**Содержание**

[Введение 3](#_Toc167972877)

Глава [1. Анализ рынка автомобилей 6](#_Toc167972878)

[1.1 Тенденции рынка автомобилей 6](#_Toc167972879)

[1.2. Анализ спроса и предложения на рынке автомобилей 6](#_Toc167972880)

[1.3. Конкурентный анализ автомобильных брендов в Кыргызстане 6](#_Toc167972881)

Глава [2. Методы прогнозирования цен на автомобили 7](#_Toc167972882)

[2.1. Статистические методы 7](#_Toc167972883)

[2.2. Методы машинного обучения 8](#_Toc167972884)

Глава [3. Реализация программного обеспечения 10](#_Toc167972885)

[3.1. Описание признаков 10](#_Toc167972886)

[3.2. Очистка данных 12](#_Toc167972887)

[3.3. Построение модели 15](#_Toc167972888)

[3.4. Выбор лучшей модели 22](#_Toc167972889)

[Заключение 24](#_Toc167972890)

[Список использованных источников 26](#_Toc167972891)

# **Введение**

Прогнозирование цен на автомобили играет важную роль как для покупателей, так и для автодилеров в автомобильной индустрии. Стабильность и точность таких прогнозов позволяют предсказывать спрос и предложение на рынке, определять оптимальные стратегии ценообразования, а также принимать рациональные решения о покупке и продаже автомобилей.

Автомобильный рынок является динамичным и подверженным различным внешним и внутренним факторам, которые могут влиять на ценовую динамику. Экономическая конъюнктура, технологические инновации, изменения в регулировании, а также предпочтения потребителей - все это оказывает значительное влияние на цены на автомобили.

Применение моделей машинного обучения для прогнозирования цен на автомобили имеет множество преимуществ. Во-первых, они позволяют учитывать большое количество факторов, влияющих на ценообразование, таких как марка и модель автомобиля, его технические характеристики, возраст, пробег, состояние, а также внешние факторы, такие как экономическая ситуация, сезонные колебания и предпочтения потребителей. Во-вторых, такие модели могут постоянно обновляться и улучшаться по мере поступления новых данных, что обеспечивает актуальность и точность прогнозов. В-третьих, применение машинного обучения может значительно снизить субъективность и ошибки, связанные с человеческим фактором, при принятии решений о ценообразовании.

**Цель курсового проекта** заключается в разработке модели прогнозирования цен на автомобили с использованием современных методов анализа данных и машинного обучения. Данная модель будет способствовать более точному определению ценовой динамики на автомобильном рынке и предоставит полезные инсайты для принятия решений как потребителям, так и участникам рынка.

Для достижения поставленной цели были определены следующие задачи:

* Изучить текущее состояние автомобильного рынка и основные факторы, влияющие на цены автомобилей.
* Проанализировать существующие методы прогнозирования цен и выбрать наиболее подходящий.
* Собрать и подготовить данные для построения модели.
* Разработать и протестировать модель прогнозирования.
* Оценить точность модели и сделать выводы о ее применимости.

**Объектом исследования** является автомобильный рынок, который включает в себя подержанные автомобили. **Предметом исследования** выступают методы прогнозирования цен на автомобили, включая их применение и анализ эффективности.

**По структуре** курсовой проект состоит из: введения, трех глав, заключения и списка использованных источников.

Над данным курсовым проектом работала команда в следующем составе:

1. **Зарлыков Келсинбек** был ответственен за руководство командой и координацию работ, а также за принятие ключевых решений по проекту. Его вклад также включает помощь в написании документации и обеспечение ее соответствия стандартам. Его работа отражена в разделе 1,2,3.

2. **Эрдовлатов Исмадияр** занимался разработкой и поддержкой кода, а также реализацией алгоритмов для визуализации данных и тестирования моделей машинного обучения. Его вклад представлен в разделе 1,2,3.

3. **Дарибаев Мирланбек** разрабатывал фронтенд часть проекта, создавал пользовательский интерфейс и обеспечивал интеграцию с бэкэндом. Его работа описана в разделе 1,2,3.

# **Глава 1. Анализ рынка автомобилей**

**1.1 Тенденции рынка автомобилей**

Основные тенденции рынка автомобилей, оказывающие влияние на рынок автомобилей в Кыргызстане. В частности, проанализированы изменения в предпочтениях потребителей относительно моделей Mercedes-Benz W210, Honda FIT, KIA K5 и Toyota Camry 70. Среди основных тенденций можно выделить рост интереса к гибридным автомобилям в связи с усиливающимися экологическими нормами и регулированиями[[1]](#footnote-1). В Кыргызстане также наблюдается повышенный интерес к технологиям автопилотирования и интернета вещей, что влияет на изменение потребительских предпочтений и требований к автомобилям. Анализ этих тенденций помогает понять, какие изменения в дизайне, функциональности и маркетинге автомобилей могут быть востребованы на рынке Кыргызстана.

## **1.2. Анализ спроса и предложения на рынке автомобилей**

Анализ спроса и предложения на рынке автомобилей в Кыргызстане рассматривается динамика продаж, рыночные доли различных автомобильных брендов, а также влияние экономических факторов, таких как уровень доходов населения, кредитные ставки и налоговая политика, на цены и объемы продаж автомобилей.

## **1.3. Конкурентный анализ автомобильных брендов в Кыргызстане**

Анализ конкурентной среды на рынке автомобилей в Кыргызстане с учетом Mercedes-Benz W210, Honda FIT, KIA K5 и Toyota Camry 70. Исследуются основные участники рынка, их стратегии продаж и маркетинга, доли рынка и ключевые конкурентные преимущества и недостатки. Также рассматривается влияние глобальных и местных факторов на позиции брендов на рынке Кыргызстана.

# **Глава 2. Методы прогнозирования цен на автомобили**

## **2.1. Статистические методы**

Статистические методы представляют собой важный инструмент для анализа и прогнозирования цен на автомобили. Они базируются на анализе статистических данных и закономерностей, выявленных в исторических и текущих данных о ценах.

Анализ временных рядов:

Один из ключевых подходов в статистическом прогнозировании - это анализ временных рядов цен на автомобили. Мы исследуем данные за определенный временной период и выявляем основные тренды, сезонные колебания и циклические изменения в ценах. Это позволяет нам прогнозировать будущие изменения на рынке автомобилей и принимать соответствующие решения.

Методы сглаживания и экстраполяции:

Для прогнозирования будущих цен на автомобили используются методы сглаживания и экстраполяции. Мы применяем различные техники, такие как скользящая средняя и экспоненциальное сглаживание, чтобы устранить шумы и выбросы в данных, а также предсказать будущие значения цен на основе имеющихся данных.

Методы регрессионного анализа:

Еще одним важным статистическим методом является регрессионный анализ. Мы строим регрессионные модели, которые описывают связь между ценами на автомобили и различными факторами, такими как характеристики автомобилей, экономические показатели и т. д. Эти модели позволяют нам оценить влияние каждого фактора на цены и прогнозировать будущие значения на основе этих оценок.

