Loginom

ModelOps в Loginom: управление моделями и экспериментами

Николай Паклин

- Николай Паклин – тимлид группы «Обучение»
- 19 лет в low-code аналитике и Loginom Company
- С командой запустил Loginom Skills, проект «Мастерская», библиотеки компонентов для Loginom
- Бизнес-тренер, преподаватель, автор книг и учебников по аналитике данных

О чем будем говорить

1. Что такое ModelOps
2. Когда и для чего важно внедрять ModelOps
3. Как управлять моделями и экспериментами в Loginom

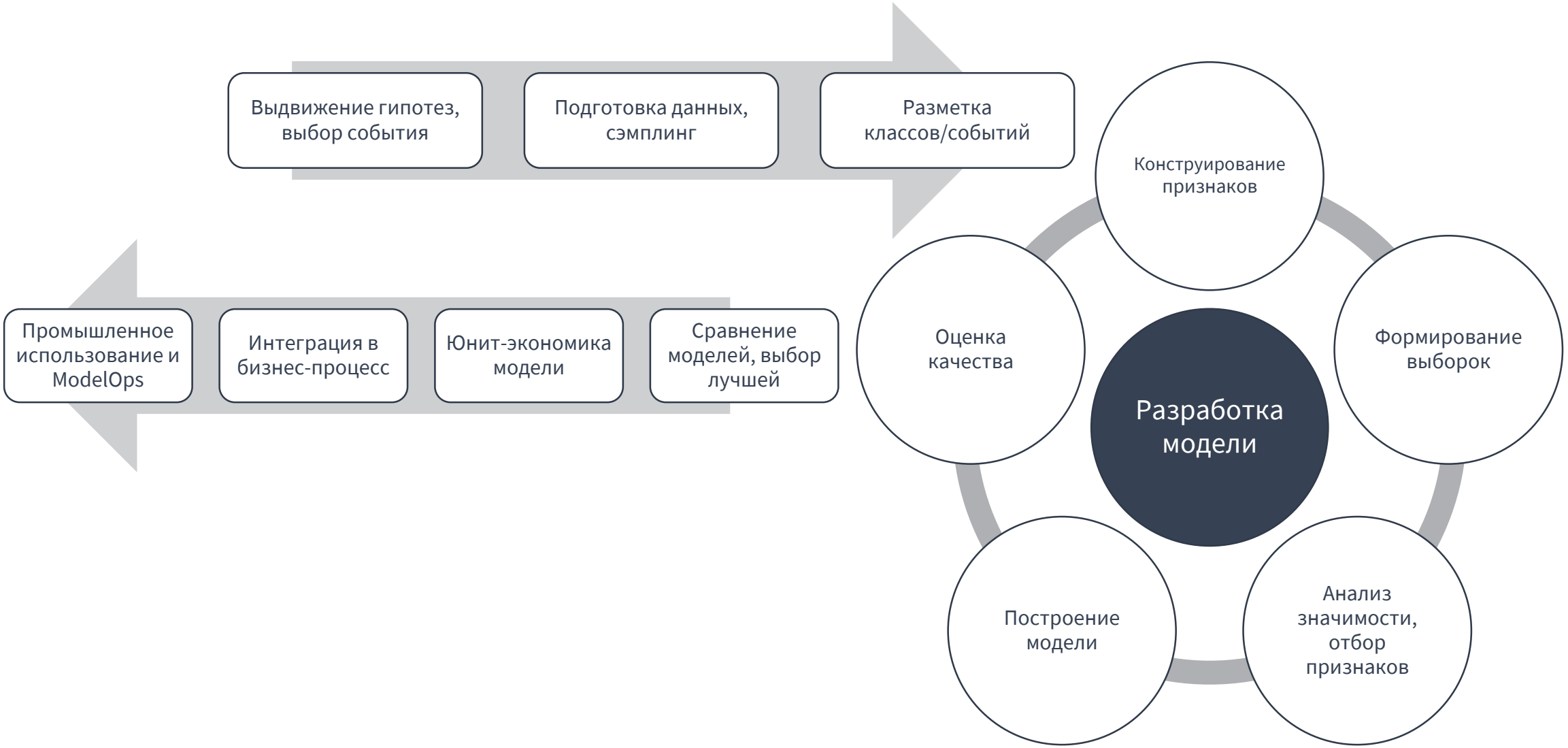
Операционализация аналитики данных

- **ModelOps** – набор практик, направленных на унификацию процессов разработки и развертывания аналитических моделей и решений *(часто идет речь о моделях machine learning - MLOps)*
- **DataOps** – набор практик для повышения качества и скорости предоставления данных, а также сокращения времени и усилий на создание и обслуживание конвейеров данных

Ключевые требования к бизнес-процессу

- В рамках одной команды важно обеспечить **воспроизводимость** вычислений
- Коллеги должны иметь возможность **повторить эксперименты** без особых усилий

Общая схема процесса построения ML-модели



Особенности процесса построения ML-модели

- Несколько версий датасетов
- Несколько моделей
- Трудность выбора лучшей модели сразу
- Проект может требовать переключаться между моделями в процессе эксплуатации

Типичные ситуации

1. Датасеты, на которых строилась модель удалены/утеряны/перезатерты
2. В папке много сценариев и не понятно, какой из них актуальный
3. Построено 20 моделей, но не ясно, где модель с лучшим ROC-AUC
4. Нужно передать коллеге исследовательскую задачу
5. Передача модели (сценария) в другой отдел для переноса в продакшн
6. Нужно понять, чем отличаются друг от друга две модели

Типичные проблемы

1. Различия в среде исполнения (разработка, тестирование, продакшн)
2. Неуправляемая случайность в данных или алгоритмах (Random Seed)
3. Проблемы с зависимостями со ссылками-пакетами, библиотеками языков программирования, зависимости в зависимостях

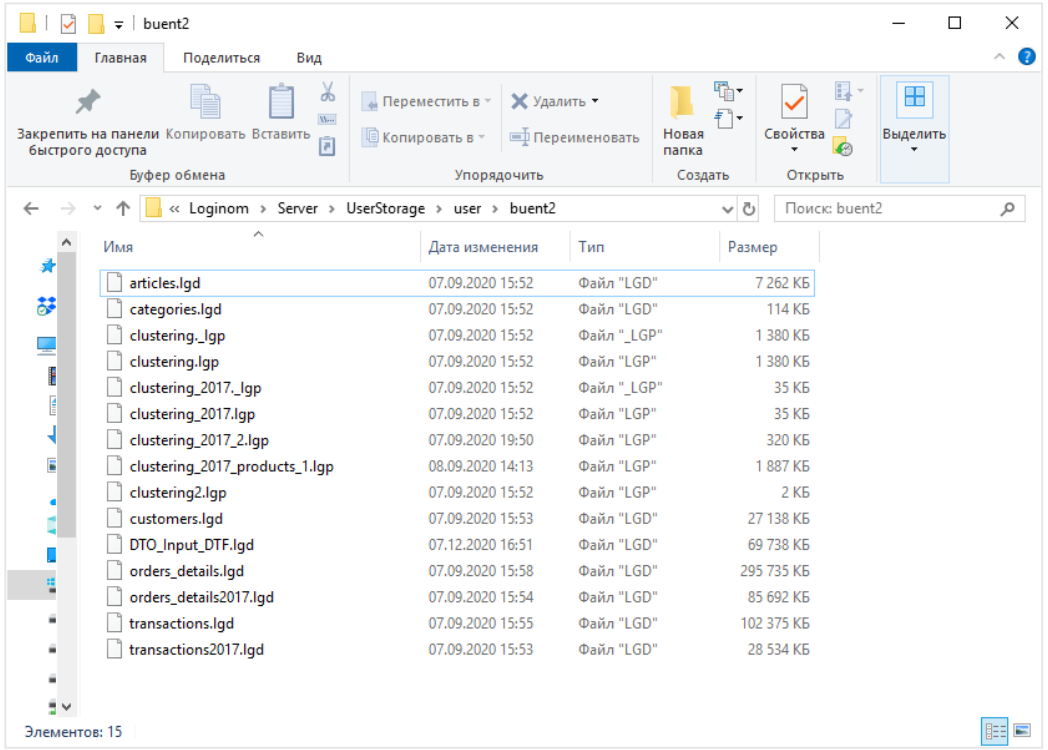
Тезис: в определенный момент нам понадобится повторить ранее проведенный вычислительный эксперимент

Архитектурная схема использования ModelOps

Управление экспериментами и моделями: уровни зрелости

Я положил тебе в сетевую папку проект, там есть все, что требуется

Как же с этим разобраться...



Типичная ситуация при ручных процессах ModelOps

Сценарий | Сценарий | Сценарий

Пакеты | LOF | Продажи - Novelty Detection | Сценарий

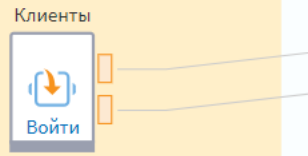
🛠 Пример демонстрирует поиск аномалий алгоритмом LOF в наборе данных по продажам в случае, когда обучение велось на "чистом" наборе данных, то есть без аномалий. Эта задача еще называется "обнаружение новизны" (Novelty Detection).

Для формирования такого "чистого" датасета мы удалили аномалии, найденные алгоритмом LOF на "грязном" датасете. Для проверки эффективности модели делается "скоринг" ранее найденных аномалий моделью обнаружения новизны.

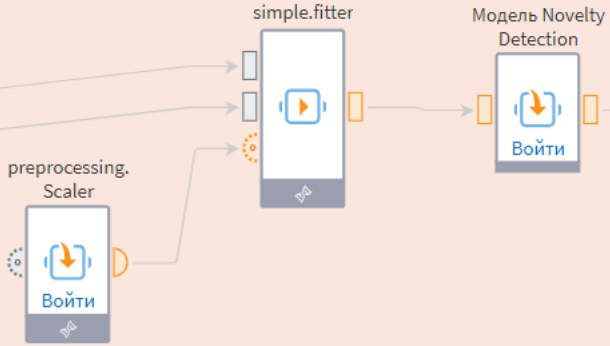
Эта задача, когда обучение ведется на одном классе, называется полуконтролируемым обучением (Semi-Supervised Learning).

Исходную выборку мы разделили на 2 части: обучающую и тестовую.
Найденные ранее аномалии находятся в тестовой выборке.

Таким способом мы моделируем задачу Novelty Detection.

