ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ МГТУ им. Н.Э. Баумана

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Инженер данных (Data engineer Pro)»

Слушатель _____Широкова Ю. С.

Москва, 2024

ЗАДАНИЕ НА ВКР

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

Группы могут быть следующими:

1. Строительно-монтажные работы (СМР).
2. Проектно-изыскательские работы (ПИР).
3. Строительный надзор.
4. Подключение коммуникаций.
5. Прочее.

По ОКПД-2 контракты в общем случае должны разделяться так:

1. Строительно-монтажные работы (СМР) - 41, 42, 43 (кроме нижеперечисленных):
    а. Проектно-изыскательские работы (ПИР) - 41.1, 71.1.
    б. Подключение коммуникаций - 43.22.
    в. Строительный надзор – четкой группы нет.

Требуется:

1. Изучить теоретические основы и методы решения поставленной задачи.
2. Написать скрипты для загрузки и обработки исходных данных. Провести разведочный анализ (в соответствии со спецификой задачей). Проверить наличие пропусков и дубликатов.
3. Разработать концептуальное описание системы хранения данных, механизмов обновления, забора и обработки в рамках поставленной задачи.
4. Разработать классификатор (от двух и более) объектов закупки по ОКПД-2. Обучить нескольких моделей для классификации объектов закупки. При построении модели необходимо 20-30% данных оставить на тестирование модели, на остальных происходит обучение моделей.
5. Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в

задании 4 (минимальный функционал: ввод данных и результат классификации).

6. Создать профиль на github.com и разместить практическую часть работы. Оформить файл README.
7. Написать пояснительную записку к проекту.

Проблема:

Далеко не всегда контракты указываются с нужным кодом, поэтому есть проблема как такие контракты «отловить» и определить в нужную группу. Поэтому задача предполагает классификацию контрактов на основе объекта закупки, который сформулирован естественным языком. Также предполагаем, что могут иметь значение цена контракта и его длительность.

Датасет с карточками госконтрактов:

https://drive.google.com/file/d/1sRHx27O3NgTivrrQHdBdTAqxdCYNmARW/view?usp=sharing

КАЛЕНДАРНЫЙ ПЛАН

№	Наименование раздела или этапа	Трудоёмкость в % от полной трудоёмкости работы	Срок выполнения
1	Изучение теоретических основ и методов решения поставленной задачи	10	01.09.2024
2	Написание скрипта для загрузки и обработки исходных данных. Проведение разведочного анализа (EDA).	30	20.09.2024
3	Разработка концептуального описания системы хранения данных, механизмов обновления, забора и обработки в рамках поставленной задачи	10	01.10.2024
4	Разработка классификатора объектов закупки по ОКПД-2.	20	15.10.2024
5	Разработка приложение, которое будет выдавать прогноз	10	20.10.2024
6	Создание профиля на github.com с README и размещение практической часть работы.	5	05.11.2024
7	Написать пояснительную записку к проекту.	15	05.11.2024

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

Тема:

Классификация госконтрактов по объектам закупки.

Описание:

В соответствии с Федеральным законом «О контрактной системе» (44-ФЗ) государственный контракт (госконтракт) — это соглашение между поставщиком и органами власти федерального, регионального или муниципального уровней. Карточки госконтрактов хранятся в специальном реестре. Карточка госконтракта содержит информацию о контракте: описание, сроки, исполнителей и т.д. В том числе хранится общероссийский классификатор продукции по видам экономической деятельности (ОКПД-2). Зачастую ОКПД-2 заявляется ошибочный. В этом заключается проблема – выделить ошибочно обозначенные госконтракты и переназначить их ОКПД-2

Входными данными в работе являются подготовленные гос. заказчиком данные карточек госконтрактов с ftp.zakupki.gov.ru. В результате выполнения задания получен классификатор госконтрактов по ОКПД-2 в соответствии с технических заданием, разработано приложение, позволяющее на основе входных данных предоставлять пользователю прогноз.

Актуальность: Создание классификатора госконтрактов по ОКПД-2 в соответствии с объектом закупки, позволит эффективно перераспределять контракты по соответствующим им группам.

2 АНАЛИТИЧЕСКАЯ ЧАСТЬ

2.1 Постановка задачи

Данные выданы заказчиком в архиве, содержащим .csv файл. Описание полей не предоставлено, анализ и сопоставление полей было выполнено во время ознакомления с объектами закупки на официальном сайте https://zakupki.gov.ru (объект закупки 10105031366210000017, объект закупки 10268021734210000064).

Датасет содержит избыточное число данных (более 20 млн. строк). Выгрузка производилась кусками (чанками) по 10000 строк, из неё отфильтровывались строки по требуемому коду ОКПД. Целевой переменной в явном виде в датасете не было, в качестве входных факторов выбраны: «Реестровый номер контракта», «ИНН поставщика», «Код региона», «Статус контракта», «Категория закупок», «Уровень бюджета», «Дата заключения контракта», «Дата завершения контракта», «Дата расторжения контракта», «Причина расторжения», «Цена контракта», «Раздел ОКВЭД», «Группа по ОКПД». «Код ОКПД».