Преимущества статистических методов:

Статистические методы обладают рядом преимуществ, включая возможность учитывать сложные взаимосвязи между переменными, а также способность обрабатывать большие объемы данных. Они позволяют нам создавать точные и надежные модели прогнозирования цен на автомобили, которые могут быть использованы для принятия стратегических решений в автомобильной индустрии.[[2]](#footnote-2)

## **2.2. Методы машинного обучения**

Методы машинного обучения являются сильным инструментом для прогнозирования цен на автомобили. Они обеспечивают возможность автоматического извлечения закономерностей из обширных объемов данных и построения моделей, способных делать точные прогнозы на основе этой информации.

Использование алгоритмов машинного обучения:

Мы широко применяем различные алгоритмы машинного обучения, такие как случайный лес, деревья решений, линейная регрессия и многие другие. Это позволяет нам создавать модели, которые учитывают сложные взаимосвязи между различными параметрами и способны предсказывать цены на автомобили с высокой точностью.

**Подготовка данных:**

Однако, для успешного применения алгоритмов машинного обучения необходима тщательная предварительная обработка данных. Этот процесс включает в себя несколько этапов:

* Очистка данных: Мы проводим удаление выбросов, заполнение пропущенных значений и обработку ошибок в данных.
* Нормализация данных: Мы приводим признаки к одному масштабу для более эффективного использования алгоритмами.
* Выбор информативных признаков: Мы отбираем наиболее значимые признаки, которые могут влиять на цены на автомобили, для дальнейшего анализа.

**Выбор модели:**

После предварительной обработки данных мы сталкиваемся с выбором наиболее подходящей модели машинного обучения для решения конкретной задачи прогнозирования. Этот выбор зависит от нескольких факторов, включая объем и структуру данных, тип задачи (регрессия или классификация), требуемую скорость обучения и предсказания, а также интерпретируемость модели.[[3]](#footnote-3)

# **Глава 3. Реализация программного обеспечения**

## **3.1. Описание признаков**

Наш датасет содержит информацию об автомобилях по следующим признакам:

**name**: Название модели автомобиля.

**company:** Производитель автомобиля.

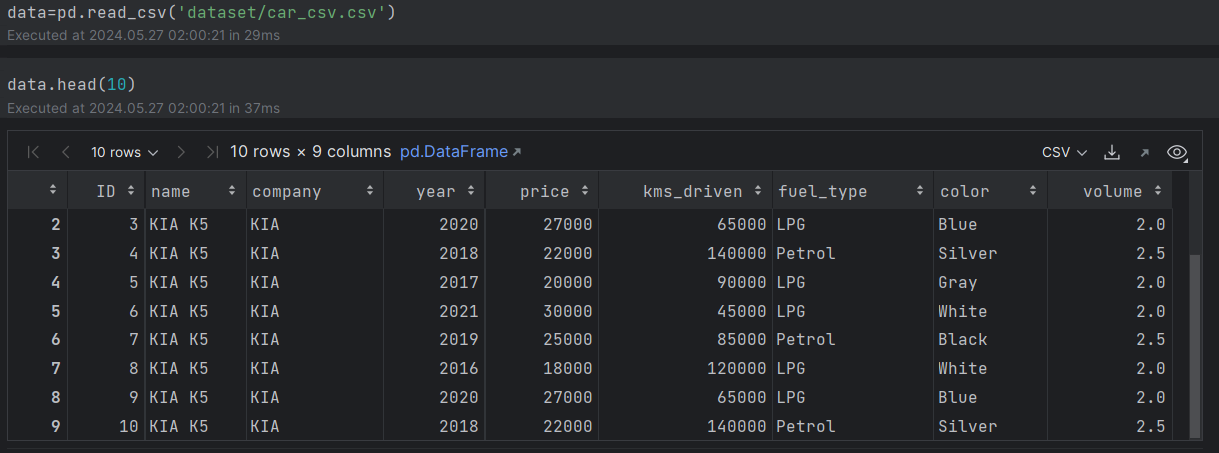
**year:** Год выпуска автомобиля.

**price:** Цена автомобиля.

**kms\_driven:** Пробег автомобиля в километрах.

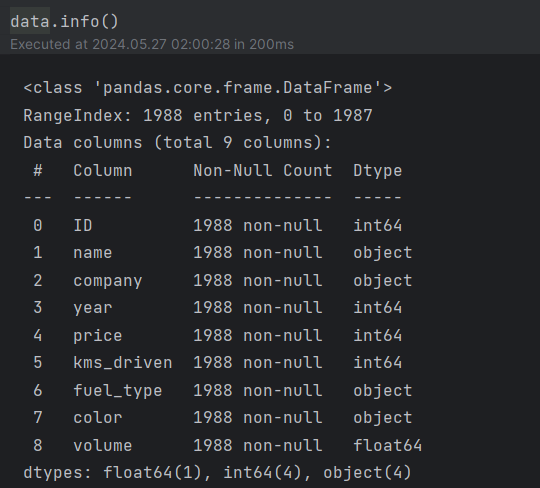
**fuel\_type**: Тип топлива, используемого автомобилем (например, бензин, дизель, электрический).

**color:** Цвет автомобиля. volume: Объем двигателя автомобиля.

****

*рис.1 Прочтение файлов из csv-файла*

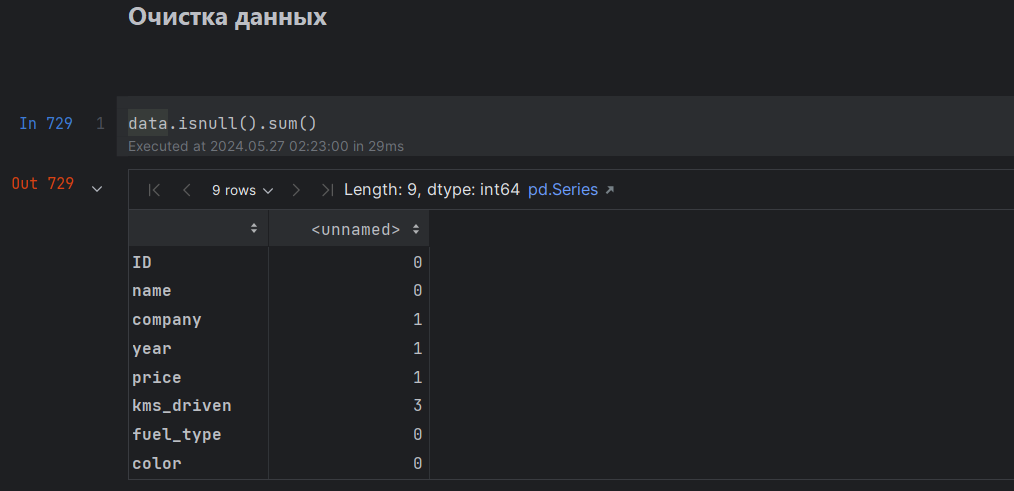
Этот код(Рис.1) загружает данные из csv файла и выводить 10 строк датасета.

