

программная часть программно-аппаратного комплекса

«ЭВИРИС»

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

Правообладатель: Общество с ограниченной
ответственностью «ГК «Иннотех» (ООО «ГК «Иннотех»)

Место нахождения и адрес правообладателя: 125167, г.
Москва, вн.тер.г. муниципальный округ Аэропорт, пр-кт
Ленинградский, д. 36, стр. 41, помещ. 23

СОДЕРЖАНИЕ

1. Термины и сокращения	3
2. Общие сведения	4
3. Права доступа.....	6
3.1. Пользовательские роли	6
3.2. Матрица совмещения ролей	6
3.3. Аутентификация пользователей.....	6
3.3.1. Начало работы с программой (вход в систему)	6
3.3.2. Завершение работы с программой (выход из системы)	8
4. Основные операции	10
4.1. Работа с проектами.....	10
4.1.1. Просмотр списка проектов.....	10
4.1.2. Создание проекта	12
4.1.3. Просмотр расширенных сведений о проекте (экран «Проект»).....	13
4.1.4. Удаление проекта.....	15
4.2. Работа с наборами данных.....	16
4.2.1. Просмотр списка наборов данных в проекте	16
4.2.2. Просмотр расширенных сведений о наборе данных (экран «Набор данных»)	18
4.3. Работа с пайплайнами подготовки.....	20
4.3.1. Подбор ключевых полей для объединения данных.....	20
4.3.2. Создание пайплайна подготовки	22
4.3.3. Просмотр перечня пайплайнов в проекте.....	25
4.3.4. Просмотр расширенных сведений о пайплайне (экран «Пайплайн»)	27
4.4. Работа с ML-моделями.....	28
4.4.1. Обучение ML-модели	28
4.4.2. Просмотр сведений о модели проекта	30
4.4.3. Применение модели в batch-режиме	31
4.4.4. Удаление модели.....	32
4.5. Просмотр отчетов по операциям в проекте	33
4.6. Действия, осуществляемые вручную.....	36
4.6.1. Загрузка наборов данных	36
4.6.1.1. Загрузка наборов данных для обучения.....	36
4.6.1.2. Загрузка наборов данных для применения	37
4.6.2. Действия, выполняемые вручную администратором программы	38

1. Термины и сокращения

Термины и сокращения, используемые в контексте настоящего документа:

Термин	Определение
Batch-применение	Применение модели в режиме, при котором прогнозирование или анализ выполняется не для отдельных экземпляров (онлайн-режим), а для группы входных данных (пакета/батча) одновременно
Модель машинного обучения	Математический алгоритм, который обучается на наборах данных для выполнения конкретной задачи, такой как классификация или регрессия, без явного программирования. Файл с моделью машинного обучения формируется в результате обучения в Пайплайне AUTO ML
Пайплайн AUTO ML	Настраиваемый в части параметров обучения и источников данных, сохраняемый, корректируемый, переиспользуемый алгоритм в системе, который предназначен для обучения моделей машинного обучения в автоматическом режиме, а также применения модели в batch-режиме
Пайплайн подготовки	Настраиваемый, сохраняемый, переиспользуемый алгоритм в системе, который предназначен для обработки наборов данных, подаваемых на вход алгоритма с сохранением результирующего набора данных на выходе
Подбор ключевых полей	Процесс автоматического подбора самых "ближайших" ключевых полей в 2-ух наборах данных для дальнейшего возможного их объединения
Проект	Выделенное рабочее пространство в программе, предназначенное для взаимодействия участников проекта в задачах: объединения данных, обучения моделей машинного, применения модели для получения прогнозов, имеющих бизнес-ценность для участников проекта
Организация-участник программы	Организация, заключившая договор с Владелец программы на работу с функциями программы
Пользователь	Лицо из организации-участника, имеющее учетную запись в программе с определенными правами доступа
Кросс-валидация	Метод оценки модели машинного обучения путем разбиения данных на несколько подмножеств или блоков (фолдов) данных с последующим обучением и проверкой модели на этих подмножествах
Фолды	Части, на которые разбиваются данные для кросс-валидации

2. Общие сведения

Программная часть программно-аппаратного комплекса «Эвирис» (далее – программа) предназначена для автоматизированного обучения (AutoML) и последующего применения моделей машинного обучения. При этом пользователи на всех этапах использования программы имеют доступ исключительно к структуре данных и метаданным, не имея доступа непосредственно к наборам данных, используемых для машинного обучения и применения ML-моделей.

В интерфейсе программы доступны следующие группы операций (действия):

- аутентификация пользователей:
 - начало работы с программой (вход в систему);
 - завершение работы с программой (выход из системы)
- работа с проектами:
 - просмотр списка проектов;
 - просмотр расширенных сведений о проекте;
 - создание проекта;
 - удаление проекта;
- работа с наборами данных:
 - просмотр списка наборов данных
 - просмотр расширенных сведений о наборе данных
 - загрузка данных для обучения (выполняется не через UI)
 - загрузка данных для применения (выполняется не через UI)
- работа с пайплайнами подготовки:
 - просмотр перечня пайплайнов в проекте;
 - подбор ключевых полей для объединения данных;
 - создание пайплайна подготовки;
 - просмотр расширенных сведений о пайплайне;
- работа с ML-моделями:
 - просмотр сведений о модели проекта;
 - обучение ML-модели;
 - применение модели в batch-режиме;
 - удаление модели;
- просмотр отчетов по операциям в проекте.

В рамках текущей версии программы часть действий, необходимых для работы программы, осуществляется вручную администратором. Такие действия описаны отдельно в п. 4.6 настоящего документа и документе «Руководство администратора».

Для начала работы пользователя в программе должны быть выполнены следующие условия:

1. Компоненты программы развернуты и настроены на инфраструктуре Владельца программы и организации-участника программы, к которой относится пользователь, соответственно.
2. В программе зафиксирована информация об организации-участнике программы, к которой относится пользователь.
3. Для пользователя в программе создана учетная запись. Логин и пароль для доступа к интерфейсу программы предоставлены пользователю.
4. Пользователю назначена хотя бы одна роль в соответствии с ролевой моделью, реализованной в программе.
5. Интерфейс программы доступен для входа и все компоненты работают в штатном режиме.

3. Права доступа

Для разграничения прав доступа в программе используется ролевая модель.

3.1. Пользовательские роли

Для контроля доступа к операциям (действиям) в программе используются следующие роли пользователей в проекте:

№	Наименование роли	Описание роли	Кому может быть предоставлена роль	Описание доступных действий и полномочий
1	Владелец проекта	Управление проектом внутри организации	Пользователь	• подключение к программе через UI; • управление проектом, контроль безопасности посредством возможности блокирования сущностей проекта
2	Разработчик модели	Разработка модели в проекте	Пользователь	• обработка данных; • обучение модели; • применение модели
3	Поставщик данных	Управление данными, используемыми в проекте	Пользователь	• загрузка данных; • удаление данных
4	Пользователь модели	Использование готовых моделей	Пользователь	• применение модели; • загрузка данных

3.2. Матрица совмещения ролей

Ограничений на совмещение ролей не накладывается.

3.3. Аутентификация пользователей

3.3.1. Начало работы с программой (вход в систему)

Для пользователей вход в систему осуществляется по ссылке: <https://<адрес сервера>/login>.

Адрес сервера может иметь вид:

- localhost:<порт>
- <IP-адрес сервера>:<порт>
- <доменное имя>

Вход в системы начинается с начальной страницы с окном «Авторизация» (Рисунок 3.3.1.1), на которой расположены поля для ввода логина и пароля, кнопка «Войти» и ссылка для перехода к экрану с контактами администратора.

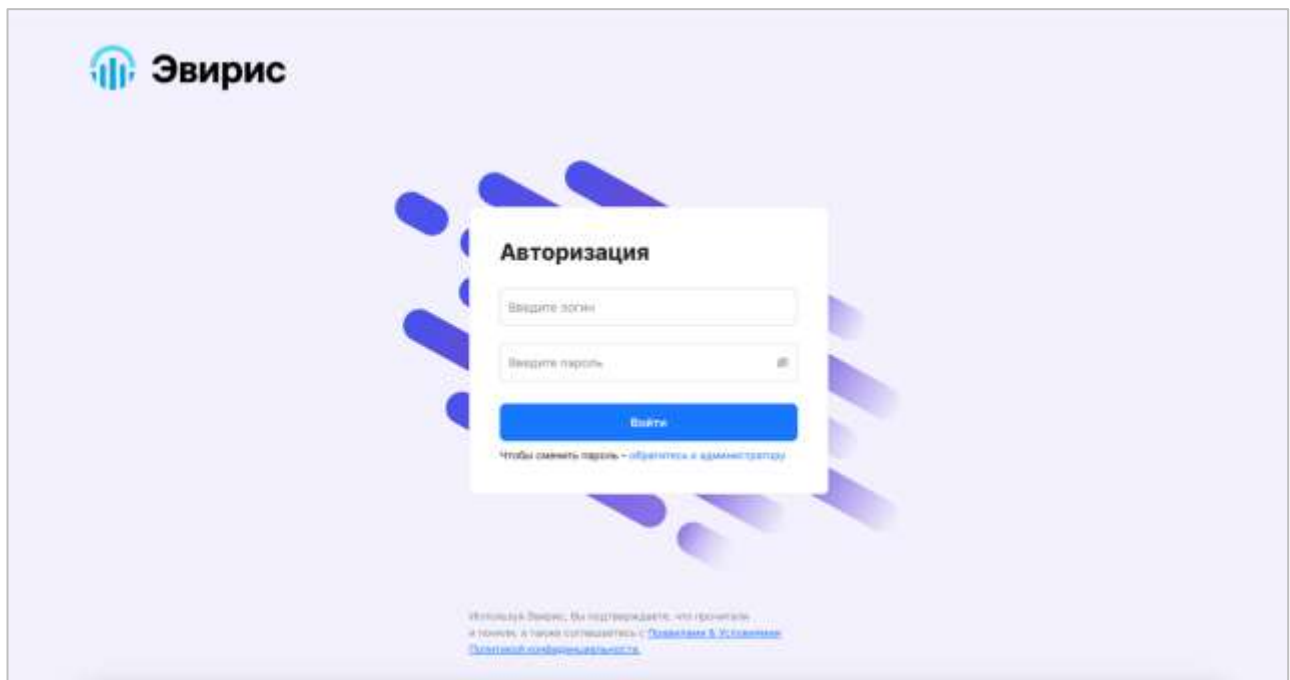


Рисунок 3.3.1.1 – Начальная страница

Для входа в систему необходимо ввести логин и пароль и нажать на кнопку «Войти» (либо клавишу Enter). После удачного входа в систему пользователь переадресуется на главную страницу (Рисунок 3.3.1.2).

