ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Классификация госконтрактов по объектам закупки

Широкова Юлия Сергеевна

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Цели и задачи

1. **Провести загрузку и предобработку сырых исходных данных**
2. **Разработать концептуальное описание системы хранения данных**
3. **Провести разведочный анализ (EDA), подготовку данных и конструирование признаков (Feature engineering)**
4. **Разработать классификатор объектов закупки по ОКПД-2**
5. **Разработать приложение с графическим интерфейсом**

Задача

на основе данных с ftp.zakupki.gov.ru научиться определять группу, к которой относится контракт с кодом ОКПД-2 41, 42, 43, 71.1.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

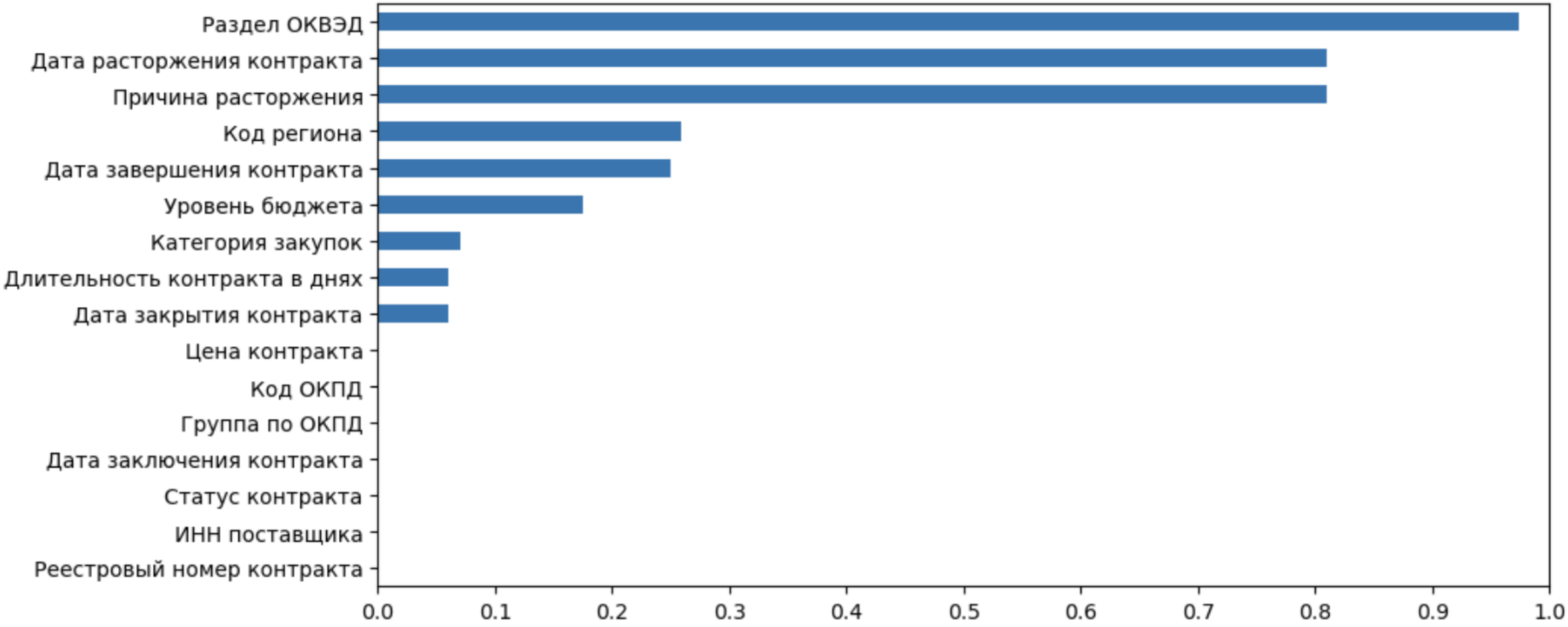
Загрузка данных

Датасет содержит избыточное число данных (более 20 млн. строк).

1. Определены необходимые для решения задачи признаки.
2. Составлен алгоритм выгрузки и предобработки данных чанками.
3. Добавлено новое поле с длительностью контракта в днях.
4. Сохранен первичный датасет в parquet формате (staging слой данных).