Часть полей содержит пропуски. Больше всего их в ОКВЭД и в полях, связанных с расторжением контракта, что очевидно так как большая часть контрактов завершается успешно.

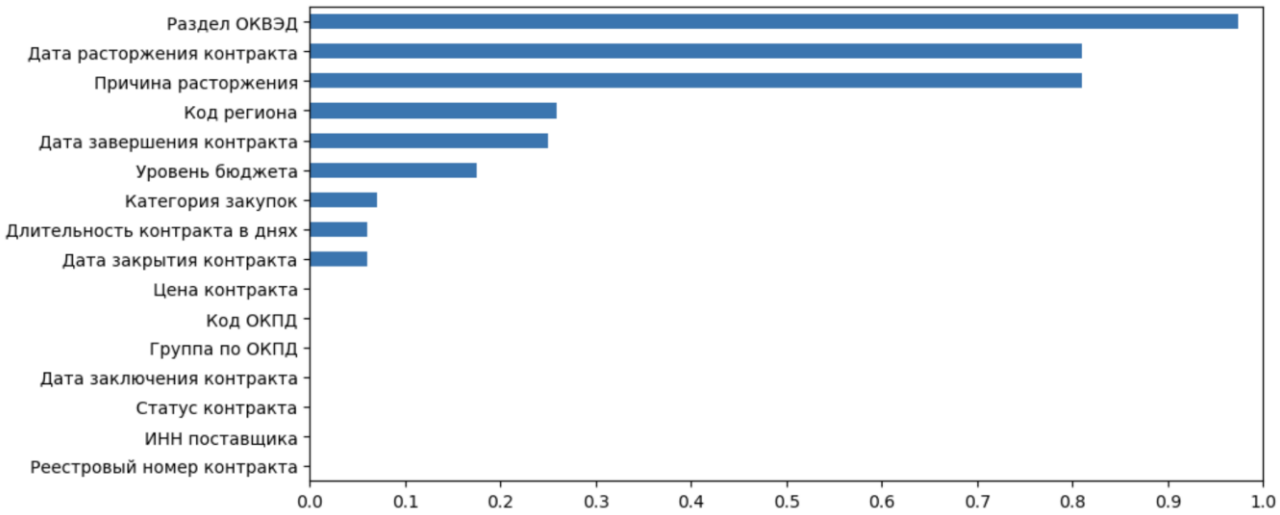

Рисунок 1 – Пропуски в данных в процентном соотношении

Вместо признаков «Даты завершения контракта» и «Даты расторжения контракта» определено новое поле с длительностью контракта в днях. Предварительный датасет для staging слоя содержит 2.5 млн. строк и 10 столбцов.

	inn	region_code	status	category	budget_level	contract_duration	price	okpd_group	okpd_code	termination_reason
1110308	1831190169	18	Контракт исполнен	Выполнение работ по проведению инженерных изыс...	федеральный бюджет	135.000	387171.000	Услуги в области архитектуры и инженерно-техни...	71.1	NaN
1395494	278109628	02	Контракт исполнен	Выполнение аварийно-восстановительных работ и ...	федеральный бюджет	349.000	660000.000	Работы строительные специализированные	43.2	NaN
370444	121508738797	NaN	Исполнение контракта прекращено	Поставка бензина АИ-92,АИ-95	федеральный бюджет	70.000	121229.700	Услуги по оптовой торговле, кроме оптовой торг...	46.7	Соглашение сторон
974771	7743529527	NaN	Контракт исполнен	Поставка бензина марки АИ - 92 для служебного ...	местный бюджет	79.000	98550.000	Услуги по розничной торговле, кроме розничной ...	47.3	NaN
1638609	344344989175	NaN	Исполнение контракта прекращено	Техническое обслуживание и ремонт автотранспорта	федеральный бюджет	253.000	150000.000	Услуги по оптовой и розничной торговле и услуг...	45.2	Соглашение сторон

(2556154, 10)

Рисунок 2 – Первичный датасет, случайная выборка из 5 элементов

2.2 Описание используемых методов

В работе используются следующие методы: градиентный бустинг, токенизация, удаление шумовых слов (Stop Words Removal) и стемминг. Большинство из них обязательны в решении задач распознавания речи, машинного перевода и классификации текстов.

Токенизация — это процесс разделения текста на отдельные элементы (токены), такие как слова, символы или числа. Токен — это наименьшая единица, которая имеет смысл в контексте обрабатываемой информации. Токенизация используется для обработки текстовых данных и анализа естественного языка.