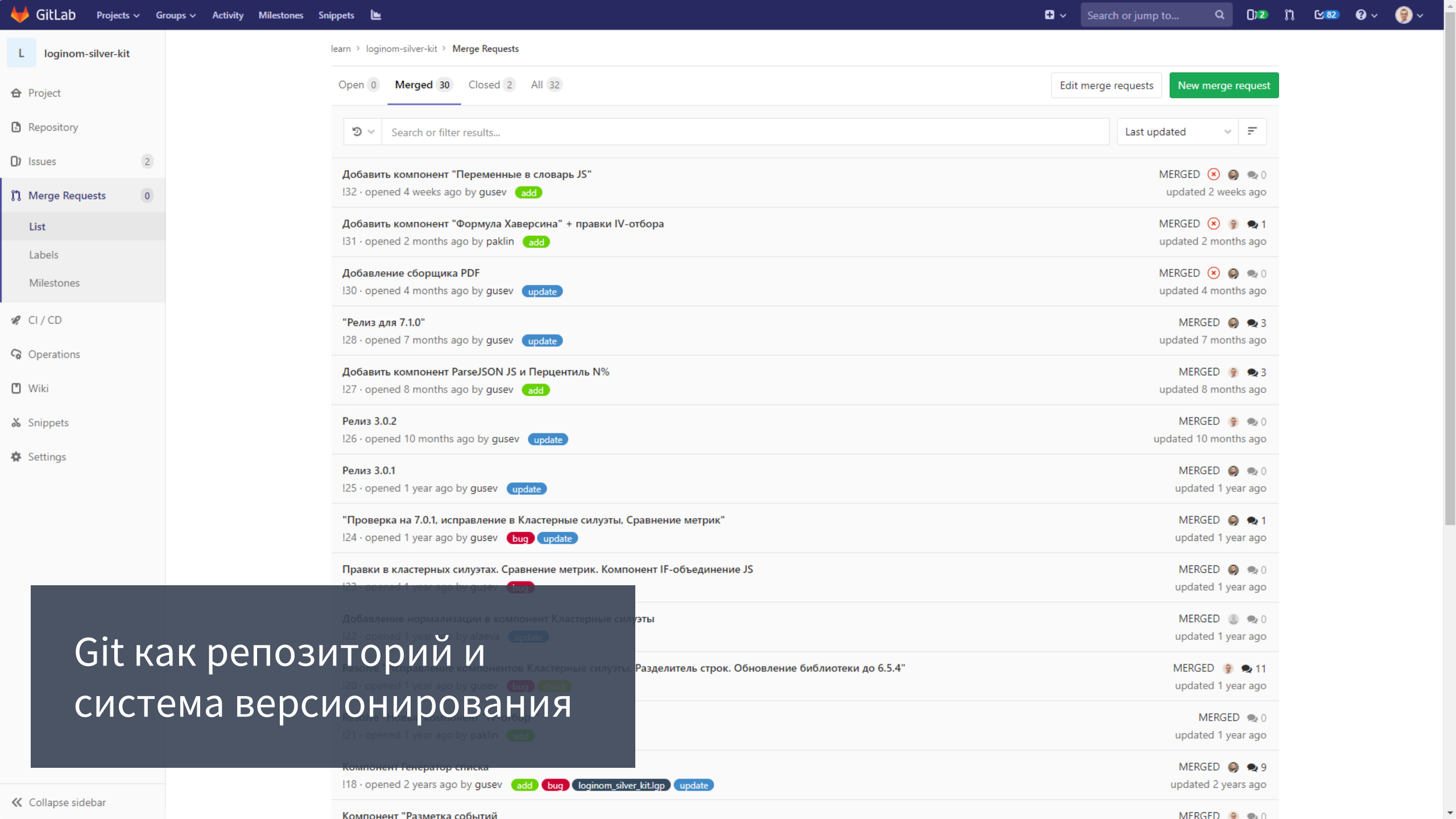


Построение модели с использованием библиотеки Loginom Python Kits.
Метод LocalOutlierFactor из библиотеки sklearn.
Используем режим Novelty = True.
Предварительно делаем z-масштабирование выборки.
Число соседей n = 20, загрязненность (contamination) 1%.



В результате скоринга через модель обнаружения новизны мы получили 230 совпадений из 230.
Такую модель можно использовать для скоринга новых записей, и относить или не относить их к аномальным.

Заметки и комментарии - не панацея, хотя помогают описать сценарий



Git как репозиторий и система версионирования

Git для ModelOps – что не так?

1. Ограничения на размер файлов в репозитории
2. Нельзя отслеживать изменения в бинарных файлах
3. Не предназначен для отслеживания метрик модели и гиперпараметров алгоритма обучения
4. Ориентирован на разработчиков ПО, а не аналитиков данных
5. Плохо согласуется с парадигмой low-code

Требования к фреймворку:

- Бесплатный в облаке, возможно, с ограничениями
- Работа в локальной сети
- Поддержка сторонних файловых хранилищ
- Подходит для систем класса low-code
- Низкий порог входа для аналитика
- Богатый Rest API

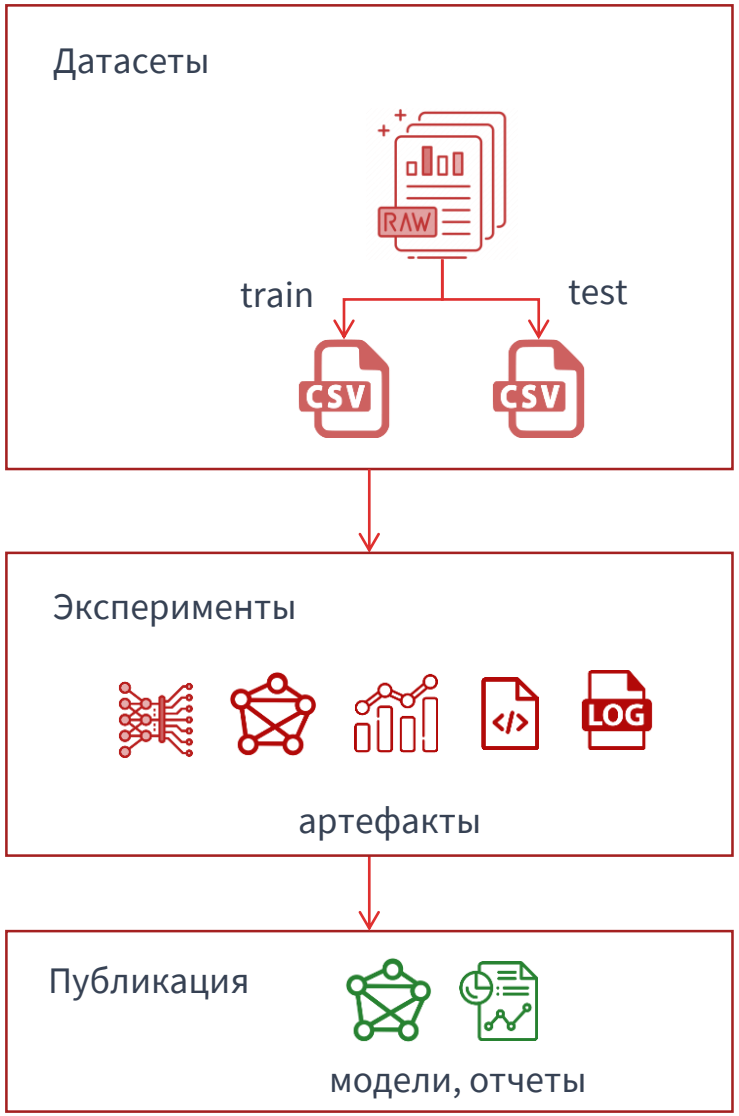
www.clear.ml

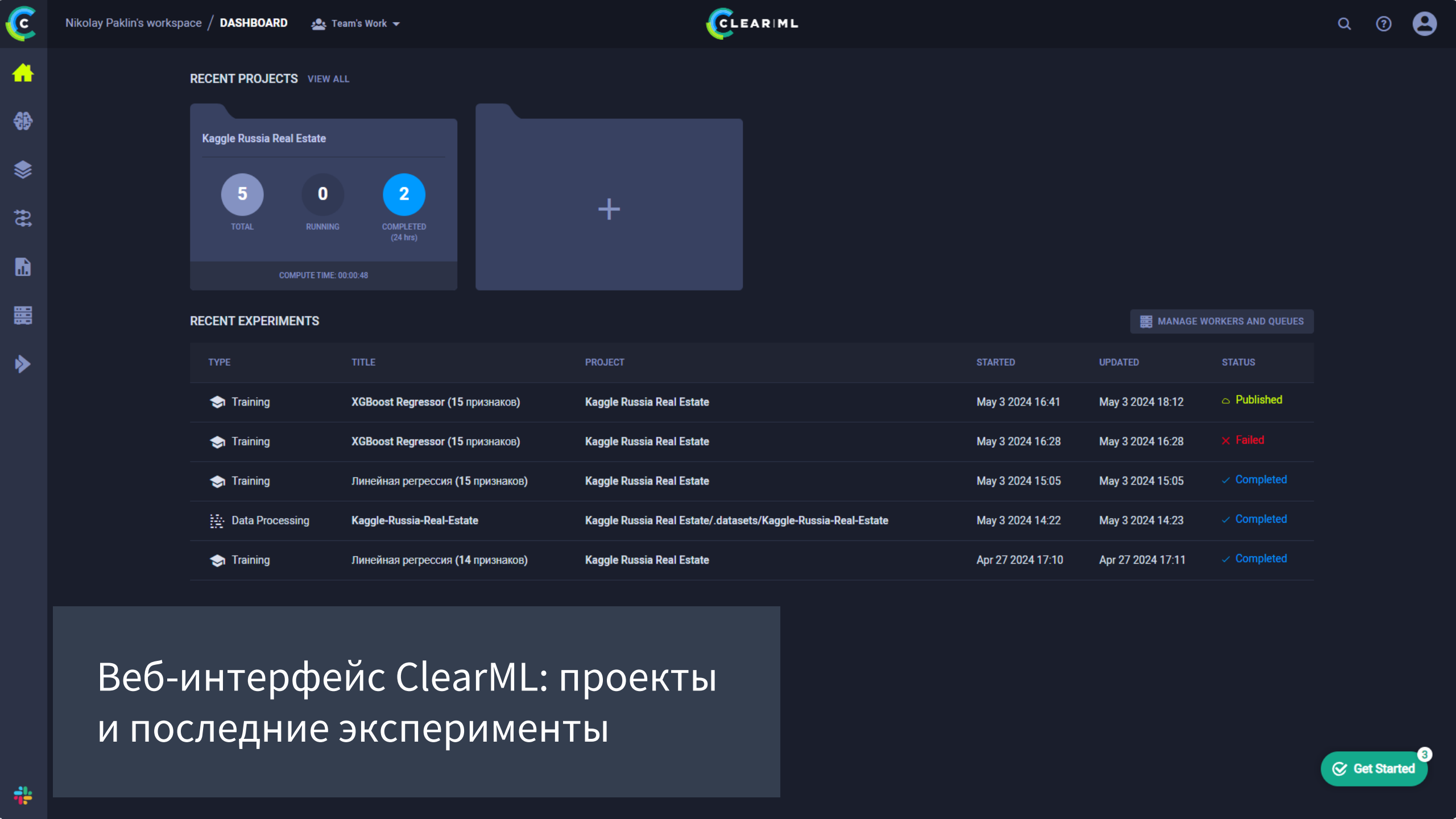
Модули ClearML

Реализация интеграции с Loginom – управление моделями и экспериментами

Терминология

1. Проект
2. Датасет, эксперимент
3. Модель, гиперпараметр
4. Метрика, артефакт
5. Отчет
6. Опубликованный объект





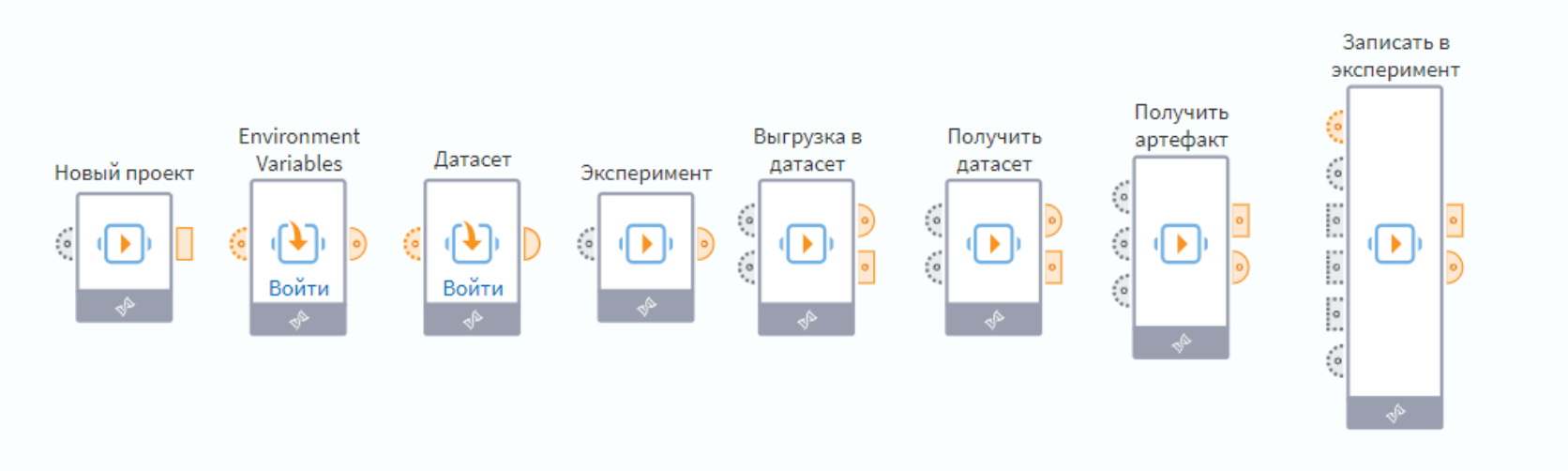
Веб-интерфейс ClearML: проекты и последние эксперименты

