ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «УНИВЕРСИТЕТ «ДУБНА»

Институт системного анализа и управления

Кафедра системного анализа и управления

Пояснительная записка:

Прогнозирование цен на подержанные автомобили с использованием машинного обучения

Группа: 4015

Команда студентов «Мы собрались»:

Короленко Марина Владимировна

Грошев Никита Викторович

Батаев Илья Сергеевич

Дубна, 2024

**Содержание**

[Введение 3](#_Toc184651738)

[Состав команды 4](#_Toc184651739)

[Постановка задачи 5](#_Toc184651740)

[Актуальность 5](#_Toc184651741)

[Цель 5](#_Toc184651742)

[Задачи 5](#_Toc184651743)

[Ожидаемый результат 5](#_Toc184651744)

[План работ 6](#_Toc184651745)

[*Mind map* проекта 7](#_Toc184651746)

[Средства реализации 8](#_Toc184651747)

[Описание модели 9](#_Toc184651748)

[Модель прогнозирования 9](#_Toc184651749)

[Данные для обучения модели 10](#_Toc184651750)

[Анализ данных 10](#_Toc184651751)

[Обучение модели 12](#_Toc184651752)

[Модель прогнозирования цен на поддержанные автомобили 12](#_Toc184651753)

[Описание ПО 13](#_Toc184651754)

[Архитектура проекта 13](#_Toc184651755)

[Диаграмма взаимодействия пользователя с системой 14](#_Toc184651756)

[Функциональные возможности 15](#_Toc184651757)

[Результаты 17](#_Toc184651758)

[Заключение 20](#_Toc184651759)

[Ссылки на источники 21](#_Toc184651760)

# Введение

С каждым годом рынок подержанных автомобилей становится всё более насыщенным, что осложняет задачу точной оценки стоимости транспортных средств. Ошибочная оценка цены может привести к финансовым потерям для автодилеров и снижению доверия со стороны покупателей. В условиях высокой конкуренции необходимо использовать передовые технологии, чтобы сделать процесс оценки максимально точным и удобным.

Машинное обучение предоставляет возможность автоматизировать процесс оценки цен на автомобили, анализируя их характеристики и определяя рыночную стоимость с минимальной погрешностью. Использование веб-приложения с интегрированной моделью машинного обучения позволяет не только автоматизировать прогнозирование цен, но и улучшить бизнес-процессы, связанные с продажей подержанных автомобилей.

# Состав команды

1. Короленко Марина − менеджер проекта/архитектор.
2. Грошев Никита – разработчик.
3. Батаев Илья – *ML*–инженер, Дизайнер.

# Постановка задачи

## Актуальность

С увеличением количества подержанных автомобилей на рынке становится все сложнее точно оценивать их стоимость. Ошибочная оценка может привести к финансовым потерям для дилеров, а также к снижению доверия со стороны покупателей. Использование машинного обучения для прогнозирования цен на подержанные автомобили может значительно улучшить точность оценок и помочь дилерам принимать более обоснованные решения.

Проект может быть применён в таких направлениях как:

* Автодилерские салоны: автоматизация процесса оценки стоимости подержанных автомобилей перед продажей.
* Аналитические агентства: мониторинг и анализ рыночных цен на автомобили для отчётов и прогнозов.
* Онлайн-площадки по продаже автомобилей: добавление функции автоматической оценки цен для удобства пользователей.

## Цель

Спроектировать и разработать веб-приложение с использованием модели машинного обучения для прогнозирования цен на подержанные автомобили на основе различных атрибутов.

## Задачи

1. Собрать и обработать данные.
2. Разработать и обучить модель машинного обучения.
3. Разработать веб интерфейс для загрузки и обработки данных.
4. Интегрировать модель машинного обучения с веб интерфейсом.

## Ожидаемый результат

Веб – приложение для автоматизации прогнозирования цен на поддержанные автомобили с использованием модели машинного обучения *XGBoost.*

# План работ

1. Анализ поставленной задачи.
2. Формирование требований к сервису.
3. Определение критериев оценки результата.
4. Создание архитектуры проекта.
5. Поиск данных.
6. Обработка данных.
7. Разработка модели машинного обучения.
8. Проектирование пользовательского интерфейса.
9. Разработка веб – приложения.
10. Интеграция модели с веб–приложением.
11. Тестирование приложения.
12. Подготовка документации.

# *Mind map* проекта

На рисунке 1 представлена интеллект–картапроекта, которая визуализирует структуру и ключевые элементы проекта, такие как:

1. Стек технологий.
2. Заинтересованные лица.
3. Способы анализа данных.
4. Способы визуализации данных.
5. *ML* модель.