В процессе токенизации текст разбивается на токены с использованием определённых правил и алгоритмов. Эти правила могут быть основаны на грамматике языка, структуре предложения или других критериях. После разделения текста на токены они могут быть обработаны дальше, например, для выполнения операций над словами, составления словаря или проведения статистических анализов.

Токенизация может быть реализована с помощью различных инструментов и технологий, таких как библиотеки и фреймворки для обработки естественного языка, а также специализированные программы и сервисы. В

настоящей работе использует наиболее распространённый пакет **nltk** – метод **word_tokenize**.

Стемминг – метод нормализации текста, который приводит слова к их базовой или корневой форме. Он помогают уменьшить количество уникальных слов в тексте, что упрощает его анализ и обработку. Метод проще чем лемматизация, которая приводит каждое слово к его основной форме, соответствующей словарному слову.

Таблица 1 – Преимущества и недостатки методов нормализации текста

Метод	**Плюсы**	**Минусы**
Стемминг	простота реализации; быстрота выполнения.	возможные неточности в выходных данных.
Лемматизация	более высокая точность результатов; учёт части речи слова для правильного анализа.	сложность реализации; медленное выполнение.

Удаление шумовых слов (Stop Words Removal) — это метод удаления стоп-слов из текста. Стоп-слова — это наиболее распространённые слова в языке, которые не несут особой смысловой нагрузки и не влияют на общий смысл текста. Они включают предлоги, союзы, артикли и другие служебные слова.

Стоп-слова удаляются из текста для улучшения качества анализа текста и повышения точности моделей машинного обучения. Например, при классификации текста на категории или определении тональности текста.

Существует несколько способов удаления стоп-слов:

1. Ручное удаление: человек самостоятельно удаляет стоп-слова из текста.
2. Статистический метод: используются статистические данные о частоте встречаемости стоп-слов в тексте для автоматического определения и удаления ненужных слов.
3. Удаление на основе корпуса nltk.corpus, в котором есть stopwords.

В работе применятся третий метод, он прост в реализации. Отсутствие сложной стилистической информации в текстовые данных делает данный метод предпочтительнее.

Градиентный бустинг CatBoost — это метод машинного обучения, разработанный компанией Yandex. CatBoost широко используется в различных задачах машинного обучения, таких как классификация, регрессия и ранжирование. Он основан на деревьях принятия решений и использует градиентный бустинг для повышения точности модели.

CatBoost отличается от других методов машинного обучения тем, что он использует категориальные признаки без дополнительных преобразований. Это позволяет обрабатывать данные с большим количеством категориальных признаков и повышает эффективность модели. Более того метод умеет работать с текстовыми данными (параметр text_features) с поддержкой GPU что делает его альтернативой применению нейросетей, особенно если нет предобученной под решаемую задачу модели. В датасете, полученном от заказчика, все признаки за вычетом стоимости относятся категориальным или текстовым, поэтому в работе выбран CatBoost.

Основные преимущества CatBoost:

- Высокая скорость обучения и предсказания.
- Хорошая обобщающая способность.
- Возможность работы с пропущенными значениями.
- Поддержка параллельных вычислений.
- Гибкость в выборе гиперпараметров.
- Встроенные методы оптимизации гиперпараметров и кросс-валидация.

2.3 Разведочный анализ данных

Разведочный анализ данных (EDA) — это анализ основных свойств данных, поиск общих закономерностей, распределений и аномалий, построение начальных моделей с использованием инструментов визуализации. Основные методы EDA включают изучение вероятностных распределений переменных,

построение и анализ корреляционных матриц, факторный анализ, дискриминантный анализ и многомерное шкалирование.

Разведочный анализ данных проводился с помощью таких библиотек как ydata_profiling, matplotlib, pandas и plotly. Библиотека ydata_profiling составляет полноценный отчет, покрывающий весь EDA, его можно найти в git-репозитории в директории data в формате html. В работе строились гистограммы и диаграммы размаха, устранялись дубликаты, аномали и выбросы, проводился корреляционный анализ, благодаря которому часть признаков в последствии было отброшено или преобразовано (feature engineering).

Методы разведочного анализа данных (EDA), использованного в настоящей работе:

1. Описательная статистика: вычисление средних значений, стандартных отклонений, минимальных и максимальных значений для каждой переменной.
2. Гистограммы: графическое представление распределения данных, позволяющее увидеть форму, асимметрию и эксцесс.
3. Коробчатые диаграммы (диаграммы размаха): графическое представление распределения данных с указанием квартилей и межквартильного размаха. По ним хорошо видный выбросы.
4. Корреляционные матрицы: анализ корреляции между всеми переменными, позволяющий выявить сильные и слабые связи. По связям удобно восстанавливать пропуски или убирать сильно коррелирующие между собой признаки.

