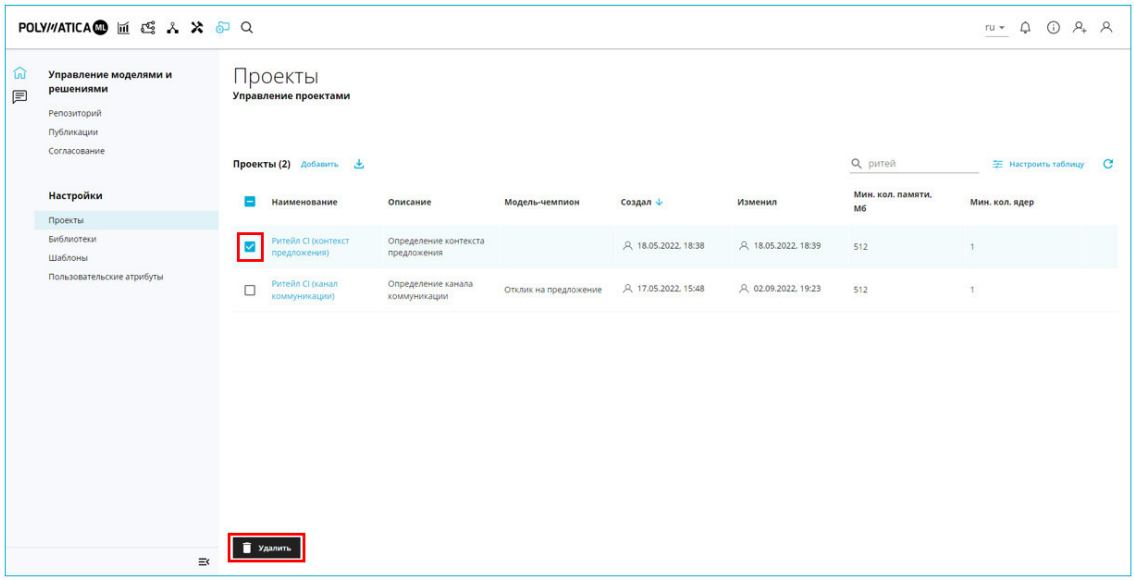


Рисунок 176 Удаление проекта ММ

> В случае, если в Проекте имеются Модели, Пользователь не сможет его удалить. Для удаления Проекта сначала необходимо удалить Модели.

5.3.3. Редактирование проекта ММ

Параметры ранее созданного Проекта можно редактировать. Помимо изменения шаблонов согласования и минимальных вычислительных ресурсов, можно задать Модель-чемпион из соответствующего списка с моделями Проекта.
Для редактирования параметров проекта необходимо:

- Выбрать интересующий проект из списка.
- В открывшемся окне **Обновление проекта** внести необходимые изменения.
- Сохранить изменения.

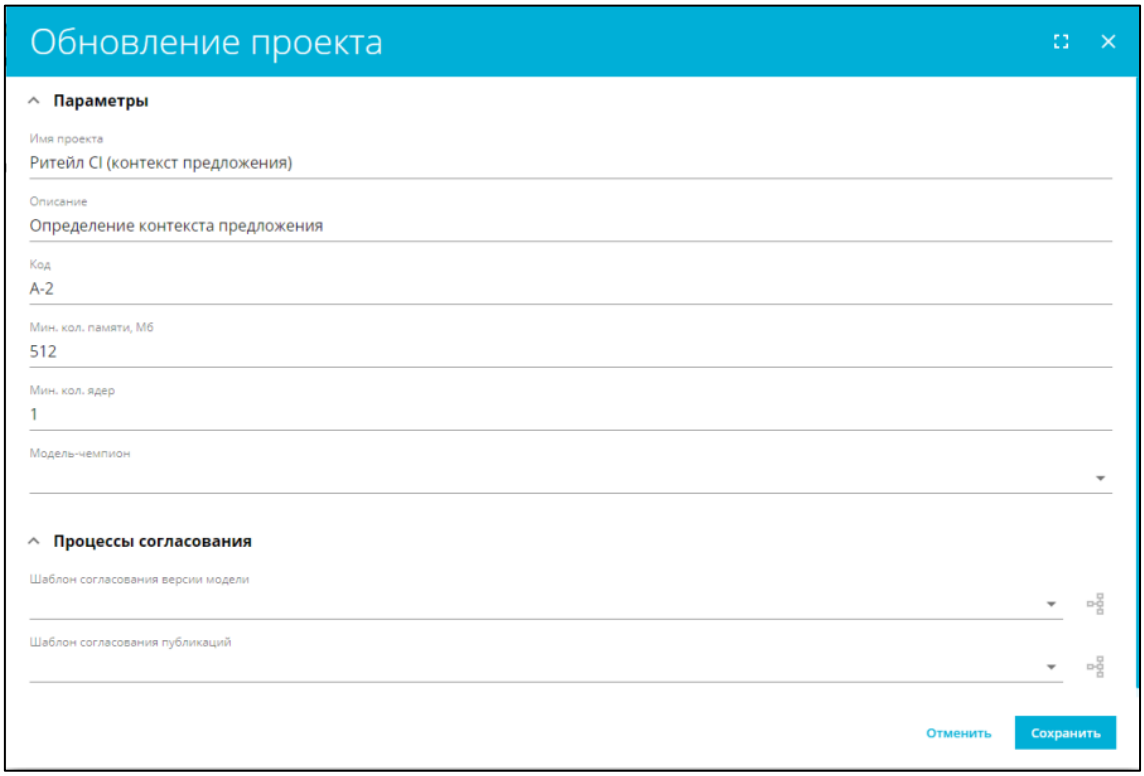


Рисунок 177 Окно Обновление проекта

5.4. Модель

Модель — метаинформация, описывающая программное решение, позволяющее произвести трансформацию или анализ переменных, которые она ожидает на вход (Входные Переменные) и вернуть результаты (Выходные Переменные).

5.4.1. Импорт модели

Зарегистрировать модель в Репозиторий можно из компонентов Построение моделей (MD) и Управление моделями и решениями (DM), а также импортировать извне.

> **ВАЖНО!** Перед тем как импортировать модель необходимо создать **Проект ММ**.

5.4.1.1. Регистрация модели из MD

Регистрация модели из компонента **Построение моделей** (Model Designer) производится при помощи узла **«Регистрация модели»** (подробнее раздел Узел «Регистрация модели»), при этом **Входные и Выходные переменные** подтягиваются автоматически.

Модель типа MD в Репозитории имеет иконку  в соответствующем столбце.

5.4.1.2. Регистрация модели из DM

Регистрация модели из компонента **Разработка решений** (Decision Manager) производится при помощи кнопки **Регистрация** в Конструкторе решений.

Модель типа DM в Репозитории имеет иконку  в соответствующем столбце.

5.4.1.3. Импорт внешней модели

> Импортировать в Модуль можно лишь модели, реализованные на Python.

Для импорта внешней модели необходимо:

1. В **Репозитории** компонента **Управление моделями и решениями** выбрать кнопку **Импортировать** и в выпадающем меню выбрать пункт **Импорт модели.**

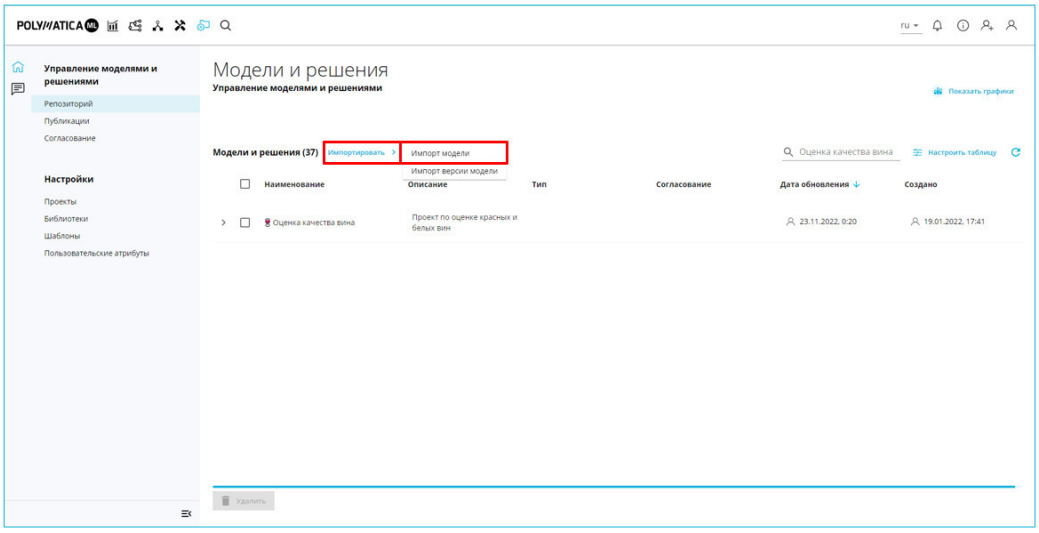


Рисунок 178 Выбор кнопки Импорта модели

2. В открывшемся экране последовательно задать:
   - Параметры модели – Имя и Описание модели, выбрать существующий Проект.
   - Параметры версии проекта – Имя и Описание версии; проект MD; Алгоритм, лежащий в основе и библиотеки, в которых реализован используемый алгоритм.

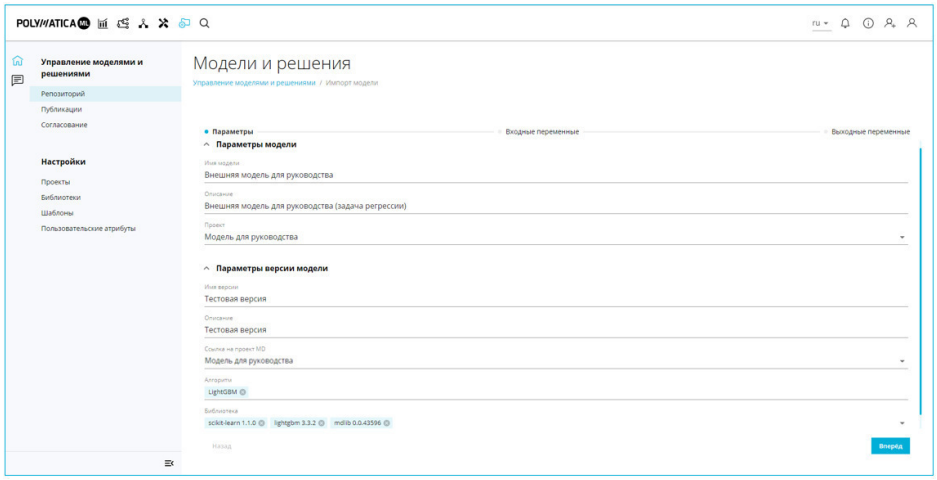


Рисунок 179 Пример экрана импорта с параметрами модели

- Входные и Выходные переменные.

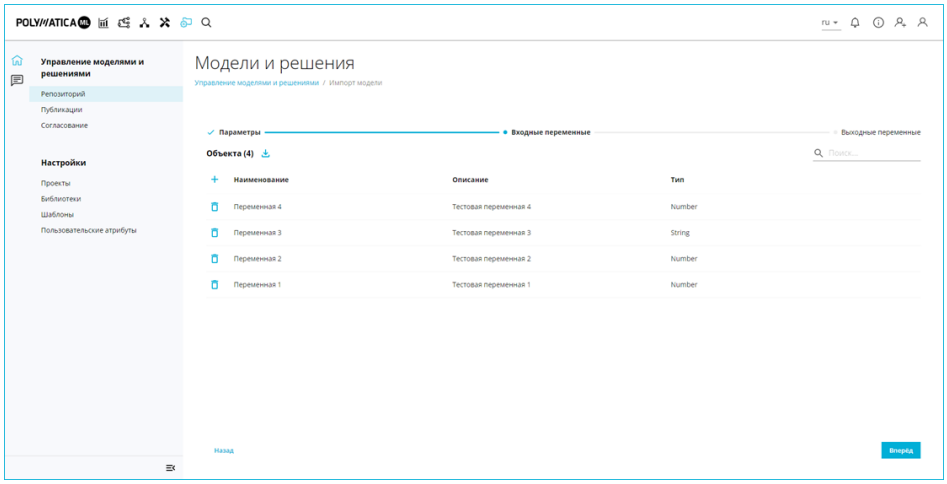


Рисунок 180 Пример экрана с входными переменными

3. Завершить импорт. Модель появится в списке Репозитория в соответствующем Проекте с типом MD.

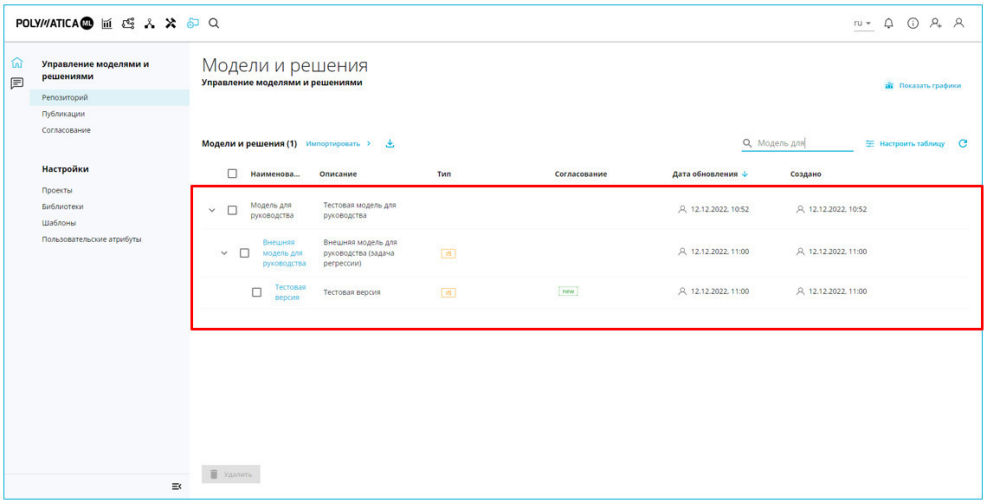


Рисунок 181 Импортированная в репозиторий внешняя модель

4. Выбрать созданную версию модели, раскрыв иерархию Проекта и созданной модели.
5. В открывшемся окне перейти во вкладку **Скоринговый код** и загрузить архив. Архив должен содержать:
    - pkl-файлы модели.
    - файл со скоринговым кодом score.py.
    - файл инициализации модели __init__.py.