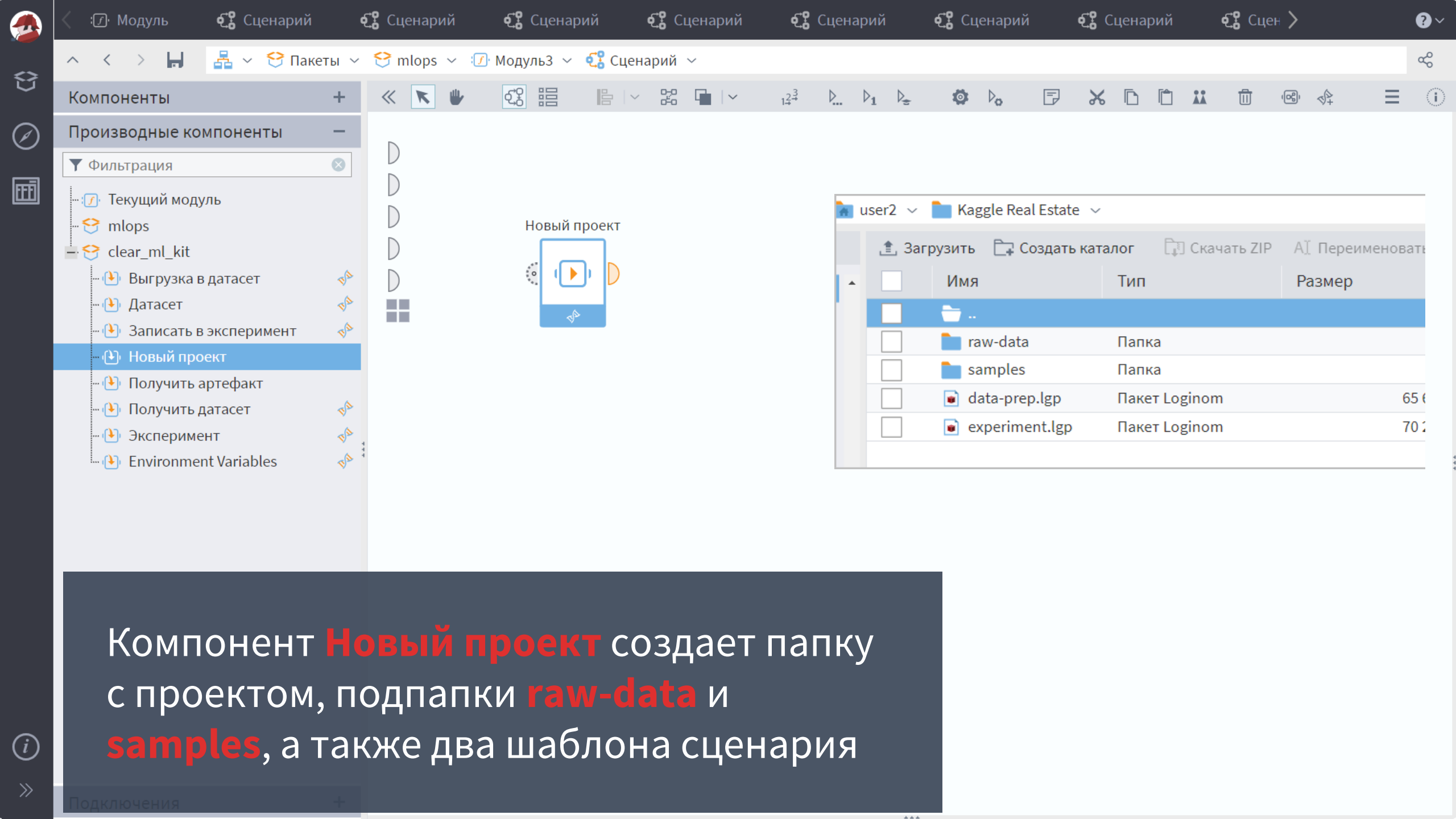
Технические подробности

Состав интеграции Loginom с ClearML:

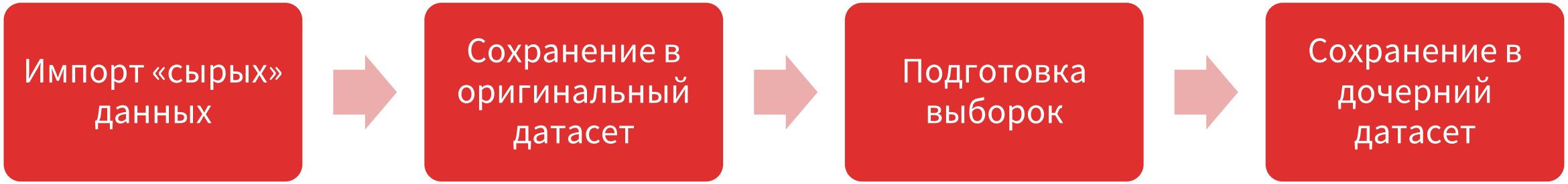
1. Библиотека компонентов **ClearML Kit**
2. Шаблоны сценариев

Для интеграции используется python-библиотека **clearml**




Модуль
Сценарий
Пакеты
mlops
Модуль3
Компоненты
Производные компоненты
Фильтрация
Текущий модуль
mlops
clear_ml_kit
Выгрузка в датасет
Датасет
Записать в эксперимент
Новый проект
Получить артефакт
Получить датасет
Эксперимент
Environment Variables
Новый проект
user2
Kaggle Real Estate
Загрузить
Создать каталог
Скачать ZIP
Имя
Тип
Размер
raw-data
Папка
samples
Папка
data-prep.lgp
Пакет Loginom
experiment.lgp
Пакет Loginom
Компонент Новый проект создает папку с проектом, подпапки raw-data и samples, а также два шаблона сценария

Модули шаблона сценария **data-prep.lgp**



Сырые данные должны находиться в папке **raw-data** и/или поставляться сценарием

Готовы компонент, требуются только ключи доступа к ClearML

Формируем сценарием выборки **train.lgd** и **test.lgd**

Готовый компонент

Сценарий

Пакеты | data_prep | Сохранение в оригинальный датасет | Сценарий

Сохранение в родительский датасет оригинальных файлов с данными (папка raw-data).

По умолчанию данный сценарий также сохраняется в эту папку
(опция отключается установкой переменной "Добавить текущий сценарий" в FALSE).

Имя проекта и датасет
Версия - опционально

Датасет

Войти

Выгрузка оригинальных файлов в датасет
(родительский датасет)