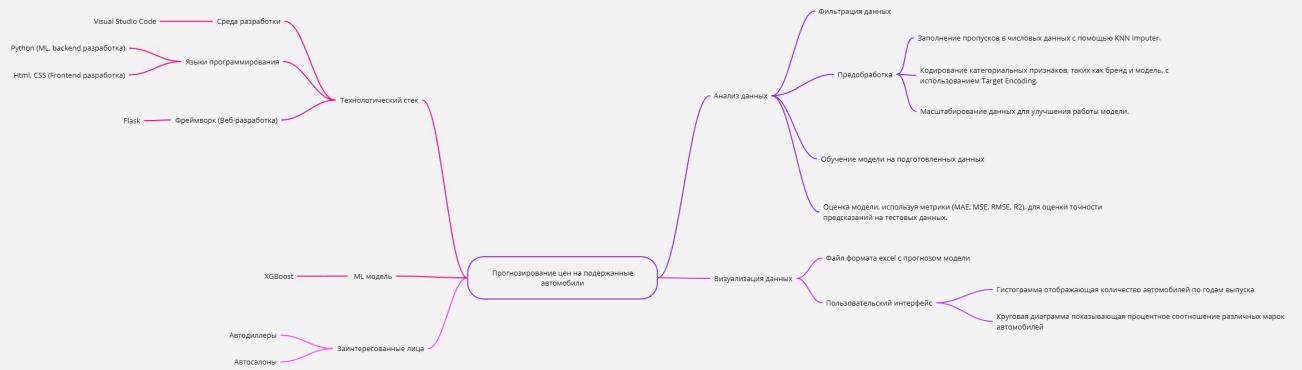


Рис. 1. Mind map проекта

# Средства реализации

В качестве средств реализации для данного проекта были выбраны следующие технологии:

1. Языки программирования: *Python, HTML, CSS*.
2. Фреймворк: *Flask*.
3. Инструменты разработки: *Visual Studio Code.*
4. Библиотеки: *sciki*–*learn, pandas, numpy, matplotlib*.
5. Система управления версиями: *GitHub*.
6. Модель машинного обучения: *XGBoost.*

# Описание модели

В рамках данного проекта использовалась модель машинного обучения для прогнозирования цен на поддержанные автомобили.

## Модель прогнозирования

Модель основана на алгоритме *XGBoost.* Это ансамблевый метод, который использует деревья решений и градиентный бустинг для повышения точности предсказаний. Основные характеристики:

1. Эффективность: Быстрое обучение и предсказание.
2. Высокая точность: Обеспечивает высокую точность предсказаний благодаря методу градиентного бустинга.
3. Регуляризация: Включает методы регуляризации для предотвращения переобучения модели.

Модель обучена на различных атрибутах автомобилей, включая: пробег, год производства, марку, модель, тип топлива, цвет машины, цвет салона, количество цилиндров, тип топлива.

На рисунке 2 представлена архитектура модели *XGBoost*.

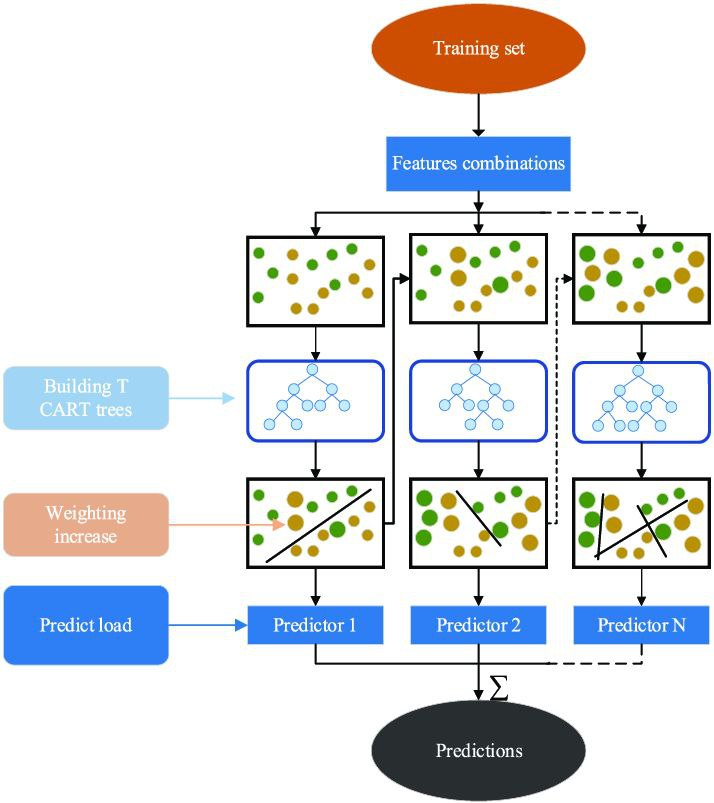


Рис.2. Архитектура модели

# Данные для обучения модели

Для обучения модели прогнозирования цен на поддержанные автомобили использовалcя датасет: *train.csv*.