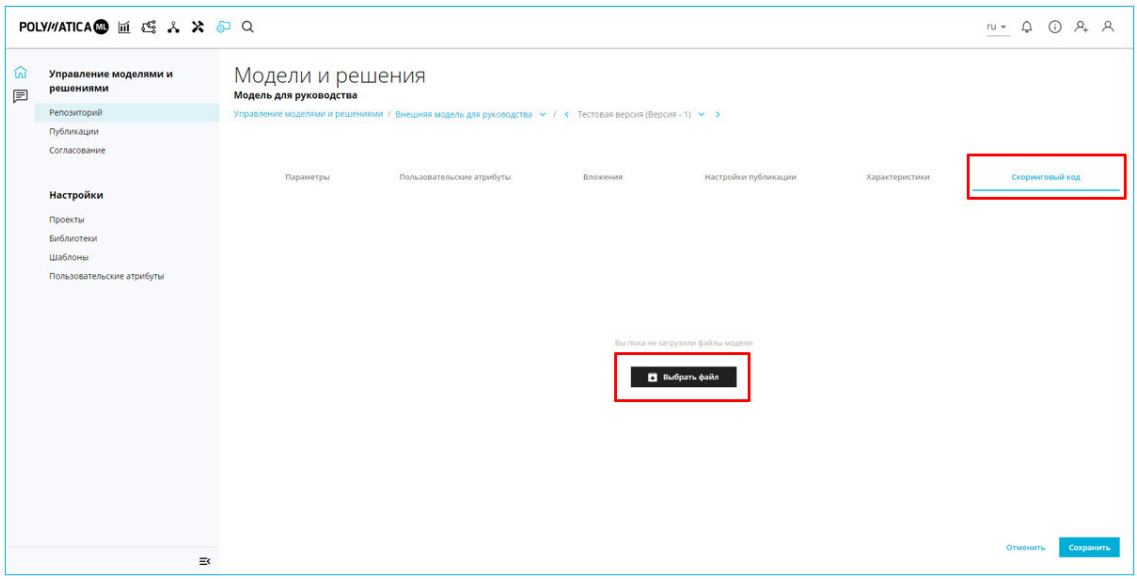


Рисунок 182 Вкладка скоринговый код внешней импортированной модели

5.4.2. Удаление модели

Для удаления Модели необходимо:

- Убедиться в том, что у удаляемой Модели нет Публикаций. Если Публикации есть – первоначально удалить их.
- Выбрать чекбокс рядом с удаляемой Моделью.
- Выбрать кнопку **«Удалить»**.

Выбранная Модель совместно с ее Версиями пропадут из списка Репозитория.

ВАЖНО! Вместе с Моделью удаляются и ее версии.

5.4.3. Работа с моделью

Для редактирования и работы с моделью необходимо выбрать ее из списка Репозитория.

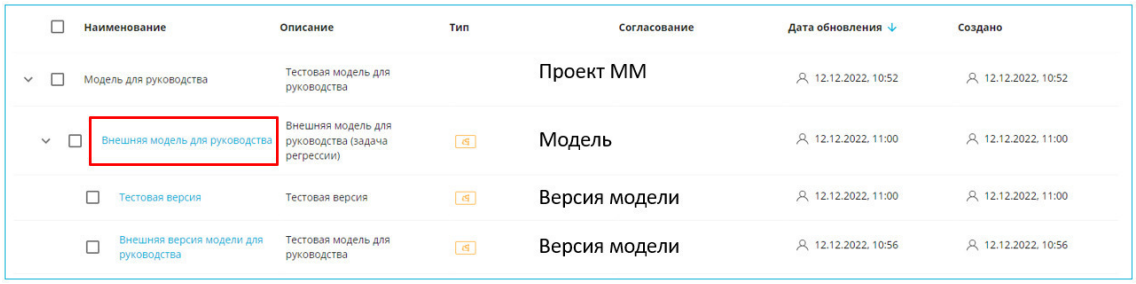


Рисунок 183 Пример выбора модели

Открывшийся экран представляет собой набор вкладок, которые позволяют настраивать Модель.

> Настройки разных типов моделей (MD или DM) несколько отличаются - у **моделей DM** отсутствует вкладка **Дополнительные функции**.

Первая вкладка **Параметры** отображает тип (MD или DM), имя и описание Модели, а также имеющиеся Версии модели.

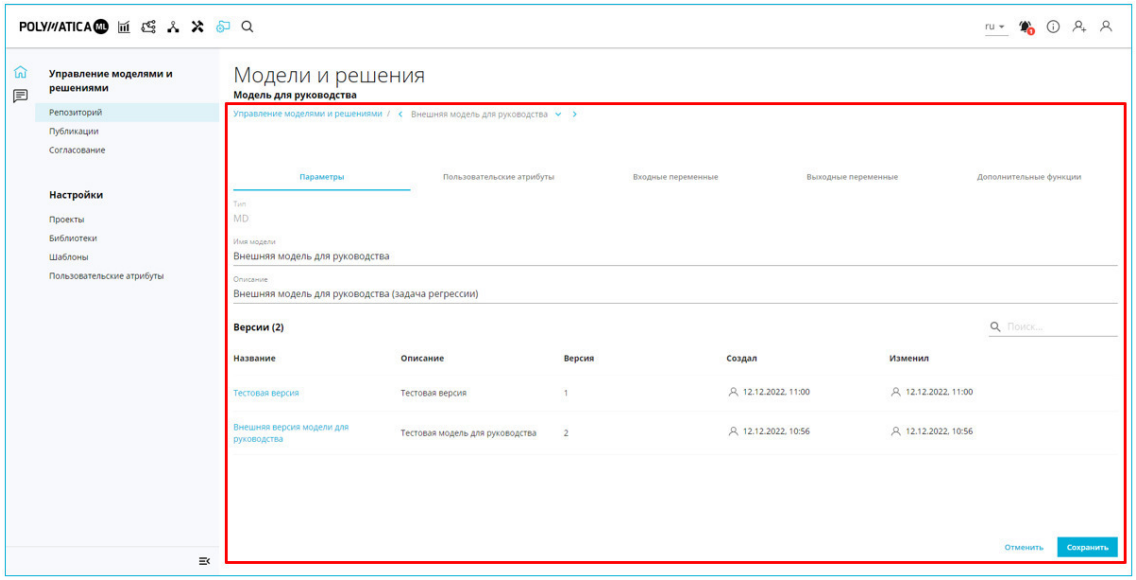


Рисунок 184 Пример экрана с настройками Модели типа MD

Вкладка **Пользовательские атрибуты** позволяет указать список атрибутов для версий текущей модели. Добавить необходимые атрибуты для отображения в данной вкладке можно в разделе **Пользовательские атрибуты** бокового экрана MD (подробнее в разделе Пользовательские атрибуты).

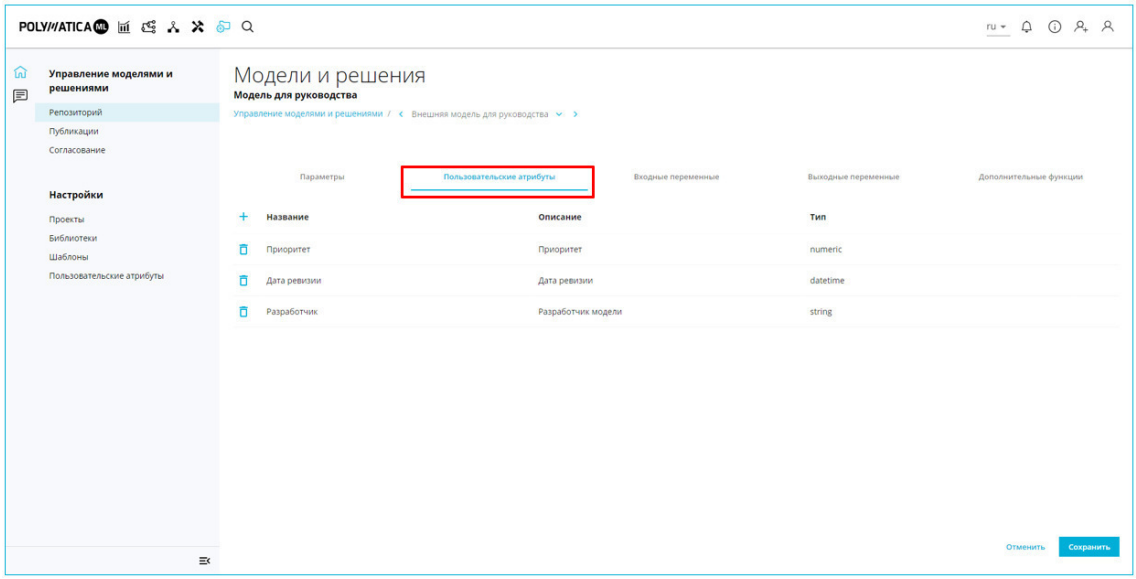


Рисунок 185 Пример вкладки Пользовательские атрибуты

Вкладки **Входные** и **Выходные переменные** имеют одинаковый интерфейс, в котором отображаются автоматически подтянутые из MD или DM входные/выходные переменные или заданные при импорте сторонней модели. Они будут одинаковы для всех версий текущей модели.

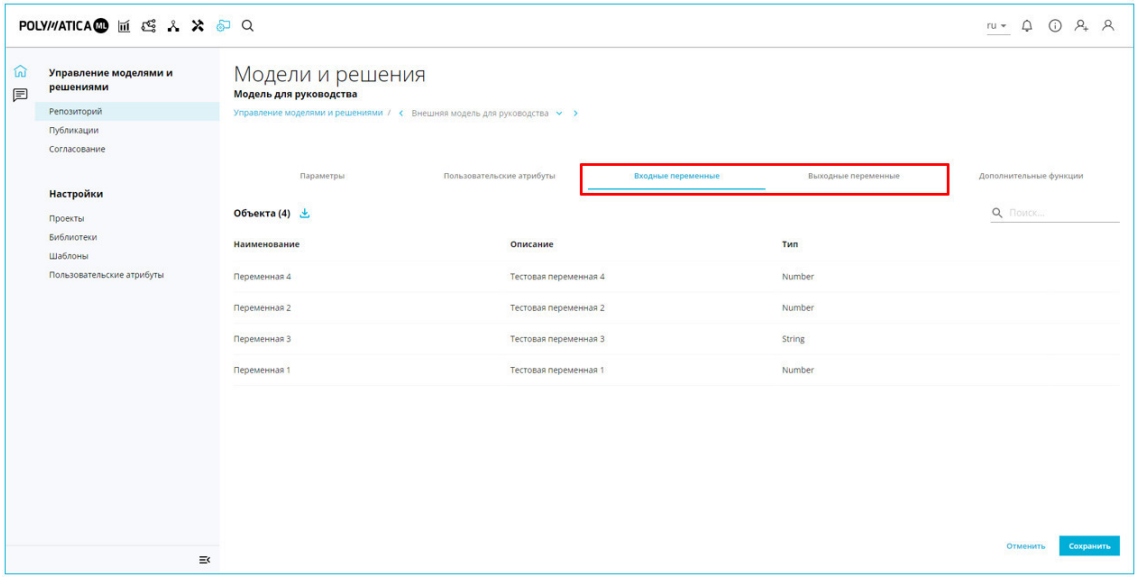


Рисунок 186 Пример вкладки Входные переменные

Последняя вкладка **Дополнительные функции** позволяет добавить pre- и post-score функции к исходному score коду версий модели MD.

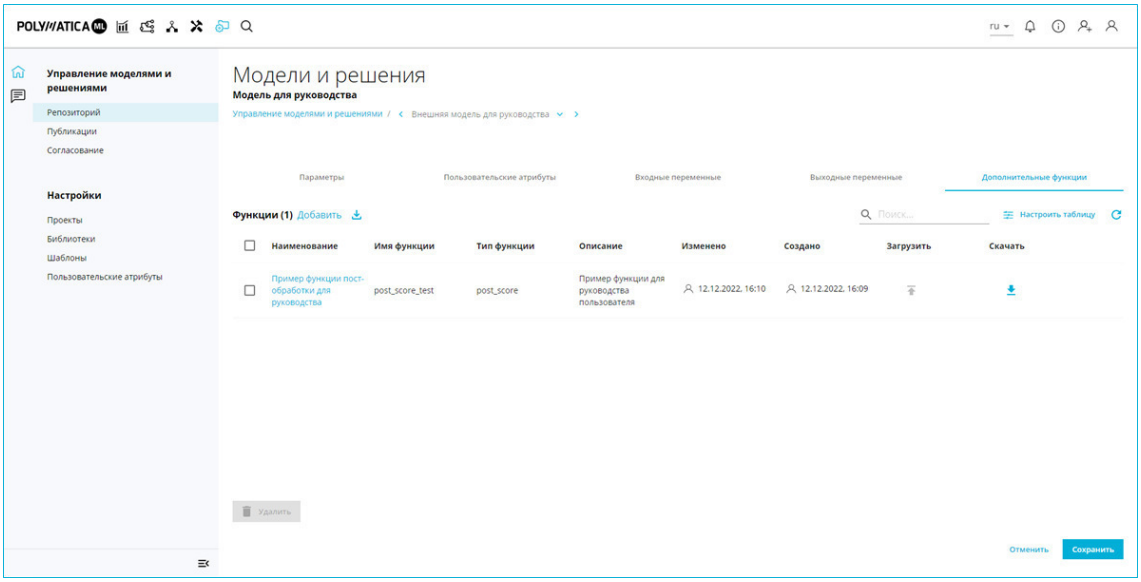


Рисунок 187 Пример вкладки Дополнительные функции

Для добавления функции необходимо:

- Выбрать кнопку **Добавить** (ниже панели с вкладками).
- В открывшемся окне **Создание функции** задать Наименование и Описание функции, выбрать ее Тип (pre-score или post-score), а также задать ее имя. Добавить изменения.

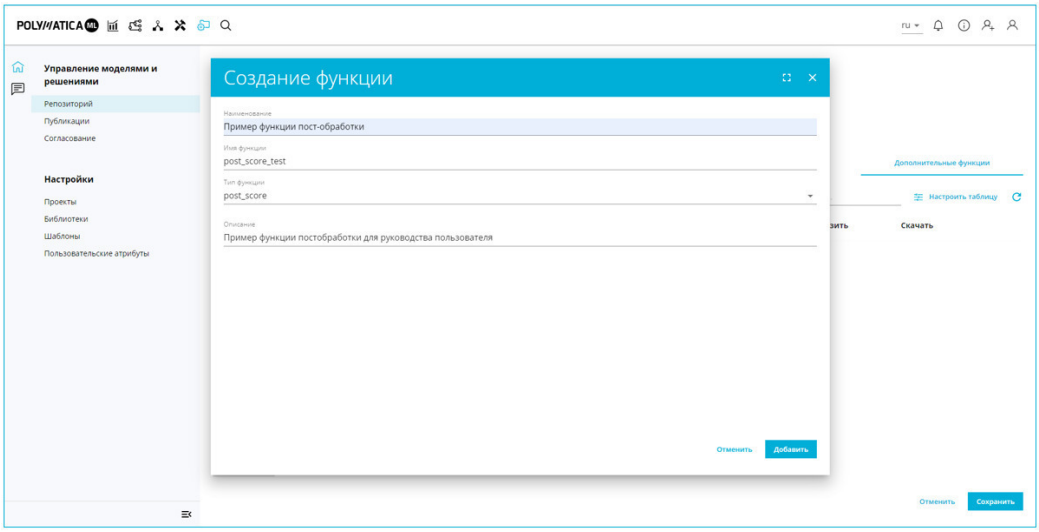


Рисунок 188 Окно Создание функции

- В списке вкладки **Дополнительные функции** отобразится созданная функция. В столбце **Загрузить** необходимо выбрать иконку , выбрать необходимый файл с расширением .py и вновь нажать на эту же иконку.