Выгрузка в датасет

Войти

Параметры авторизации к ClearML

Crendetials

Проверьте переменные

Environment Variables

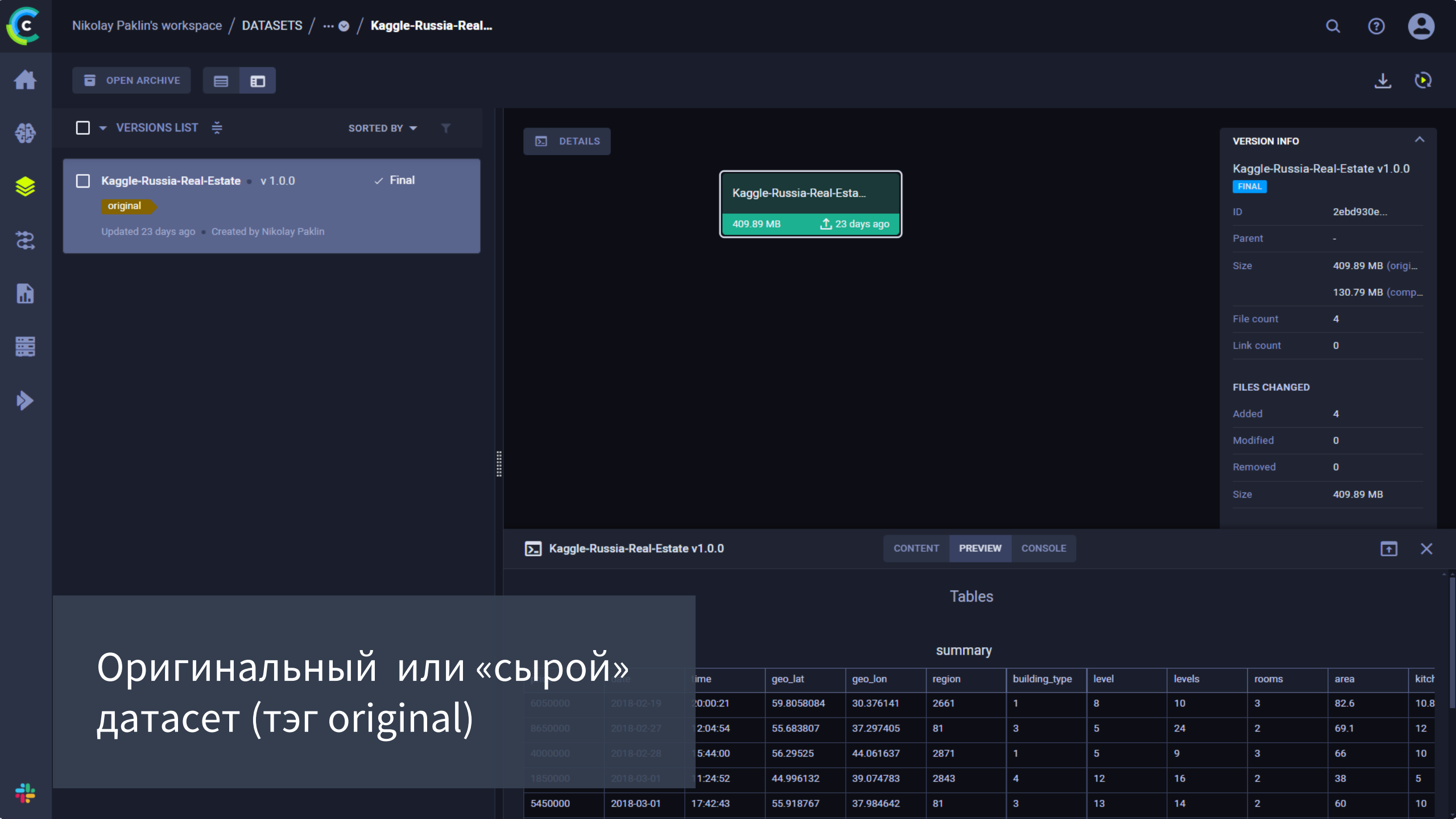
Войти

Компонент загружает все файлы из папки **raw-data** на сервер ClearML

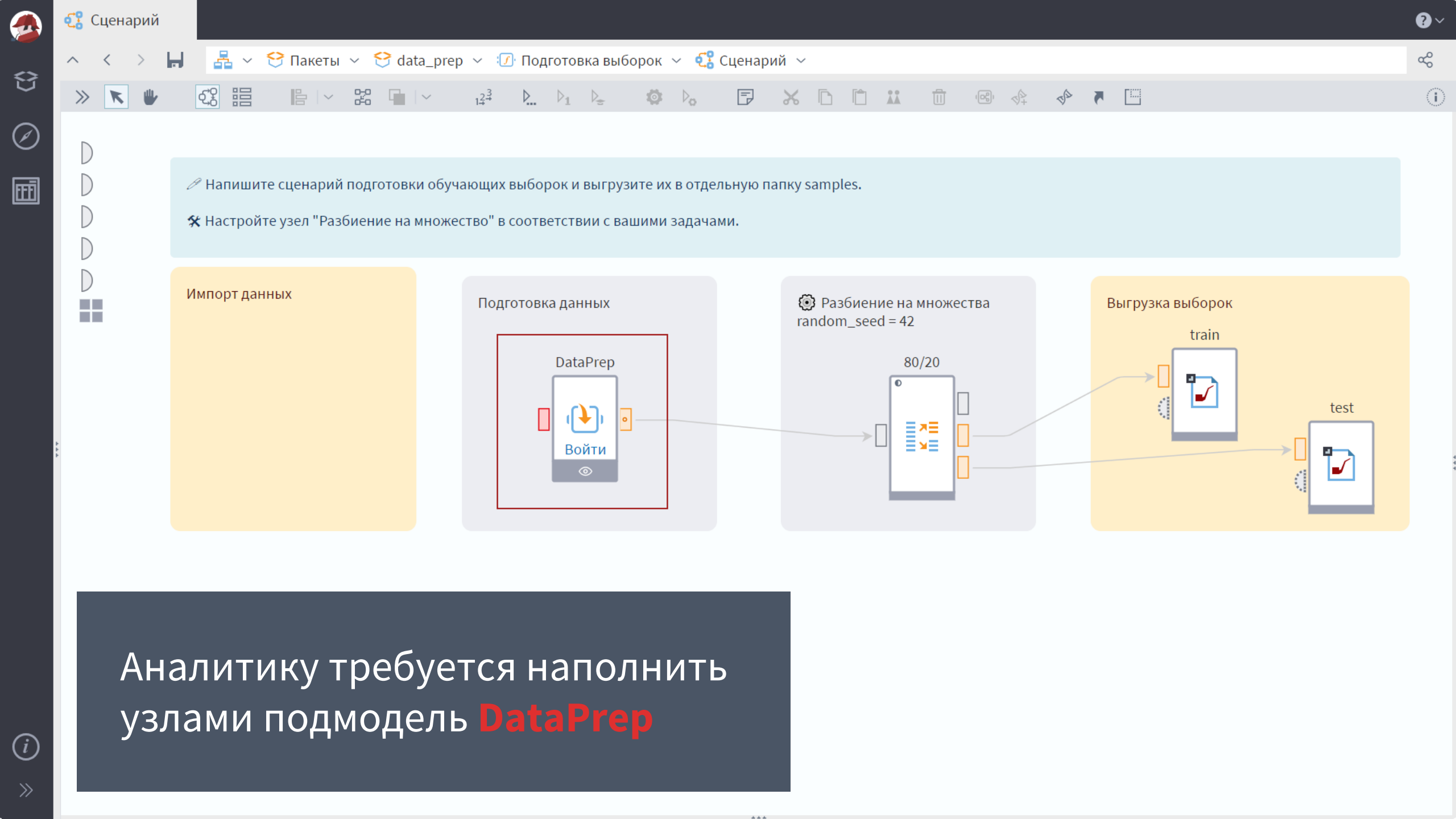
Environment Variables • Быстрый просмотр

Environment Variables

№	Метка	Значение
1	АВТОРИЗАЦИЯ	
2	CLEARML_WEB_HOST	https://app.clear.ml
3	CLEARML_API_HOST	https://api.clear.ml
4	CLEARML_FILES_HOST	https://files.clear.ml
5	Ключ доступа	PWD228F9DP
6	Секретный ключ	ecGEv32iMOhrPBXe0m9P
7	ОБЩИЕ ПАРАМЕТРЫ	
8	Каталог для загрузки артефактов	*<null>*
9	Список расширений	.~lgp,.lck
10	Список файлов	*<null>*
11	Игнорировать	true



		ime	geo_lat	geo_lon	region	building_type	level	levels	rooms	area	kitch
6050000	2018-02-19	20:00:21	59.8058084	30.376141	2661	1	8	10	3	82.6	10.8
8650000	2018-02-27	2:04:54	55.683807	37.297405	81	3	5	24	2	69.1	12
4000000	2018-02-28	5:44:00	56.29525	44.061637	2871	1	5	9	3	66	10
1850000	2018-03-01	1:24:52	44.996132	39.074783	2843	4	12	16	2	38	5
5450000	2018-03-01	17:42:43	55.918767	37.984642	81	3	13	14	2	60	10


Сценарий
Пакеты
data_prep
Подготовка выборок
Сценарий
Напишите сценарий подготовки обучающих выборок и выгрузите их в отдельную папку samples.
Настройте узел "Разбиение на множество" в соответствии с вашими задачами.
Импорт данных
Подготовка данных
DataPrep
Войти
Разбиение на множества
random_seed = 42
80/20
Выгрузка выборок
train
test
Аналитику требуется наполнить
узлами подмодель DataPrep

Сценарий

Пакеты > data_prep > Сохранение в дочерний датасет > Сценарий

Сохранение выборок train.lgd и test.lgd в дочерний датасет (папка samples).
Это действие можно повторять неоднократно - новые версии выборок будут наследоваться от родительского датасета.

По умолчанию данный сценарий также сохраняется в эту папку
(опция отключается установкой переменной "Добавить текущий сценарий" в FALSE).

Если требуется отнаследоваться от дочернего датасета, корректно настройте переменную "Родительский датасет" внутри подмодели "Датасет"

Имя проекта и датасет (ссылка)
Версия - опционально

Датасет*

Войти

Переменные ClearML (ссылка)

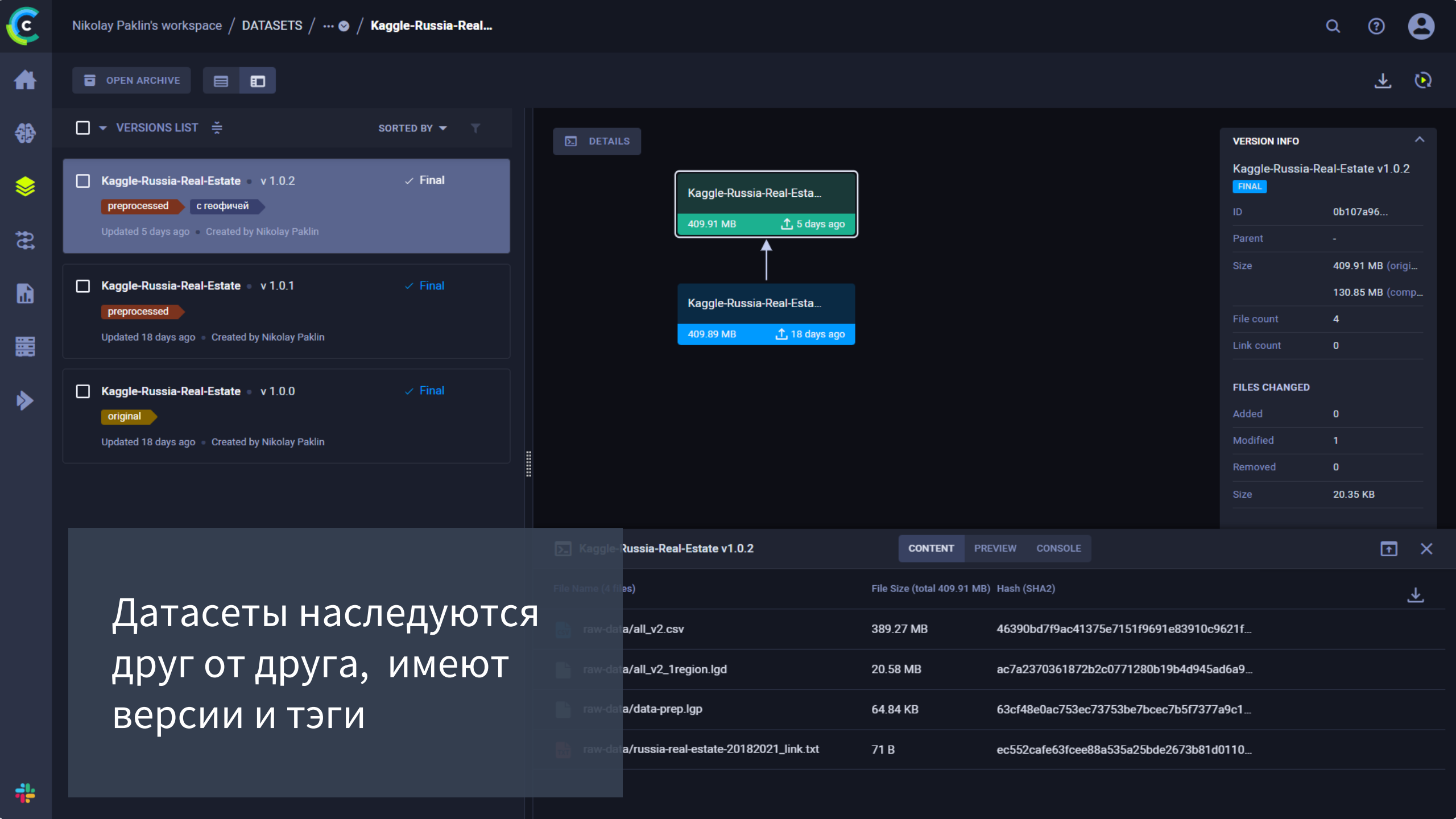
Environment Variables

Не забудьте сохранить текущий сценарий перед записью в датасет!

Выгрузка выборок в дочернюю версию датасета

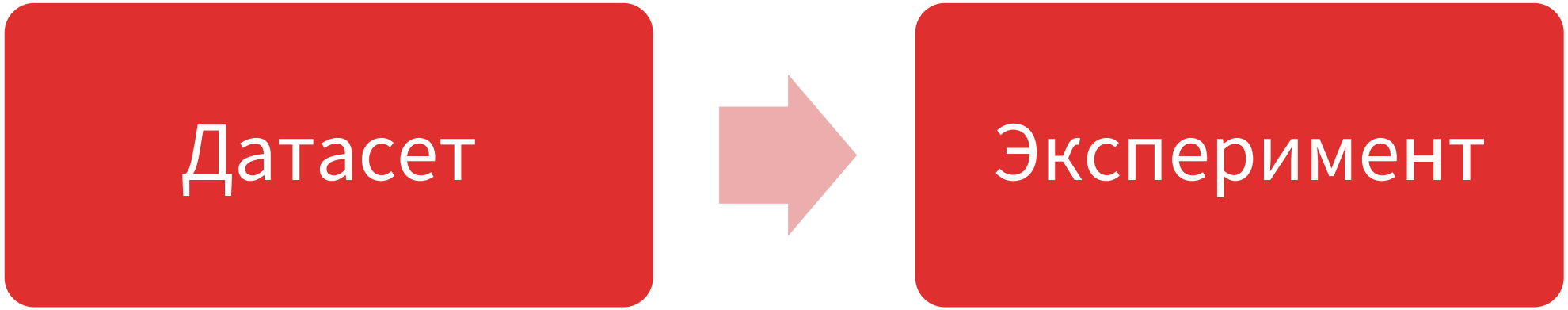
Выгрузка в датасет

Компонент загружает все файлы из папки **samples** на сервер ClearML как дочерний датасет