****

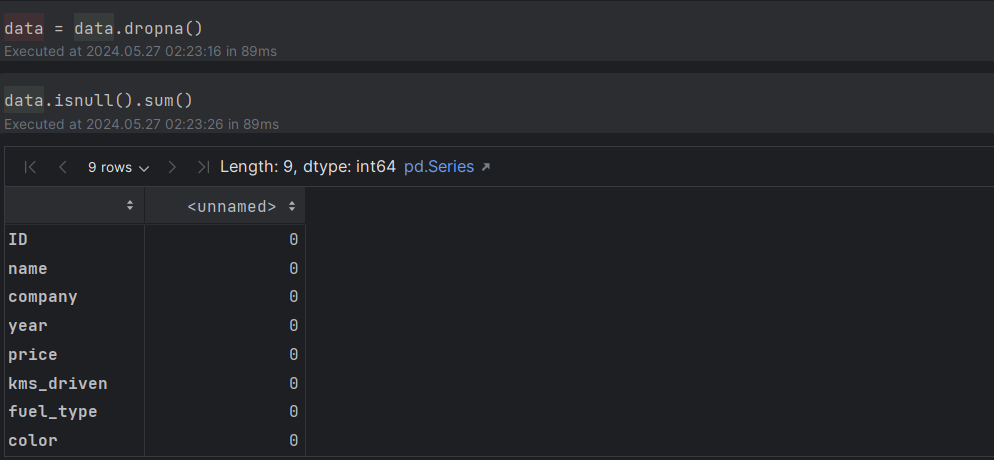
*рис.2 Структура дата-фрейма*

Метод data.info() из библиотеки pandas в Python предоставляет сводку о структуре дата фрейма

## **3.2. Очистка данных**

*****рис.3 Проверка на пропущенных значений*

На этом коде (Рис.3) мы с помощью метода isnull() проверяем на наличие пропущенных значенией в датасете.

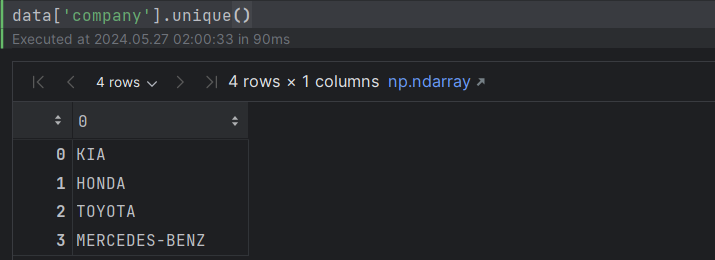
****

*рис.4 Удаление пропущенных значений*

Этот код(Рис.4) удаляет пропущенные значение с помощью метода dropna() и и снова проверяет, остались ли пропущенные значение.

****

*рис.5* *Проверка на наличие дубликатов*

****

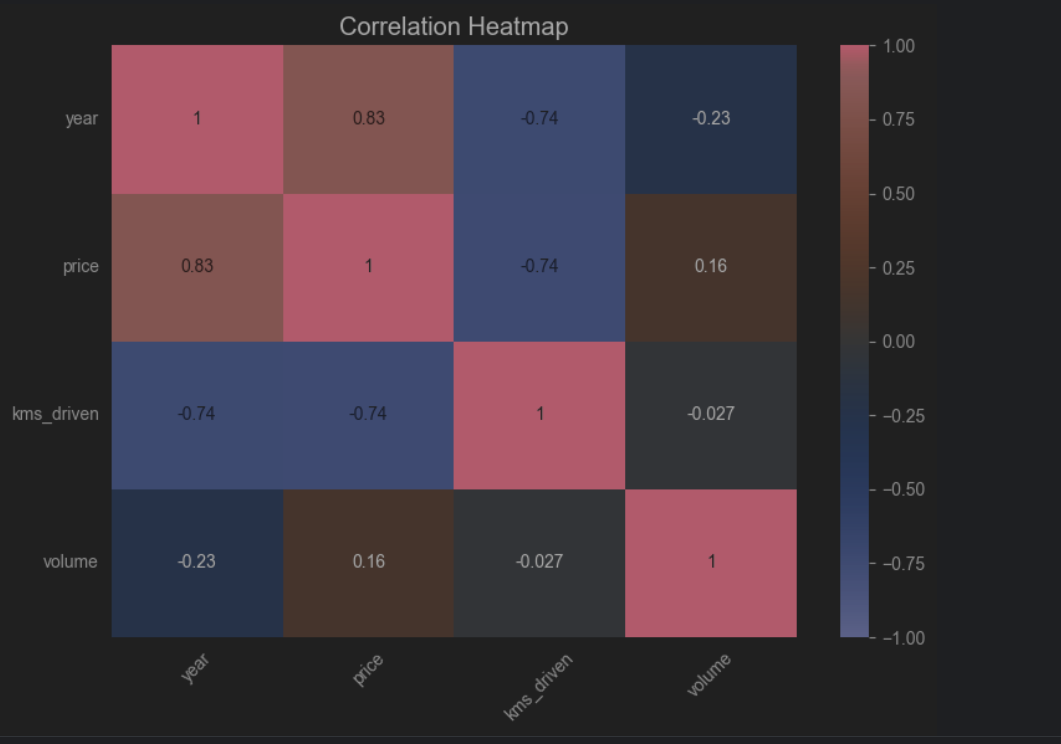
*рис.6 Уникальные значение*

Этот фрагмент кода извлекает уникальные значения из столбца 'company' в датафрейме `data`.

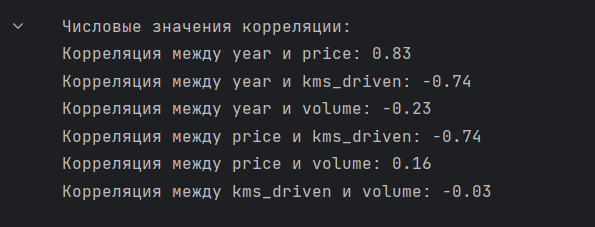
Это необходимо для определения всех уникальных производителей автомобилей, представленных в наборе данных.

Уникальные значения могут быть полезны для анализа распределения автомобилей по компаниям

или для выполнения дальнейших фильтраций и агрегаций на основе компаний.

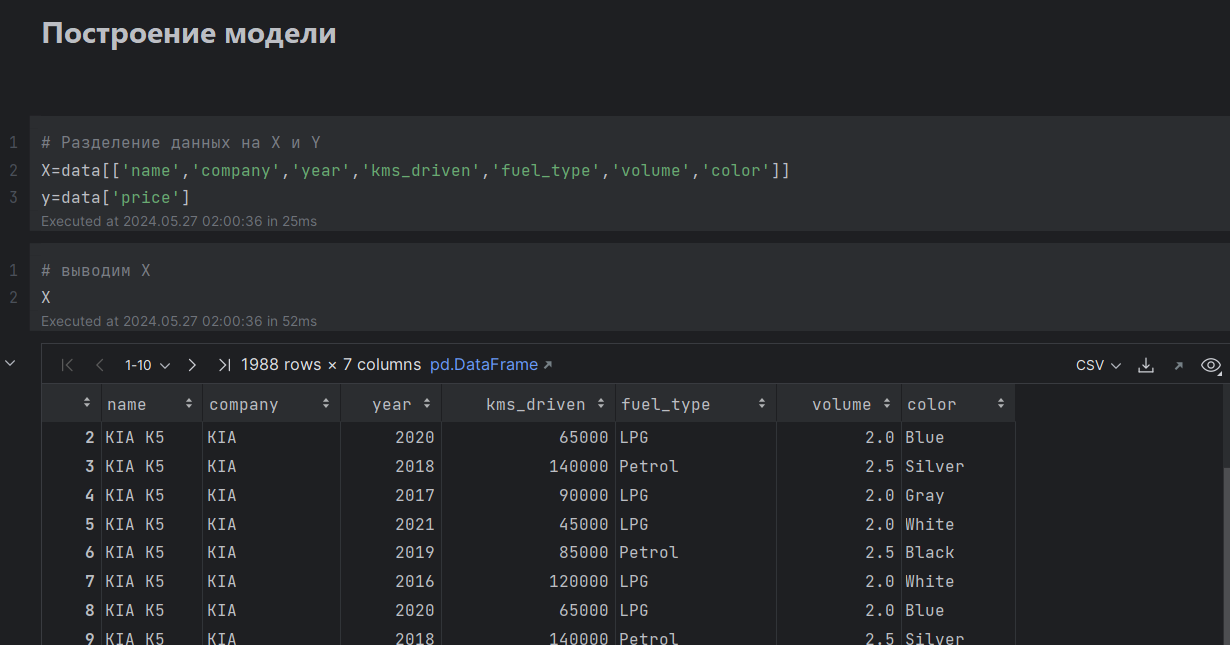
****

*рис.7 Визуализация корреляции между признаками*

****

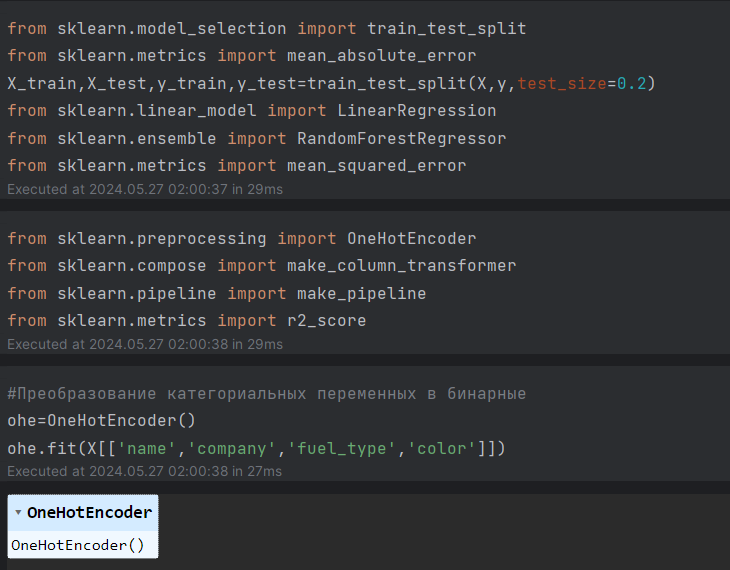
*рис.8* *Корреляция признаков*

## **3.3. Построение модели**

****

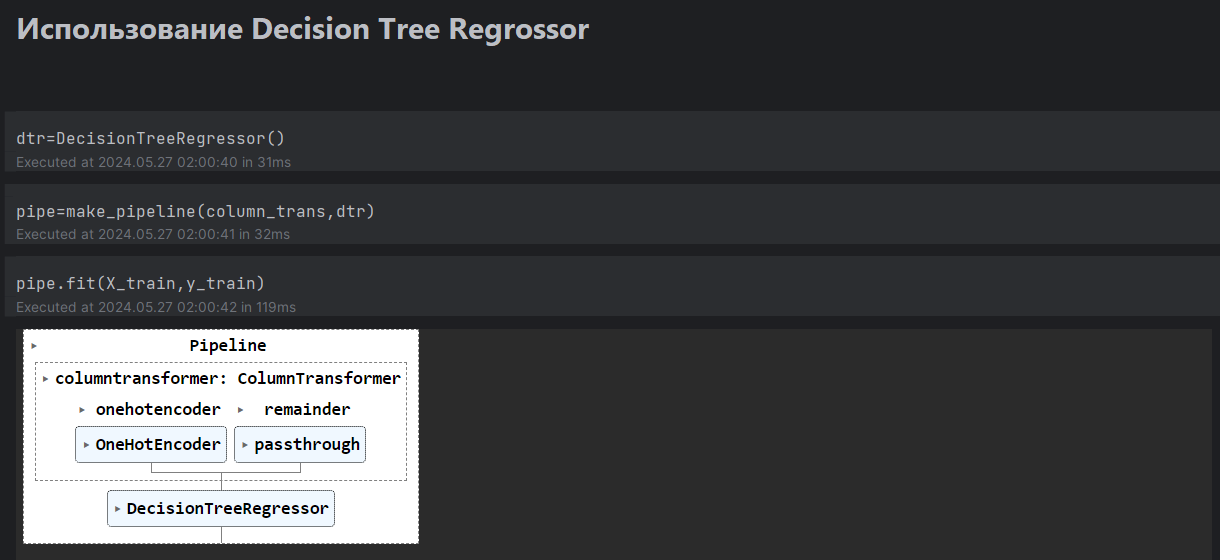
*рис.9 Разделение данных на X и Y*

Разделяем данные на признаки (X) и целевую переменную (Y). X содержит информацию о модели, производителе, годе выпуска, пробеге, типе топлива, объеме и цвете автомобилей. Y содержит цены на автомобили, которые мы хотим прогнозировать.

*****рис.10 Преобразование категориальных данных*

Этот блок кода содержит импорты необходимых модулей и функций для выполнения следующих действий:

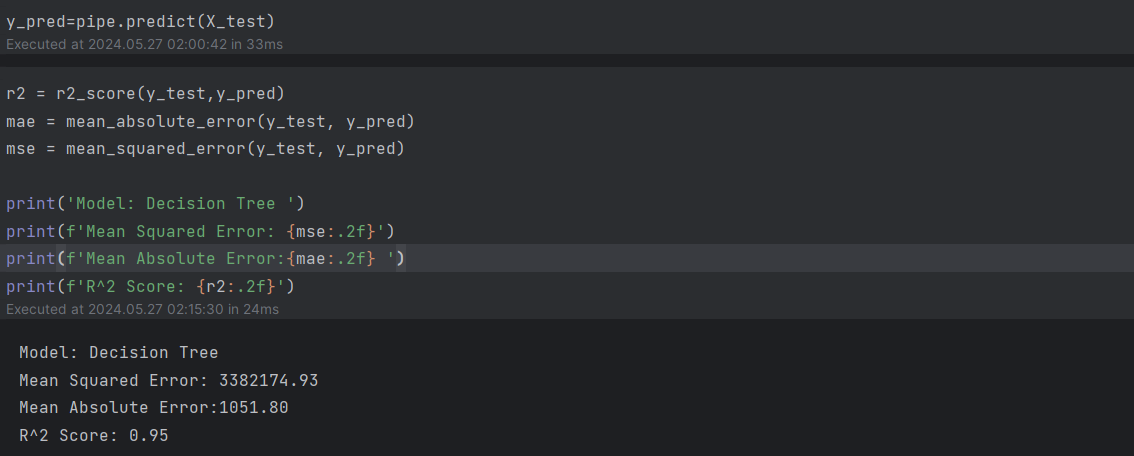
* Импорт функции `train\_test\_split` из модуля `sklearn.model\_selection` для разделения данных на обучающий и тестовый наборы.
* Импорт функции `mean\_absolute\_error` из модуля `sklearn.metrics` для вычисления средней абсолютной ошибки.
* Импорт класса `LinearRegression` из модуля `sklearn.linear\_model` для построения модели линейной регрессии.
* Импорт класса `Random Forest Regressor` из модуля `sklearn.ensemble` для построения модели случайного леса.
* Импорт функции `mean\_squared\_error` из модуля `sklearn.metrics` для вычисления среднеквадратической ошибки.
* А также преобразование категориальных переменных в бинарные значение.

****

*рис.11 Использование Decision Tree Regression*

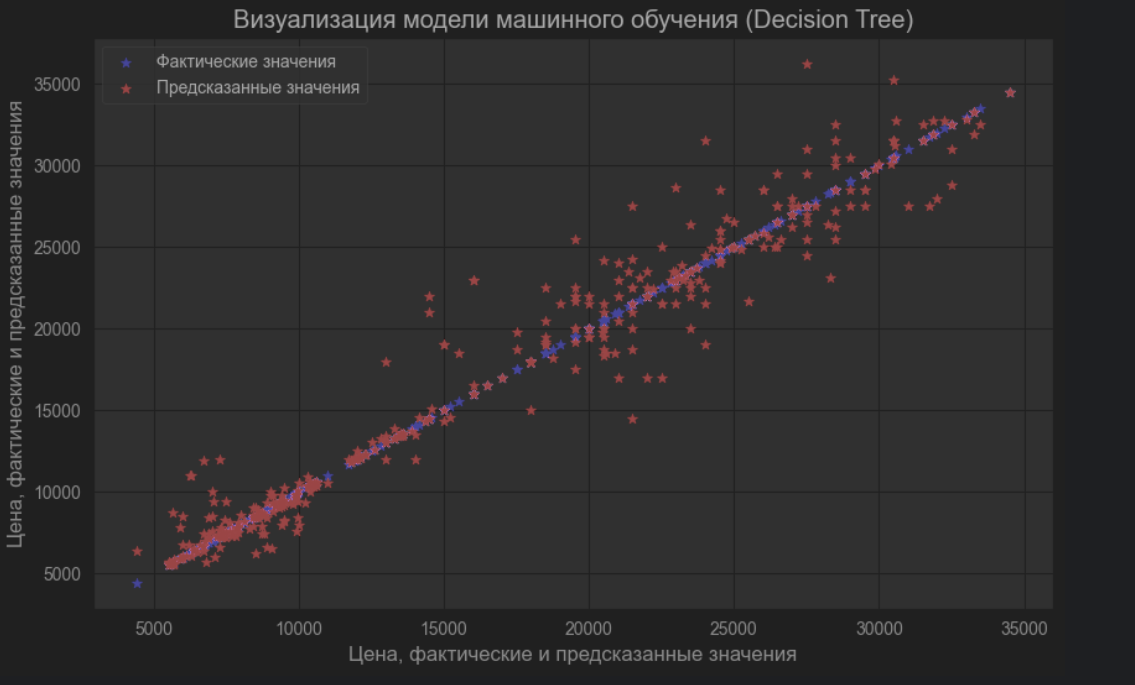
Этот код (Рис.11) выполняет следующие действия:

* Создает экземпляр модели регрессии на основе дерева решений с помощью класса `DecisionTreeRegressor`.
* Импортирует функцию `make\_pipeline` из модуля `sklearn.pipeline` для создания конвейера (пайплайна) для последовательного преобразования данных и применения модели.
* Создает конвейер (пайплайн) с использованием преобразований столбцов (column\_trans) и модели регрессии на основе дерева решений (dtr).
* Обучает конвейер на обучающем наборе данных (`X\_train` и `y\_train`) с помощью метода `fit()`.

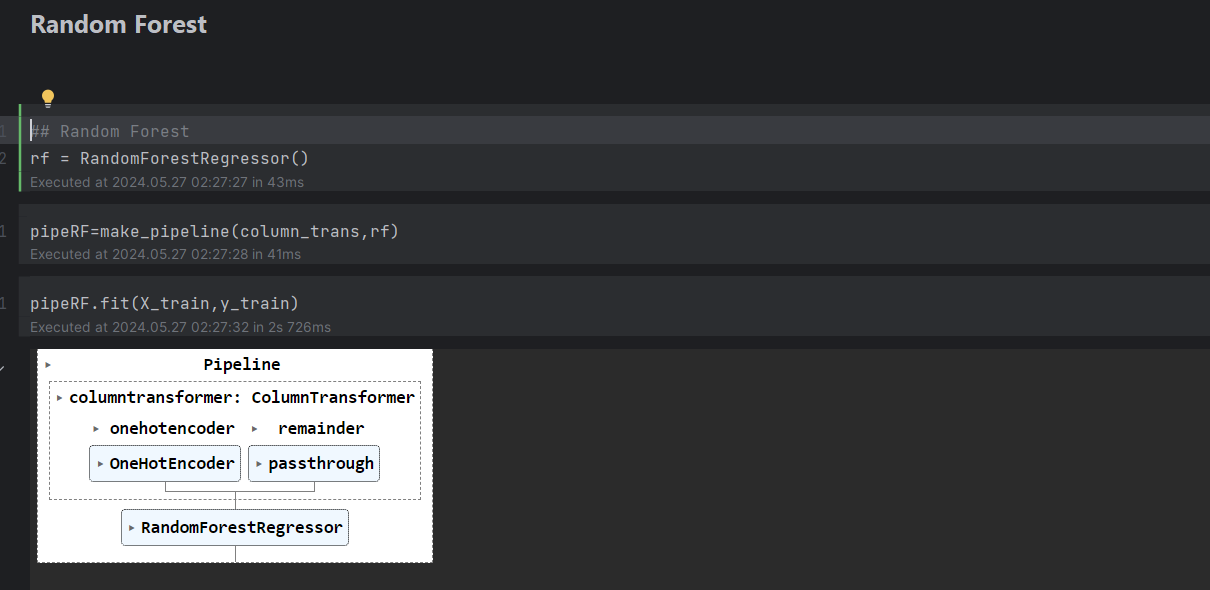
*****рис.12* *Предсказание на тестовых данных*

Этот блок кода(Рис.12) выполняет следующие действия:

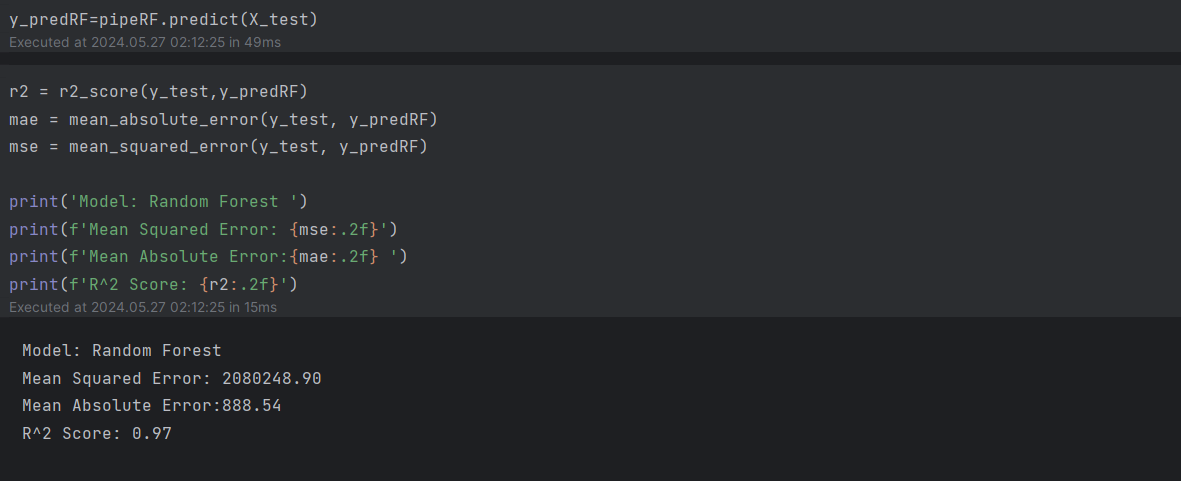
* Выполняет прогнозирование целевой переменной `y\_pred` на тестовом наборе данных `X\_test` с помощью метода `predict()` конвейера `pipe`.
* Импортирует функцию `r2\_score` из модуля `sklearn.metrics` для вычисления коэффициента детерминации (R-квадрат).
* Вычисляет коэффициент детерминации (R-квадрат), среднюю абсолютную ошибку (MAE) и среднеквадратическую ошибку (MSE) с помощью функций `r2\_score`, `mean\_absolute\_error` и `mean\_squared\_error` соответственно.
* Выводит результаты, включая название модели, значения MSE, MAE и R-квадрат.

****

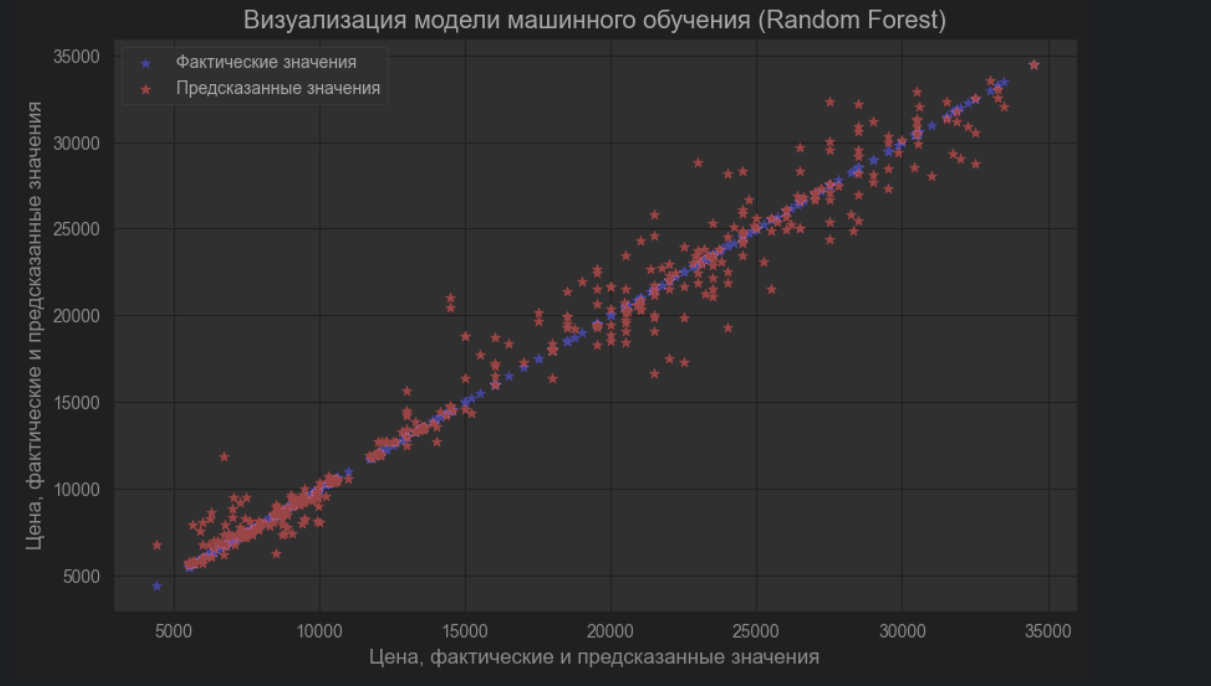
*рис.13* *Визуализация модели машинного обучение дерева решений*

****

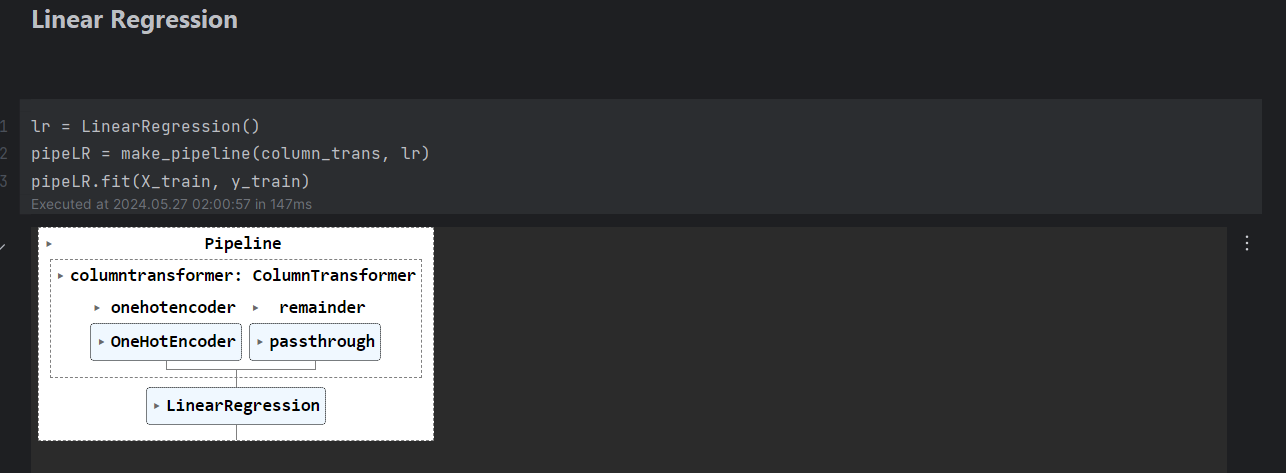
*рис.14* *Использование Random Forest*

****

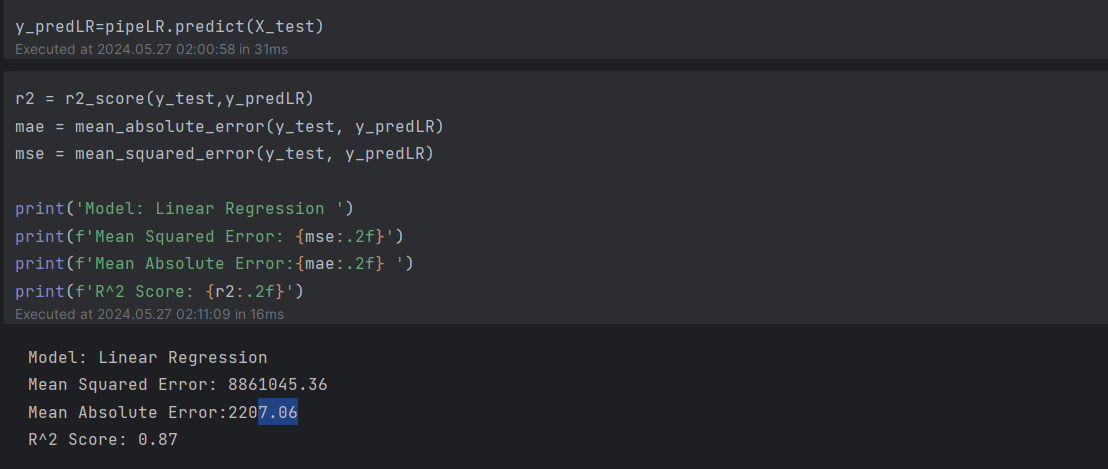
*рис.15 Производительность модели машинного обучения*

****

*рис.16* *Визуализация модели машинного обучение случайный лес*

****

*рис.17* *Использование Linear regression*

****

*рис.18* *Производительность модели машинного обучения*

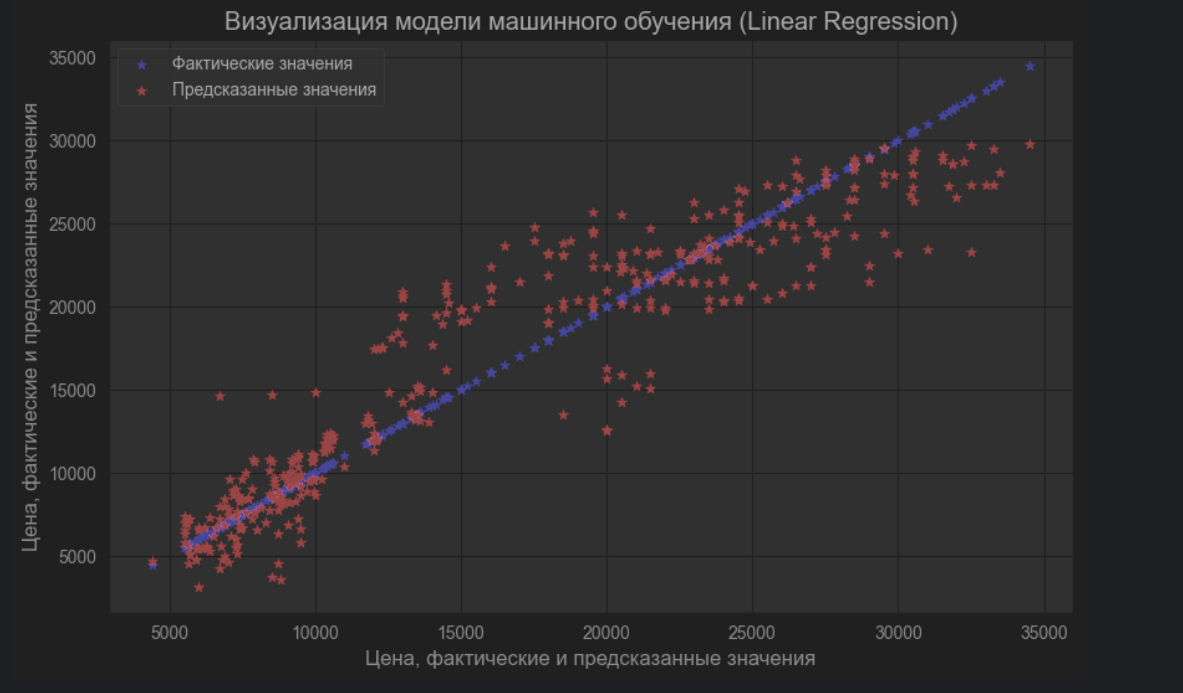
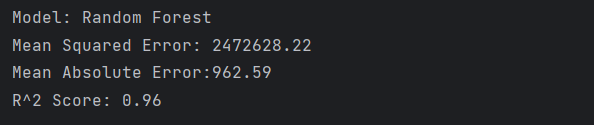
****

рис.19 *Визуализация модели машинного обучение Linear Regression*

## **3.4. Выбор лучшей модели**

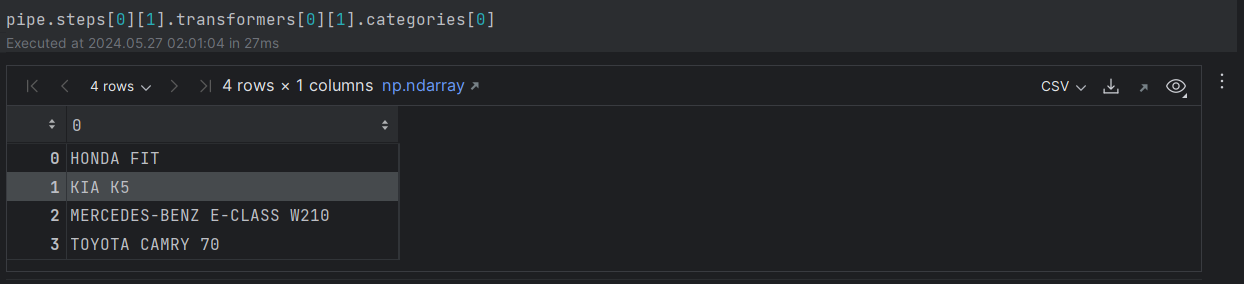
****

*рис.20 Выбор лучшей модели*

После того как мы обучили наши данные с помощью трех моделей машинного обучения, самый высокий результат показал Random Forest.