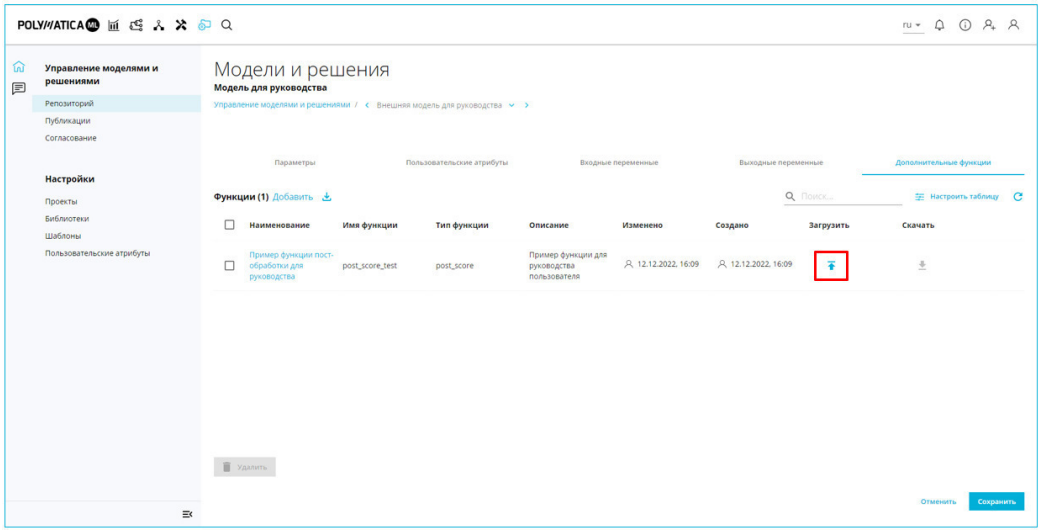


Рисунок 189 Пример загрузки функции пост-обработки

- После загрузки файла с функцией активизируется иконка в столбце **Скачать**.

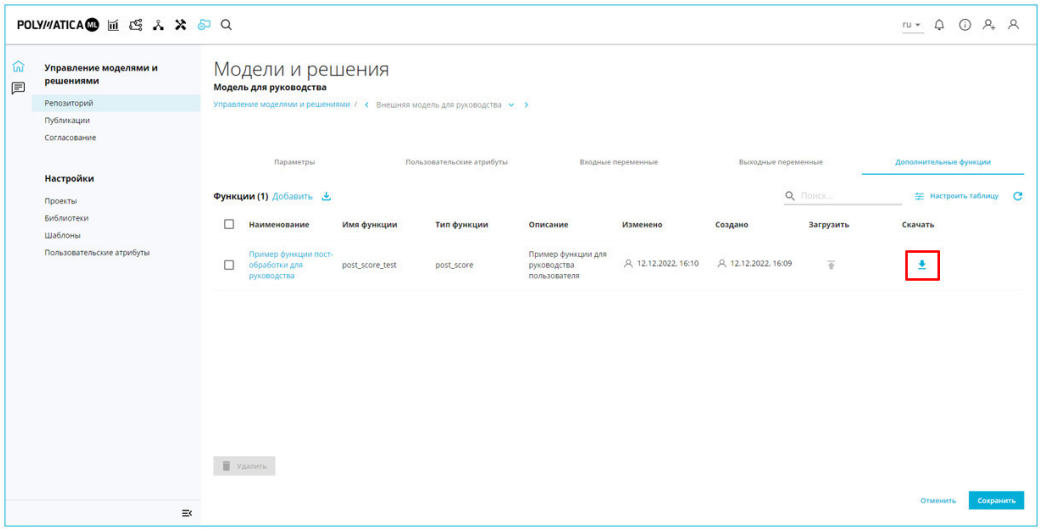


Рисунок 190 Пример выгрузки функции пост-обработки

5.5. Версия модели

Версия модели является итеративной реализацией модели.

5.5.1. Импорт версии модели

Зарегистрировать **Версию модели** в **Репозиторий** можно аналогично **Модели** - из компонентов **Построение моделей (MD)** и **Управление моделями и решениями (DM)**, а также импортировать извне.

5.5.2. Удаление версии модели

Для удаления Версии модели необходимо:

- Убедиться в том, что у удаляемой Версии модели нет Публикаций. Если Публикации есть – снять их с публикации и удалить.
- Выбрать чекбокс рядом с удаляемой Версию модели.
- Выбрать кнопку **«Удалить»**.

5.5.3. Работа с версией модели

Для редактирования и работы с моделью необходимо выбрать ее из списка Репозитория.

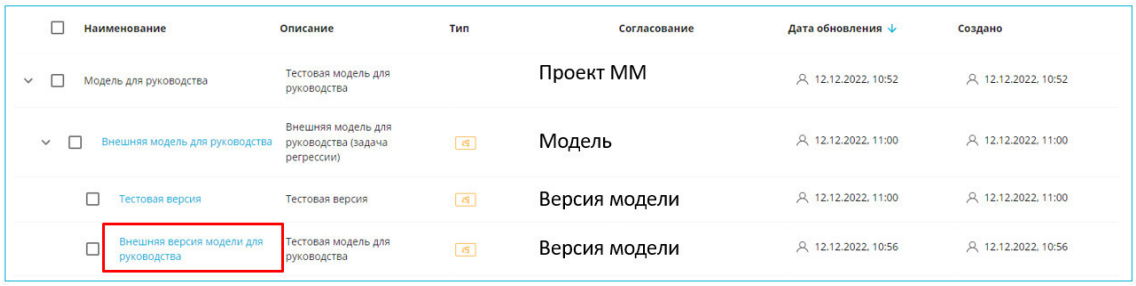


Рисунок 191 Пример выбора версии модели

Открывшийся экран представляет собой набор вкладок, которые позволяют настраивать Версию модели. Они различаются для Версий модели MD и DM.

Набор вкладок для Версии модели MD:

- Параметры
- Пользовательские атрибуты
- Вложения
- Настройки публикации
- Характеристики
- Скоринговый код

Набор вкладок для Версии модели DM:

- Параметры
- Пользовательские атрибуты
- Подключения
- Вложения
- Настройки публикации
- Скоринговый код

На вкладке **Параметры** отображены Тип, Имя, Описание и номер версии модели, а также используемые алгоритмы и библиотеки. Помимо этого здесь есть возможность перейти в проект MD, из которого была зарегистрирована версия (кнопка Перейти в **Model Designer**), а также отправить версию модели на согласование (кнопка **Ожидает отправки на согласование**, подробнее о процессе согласования в разделе Согласование версий модели и публикаций).

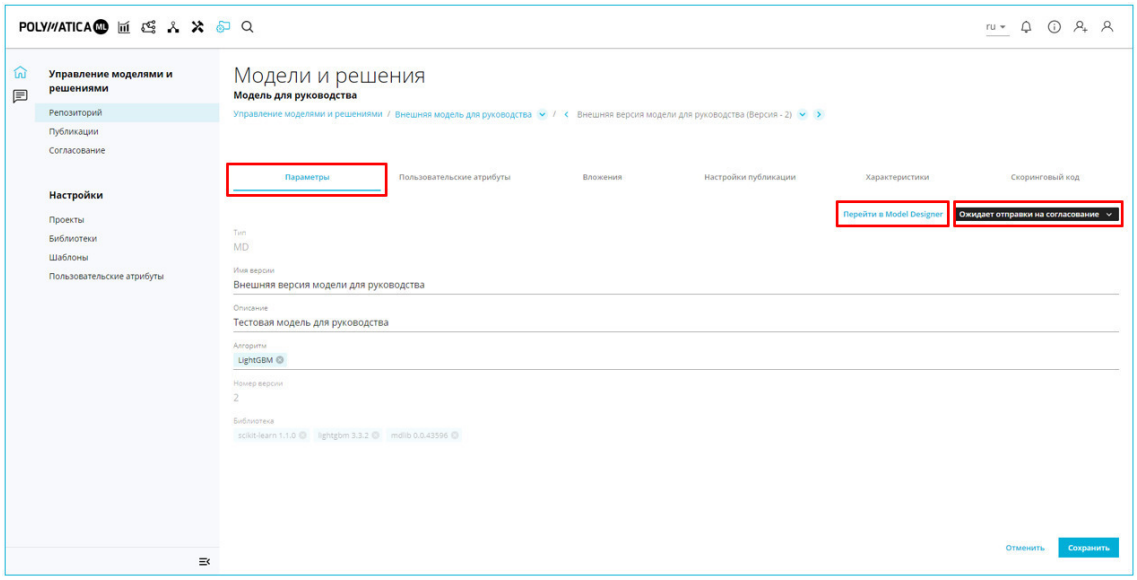


Рисунок 192 Вкладка Параметры

Вкладка **Пользовательские атрибуты** позволяет задать конкретные значения для выбранных атрибутов. Для этого нужно нажать на пустое поле столбца **Значение** и ввести необходимое значение.

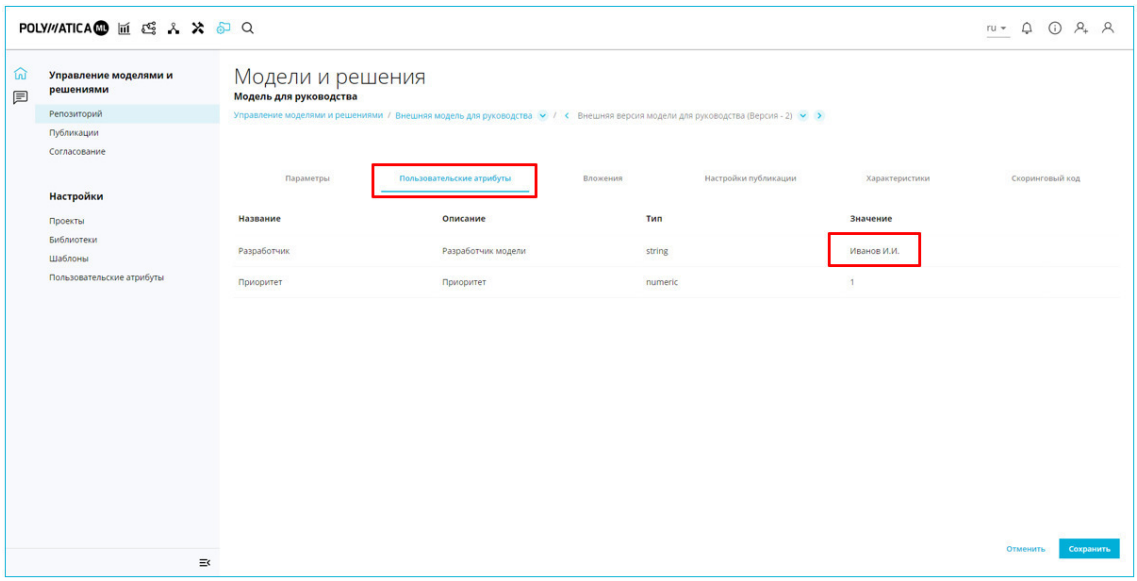


Рисунок 193 Вкладка Пользовательские атрибуты

Вкладка **Вложения** позволяет добавить необходимые артефакты (документацию, инструкции, список и объяснение атрибутов и т.д.) к версии модели. Для этого необходимо выбрать иконку и загрузить необходимые файлы. Далее любой Пользователь, имеющий доступ к данной версии модели, сможет скачать необходимый файл, выбрав иконку рядом.

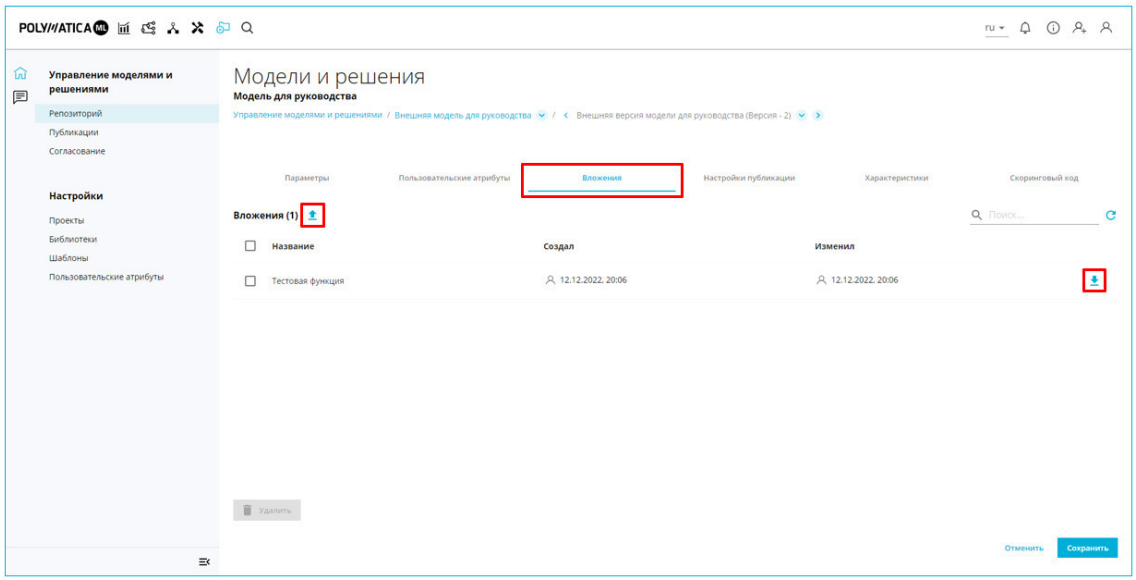


Рисунок 194 Вкладка Вложения

Вкладка **Настройки публикации** позволяет опубликовать версию модели в виде:

- Сервиса.
- Пакетного обработчика (Пакетная публикация).
- Скрипта (только для типа модели MD).
- Оценить качество модели (Оценка характеристик) (только для типа модели MD).
- Polymatica Analytics (только для типа модели MD).

Подробнее про публикацию модели в разделе Публикация версии модели.

Также в меню ⋮ рядом с публикацией можно инициировать процесс согласования публикации версии модели (подробнее в разделе Согласование версий модели и публикаций).