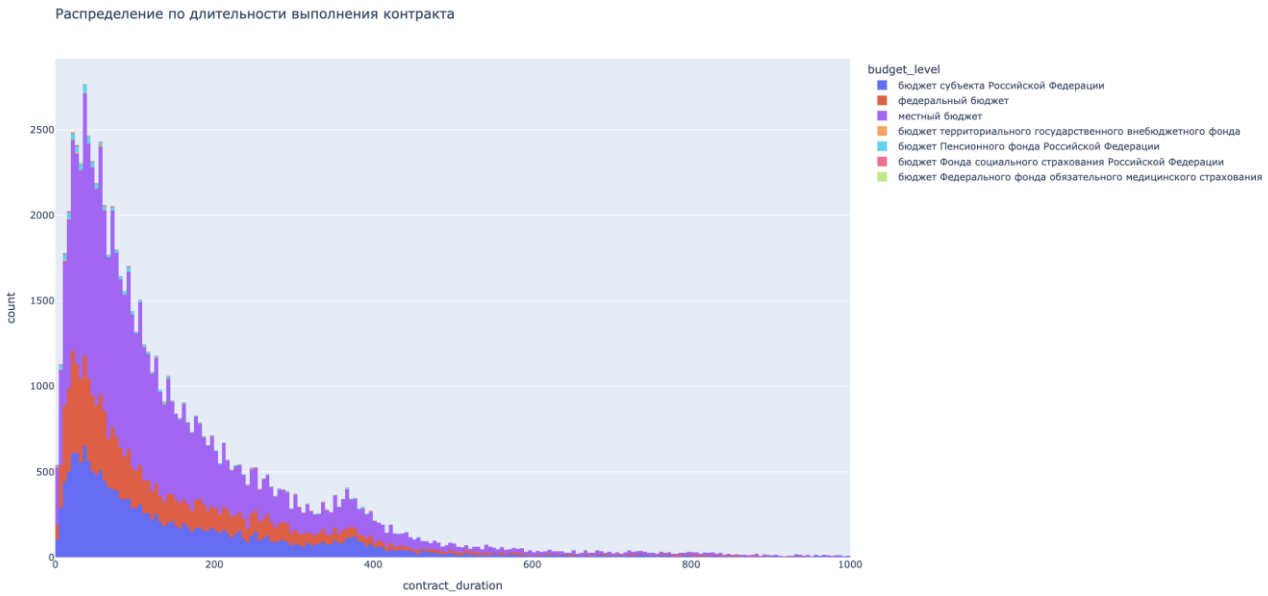

Рисунок 3 – Гистограмма длительности контракта

Рисунок 4 – Диаграмма размаха по стоимости выполнения контракта

Концептуальное описание системы хранения данных. Данные следует подвергнуть нормализации и хранить в SQL базе данных, например в PostgreSQL. В самой БД нужно выделить staging слой сырых данных, поступающих с сайта zakupki.gov и слой детальных данных (DDS) с очищенной информацией, информации об ИНН и регионах лучше загружать в кеширующий сервис (Redis). Обновление данных в БД можно сделать в системе оркестрации AirFlow через даги на получения обновлений (API к госзакупкам), записи данных в raw слой и трансформацию в DDS слой. Дообучение модели следует реализовать через систему версионирования данных Data Version Control (DVC).

3 ПРАКТИЧЕСКАЯ ЧАСТЬ

3.1 Предобработка данных

Удалены выбросы по стоимости контракта, отфильтрованы все контракты со стоимость выше 250 млн. руб. или меньше нуля. Удалены полные дубликаты (когда в 2 и более строках все поля совпадают). Обнаружены контракты с отрицательной длительность, они так же исключены.

Применена бинификация, созданы два новых признака:

1. region_size – определяет степень крупности региона на основе количества контрактов по полю region_code.
2. executor_size – определяет статус исполнителя контракта на основе количества контрактов по полю region_code.

Используя высокую корреляционную связь межу inn и region_code, восстановлены пропуски в region_code. Для заполнения пропусков выбран метод SimpleImputer библиотеки sklearn.

Задана целевая переменная target_group в зависимости от кодов okpd_code, которые она включает.

Проведена предобработка текстового признака category. Создана функция prepare_text. При её помощи подготовлен текст: убраны все символы кроме кириллицы, удалены стоп-слова и выполнит стемминг методом SnowballStemmer.

3.2 Разработка и обучение модели

Решение задачи мультиклассовой классификации для несбалансированной выборки выполнялось в модели CatBoost. Поиск оптимальных гиперпараметров проводился методом randomized_search. В качестве альтернативной модель рассматривалась логистическая регрессия, но sklearn не поддерживает GPU, поэтому она в работе не рассмотрена.

3.3 Тестирование модели

Тестирование модели совмещено с поиском оптимальных гиперпараметров и выполнялось в облачном сервисе Yandex DataSphere.

