МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева»

(Самарский университет)

Институт информатики и кибернетики Кафедра программных систем

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**«ВЕБ-ПРИЛОЖЕНИЕ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ ЛЕГКОВОГО АВТОМОБИЛЯ»**

по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Обучающийся Д.О. Колбанов

(подпись, дата)

Руководитель ВКР

доцент кафедры программных систем,

к.т.н., доцент О.А. Гордеева

(подпись, дата)

Нормоконтролер О.А. Гордеева

(подпись, дата)

Самара 2025

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования

«Самарский национальный исследовательский университет имени академика С.П. Королева»

Кафедра программных систем

УТВЕРЖДАЮ

Заведующий кафедрой

С.В. Востокин

« » 2024 г.

**задание на выпускную квалификационную работу (ВКР)**

обучающемуся Колбанову Дмитрию Олеговичу группы 6401-020302D

1. Тема работы: Веб-приложение прогнозирования стоимости легкового автомобиля Исходные данные прогнозирование стоимости автомобиля, регрессионный анализ, машинное обучение, регрессия, градиентный бустинг
2. Перечень вопросов, подлежащих разработке в ВКР:
   1. Провести анализ предметной области
   2. Сделать обзор систем-аналогов в области прогнозирования стоимости автомобилей
   3. Собрать данные для обучения моделей машинного обучения
   4. Провести сравнительный анализ примененных алгоритмов прогнозирования стоимости легкового автомобиля
   5. Разработать и реализовать информационное и программное обеспечение
   6. Провести тестирование и отладку разработанного веб-приложения
3. Дата выдачи задания: «29» января 2024г.
4. Срок представления на кафедру законченной ВКР: «01» июня 2024г.

Руководитель ВКР

доцент кафедры программных систем**,**

к.т.н., доцент О.А. Гордеева

« 29 » 01 2024 г.

Задание принял к исполнению Д.О. Колбанов

« 29 » 01 2024 г.

РЕФЕРАТ

Пояснительная записка с, рисунков, таблиц, источников,

приложения.

Графическая часть: слайда презентации PowerPoint.

ПРОГНОЗИРОВАНИЕ СТОИМОСТИ АВТОМОБИЛЯ, РЕГРЕССИОННЫЙ АНАЛИЗ, МАШИННОЕ ОБУЧЕНИЕ, РЕГРЕССИЯ, ГРАДИЕНТНЫЙ БУСТИНГ.

Цель работы – разработать веб-приложение для прогнозирования стоимости легкового автомобиля с использованием нескольких алгоритмов машинного обучения и провести сравнительный анализ их эффективности.

В процессе работы были разработаны алгоритмы и соответствующая программа, позволяющая пользователю получить прогноз рыночной стоимости легкового автомобиля на основе входных параметров. Система выполняет предобработку данных, рассчитывает стоимость, предоставляет результаты в текстовом виде.

Клиентская часть системы разработана на языке JavaScript с использованием библиотеки Vue.js. Серверная часть системы разработана на языке Python с использованием фреймворка Flask. Доступ к данным осуществляется с помощью СУБД SQLite.

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 6](#_Toc197617151)

[1 Описание и анализ предметной области 7](#_Toc197617152)

[1.1 Основные понятия и определения 7](#_Toc197617153)

[1.2 Актуальность задачи 8](#_Toc197617154)

[1.3 Описание систем-аналогов 8](#_Toc197617155)

[1.3.1 Сервис Avito Оценка 9](#_Toc197617156)

[1.3.2 Сервис Auto.ru «Оценка» 10](#_Toc197617157)

[1.3.3 Конкурентный анализ систем-аналогов 11](#_Toc197617158)

[1.4 Описание автоматизируемого процесса 13](#_Toc197617159)

[1.5 Постановка задачи 21](#_Toc197617160)

[1.6 Выводы по главе. 21](#_Toc197617161)

[2 Проектирование системы 22](#_Toc197617162)

[2.1 Описание средств реализации 22](#_Toc197617163)

[2.1.1 Описание операционной системы 22](#_Toc197617164)

[2.1.2 Описание языка программирования 22](#_Toc197617165)

[2.1.3 Описание среды разработки 23](#_Toc197617166)

[2.1.4 Описание используемой библиотеки 24](#_Toc197617167)

[2.2 Описание метода прогнозирования 25](#_Toc197617168)

[2.2.1 Линейная регрессия 25](#_Toc197617169)

[2.2.2 Градиентный спуск 25](#_Toc197617170)

[2.2.3 Регуляризация линейной регрессии 26](#_Toc197617171)

[2.2.4 Метод ближайших соседей 27](#_Toc197617172)

[2.2.5 Случайный лес 28](#_Toc197617173)

[2.2.6 Градиентный бустинг 28](#_Toc197617174)

[2.2.7 Принцип построения ансамбля 29](#_Toc197617175)

[2.2.8 Реализация градиентного бустинга в CatBoost 30](#_Toc197617176)

[2.2.9 Метрика качества 31](#_Toc197617177)

[2.3 Проект системы 32](#_Toc197617178)

[2.3.1 Структурная схема системы 32](#_Toc197617179)

[2.3.2 Диаграмма вариантов использования 33](#_Toc197617180)

[2.4 Выводы по главе 34](#_Toc197617181)

[3 Реализация системы 35](#_Toc197617182)

[3.1 Подготовка данных 35](#_Toc197617183)

[3.2 Описание проведенных исследований 36](#_Toc197617184)

[3.2.1 Предмет исследования 36](#_Toc197617185)

[3.2.2 Результаты исследования 36](#_Toc197617186)

[3.3 Описание экранных форм разработанного программного приложения 38](#_Toc197617187)

[3.4 Выводы по главе 40](#_Toc197617188)

ВВЕДЕНИЕ

В настоящее время рынок подержанных и новых легковых автомобилей характеризуется высокой динамичностью и большим разнообразием моделей, что усложняет процесс объективной оценки их стоимости [1]. Одновременно с этим широкое распространение современных веб‑технологий и методов машинного обучения открывает новые возможности для автоматизации процесса прогнозирования цен, повышая прозрачность рынка и помогая как продавцам, так и покупателям принимать более обоснованные решения.

Прогнозирование стоимости автомобиля представляет собой задачу регрессионного анализа, где в качестве входных признаков используются технические и эксплуатационные характеристики (марка, модель, год выпуска, пробег, тип двигателя и др.). Традиционные методы ручной оценки требуют глубоких экспертных знаний и существенно зависят от субъективного опыта оценщика. Автоматизированная система на основе машинного обучения позволяет построить модель, учитывающую многомерность данных и выявляющую сложные взаимосвязи между признаками, что повышает точность прогноза и скорость обработки запросов.