Данные были взяты с *kaggle: https://www.kaggle.com/competitions/playground-series-s4e9/data.*

Основные атрибуты:

* Пробег;
* год выпуска;
* марка;
* цвет автомобиля;
* модель;
* тип топлива;
* цвет салона;
* количество цилиндров;
* история аварий.

## Анализ данных

1. Минимальное значение: 2000.

Это говорит о том, что в вашем наборе данных есть автомобили с очень низкими ценами (возможно, старые или поврежденные автомобили).

1. Максимальное значение: 2,954,083. Максимальное значение является значительно большим по сравнению с медианой и средним, что может свидетельствовать о наличии выбросов — редких, но экстремальных значений (возможно, элитные автомобили или редкие модели).
2. Среднее значение: 43,878.02. Среднее значение несколько выше медианы, что говорит о скошенности распределения вправо (в сторону более высоких цен). Это может быть связано с тем, что небольшое количество дорогих автомобилей "тянет" среднее вверх.
3. Медиана: 30,825. Медиана показывает, что половина автомобилей стоит меньше 30,825, а другая половина — больше. Поскольку медиана ниже среднего, это еще раз подтверждает наличие выбросов.
4. Дисперсия: 6,212,517,088.42. Дисперсия — это квадрат стандартного отклонения и также показывает степень разброса данных. Она говорит о том, что цены очень разнородны.
5. Количество ненулевых значений: 188,533. Данный показатель говорит о том, что все значения в наборе данных ненулевые и все строки содержат информацию о ценах автомобилей.
6. Квартиль 25 (Q1): 17,000. 25% автомобилей стоят меньше 17,000. Это нижний квартиль, который дает представление о том, как распределены наиболее доступные автомобили в наборе данных.
7. Мода: 15,000. Мода — это наиболее часто встречающееся значение в наборе данных. Значение 15,000 означает, что наибольшее количество автомобилей имеют цену именно 15,000, что может свидетельствовать о наличии большого числа доступных автомобилей.

.

# Обучение модели

## Модель прогнозирования цен на поддержанные автомобили

Процесс обучения модели включает следующие этапы:

1. Поиск и предобработка данных.
2. Разделение данных на тренировочную и тестовую выборки.
3. Настройка параметров модели.
4. Обучение модели.
5. Оценка точности модели на тестовой выборке.

Оценка качества модели: Использование метрик *MSE, RMSE* и *R*^2 для оценки качества модели. На рисунке 3 представлено значение метрики *RMSE.*

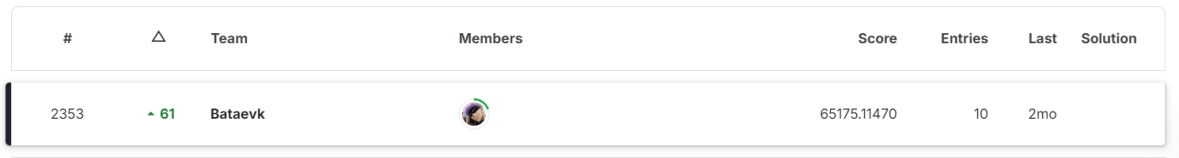


Рис. 3. Значение *RMSE*

# Описание ПО

Название ПО – веб−приложение для прогнозирования цен на поддержанные автомобили.

## Архитектура проекта

На рисунке 3 представлена архитектура данного проекта.