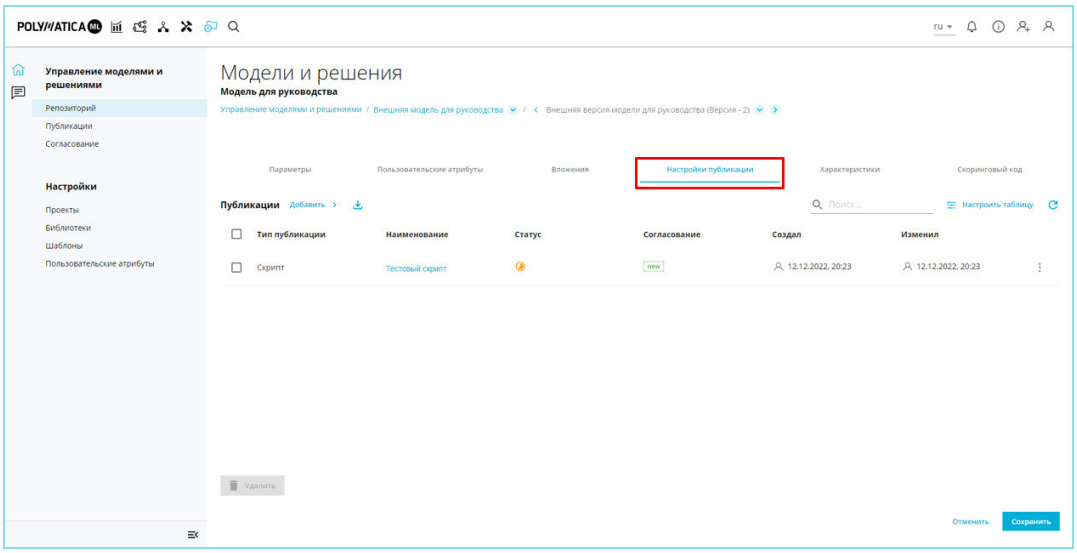


Рисунок 195 Вкладка Настройки публикации

Вкладка **Характеристики** актуальна только для моделей типа MD и содержит в себе результаты оценки качества модели при расчетах в MD (в столбце **Тип выборки** - Обучение, Тест, Валидация), а также при публикации модели типа Оценка характеристик (в столбце **Тип выборки** - Скоринг).

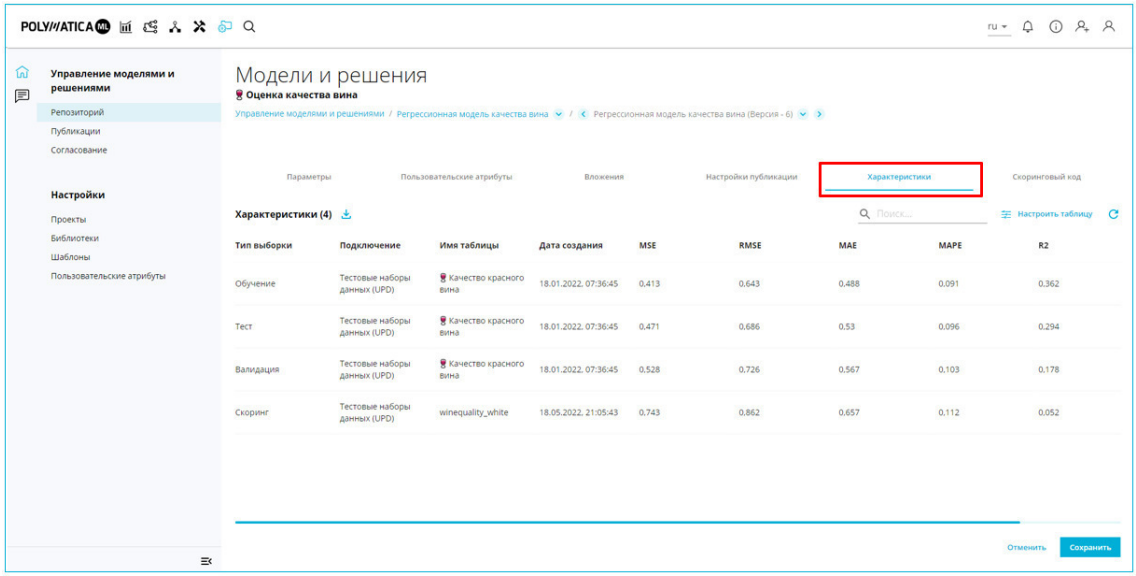


Рисунок 196 Вкладка Характеристики

Во вкладке **Скоринговый код** отображен код модели. Для раскрытия архива с моделью необходимо выбрать иконку 

. В раскрывшемся списке можно увидеть pkl-файлы (просмотр данных файлов не предусмотрен), файл инициализации модели, а также файл со скоринговым кодом score.py. Два последних файла можно открыть и посмотреть их содержимое (изменить содержимое в данном редакторе невозможно).

В файле score.py можно заменить функции pre_score и post_score. Именно сюда добавляются пользовательские функции, загруженные в модель (подробнее в разделе Работа с моделью).

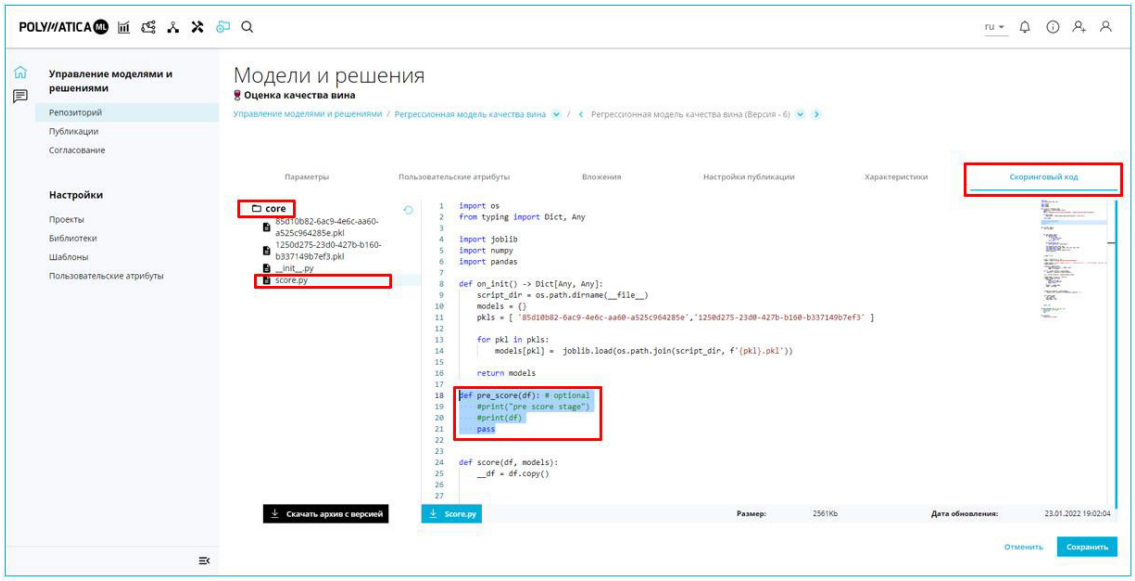


Рисунок 197 Вкладка Скоринговый код

Вкладка **Подключения** актуальна только для моделей типа DM. В ней отображен список всех Подключений (к БД), которые используются в решении.

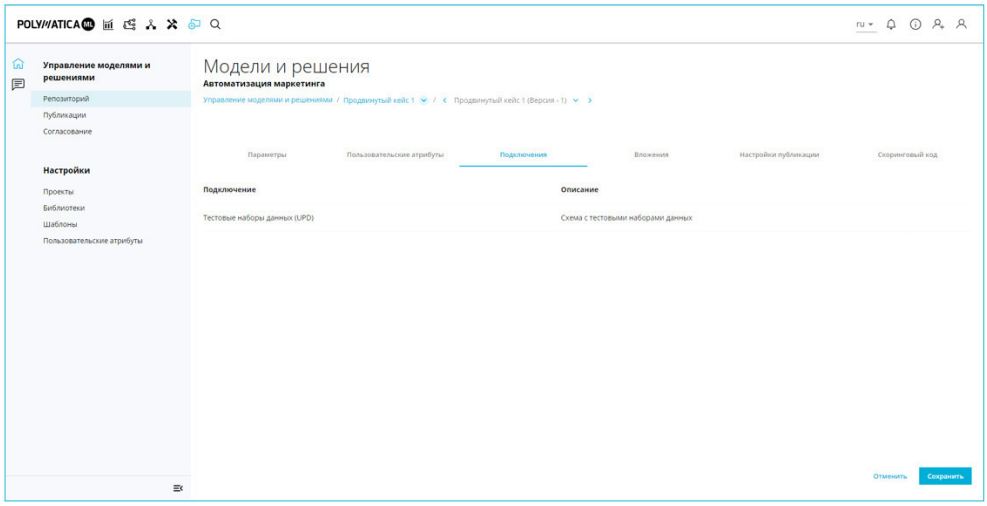


Рисунок 198 Вкладка Подключения

5.6. Публикация версии модели

Модель может быть опубликована в виде:

- Сервиса.
- Пакетной публикации.
- Скрипта (только для типа модели MD).
- Оценить качество модели (Оценка характеристик) (только для типа модели MD).
- Polymatica Analytics (только для типа модели MD).

5.6.1. Сервис

Для публикации модели в виде Сервиса необходимо:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Настройки публикации** нажать на кнопку **Добавить** и в выпадающем списке выбрать пункт **Сервис**.

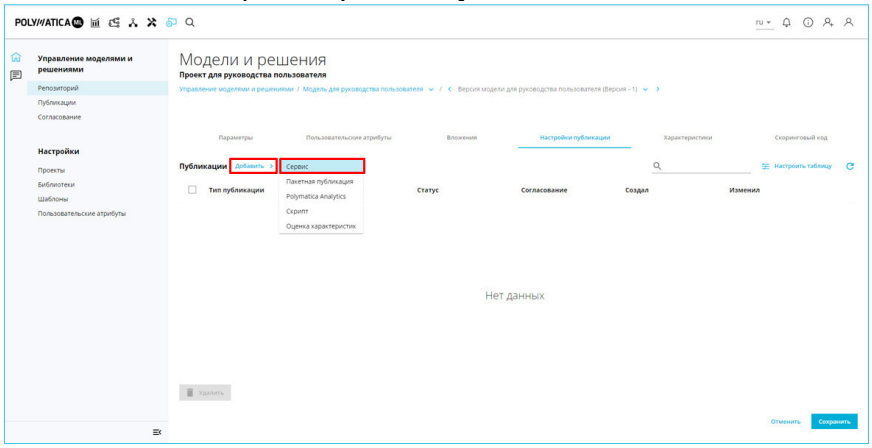


Рисунок 199 Создание публикации типа Сервис

3. В открывшемся окне **Создание экземпляра** во вкладке **Параметры** задать следующие настройки:
    - **Наименование**.
    - **Уровень логирования** — какие события добавлять в лог (All, Debug, Info, Warn, Error, Fatal, Off).
    - **Количество реплик**.
    - Название эндпоинта (заглавными латинскими буквами без пробелов)
    - **Максимальное количество ядер**.
    - **Максимальное количество оперативной памяти** (в Мб).
    - **Pre score** — функция предобработки (при необходимости).
    - **Post score** — функция постобработки (при необходимости).

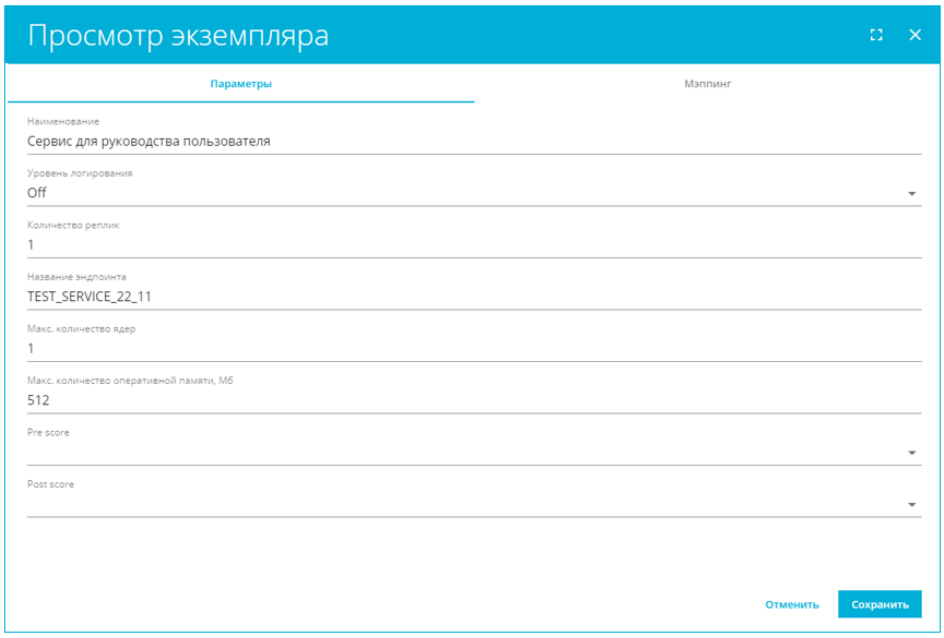


Рисунок 200 Окно Создание экземпляра

4. При необходимости изменить имена переменным (столбец **Имя переменной мэппинга)** во вкладке **Мэппинг**.
5. Сохранить изменения.
6. После успешной сборки модели (иконка ✅ в столбце Статус) выбрать меню ⋮ и в выпадающем списке выбрать пункт **Запустить сервис**.

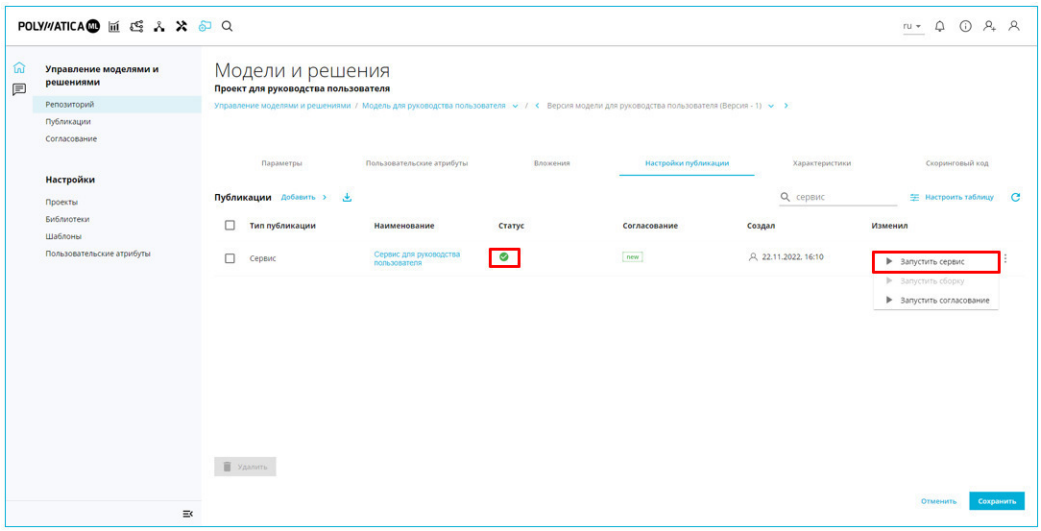


Рисунок 201 Запуск сервиса

7. Опубликованный сервис появится в разделе **Публикации** боковой панели компонента **Управление моделями и решениями (ММ).** После изменения статуса публикации на Работает (иконка ●) выбрать меню ⋮ и в выпадающем списке выбрать пункт **Перейти к сервису**.