Датасеты наследуются друг от друга, имеют версии и тэги

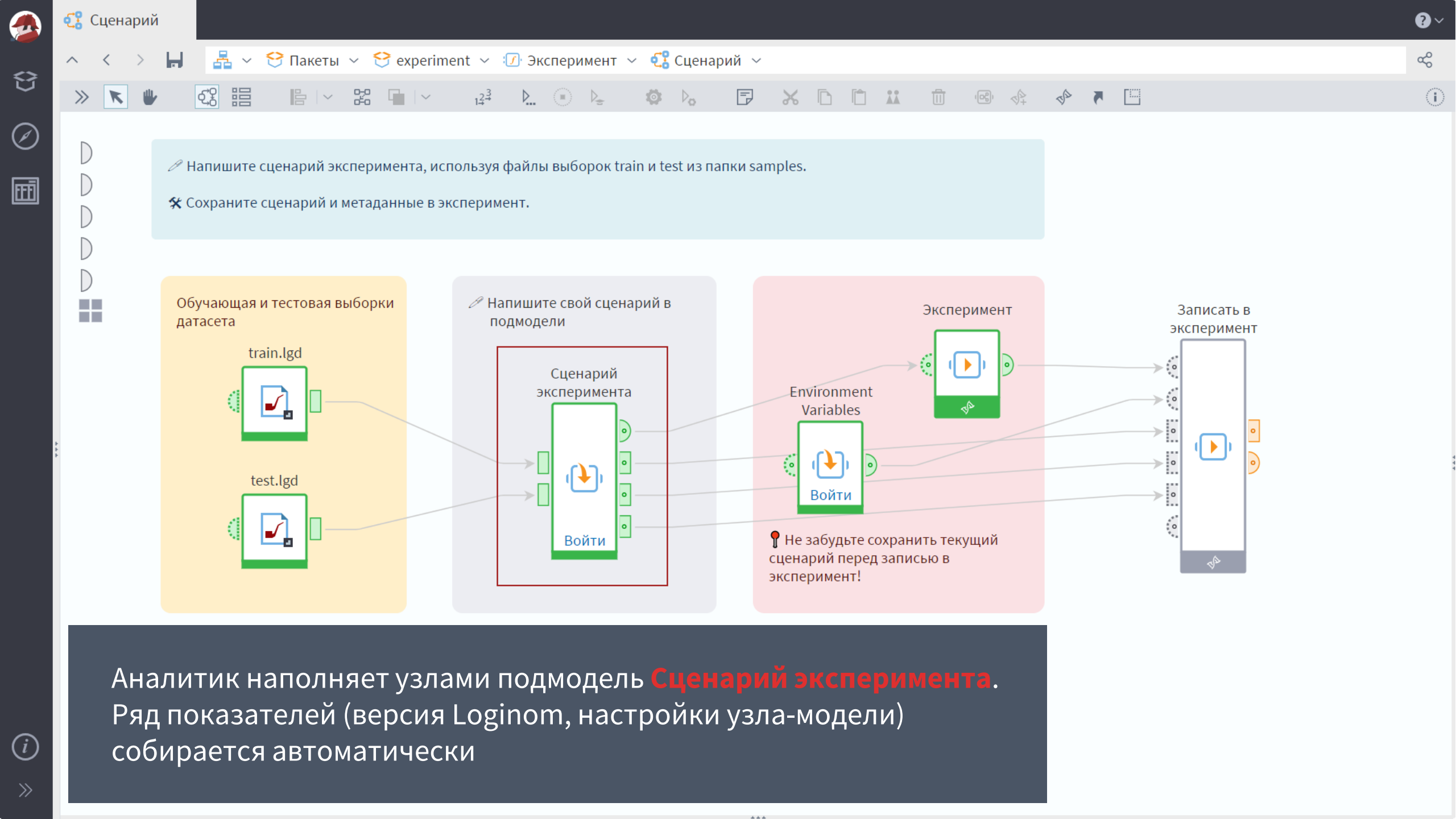
Модули шаблона сценария **experiment.lgp**

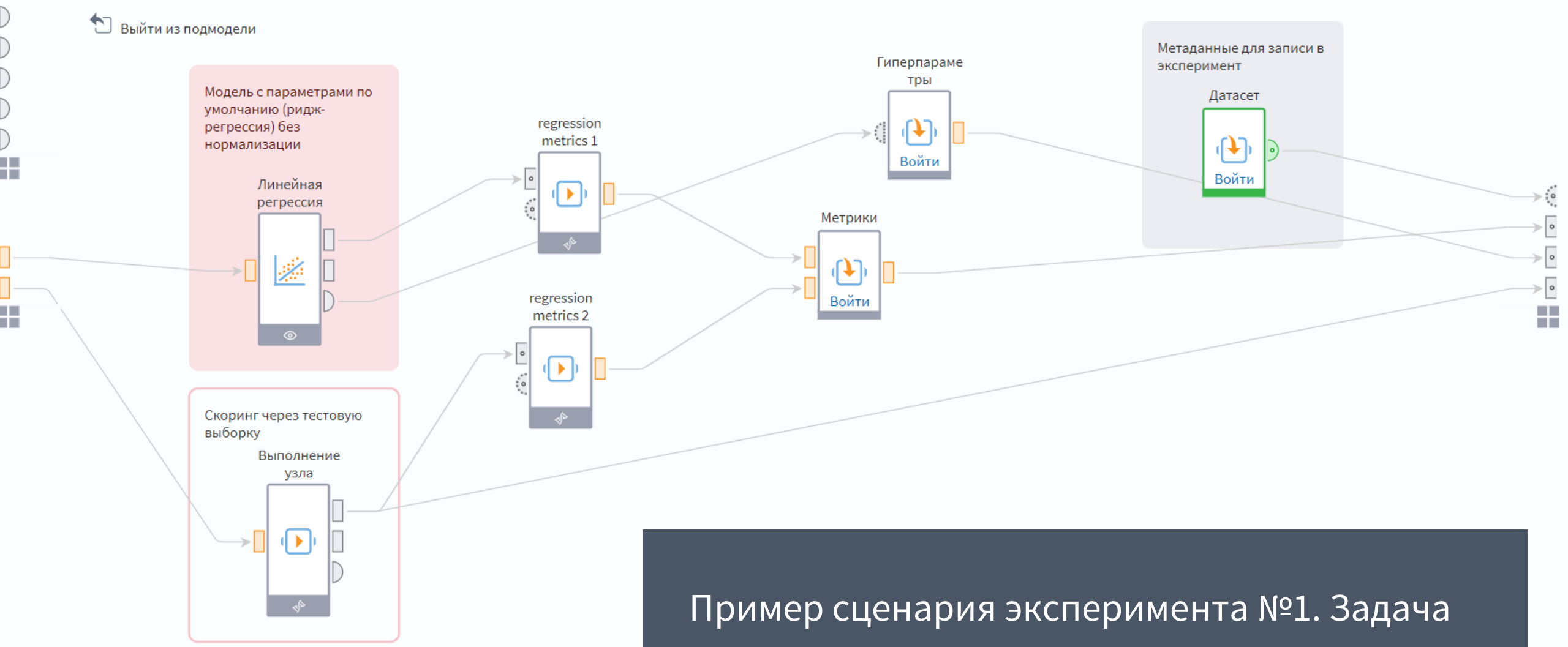


Загрузка локальных копий дочерних выборок **train.lgd** и **test.lgd** или проверка хэша уже имеющихся

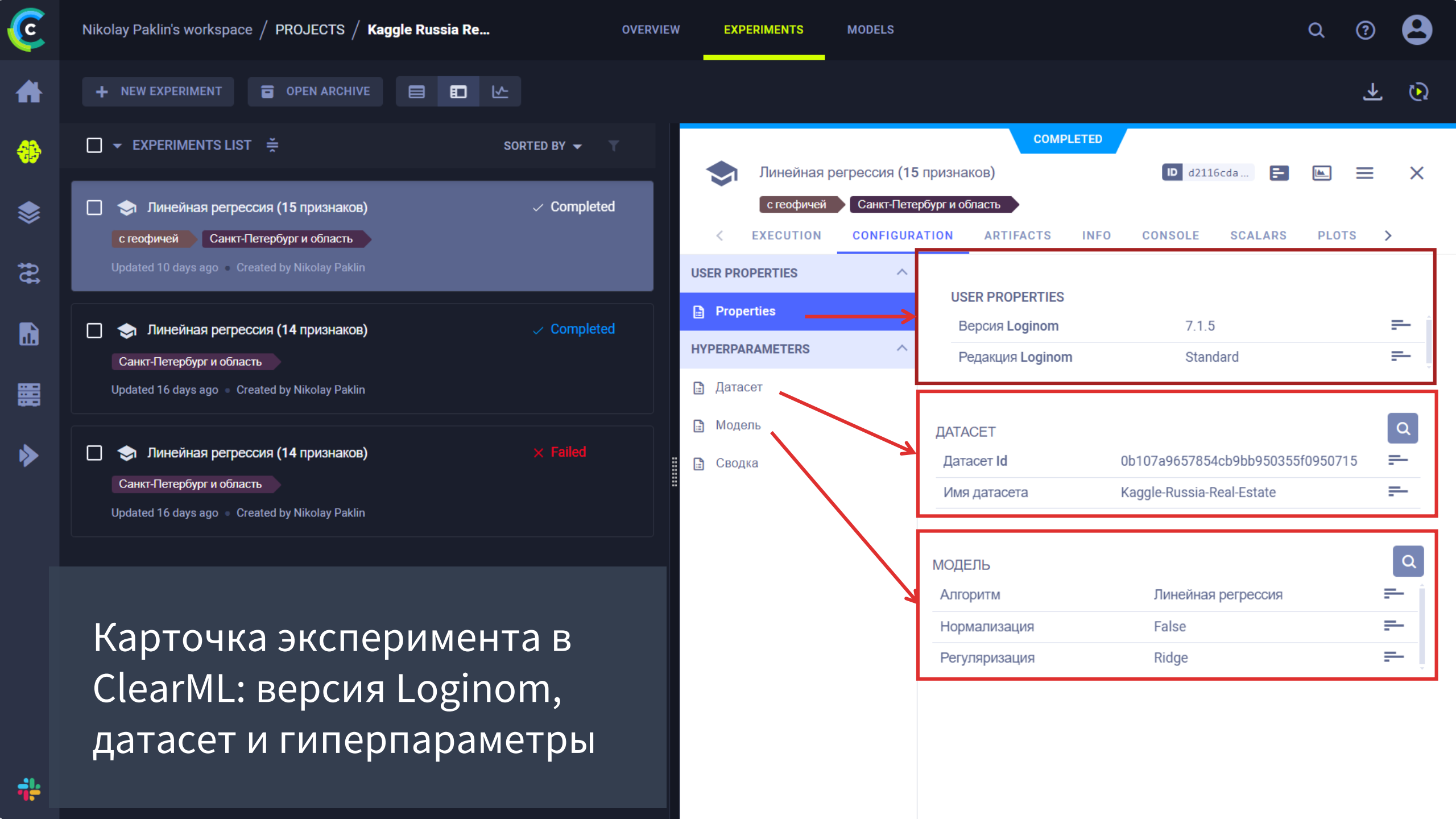
Построение модели и запись в эксперимент:

- сценарий эксперимента
- среда эксперимента (версия Loginom, пакеты-ссылки, версии библиотек…)
- гиперпараметры
- метрики
- диаграмма