Выбраны начальные параметры, задана сетка поиска. Чтобы ускорить подбор гиперпараметров в randomized_search используется ограничитель n_iter, который был увеличен с 10 до 30. Датасет на входе в поиск дополнительно делился на тренировочный и валидационный, количество фолдов равно 5.

Результат работы randomized_search на каждой итерации приведен ниже.

```
0:      loss: 0.6382795  best: 0.6382795 (0)      total: 26.6s     remaining: 12m 51s
1:      loss: 0.5839721  best: 0.6382795 (0)      total: 29.9s     remaining: 6m 58s
2:      loss: 0.6384059  best: 0.6384059 (2)      total: 44.5s     remaining: 6m 40s
3:      loss: 0.6382795  best: 0.6384059 (2)      total: 59.5s     remaining: 6m 26s
4:      loss: 0.6384059  best: 0.6384059 (2)      total: 1m 14s    remaining: 6m 10s
5:      loss: 0.6382795  best: 0.6384059 (2)      total: 1m 29s    remaining: 5m 57s
6:      loss: 0.6384059  best: 0.6384059 (2)      total: 1m 43s    remaining: 5m 41s
7:      loss: 0.5933759  best: 0.6384059 (2)      total: 1m 47s    remaining: 4m 56s
8:      loss: 0.6405260  best: 0.6405260 (8)      total: 2m 4s     remaining: 4m 51s
9:      loss: 0.5934246  best: 0.6405260 (8)      total: 2m 8s     remaining: 4m 17s
10:     loss: 0.5934246  best: 0.6405260 (8)      total: 2m 12s    remaining: 3m 48s
11:     loss: 0.6405260  best: 0.6405260 (8)      total: 2m 29s    remaining: 3m 44s
12:     loss: 0.6406156  best: 0.6406156 (12)     total: 2m 47s    remaining: 3m 38s
13:     loss: 0.5935790  best: 0.6406156 (12)     total: 2m 50s    remaining: 3m 14s
14:     loss: 0.5935790  best: 0.6406156 (12)     total: 2m 53s    remaining: 2m 53s
15:     loss: 0.5938027  best: 0.6406156 (12)     total: 2m 57s    remaining: 2m 34s
16:     loss: 0.6437347  best: 0.6437347 (16)     total: 3m 17s    remaining: 2m 31s
17:     loss: 0.5938027  best: 0.6437347 (16)     total: 3m 20s    remaining: 2m 13s
18:     loss: 0.6437347  best: 0.6437347 (16)     total: 3m 41s    remaining: 2m 8s
19:     loss: 0.5933044  best: 0.6437347 (16)     total: 3m 44s    remaining: 1m 52s
20:     loss: 0.5933044  best: 0.6437347 (16)     total: 3m 48s    remaining: 1m 38s
21:     loss: 0.6459471  best: 0.6459471 (21)     total: 4m 15s    remaining: 1m 33s
22:     loss: 0.5933044  best: 0.6459471 (21)     total: 4m 19s    remaining: 1m 19s
23:     loss: 0.5933044  best: 0.6459471 (21)     total: 4m 23s    remaining: 1m 5s
24:     loss: 0.6459471  best: 0.6459471 (21)     total: 4m 50s    remaining: 58.1s
25:     loss: 0.6459736  best: 0.6459736 (25)     total: 5m 17s    remaining: 48.8s
26:     loss: 0.6459471  best: 0.6459736 (25)     total: 5m 44s    remaining: 38.3s
27:     loss: 0.5933044  best: 0.6459736 (25)     total: 5m 48s    remaining: 24.9s
28:     loss: 0.6475247  best: 0.6475247 (28)     total: 6m 25s    remaining: 13.3s
29:     loss: 0.5933044  best: 0.6475247 (28)     total: 6m 29s    remaining: 0us
Estimating final quality...
```

Рисунок 5 – Вывод работы поискового алгоритма метода randomized_search

Выбраны оптимальные гиперпараметры, с ними модель показала себя очень хорошо на тестовой выборке. Средневзвешенное значение Weighted F1 составило 0,96.

Таблица 2 – Результат работы модели на тестовой выборке

	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

Другие модели не рассматривались, реальной альтернативой является применение нейросетей, но это задача трудоёмка и в ВКР не представлена.

3.4 Разработка приложения

В соответствии с ТЗ разработано приложение на Python с графическим интерфейсом, которое выдает прогноз на основе входных данных, поступающих от пользователя. Пользователь должен зайти на сайт классификатора и указать данные о закупе, после чего отправить запрос. Данные проходят проверку на этапе отправки, если проблем нет, то они поступают на сервер. Модель получает входные признаки и выдает прогноз.

Краткая инструкция использования.

1. Перейти на сайт классификатора.
2. Заполнить требуемые данные.
    a. Ввести поля ИНН, стоимость, код региона и длительность.
    b. Выбрать из доступных вариантов поля статуса уровень бюджета.
    c. Внести текстовую информацию об объекте закупки.
3. Нажать кнопу «Проверить» и получить прогноз.