Практическая значимость работы заключается в создании доступного и наглядного инструмента, который может быть использован автодилерами, сервисами частных объявлений и конечными пользователями для быстрой оценки рыночной стоимости автомобиля. Полученные результаты могут быть внедрены в информационно‑аналитические системы автокомпаний и онлайн‑площадок для повышения конкурентоспособности и доверия клиентов.

В процессе работы необходимо разработать алгоритмы и соответствующую программу, позволяющую пользователю получить прогноз рыночной стоимости легкового автомобиля на основе входных параметров. Система должна выполнять предобработку данных, рассчитывать стоимость, предоставлять результаты в текстовом виде.

1. Описание и анализ предметной области
   1. Основные понятия и определения

Прогнозирование рыночной стоимости автомобиля — процесс определения будущей или текущей стоимости транспортного средства на основе анализа исторических данных, рыночных тенденций и технических характеристик [2]. Прогноз строится с использованием математических моделей и алгоритмов машинного обучения.

Регрессионный анализ — это статистический метод, позволяющий исследовать связь переменных. Основная цель регрессионного анализа — предсказать значение одной переменной на основе одной или нескольких других переменных. При этом регрессионный анализ помогает выявить, насколько тесно эти переменные связаны и как одна из них влияет на другую. [3]. В задачах прогнозирования стоимости автомобилей регрессионный анализ позволяет установить связь между характеристиками автомобиля (например, пробег, год выпуска, марка) и его ценой.

Машинное обучение — это область знаний и исследований в области искусственного интеллекта, которая занимается разработкой алгоритмов и статистических моделей, которые могут аппроксимировать данные, обучаться на них, обобщать их на невидимые зависимости и, таким образом, выполнять задачи без явных инструкций [4].

Модель машинного обучения — это объект (хранящийся локально в файле), который был обучен для распознавания определенных типов шаблонов. Модель обучается на основе набора данных по предоставленному ей алгоритму, который она может использовать для анализа и обучения на основе этих данных [5].

Регрессия в машинном обучении относится к задаче обучения с учителем, при котором целью является прогнозирование непрерывного числового значения на основе одной или нескольких независимых характеристик [6]. Прогнозирование стоимости автомобиля является задачей регрессии.

* 1. Актуальность задачи

Прогнозирование рыночной стоимости автомобиля является одной из ключевых задач в современной автомобильной индустрии и секторах, связанных с продажей, страхованием и кредитованием автотранспорта. Рынок подержанных автомобилей отличается высокой динамичностью: на цену влияет широкий спектр факторов — возраст и пробег машины, марка и модель, географическое расположение, сезонность, техническое состояние, наличие ДТП, а также колебания спроса и предложения. В условиях растущей конкуренции и всё более требовательных потребителей своевременная и точная оценка стоимости позволяет продавцам и покупателям принимать обоснованные решения, минимизировать финансовые риски и повышать прозрачность сделок на вторичном рынке.

В России покупка автомобиля на вторичном рынке приобрела особую значимость из-за политической и экономической нестабильности, санкций и ограниченного предложения новых моделей у официальных дилеров. Недостаток выбора на первичном рынке вынуждает потребителей обращаться к б/у автомобилям, что делает задачу автоматизированного и объективного прогнозирования цен ещё более востребованной.

Существуют различные подходы к оценке стоимости автомобилей: от экспертных систем и аналитики исторических объявлений до машинного обучения и нейронных сетей, обрабатывающих большие объёмы данных. Однако многие существующие решения либо требуют финансовых вложений, либо ограничены в удобстве использования. В связи с этим разработка веб-приложения, объединяющего в себе современные модели прогнозирования, автоматизированный сбор и предобработку данных, а также удобный интерфейс для различных категорий пользователей, является актуальной и востребованной задачей как в научно-исследовательской, так и в практической плоскости.

* 1. Описание систем-аналогов

В настоящее время существуют различные сервисы для прогнозирования стоимости автомобиля, предоставляющие различный функционал. Каждая система имеет свои достоинства и недостатки. Рассмотрим некоторые из них.

* + 1. Сервис Avito Оценка

Сервис Avito Оценка [7] предоставляет пользователю встроенный в площадку Avito онлайн-инструмент для быстрой и бесплатной оценки рыночной стоимости автомобиля. При переходе в карточку любого объявления система автоматически анализирует параметры конкретного лота — марку, модель, год выпуска, пробег, регион и комплектующие опции — и сравнивает их с тысячами совершённых сделок на той же площадке. Благодаря интеграции непосредственно в интерфейс Avito, пользователь получает рекомендации по справедливой цене без необходимости вручную вводить все данные и переключаться на сторонние сайты.

Результат оценки выводится в виде цветовой метки и текста непосредственно в списке объявлений и на странице лота: пользователь видит, попадает ли цена в «зону справедливой стоимости» или нет, а при клике может посмотреть график средней и разброса цен по аналогам. Это позволяет сразу ориентироваться на выгодные предложения.

Результат прогнозирования стоимости автомобиля, полученный данным сервисом представлен на рисунке 1.

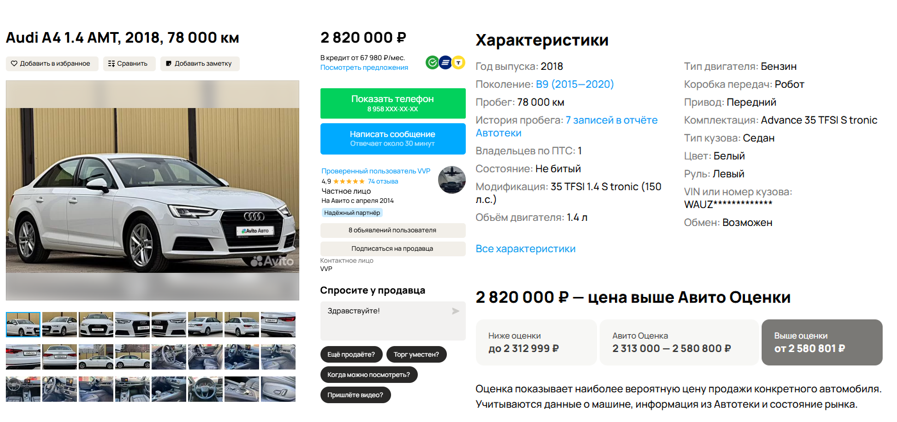


Рисунок 1 – Пример результата прогнозирования стоимости автомобиля сервисом Avito Оценка

* + 1. Сервис Auto.ru «Оценка»

Сервис Auto.ru «Оценка» [8] — это бесплатный веб-калькулятор, доступный на портале Auto.ru, предназначенный для моментальной оценки б/у автомобилей. Пользователь вводит либо госномер автомобиля, либо вручную основные параметры (марка, модель, год, пробег, тип кузова и комплектация), после чего система автоматически подбирает аналоги по базе объявлений и сделок площадки. Для удобства доступен быстрый ввод VIN-кода, что позволяет получить более точные данные по комплектации.