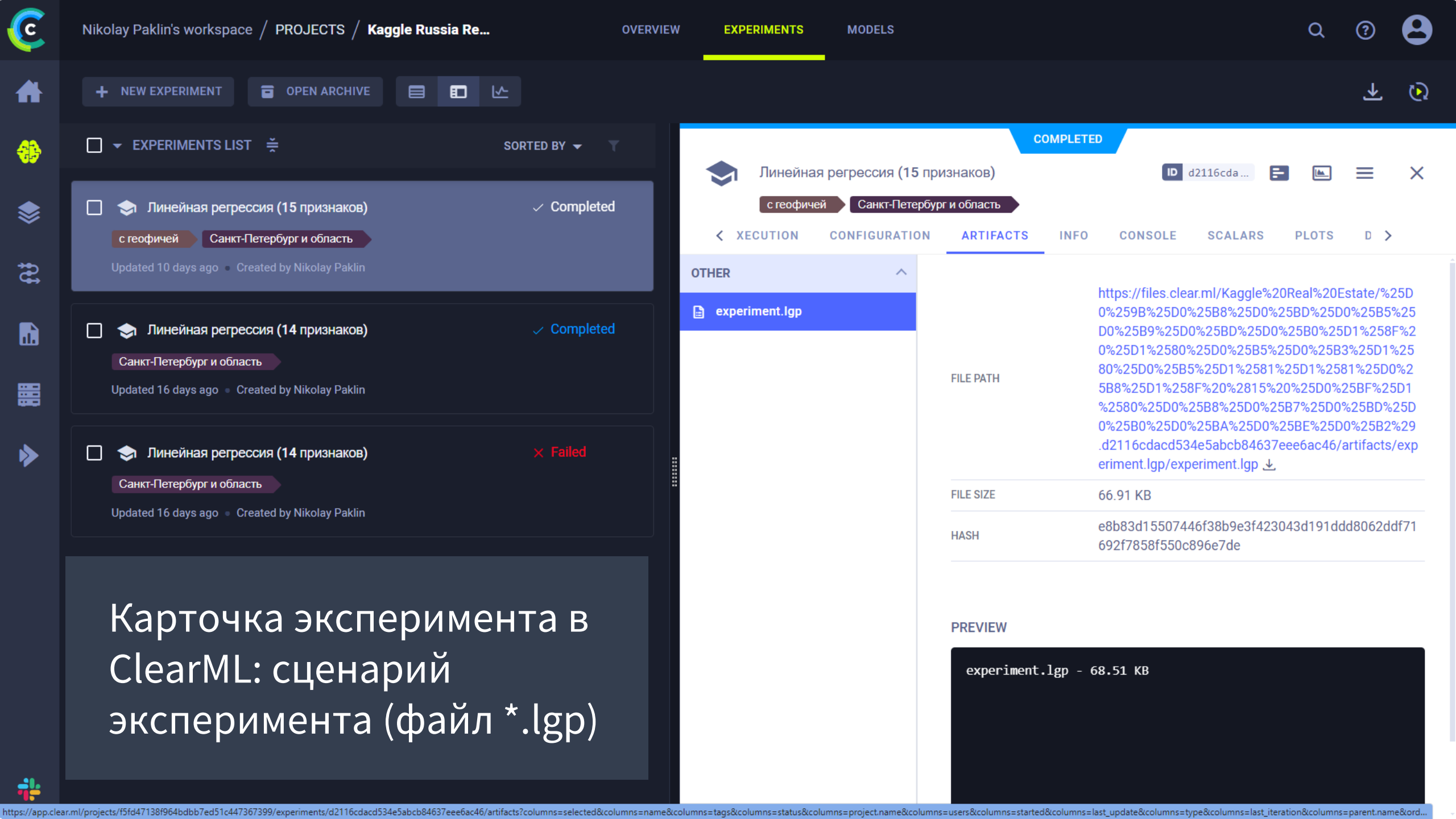

Сценарий
Пакеты
experiment
Эксперимент
Сценарий
Напишите сценарий эксперимента, используя файлы выборок train и test из папки samples.
Сохраните сценарий и метаданные в эксперимент.
Обучающая и тестовая выборки датасета
train.lgd
test.lgd
Напишите свой сценарий в подмодели
Сценарий эксперимента
Войти
Environment Variables
Эксперимент
Войти
Не забудьте сохранить текущий сценарий перед записью в эксперимент!
Записать в эксперимент
Аналитик наполняет узлами подмодель Сценарий эксперимента.
Ряд показателей (версия Loginom, настройки узла-модели)
собирается автоматически



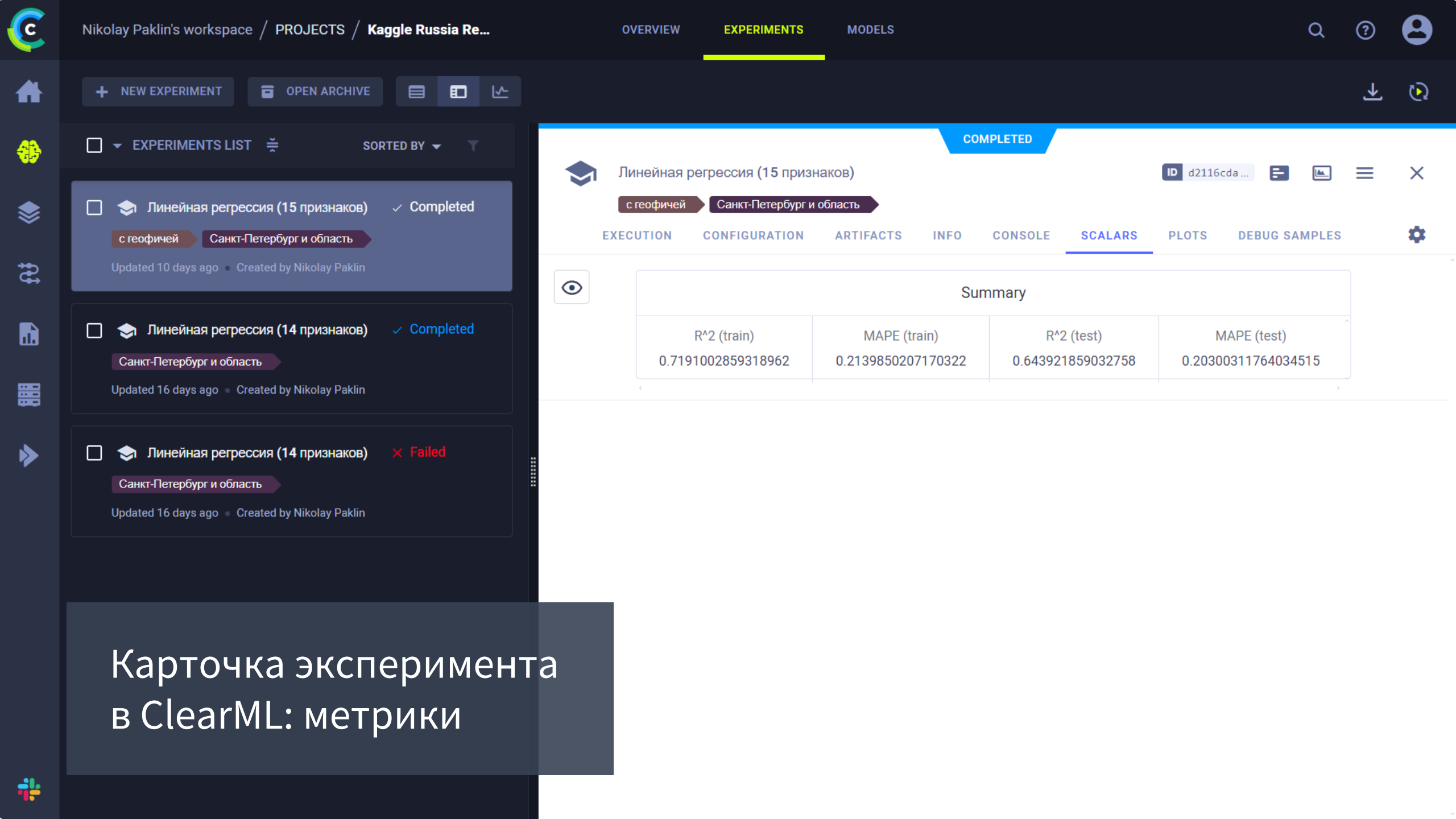
Пример сценария эксперимента №1. Задача оценки стоимости недвижимости, датасет **Kaggle Russia Real Estate 2018-2021**, модель линейной ридж-регрессии