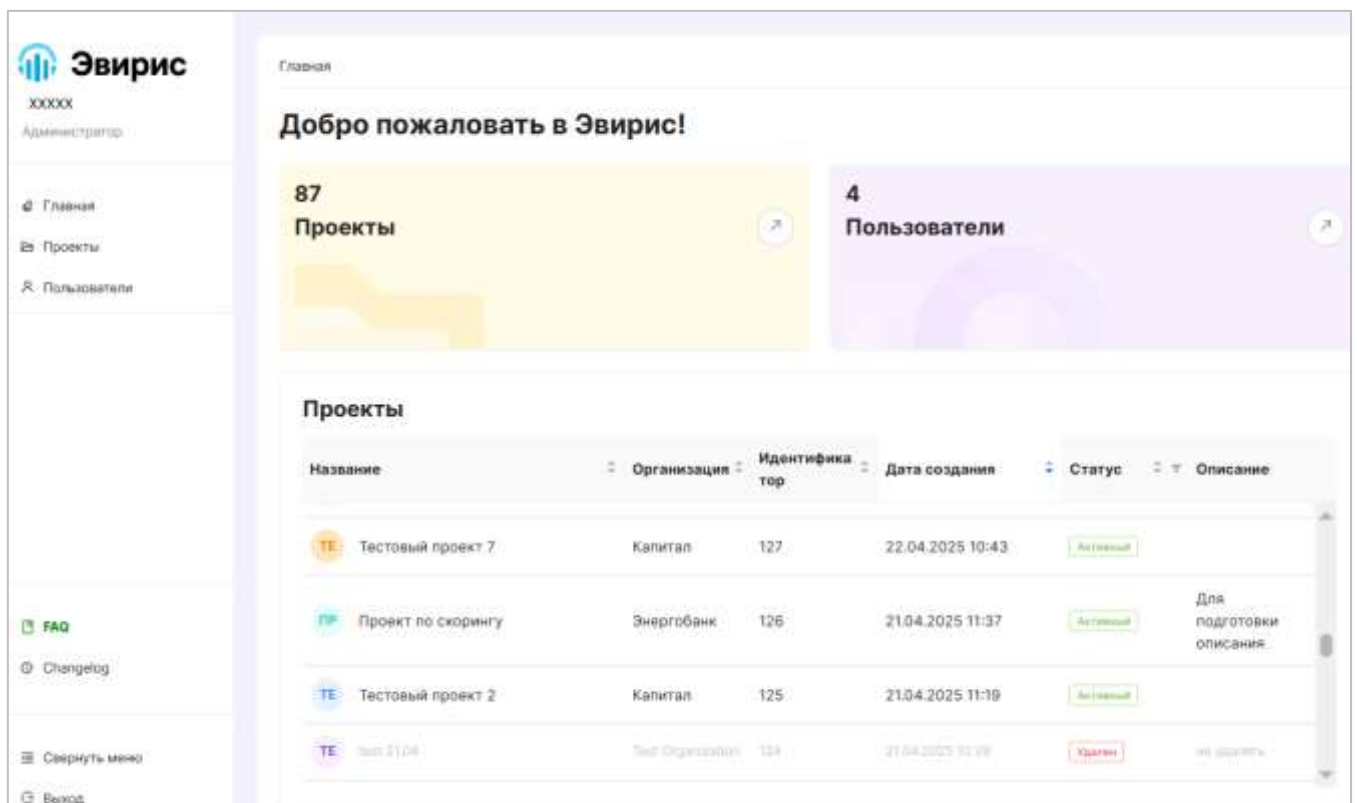


Рисунок 3.3.1.2 – Главная страница

На главной странице расположено два виджета «Проекты» и «Пользователи», а также таблица с перечнем проектов, созданных в программе.

На главной странице доступны следующие операции:

- переход на страницу «Проекты» по щелчку на виджете «Проекты»;
- переход на страницу «Пользователи» по щелчку на виджете «Пользователи»;
- переход на страницу «Проект» по щелчку на строке таблицы с проектами.

Если данные для входа будут введены неверно, то в окне «Авторизация» отобразится ошибка входа (Рисунок 3.3.1.3).

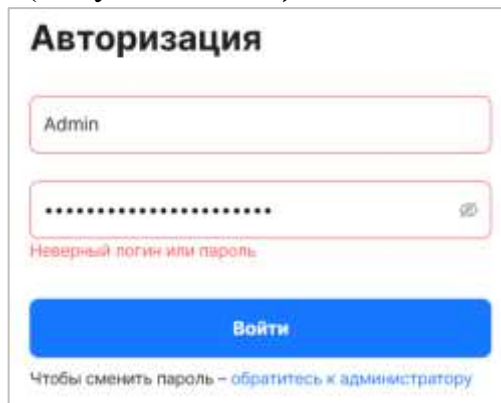


Рисунок 3.3.1.3 – Ошибка входа

При нажатии на ссылку «обратитесь к администратору» появляется экран с электронной почтой администратора системы (Рисунок 3.3.1.4). На указанный адрес электронной почты необходимо направлять обращения в случае возникновения ошибок в работе программы или иных вопросов.

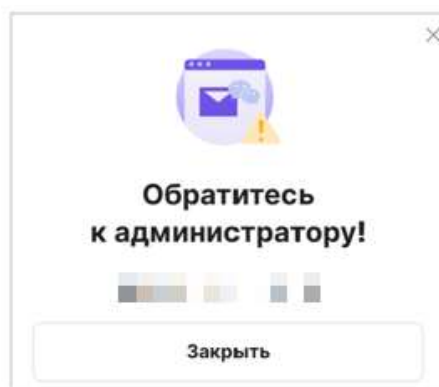


Рисунок 3.3.1.4 – Окно с контактами администратора

3.3.2. Завершение работы с программой (выход из системы)

Для завершения сессии (сеанса) работы в интерфейсе программы необходимо выбрать пункт «Выход» на боковой панели (Рисунок 3.3.2.1), а затем подтвердить выход в открывшемся информационном окне (Рисунок 3.3.2.2).

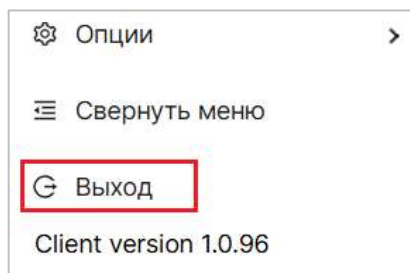


Рисунок 3.3.2.1 – Выбор пункта меню «Выход»

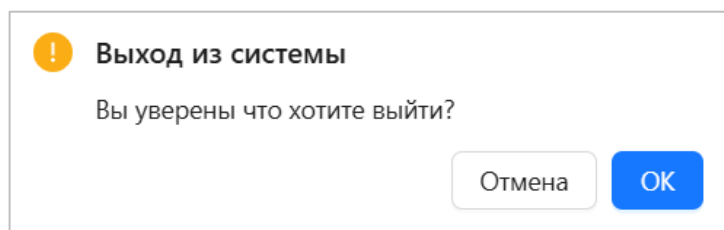


Рисунок 3.3.2.2 – Информационное окно для подтверждения выхода из системы

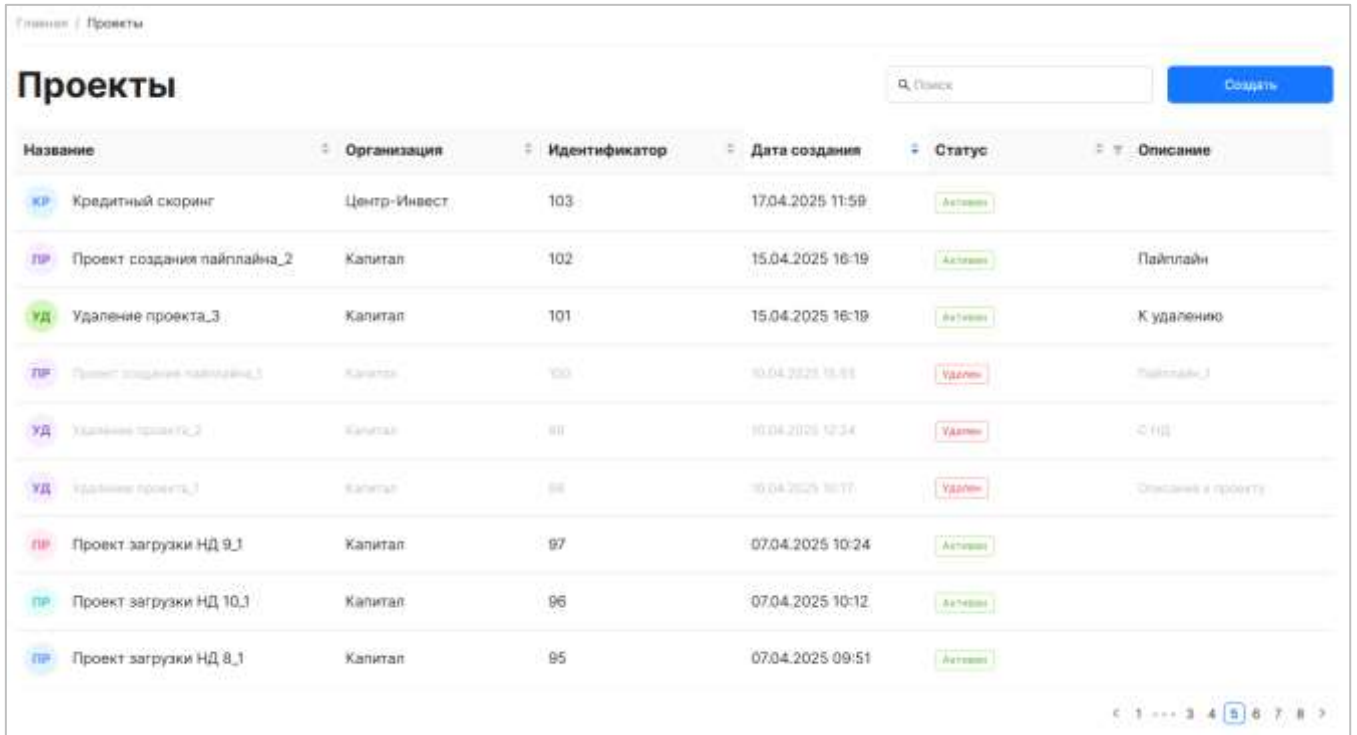
4. Основные операции

4.1. Работа с проектами

4.1.1. Просмотр списка проектов

Для просмотра списка проектов реализован экран «Проекты» (Рисунок 4.4.1.1). К экрану можно перейти, выбрав соответствующий пункт «Проекты» на боковой панели или виджет «Проекты» на главной странице.

Также список проектов отображается на главной странице.



Название	Организация	Идентификатор	Дата создания	Статус	Описание
КР Кредитный скоринг	Центр-Инвест	103	17.04.2025 11:59	Активен	
ПР Проект создания пайплайна_2	Капитал	102	15.04.2025 16:19	Активен	Пайплайн
УД Удаление проекта_3	Капитал	101	15.04.2025 16:19	Активен	К удалению
ПР Проект создания пайплайна_1	Капитал	100	10.04.2025 16:51	Удален	Пайплайн_1
УД Удаление проекта_2	Капитал	99	10.04.2025 12:34	Удален	С 103
УД Удаление проекта_1	Капитал	98	10.04.2025 10:17	Удален	Описания в проекте
ПР Проект загрузки НД 9_1	Капитал	97	07.04.2025 10:24	Активен	
ПР Проект загрузки НД 10_1	Капитал	96	07.04.2025 10:12	Активен	
ПР Проект загрузки НД 8_1	Капитал	95	07.04.2025 09:51	Активен	

Рисунок 4.4.1.1 – Экран «Проекты»

Список проектов представляет собой таблицу, в строках которой размещены сведения о проектах, созданных в программе. В столбцах этой таблицы представлена информация о следующих параметрах проекта:

- название;
- название организации-владельца;
- идентификатор;
- дата и время создания;
- статус (может принимать значения «Активный» и «Удален»);
- описание.

По умолчанию список проектов отсортирован по полю «Дата создания» в порядке убывания (новые записи сверху). Сортировку списка проектов можно изменить с помощью переключателей в заголовках столбцов таблицы (Рисунок 4.4.1.2) или щелчком левой кнопки мыши в области заголовка столбца таблицы.

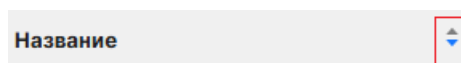


Рисунок 4.4.1.2 – Переключатели для изменения сортировки

Также предусмотрена возможность фильтрации списка проектов по текущему статусу. Для установки фильтра предусмотрен переключатель в заголовке столбца «Статус» (Рисунок 4.4.1.3). В появившемся окне необходимо выбрать требуемое значение статуса и нажать на кнопку «Применить».

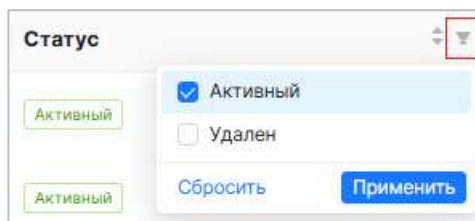


Рисунок 4.4.1.3 – Установка фильтра по статусу проекта

Фильтр можно сбросить, нажав на соответствующую кнопку в окне настройки фильтра.

На экране «Проекты» реализована возможность регистронезависимого поиска по вхождению поискового значения в значения параметров проекта. Для начала поиска пользователю необходимо указать значение для поиска в поисковой строке (Рисунок 4.4.1.4).