Результат работы выводится в виде конкретного диапазона рыночной цены и «справедливой стоимости» в процентах от среднего значения по аналогам. Пользователь может сформировать и скачать PDF-отчёт с подробной разбивкой: график динамики цен, таблицу сравнения аналогов и выводы экспертов Auto.ru.

Недостатком данного сервиса является платный доступ к расширенным функциям анализа.

На рисунке 2 представлен пример прогнозирования стоимости автомобиля, полученный данным сервисом.

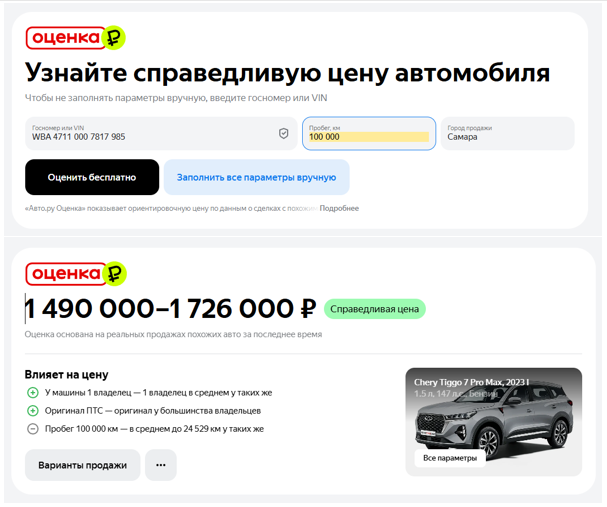


Рисунок 2 – Пример результата прогнозирования стоимости автомобиля сервисом Auto.ru «Оценка»

* + 1. Конкурентный анализ систем-аналогов

Проведем сравнительный анализ для вышеописанных сервисов прогнозирования стоимости автомобиля, выделив их основные отличительные особенности, см. таблицу 1.

Таблица 1 – Сравнительная характеристика систем-аналогов

|  |  |  |  |
| --- | --- | --- | --- |
| Показатель | Система | | Требования к  разрабатываемой системе |
| Avito Оценка | Auto.ru «Оценка» |
| Полностью бесплатное использование | + | − | + |
| Использование машинного обучения | − | + | + |
| Удобный интерфейс | + | + | + |
| Ввод характеристик автомобиля | − | + | + |
| Экспорт данных | − | + | − |

* 1. Описание автоматизируемого процесса

В рамках выпускной квалификационной работы необходимо разработать веб-приложение прогнозирования стоимости легкового автомобиля.

К основным этапам разработки веб-приложений с использованием алгоритмов машинного обучения относятся следующие: разработка модели прогнозирования, проектирование баз данных, разработка клиентской части, разработка серверной части.

Разработка модели прогнозирования включает в себя сбор, анализ и обработку данных для обучения, а также непосредственно обучение нескольких различных моделей машинного обучения, подбор гипер-параметров и выбор наилучшего алгоритма.

Разработка клиентской части приложения включает в себя такие задачи как разработка интерфейса пользователя и его прототипов. Необходимо учитывать, что интерфейс пользователя является важной частью системы, так как именно через работу с интерфейсом пользователь может взаимодействовать с системой.

Разработка серверной части приложения включает в себя следующие подзадачи: реализация основной бизнес-логики системы, разработка способов взаимодействия сервера с клиентом и базой данных. В рамках автоматизируемого процесса прогнозирования стоимости легкового автомобиля, необходимо реализовать предварительную обработку данных, а также расчет прогноза стоимости.

Для решения задачи прогнозирования стоимости легкового автомобиля и последующего исследования были выбраны линейная регрессия, k-ближайших соседей, случайный лес и градиентный бустинг. Этап исследования включает в себя вычисление метрик, отражающих точность прогнозирования стоимости – MAE, R2, MAPE.

* 1. Постановка задачи

Цель работы: во время выпускной квалификационной работы необходимо разработать веб-приложение прогнозирования стоимости легкового автомобиля.

Задачи:

* провести анализ предметной области;
* cделать обзор систем-аналогов в области прогнозирования стоимости автомобилей;
* собрать данные для обучения моделей машинного обучения;
* провести сравнительный анализ примененных алгоритмов прогнозирования стоимости легкового автомобиля;
* разработать и реализовать информационное и программное обеспечение;
* провести тестирование и отладку разработанного веб-приложения.

Разрабатываемое веб-приложение должно выполнять следующие функции:

1. серверная часть системы;
   * прогнозирование стоимости автомобиля;
2. клиентская часть системы;
   * ввод характеристик автомобиля;
   * просмотр результатов прогноза.
   1. Выводы по главе.

В данной главе был произведен анализ предметной области: изучены и описаны основные определения в области прогнозирования стоимости легкового автомобиля, актуальность исследования.

Произведен обзор существующих систем-аналогов, сформулирована постановка задачи ВКР и определены основные функции разрабатываемой системы.

1. Проектирование системы
   1. Описание средств реализации

В настоящее время существует огромное количество программных продуктов, позволяющих в эффективно и качественно разработать программный комплекс для различных предметных областей. Для правильного и обоснованного выбора во внимание принимались различные критерии для оценки качества программного продукта.

* + 1. Описание операционной системы

В качестве операционной системы (ОС) для клиентской части выбрана Windows 10 – операционная система, разработанная компанией Microsoft в 2015 году. Имеет следующие преимущества:

* обладает удобным интерфейсом для облегчения установки и поддержки любого программного обеспечения;
* справляется с перепадами напряжения в сети и обеспечивает отказоустойчивость;
* практически любое программное обеспечение выпускается помимо других ОС для ОС Windows 10.
  + 1. Описание языка программирования

Для реализации клиентской части системы выбран язык программирования JavaScript с использованием фреймворка Vue.js.

JavaScript – это интерпретируемый язык программирования, который широко используется для создания динамических веб-сайтов и веб-приложений. Он обеспечивает взаимодействие пользователя с содержимым страницы, обновление данных без перезагрузки страницы и дополнительные функциональные возможности.

Vue.js — JavaScript фреймворк для создания пользовательских интерфейсов. Он создан на стандартах HTML, CSS и JavaScript и предоставляет декларативную и компонентную модель программирования, которая помогает эффективно разрабатывать пользовательские интерфейсы любой сложности

JavaScript был выбран для реализации фронтенда проекта из-за его широкого использования в веб-разработке и гибкости. Vue.js, был выбран из-за своей простоты, производительности и поддержки различных функциональных возможностей для веб-приложений.

Для реализации серверной части системы был выбран язык программирования Python и фреймворк Flask. Python — мультипарадигмальный высокоуровневый язык программирования общего назначения с динамической строгой типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости кода и его качества, а также на обеспечение переносимости написанных на нём программ. Язык является полностью объектно-ориентированным. Синтаксис ядра языка минималистичен, за счёт чего на практике редко возникает необходимость обращаться к документации. Сам же язык известен как интерпретируемый и используется в том числе для написания скриптов. Также Python имеет большое количество библиотек и фреймворков для анализа данных и машинного обучения.

* + 1. Описание среды разработки

В качестве среды программирования была выбрана Visual Studio Code. Visual Studio Code (VS Code) — это редактор исходного кода. Его разработал Microsoft для всех популярных операционных систем: Windows, Linux и macOS. Визуальный редактор кода позволяет:

* работать с IntelliSense — автоматическим дописыванием функций при вводе первых букв;
* выполнять отладку — искать и устранять ошибки в написанном коде;
* удобно писать код — автоматически заполнять нужную информацию, подсвечивать элементы синтаксиса в зависимости от выбранного языка, расставлять нужные отступы;
* контролировать версии кода, в том числе с помощью системы управления версиями Git;
* рефакторить код для улучшения его работы и читабельности.
  + 1. Описание используемой библиотеки

Для написания парсера, собирающего об автомобилях, использовалась библиотека Selenium. Selenium WebDriver – это программная библиотека для управления браузерами. WebDriver представляет собой драйверы для различных браузеров и клиентские библиотеки на разных языках программирования, предназначенные для управления этими драйверами.

По сути своей использование такого веб-драйвера сводится к созданию бота, выполняющего всю ручную работу с браузером автоматизировано.

Библиотеки WebDriver доступны на языках Java, .Net (C#), Python, Ruby, JavaScript, драйверы реализованы для браузеров Firefox, InternetExplorer, Safari, а также Chrome и Opera.

Для обработки данных использовалась библиотека pandas. Pandas — программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временны́ми рядами. Pandas имеет хорошие показатели скорости за счет оптимизации кода, интуитивно понятный интерфейс, интеграцию с другими Python-библиотеками.

Для реализации метода прогнозирования использовались библиотеки Scikit-learn и CatBoost.

Scikit-learn — библиотека, предназначенная для машинного обучения, написанная на языке программирования Python и распространяемая в виде свободного программного обеспечения. В её состав входят различные алгоритмы, в том числе предназначенные для задач классификации, регрессионного и кластерного анализа данных, включая метод опорных векторов, метод случайного леса, алгоритм усиления градиента, метод k-средних и DBSCAN. Библиотека была разработана для взаимодействия с численными и научными библиотеками языка программирования Python NumPy и SciPy.

CatBoost — открытая программная библиотека, разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга. CatBoost обладает рядом преимуществ, таких как автоматическая обработка категориальных признаков, встроенная регуляризация, автоматический отбор признаков и высокая производительность.

* 1. Описание метода прогнозирования

Существует большое количество методов машинного обучения, решающих задачу регрессии. Были взяты несколько различных моделей, обучены на подготовленных данных и оценены по метрике MAPE (Mean Absolute Percentage Error). Ниже рассмотрим каждый алгоритм:

* + 1. Линейная регрессия

Линейная регрессия (Linear regression) — один из простейший алгоритмов машинного обучения, описывающий зависимость целевой переменной от признака в виде линейной функции y = kx + b [9]. В данном случае была представлена простая или парная линейная регрессия, а уравнение вида

называется множественной линейной регрессией, где *b* — смещение модели, *w* — вектор её весов, а *x* — вектор признаков одного обучающего образца.

Выбор регрессионной линии (плоскости), описывающей взаимосвязь данных наилучшим образом, заключается в минимизации функции потерь, представленной в виде среднеквадратичной ошибки. Проще говоря, линия должна проходить через данные таким образом, чтобы в среднем разница квадратов ожидаемых и реальных значений была минимальна. Данный метод называется методом наименьших квадратов.

Линейная регрессия обучается с помощью итеративной оптимизации с постепенным снижением ошибки модели на основе градиентного спуска и его разновидностей.

* + 1. Градиентный спуск

Градиентный спуск — это итеративный алгоритм оптимизации, используемый для минимизации функции, чаще всего функции потерь в контексте машинного обучения. Он работает путем нахождения направления, в котором функция потерь уменьшается наиболее быстро, и делает шаги в этом направлении для постепенного уменьшения значения функции потерь [10].

Градиент функции — это вектор, состоящий из частных производных, который указывает направление наискорейшего роста функции. В контексте оптимизации, мы интересуемся направлением наискорейшего убывания, то есть движемся в противоположном направлении градиента. Частные производные вычисляются для каждого параметра модели.

На каждой итерации алгоритма параметры модели обновляются в направлении, противоположном градиенту функции потерь. Размер шага, который делает алгоритм в этом направлении, определяется скоростью обучения (learning rate). Оптимальная скорость обучения — ключевой параметр, поскольку слишком большой шаг может привести к пропуску минимума, а слишком маленький делает процесс оптимизации медленным.

Формула для обновления параметра θ на каждой итерации выглядит следующим образом:

где *η* - скорость обучения, а *∇θ⋅J(θ)* - градиент функции потерь *J* по параметру *θ*.

* + 1. Регуляризация линейной регрессии

Регуляризация в статистике, машинном обучении, теории обратных задач — метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Чаще всего эта информация имеет вид штрафа за сложность модели [11].

В линейной регрессии методы регуляризации работают путем добавления штрафных коэффициентов к исходной функции потерь модели таким образом, что высокие значения коэффициентов снижаются. А признаки с очень низкими значениями коэффициентов (после штрафования) могут быть вообще отброшены. Это помогает уменьшать сложность модели. Рассмотрим Ridge и Lasso регуляризации.

Гребневая регрессия (Ridge regression) или регуляризация Тихонова применяется в случае мультиколлинеарности через добавление L2-регуляризации к функции потерь во время обучения и сильнее всего занижает веса для признаков с высокой корреляцией: их значения будут приближаться к нулю, но никогда его не достигнут.

Лассо-регрессия (Lasso regression или Least Absolute Shrinkage & Selection Operator) обычно используется для отбора признаков через добавление L1-регуляризации к функции потерь во время обучения. Проще говоря, лассо-регрессия стремится уменьшить число параметров модели путем зануления весов для неинформативных и избыточных признаков, что на выходе даст разреженную модель (с небольшим числом ненулевых весов признаков).

где *λ* (лямбда) — это гиперпараметр, который контролирует силу регуляризации. Он выбирается заранее и может быть настроен в процессе обучения модели, *n* - количество признаков в модели, *wi*​ - вес (коэффициент) i-го признака.

* + 1. Метод ближайших соседей

Метод ближайших соседей (KNN) в задаче регрессии — это непараметрический метод, используемый для прогнозирования непрерывных значений [12].

Основная идея заключается в прогнозировании целевого значения для новой точки данных путём усреднения целевых значений K ближайших соседей в пространстве объектов. Расстояние между точками данных обычно измеряется с использованием евклидова расстояния, хотя могут использоваться и другие показатели расстояния.

Работа KNN-регрессии включает несколько шагов:

* выбор количества соседей (K). Этот выбор сильно влияет на производительность модели. Меньшее значение K делает модель более подверженной шуму, в то время как большее значение K приводит к более плавным прогнозам;
* вычисление расстояний. Для новой точки данных вычисляют расстояние между этой точкой и всеми точками в обучающем наборе;
* поиск K ближайших соседей. Определяют K точек в обучающем наборе, которые находятся ближе всего к новой точке данных;
* прогнозирование целевого значения. Вычисляют среднее значение целевых значений K ближайших соседей и используют это в качестве прогнозируемого значения для новой точки данных.
  + 1. Случайный лес

Решающие деревья являются хорошим семейством базовых классификаторов для бэггинга, поскольку они достаточно сложны и могут достигать нулевой ошибки на любой выборке. Метод случайных подпространств позволяет снизить коррелированность между деревьями и избежать переобучения. Базовые алгоритмы обучаются на различных подмножествах признакового описания, которые также выделяются случайным образом [13].

Алгоритм построения случайного леса, состоящего из N деревьев, выглядит следующим образом: для каждого n = 1, …, N:

1. сгенерировать выборку Xn с помощью бутстрэпа;
2. построить решающее дерево bn по выборке Xn:

* по заданному критерию мы выбираем лучший признак, делаем разбиение в дереве по нему и так до исчерпания выборки;
* дерево строится, пока в каждом листе не более nmin объектов или пока не достигнем определенной высоты дерева;
* при каждом разбиении сначала выбирается m случайных признаков из n исходных, и оптимальное разделение выборки ищется только среди них.

Итоговая модель выглядит следующим образом:

* + 1. Градиентный бустинг

Бустинг – это ансамблевый метод машинного обучения, целью которого является объединение нескольких слабых моделей предсказания для создания одной сильной. Слабая модель – это такая, которая выполняет предсказания немного лучше, чем наугад, в то время как сильная модель обладает высокой предсказательной способностью. Цель бустинга – улучшить точность предсказаний [14].

Бустинг работает путём последовательного добавления моделей в ансамбль. Каждая следующая модель строится таким образом, чтобы исправлять ошибки, сделанные предыдущими моделями. Это достигается путём фокусировки на наиболее проблемных данных, которые были неверно классифицированы или предсказаны ранее.

Одной из основных особенностей бустинга является динамическое взвешивание обучающих данных. После каждого этапа обучения модели в ансамбле, данные, на которых были допущены ошибки, получают больший вес. Это означает, что последующие модели уделяют больше внимания именно этим трудным случаям.

Когда используются решающие деревья, каждое последующее дерево строится с учетом ошибок, сделанных предыдущими деревьями. Новые деревья учатся на ошибках, улучшая общую точность ансамбля.

Несмотря на свою эффективность, бустинг может быть склонен к переобучению, особенно если в ансамбле слишком много моделей или они слишком сложные. Для контроля переобучения используется ранняя остановка (early stopping).

* + 1. Принцип построения ансамбля

Ансамбль в градиентном бустинге обычно состоит из последовательности слабых предсказательных моделей. Чаще всего используются решающие деревья из-за их способности моделировать нелинейные зависимости и взаимодействия между признаками. Каждое новое дерево в ансамбле строится так, чтобы уменьшить оставшуюся ошибку предыдущих деревьев [15].

В градиентном бустинге каждая следующая модель обучается с учетом ошибок, допущенных всеми предыдущими моделями в ансамбле. Это достигается путем фокусировки на самых трудных для предсказания случаях, которые были неправильно классифицированы или предсказаны ранее.

Суть метода заключается в том, что веса для каждого наблюдения в обучающем наборе данных корректируются на каждом шаге. Наблюдения, которые были неправильно предсказаны предыдущей моделью, получают больший вес, тем самым увеличивая вероятность их правильного предсказания последующими моделями.

* + 1. Реализация градиентного бустинга в CatBoost

CatBoost использует решающие деревья глубины 1 или 2 в качестве базовых моделей [16]. Эти неглубокие деревья имеют следующие характеристики:

* каждый узел дерева делает бинарное разбиение на основе значения одной из признаков;
* эти короткие деревья обладают небольшой глубиной, что делает их более устойчивыми к переобучению.

CatBoost включает механизм регуляризации, чтобы предотвратить переобучение модели. Он использует L2-регуляризацию, представляет собой метод добавления штрафа к функции потерь модели с целью предотвратить переобучение. Этот метод помогает улучшить обобщающую способность модели и снизить риск переобучения, особенно в случаях, когда много признаков или они коррелированы между собой.

CatBoost автоматически выполняет отбор признаков путем оценки их важности для модели. Это позволяет модели сосредотачиваться на наиболее информативных признаках и уменьшить шум от менее значимых.

CatBoost применяет градиентный бустинг для обучения ансамбля решающих деревьев. Градиентный бустинг минимизирует функцию потерь с использованием градиентного спуска, постепенно улучшая качество модели.

На каждой итерации градиентного бустинга добавляется новое решающее дерево, которое исправляет ошибки предыдущих деревьев.

* + 1. Метрика качества

MAPE выражает среднее абсолютное отклонение прогнозируемых значений от фактических значений в процентах, что делает эту метрику очень наглядной для интерпретации результатов [17].

Формула MAPE определяется как:

где:

*N* – количество наблюдений,

*yi* – фактическое значение целевой переменной,

*f(xi)* – прогнозируемое значение, полученное моделью,

*|yi−f(xi)|* – абсолютная ошибка прогноза для (i)-го наблюдения,

*|yi|* – абсолютное значение фактического значения целевой переменной для нормализации ошибки.

В результате оценки моделей на тестовых данных – случайной выборки из 3400 автомобилей получились следующие значения:

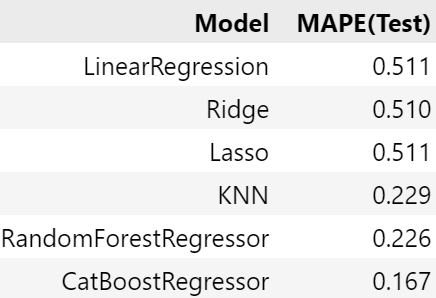


Рисунок 3 – Результаты оценки моделей

Наилучший результат показала модель CatBoostRegressor, которая представляет собой реализацию градиентного бустинга над решающими деревьями библиотеки CatBoost. Для прогнозирования стоимости автомобиля будет использоваться эта модель.

* 1. Проект системы
     1. Структурная схема системы

Структурный подход при разработке системы подразумевает разбиение системы на функциональные подсистемы, тем самым система представляет из себя совокупность отдельных компонентов, взаимодействующих между собой [18]. Таким образом, сохраняется целостность представления о системе, в которой все выделенные подсистемы связаны.

Структурная схема системы представляет из себя диаграмму, отражающую взаимосвязь компонентов системы для понимания принципов работы разрабатываемой системы.

На рисунке 4 приведена структурная схема разрабатываемой системы, в ее состав входят клиентская и серверная части, которые взаимодействуют между собой с помощью протокола HTTP.

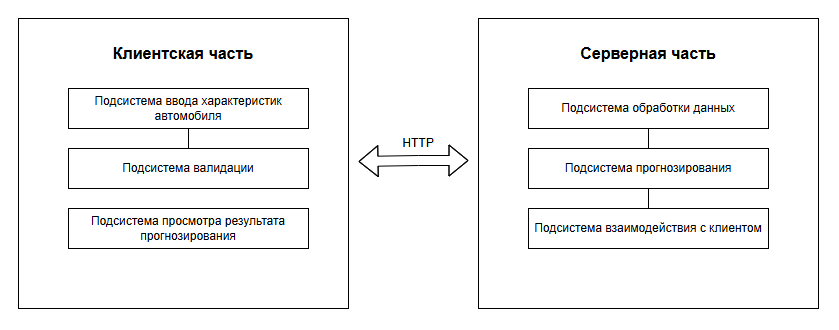
****

Рисунок 4 – Структурная схема системы

В состав клиентской части входят следующие подсистемы:

1. подсистема ввода характеристик автомобиля, позволяющая пользователю вводить данные об автомобиле;
2. подсистема валидации, отвечающая за проверку корректности введенных значений;
3. подсистема просмотра результатов прогнозирования, которая позволяет пользователю просмотреть результат работы модели.

В состав серверной части входят следующие подсистемы:

1. подсистема обработки данных, которая подготавливает данные, полученные от клиента, в формат, пригодный для работы модели;
2. подсистема прогнозирования, рассчитывающая стоимость автомобиля по введенным данным, используя модель машинного обучения;
3. подсистема взаимодействия с клиентом, которая отвечает за взаимодействие с клиентом.
   * 1. Диаграмма вариантов использования

Диаграмма вариантов использования – диаграмма, отражающая взаимодействие между актерами и вариантами использования, позволяющая описать концептуальную модель системы [19].

Данная диаграмма используется для описания функционала и поведения системы, не погружаясь в особенности реализации функционала.

В качестве актеров выступает некоторое множество ролей, взаимодействующих с системой различными вариантами использования.

Вариант использования представляет из себя спецификацию последовательности действий, которые система выполняет при взаимодействии с актерами, таким образом выделяется набор действий, которые совершает систем, взаимодействуя с актером.

На рисунке 5 приведена диаграмма вариантов использования для пользователя в разрабатываемой системе.

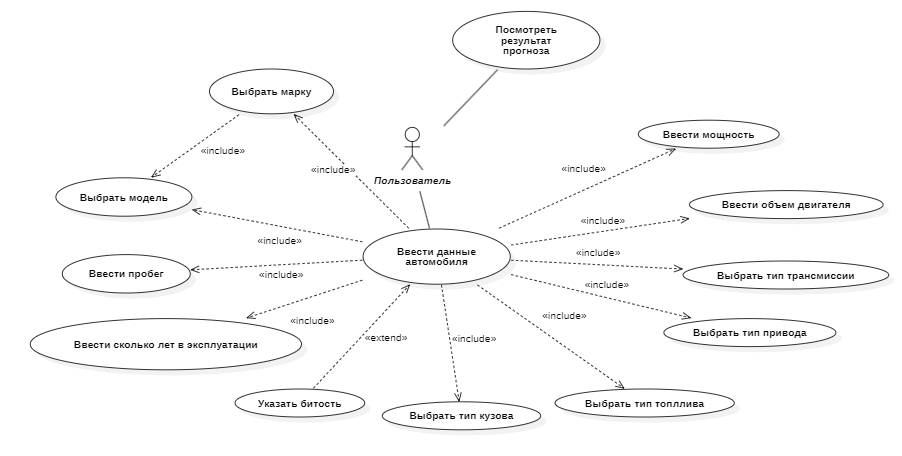
****

Рисунок 5 – Диаграмма вариантов использования

* 1. Выводы по главе

В данной главе было выполнено проектирование разрабатываемого веб-приложения прогнозирования стоимости автомобиля. Были исследованы и протестированы различные методы прогнозирования и выбран наилучший по представленным метрикам. Был выбран и описан комплекс программных средств для реализации.

1. Реализация системы
   1. Подготовка данных

Сбор данных является ключевым этапом разработки модели машинного обучения, так как качество прогнозов напрямую зависит от объема и репрезентативности данных. На основе собранных характеристик автомобилей модель обучается выявлять закономерности между входными параметрами и стоимостью, что позволяет ей делать точные предсказания для новых данных.

Для обучения модели прогнозирования стоимости автомобилей использовались данные, собранные с платформы Avito — одного из крупнейших российских сайтов объявлений о продаже автомобилей.

С помощью библиотеки Selenium был разработан парсер, который автоматически собирает информацию из объявлений о продаже автомобилей.

В результате парсинга были собраны данные о 19994 автомобилях в Самарской области.

Из карточки каждого автомобиля были извлечены ключевые характеристики:

* бренд и модель;
* эксплуатационные характеристики (год выпуска, пробег, состояние);
* технические характеристики (мощность, объем двигателя и т.д.);
* цена – таргет.

Пример карточки приведен на рисунке 4.



Рисунок 4 – карточка объявления автомобиля

Процесс обработки данных включал следующие этапы:

* очистка данных;
* преобразование категориальных признаков;
* анализ значимости признаков;
* разбиение на обучающую и тестовую выборки.
  1. Описание проведенных исследований
     1. Предмет исследования

В рамках выпускной квалификационной работы предметом исследования является значимость признаков в прогнозировании стоимости автомобиля. Рассмотрим исследование признака is\_crashed («битый / не битый»), используя сравнительную статистику.

* + 1. Результаты исследования

Для каждого битого автомобиля из выборки подбирался аналогичный не битый автомобиль той же модели и года выпуска, с максимально близким пробегом (допустимое отклонение ±5%). Таким образом, сравнивались пары автомобилей, различающиеся только по признаку аварийности, при прочих равных характеристиках. На основании анализа 89 пар автомобилей была построена сводная таблица с основными статистическими характеристиками разницы в стоимости между битым и не битым автомобилем. Как видно из данных таблицы 2, средняя разница составила 270 328 ₽, при этом максимальная разница достигала 1 249 000 ₽, а медианная разница — 190 000 ₽. Это подтверждает, что битость оказывает существенное влияние на рыночную цену автомобилей. Также важно отметить, что различия в пробеге между сравниваемыми автомобилями были минимальными (в среднем около 3 500 км), что делает результаты сравнения более достоверными.