Ниже приведен пример работы приложения.

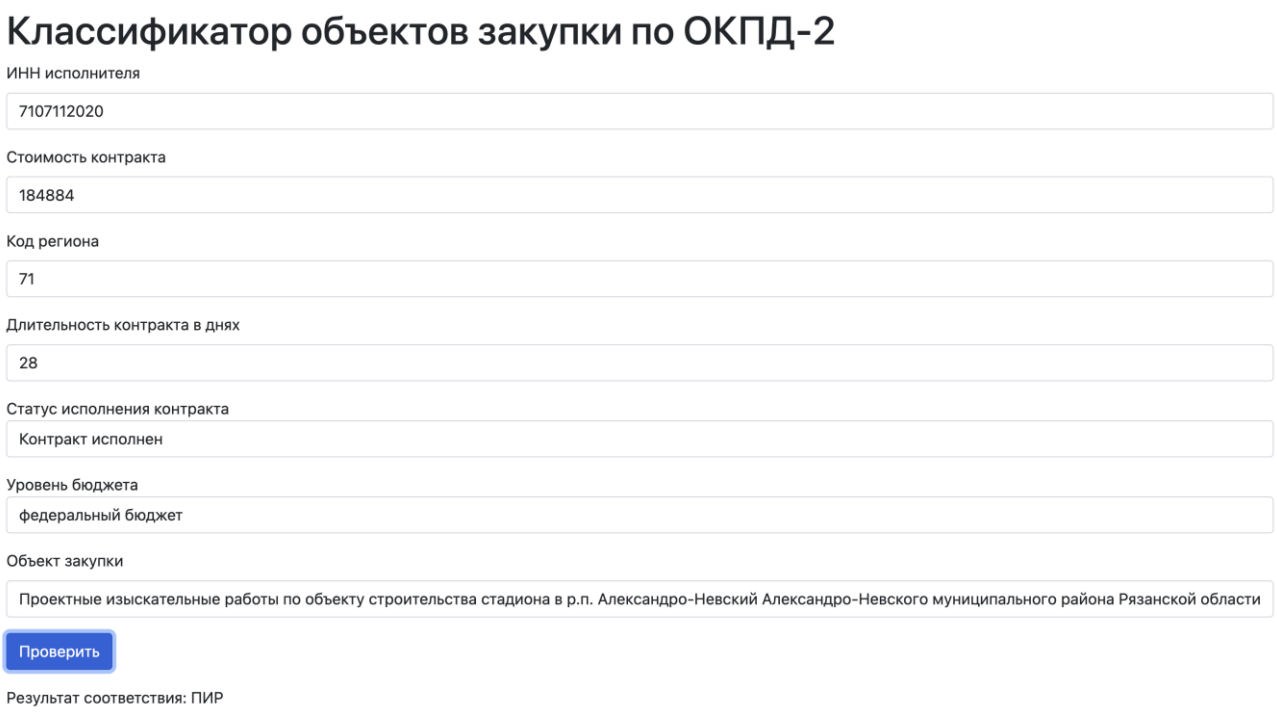

Рисунок 6 – Демонстрации работы приложения (https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=3620900301117000007)

3.5 Создание удаленного репозитория и загрузка результатов работы на него

Указывается страница слушателя, созданный репозиторий, коммиты в репозитории.

4 ЗАКЛЮЧЕНИЕ

В результате выполнения ВКР:

1. Изучены теоретические основы и методы решения NLP задач.
2. Разработаны скрипты для загрузки и обработки исходных данных.
3. Проведен разведочный анализ (EDA). Данные очищены от аномалий выбросов, заполнены пропуски и удалены дубликаты.
4. Спроектирована система хранения данных. Проработан концептуальный механизм обновления, забора и обработки данных.
5. Разработана классификатор объектов закупки по ОКПД-2.
6. Разработана приложение на Flax, которое выдает прогноз.
7. Создан профиль на github.com, на котором размещена практическую часть работы.

5 БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. Пособие. М.: Изд-во НИУ ВШЭ. 2017. 269 с. URL: https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf

2. Белова К.М., Судаков В.А. Исследование эффективности методов оценки релевантности текстов // Препринты ИПМ им. М.В.Келдыша. 2020. № 68. 16 с. http://doi.org/10.20948/prepr-2020-68 URL: http://library.keldysh.ru/preprint.asp?id=2020-68

3. Список стоп-слов для русского языка [Электронный ресурс]: https://countwordsfree.com/stopwords/russian (дата обращения: 20.10.2024)

4. Bengfort B., Bilbro R., Ojeda T. Applied Text Analysis with Python. США, Калифорния: O'Reilly Media. 2018. 332 с.

5. Machine Learning Algorithm Classification for Beginners [Электронный ресурс]: https://serokell.io/blog/machine-learning-algorithm-classification-overview.