Рисунок 4.4.1.4 – Поисковая строка

По мере ввода значения в поисковую строку перечень отображаемых проектов будет корректироваться – будут отображаться записи, соответствующие текущему поисковому запросу.

В текущей версии программы реализован поиск по столбцам «Название», «Организация», «Идентификатор», «Статус», «Описание». Поиск по столбцу «Дата создания» не осуществляется.

Внизу экрана «Проекты» предусмотрена пагинация для переключения между страницами списка проектов. Пользователь может выбрать конкретный номер страницы или воспользоваться переключателями «Вперед», «Назад».

Если в таблице с перечнем проектов навести курсор на строку с проектом и щелкнуть левой кнопкой мыши, то откроется карточка проекта (экран «Проект») с подробной информацией о нем (см. п.4.1.3).

Если на экране «Проекты» нажать на кнопку «Создать» появится экран «Создание нового проекта» (см. п.4.1.2).

4.1.2. Создание проекта

Для создания проекта в окне «Создание нового проекта» (Рисунок 4.1.2.1) необходимо:

1. Ввести вручную название проекта (обязательный параметр, не менее 10 и не более 50 символов).
2. Выбрать из списка организацию-владельца проекта (обязательный параметр).
3. Ввести вручную описание проекта (необязательный параметр, не более 255 символов).
4. Нажать на кнопку «Создать».

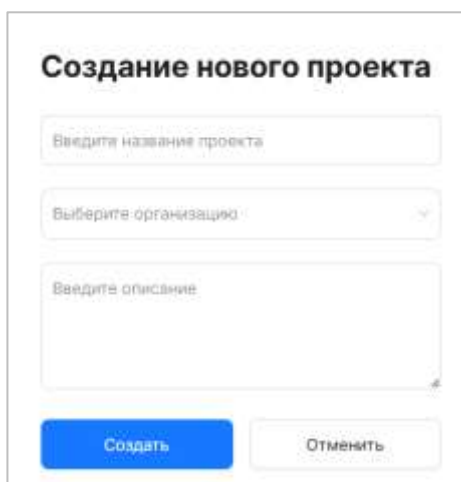


Рисунок 4.1.2.1 – Окно для создания нового проекта

Наименование проекта должно быть уникальным в рамках выбранной организации-владельца. Если попытаться создать проект с неуникальным названием, то появится уведомление об ошибке: «Проект с таким именем уже существует в рамках данной организации».

Когда запрос на создание проекта пройдет первичные проверки и будет передан на исполнение, в правом нижнем углу экрана появится уведомление: «Запрос успешно принят в обработку».

После успешного создания проекта в системе в правом нижнем углу экрана появится уведомление: «{Название проекта} успешно был создан».

Если при создании проекта возникает ошибка, то в правом нижнем углу экрана появляется уведомление: «{Название проекта} не удалось создать. Попробуйте еще раз. Если ошибка повторится – обратитесь к администратору».

После успешного создания новый проект будет отображаться в списке проектов.

4.1.3. Просмотр расширенных сведений о проекте (экран «Проект»)

Экран с подробной информацией о проекте (экран «Проект», Рисунок 4.1.3.1) открывается при щелчке левой кнопкой мыши на соответствующей строке перечня проектов. На экране «Проект» отображаются:

- название проекта;
- наименование организации-владельца проекта;
- идентификатор проекта;
- дата создания проекта;
- текущий статус проекта;
- описание проекта;
- перечень участников проекта;
- сведения о модели в проекте (Рисунок 4.1.3.2);
- сведения о наборах данных, доступных в проекте (Рисунок 4.1.3.3);
- сведения о пайплайнах, созданных в проекте (Рисунок 4.1.3.4);
- кнопка «Удалить проект».

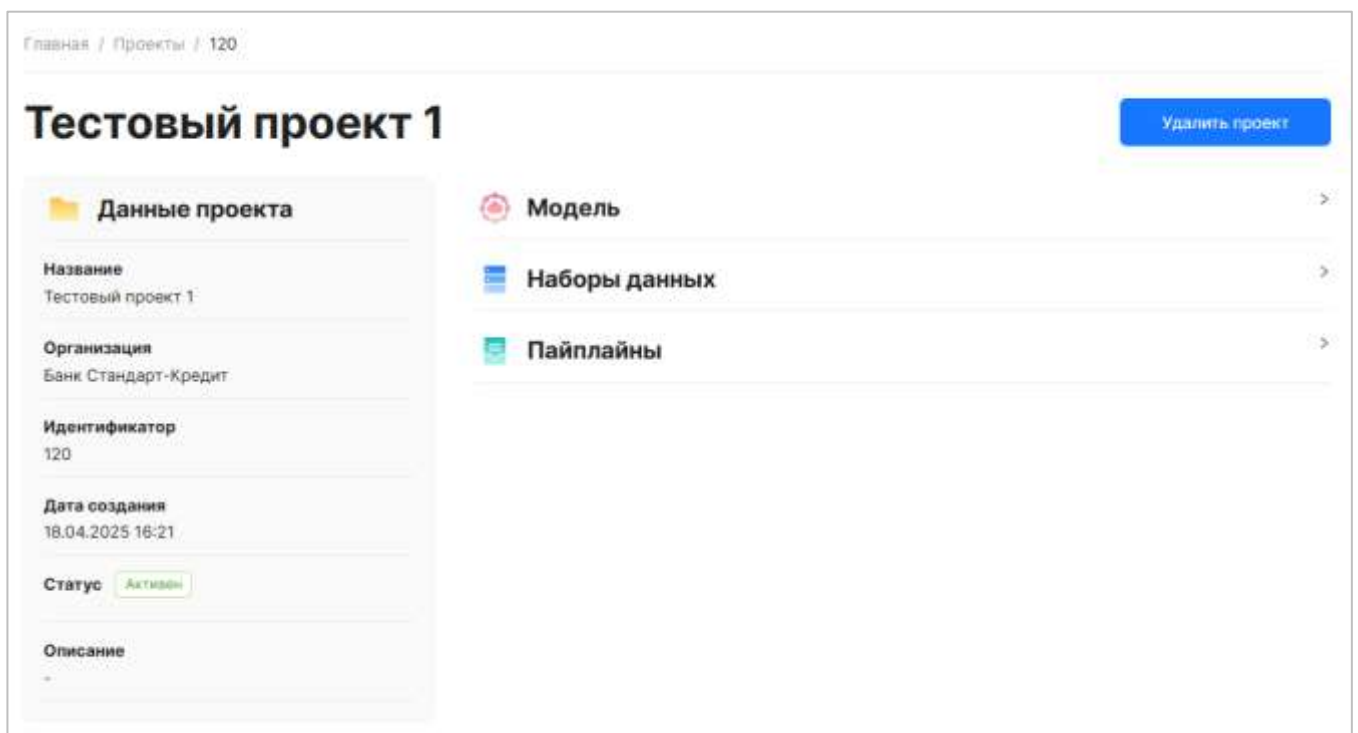


Рисунок 4.1.3.1 – Экран «Проект»

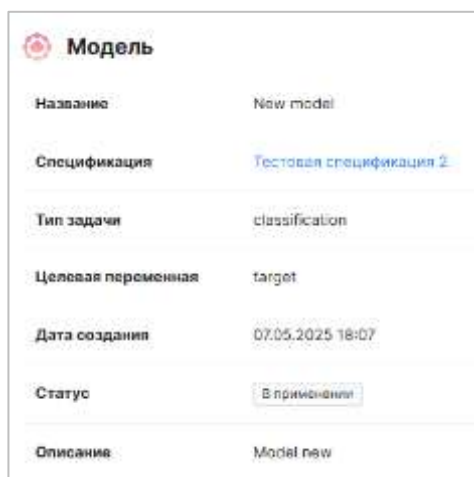


Рисунок 4.1.3.2 – Экран «Проект», раздел «Модель»

Название	Спецификация	Тип	Статус	Дата загрузки
6.datasetone.score1_2025-04-30_13-22-42	Спецификация 126	ДАННЫЕ ДЛЯ ОБУЧЕНИЯ	Загружен. Готов к применению	30.04.2025 13:24
project_126.datasetthree	Спецификация 122	ДАННЫЕ ДЛЯ ОБУЧЕНИЯ	Загружен. Готов к применению	29.04.2025 16:12
project_126.datasettwo	Спецификация 108	ДАННЫЕ ДЛЯ ОБУЧЕНИЯ	Загружен. Готов к применению	28.04.2025 14:27
Основной НД	Спецификация 107	ДАННЫЕ ДЛЯ ОБУЧЕНИЯ	Загружен. Готов к применению	28.04.2025 14:26

Рисунок 4.1.3.3 – Экран «Проект», раздел «Наборы данных»

Название	Тип	Дата создания	Входные спецификации	Выходная спецификация	Статус
pipeline_name	PREPARE	12.05.2025 18:29	Спецификация 32, Спецификация 34	Спецификация 141	Сохранен
Test_Pipeline_130	PREPARE	28.04.2025 16:39	Спецификация 112, Спецификация 113	Спецификация 114	Сохранен
automi_pipeline_classification_target_2025-04-28_14-06-09	AUTOML	28.04.2025 17:06	Спецификация 114	-	Удален

Рисунок 4.1.3.4 – Экран «Проект», раздел «Пайплайны»

Если в таблице с перечнем наборов данных или пайплайнов навести курсор на строку и щелкнуть левой кнопкой мыши, то откроется карточка набора данных (экран «Набор данных») или пайплайна (экран «Пайплайн») соответственно. По артефактам проекта также можно переключаться с помощью меню боковой панели (Рисунок 3.2.3.5). Для просмотра определенных сведений о проекте необходимо выбрать щелчком левой кнопки мыши соответствующий пункт меню.

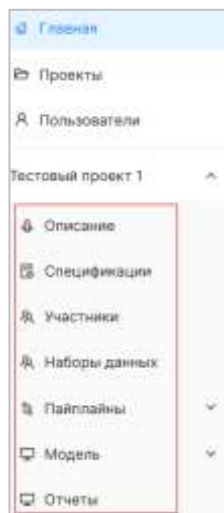


Рисунок 4.1.3.5 – Меню проекта в боковой панели

4.1.4. Удаление проекта

Удаление проекта можно инициировать из карточки соответствующего проекта (экран «Проект») нажатием кнопки «Удалить проект». После чего появится окно для подтверждения удаления (Рисунок 4.1.4.1).

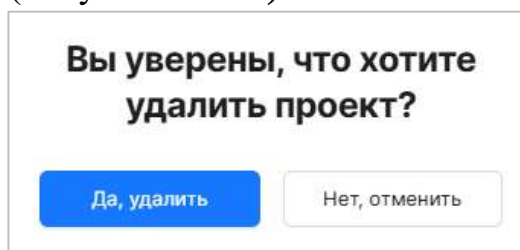


Рисунок 4.1.4.1 – Первое подтверждение удаления проекта

Если нажать «Нет, отменить», то процесс удаления будет остановлен.

Если нажать «Да, удалить», то появится второе окно для подтверждения с удалением с дополнительной информацией о каскадном удалении всех сущностей проекта (Рисунок 4.1.4.2).

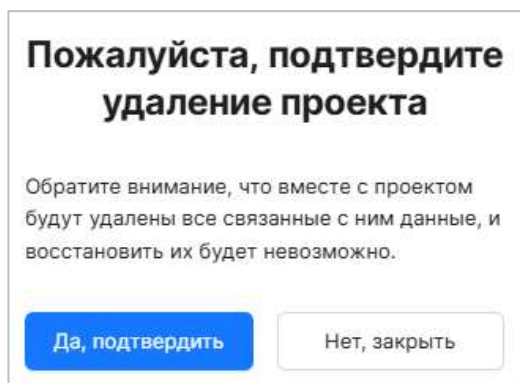


Рисунок 4.1.4.2 - Второе подтверждение удаления проекта

Если нажать «Нет, закрыть», то процесс удаления будет остановлен.

Если нажать «Да, подтвердить», то стартует процесс удаление проекта и всех его сущностей. Помимо проекта будут удалены наборы данных, спецификации и результаты профилирования, пайплайны подготовки данных и пайплайны AutoML, ML-модель, результаты применения модели.