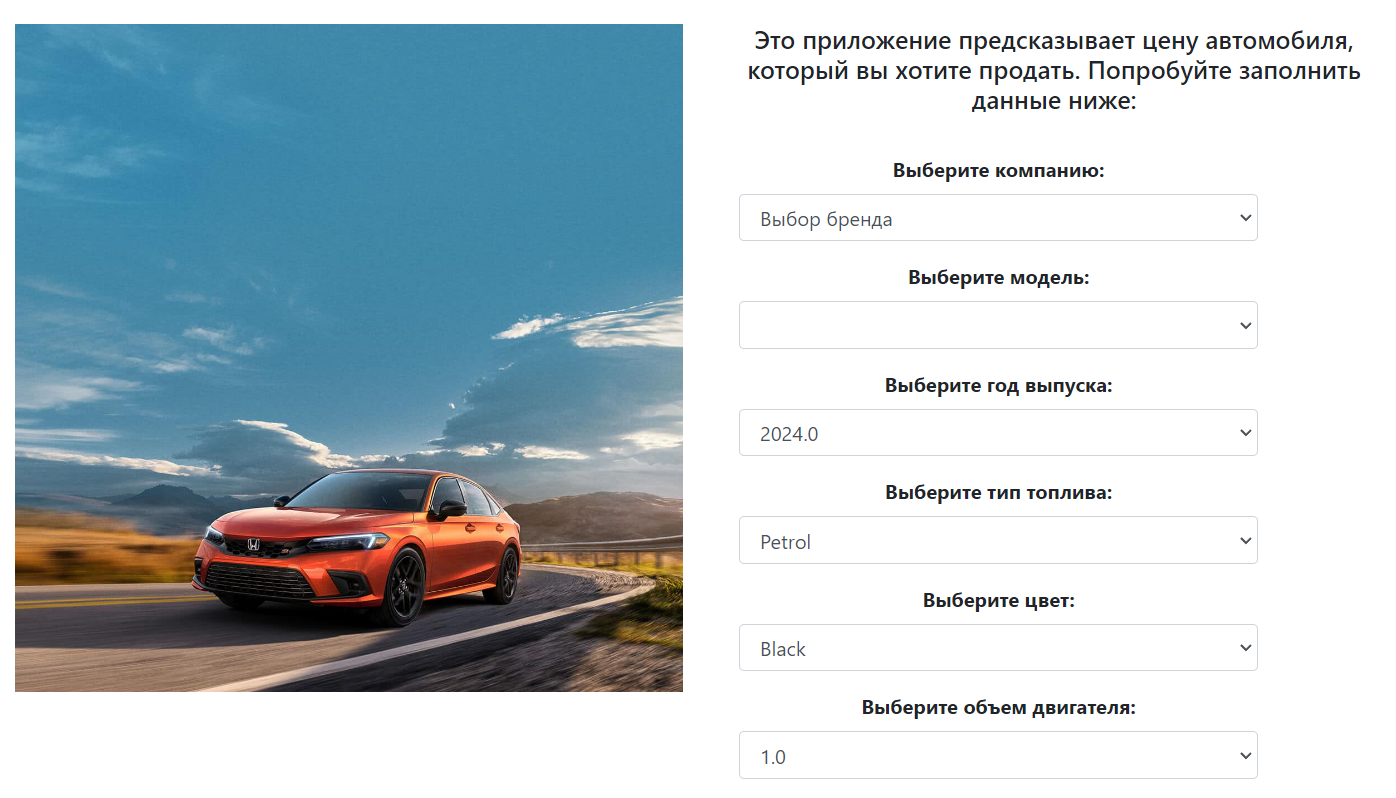
****

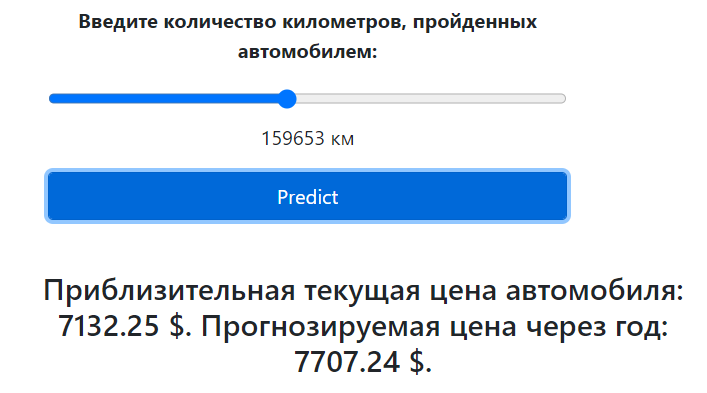
*рис.21* *Сериализация прогнозируемых данных в pkl*

****

*рис.22* *Вывод уникальных марок машин*

Этот код (Рис.22) выводить марки наших автомобилей.

****

****

*рис.23 Веб-сайт проекта*

Рисунок 23 представляет скриншот программы, которая использует различные характеристики автомобиля для прогнозирования его цены. Это хороший способ визуализировать вашу программу и показать ее функциональность.

# **Заключение**

Итогом нашего разработки курсового проекта по прогнозированию цен на автомобили стала разработка модели, способной точно предсказывать ценовую динамику на автомобильном рынке. На протяжении исследования мы прошли через несколько ключевых этапов, включающих сбор, обработку и анализ данных, а также применение различных методов машинного обучения для построения прогнозной модели.

В ходе работы были собраны данные о различных характеристиках автомобилей, таких как марка, модель, год выпуска, пробег, а также внешние факторы, влияющие на цены. Эти данные прошли тщательную очистку и подготовку, включая обработку пропущенных значений, нормализацию и кодирование категориальных признаков, что позволило обеспечить их пригодность для анализа и моделирования.

На этапе исследовательского анализа данных (EDA) были использованы различные методы визуализации для понимания распределения и динамики цен. Мы построили гистограммы, тепловые карты корреляций, что помогло выявить ключевые закономерности и факторы, влияющие на цены автомобилей. Визуализация позволила нам глубже понять структуру данных и выявить важные тренды, такие как сезонные колебания и долгосрочные изменения.

Для прогнозирования цен на автомобили были применены различные методы машинного обучения, включая линейную регрессию, случайные леса и дерево решений. Каждая модель прошла настройку гиперпараметров и была оценена на основе метрик точности, таких как среднеквадратичная ошибка (MSE), средняя абсолютная ошибка (MAE) и коэффициент детерминации (R²). Это позволило выбрать наиболее точную и стабильную модель для прогнозирования цен.

Наша модель учитывает широкий спектр факторов, таких как марка, модель, год выпуска, пробег, состояние автомобиля и другие характеристики. Результаты исследования показали, что использование современных технологий и аналитических подходов позволяет значительно улучшить точность прогнозов цен на автомобили. Разработанная модель демонстрирует высокую точность и стабильность, что делает ее ценным инструментом для различных участников автомобильного рынка.

Потребители могут использовать модель для принятия обоснованных решений о покупке или продаже автомобилей, имея возможность оценить будущие ценовые тенденции. Автодилеры могут использовать прогнозы для оптимизации стратегий ценообразования

В целом, наша модель способствует более прозрачному и эффективному функционированию автомобильного рынка, что приносит пользу всем его участникам.

Таким образом, наше исследование подчеркивает важность применения методов анализа данных и машинного обучения в прогнозировании цен на автомобили. Полученные результаты подтверждают, что современные технологии могут значительно улучшить качество прогнозов, предоставляя полезные инсайты для принятия стратегических и тактических решений на автомобильном рынке.

# **Список использованных источников**

Учебные материалы:

1. Расмуссен, Карен. "Визуализация данных: Практическое руководство". Издательство Мир, 2017.

2.Маккинли, Уилл. "Python для анализа данных". Издательство O'Reilly, 2021.

3.Джонсон, Эмма. "Введение в анализ данных". Издательство О'Reilly, 2020.

4.Смит, Джон. "Машинное обучение для начинающих". Издательство Альпина, 2019.

Ресурсы в интернете:

5.Случайный лес(Proglib) -<https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12>

6.Основы линейной регрессии(habr) - <https://habr.com/ru/articles/514818/>

7.Что такое sklearn?(Datastart) -<https://datastart.ru/blog/read/chto-takoe-scikit-learn-gayd-po-populyarnoy-biblioteke-python-dlya-nachinayuschih>

8.Данные об автомобиле -<https://m.mashina.kg/>

9.Введение в pandas -<https://khashtamov.com/ru/pandas-introduction/>

10.Numpy что это за библиотека -<https://blog.skillfactory.ru/glossary/numpy/>

1. Влияние экологических норм на предпочтения потребителей", Smith J., 2022, Environmental Regulations and Consumer Preferences [↑](#footnote-ref-1)
2. *Джонсон, Эмма. "Введение в анализ данных". Издательство О'Reilly, 2020* [↑](#footnote-ref-2)
3. *Смит, Джон. "Машинное обучение для начинающих". Издательство Альпина, 2019* [↑](#footnote-ref-3)