Таблица 2 – Статистические характеристики разницы в стоимости между битым и не битым автомобилем

|  |  |
| --- | --- |
| Показатель | Значение |
| Количество пар автомобилей | 89 |
| Средняя разница в цене | 270 328 ₽ |
| Минимальная разница в цене | 10 000 ₽ |
| Максимальная разница в цене | 1 249 000 ₽ |
| Медианная разница в цене | 190 000 ₽ |
| Средний пробег битого авто | 142 785 км |
| Средний пробег не битого авто | 139 212 км |

Как видно на рисунке 5, распределение разницы в стоимости между битым и не битым автомобилем характеризуется концентрацией значений в диапазоне от 150 000 ₽ до 300 000 ₽. Анализ плотности распределения показывает, что наибольшее количество пар автомобилей имеет разницу в цене, близкую к средней величине, которая составляет 270 328 ₽.

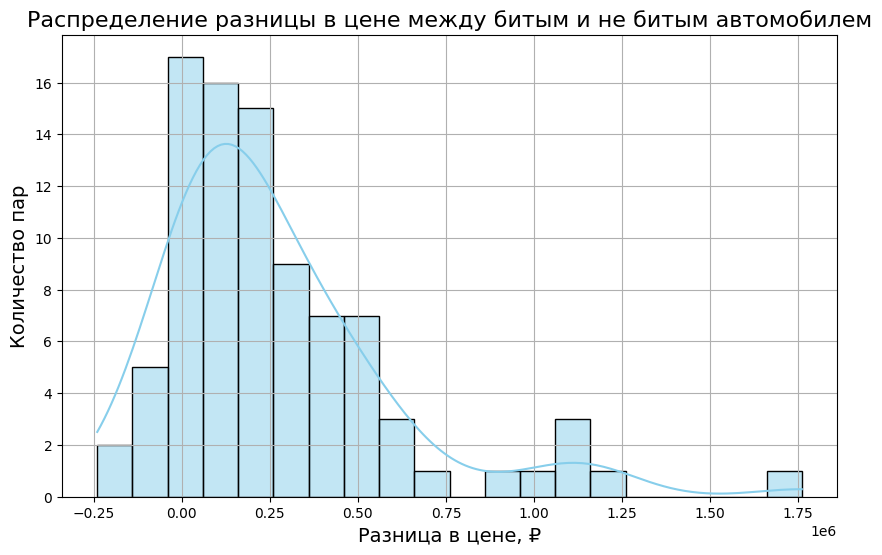


Рисунок 5 – график распределения разницы в ценах

Распределение имеет асимметричную форму с выраженным правосторонним хвостом, что свидетельствует о наличии автомобилей с экстремально высокой разницей в стоимости. Такие случаи, как правило, соответствуют дорогостоящим моделям автомобилей, для которых битость существенно влияет на потерю рыночной стоимости.

На основании полученных данных можно сделать вывод, что признак битости автомобиля является значимым фактором при прогнозировании его рыночной стоимости. Наличие повреждений приводит к заметному снижению стоимости транспортного средства, что необходимо учитывать при разработке моделей оценки цены автомобиля.

* 1. Описание экранных форм разработанного программного приложения

Пользовательский интерфейс – одна из разновидностей интерфейсов, который является совокупностью средств и методов взаимодействия пользователя с вычислительными устройствами (персональным компьютером).

Интерфейс характеризуется удобством, эффективностью, понятностью и дружественностью.

Дружественный интерфейс предоставляет пользователю наиболее удобный способ взаимодействия с программным обеспечением путем обеспечения логичности и простоты в расположении элементов управления.

Разработанная система представляет собой веб-приложение, которое может работать на компьютере под управлением любой операционной системы при наличии браузера.

При запуске приложения открывается страница с формой для ввода характеристик автомобиля.

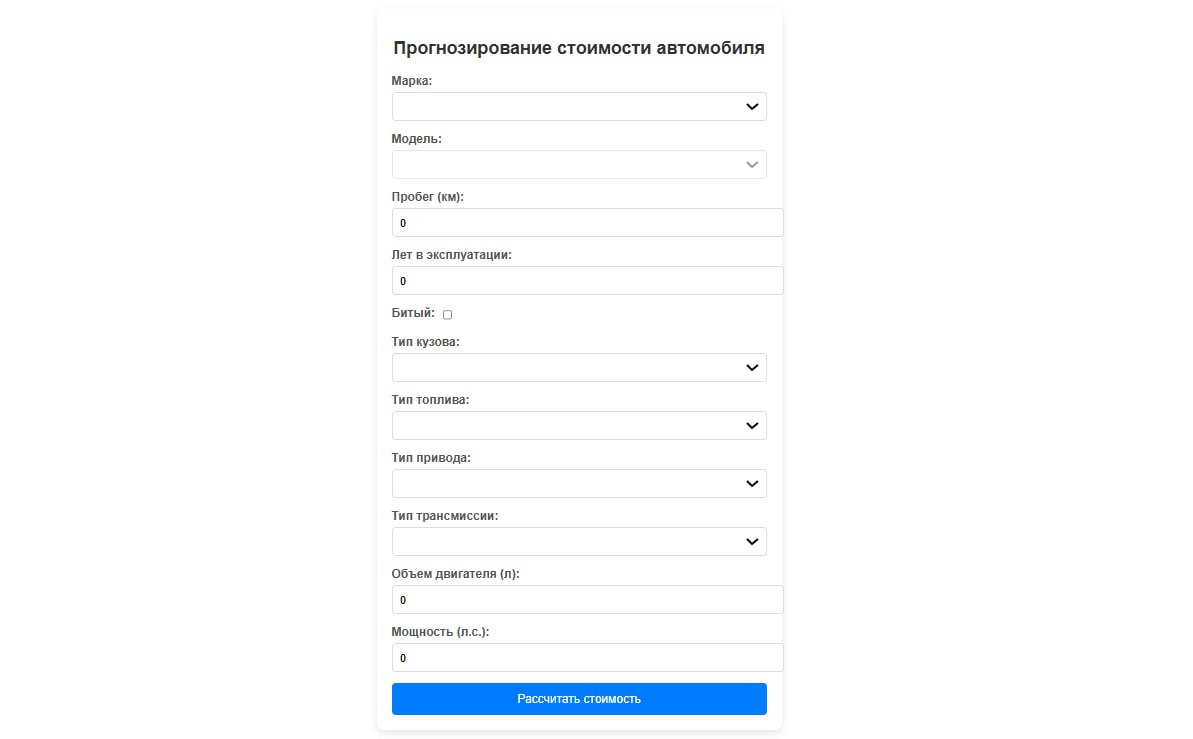


Рисунок 6 – форма ввода характеристик

Здесь пользователь сможет ввести параметры интересующего автомобиля, после чего нажать кнопку «Рассчитать стоимость» и получить прогноз.

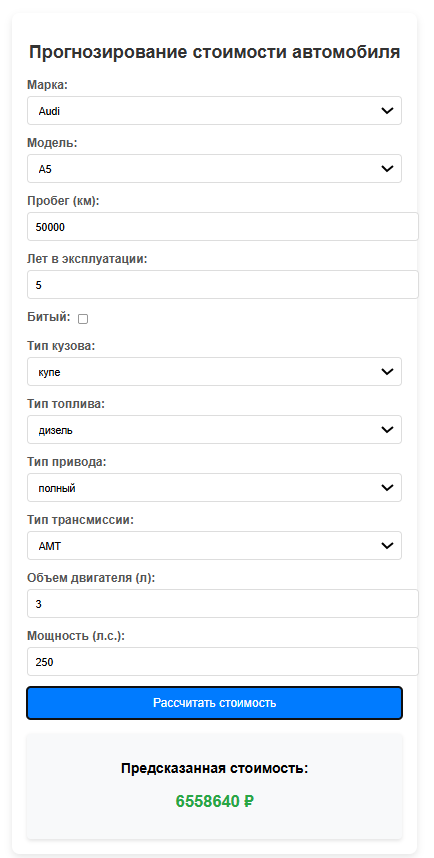


Рисунок 7 – результат прогноза

* 1. Выводы по главе

В данной главе был рассмотрен процесс подготовки данных для построения модели прогнозирования стоимости автомобиля. С использованием парсинга объявлений с платформы Avito был сформирован объемный и репрезентативный набор данных, содержащий ключевые технические и эксплуатационные характеристики автомобилей, а также информацию об их состоянии.

Также было проведено исследование значимости признаков, в частности бинарного признака состояния автомобиля "битый/не битый". Проведенный сравнительный анализ пар автомобилей с одинаковыми характеристиками (модель, год выпуска, пробег) показал, что наличие повреждений оказывает существенное влияние на рыночную стоимость автомобиля.

Таким образом, проведенное исследование позволило обоснованно утверждать, что признак битости является одним из наиболее важных факторов, влияющих на стоимость автомобиля, и должен обязательно учитываться при построении моделей машинного обучения для оценки рыночной цены транспортных средств.

ЗАКЛЮЧЕНИЕ

В процессе выполнения выпускной работы было разработано веб-приложение прогнозирования стоимости легкового автомобиля.

В первом разделе были приведены основные понятия и определения предметной области прогнозирования стоимости легкового автомобиля, актуальность исследования, приведены характеристики систем-аналогов, на основании этого была сформулирована постановка задачи и основные требования к системе.

Во втором разделе было выполнено проектирование разрабатываемого веб-приложения прогнозирования стоимости автомобиля. Были исследованы и протестированы различные методы прогнозирования и выбран наилучший по представленным метрикам, а также был выбран и обоснован комплекс программных средств.

В третьем разделе описан процесс сбора и подготовки данных, интерфейс пользователя, а также приведены результаты проведенных исследований, проведен анализ значимости признаков на прогнозирование стоимости автомобиля.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. В России бум продаж подержанных машин. [Электронный ресурс] // Газета.Ru – URL: https://www.gazeta.ru/auto/2025/02/20/20585066.shtml (дата обращения: 22.04.2025).
2. Обзор методов прогнозирования. [Электронный ресурс] // Ivan Shamaev – URL: https://ivan-shamaev.ru/overview-forecast-methods/ (дата обращения: 24.04.2025).
3. Что такое регрессионный анализ [Электронный ресурс] // Яндекс Практикум – URL: https://practicum.yandex.ru/blog/chto-takoe-regressionnyj-analiz/ (дата обращения: 22.04.2025).
4. Машинное обучение: общие принципы и концепции [Электронный ресурс] // Хабр: [сайт]. – URL: https://habr.com/ru/articles/862704/ (дата обращения: 22.04.2025).
5. Что такое модель машинного обучения? [Электронный ресурс]. // Microsoft Learn: [сайт]. – URL: https://learn.microsoft.com/ru-ru/windows/ai/windows-ml/what-is-a-machine-learning-model (дата обращения: 22.04.2025).
6. Regression in machine learning [Электронный ресурс]. // GeeksforGeeks: [сайт]. – URL: https://www.geeksforgeeks.org/regression-in-machine-learning/ (дата обращения: 22.04.2025).
7. Авито Оценка [Электронный ресурс]. – URL: https://www.avito.ru/evaluation/cars (дата обращения: 26.04.2025).
8. Auto.ru «Оценка» [Электронный ресурс]. – URL: https://miratext.ru/seo\_analiz\_text (дата обращения: 26.04.2025).
9. Линейная регрессия. Основная идея, модификации и реализация с нуля на Python [Электронный ресурс]. URL: https://habr.com/ru/articles/804135/ (дата обращения: 20.12.2024).
10. Алгоритм градиентного спуска в машинном обучении [Электронный ресурс]. URL: https://www.geeksforgeeks.org/gradient-descent-algorithm-and-its-variants/ (дата обращения: 20.12.2024).
11. Регуляризация – Викиконспекты [Электронный ресурс]. URL: https://neerc.ifmo.ru/wiki/index.php?title=Регуляризация (дата обращения: 20.12.2024).
12. Регрессия методом k-ближайших соседей (KNN) с помощью Scikit-Learn [Электронный ресурс]. URL: https://www.geeksforgeeks.org/k-nearest-neighbors-knn-regression-with-scikit-learn/ (дата обращения: 20.12.2024).
13. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес [Электронный ресурс]. URL: https://habr.com/ru/companies/ods/articles/324402/ (дата обращения: 20.12.2024).
14. Что такое бустинг? [Электронный ресурс]. URL: https://aws.amazon.com/ru/what-is/boosting/ (дата обращения: 21.12.2024).
15. Градиентный бустинг [Электронный ресурс]. URL: https://education.yandex.ru/handbook/ml/article/gradientnyj-busting (дата обращения: 21.12.2024).
16. CatBoost [Электронный ресурс]. URL: https://habr.com/ru/companies/otus/articles/778714/ (дата обращения: 21.12.2024).
17. Метрики в машинном обучении: понимание, применение и интерпретация [Электронный ресурс]. URL: https://shakhbanov.org/metriki-v-mashinnom-obuchenii/ (дата обращения: 21.12.2024).
18. Буч Г. Язык UML. Руководство пользователя: Пер. с англ. / Г. Буч, Д. Рамбо, Б. Джекобсон. М.: ДМК-Пресс, 2001. 432 с.
19. Коварцев А.Н. Автоматизация тестирования программного обеспечения учебное пособие. Самара: СГАУ, 2010. 122 с.

ПРИЛОЖЕНИЕ А

Руководство пользователя

ПРИЛОЖЕНИЕ Б

Код программы