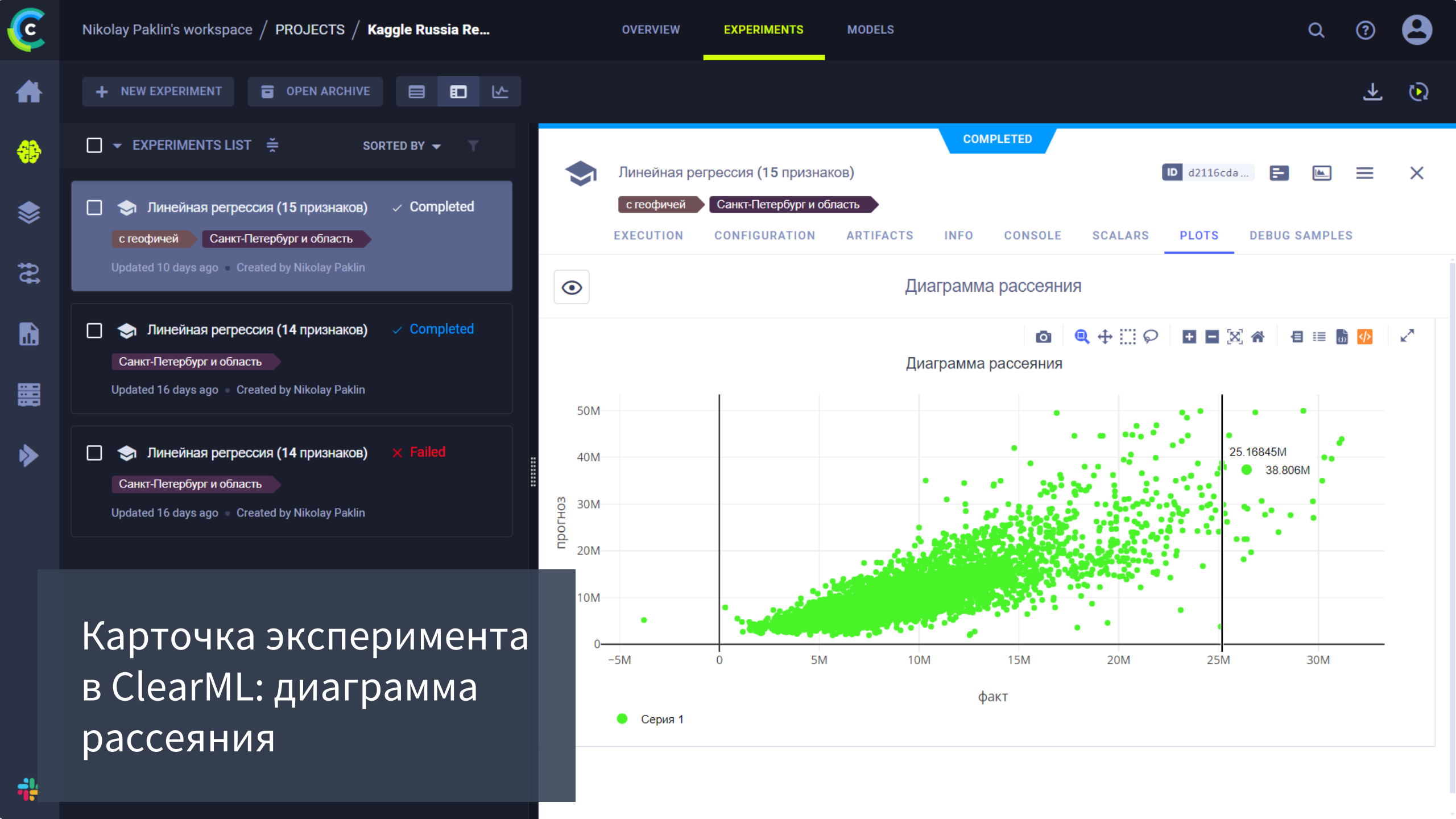


Карточка эксперимента в ClearML: версия Loginom, датасет и гиперпараметры

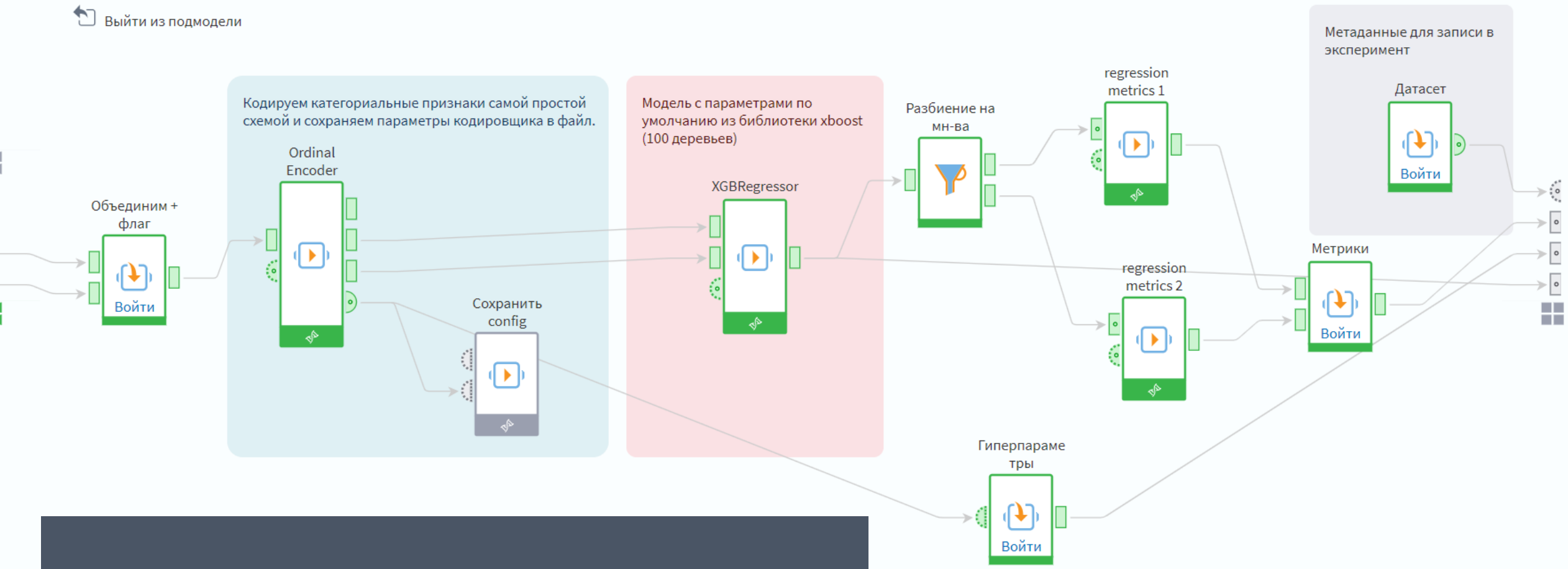


Карточка эксперимента в ClearML: сценарий эксперимента (файл *.lgp)


Nikolay Paklin's workspace / PROJECTS / Kaggle Russia Re...
OVERVIEW
EXPERIMENTS
MODELS
NEW EXPERIMENT
OPEN ARCHIVE
EXPERIMENTS LIST
SORTED BY
Линейная регрессия (15 признаков)
Completed
с геофичей
Санкт-Петербург и область
Updated 10 days ago • Created by Nikolay Paklin
Линейная регрессия (14 признаков)
Completed
Санкт-Петербург и область
Updated 16 days ago • Created by Nikolay Paklin
Линейная регрессия (14 признаков)
Failed
Санкт-Петербург и область
Updated 16 days ago • Created by Nikolay Paklin
COMPLETED
Линейная регрессия (15 признаков)
ID d2116cda...
с геофичей
Санкт-Петербург и область
EXECUTION
CONFIGURATION
ARTIFACTS
INFO
CONSOLE
SCALARS
PLOTS
DEBUG SAMPLES
Summary
R^2 (train)
0.7191002859318962
MAPE (train)
0.2139850207170322
R^2 (test)
0.643921859032758
MAPE (test)
0.20300311764034515
Карточка эксперимента в ClearML: метрики



Карточка эксперимента в ClearML: диаграмма рассеяния



Эксперимент №2. Модель **XGBoostRegressor**, построена в Loginom без кода на основе библиотеки компонентов **Python Kits**

EXPERIMENTS + | Values | Hide Identical Fields | Type to search

Kaggle Russia Real Estate

XGBoost Regressor (15 признаков)

ID 21212ee1...

с геофичей | Санкт-Петербург и область

Published • 0 iterations • Last updated 10 days ago

▼ ДАТАСЕТ

Датасет Id	0b107a9657854cb9bb950355f0950715
Имя датасета	Kaggle-Russia-Real-Estate

› КОДИРОВЩИК

▼ МОДЕЛЬ

Алгоритм	XGBoost Regressor

▼ PROPERTIES

Версия Loginom	7.1.5
Редакция Loginom	Standard
Ссылка-пакет 1	loginom_clearml_kit ^1.0.0
Ссылка-пакет 2	loginom_silver_kit ^3.1.2
Ссылка-пакет 3	loginom_categoty_kit ^3.1.0
Ссылка-пакет 4	loginom_sklearn_kit ^3.1.0
Ссылка-пакет 5	data_prep ^1.0.0

Kaggle Russia Real Estate

Линейная регрессия (15 признаков)

ID d2116cda...

с геофичей | Санкт-Петербург и область

Completed • 0 iterations • Last updated 10 days ago

▼ ДАТАСЕТ

Датасет Id	0b107a9657854cb9bb950355f0950715
Имя датасета	Kaggle-Russia-Real-Estate

▼ МОДЕЛЬ

Алгоритм	Линейная регрессия
Нормализация	False
Регуляризация	Ridge

› СВОДКА

▼ PROPERTIES

Версия Loginom	7.1.5
Редакция Loginom	Standard

Сравнение экспериментов: версии Loginom, датасеты и гиперпараметры

Nikolay Paklin's workspace / PROJECTS / ··· /

DETAILS | HYPERPARAMETERS | SCALARS | PLOTS | DEBUG SAMPLES

EXPERIMENTS + | Graph

Group by: Metric + Variant

Horizontal Axis: Iterations

Smoothing: 0

Exponential Moving Ave...

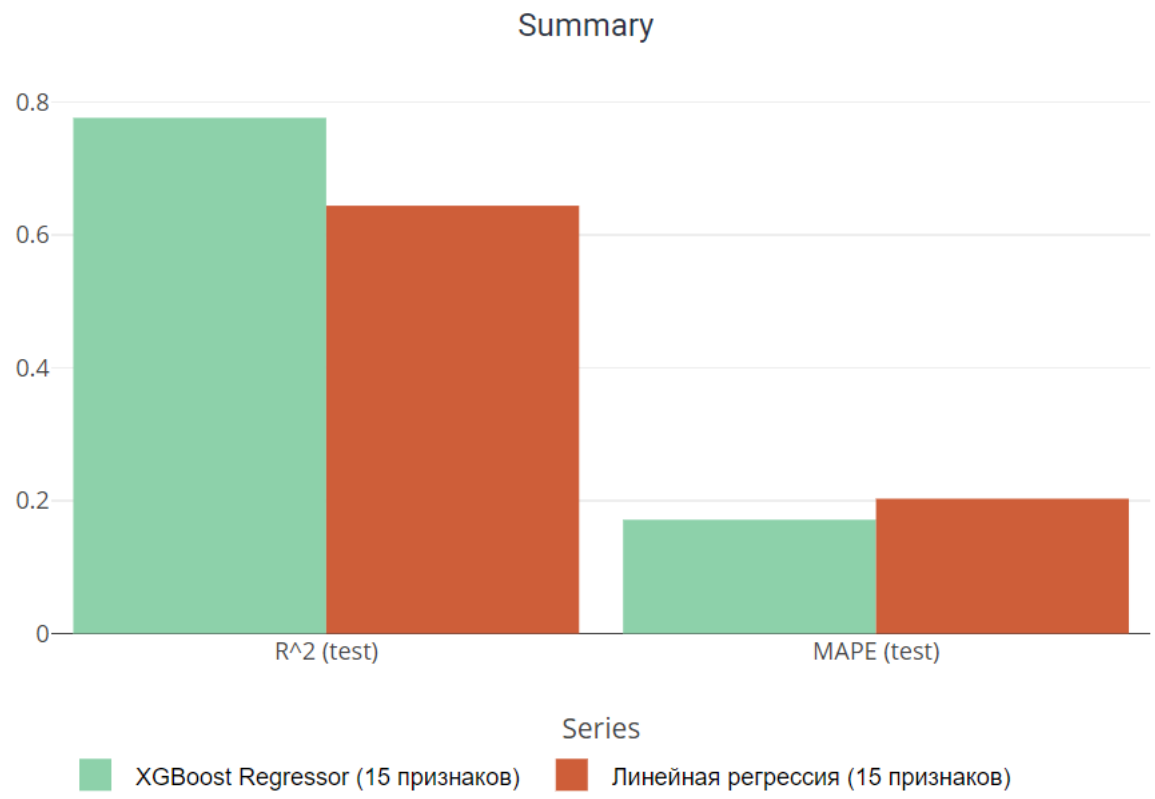
Find scalars

Hide all

Summary

MAPE (test)

R^2 (test)



Сравнение экспериментов: метрики

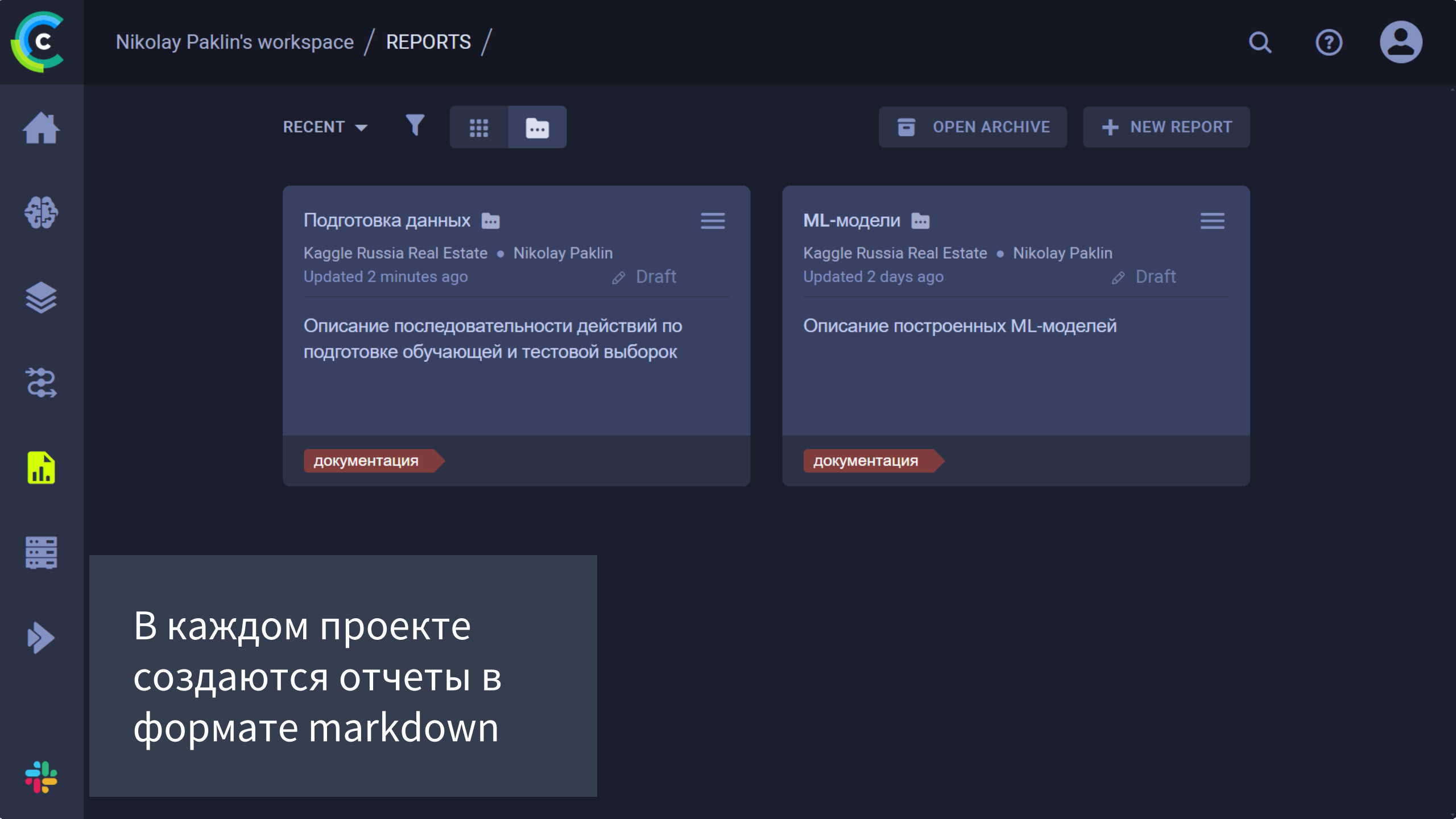
EXPERIMENTS +

Find plots

Hide all

Диаграмма рассеяния - Серия 1

Сравнение экспериментов: диаграммы рассеяния


Nikolay Paklin's workspace / REPORTS /
RECENT
OPEN ARCHIVE
NEW REPORT
Подготовка данных
Kaggle Russia Real Estate • Nikolay Paklin
Updated 2 minutes ago
Draft
Описание последовательности действий по подготовке обучающей и тестовой выборок
документация
ML-модели
Kaggle Russia Real Estate • Nikolay Paklin
Updated 2 days ago
Draft
Описание построенных ML-моделей
документация
В каждом проекте создаются отчеты в формате markdown