После успешного выполнения первого этапа удаления пользователю будет направлено уведомление с текстом: «Первый этап удаления проекта {Наименование проекта} завершен. Стартовало физическое удаление данных проекта».

После полного удаления пользователю будет направлено уведомление с текстом: «{Наименование проекта} был успешно удален». Проект и все его содержимое (наборы данных, пайплайны и т.д.) получают статус «Удален». Восстановление удаленных объектов не предусмотрено.

Если при удалении проекта и его наполнения возникают ошибки, то пользователю отображается соответствующее уведомление.

4.2. Работа с наборами данных

4.2.1. Просмотр списка наборов данных в проекте

Просмотр списка наборов данных, доступных в проекте, возможен на экране «Проект» (Рисунок 3.2.3.3) или на отдельном экране «Наборы данных», который можно открыть через соответствующее меню боковой панели (Рисунок 4.2.1.1, 4.2.1.2).

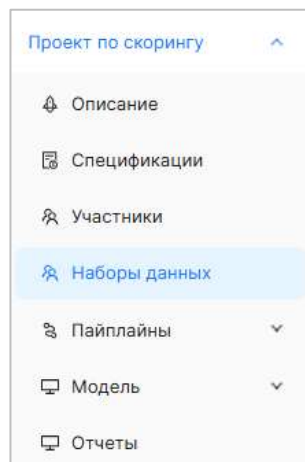


Рисунок 4.2.1.1 – Пункт меню «Наборы данных» на боковой панели

Главная / Проекты / 126 / Наборы данных

Наборы данных

Поиск

Название	Спецификация	Тип	Статус	Дата загрузки
preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-58-29	preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-58-29	Данные для обучения	Загружен. Готов к применению	05.05.2025 19:00
preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-52-21	preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-52-21	Данные для обучения	Загружен. Готов к применению	05.05.2025 18:04
preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-45-39	preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-45-39	Данные для обучения	Загружен. Готов к применению	05.05.2025 18:47
preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-02-44	preparing_dataset_regression_project_126.dataseton e.score1_2025-05-05-19-02-44	Данные для обучения	Загружен. Готов к применению	05.05.2025 14:04
preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-16-03-21	preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-16-03-21	Данные для обучения	Загружен. Готов к применению	30.04.2025 16:06
preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-16-03-49	preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-16-03-49	Данные для обучения	Загружен. Готов к применению	30.04.2025 16:04
preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-13-22-45	preparing_dataset_regression_project_126.dataseton e.score1_2025-04-30-13-22-45	Данные для обучения	Загружен. Готов к применению	30.04.2025 13:24
project_126.dataseton	project_126.dataseton	Данные для обучения	Загружен. Готов к применению	29.04.2025 16:13
project_126.dataseton	project_126.dataseton	Данные для обучения	Загружен. Готов к применению	29.04.2025 14:27
Оскаковой ИД	project_126.dataseton	Данные для обучения	Загружен. Готов к применению	29.04.2025 14:26

Рисунок 4.2.1.2 – Экран «Наборы данных»

Экран «Наборы данных» содержит следующую информацию о каждом наборе данных:

- название;
- название спецификации;
- тип данных (возможные значения: Данные для обучения, Данные для применения);
- статус (возможные значения: «Загружен», «Готов к применению», «Удален»);
- дата и время создания.

По умолчанию список наборов данных отсортирован по полю «Дата загрузки» в порядке убывания (новые записи сверху). Сортировку списка наборов данных можно изменить с помощью переключателей в заголовках столбцов таблицы или щелчком левой кнопкой мыши в области заголовка столбца таблицы.

Также предусмотрена возможность фильтрации списка наборов данных по текущему статусу. Для установки фильтра предусмотрен переключатель в заголовке столбца «Статус».

На экране «Наборы данных» реализована возможность регистронезависимого поиска по вхождению поискового значения в значения параметров набора данных. Для начала поиска пользователю необходимо указать значение для поиска в поисковой строке.

По мере ввода значения в поисковую строку перечень отображаемых наборов данных будет корректироваться – будут отображаться записи, соответствующие текущему поисковому запросу.

На данный момент реализован поиск по столбцам «Название», «Спецификация», «Тип», «Статус».

Внизу экрана «Наборы данных» предусмотрена пагинация для переключения между страницами списка. Пользователь может выбрать конкретный номер страницы или воспользоваться переключателями «Вперед», «Назад».

Сортировка, фильтрация, поиск, пагинация на экране «Наборы данных» реализованы аналогично соответствующему функционалу на экране «Проекты» (см. п. 4.4.1).

Если в таблице с перечнем наборов данных проекта навести курсор на строку с набором данных и щелкнуть левой кнопкой мыши, то откроется карточка набора данных (экран «Набор данных») с подробной информацией о нем (см. п. 4.2.2).

Если на экране «Наборы данных» выбрать два набора данных и нажать на кнопку «Подобрать ключевые поля» выполняется подбор ключевых полей для объединения выбранных пользователем наборов данных (см. п. 4.3.1).

4.2.2. Просмотр расширенных сведений о наборе данных (экран «Набор данных»)

Экран с подробной информацией о наборе данных (экран «Набор данных», Рисунки 4.2.2.1, 4.2.2.2) открывается при щелчке левой кнопкой мыши на соответствующей строке перечня наборов данных. На экране «Набор данных» отображаются:

- название набора данных;
- тип данных;
- дата\время первой загрузки;
- дата\время последнего обновления;
- наименование спецификации;
- статус;
- описание;
- вкладка «Профиль данных»;
- вкладка «Атрибуты».

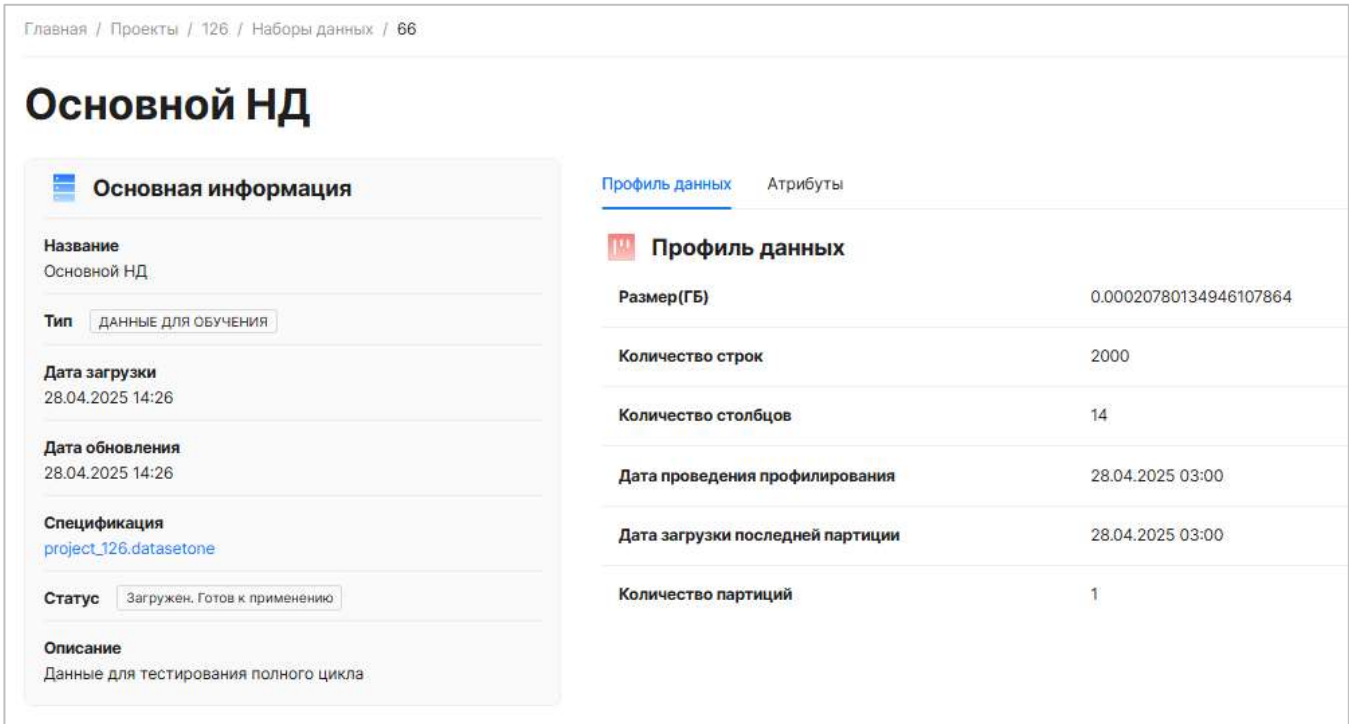


Рисунок 4.2.2.1 – Экран «Набор данных» с активной вкладкой «Профиль данных»

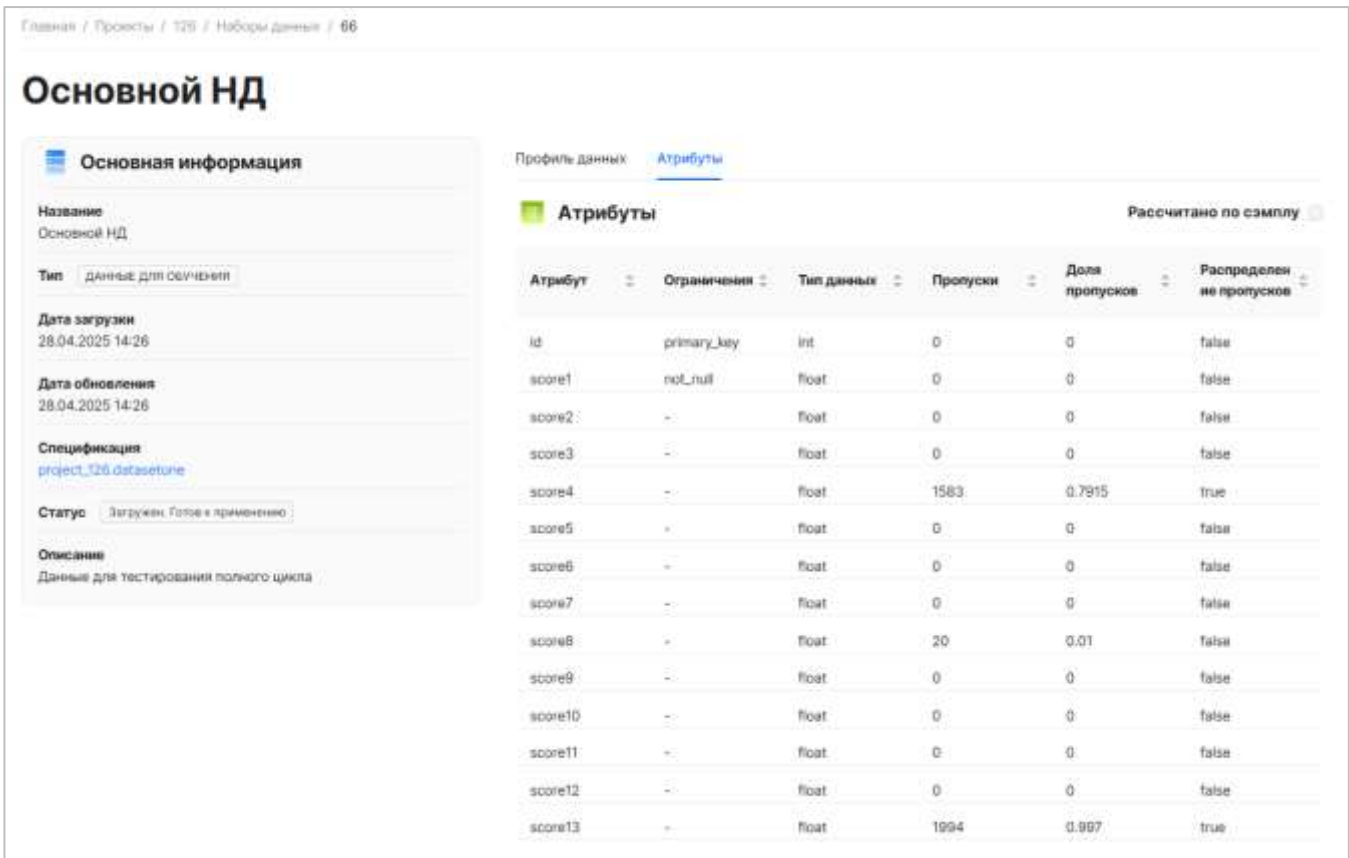


Рисунок 4.2.2.2 - Экран «Набор данных» с активной вкладкой «Атрибуты»

На вкладке «Профиль данных» отображаются:

- текущий размер набора данных в гигабайтах;

- текущее количество строк в наборе данных;
- текущее количество столбцов в наборе данных;
- дата и время проведения последнего профилирования;
- дата и время загрузки последней партии набора данных;
- количество партий.

При расчете данных метрик учитывается весь накопленный объем по набору данных на момент расчета. Т. е. если в течении некоторого периода производится дозагрузка данных по определенному набору данных, то на вкладке «Профиль данных» отображаются актуальные сведения по накопленному за это период объему данных. На вкладке «Атрибуты» в табличной форме для каждого атрибута указывается:

- наименование;
- ограничения (здесь указывается является ли атрибут первичным ключом (primary_key), обязательным для заполнения (not_null) или должно обладать уникальным значением (unique));
- тип данных;
- пропуски в каждом признаке/атрибуте (столбец «Пропуски»). Рассчитывается количество незаполненных строк (учитываются значения типа null, NaN);
- доля пропусков в каждом признаке/атрибуте (столбец «Доля пропусков»). Рассчитывается доля пустых строк в анализируемом общем количестве строк.
- распределение пропусков в каждом признаке/атрибуте (столбец «Распределение пропусков»). Определяется $\geq 50\%$ пропусков в атрибуте. Если условие выполняется, то значение метрики - true, если не выполняется – false.

На вкладке «Атрибуты» также присутствует недоступный для редактирования пользователем признак «Рассчитано по сэмплу». Этот признак заполняется автоматически и иллюстрирует были рассчитаны метрики по всему накопленному объему набора данных или по выделенному сэмплу. Если на момент расчета метрик набор данных содержит в целом > 5 млн строк и количество атрибутов > 50 , то для расчета метрик выделяется сэмпл, а признак «Рассчитано по сэмплу» имеет значение «Истина».

4.3. Работа с пайплайнами подготовки

4.3.1. Подбор ключевых полей для объединения данных

Подбор ключевых полей для последующего объединения данных можно выполнить на странице «Наборы данных» (экран «Наборы данных»). Для перехода на данную страницу необходимо выбрать проект, а затем выбрать на боковой панели пункт меню «Наборы данных» или пункт меню «Пайплайны»\«Подбор ключевых полей».

Для запуска процесса подбора ключевых полей на странице «Наборы данных» необходимо выбрать два набора данных и нажать кнопку «Подобрать ключевые поля» (Рисунок 4.3.1.1).

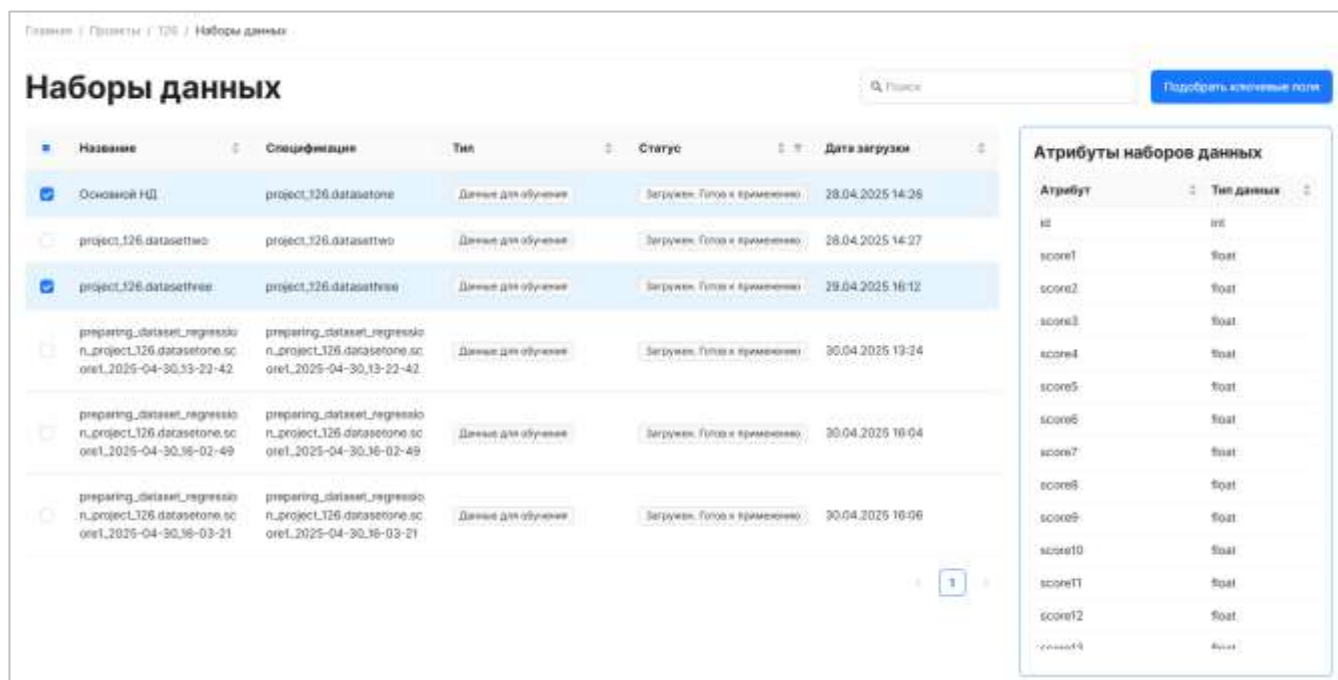


Рисунок 4.3.1.1 – Выбор наборов данных для подбора ключевых полей

После выбора набора данных в правой части экрана список его атрибутов с указанием типа данных (если выбрать два набора данных, то список содержит атрибуты обоих наборов).

Когда запрос на подбор ключевых полей проходит первичные проверки и передается на исполнение, в правом нижнем углу экрана отображается уведомление: «Запрос успешно принят в обработку».

После успешного завершения подбора ключевых полей в правом нижнем углу экрана отображается уведомление: «Подбор ключевых полей» может быть найден в меню «Отчеты» (см. п. 4.5).

Для перехода на страницу с перечнем отчетов необходимо выбрать пункт меню «Отчеты» на боковой панели. Отчеты по результатам подбора ключевых полей имеют тип «match_keys» и представляет собой json-файл, который необходимо скачать для просмотра.

Отчет содержит вероятность соответствия пар полей из указанных пользователем наборов данных (Рисунок 4.3.1.2).

В отчете используются следующие обозначения:

- first_field – первое поле в паре сравниваемых полей;
- second_field – второе поле в паре сравниваемых полей;
- dataset_id – идентификатор набора данных;

- dataset_name – название набора данных;
- dataset_field – название поля набора данных;
- weight – вероятность соответствия пары полей, значение может быть от 0 до 100.

```
{
  "match_keys": [
    {
      "first_field": {
        "dataset_field": "score5",
        "dataset_id": "66",
        "dataset_name": "project_126.datasetone"
      },
      "second_field": {
        "dataset_field": "score15",
        "dataset_id": "76",
        "dataset_name": "project_126.datasetthree"
      },
      "weight": 10.725720384204909
    },
    {
      "first_field": {
        "dataset_field": "score7",
        "dataset_id": "66",
        "dataset_name": "project_126.datasetone"
      },
      "second_field": {
        "dataset_field": "score15",
        "dataset_id": "76",
        "dataset_name": "project_126.datasetthree"
      },
      "weight": 10.672358591248667
    }
  ]
}
```

Рисунок 4.3.1.2 – Фрагмент отчета «Подбор ключевых полей»

4.3.2. Создание пайплайна подготовки

Создание пайплайна подготовки можно выполнить на странице (экране) «Создание пайплайна подготовки» (Рисунок 4.3.2.1). Для перехода на данную страницу необходимо выбрать проект, а затем выбрать на боковой панели пункт меню «Пайплайны»\«Создание пайплайна». Также переход к экрану «Создание пайплайна подготовки» осуществляется, если нажать на кнопку «Создать пайплайн» на странице (экране) «Пайплайны».

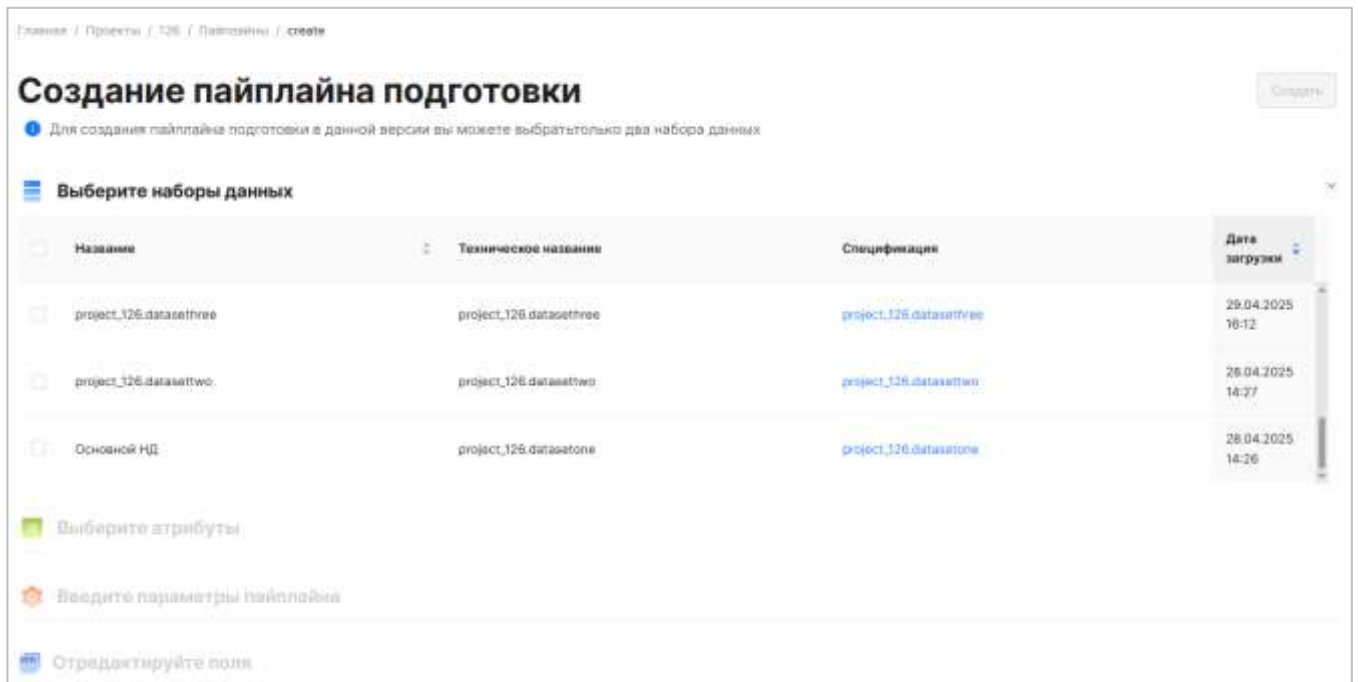


Рисунок 4.3.2.1 – Начальный вид экрана «Создание пайплайна подготовки»

Для запуска процесса создания пайплайна подготовки необходимо выполнить следующие шаги:

1. Выбрать 2 набора данных в рабочей области «Выберите наборы данных».
2. Выбрать атрибуты для объединения наборов данных в рабочей области «Выберите атрибуты» (Рисунок 4.3.2.2). Эта рабочая область станет активной после выполнения шага 1. Атрибуты необходимо выбрать для каждого из двух указанных наборов данных.