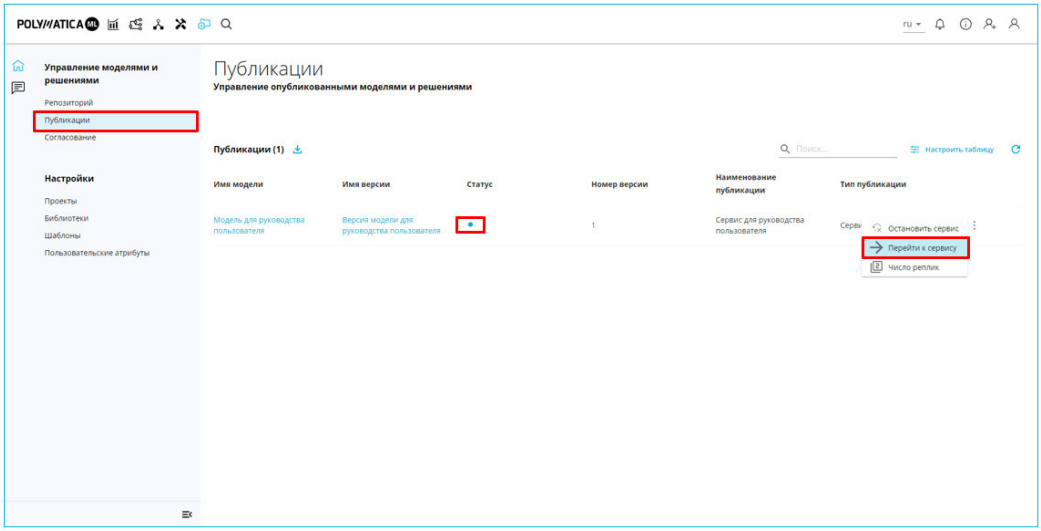


Рисунок 202 Переход к сервису

Для снятия с публикации модели в виде Сервиса необходимо:

- В разделе **Публикации** боковой панели компонента **Управление моделями и решениями (ММ)** выбрать меню рядом с созданной публикацией и в выпадающем списке нажать на **Остановить сервис.**

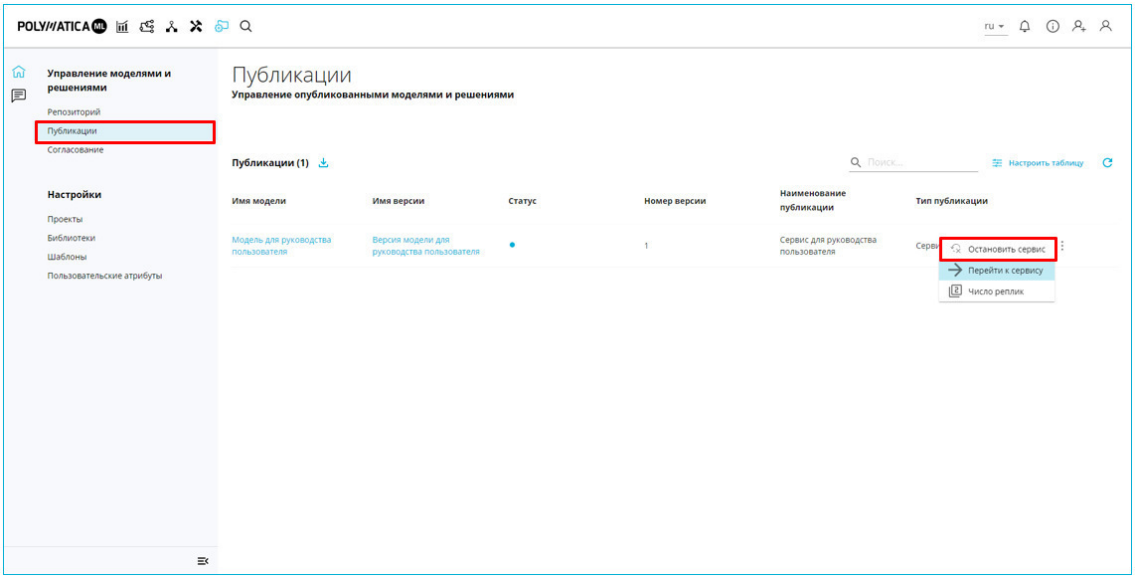


Рисунок 203 Снятие сервиса с публикации

5.6.2. Пакетная публикация

Для создания **пакетной публикации** модели необходимо:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Настройки публикации** нажать на кнопку **Добавить** и в выпадающем списке выбрать пункт **Пакетная публикация**.

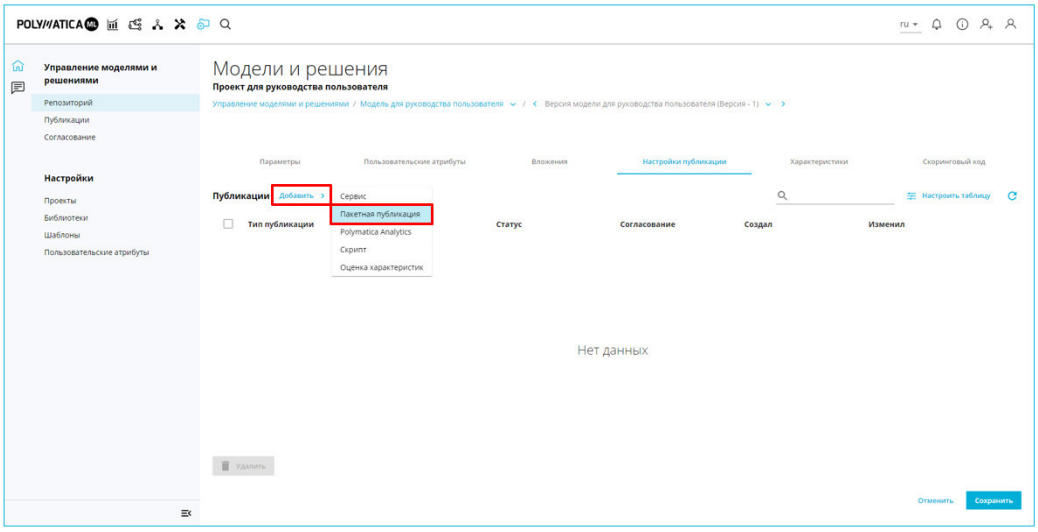


Рисунок 204 Создание пакетной публикации

3. В открывшемся окне **Создание экземпляра** во вкладке **Параметры** задать следующие настройки:
    - **Наименование**.
    - **Подключение** — указать подключение, в котором находится необходимый для скоринга набор данных.
    - **input_table_name** — название входного набора данных.

- **Подключение** — указать подключение, в которое необходимо положить результаты скоринга.
- **output_table_name** — название выходного набора данных.
- **exit_code_column** (при необходимости записи ошибок и информации об обработке в отдельную колонку).
- **Уровень логирования** — какие события добавлять в лог (All, Debug, Info, Warn, Error, Fatal, Off).
- Чекбокс **Обработка в несколько потоков** (при выборе чекбокса отобразится параметр **Количество реплик**, в котором можно указать необходимое число потоков)
- **Максимальное количество ядер**.
- **Максимальное количество оперативной памяти** (в Мб).
- **Pre score** — функция предобработки (при необходимости).
- **Post score** — функция постобработки (при необходимости).

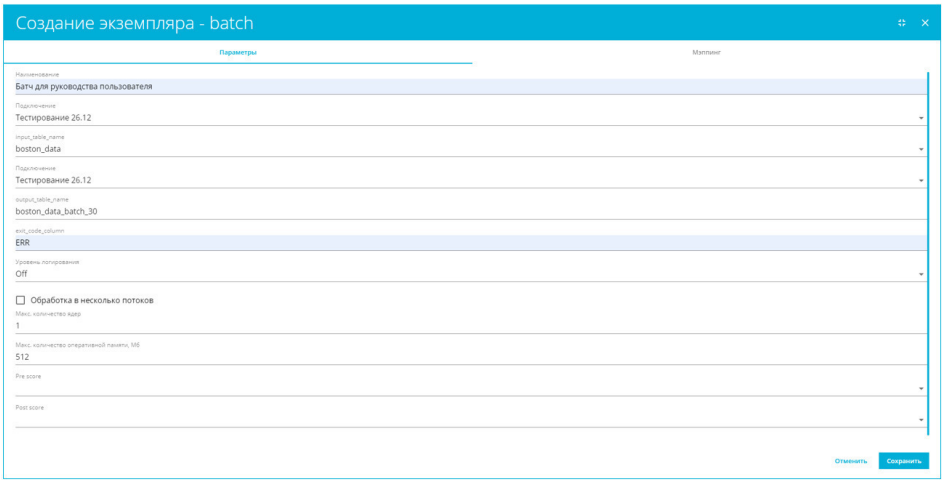


Рисунок 205 Окно Создание экземпляра

4. Задать имена переменным во вкладке **Мэппинг**.

> Значения столбца **Имя переменной мэппинга** для входных переменных должно совпадать с наименованием соответствующего столбца в таблице. Для выходных переменных можно оставить определенное системой имя.

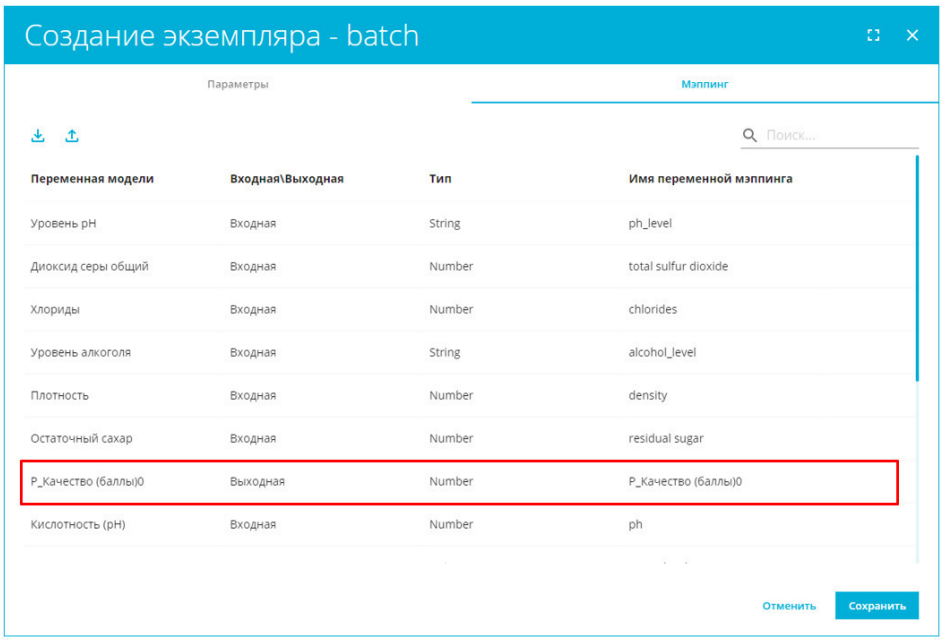


Рисунок 206 Мэппинг переменных

5. Сохранить изменения.
6. После успешной сборки модели (иконка в столбце Статус) выбрать меню и в выпадающем списке выбрать пункт **Запустить пакетную обработку**.

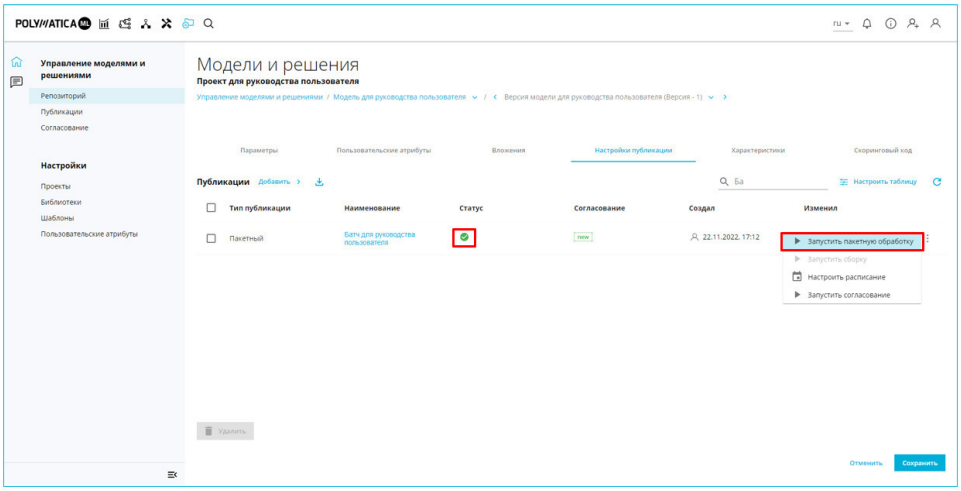


Рисунок 207 Запуск пакетной обработки

7. Публикация появится в разделе **Публикации** боковой панели компонента **Управление моделями и решениями (ММ).** После завершения пакетной обработки публикация пропадет из раздела **Публикации**. С результатами расчетов Пользователь может ознакомиться в выходном файле (параметр **output_table_name**) в соответствующей БД (параметр **Подключение**). В выходной таблице помимо исходных столбцов появятся столбец **exit_code_column (**значение которого в случае отсутствия ошибок должно быть равным 1), а также выходные столбцы модели (количество зависит от решаемой задачи).

Запуск пакетной обработки можно поставить на расписание. Для этого необходимо:

- Рядом с интересующей публикацией в меню ⋮ выбрать пункт **Настроить расписание**.
- Выбрать **Добавить Правило**.
- В открывшемся окне **Настройка расписания** задать:
  - Название, которое далее будет отображаться в системе.
  - Выбрать дату и время первого запуска.
  - Настроить повторение (период — ежедневно, еженедельно, ежемесячно и время повторного запуска).
  - Окончание расписания (Всегда запускать/Число выполнений (указать число запусков)/Дата окончания (указать последнюю дату запуска)).
  - Сохранить настройки.

5.6.3. Оценка характеристик

Для **оценки качества модели** необходимо:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Настройки публикации** нажать на кнопку **Добавить** и в выпадающем списке выбрать пункт **Оценка характеристик**.

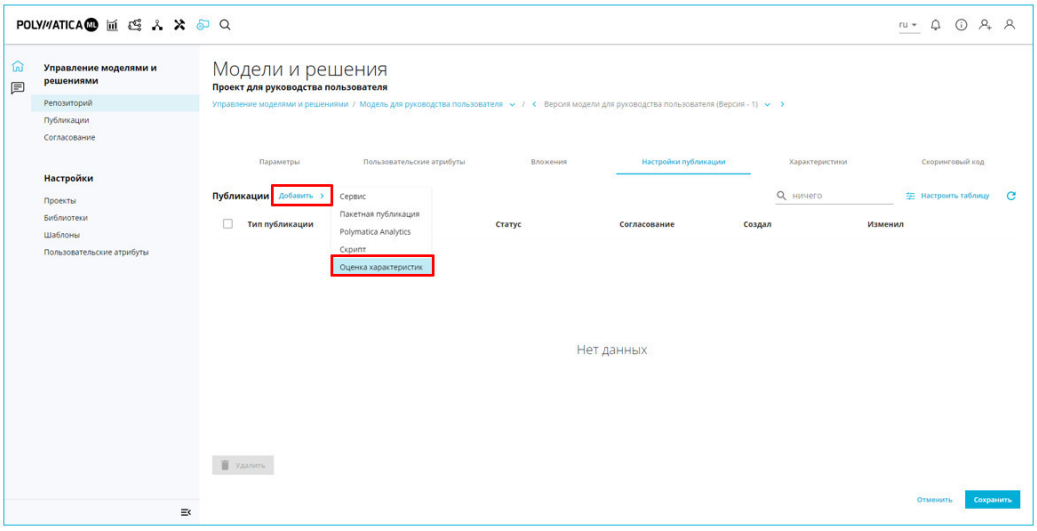


Рисунок 208 Создание публикации типа Оценка

3. В открывшемся окне **Создание экземпляра** во вкладке **Параметры** задать следующие настройки:
   - **Наименование.**
   - **Подключение** — указать подключение, в котором находится необходимый для скоринга набор данных.

