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**

**на тему: «Прогнозирование срока окупаемости внедрения автоматизированных систем на складах временного хранения»**

**по курсу**

«Data Science»

Слушатель Кривонос Анастасия Дмитриевна

Москва, 2023

# Содержание

[Содержание 2](#_Toc133356556)

[Введение 3](#_Toc133356557)

[1. Аналитическая часть 4](#_Toc133356558)

[1.1. Постановка задачи 4](#_Toc133356559)

[1.2 Описание используемых методов 7](#_Toc133356560)

[1.3 Разведочный анализ данных 9](#_Toc133356561)

[2 Практическая часть 15](#_Toc133356562)

[2.1 Предобработка данных 15](#_Toc133356563)

[2.2 Разработка и обучение модели 16](#_Toc133356564)

[2.3 Тестирование модели 16](#_Toc133356565)

[Размер тестовой выборки: (162, 13) 16](#_Toc133356566)

[2.4 Нейронная сеть 16](#_Toc133356567)

[2.5 Разработка приложения 18](#_Toc133356568)

[2.6. Создание удаленного репозитория 19](#_Toc133356569)

[Заключение 20](#_Toc133356570)

# Введение

Данная работа выполнена в рамках курса Data Science.

В качестве анализируемой задачи принята тема «Способы минимизации логистических издержек на складе временного хранения».

Система анализа данных включает в себя следующий алгорит анализа, реализованный в работе:

1. Осуществлен сбор данных: собраны практические данные СВХ, на основании создана база данных, сформирована электронная таблица.
2. Произведена очистка и подготовка данных: удалены отсутствующие или нерелевантные значения, обработаны выбросы и преобразованы переменные по мере необходимости.
3. Осуществлён исследовательский анализ данных: выполнены начальные исследования данных для выявления закономерностей, тенденции и взаимосвязи, использованы такие методы, как сводная статистика, визуализация и проверка гипотез.
4. Проведено прогнозное моделирование: создана прогнозная модель для прогнозирования будущих результатов на основе прошлых данных, используя такие методы, как регрессионный анализ, алгоритмы машинного обучения.
5. Осуществлена визуализация данных: созданы визуальные представления данных, такие как диаграммы, графики, полученные путем обработки имеющихся данных.
6. Визуализированы результаты анализа, подготовлена презентация.

# Аналитическая часть

## Постановка задачи

Современный глобальный мир повсеместно нацелен на поступательное развитие, расширение ассортимента товаров, что связано в основном с внедрением передовых технологий в деятельность всех структурных элементов экономики, в том числе развитием современных складских технологий. Указанное требует от складов выполнения большего количества операций путем их автоматизации и цифровизации. Прогрессивные цифровые технологии складской логистики позволяют реализовать поставленные задачи: точный учет поставок товаров; таможенный контроль; отслеживание поставок; цифровизация услуг для клиентов и их высокое качество; расширение диапазона услуг предпродажной подготовки товаров к реализации до их выпуска.

Современная логистика требует постоянного совершенствования и оптимизации процессов, чтобы обеспечить максимальную эффективность и минимальные затраты. Одним из ключевых направлений развития является использование новых технологий, таких как системы автоматического взвешивания и контроля грузов. Эти инновационные решения позволяют значительно ускорить и упростить процессы складирования и транспортировки товаров, а также повысить точность и надежность контроля за грузами. В данной работе предлагается рассмотреть преимущества внедрения таких систем и возможности их использования для оптимизации затрат логистических предприятий, в том числе путем проведения расчетов и анализа данных.

***Сбор данных***

Системы автоматического взвешивания и контроля грузов являются важными инструментами для оптимизации процессов складирования и транспортировки товаров. Системы автоматического взвешивания позволяют быстро и точно определить вес груза, что особенно важно при работе с большими объемами товаров и безопасностью поставок.

Актуальность данного анализа также связана с рассмотрением в настоящий момент проекта изменений №296887-8 в Федеральный закон от 03.08.2018 № 289-ФЗ  «О таможенном регулировании в Российской Федерации», вынесенного на рассмотрение депутатом В.М. Резником – членом Комитета Государственной Думы по бюджету и налогам (далее – Комитет Госдумы), срок предоставления поправок - 05.05.2023, в том числе устанавливает новые требования к обустройству территорий, прилегающих к складам временного хранения (далее – СВХ) и самих СВХ в отношении, в том числе:

* оснащения системой видеонаблюдения, обеспечивающей возможность визуального контроля всей территории склада, дистанционного вращения в вертикальной и горизонтальной плоскостях, фокусирования, приближения и удаления участков и объектов видеонаблюдения;
* оснащения стационарной системой радиационного контроля с детекторами гамма- и нейтронного излучения, оснащенной комплектом видеонаблюдения с выводом информации на автоматизированное рабочее место;
* наличие автоматизированной системы учета товаров, совместимой с программными продуктами государственных органов;
  + обеспечение информационного взаимодействия между владельцем склада временного хранения и таможенным органом через личный кабинет, в том числе обеспечение передачи таможенному органу в электронном виде через личный кабинет сведений, содержащихся в отчетности о товарах, находящихся на складе временного хранения, и получения в электронном виде от таможенного органа сведений о выпуске товаров, находящихся на складе временного хранения;
  + оснащения системой считывания регистрационного номера автомобиля в местах въезда (выезда) на (с) территорию склада временного хранения с предоставлением доступа к указанной системе таможенному органу и возможностью передачи информации в информационную систему таможенных органов;
  + оснащения инспекционно-досмотровым комплексом.

Примерами систем автоматического взвешивания российской разработки можно привести системы производства компании "МЕТРА"; автоматизированную систему управления АСУ «Весовой поток» - это модули системы интеллектуального управления логистикой Ярд 2.0 – современного программно-аппаратного комплекса, разработанные компаниями «Цифровые контрольные технологии»; системы динамического взвешивания WIM VanJee  производства компании Sensotec и другие - данные системы на датчиках, которые устанавливаются на дороге (дорожном покрытии) и позволяют автоматически взвешивать грузовые автомобили, распознавать государственные номера с обеспечением фотофиксации и видеофиксации взвешивания, управления шлагбаумами и светофорами.

В общем понимании программно-аппаратные комплексы для автоматизации взвешивания и распознавания данных предназначены для уменьшения влияния человеческого фактора, упорядочивания и ускорения процесса взвешивания автомобилей, контроля за работой сотрудников, предотвращения мошенничества при взвешивании, проведения автоматического взвешивания без участия оператора.

В базовую систему входят следующие модули:

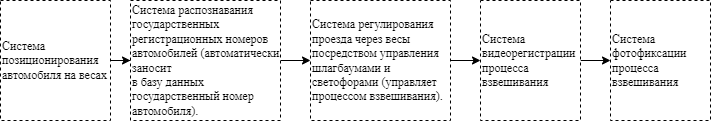


Рис. 1. Модули системы программно-аппаратного комплекса для автоматизации взвешивания и распознавания данных

Также схематично представим возможный алгоритм прохождения цепочки поставки до момента поступления на СВХ с применением автоматизированных методик.

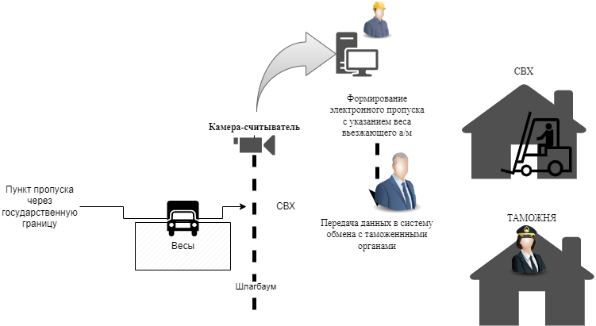


Рис.2. Алгоритм прохождения цепочки поставки до момента поступления на СВХ с применением автоматизированных методик

Для того чтобы произвести необходимые расчеты необходимо выделить ключевые параметры, влияющие на формирование издержек склада, определить какие расходы повлечет за собой необходимость внедрения автоматизированных законодательных инициатив, в течение какого срока данные расходы будут компенсированы и за счет каких производимых изменений производственной цепочки.

В качестве входных данных приняты следующие данные:

1. Номер СВХ;
2. Наименование владельца СВХ;
3. Местонахождение владельца СВХ;
4. ИНН СВХ;
5. Площадь СВХ;
6. Региональный таможенный орган;
7. Текущие издержки на КПП1;
8. Текущие издержки на КПП2;
9. Текущие издержки на ДИСПЕТЧЕРСКАЯ;
10. Количество сотрудников КПП 1;
11. Количество сотрудников КПП 2;
12. Количество сотрудников Диспетчерской;
13. ФОТ КПП1\мес.;
14. ФОТ КПП1\мес.;
15. ФОТ ДИСПЕТЧЕРСКАЯ\мес.;
16. Стоимость заезда товарной партии на СВХ;
17. Стоимость закупки и внедрения АСУ "ВЕСЫ 1";
18. Стоимость интегрированного ПО "ВЕСЫ 1" (мин пакет 10 рабочих мест);
19. Стоимость ежемесячного обслуживания "ВЕСЫ 1" (мин пакет 10 рабочих мест);
20. Срок окупаемости внедрения ПО/мес.

Общее количество параметров для анализа – 20.

Данные объединены исходя из приведенных в датасете числовых полей.

С помощью команды data.shape получена размерность датасета: результат кортеж (541, 26), что означает, что DataFrame содержит 541 строк и 26 колонок

Итоговый датасет нормализованный, пропуски отсутствуют. Элементы массива соответствуют типу float64.

На выходе необходимо предложить модель минимизации издержек СВХ при законодательной необходимости автоматизации производства, а также сделать расчет срока окупаемости внедрения АСУ. Кейс основан на реальных данных деятельности СВХ России по типу транспорта - автодорожный, а также на опубликованных государственных инициативах на сайте Госдумы.

**Актуальность:** Созданная прогнозная модель поможет рассчитать эффективность автоматизации производства, «замораживание» денежных средств, определить срок возврата привлеченных средств.

## 1.2 Описание используемых методов

***В ходе анализа*** произведена очистка и подготовка данных: удалены отсутствующие или нерелевантные значения, обработаны выбросы и преобразованы переменные по мере необходимости.

Для достижения оптимальных результатов необходимо использовать экономико-математическое моделирование, а также анализ данных, которые позволяют учитывать множество факторов и прогнозировать изменения. На любой экономический показатель чаще всего оказывает влияние не один, а несколько факторов. Математическая модель оптимизации может быть разработана на основе метода поиска оптимального решения, например, метода линейного программирования, метода решения задач регрессии для прогнозирования параметров для определения оптимального внедрения автоматизированных методик с минимальными затратами. Для решения задачи регрессии использовались: линейная регрессия, случайный лес, метод k ближайших соседей.

Для решения поставленной задачи выбраны методы:

* решение задачи регрессии для прогнозирования параметров: Срок окупаемости внедрения ПО/мес.

На рисунке 3 представлены сценарии работы основных элементов системы СВХ (на текущий момент и после автоматизации производства).

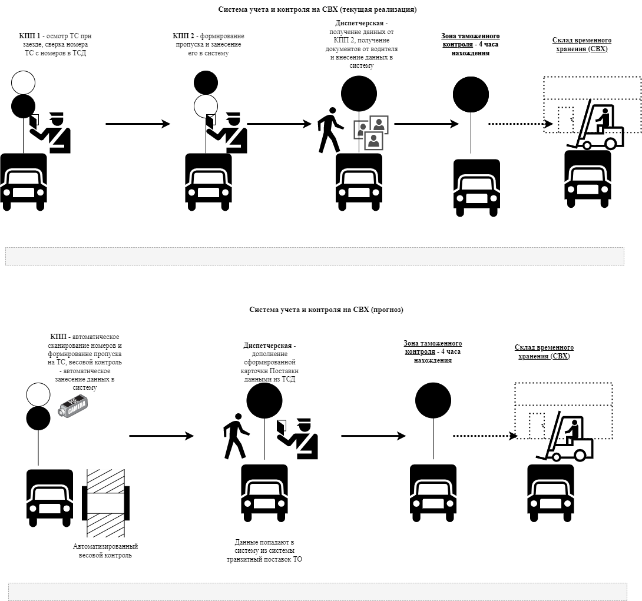
****

Рис.3. Сценарии работы основных элементов системы СВХ

Оценка качества моделей указана на рисунке 4.

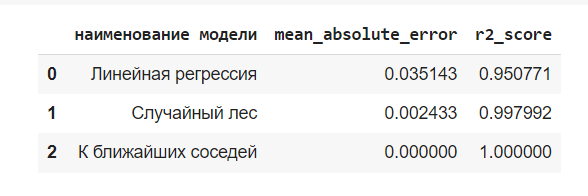


Рис. 4. Оценка качества моделей

Средняя абсолютная ошибка (mean\_absolute\_error) для всех моделей разная и находится в пределах 0,04-0.

Коэффициент детерминации (r2\_score) принимает положительные значения для всех моделей.

## 1.3 Разведочный анализ данных

Для разведочного анализа данных использованы методы описательной статистики.

Датасет был проверен на наличие пропусков в значениях (команда df.isna().sum()). Установлено, что данные не полные. Для корректной работы модели при помощи команды data.columns[-4:] удалены 4 последних столбца и строки с пропущенными значениями.

Датасет проверен на наличие дубликатов (команда data.drop\_duplicates().shape). Дубликатов не обнаружено.

Команда Describe () позволила выявить наличие множества дискретных величины.

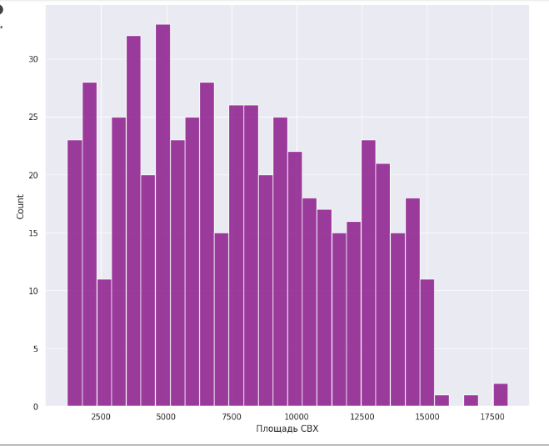


Рис.5. – Гистограмма «Соотношение Площадь СВХ»



Рис.6. Гистограмма «Издержки на КПП 1»

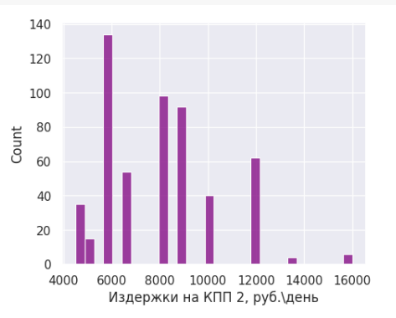


Рис. 7. – Гистограмма «Издержки на КПП 2»

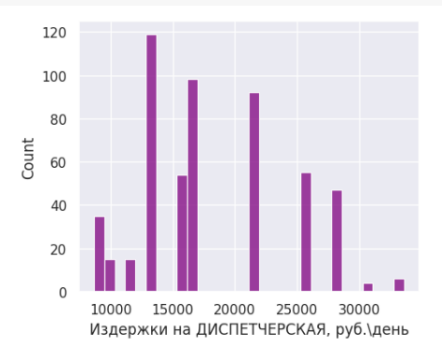


Рис. 8. Гистограмма «Издержки на ДИСПЕТЧЕРСКАЯ»

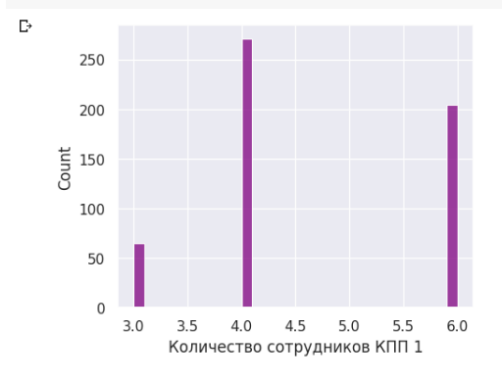


Рис. 9. Гистограмма «Количество сотрудников КПП 1»



Рис. 10. Гистограмма «Количество сотрудников КПП 2»



Рис. 11. Гистограмма «Количество сотрудников Диспетчерской»

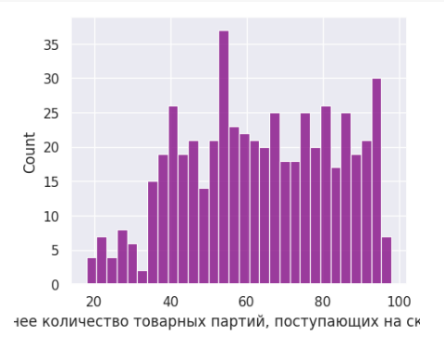


Рис. 12. Гистограмма «Среднее количество товарных партий, поступающих на СВХ»

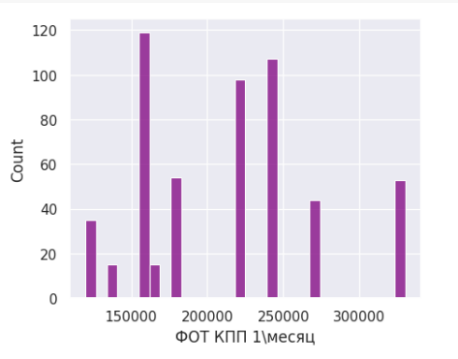


Рис. 13. Гистограмма «ФОТ КПП 1\месяц»

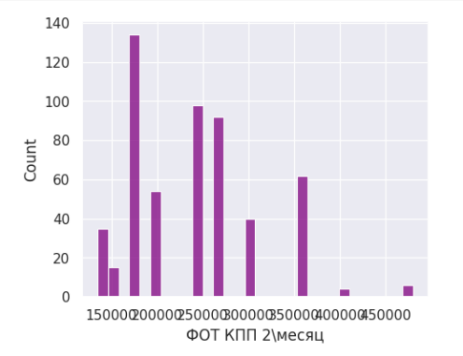


Рис. 14. Гистограмма «ФОТ КПП 2\месяц»

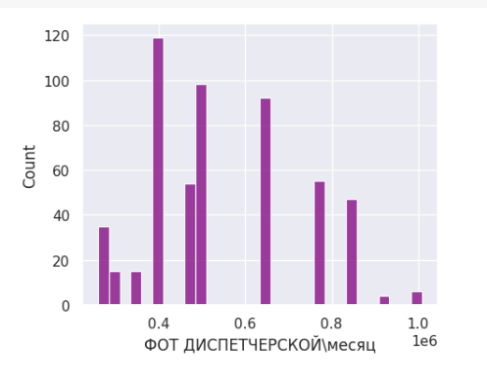


Рис. 15. Гистограмма «ФОТ ДИСПЕТЧЕРСКОЙ\месяц»

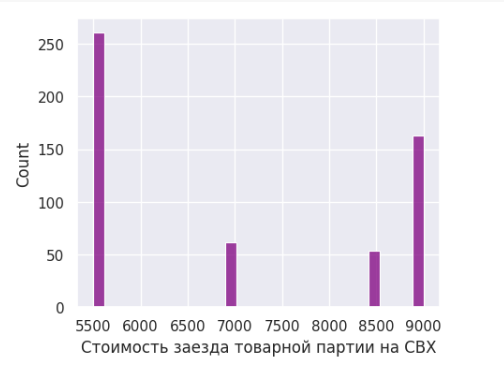


Рис. 16. Гистограмма «Стоимость заезда товарной партии на СВХ»

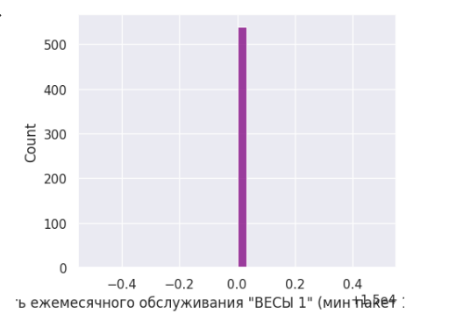


Рис. 17. Гистограмма «Стоимость ежемесячного обслуживания «ВЕСЫ 1»»

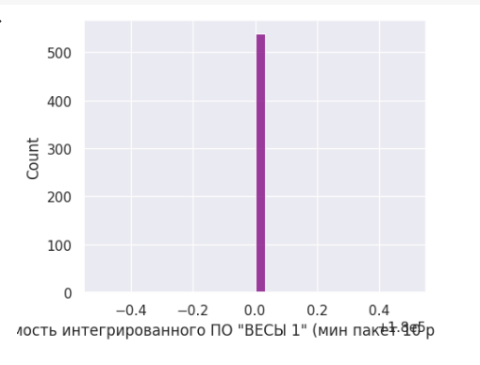


Рис. 18. Гистограмма «Стоимость интегрированного ПО «ВЕСЫ 1»»



Рис. 19. Гистограмма «Стоимость закупки и внедрения АСУ «ВЕСЫ 1»»



Рис. 20. Гистограмма «Срок окупаемости внедрения ПО\мес.»»

Все признаки далеки от нормального распределения, соответственно выбросов не наблюдается. Построение диаграмм «ящик с усами» также подтвердило практически 100% отсутствие выбросов, лишь по одному параметру «Издержки на КПП 2» наблюдается ввыбросы в зоне малых значений, однако в связи с тем, что данный параметр подлежит замене внедрением АСУ дальнейший анализ его не целесообразен.

Примеры для приведены на рисунках 21 и 22.

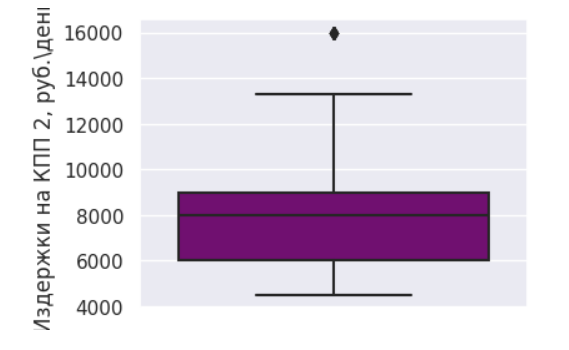


Рис. 21. Диаграмма «Издержки на КПП 2»

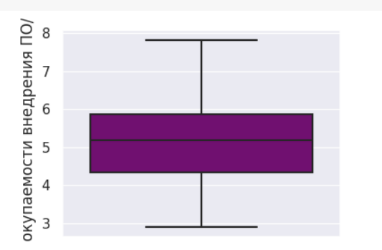


Рис. 22. Диаграмма «Срок окупаемости внедрения ПО»

Видно, что параметры «Стоимость интегрированного ПО "ВЕСЫ 1" (мин пакет 10 рабочих мест)» и «Стоимость ежемесячного обслуживания "ВЕСЫ 1" (мин пакет 10 рабочих мест)» - константные данные, поэтому их нужно удалить с помощью команды num\_col.remove из списка численных признаков.

На рисунке 23 приведена тепловая карта коэффициентов корреляции, значения которой показывают, что все полученные коэффициенты корреляции находятся в промежутке значений от - 0,86 до 0,99. Исходя из этого можно сделать вывод, что наблюдается сильная корреляция, переменные являются зависимыми.

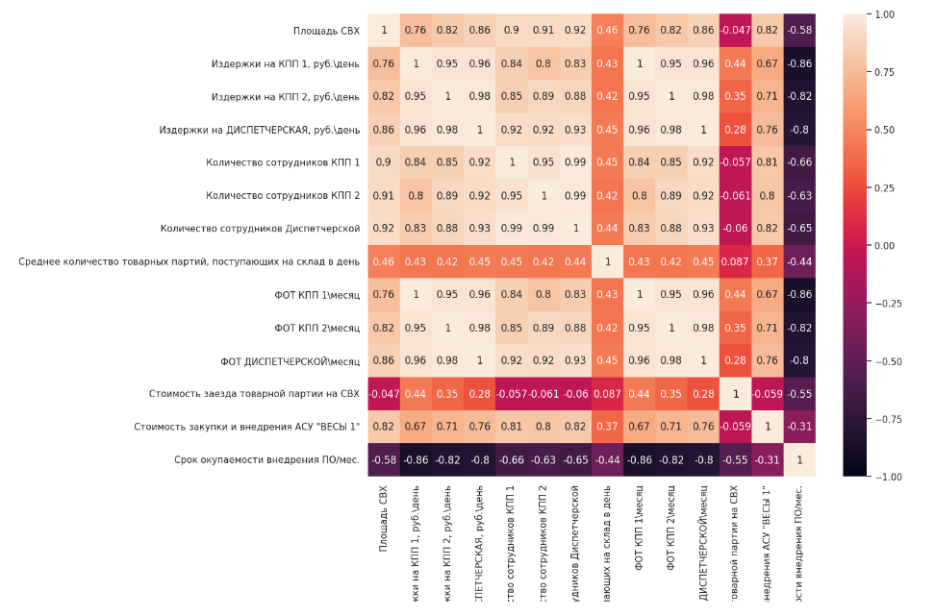


Рис. 23. Тепловая карта коэффициентов корреляции

# 2. Практическая часть

## 2.1 Предобработка данных

Предобработка данных осуществлялась на основании разведочного анализа данных, который показал отсутствие выбросов. При помощи команды data.info() определены признаки, их типы, количество данных (рис. 22).

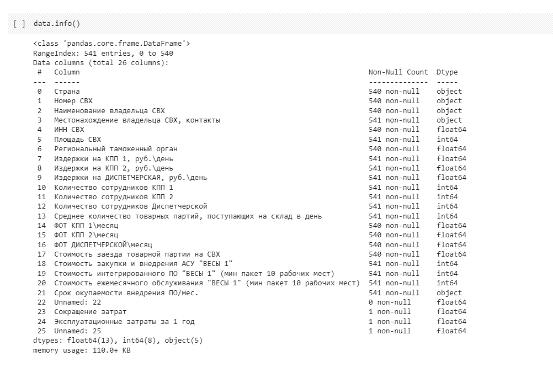


Рис. 24. Признаки, типы, количество данных датасета

Затем данные проверены на уникальность и с помощью команды data.isna().sum() полноту. Были удалены битые данные, не несущие информации - 4 последних стобца, а также удалены строки с пропущенными значениями командой data.dropna().

Данные проверены на дубликаты с помощью команды data.drop\_duplicates().shape, установлено, что дубликатов нет.   
При проведении обработки признаков установлено, что 21-й признак является объектом хотя должен быть числом, исправили это, также вычислены битые строки с помощью команды data['Срок окупаемости внедрения ПО/мес.'].unique() и удалены.

Затем по датасету определены основные характеристики числовых признаков.

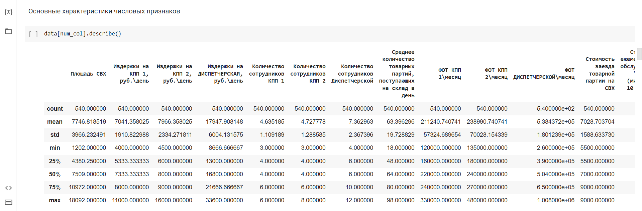


Рис. 25. Основные характеристики числовых признаков

Проведение дальнейшего анализ влияния факторов весов не целесообразно, так как лучшее качество у модели, которая не определяет веса для параметров.

Для дальнейшей разработки и обучения модели была выполнена нормализация данных с помощью MinMaxScaler.

## 2.2 Разработка и обучение модели

В качестве модели выбрана линейная регрессия, случайный лес, К ближайших соседей. Обучение и тестирование производилось для прогнозируемого параметра: Срок окупаемости внедрения ПО/мес.

При построении модели «Случайный лес» был осуществлен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой.

При построении модели «К ближайших соседей» определены значения целевой переменной для прогнозного объекта, используя ближайшие объекты из обучающей выборки. Чтобы сделать прогноз для нового входного сигнала, алгоритм вычисляет расстояние (обычно евклидово расстояние) между новым входным сигналом и всеми обучающими экземплярами и выбирает K ближайших соседей. Затем прогноз делается на основе целевого значения большинства из K соседей. Значение K - это гиперпараметр, который может быть настроен в зависимости от конкретной задачи и набора данных. Меньшее значение K приведет к более сложной границе принятия решения и может быть более чувствительным к шуму в данных, в то время как большее значение K приведет к более плавной границе принятия решения, но может быть менее чувствительным к локальным изменениям в данных.

На рисунке 4 приведен анализ качества разработанных моделей. При построении модели «К ближайших соседей» является наиболее точным.

## 2.3 Тестирование модели

Из датасета была выделена прогнозируемая (зависимая) переменная. Далее, выборка была разделена на обучающую и тестовую выборки, в соответствии с условием задачи 70% (на обучение) /30% (на тестирование).

Размер обучающей выборки: (378, 13)

## Размер тестовой выборки: (162, 13)

## 2.4 Нейронная сеть

В качестве нейронной сети был принят многослойный персептрон.

Гиперпараметры модели:

- количество скрытых слоев = 3;

- количество нейронов на слое = 256 и 64;

- активационная функция «relu»;

- количество нейронов на выходном слое = 1;

- оптимизатор «Adam»;

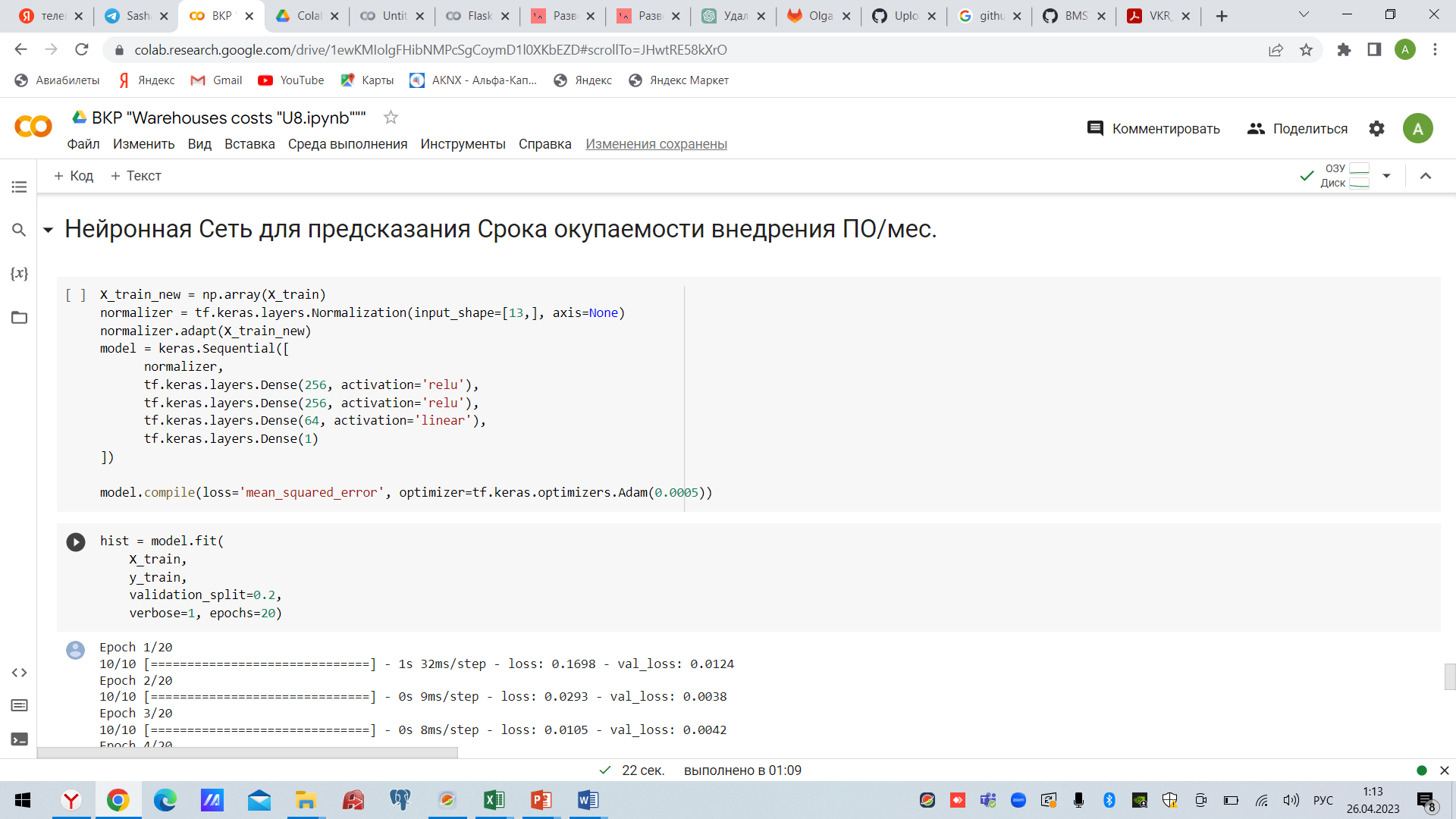


Рис. 26. Основные характеристики модели

Обучение модели происходило за 20 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE и общих результатов обучения модели). Результат изменения MSE модели указан на рисунке 31. MSE уменьшается со временем по мере выполнения алгоритма. Это означает, что модель приближаемся к оптимальному решению.

Разработка нейронной сети, способной точно предсказывать интересующий параметр при небольшом размере сети, является значительным достижением. Это указывает на то, что корреляция между входными данными и представляющим интерес параметром относительно проста, и для учета этой взаимосвязи не требуется сложной модели. За счет создания модели с небольшим количеством нейронов снижаются вычислительные затраты и требования к памяти, что делает ее более практичной для развертывания в средах с ограниченными ресурсами. В целом, это представляет собой эффективный подход к решению рассматриваемой проблемы.

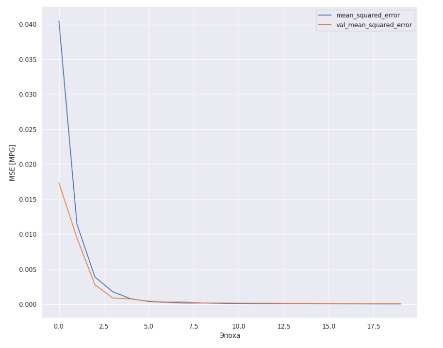


Рис. 27. Изменение MSE за время обучения модели

Ошибка в основном распределяется между -0.022 и 0.05, где 0 является наибольшим. Это показывает, что модель обучения подходящая. Гистограмма распределения ошибки приведена на рисунке 32.

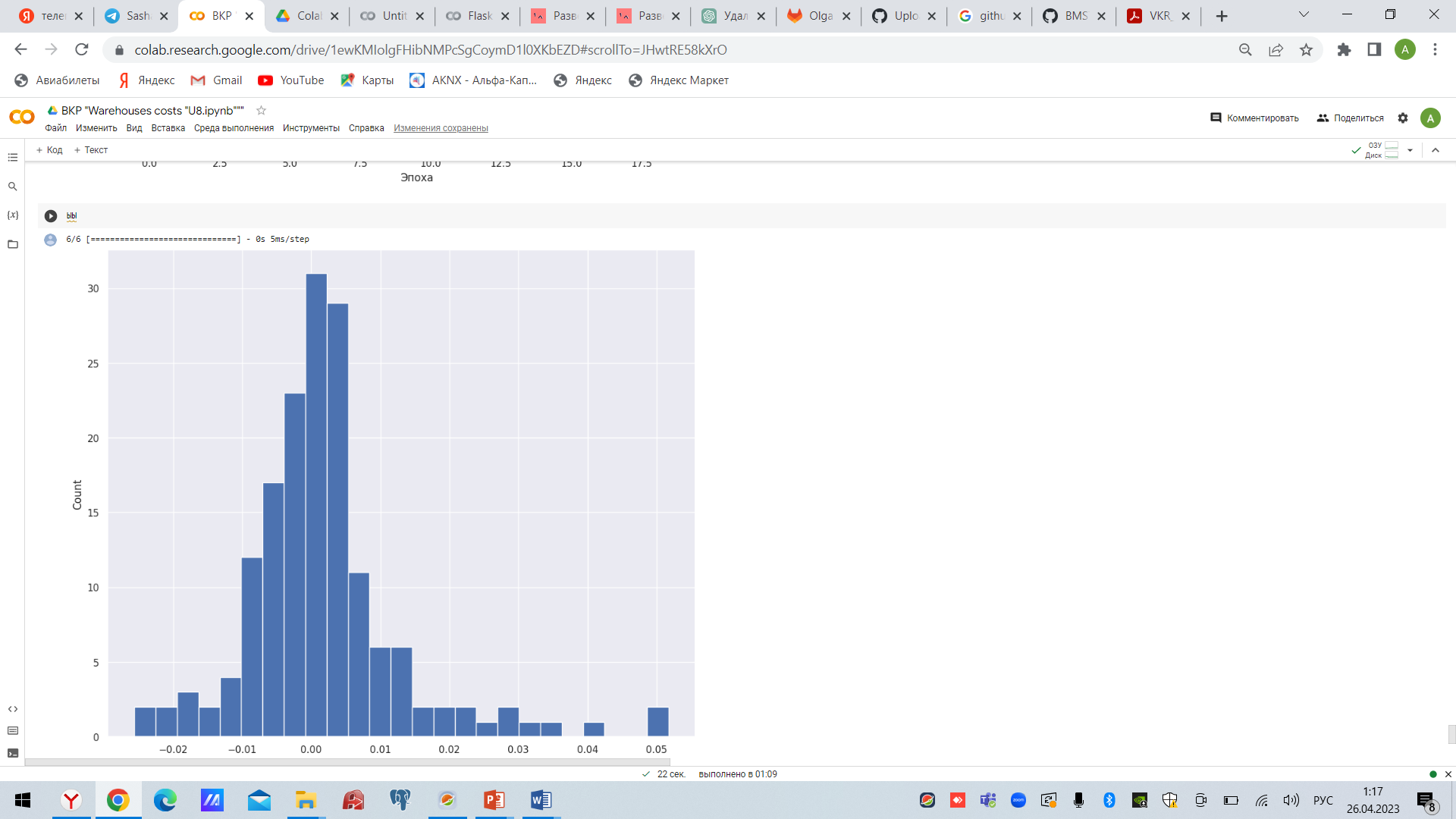


Рис. 28. Распределение ошибки

## 2.5 Разработка приложения

Приложение разработано в PyCharm для запуска из командной строки - расчет срока окупаемости внедрения ПО в зависимости от заданных условий.

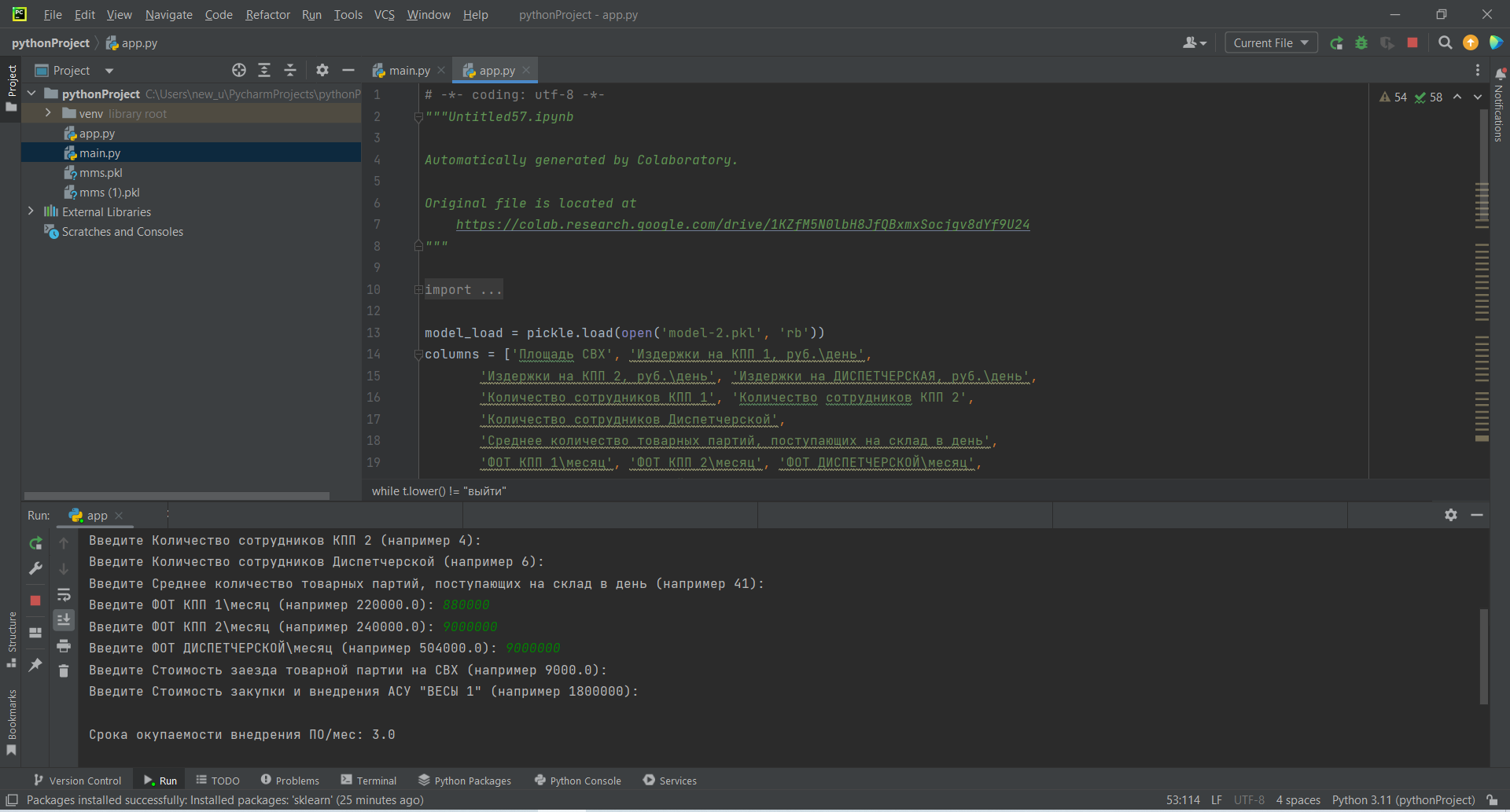
**

Рис. 29. Макет приложения

## 2.6. Создание удаленного репозитория

Страница слушателя на GitHub

Созданный репозиторий:

<https://github.com/anastasiiak89/Warehouses-costs-VESI>

Страница слушателя, созданный репозиторий, коммиты в репозитории.

<https://github.com/anastasiiak89/Warehouses-costs-VESI>

# Заключение

В ходе выполнения ВКР были изучены способы анализа и предобработки данных. Построенная модель показала, что исходный датасет содержит реальные значений для отработки обучения и тренировки моделей.

Полученная модель нейронной сети оптимальна.

Введение требований по наличию современных систем контроля и отслеживания за перемещением товаров, транспортных средств, лиц по территории СВХ, в том числе систем таможенного контроля, позволит повысить уровень автоматизации отдельных операций, приведет к возможности опосредованного участия должностных лиц в контроле за размещением товаров на территории склада временного хранения. Внедрение систем автоматического взвешивания и контроля грузов позволяет сократить затраты на транспортировку и складирование товаров, а также повысить уровень удовлетворенности клиентов.