6. Kalchbrenner N., Grefenstette E., & Blunsom P. A Convolutional Neural Network for Modelling Sentences, In Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL '2014, Baltimore, MD, USA, 2014, Vol. 1, pp. 655-665.

7. Malinowski M., Rohrbach M. & Fritz M. 2015. Ask Your Neurons: A Neural based Approach to Answering Questions about Images, IEEE Int. Conf. on Computer Vision, 2015, pp. 1-9.

8. Yu L., Hermann K.M., Blunsom P. & Pulman S. Deep Learning for Answer Sentence Selection, NIPS Deep Learning Workshop, 2014, 9.

9. Bird, S., Klein, E., & Loper, E. Обработка естественного языка с помощью

10. Python. O'Reilly Media, 2019.

11. Bird, S., & Loper, E. NLTK: Набор инструментов для обработки естественного языка. // Материалы ACL 2004 по интерактивным постерам и демонстрационным сессиям. – 2004. – С. 31–34.

12. Eisenstein, J. Введение в обработку естественного языка. Издательство Массачусетского технологического института, 2019.

13. Honnibal, M., & Montani, I. spaCy 2: Понимание естественного языка с помощью вложений Блума, сверточных нейронных сетей и инкрементального синтаксического анализа.

14. Manning, C. D., Raghavan, P., & Schütze, H. Введение в информационный поиск. Издательство Кембриджского университета, 2008.

15. Řehůřek, R., & Sojka, P. Программный каркас для тематического моделирования с большими корпусами // Материалы МКР 2010 Workshop on New

16. Challenges for NLP Frameworks. – 2010. – С. 45–50,

17. Репозиторий тестовых программ [Электронный ресурс]. URL: https://github.com/Daniliuk/Test_nlp (дата обращения: 18.10.2024)

18. SentiStrength [Электронный ресурс]: SentiStrength - sentiment strength detection in short texts. - Режим доступа: http://sentistrength.wlv.ac.uk/#About (дата обращения: 21.09.2024).

19. SentiWordNet - lexical resource for opinion mining. [Электронный ресурс]. - Режим доступа: https://github.com/aesuli/sentiwordnet (дата обращения: 25.10.2024).

20. simpletransformers [Электронный ресурс]. - Режим доступа: https://github.com/ThilinaRajapakse/simpletransformers (дата обращения: 21.10.2024).

21. Transformer - новая архитектура нейросетей для работы с последовательностями [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/341240/(дата обращения: 25.10.2024).

22. What is a Transformer? [Электронный ресурс]. - Режим доступа: https://medium.com/inside-machine-learning/what-is-a-transformer-d07ddfbec04 (дата обращения: 21.09.2024).

23. Алексеев А. А., Лазарева И. М. Морфологический анализ учебных текстов //Актуальные направления научных исследований XXI века: теория и практика. - 2015. - Т. 3. - №. 7-3. - С. 289-292.

24. Анализ точности однофакторного уравнения регрессии [Электронный ресурс]. - Режим доступа: https://studme.org/140829/matematika_himiya_fizik/analiz_tochnosti_odnofaktornogo_uravneniya_regressii (дата обращения: 21.10.2024).

25. Андреев, И. А., Армер, А. И., Крашенинникова, Н. А., Мошкин, В. С. Подход к решению задачи членения слитной речи на речевые единицы //Информационные технологии и нанотехнологии (ИТНТ-2017). - 2017. - С. 473-476.

26. Андреев, И. А., Башаев, В. А., Клейн, В. В., Мошкин, В. С. Определение вероятности терминологичности словоупотреблений в текстах конкретной предметной области //Интегрированные модели и мягкие вычисления в искусственном интеллекте. - 2015. - С. 764-773.

27. Гречачин В. А. К вопросу о токенизации текста //Международный научно-исследовательский журнал. - 2016. - №. 6 (48) Часть 4. - С. 25-27.

28. Алгоритм Word2Vec [Электронный ресурс]. - Режим доступа: https://neurohive.io/ru/ (дата обращения: 30.10.2024)

29. Кафтанников И. Л., Парасич А. В. Проблемы формирования обучающей выборки в задачах машинного обучения //Вестник ЮжноУральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. - 2016. - Т. 16. - №. 3. - С. 15-24.

30. PyTez - NLPub [Электронный ресурс]. - Режим доступа: https://nlpub.mipt.ru/PyTez (дата обращения: 21.10.2024).

31. Pass form from front-end to flask backend for processing and return variable value back to front end [Электронный ресурс]. - Режим доступа:

https://www.reddit.com/r/flask/comments/1dq1j39/pass_form_from_frontend_to_flask_backend_for/?rdt=45369 (дата обращения: 05.11.2024).

32. Examples and usage guidelines for form control styles, layout options, and custom components for creating a wide variety of forms [Электронный ресурс]. - Режим доступа: https://getbootstrap.com/docs/4.0/components/forms/ (дата обращения: 05.11.2024).