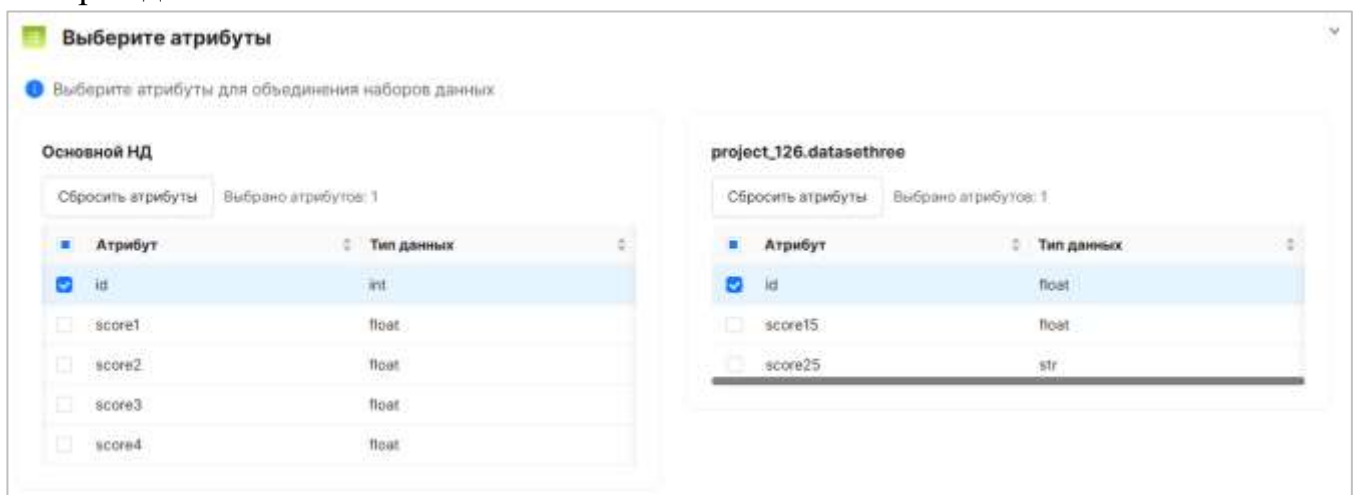


Рисунок 4.3.2.2 – Выбор атрибутов для объединения

3. Ввести название и описание пайплайна, тип задачи, целевую переменную в рабочей области «Введите параметры пайплайна» (Рисунок 4.3.2.3).

Рисунок 4.3.2.3 – Ввод параметров пайплайна

Тип решаемой задачи необходимо выбрать из списка. Допустимые значения: регрессия и классификация.

Целевую переменную необходимо выбрать из перечня атрибутов указанных наборов данных.

Указание описания пайплайна является необязательным.

4. Задать набор трансформаций с данными в рабочей области «Отредактируйте поля» (необязательно).

Набор трансформаций с данными – это набор (последовательность) инструкций, записанных в специальном формате и применяющихся к данным.

Все инструкции имеют следующий формат:

{Название функции}:{Название датасета}.{Название поля}, {Название датасета}.{Название поля}, где:

- «:» - отделяет название функции от аргументов;
- «,» - отделяет аргументы друг от друга.

Поддерживаемые функции указаны в таблице ниже.

Поддерживаемые функции трансформаций с данными:

Название функции	Количество аргументов	Описание
AVG	1	Среднее значение (по всем значениям или по сгруппированным через GROUP BY). Вычисляется путем деления суммы этих значений на число непустых значений
EXCLUDE	> 1	Поля наборов данных, которые следует исключить. По умолчанию, в результирующем наборе данных должны оказаться все поля всех используемых наборов. Если часть полей требуется исключить, то используется эта функция
GROUP BY	> 1	Группировка по нескольким полям
LOWER	1	Для полей со строковыми значениями осуществляет преобразования символов верхнего регистра в символы нижнего регистра

Название функции	Количество аргументов	Описание
MAX	1	Максимальное значение (по всем значениям или по сгруппированным через GROUP BY)
MIN	1	Минимальное значение (по всем значениям или по сгруппированным через GROUP BY)
UPPER	1	Для полей со строковыми значениями осуществляет преобразования символов нижнего регистра в символы верхнего регистра

Примеры инструкций:

AVG: tele2_calls_1_kvartal_2024.calls_count

EXCLUDE: vtb_1_kvartal_2024.temp_field, tele2_calls_1_kvartal_2024.temp_field

GROUP BY: vtb_1_kvartal_2024.month, vtb_1_kvartal_2024.year

LOWER: vtb_1_kvartal_2024.email

MAX: tele2_calls_1_kvartal_2024.call_duration

MIN: tele2_calls_1_kvartal_2024.call_duration

UPPER: vtb_1_kvartal_2024.address

5. Нажать на кнопку «Создать».

Когда запрос на создание пайплайна подготовки пройдет первичные проверки и будет передан на исполнение, в правом нижнем углу экрана появится уведомление: «Запрос успешно принят в обработку», а пользователь будет перенаправлен на страницу «Пайплайны».

После успешного завершения создания пайплайна подготовки в правом нижнем углу экрана появится уведомление: «Пайплайн подготовки успешно создан».

Созданный пайплайн подготовки можно просмотреть через список на странице (экране) «Пайплайны» (меню боковой панели Проект\Пайплайны\Созданные).

Отчет по итогам создания пайплайна можно скачать на странице «Отчеты» (меню боковой панели Проект\Отчеты) или на странице пайплайна, нажав кнопку Скачать отчет».

На странице «Отчеты» отчет по результатам создания пайплайна подготовки имеет тип «rpt» и представляет собой текстовый документ, который необходимо скачать для просмотра. Структура отчета описана в п. 4.5.

4.3.3. Просмотр перечня пайплайнов в проекте

Просмотр списка пайплайнов, доступных в проекте, возможен на экране «Проект» (Рисунок 4.1.3.4) или на отдельном экране «Пайплайны» (Рисунок 4.3.3.1), который можно открыть через меню боковой панели Проект\ Пайплайны\ Созданные.

Главная / Проекты / 139 / Пайплайны

Пайплайны

Поиск Создать пайплайн

Название	Тип	Дата создания	Входные спецификации	Выходная спецификация	Статус
automl_pipeline_classification_target_2025-05-16_09-49-31	AUTOML	16.05.2025 12:49	preparing_dataset_classification_project_139.home_credit_part1.target_2025-05-13_12-53-11	-	Сохранен
Пайплайн 139_2	PREPARE	13.05.2025 16:08	project_139.home_credit_part1, project_139.home_credit_part2	preparing_dataset_classification_project_139.home_credit_part1.target_2025-05-13-12-53-11	Сохранен
Пайплайн 139_1	PREPARE	13.05.2025 15:51	project_139.home_credit_part1, project_139.home_credit_part2	preparing_dataset_regression_project_139.home_credit_part1.target_2025-05-13-12-38-11	Сохранен

< 1 >

Рисунок 4.3.3.1 – Экран «Пайплайны»

Экран «Пайплайны» содержит следующую информацию о каждом пайплайне:

- название;
- тип;
- наименования входных спецификаций;
- наименование выходной спецификации;
- дата и время создания;
- статус.

По умолчанию список пайплайнов отсортирован по полю «Дата создания» в порядке убывания (новые записи сверху). Сортировку списка пайплайнов можно изменить с помощью переключателей в заголовках столбцов таблицы или щелчком левой кнопкой мыши в области заголовка столбца таблицы.

Также предусмотрена возможность фильтрации списка наборов данных по текущему статусу. Для установки фильтра предусмотрен переключатель в заголовке столбца «Статус».

На экране «Пайплайны» реализована возможность регистронезависимого поиска по вхождению поискового значения в значения параметров набора данных. Для начала поиска пользователю необходимо указать значение для поиска в поисковой строке. По мере ввода значения в поисковую строку перечень отображаемых наборов данных будет корректироваться – будут отображаться записи, соответствующие текущему поисковому запросу.

На данный момент реализован поиск по столбцам «Название», «Тип», «Статус». Поиск по столбцу

Внизу экрана «Пайплайны» предусмотрена пагинация для переключения между страницами списка. Пользователь может выбрать конкретный номер страницы или воспользоваться переключателями «Вперед», «Назад».

Сортировка, фильтрация, поиск, пагинация на экране «Пайплайны» реализованы аналогично соответствующему функционалу на экране «Проекты» (см. п. 4.1.1).

Если в таблице с перечнем наборов данных проекта навести курсор на строку с набором данных и щелкнуть левой кнопкой мыши, то откроется карточка пайплайна (экран «Пайплайны») с подробной информацией о нем (см. п. 4.3.4).

Если на экране «Пайплайны» нажать кнопку «Создать пайплайн» будет произведена переадресация пользователя на страницу «Создание пайплайна подготовки» (см. п. 4.3.2).

4.3.4. Просмотр расширенных сведений о пайплайне (экран «Пайплайн»)

Экран с подробной информацией о пайплайне (экран «Пайплайн», Рисунок 4.3.4.1) открывается при щелчке левой кнопкой мыши на соответствующей строке перечня пайплайнов.

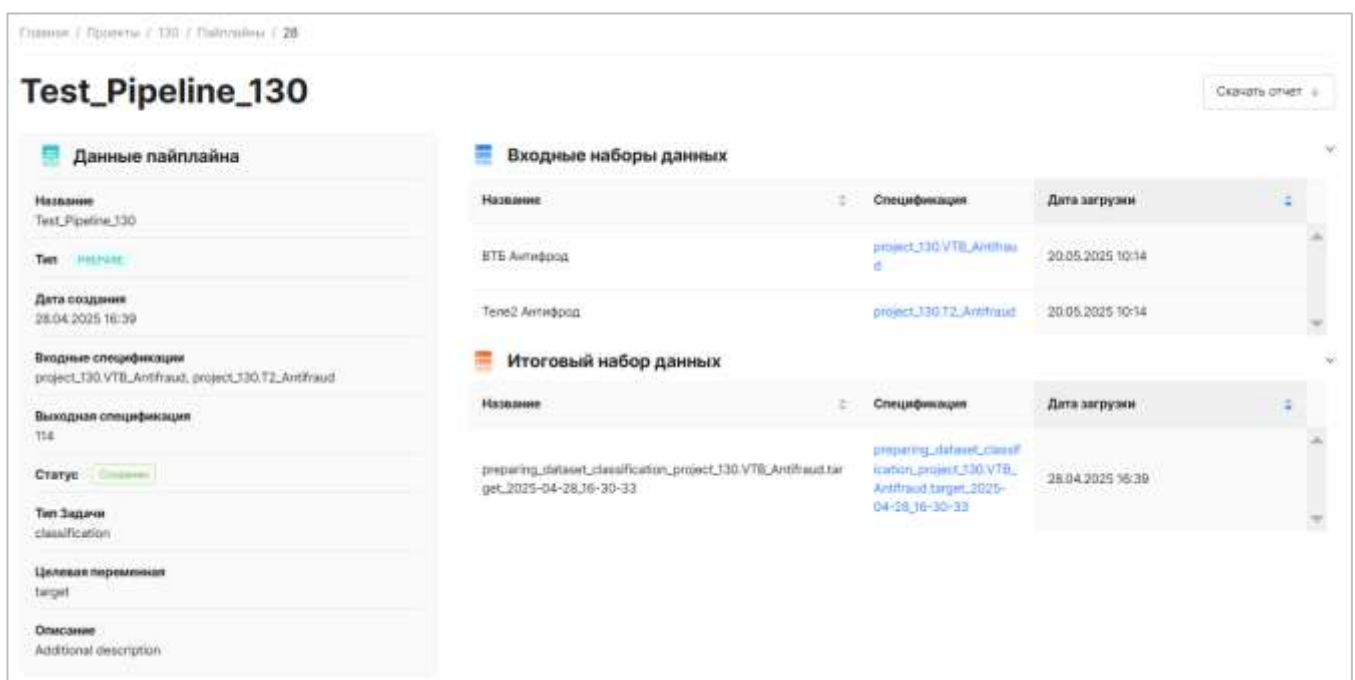


Рисунок 4.3.4.1 - Экран «Пайплайн»

На экране «Пайплайн» отображаются:

- название пайплайна;
- тип;
- дата\время создания;
- перечень входных спецификаций;
- выходная спецификация;

- статус;
- тип задачи;
- целевая переменная;
- описание;
- сведения входных наборах данных (рабочая область «Входные наборы данных»);
- сведения об итоговом наборе данных (рабочая область «Итоговый набор данных»).