- **input_table_name** — название входного набора данных.
- **Подключение** — указать подключение, в которое необходимо положить результаты скоринга.
- **output_table_name** — название выходного набора данных.
- **Уровень логирования** — какие события добавлять в лог (All, Debug, Info, Warn, Error, Fatal, Off).
- **Максимальное количество ядер.**
- **Максимальное количество оперативной памяти** (в Мб).
- **Pre score** — функция предобработки (при необходимости).
- **Post score** — функция постобработки (при необходимости).
- **Целевая переменная** — переменная в исходном наборе данных, относительно которой будет рассчитываться оценка.

Задание параметра **Целевая переменная** зависит от решаемой задачи.
Для задачи регрессии: Для единственной выходной переменной необходимо выбрать **Тип метрики = interval**.

Рисунок 209 Пример задания Целевой переменной для задачи регрессии

Для задачи бинарной классификации: Для двух переменных с префиксом P_ необходимо выбрать **Тип метрики = binary_prob** (т.к. они отражают вероятность принадлежности наблюдения к классу), для переменной с префиксом C_ необходимо выбрать **Тип метрики = nominal_class** (т.к. она отражает принадлежность к одному из классов).

Рисунок 210 Пример задания Целевой переменной для задачи бинарной классификации

Для задачи многоклассовой классификации: Для всех переменных (количество совпадает с количеством классов) с префиксом P_ необходимо выбрать **Тип метрики = nominal_prob** (т.к. они

отражают вероятность принадлежности наблюдения к классу), для переменной с префиксом C_ необходимо выбрать **Тип метрики = nominal_class** (т.к. она отражает принадлежность к одному из классов).

Рисунок 211 Пример задания Целевой переменной для задачи многоклассовой классификации

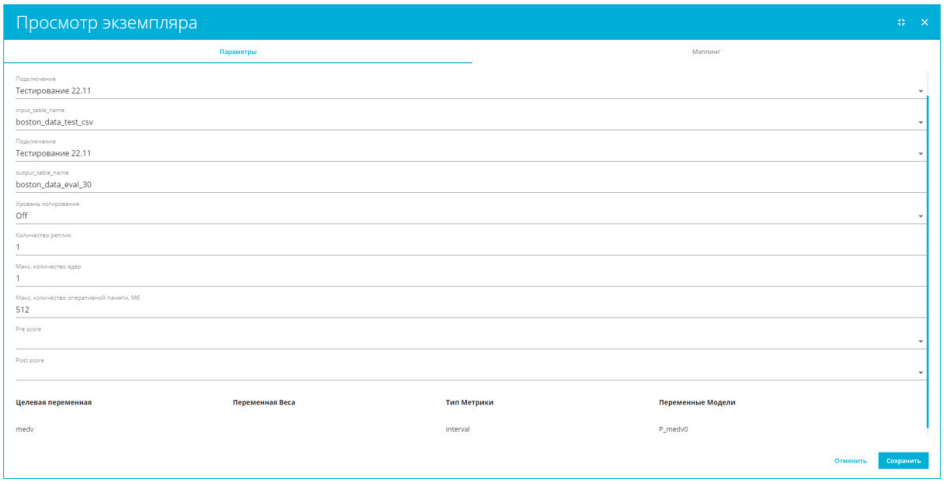


Рисунок 212 Окно Создание экзепляра

4. Задать имена переменным во вкладке **Мэппинг**.

Значения столбца **Имя переменной мэппинга** для входных переменных должно совпадать с наименованием соответствующего столбца в таблице. Для выходных переменных можно оставить определенное системой имя.

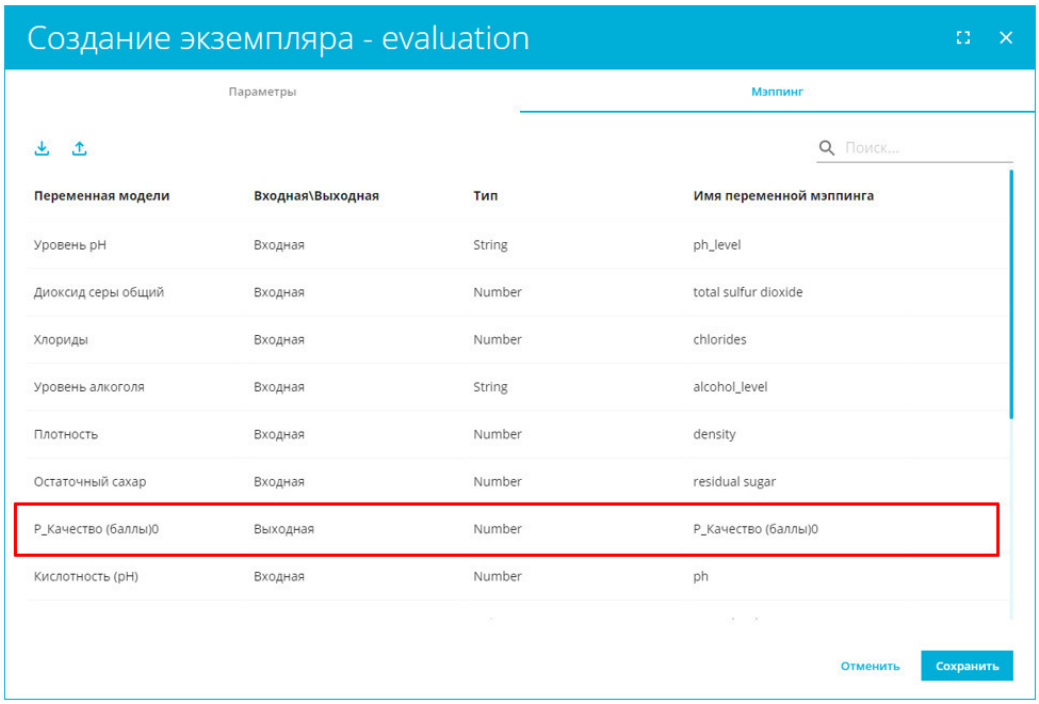


Рисунок 213 Задание мэппинга переменных

5. Сохранить изменения.
6. После успешной сборки модели (иконка в столбце Статус) выбрать меню и в выпадающем списке выбрать пункт **Запустить оценку метрик**.
7. Публикация появится в разделе **Публикации** боковой панели компонента **Управление моделями и решениями (ММ).** После завершения оценки публикация пропадет из раздела **Публикации**. С результатами расчетов Пользователь может ознакомиться в во вкладке **Характеристики модели**.

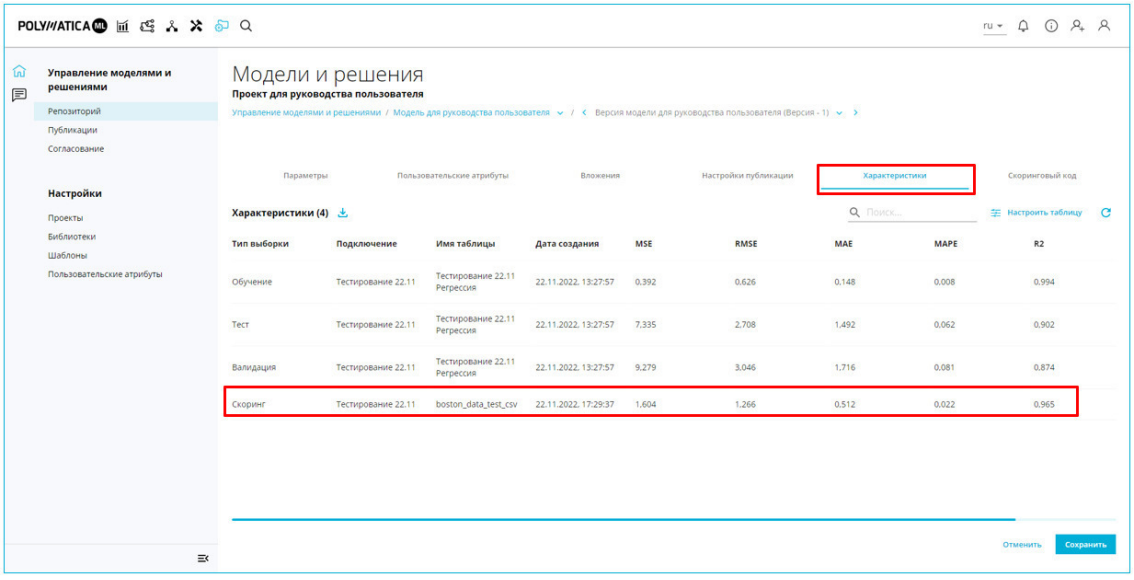


Рисунок 214 Результаты расчета оценки модели

Запуск оценки характеристик модели можно поставить на расписание. Для этого необходимо:

- Рядом с интересующей публикацией в меню ⋮ выбрать пункт **Настроить расписание**.
- Выбрать **Добавить Правило**.
- В открывшемся окне **Настройка расписания** задать:
    - Название, которое далее будет отображаться в системе.
    - Выбрать дату и время первого запуска.
    - Настроить повторение (период — ежедневно, еженедельно, ежемесячно и время повторного запуска).
    - Окончание расписания (Всегда запускать/Число выполнений (указать число запусков)/Дата окончания (указать последнюю дату запуска)).
    - Сохранить настройки.

5.6.4. Скрипт

Для публикации модели в виде **Скрипта** необходимо:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Настройки публикации** нажать на кнопку **Добавить** и в выпадающем списке выбрать пункт **Скрипт**.
3. В открывшемся окне **Создание экземпляра** во вкладке **Параметры** задать **Наименование**.
4. Сохранить изменения.
5. После успешной сборки модели (иконка

в столбце Статус) выбрать меню

⋮

и в выпадающем списке выбрать пункт **Скачать**. В скаченном архиве можно найти все необходимые для запуска файлы, а также инструкцию по использованию (файл **README.md**).

5.6.5. Публикация для скоринга мультисфер (Polymatica Analytics)

Для публикации модели для скоринга мультисфер (**Polymatica Analytics**) необходимо:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.

2. Во вкладке **Настройки публикации** нажать на кнопку **Добавить** и в выпадающем списке выбрать пункт **Polymatica Analytics**.
3. В открывшемся окне **Создание экземпляра** во вкладке **Параметры** задать следующие настройки:

- **Наименование**.
- **Уровень логирования** — какие события добавлять в лог (All, Debug, Info, Warn, Error, Fatal, Off).
- **Количество реплик**.
- Название эндпоинта (заглавными латинскими буквами без пробелов)
- **Максимальное количество ядер**.
- **Максимальное количество оперативной памяти** (в Мб).
- **Pre score** — функция предобработки (при необходимости).
- **Post score** — функция постобработки (при необходимости).

1. При необходимости поправить **Имя переменной мэппинга** во вкладке **Мэппинг**.
2. Сохранить изменения.
3. После успешной сборки модели (иконка в столбце Статус) выбрать меню и в выпадающем списке выбрать пункт **Запустить**.
4. Публикация появится в разделе **Публикации** боковой панели компонента **Управление моделями и решениями (ММ).** После изменения статуса публикации на Работает (иконка) выбрать меню и в выпадающем списке выбрать пункт **Перейти к скорингу мультисфер**.

5.7. Согласование версий модели и публикаций

5.7.1. Шаблон согласования

Шаблон представляет собой BPMN-процесс, в соответствии с которым будет выполняться согласование версии модели или публикации.

Шаблоны могут быть произвольными и базируются на ролевых моделях пользователей. Создаются шаблоны Администратором.

При выборе пункта **Шаблоны** боковой панели открывается одноименный раздел со списком Шаблонов согласования.

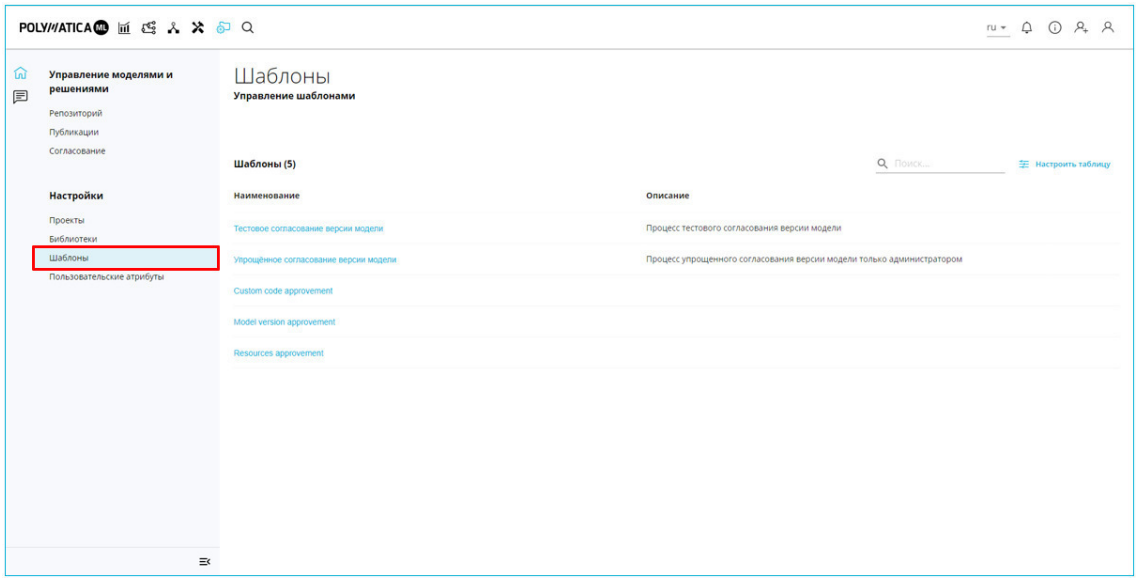


Рисунок 215 Раздел Шаблоны согласования

При нажатии на наименование шаблона открывается окно просмотра шаблона.

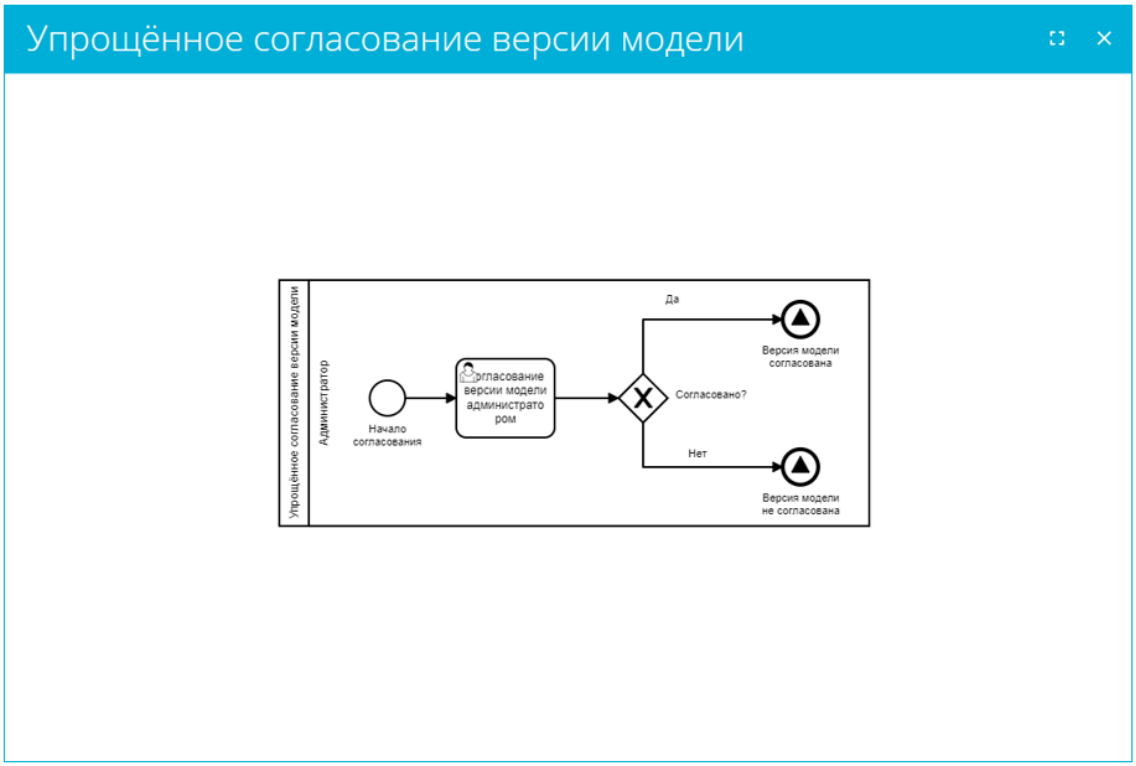


Рисунок 216 Окно просмотра шаблона согласования

5.7.2. Процесс согласования

Процесс согласования происходит в соответствии с шаблонами, которые были указаны Пользователем в **Проекте ММ**.

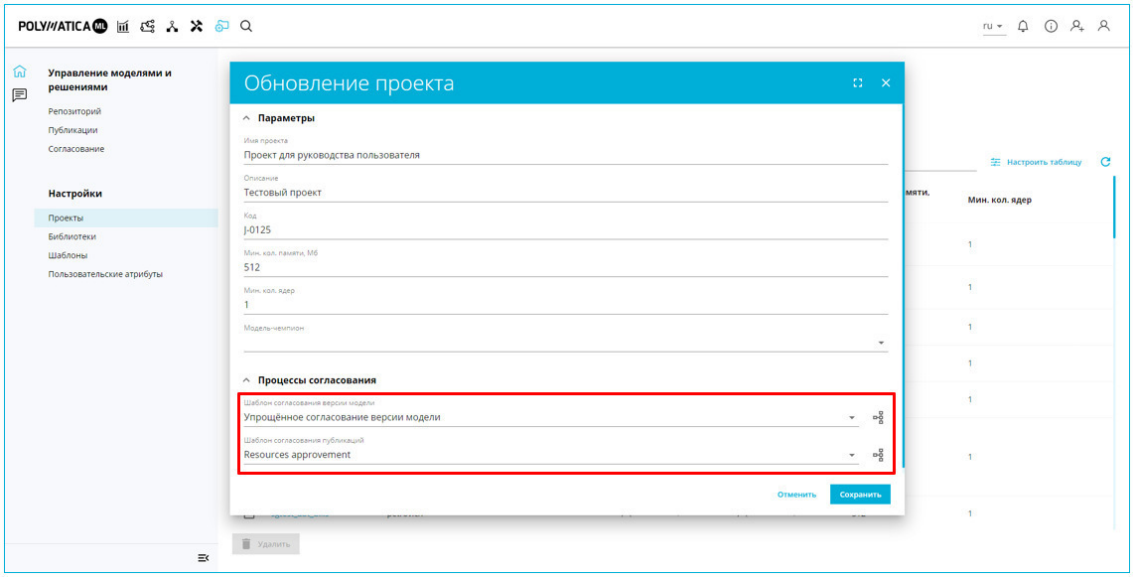


Рисунок 217 Задание шаблонов согласования в проекте ММ

Для запуска согласования версии модели:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Параметры** нажать на кнопку **Ожидает отправки на согласование** и выпадающем списке выбрать пункт **Отправить на согласование**.

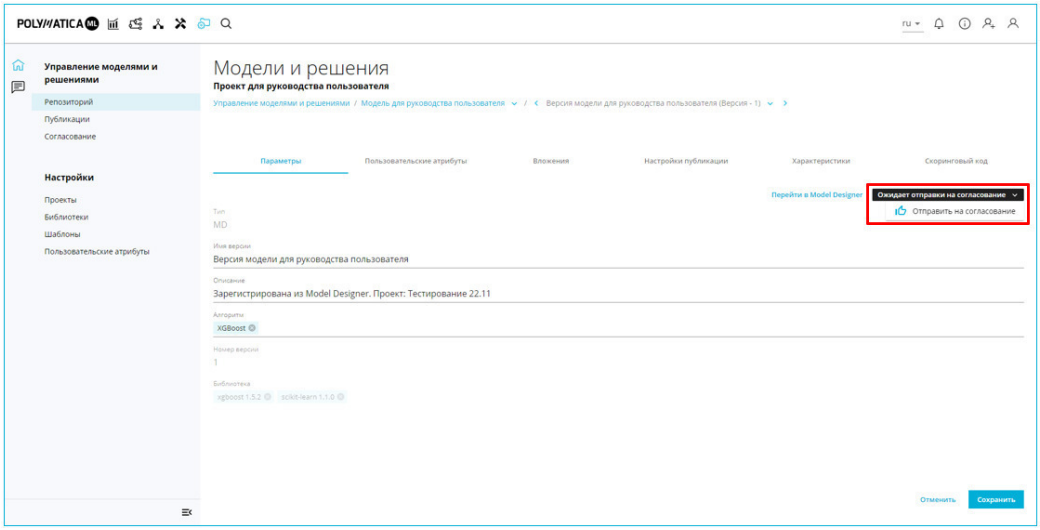


Рисунок 218 Запуск процесса согласования

3. Процесс согласования запущен. Статус версии модели изменился на **В процессе согласования**. В соответствии с шаблоном согласования всем заинтересованным пользователям придет сообщение о **Согласовании версии модели.** Для согласования (или отправки на доработку) Пользователю с соответствующими правами необходимо перейти в раздел **Согласование**, выбрать соответствующее сообщение, назначить себя в качестве Согласующего, проверить характеристики модели и согласовать (или отправить на доработку версию модели).

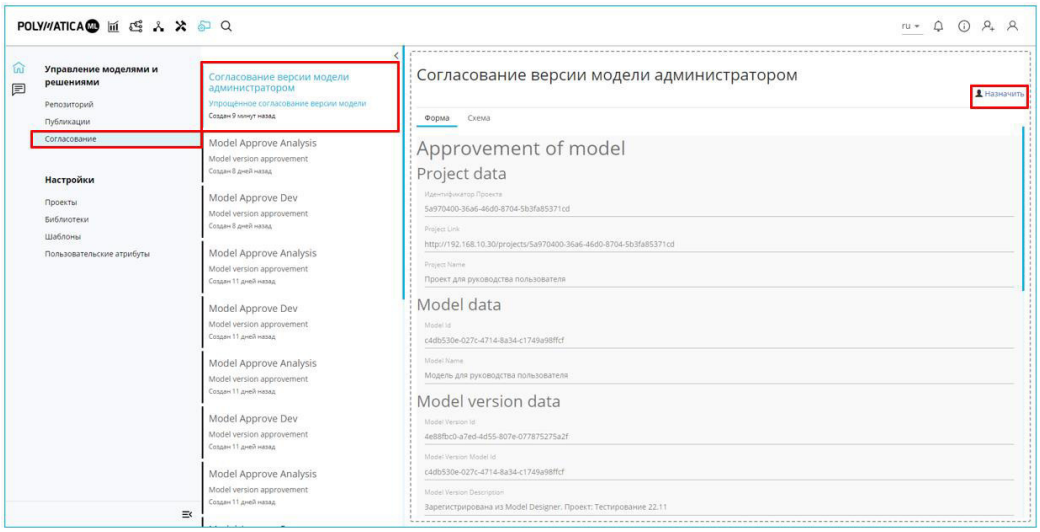


Рисунок 219 Процесс согласования Версии модели

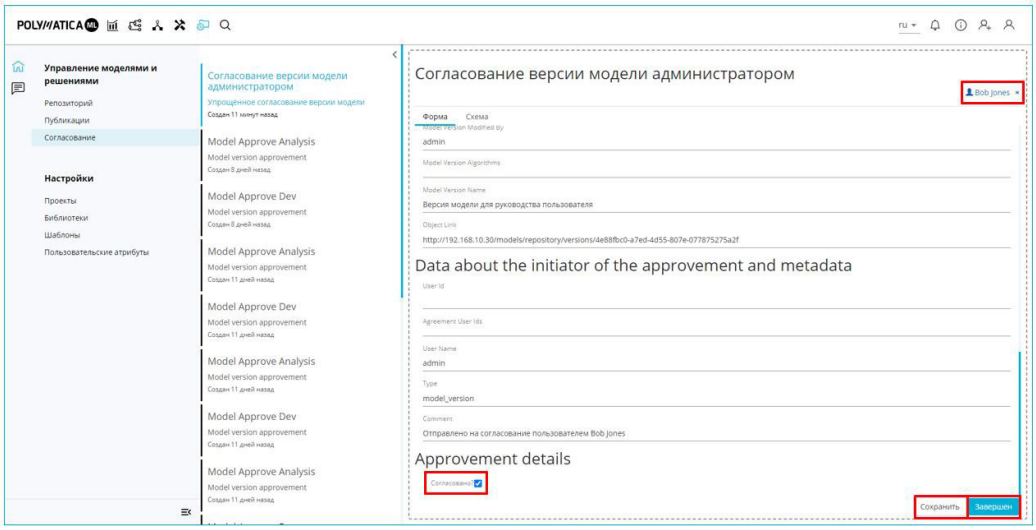


Рисунок 220 Процесс согласования Версии модели

4. В зависимости от решения Согласующего в Репозитории в строке с Версией модели в столбце Согласование отобразится соответствующий статус (Согласовано/Не согласовано).

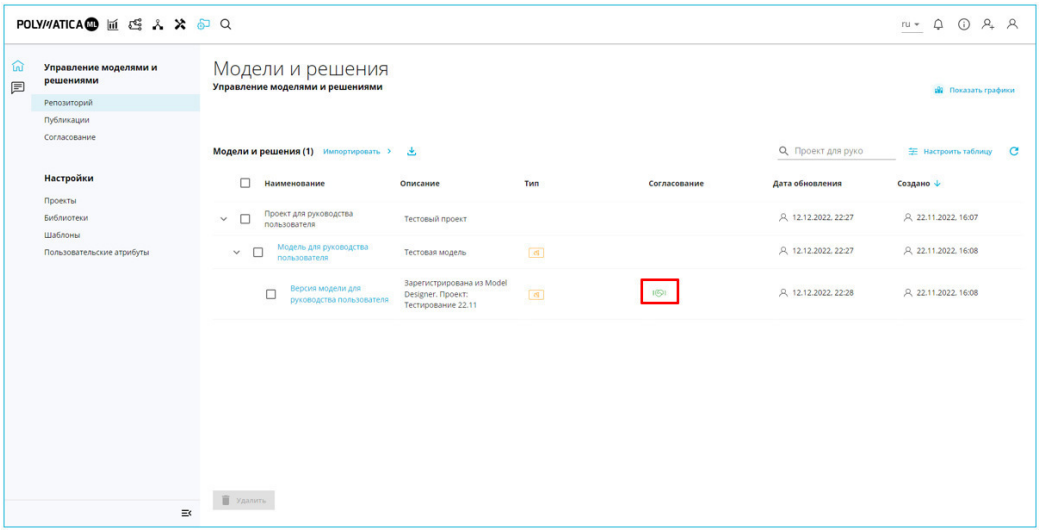


Рисунок 221 Статус согласования Версии модели

Для запуска согласования публикации:

1. Выбрать наименование интересующей Версии модели в **Репозитории**.
2. Во вкладке **Настройки публикации** выбрать интересующую публикацию и в меню ⋮ выбрать пункт **Запустить согласование**.

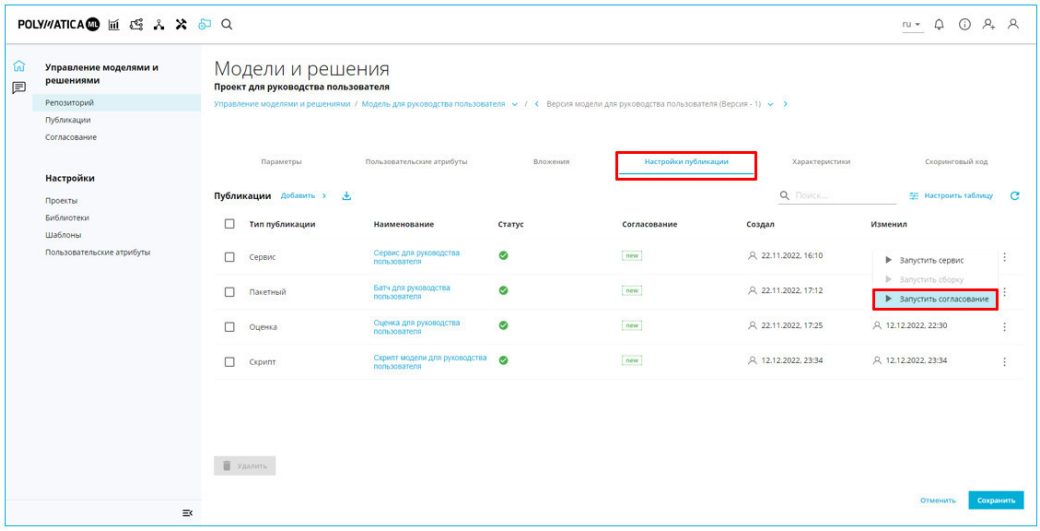


Рисунок 222 Запуск процесса согласования публикации

3. Процесс согласования запущен. Статус версии модели изменился на **В процессе согласования**. В соответствии с шаблоном согласования всем заинтересованным пользователям придет сообщение о **Согласовании версии модели.** Для согласования (или отправки на доработку) Пользователю с соответствующими правами необходимо перейти в раздел **Согласование**, выбрать соответствующее сообщение, назначить себя в качестве Согласующего, проверить характеристики модели и согласовать (или отправить на доработку версию модели).

4. В зависимости от решения Согласующего во вкладке **Настройки публикации** в строке с Публикацией в столбце **Согласование** отобразится соответствующий статус (Согласовано/Не согласовано).

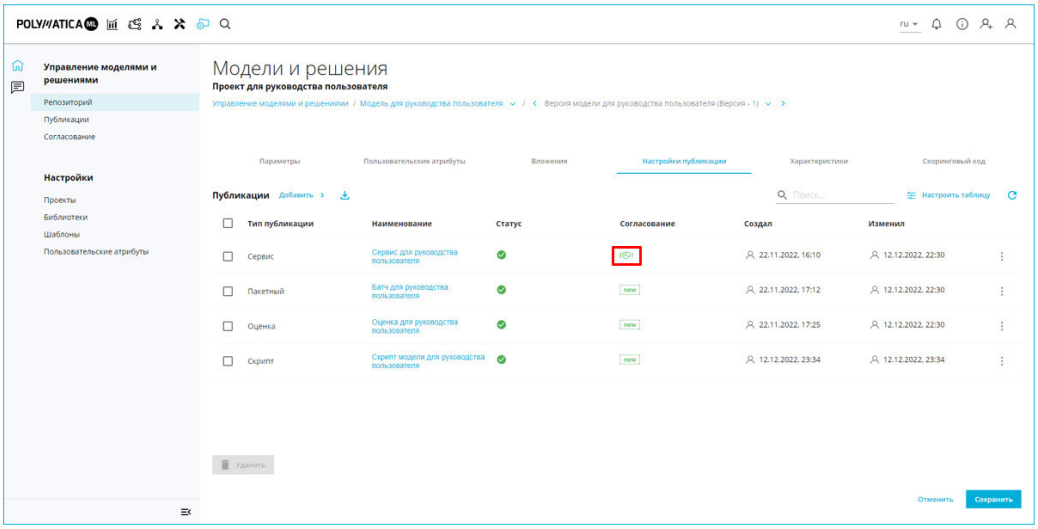


Рисунок 223 Статус согласования публикации

5.8. Библиотеки

Библиотеки используются при сборке публикаций версии модели.

Для ознакомления со списком библиотек необходимо выбрать одноименный пункт боковой панели компонента **Управление моделями и решениями (ММ)**.

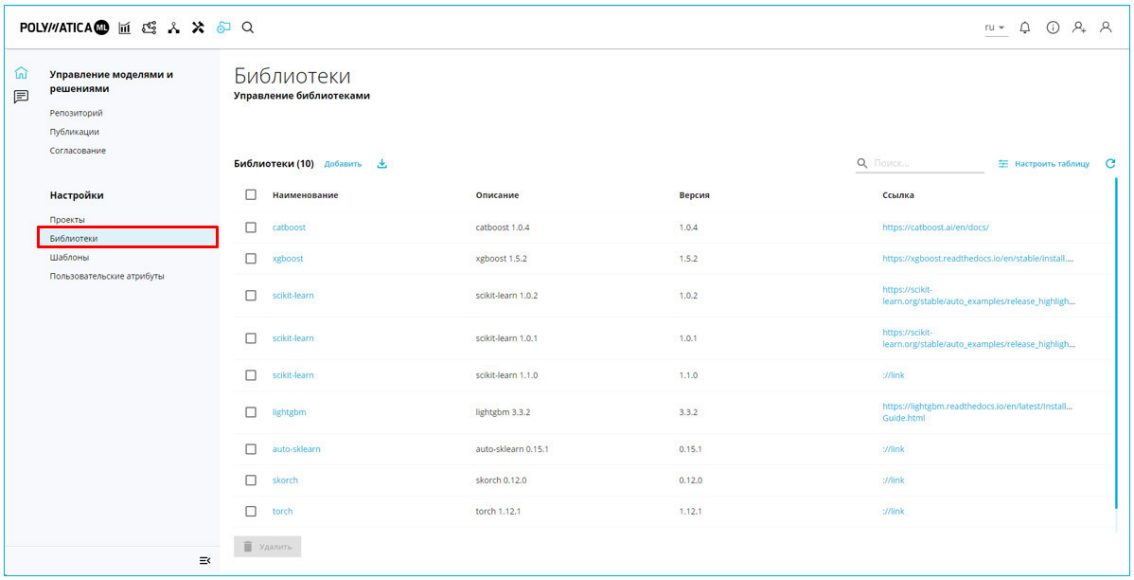


Рисунок 224 Раздел Библиотеки бокового меню

Для создания Библиотеки необходимо:

1. Выбрать кнопку **Добавить** в верхней части таблицы.
2. В открывшемся окне **Создание библиотеки** ввести следующие параметры:

- Наименование (в соответствии с реальным названием используемой библиотеки).
- Описание.
- Версию библиотеки.
- Ссылку.

3. Выбрать кнопку **Сохранить.**

Для удаления Библиотеки необходимо:

1. Выбрать чекбокс рядом с наименованием удаляемой библиотеки.
2. Выбрать кнопку **Удалить**.

Для обновления параметров библиотеки необходимо:

1. Нажать на наименование библиотеки.
2. В открывшемся окне **Просмотр библиотеки** ввести необходимые изменения.
3. Сохранить редактированные параметры, выбрав кнопку **Сохранить.**

5.9. Пользовательские атрибуты

Атрибуты предусмотрены для указания дополнительной информации о версии модели - инициалы автора, менеджера модели, дата разработки/важных изменений, приоритет версии и все то, что посчитает важным Пользователь.

Для создания нового пользовательского атрибута необходимо:

1. Выбрать одноименный пункт в боковой панели компонента **Управление моделями и решениями (ММ)**.
2. На открывшемся экране **Пользовательские атрибуты** выбрать кнопку **Добавить** в верхней части списка атрибутов.

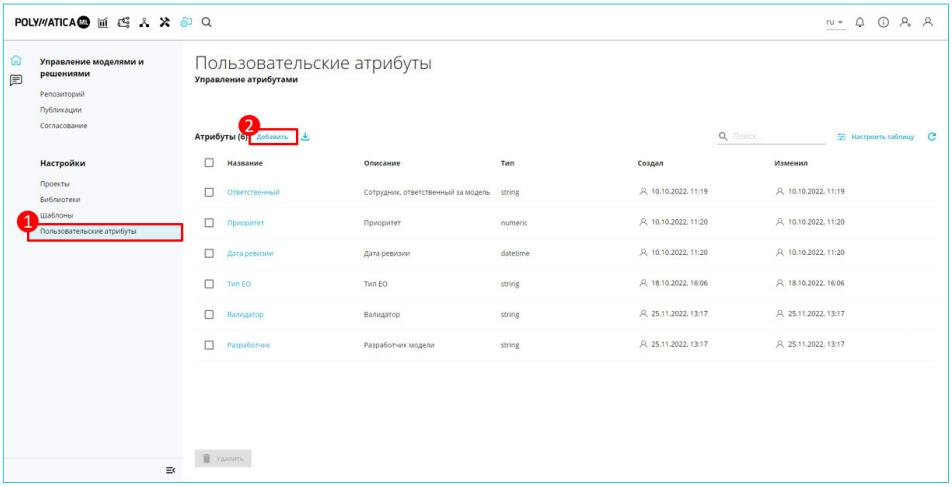


Рисунок 225 Создание нового Пользовательского атрибута

3. В открывшемся окне **Создание атрибута** задать название, описание и тип. Сохранить внесенные изменения
4. Созданный атрибут появится в списке экрана Пользовательские атрибуты и его можно будет указать в Модели в Репозитории и задать конкретное значение для Версий модели.

Для удаления пользовательского атрибута необходимо:

1. Выбрать чекбокс рядом с удаляемым атрибутом.
2. Нажать кнопку **Удалить** в нижней части экрана.

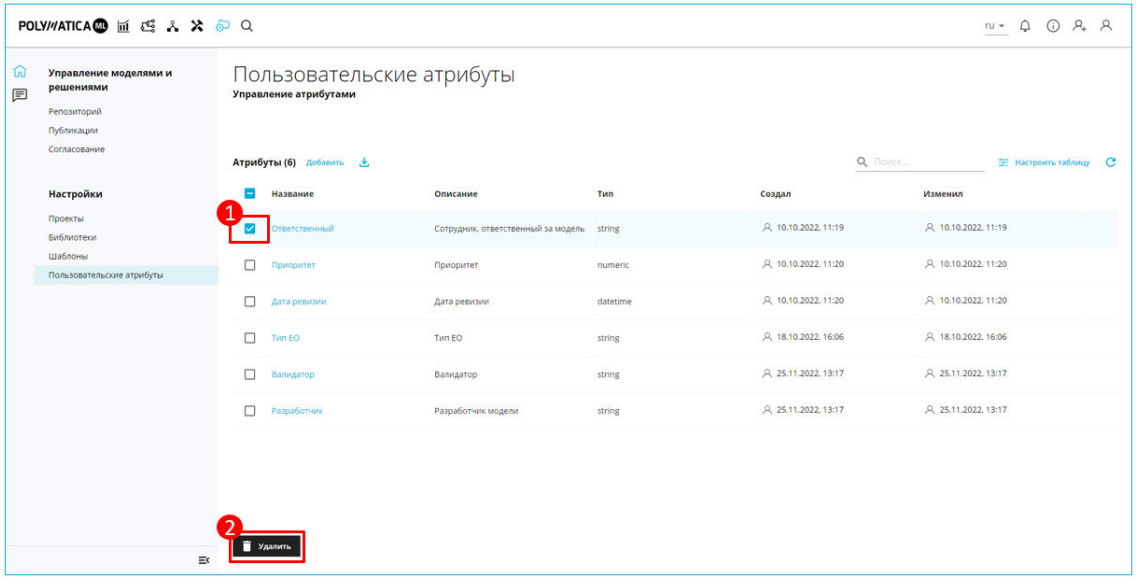


Рисунок 226 Удаление пользовательского атрибута

6. Глобальный поиск

6.1. Интерфейс экрана Глобального поиска

Экран Поиска открывается при выборе иконки в левом верхнем меню и представляет собой список доступных пользователю объектов системы.

Рисунок 227 Экран глобального поиска

При выборе наименования объекта откроется соответствующий раздел Модуля с выбранным объектом для дальнейшей работы. Помимо наименования самих объектов в таблице также отображены:

- Тип объекта – DataSet, MDProject, MMProject.
- Версия
- Описание объекта
- Дата создания и изменения объекта
- Автор
- GUID

Таблица с доступными пользователю объектами системы имеет гибкие настройки отображения. Пользователь может:

- изменить ширину любого столбца (для этого необходимо перетащить границу его заголовка до нужной ширины);
- сортировать таблицу (для этого необходимо выбрать иконку рядом с заголовком сортируемого столбца);
- скрывать/отображать столбцы и изменять их порядок в окне **Вид таблицы** (для открытия окна необходимо выбрать Настроить Таблицу в правом верхнем углу таблицы; при выборе иконки столбец скроется, при наведении на иконку активируется возможность перемещения столбца);
- сбросить внесенные изменения также в окне **Вид таблицы** (для этого выбрать кнопку **«Сбросить»**).

Для быстрого поиска объекта в таблице предусмотрено поле Поиск... в правой верхней части таблицы.
Объекты таблицы можно выгрузить в формате Excel. Для этого нужно выбрать иконку **Экспорта в excel** .
Помимо этого, реализована работа с тэгами для быстрого поиска объектов.

6.2. Работа с тэгами

Скрытая левая боковая панель содержит тэги, при помощи которых можно быстро найти интересующий объект.

Для раскрытия панели необходимо выбрать > в нижней части интерфейса.

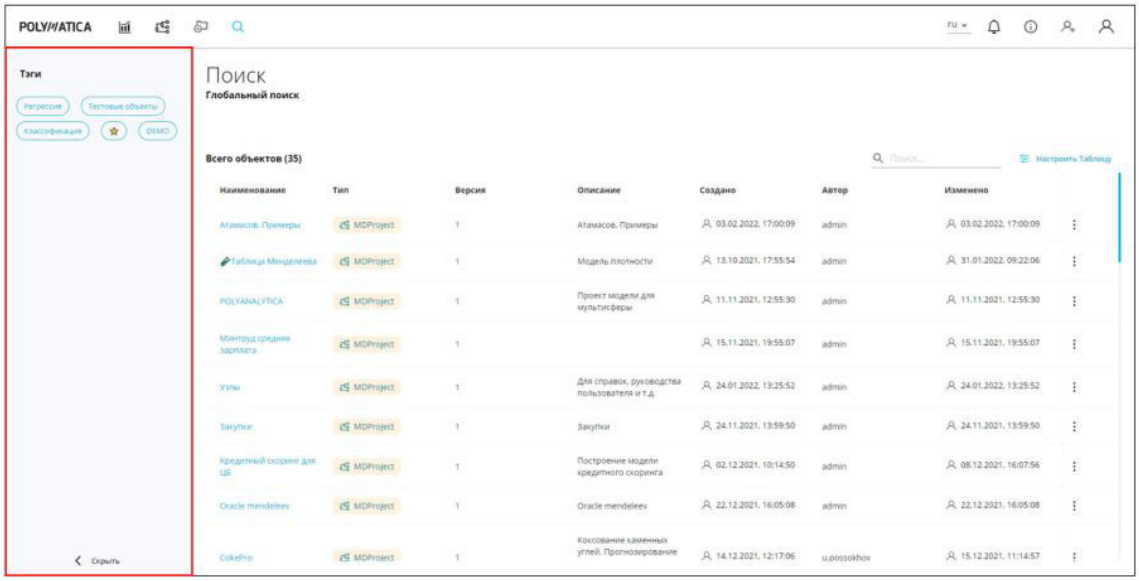


Рисунок 228 Боковая панель со списком тэгов

6.2.1. Добавление тэга

Для задания объекту тэга необходимо:

- В строке с интересующим объектом выбрать меню в виде трех вертикальных точек ⋮ и в раскрывшейся панели выбрать пункт **Работа с Тэгами.**
- В открывшемся окне **Работа с Тэгами** в строку ввода ввести тэг и выбрать кнопку **Добавить**.

6.2.2. Фильтрация при помощи тэгов

Для фильтрации объектов при помощи тэгов необходимо:

- Раскрыть левую боковую панель.
- В списке с тэгами найти интересующий и выбрать его (можно выбрать несколько тэгов у одного объекта).

- Таблица с доступными объектами отсортируется в соответствии с выбранными тэгами.

6.3. Связанные объекты

Помимо прочего на экране глобального поиска можно узнать связь между объектами. Для этого необходимо:

- В строке с интересующим объектом выбрать меню ⋮ и в раскрывшейся панели выбрать пункт **Связанные Объекты.**
- В открывшемся окне содержится список со связанными объектами. При выборе наименования объекта из данного списка откроется соответствующий раздел Модуля с выбранным объектом для дальнейшей работы.