33. Анализ текстовых данных с помощью NLTK и Python [Электронный ресурс]. - Режим доступа: https://habr.com/en/companies/otus/articles/774498/ (дата обращения: 05.11.2024).

ПРИЛОЖЕНИЕ А

Приложение 2 (https://drive.google.com/file/d/1sRHx27O3NgTivrrQHdBdTAqxdCYNmARW/view?usp=sharing)							
Реестровый номер контракта	Объект закупки	Цена контракта	Длительность контракта в днях	ОКПД 2	Ссылка на zakupki.gov.ru	Группа по ОКПД	Правильная группа
1524800801121000037	Поставка канализационной насосной станции для ОГИБДД МО МВД России "Городецкий"	76 500	15	42.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1524800801121000037	Строительно-монтажные работы (СМР)	Прочее
2222602079621000090	Поставка уличных спортивных тренажеров	91 482	77	42.9	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=2222602079621000090	Строительно-монтажные работы (СМР)	Прочее
2231129560021000003	Приобретение товаров для оснащения спортивных объектов на открытых площадках	41 980	23	42.9	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=2231129560021000003	Строительно-монтажные работы (СМР)	Прочее
3310202094021000024	Поставка ограждения для детской площадке	98 777	21	42.9	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=3310202094021000024	Строительно-монтажные работы (СМР)	Прочее
1543316134221000403	'Выполнение монтажных и пусконаладочных работ автоматической установки пожарной сигнализации (АУПС) и системы оповещения и управления эвакуацией людей при пожаре (СОУЭ) корпуса № 43 ФБУН ГНЦ ВБ "Вектор" Роспотребнадзора'	1 432 267	23	43.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1543316134221000403	Строительно-монтажные работы (СМР)	Прочее
1504208141016000050	Поставка стоек ограждений с вытяжной лентой для нужд ФГКУ "12 ЦНИИ" Минобороны России в 2016 году	71 040	74	42.1	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1504208141016000050	Строительно-монтажные работы (СМР)	Прочее
1540601118616000143	Поставка шлагбаума для нужд ГБОУ ВПО НГМУ Минздрава России для обеспечения жизнедеятельности учреждения в целях	79 650	11	42.9	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1540601118616000143	Строительно-монтажные работы (СМР)	Прочее

	выполнения государственного задания						
2771596681416000055	Оказание услуг технического надзора (контроля)	43 887	11	41.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=2771596681416000055	Строительно-монтажные работы (СМР)	Строительный надзор
3131810616016000029	'Выполнение проектных (изыскательских)работ по объекту: Строительство ВЛ-0,4 кВ по ул. Садовая. Адрес строи-тельства: Республика Мордовия, Старошайговский район, с. Старое Шайгово, ул. Садовая.'	22 274	67	42.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=3131810616016000029	Строительно-монтажные работы (СМР)	Проектно-изыскательские работы (ПИР)
3620900301117000007	'Проектные изыскательные работы по объекту строительства стадиона в р.п. Александро-Невский Александро-Невского муниципального района Рязанской области'	184 884	28	43.9	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=3620900301117000007	Строительно-монтажные работы (СМР)	Проектно-изыскательские работы (ПИР)
2616106092518000062	Технологическое подключение к сетям водоотведения быстровозводимого модульного здания пожарного депо	1 832 977	14	43.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=2616106092518000062	Строительно-монтажные работы (СМР)	Подключение коммуникаций
3504301461717000061	Технологическое присоединение объекта капитального строительства к сети газораспределения жилого дома №38а по ул. Селецкой городского округа Серпухов	10 517	258	42.2	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=3504301461717000061	Строительно-монтажные работы (СМР)	Подключение коммуникаций
1771454974421000086	Осуществление строительного контроля на объекте: "Реконструкция аэродрома Охотск, Хабаровский край"	79 854 000	1 113	71.1	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1771454974421000086	Проектно-изыскательские работы (ПИР)	Строительный надзор
1771454974421000082	Оказание услуг по осуществлению строительного контроля на объекте "Реконструкция ИВПП-2 аэропорта Якутск (III очередь строительства), Республика Саха (Якутия)"	25 000 000	720	71.1	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=1771454974421000082	Проектно-изыскательские работы (ПИР)	Строительный надзор

2650115377817000019	Выполнение работ по строительству "под ключ" жилых домов в г. Северо-Курильске по объекту: Строительство жилых домов в г. Северо-Курильске на острове Парамушир (в том числе проектные и изыскательские работы) и (или) приобретение квартир в новых жилых домах	116 677 933	806	71.1	https://zakupki.gov.ru/epz/contract/contractCard/common-info.html?reestrNumber=2650115377817000019	Проектно-изыскательские работы (ПИР)	Строительно-монтажные работы (СМР)