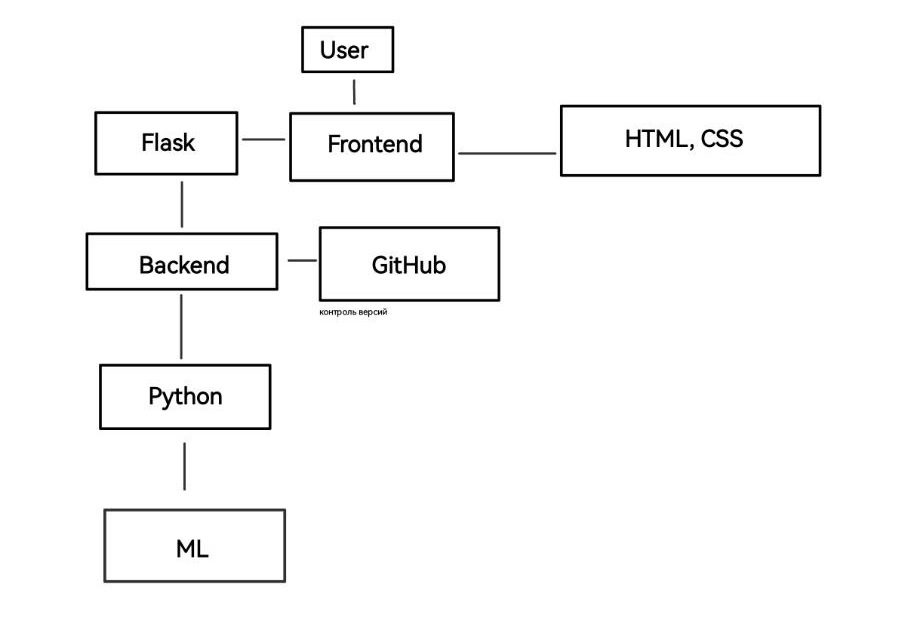


Рис. 4. Архитектура проекта

Данная архитектура включает в себя несколько ключевых компонентов и их взаимодействие:

1. пользователь (*User*): конечный пользователь взаимодействует с системой через пользовательский интерфейс.
2. Фронтенд (*Frontend*): пользовательский интерфейс, построенный с использованием *HTML* и *CSS*.
3. *Flask*: веб-фреймворк, который соединяет фронтенд с бекендом, обрабатывает запросы от пользователя.
4. *Backend*: основная логика приложения, написанная на языке программирования *Python*. Отвечает за обработку данных, выполнение алгоритма прогнозирования.
5. *GitHub*: контроль версий проекта.
6. Модель машинного обучения: *XGBoost.*

## Диаграмма взаимодействия пользователя с системой

На рисунке 4 представлена диаграмма отображающая процесс взаимодействия пользователя (автомобильного дилера) с системой.

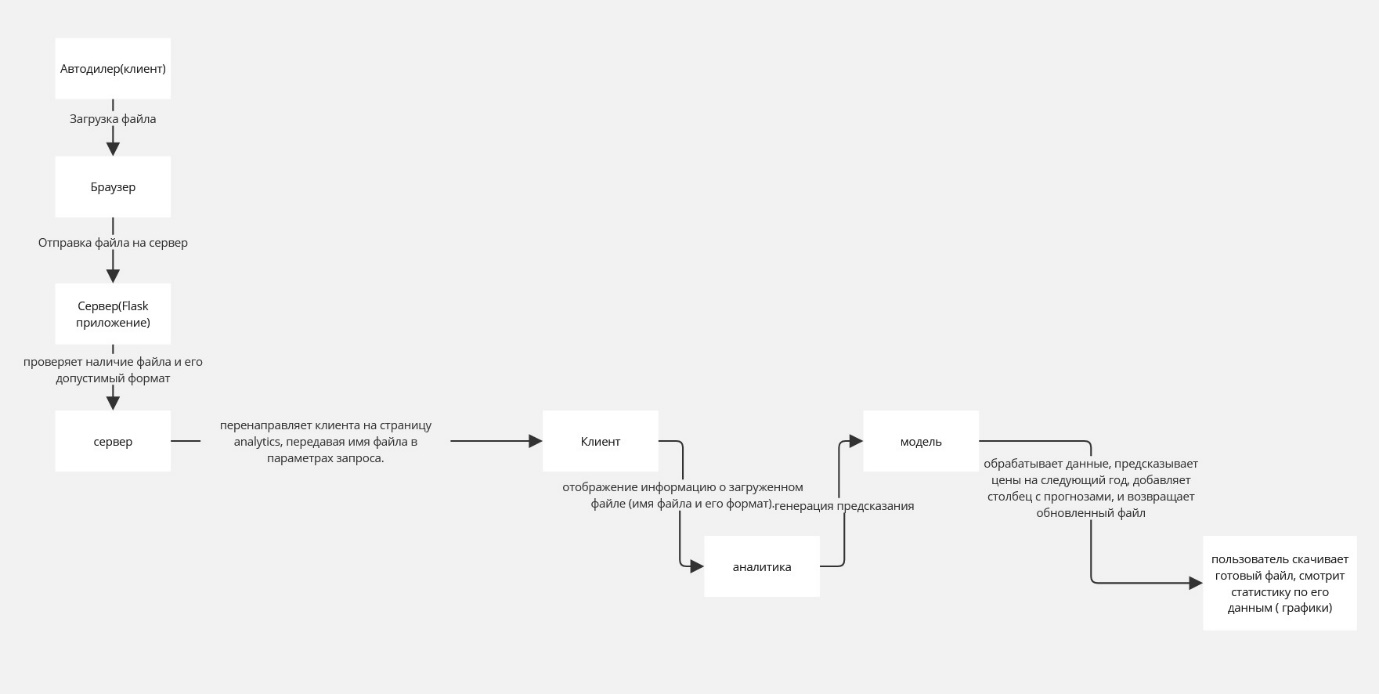


Рис. 5. Диаграмма

Процесс взаимодействия пользователя с системой:

1. Скачивание шаблона:

Пользователь переходит на страницу, где может скачать шаблон файла формата CSV для заполнения данных об автомобилях.

1. Заполнение шаблона:

Пользователь заполняет шаблон необходимой информацией (марка, модель, год выпуска, пробег, тип топлива и т.д.) и сохраняет файл.

1. Загрузка данных в систему:

Пользователь возвращается на сайт и загружает заполненный файл через соответствующую форму.

1. Обработка данных:

Система проверяет корректность загруженного файла. Происходит предобработка данных: очистка, масштабирование, преобразование и т.д.

1. Прогнозирование цен:

Загрузив данные, система с помощью модели машинного обучения *XGBoost* прогнозирует стоимость автомобилей.

1. Просмотр результата:

* Пользователь получает таблицу с данными, в которой добавлена колонка "*price*"*.*
* На странице отображается статистика: круговая диаграмма распределения автомобилей по маркам, средняя стоимость автомобилей и другие ключевые показатели.

1. Скачивание обновлённого файла:

Пользователь может скачать новый файл CSV с результатами прогнозирования, нажав на кнопку "Скачать".

1. Закрытие сессии:

Пользователь завершает работу с системой.

## Функциональные возможности

Основные функциональные возможности проекта:

1. **Загрузка шаблона CSV для ввода данных:**  
   Пользователи могут скачать шаблон в формате CSV, который содержит необходимые поля для ввода информации о подержанных автомобилях (например, марка, модель, год выпуска, пробег, тип топлива и т.д.).
2. **Заполнение и загрузка данных:**  
   После скачивания и заполнения шаблона пользователи могут загрузить готовый файл обратно в систему для дальнейшей обработки.
3. **Прогнозирование цен с использованием модели машинного обучения *XGBoost:***  
   После загрузки данных система применяет обученную модель машинного обучения *XGBoost* для прогнозирования рыночной стоимости автомобилей на основе заданных атрибутов.
4. **Просмотр результатов прогнозирования**:  
   Пользователи получают таблицу с прогнозируемыми ценами для каждого автомобиля, которая добавляется как новая колонка в загруженный файл.
5. **Визуализация статистики**:  
   На основе загруженных данных и результатов прогнозирования отображаются визуализации, такие как круговые диаграммы распределения автомобилей по маркам, средние цены и другие ключевые метрики.
6. **Скачивание обработанных данных:**  
   После получения результатов прогнозирования пользователи могут скачать обновлённый CSV файл, который содержит оригинальные данные с добавленными прогнозируемыми ценами.
7. **Интуитивно понятный пользовательский интерфейс:**  
   Простая и понятная навигация по веб-приложению позволяет пользователям легко загружать файлы, просматривать результаты и работать с визуализациями без сложных настроек.

Эти функциональные возможности делают проект удобным и эффективным инструментом для автодилеров, обеспечивая автоматизацию процесса оценки стоимости подержанных автомобилей и улучшая качество бизнес-процессов.

# Результаты

В ходе работы над проектом было разработано веб-приложение, автоматизирующее процесс прогнозирования цен на подержанные автомобили.

* Интеграция модели машинного обучения *XGBoost* позволяет производить точное прогнозирование на основе загруженных данных.
* Предоставлен удобный интерфейс, позволяющий пользователям:
  1. Скачать шаблон для заполнения.
  2. Загружать данные и получать результаты прогнозирования в формате CSV.
  3. Анализировать статистику с помощью визуализации (круговая диаграмма и т.д.).
* Веб-приложение помогает автодилерам минимизировать финансовые риски, улучшить процесс ценообразования и повысить доверие со стороны клиентов.
* Система поддерживает экспорт обработанных данных, что упрощает дальнейшее использование полученных результатов.

На рисунках 5−9 представлен интерфейс проекта.

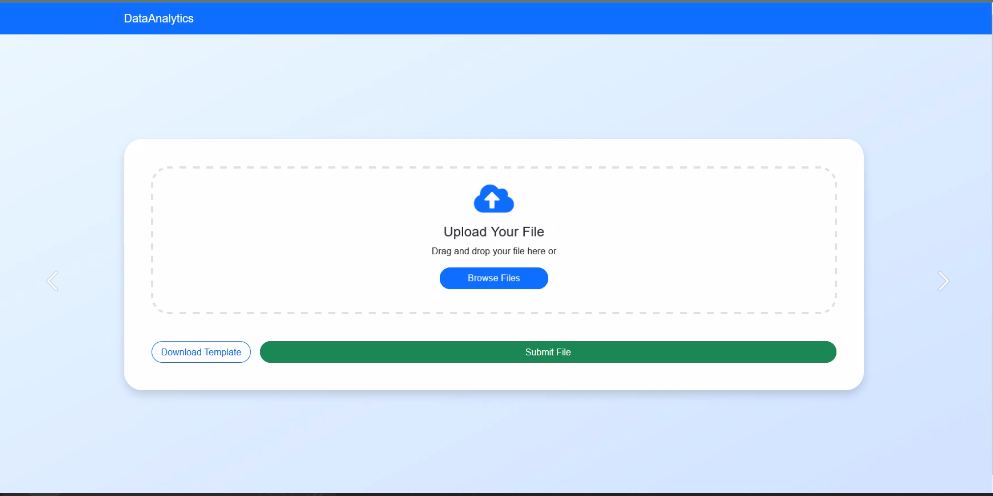


Рис. 6. Начальная страница

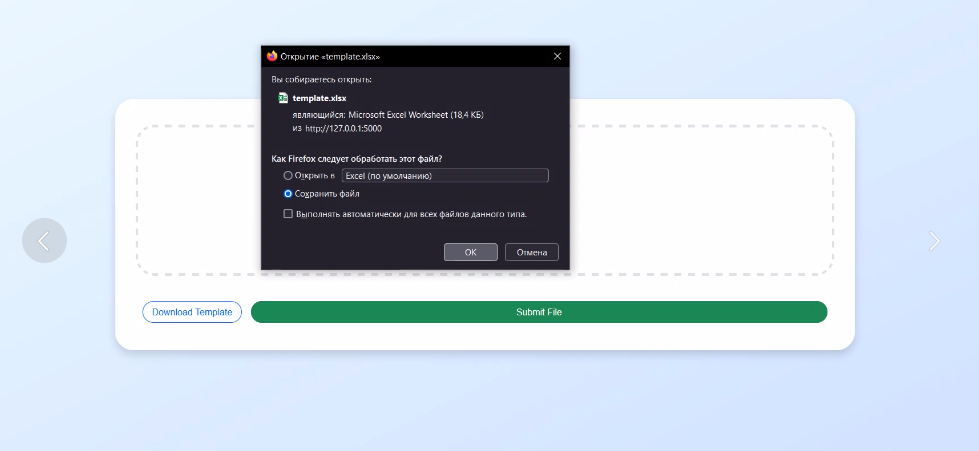


Рис. 7. Скачивание шаблона

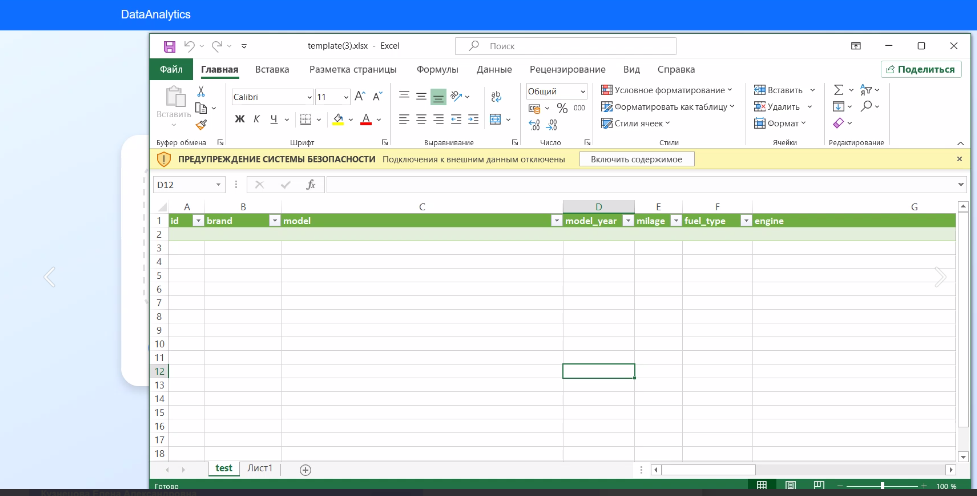


Рис. 8. Шаблон

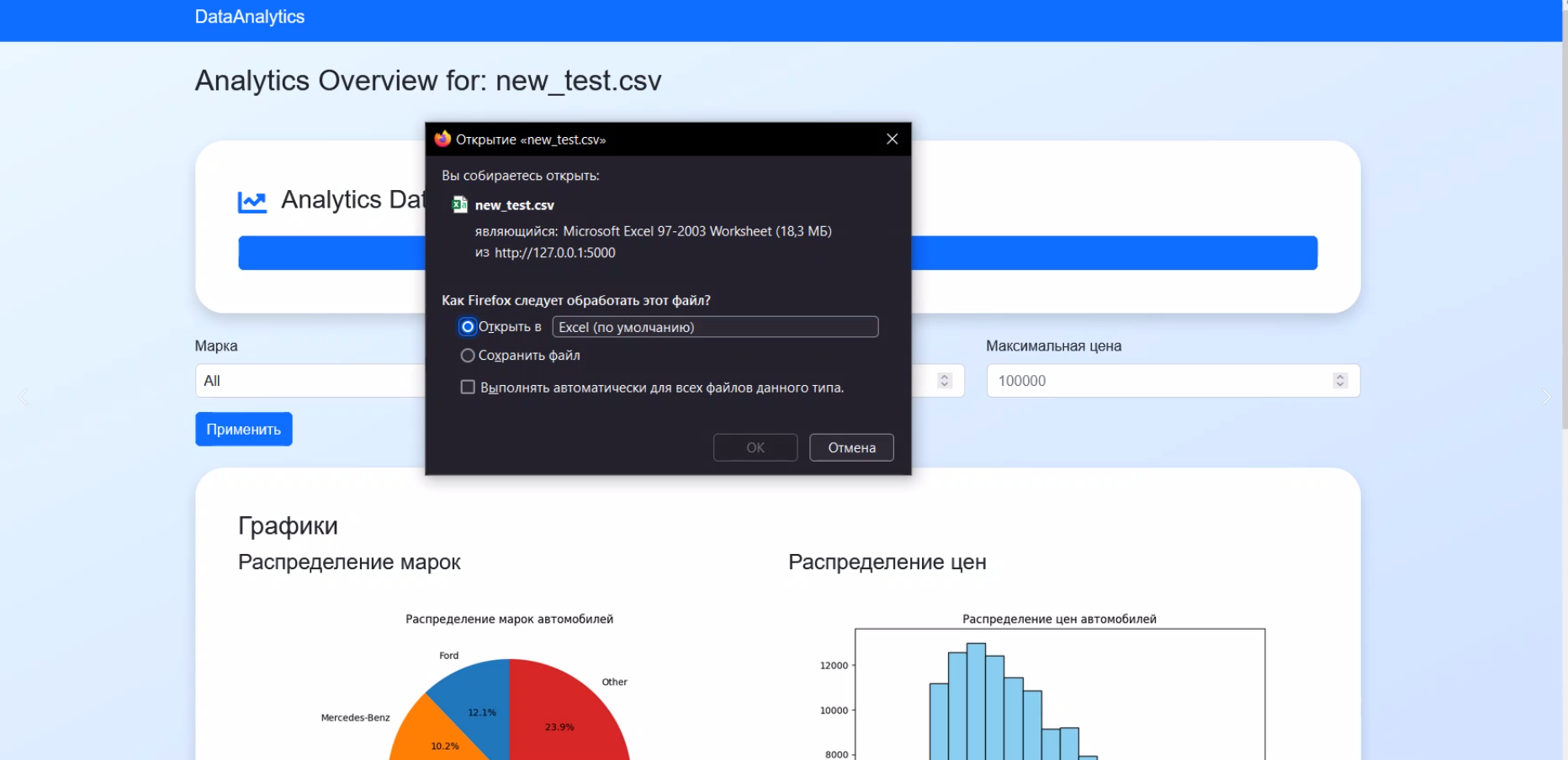


Рис. 9. Скачивание обновленного файла, добавлен “price”

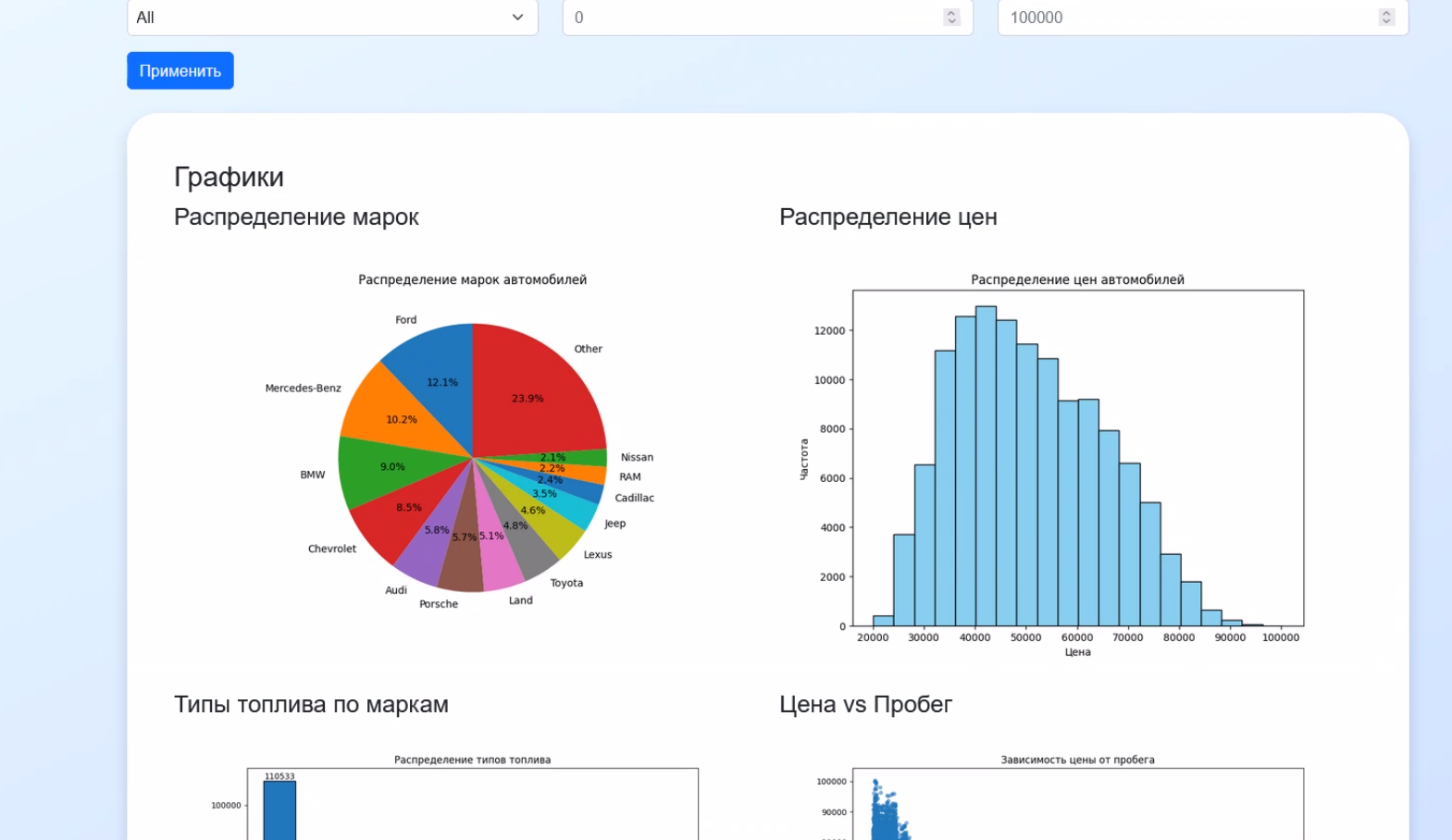


Рис. 10. Аналитика

Проект можно посмотреть на *GitHub*: https://github.com/Bataevk/Regression-of-Used-Car-Prices-by-bataevk

# Заключение

Реализация данного проекта позволяет решить одну из актуальных проблем автомобильного рынка, предоставляя удобный инструмент для прогнозирования цен на подержанные автомобили. Автоматизация процессов на основе машинного обучения повышает эффективность работы автодилеров и помогает принимать более обоснованные решения.

# Ссылки на источники

1. *Flask Documentation // Flask URL*: https://flask.palletsprojects.com/en/stable/ (дата обращения: 13.10.2024).
2. *Regression of Used Car Prices // Kaggle URL*: https://www.kaggle.com/competitions/playground-series-s4e9/data (дата обращения: 26.09.2024).