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Процентное содержание пропусков в данных



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Методы используемые в работе

Подзаголовок слайда

- **Токенизация** — это процесс разделения текста на отдельные элементы (токены), такие как слова, символы или числа.
- **Стемминг** – метод нормализации текста, который приводит слова к их базовой или корневой форме. В работе используется SnowballStemmer.
- **Stop Words Removal** — это метод удаления стоп-слов (наиболее распространённые слова в языке, которые не несут особого смысла) из текста.
- **Градиентный бустинг CatBoost** — это метод машинного обучения, разработанный компанией Yandex, основанный на деревьях принятия решений для повышения точности модели. Он отличается от других методов машинного обучения тем, что он использует категориальные признаки без дополнительных преобразований и умеет работать с текстовыми данными.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных (EDA)

Overview | Alerts 11 | Reproduction

Dataset statistics

Number of variables	10
Number of observations	250000
Missing cells	334949
Missing cells (%)	13.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	302.7 MiB
Average record size in memory	1.2 KiB

Variable types

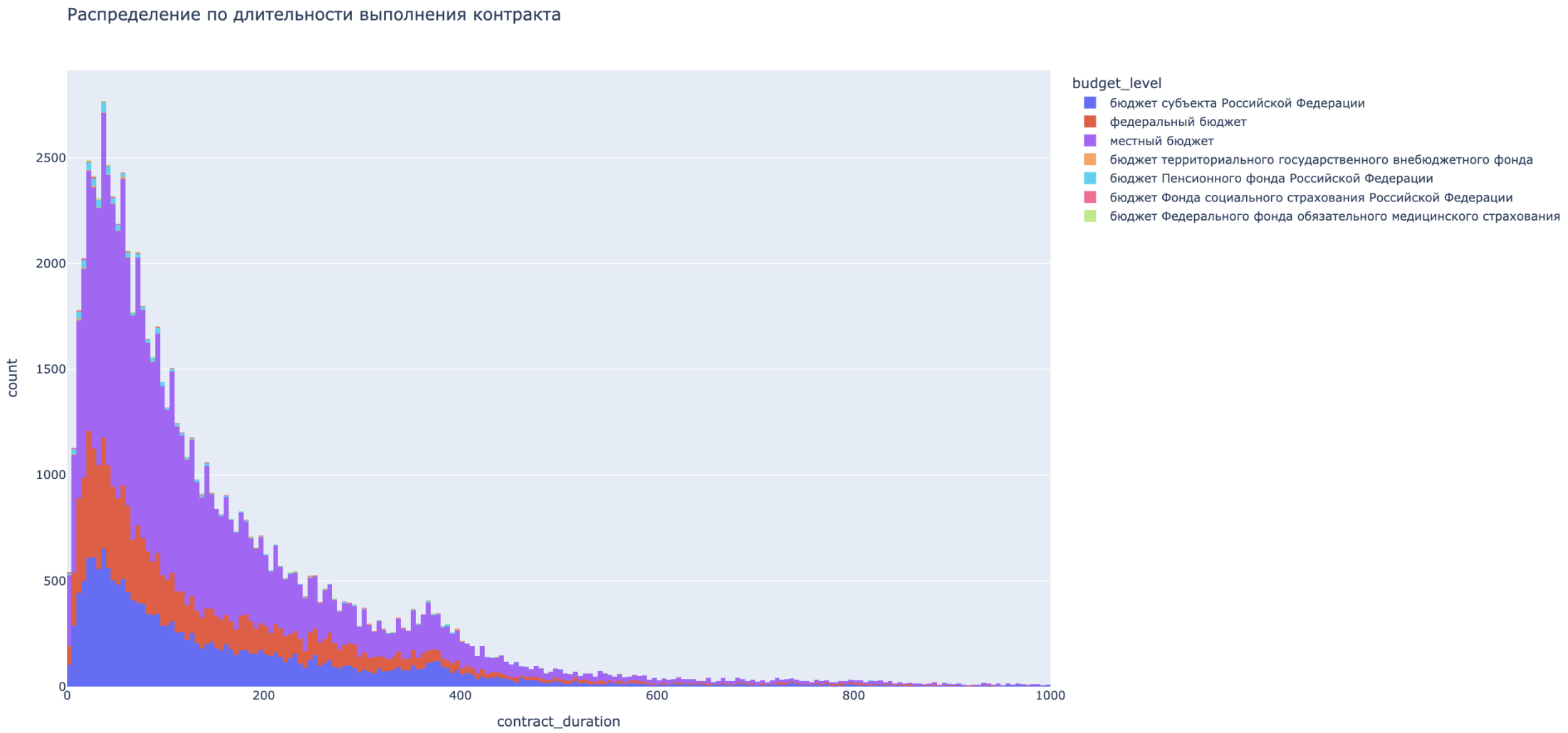
Numeric	5
Categorical	4
Text	1

Большая часть EDA выполнена с использованием библиотеки **ydata_profiling.**

В работе строились гистограммы и диаграммы размаха, проводился корреляционный анализ данных.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