Подготовка данных

документация

PUBLISH

Description

Описание последовательности действий по подготовке обучающей и тестовой выборок

EDIT

Описание сценария

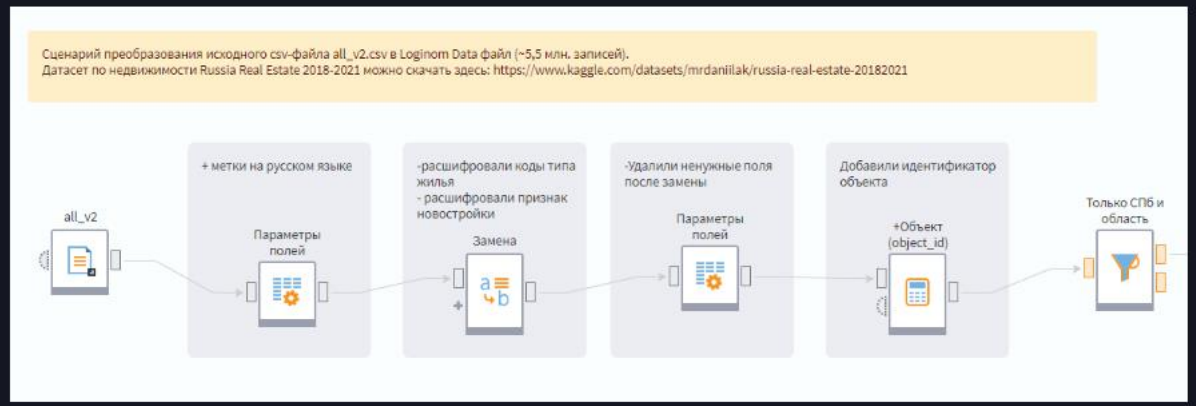
Импорт" "сырых" данных

Датасет объемом ~5,5 млн. записей состоит из 540 тыс уникальных объектов с популярных порталов по продаже недвижимости в России из разных регионов. Набор данных содержит информацию о месторасположении дома, материале, из которого он построен (кирпичный, панельный, деревянный и т.д.), количестве этажей, площади квартиры и ее стоимости.

Набор данных состоит из 13 полей:

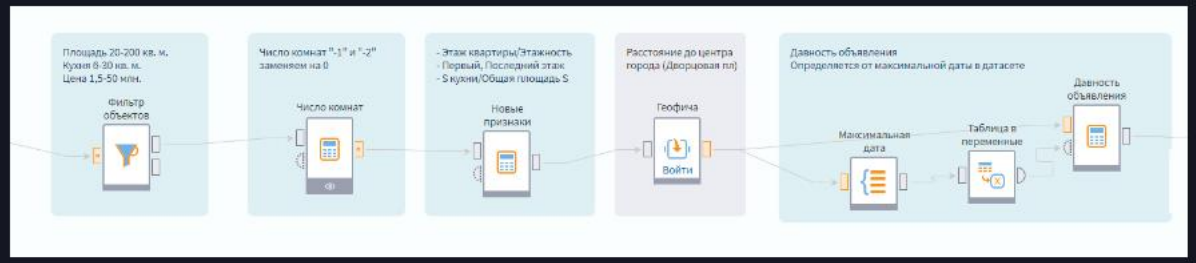
В датасете присутствуют выбросы и ошибки ввода данных.

Сенарий подготовки данных представлен на рисунке ниже.



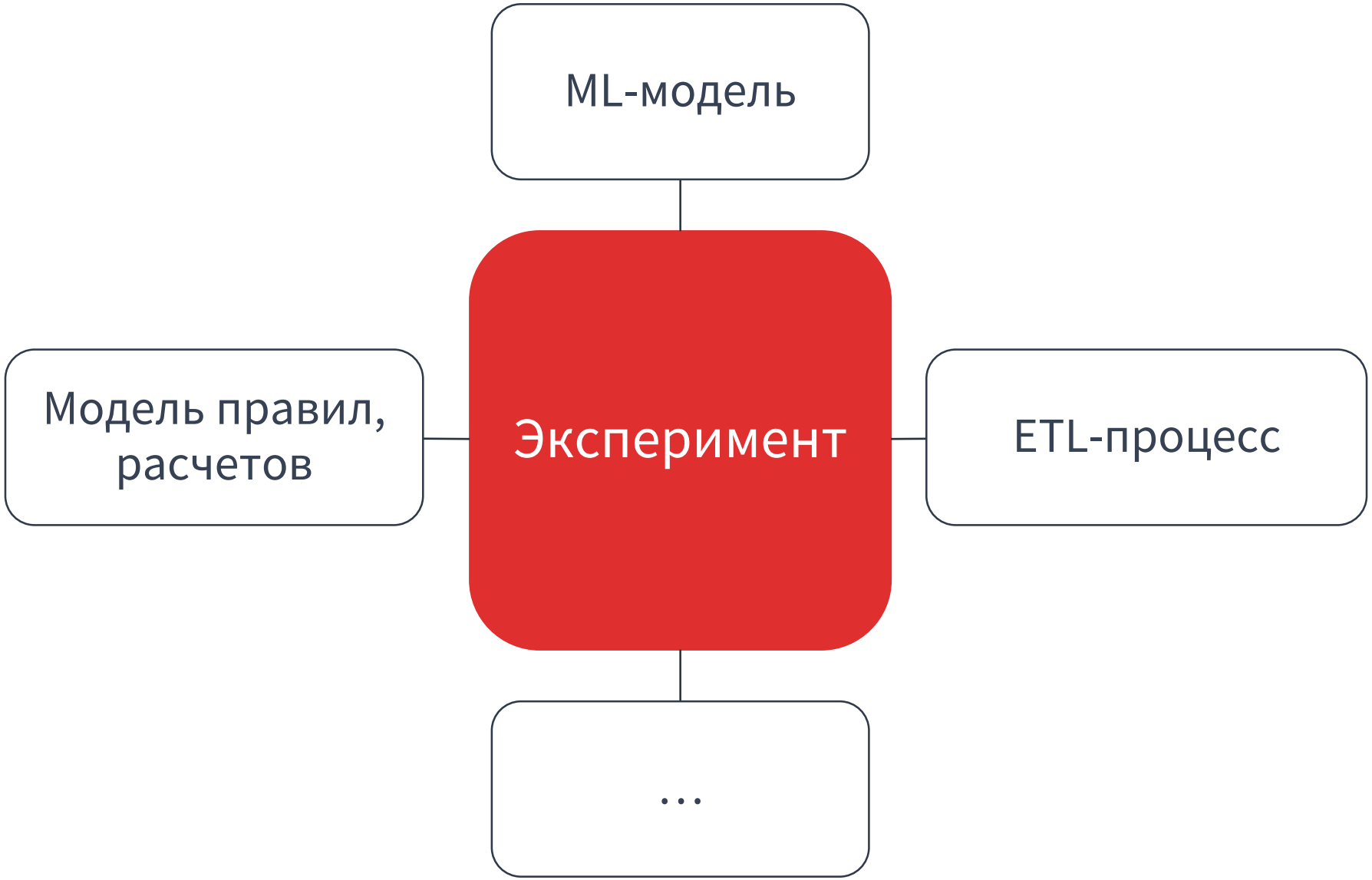
Он вносит косметические изменения в "сырой" датасет и отфильтровывает записи по региону "СПб и область".

Подготовка выборок Для подготовки обучающей и тестовой выборок используется слебующий сценарий.



В нем мы последовательно рассчитываем дополнительные показатели, но первый шаг - очистка от выбросов. Квартиры с общими площадями менее 20 кв. и более 200 метров признаются выбросами.

Пример отчета. После публикации его увидят все пользователи


ML-модель
Модель правил, расчетов
Эксперимент
ETL-процесс
…

ЛЕТО

2024

- Релиз библиотеки компонентов **ClearML Kit**
- Вебинар по ModelOps

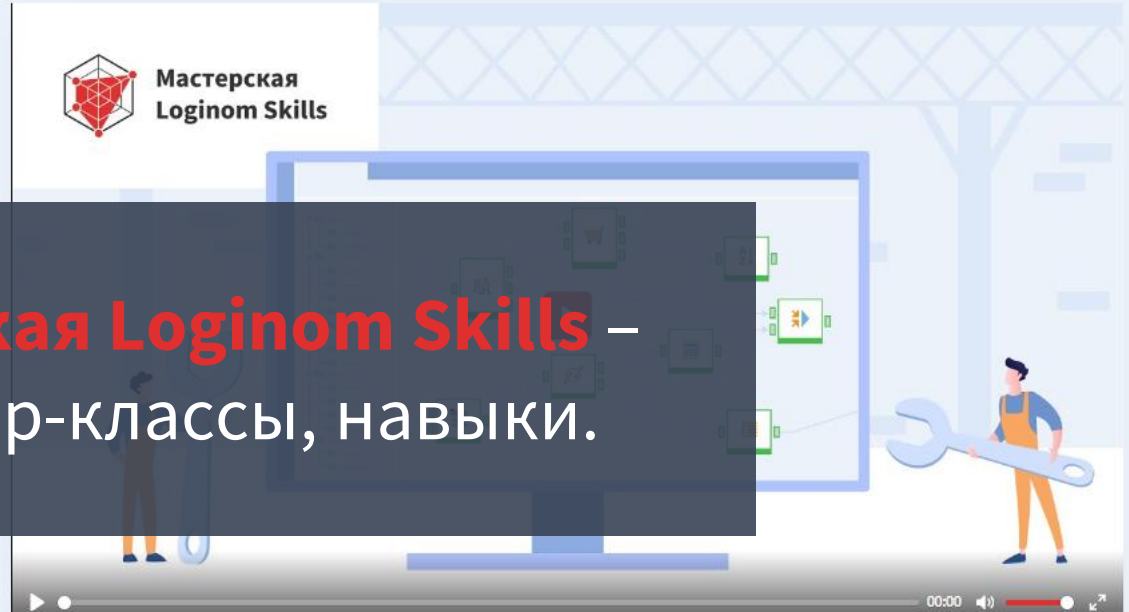
LOGINOM SKILLS

Вход в кабинет

Мастерская Loginom Skills

Онлайн-воркшопы с экспертом

Аналитика данных | Машинное обучение и моделирование | Эффективный зерокодинг



Проект **Мастерская Loginom Skills** – воркшопы, мастер-классы, навыки.

https://t.me/loginom_skills

Спасибо за внимание!