В рабочей области «Входные наборы данных» отображаются в табличном виде название, спецификация и дата\время загрузки для наборов данных, которые были использованы для создания пайплайна подготовки. Если щелкнуть левой кнопки мыши на строке таблицы, то откроется страница (экран) «Набор данных» (п.4.2.2) с подробной информацией о выбранном наборе данных.

Рабочая область «Итоговый набор данных» содержит аналогичную информацию о наборе данных, который был сформирован в процессе создания пайплайна подготовки. Из рабочей области также возможен переход к экрану «Набор данных». На экране «Пайплайн» также предусмотрена возможность скачивать отчет, сформированный по итогам создания пайплайна. Для этого предусмотрена кнопка «Скачать отчет». Структура отчета описана в п. 4.5.

4.4. Работа с ML-моделями

4.4.1. Обучение ML-модели

Обучение ML-модели можно выполнить на странице (экране) «Обучение модели» (Рисунок 4.4.1.1). Для перехода на данную страницу необходимо выбрать проект, а затем выбрать на боковой панели пункт меню «Модели»\«Обучение модели».

Главная / Проекты / 126 / Модель / Обучение модели

Обучение модели

Для обучения модели вы можете выбрать только один набор данных

Название модели
Введите название

Описание модели
Введите описание

Тип задачи
Выберите тип задачи

Для выбора целевой переменной - выберите набор данных

Целевая переменная
Выберите целевую переменную

Размер тестовой выборки
Введите размер тестовой выборки

Количество фолдов для кросс-валидации
Введите количество фолдов

Выберите набор данных для обучения модели

Название	Спецификация	Дата загрузки
preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-58-29	preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-58-29	05.05.2025 17:00
preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-52-22	preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-52-22	05.05.2025 16:54
preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-45-33	preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_13-45-33	05.05.2025 16:47
preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_11-02-44	preparing_dataset_regression_project_126_datasetone.score1_2025-05-05_11-02-44	05.05.2025 14:04
preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_18-03-21	preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_18-03-21	30.04.2025 16:06
preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_18-02-49	preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_18-02-49	30.04.2025 16:04
preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_13-22-42	preparing_dataset_regression_project_126_datasetone.score1_2025-04-30_13-22-42	30.04.2025 13:24
project_126_datasetone	project_126_datasetone	29.04.2025 16:12
project_126_datasetone	project_126_datasetone	28.04.2025 14:27
Окноной ИД	project_126_datasetone	28.04.2025 14:26

Рисунок 4.4.1.1 – Начальный вид экрана «Обучение модели»

Для запуска процесса обучения модели необходимо выполнить следующие шаги:

1. Ввести название модели.
2. Ввести описание модели (необязательно).
3. Выбрать из списка тип решаемой задачи. Для выбора доступно два значения: регрессия и классификация.
4. Выбрать один набор данных для обучения модели (рабочая область «Выберите набор данных для обучения модели»).
5. Выбрать из перечня атрибутов выбранного на предыдущем шаге набора данных целевую переменную.
6. Указать размер тестовой выборки. Процент тренировочных данных - обязательное поле со значением от 1 до 99.
7. Указать количество фолдов для кросс-валидации. Должно быть положительным целым числом больше нуля
8. Нажать на кнопку «Обучить модель».

Когда запрос на обучение модели пройдет первичные проверки и будет передан на исполнение, в правом нижнем углу экрана появится уведомление: «Запрос успешно принят в обработку», а пользователь будет перенаправлен на страницу проекта, в рамках которого было запущено обучение модели.

После успешного завершения обучения модели в правом нижнем углу экрана появится уведомление: «Модель успешно создана».

Созданный пайплайн automl можно просмотреть через список пайплайнов на странице (экране) «Пайплайны» (меню боковой панели «Проект»\«Пайплайны»\«Созданные»).

Отчет по итогам создания пайплайна можно скачать на странице «Пайплайн» (см. п. 4.3.4) или на странице «Отчеты» (меню боковой панели Проект\Отчеты). Отчеты по результатам обучения модели на странице «Отчеты» имеет тип «automl» и представляет собой текстовый документ, который необходимо скачать для просмотра. Структура отчета описана в п. 4.5.

4.4.2. Просмотр сведений о модели проекта

Просмотр сведений о ML-модели, созданной в проекте, возможен на экране «Проект» (Рисунок 4.1.3.2) или на отдельном экране «Модель» (Рисунок 4.4.2.1), который можно открыть через меню боковой панели «Проект»\ «Модель»\ «Созданная».

В проекте может быть обучена только одна ML-модель.

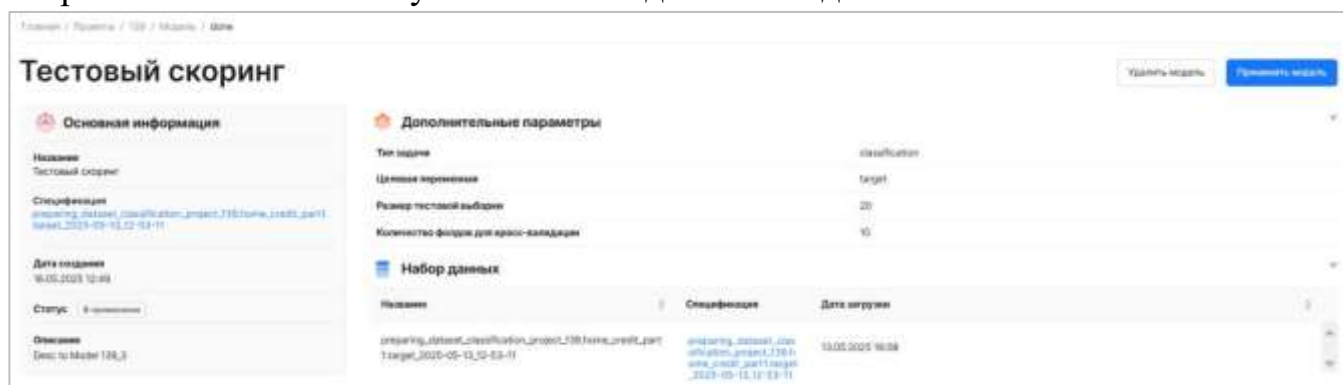


Рисунок 4.4.2.1 -Экран «Модель»

На экране проекта указаны следующие сведения о модели:

- название;
- спецификация;
- тип решаемой задачи;
- целевая переменная;
- дата и время создания;
- статус;
- описание.

На экране модели содержатся следующие рабочие области:

- «Основная информация»;
- «Дополнительные параметры»;
- «Набор данных».

В рабочей области «Основная информация» указаны следующие сведения о модели:

- название;
- спецификация;

- дата и время создания;
- статус;
- описание.

В рабочей области «Дополнительные параметры» указаны следующие сведения о модели:

- тип решаемой задачи;
- целевая переменная;
- размер тестовой выборки;
- количество фолдов для кросс-валидации.

В рабочей области «Набор данных» в табличном виде указаны название, спецификация, дата и время создания входного набора данных. Если щелкнуть левой кнопки мыши на строке таблицы, то откроется страница (экран) «Набор данных» (п. 3.3.2) с подробной информацией о наборе данных.

Также на экране «Модель» предусмотрены кнопки «Удалить модель» и «Применить модель».

Процесс применения модели описан в п. 4.4.3.

Процесс удаления модели описан в п. 4.4.4.

4.4.3. Применение модели в batch-режиме

Применение ML-модели можно запустить на странице (экране) «Применение модели» (Рисунок 4.4.3.1). Для перехода на данную страницу необходимо выбрать проект, а затем выбрать на боковой панели пункт меню «Модели»\«Применение модели» или нажать на кнопку «Применить модель» на странице модели (Рисунок 4.4.2.1).

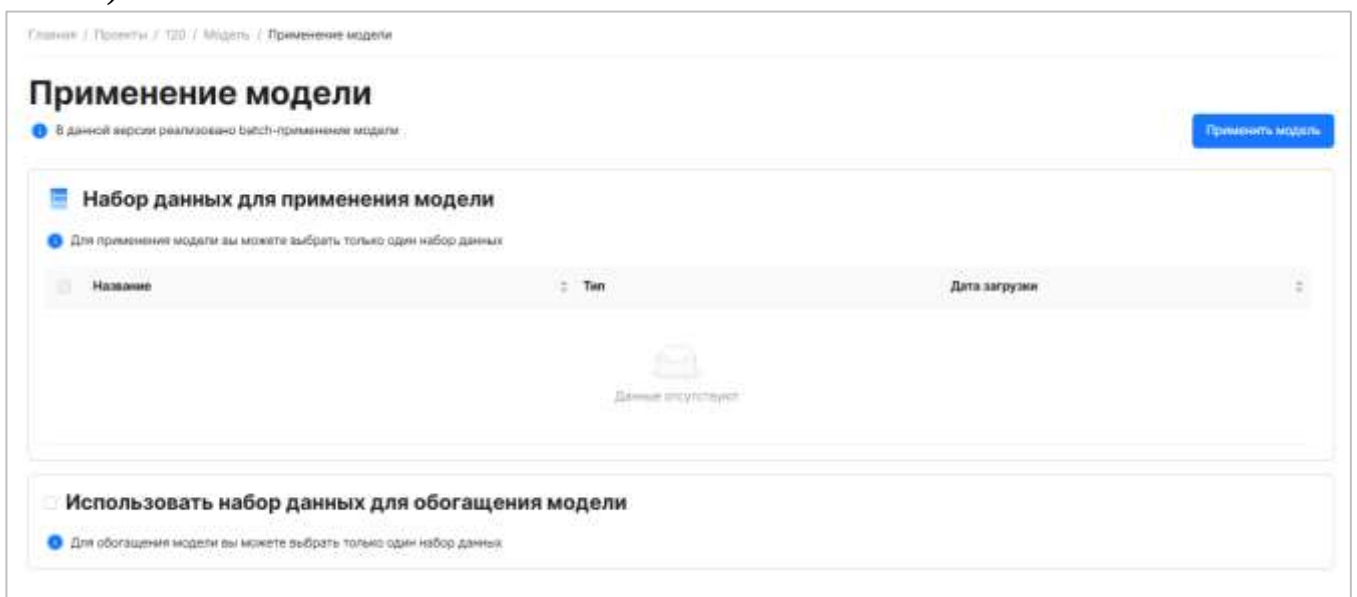


Рисунок 4.4.3.1 – Экран «Применение модели»

Для запуска процесса применения модели необходимо выполнить следующие шаги:

1. Выбрать набор данных для применения модели. Наборы данных для применения должно быть загружены предварительно (см. п. 4.6.1).
2. Указать требуется ли использовать обогащение при применении модели (для этого следует установить флажок в соответствующем чекбоксе) и при необходимости выбрать набор данных для обогащения из появившегося списка. Для обогащения могут быть использованы наборы данных, которые участвовали в создании пайплайна подготовки для применяемой модели.
3. Нажать на кнопку «Применить модель».

Когда запрос на применение модели проходит первичные проверки и передается на исполнение, в правом нижнем углу экрана отображается уведомление: «Запрос успешно принят в обработку», а пользователь перенаправляется на страницу проекта, в рамках которого было запущено обучение модели.

После успешного завершения применения модели в правом нижнем углу экрана отображается уведомление: «Модель {Название модели} успешно применена. Можно скачать результат».

Отчет по итогам обучения модели можно скачать на странице «Отчеты» (меню боковой панели «Проект»\«Отчеты»). Отчеты по результатам применения модели на странице «Отчеты» имеет тип «batch» и представляет собой текстовый документ, который необходимо скачать для просмотра. Структура отчета описана в п. 4.5.

4.4.4. Удаление модели

Для запуска процесса удаления модели необходимо нажать на кнопку «Удалить модель» на странице этой модели (экран «Модель», Рисунок 4.4.2.1).

После чего появится окно для подтверждения удаления (Рисунок 4.4.4.1).

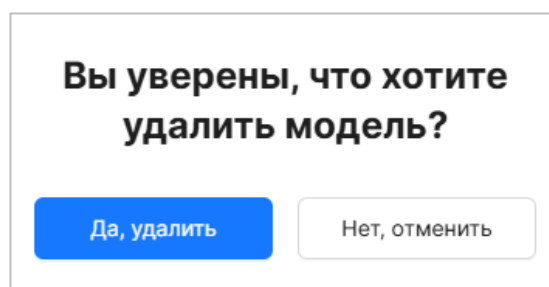


Рисунок 4.4.4.1 – Подтверждение удаления модели

Если нажать «Нет, отменить», то процесс удаления будет остановлен.

Если нажать «Да, удалить», то стартует процесс удаление модели.

Если нажать «Да, подтвердить», то стартует процесс удаление модели и связанных с ней артефактов: пайплайна AutoML (тип пайплайна «AutoML») и результатов применения модели.

После успешного выполнения первого этапа удаления пользователю будет направлено уведомление с текстом: «Первый этап удаления модели {Наименование модели} завершен. Стартовало физическое удаление данных модели».

После полного удаления пользователю направляется уведомление с текстом: «{Наименование модели} была успешно удалена!». Модель, пайплайн AutoML и результаты применения модели получают статус «Удалена» или «Удален». Восстановление удаленных объектов не предусмотрено.

Если при удалении модели и связанных с ней артефактов возникли ошибки, то пользователю будет выведено соответствующее уведомление.

4.5. Просмотр отчетов по операциям в проекте

Отчеты по результатам выполнения длительных операций в рамках проекта можно скачать для просмотра на странице (экране) «Отчеты» (Рисунок 4.5.1). Для перехода на страницу отчеты необходимо перейти в меню на боковой панели «Проект»\«Отчеты»).

Название	Тип отчета	Дата создания	Скачать
automl_pipeline_regression_score1_2025-05-02-06-33-30	automl	02.05.2025 13:23	Скачать
datasets_match_keys_2025-05-18-03-45-53	match_keys	18.05.2025 13:38	Скачать
datasets_match_keys_2025-05-18-07-06-06	match_keys	18.05.2025 10:41	Скачать
automl_pipeline_regression_score1_2025-05-06-10-17-47	automl	06.05.2025 13:57	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-05-05-13-58-28	prepare	05.05.2025 17:00	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-05-05-13-52-22	prepare	05.05.2025 16:54	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-05-05-13-45-33	prepare	05.05.2025 16:47	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-05-05-11-02-44	prepare	05.05.2025 14:04	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-04-30-18-00-57	prepare	30.04.2025 18:00	Скачать
preprocessing_pipeline_preparing_dataset_regression_project_136-dataset-score1_2025-04-30-18-00-48	prepare	30.04.2025 18:04	Скачать

Рисунок 4.5.1. – Экран «Отчеты»

На данной странице можно скачать отчет по результатам следующих операций:

- подбор ключевых полей для объединения данных (тип отчета match_keys) (отчет описан в п. 4.3.1);
- создание пайплайна подготовки (тип отчета prepare);
- создание пайплайна обучения модели (тип отчета automl);
- применение модели в batch-режиме (тип отчета batch).

Перечень отчетов представлен в табличном виде. Для каждого отчета указаны название, тип, дата и время создания. Для скачивания отчета необходимо нажать на ссылку «Скачать» в соответствующей строке таблицы.

По умолчанию список отчетов отсортирован по полю «Дата создания» в порядке убывания (новые записи сверху).

Отчет по итогам создания пайплайна подготовки содержит информацию о входящих данных и выполненных преобразованиях (Рисунок 4.5.2).

В отчете используются следующие обозначения:

- entity_type – тип созданной сущности;
- id – идентификатор созданного пайплайна;
- name – имя пайплайна;
- description – описание;
- status – текущий статус;
- project_id – идентификатор проекта, в котором создан пайплайн;
- created_at – дата и время создания;
- updated_at – дата и время обновления;
- main_data – раздел для описания выполненных преобразований с данными;
- type – тип пайплайна;
- configuration.transformer.path – путь до трансформера в S3-хранилище программы;
- input_specs – идентификаторы спецификаций входных наборов данных;
- output_specs – идентификатор спецификации итогового набора данных.

```
{
  "entity_type": "pipeline",
  "id": 44,
  "name": "preprocessing_pipeline_preparing_dataset_regression_project_126.datasetone.score1_2025-05-05_13-58-29",
  "description": "Пайплайн для подготовки данных",
  "status": 2,
  "project_id": 126,
  "created_at": "2025-05-05T14:00:26.471125",
  "updated_at": "2025-05-05T14:00:26.471130",
  "main_data": [
    "Eliminated: datasetone_score11 (reasons: high_correlation)",
    "Eliminated: datasetone_score12 (reasons: high_correlation)",
    "Eliminated: datasetone_score13 (reasons: high_null_values)"
  ],
  "model_specific_data": null,
  "additional_attributes": {
    "type": "PREPARE",
    "configuration": {
      "transformer": {
        "path": "project-126/models/preparing_pipeline_regression_project_126.datasetone.score1_2025-05-05_13-58-29"
      }
    }
  },
  "input_specs": "107,122",
  "output_specs": "136"
}
```

Рисунок 4.5.2. - Отчет «Создание пайплайна подготовки»

Отчет по итогам создания пайплайна automl содержит информацию о процессе обучения и модели-победителя.

В отчете используются следующие обозначения:

- entity_type – тип созданной сущности;
- id – идентификатор созданного пайплайна automl;
- name – имя пайплайна;
- description – описание;
- status – текущий статус;
- project_id – идентификатор проекта, в котором создан пайплайн;
- created_at – дата и время создания;
- updated_at – дата и время обновления;
- main_data.best_model – раздел для описания параметров модели-победителя;
- main_data.trained_models – раздел для описания параметров моделей, обученных в ходе процесса automl;
- type – тип пайплайна;
- configuration. dataset – раздел для информации о наборе данных, использованном при обучении;
- fold_count – количество фолдов для кросс-валидации;
- target – целевая переменная;
- training_percent – размер тестовой выборки;
- typeTask – тип задачи;
- input_specs – идентификатор спецификации входного набора данных;
- output_specs – идентификатор спецификации итогового набора данных.

Отчет по итогам применения модели содержит информацию о входящих данных и преобразованиях (Рисунок 4.5.3).

```
{
  "model_name": "Application_1",
  "mergerdata": true,
  "created_at": "2025-05-05T14:00:26.471125",
  "result": {
    "id1": 0.8235,
    "id2": 0.4567,
    "id3": 0.9083
  }
}
```

Рисунок 4.5.3. - Отчет «Применение модели»

В отчете используются следующие обозначения:

- model_name – наименование модели;

- `mergerdata` – признак, указывающий на использование обогащения при применении модели;
- `created_at` – дата и время применения модели;
- `result` – раздел для описания результатов применения.

4.6. Действия, осуществляемые вручную

4.6.1. Загрузка наборов данных

4.6.1.1. Загрузка наборов данных для обучения

Для каждого проекта и вида набора данных (например, данные о звонках, о выданных кредитах и т. д.) администратор организации-поставщика данных должен написать отдельный скрипт для загрузки данных в S3-хранилище программы. Данные могут быть загружены только в формате `parquet`.

В скрипте обязательно должны быть заданы\указаны:

- параметры для проверки доступа;
- имя, под которым должен быть сохранен `parquet`-файл в S3-хранилище;
- путь, по которому необходимо загрузить данные в S3-хранилище.

Проектные данные пользователи должны загружать изначально в папку `project-{project_id}/stg/{dataset_name}/{upload_date}`, где:

- `project_id` – идентификатор проекта;
- `project-`, `stg` – константы;
- `dataset_name` – имя, под которым должен быть сохранен `parquet`-файл в S3-хранилище;
- `upload_date` – дата загрузки в формате `YYYYMMDD` (в текущей версии программы рассматривается загрузка одного вида набора данных один раз в сутки).

Вместе с `parquet`-файлом набора данных опционально может быть загружен `json`-файл с информацией о первичном ключе (может быть составным), обязательных и уникальных полях загружаемого НД. Имя `json`-файла, под которым должен быть сохранен `json`-файл в S3-хранилище, должно быть равно `dataset_name`. `Json`-файл с дополнительной информацией должен загружаться в одну папку в `stg`-слое S3-хранилища совместно с набором данных. Информация о первичном ключе, обязательных и уникальных полях набора данных может быть зафиксирована только в рамках 1-й загрузки и не может быть скорректирована в дальнейшем.

В файле `dataset_name.json` пользователь также может указать общее описание назначения и содержания данных, а также наименование набора данных для отображения в интерфейсе.

Структура файла `dataset_name.json`, который может быть успешно обработан, фиксирована (Рисунок 4.6.1.1.1).

```
{
  "type": "object",
  "properties": {
    "primary_key": {
      "type": "array",
      "items": {
        "type": "string"
      }
    },
    "add_constraints": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "column_name": {
            "type": "string"
          },
          "not_null": {
            "type": "boolean"
          },
          "unique": {
            "type": "boolean"
          }
        },
        "required": [
          "column_name"
        ]
      }
    },
    "description_user": {
      "type": "string",
      "description": "Описание данных от пользователя"
    },
    "dataset_name_ui": {
      "type": "string",
      "description": "Наименование НД для отображения на UI"
    }
  }
}
```

Рисунок 4.6.1.1.1 – Структура файла с дополнительной информацией о загружаемом наборе данных

После завершения загрузки набора данных в stg-слой S3-хранилища в программе будут выполнены дополнительные проверки и профилирование набора данных. Набор данных, успешно прошедший все проверки, будет перемещен из stg-слоя в слой il.

После успешного завершения всех этапов загрузки пользователь может увидеть информацию о загруженном наборе данных в интерфейсе программы на экранах «Проект», «Наборы данных», «Набор данных».

4.6.1.2. Загрузка наборов данных для применения

Для каждого проекта и вида набора данных (например, данные о звонках, о выданных кредитах и т. д.) администратор организации должен написать отдельный скрипт для загрузки данных в S3-хранилище программы. Данные могут быть загружены только в формате parquet.

В скрипте обязательно должны быть заданы\указаны:

- параметры для проверки доступа;

- имя, под которым должен быть сохранен parquet-файл в S3-хранилище;
- путь, по которому необходимо загрузить данные в S3-хранилище.

Проектные данные пользователи должны загружать изначально в папку project-{project_id}/app/{dataset_name}/{upload_date}, где:

- project_id – идентификатор проекта;
- project-, app – константы;
- dataset_name – имя, под которым должен быть сохранен parquet-файл в S3-хранилище;
- upload_date – дата загрузки в формате YYYYMMDD.

Вместе с parquet-файлом набора данных опционально может быть загружен json-файл с описанием назначения и содержания загружаемого набора данных для применения и его именем для отображения в интерфейсе. Имя json-файла, под которым должен быть сохранен json-файл в S3-хранилище, должно быть равно dataset_name. Json-файл с дополнительной информацией должен загружаться в одну папку в app слое S3-хранилища совместно с набором данных.

Структура файла dataset_name.json, который может быть успешно обработан, фиксирована (Рисунок 4.6.1.2.1).

```
{
  "type": "object",
  "properties": {
    "description_user": {
      "type": "string",
      "description": "Описание данных от пользователя"
    },
    "dataset_name_ui": {
      "type": "string",
      "description": "Наименование НД для отображения на UI"
    }
  }
}
```

Рисунок 4.6.1.2.1 – Структура файла с дополнительной информацией о загружаемом для применения наборе данных

После завершения загрузки набора данных в app-слой S3-хранилища в программе будут выполнены дополнительные проверки. Набор данных для применения, успешно прошедший все проверки, не перемещается из app-слоя S3-хранилища.

После успешного завершения всех этапов загрузки пользователь сможет увидеть информацию о загруженном наборе данных в интерфейсе программы на экранах «Проект», «Наборы данных», «Набор данных».

4.6.2. Действия, выполняемые вручную администратором программы

Перечень действие, которые Администратор программы выполняет вручную:

- создание организации-участника программы и назначение прав;

- создание пользователя и назначение прав;
- формирование и корректировка набора участников проекта;
- удаление пользователя и организации;
- редактирование прав;
- смена пароля;
- удаление набора данных, пайплайна;
- устранение нештатной ситуации.