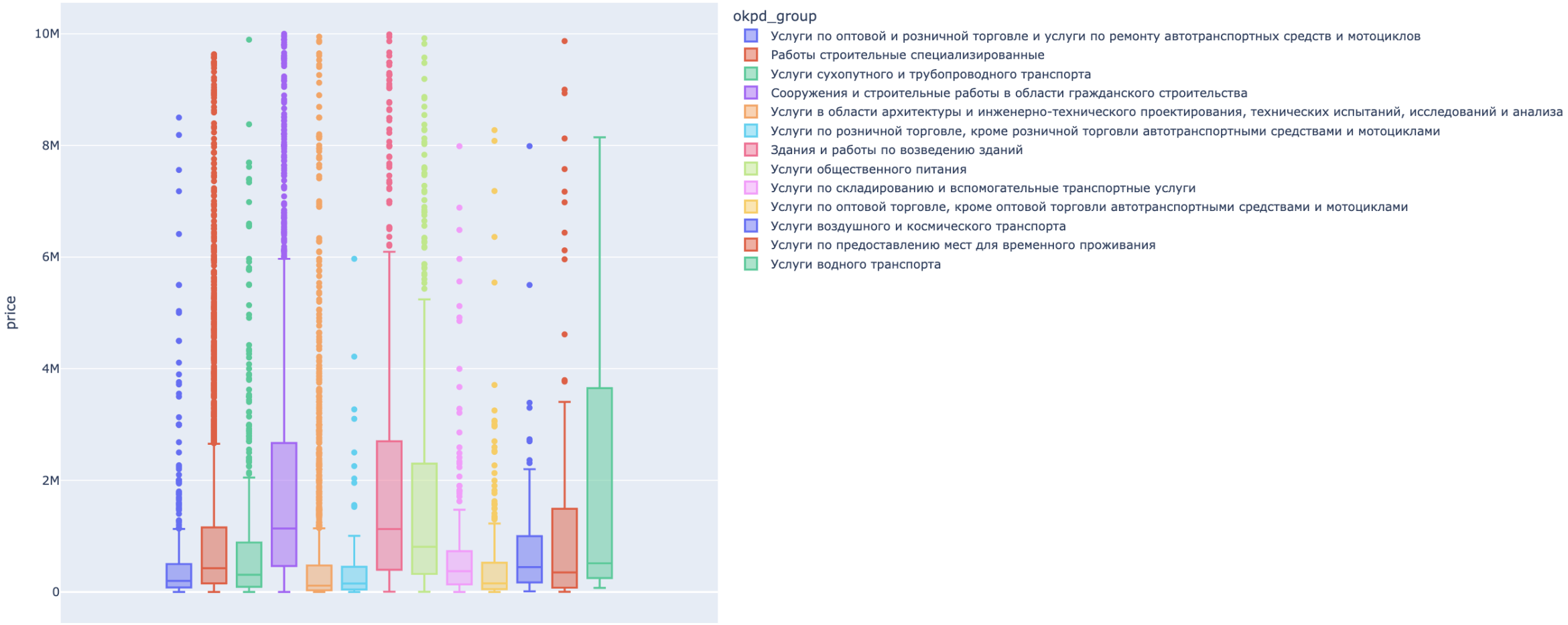
Гистограмма по длительности контракта в днях



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Диаграмма размаха по стоимости контракта

Диаграмма размаха для стоимости контракта



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Концептуальное описание системы хранения данных

Там выглядит схема хранения и обработки данных по мнению Алисы (Yandex.GPT) по запросу 😊 :

«Нарисуй схему в которой: пользователь обращается к веб сервису на flask python, сервис соединён с базой данный PostgreSQL и кэшем Redis. При этом слева есть API интерфейс с системой оркестрации AirFlow»

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка, обучение и тестирование модели

Обучение модели проводилось в облачном сервисе Yandex DataSphere

```
# Начальные параметры
params={
    'task_type': 'GPU',
    'devices': '0:1',
    'random_seed': RANDOM_STATE,
    'objective': 'MultiClass',
    'loss_function': 'MultiClass',
    'eval_metric': 'TotalF1',
    'early_stopping_rounds': 100,
    'logging_level': 'Silent',
}
```

```
# Сетка параметров для поиска
param_grid = {
    'depth': range(6,11,1),
    'bootstrap_type': ['Poisson', 'Bayesian'],
    'random_strength': np.linspace(0.7, 1, 11),
    'learning_rate': [0.001, 0.01],
}
```

```
Результ после 30 итераций рандомизированого поиска:
{'random_strength': 0.94, 'depth': 10,
'bootstrap_type': 'Poisson', 'learning_rate': 0.01}
```

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Результаты модели на тестовой выборке

```
model = CatBoostClassifier(**params)
```

	precision	recall	f1-score	support
ДРУГИЕ	0.95	0.94	0.94	112711
ПИР	0.97	0.98	0.98	215208
СМР	0.91	0.91	0.91	45174
accuracy			0.96	373093
macro avg	0.94	0.94	0.94	373093
weighted avg	0.96	0.96	**0.96**	373093

Классификатор объектов закупки по ОКПД-2

ИНН исполнителя

7107112020

Стоимость контракта

184884

Код региона

71

Длительность контракта в днях

28

Статус исполнения контракта

Контракт исполнен

Уровень бюджета

федеральный бюджет

Объект закупки

Проектные изыскательные работы по объекту строительства стадиона в р.п. Александро-Невский Александро-Невского муниципального района Рязанской области

Проверить

Результат соответствия: ПИР

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru