

Оглавление

[ВВЕДЕНИЕ 2](#_Toc656862967)

[1 Теоретические сведения и используемые технологии 3](#_Toc1374139238)

[1.1 Python 3](#_Toc1198366121)

[1.1.1 Общая информация 3](#_Toc1899648672)

[1.1.2 Используемые библиотеки Python 4](#_Toc834472679)

[1.2 Google colab 4](#_Toc846114251)

[1.2.1 Общая информация 5](#_Toc1861781544)

[1.2.2 Архитектура 5](#_Toc615717806)

[2. Аналитическая глава 6](#_Toc1262744185)

[2.1 Анализ ансамблевых методов 6](#_Toc604218854)

[2.1.1 Виды ансамблевых методов 6](#_Toc1377623449)

[2.1.4 Применение ансамблевых методов в анализе цен на образовательные услуги 8](#_Toc2101437924)

[2.1.5 Заключение аналитической главы 8](#_Toc229070276)

[3. Практическая часть 9](#_Toc1021662376)

[3.1 Сбор и подготовка данных 9](#_Toc909992014)

[3.1.1 Источники данных 9](#_Toc1153208238)

[3.1.2 Методы предобработки 9](#_Toc2141783316)

[3.2 Визуализация обработанных данных 12](#_Toc1674147207)

[1.3 Создание моделей 14](#_Toc1075736932)

[3.3.1 Линейная регрессия 14](#_Toc809068264)

[3.3.2 Использование ансамблевых методов 16](#_Toc879821496)

[Заключение 18](#_Toc1595008200)

[Список литературы 19](#_Toc1322815080)

# ВВЕДЕНИЕ

В современном мире данные зачастую имеют решающее значение для принятия решений и стратегического планирования. Поскольку с каждым днем генерируется все больше и больше информации, возникает потребность в эффективных инструментах для анализа и обработки данных. Анализ данных стал горячей областью, охватывающей многие области, такие как бизнес, здравоохранение, финансы и наука.

Целью данной курсовой работы является анализ набора данных, описывающих высшие учебные заведения. Анализ включает в себя предварительную обработку данных, а также пре- и постанализ с визуализацией результатов. Для исследования выбирается набор данных, соответствующий теме и предлагающий возможность углубленного анализа.

Одним из ключевых этапов работы станет предварительная обработка данных, устранение пробелов и аномалий, нормализация данных и преобразование категориальных признаков в числовую форму. Эти шаги имеют решающее значение, поскольку качество входных данных напрямую влияет на результаты анализа, и даже небольшие ошибки на этом этапе могут привести к неверным выводам.

После обработки данных они анализируются с использованием различных методов. В рамках углубленного анализа планируется предоставить обзор набора данных с использованием описательной статистики. Методы визуализации, такие как графики, диаграммы и тепловые карты, также используются для облегчения восприятия и интерпретации данных. Визуализация помогает выявить скрытые закономерности, аномалии и интересные моменты, которые в противном случае могли бы остаться незамеченными, если бы использовались только числовые значения.

# 1 Теоретические сведения и используемые технологии

## Python

**1.1.1 Общая информация**

Python – высокоуровневый язык программирования, разрабатываемый с конца 1980-х годов Гвидо Ван Россумом, полноценный релиз состоялся в 1991 году. Данный язык программирования обладает рядом преимуществ над другими, что в последствии сделало его одним из самых популярных языков программирования.

Одной из основных характеристик Pythonа является условная простота составления и читаемость кода - этот язык программирования имеет простой и лаконичный синтаксис, что облегчает освоение и работу.

Python является достаточно многофункциональным языком, поддерживающим множество парадигм программирования, таких как объектно-ориентированное, императивное и функциональное программирование и поддерживает большое количество библиотек необходимых для выполнения задач, таких как вычисления, веб-программирования и работа с данными.

Язык программирования Python является кроссплатформенным, его можно запускать на разных операционных системах (windows, linux, macOS) и он будет работать.

Этот язык используется для анализа и обработки данных, позволяет реализовывать визуализацию и выполнять сложные вычисления, что является важным аспектом для выполнения данной работы.

**1.1.2 Используемые библиотеки Python**

Для возможности проведения анализа данных требуется выбрать среду разработки и набор всех необходимых библиотек для Python для правильного выполнения задач.

Библиотеки использованные в ходе выполнения работы:

1. **Pandas**

Библиотека для Python, предназначенная для анализа и обработки данных. Обладает возможностью загружать файлы в формате scv.

1. **Matplotlib**

Библиотека предназначенная для визуализации данных, позволяя строить графики, диаграммы, гистограммы.

1. **Sklearn**

Обширная библиотека, используемая для машинного обучения. Представляет спектр инструментов для разработки моделей, проведения анализа и прогнозирования.

1. **Numpy**

Библиотека для работы с массивами данных, которые используют модели.

1. Seaborn

Библиотека для дополнительной визуализации данных. Построена на основе matplotlib. Используется для построения парных графиков, тепловых карт.

# 1.2 Google colab

# 1.2.1 Общая информация

Для удобного и эффективного выполнения работы была выбрана платформа Google colab - это бесплатный облачный сервис от Google, который позволяет пользователям писать и выполнять Python-код в браузере. Он предоставляет удобную среду для разработки и анализа данных, особенно в области машинного обучения и научных вычислений.

**1.2.2 Архитектура**

Google colab обладает рядом возможностей для поддержания эффективного выполнения задач.

1. Облачная среда: Google Colab позволяет запускать код на удаленных серверах Google, что освобождает пользователей от необходимости устанавливать программное обеспечение и настраивать окружение на своих локальных машинах.
2. Поддержка Jupyter Notebooks: Colab основан на Jupyter Notebook, что позволяет пользователям создавать интерактивные документы, содержащие код, текст, изображения и визуализации.
3. Поддержка Python: Google Colab в первую очередь предназначен для работы с Python, что делает его идеальным инструментом для научных исследований, анализа данных и разработки моделей машинного обучения.
4. Доступ к GPU и TPU: Один из ключевых преимуществ Google Colab — это возможность использовать графические процессоры (GPU) и тензорные процессоры (TPU) для ускорения вычислений, что особенно полезно при обучении глубоких нейронных сетей.
5. Интеграция с Google Drive: Пользователи могут легко сохранять свои ноутбуки и данные в Google Drive, что обеспечивает удобный доступ и совместное использование.
6. Установка библиотек: В Google Colab можно устанавливать дополнительные библиотеки с помощью pip, что позволяет использовать широкий спектр инструментов и пакетов.

# 2. **Аналитическая глава**

## 2.1 **Анализ ансамблевых методов**

Ансамблевые методы — это подходы в машинном обучении, которые объединяют несколько моделей для улучшения общей производительности и точности предсказаний. Идея заключается в том, что комбинация нескольких моделей может привести к более надежным и устойчивым результатам, чем использование одной модели. Ансамблевые методы часто применяются для задач классификации и регрессии, применяется несколько принципов и видов методов.

К принципам ансамблевых методов относятся :

1. Комбинирование моделей: Вместо того чтобы полагаться на одну модель, ансамблевые методы используют несколько моделей (или "ансамблей"), которые могут быть как однородными (например, несколько деревьев решений), так и разнородными (например, комбинация деревьев решений и линейных моделей).
2. Устойчивость к переобучению: Ансамблевые методы могут уменьшить риск переобучения, поскольку они учитывают разные аспекты данных, что делает их более устойчивыми к шуму и выбросам.
3. Улучшение точности: Объединение предсказаний нескольких моделей может привести к более точным результатам, поскольку ошибки одной модели могут быть компенсированы другими.

**2.1.1** Виды ансамблевых методов

Самые основные виды:

1. Bagging (Bootstrap Aggregating): Метод, который использует несколько моделей одного типа, обученных на различных выборках данных. Пример: Random Forest — ансамбль деревьев решений, где каждое дерево обучается на случайной подвыборке данных с возвращением (bootstrap sampling).
2. Boosting: Метод, который обучает модели последовательно, каждая новая модель фокусируется на ошибках предыдущих.

Примеры: AdaBoost, Gradient Boosting, XGBoost — все они создают ансамбль путем последовательного добавления моделей, которые исправляют ошибки предыдущих.

1. Stacking: Метод, который объединяет предсказания нескольких базовых моделей с помощью мета-модели.

Базовые модели обучаются на одном наборе данных, а мета-модель (например, линейная регрессия) обучается на выходах базовых моделей для получения окончательного предсказания.

Применение ансамблевых методов имеет множество преимуществ и ряд недостатков, которые нужно учитывать при выполнении задач :

Преимущества ансамблевых методов

1. Увеличение точности: Комбинирование нескольких моделей часто приводит к более высокой точности предсказаний по сравнению с использованием одной модели. Ошибки одной модели могут быть компенсированы другими.
2. Снижение риска переобучения: Ансамблевые методы, особенно bagging, помогают уменьшить риск переобучения, так как они усредняют предсказания нескольких моделей, что делает их более устойчивыми к шуму в данных.
3. Устойчивость к выбросам и шуму: Ансамблевые методы могут быть более устойчивыми к выбросам и шуму в данных, так как разные модели могут по-разному реагировать на такие аномалии.
4. Гибкость: Можно комбинировать различные типы моделей (например, деревья решений, линейные модели, нейронные сети), что позволяет использовать сильные стороны каждой из них.
5. Обработка больших данных: Ансамблевые методы могут эффективно обрабатывать большие объемы данных, особенно в случае параллельной обработки (например, в bagging).
6. Улучшение обобщающей способности: За счет объединения различных моделей ансамблевые методы могут лучше обобщать данные, что полезно для предсказания на новых, невидимых данных.

Недостатками данных методов является :

Сложность и интерпретируемость: Ансамблевые методы, особенно сложные (например, stacking), могут быть трудными для интерпретации. Это может затруднить понимание, как модель принимает решения.

Время обучения: Обучение нескольких моделей может занять значительно больше времени по сравнению с обучением одной модели, особенно если используются сложные алгоритмы.

Потребление ресурсов: Ансамблевые методы могут требовать больше вычислительных ресурсов (памяти и процессорного времени), что может быть проблемой при работе с большими данными.

Риск переобучения в boosting: Хотя boosting может снизить переобучение, при неправильной настройке (например, слишком большое количество итераций) он опять же может привести к переобучению.

Необходимость в настройке гиперпараметров: Ансамблевые методы часто требуют тщательной настройки гиперпараметров, что может усложнить процесс разработки модели.

Сложность в отладке: Если ансамбль не работает должным образом, может быть сложно определить, какая модель или комбинация моделей вызывает проблемы.

**2.1.4** Применение ансамблевых методов в анализе цен на образовательные услуги

В ходе выполнения данной работы по анализу цен на образовательные услуги ансамблевые методы использовались для обучения моделей анализа и прогноза. В ходе работы был применен метод Случайный лес и Градиентный бустинг, для повышения анализа и точности модели, что должно послужить увеличению производительности модели и способности к прогнозированию.

**2.1.5** Заключение аналитической главы

Не смотря на то, что ансамблевые методы являются эффективными в плане поддержания производительности при машинном обучении, всегда нужно учитывать их сильные и слабые стороны и правильно подбирать метод и сопутствующие параметры в зависимости от контекста задач, так как выбор может сильно повлиять на результаты.

# 3. **Практическая часть**

## 3.1 Сбор и подготовка данных

**3.1.1 Источники данных**

Для выполнения задания по проведению анализа цен на услуги образования нужно было подобрать подходящий датасет. Для выполнения задания был выбран датасет Harvard Tuition graduate, включающий в себя колонки направлений и стоимости. ( [Harvard Tuition](https://www.kaggle.com/datasets/harvard-university/harvard-tuition/data) )

**3.1.2 Методы предобработки**

Для качественного проведения анализа данных, входные данные необходимо провести предобработку данных, так как уровень подготовленности данных прямо вляет на точность и производительность машинного обучения. Просмотр данных датасета показал проблему с пропущенными данными, отмеченными как NaN.

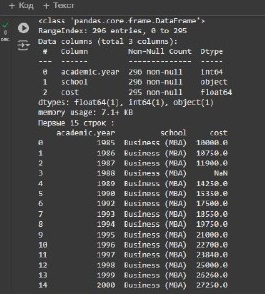


Рисунок 1 – Изначальные данные

В датасете имеются неоднородные, либо пропущенные данные, которые необходимо обработать.

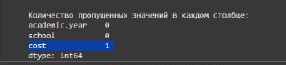


Рисунок 2 – Пропущенные значения

В строке cost, указывающей цену, NaN строка превратилось в значение Unknown. Для исправления данной ошибки необходимо удалить эту строку.

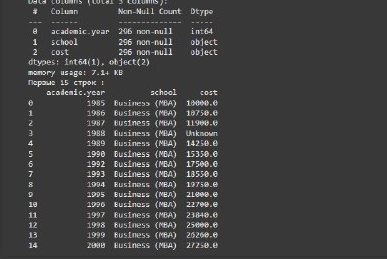


Рисунок 3 – Unknown строка.

После удаления проблемных строк однородность данных должна быть восстановлена. Можно проводить дальнейший анализ данных.

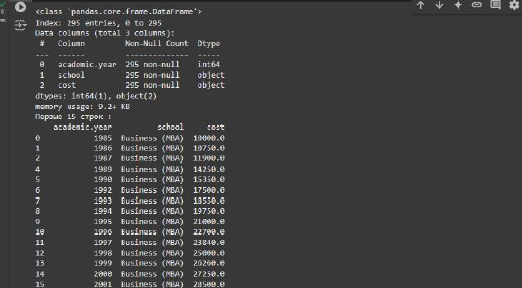


Рисунок 4 – Исправление целостности

В результате проведенных действий, датасет готов к дальнейшему применению и анализу интеллектуальной моделью .

## 3.2 Визуализация обработанных данных

Для точного и наглядного ознакомления и анализа обработанных данных нужно провести визуализацию данных, используя линейные диаграммы, гистограммы, матрицы.

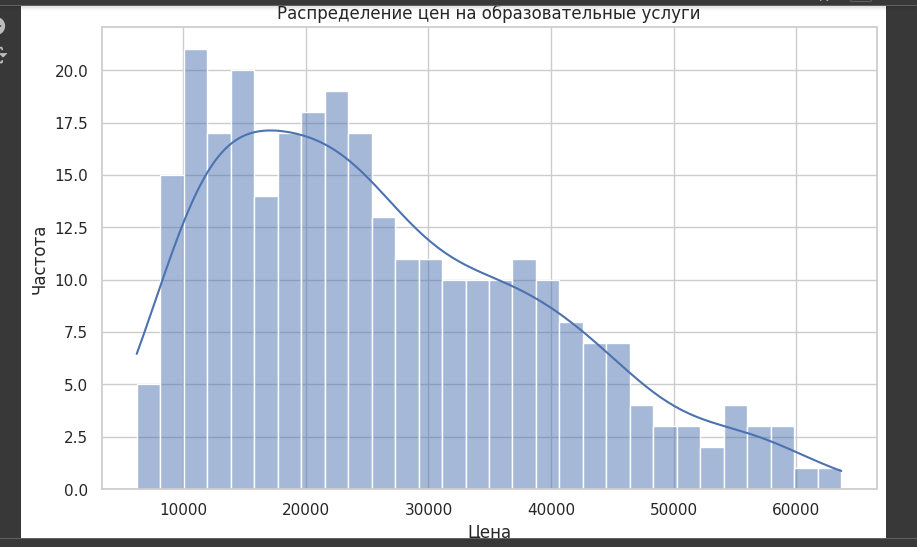


Рисунок 1 – Диаграмма частоты цен

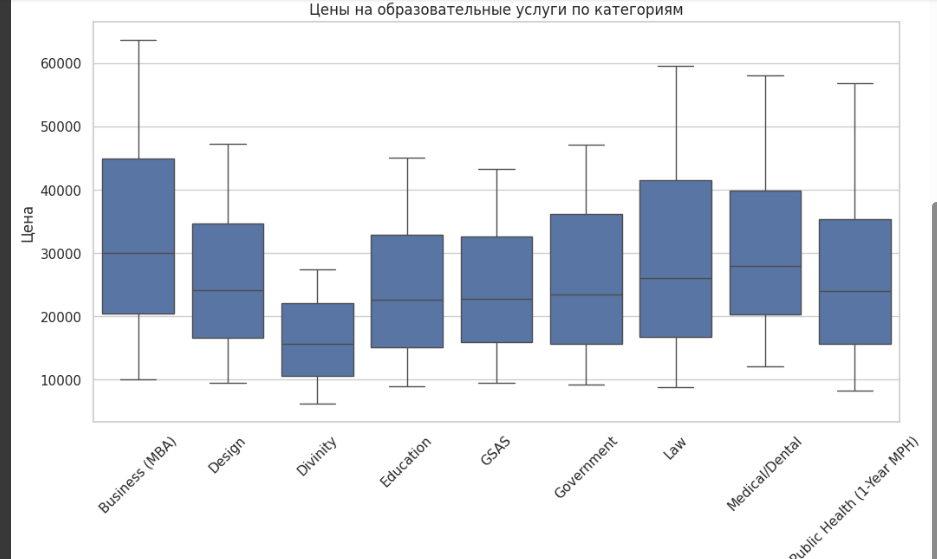


Рисунок 2 – Диаграмма размаха

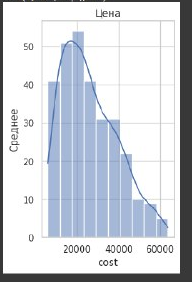


Рисунок 3 – Гистограмма по средней цене.

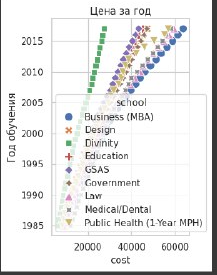
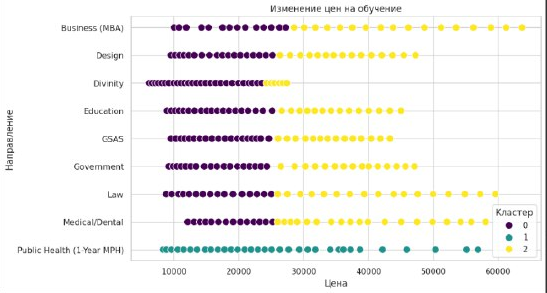


Рисунок 4 - Диаграмма

Рисунок 5. Кластеризация

## Создание моделей

**3.3.1 Линейная регрессия**

Для выявления зависимости между ценами на образовательные услуги и годами обучения (изменение цен со временем) была создана модель линейной регрессии. Работа модели проводилась на основе обработанных данных, что повышает точность анализа и вычислений. В процессе анализа, моделью так же было проведено прогнозирование.

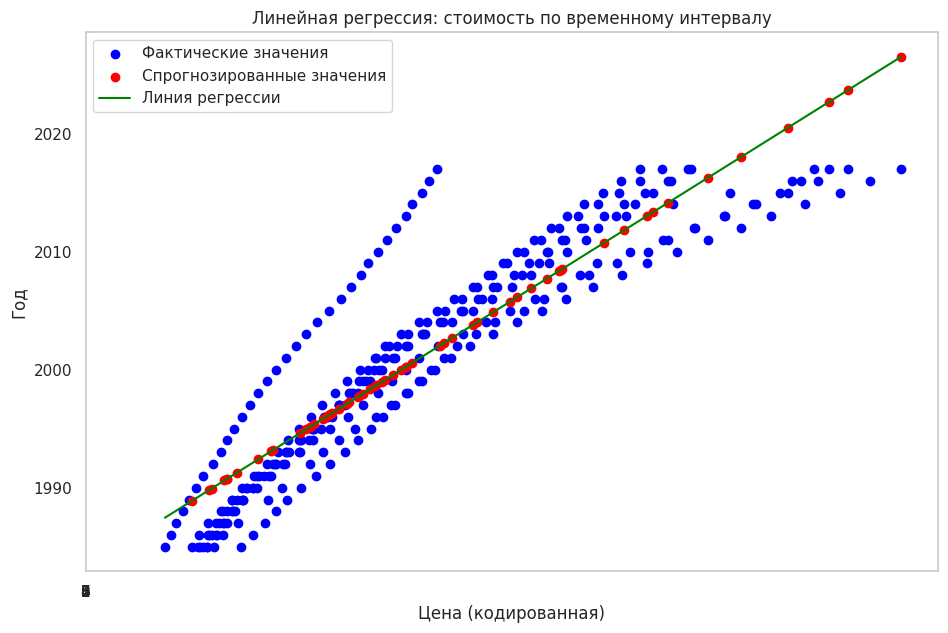


Рисунок 1 – Линейная регрессия

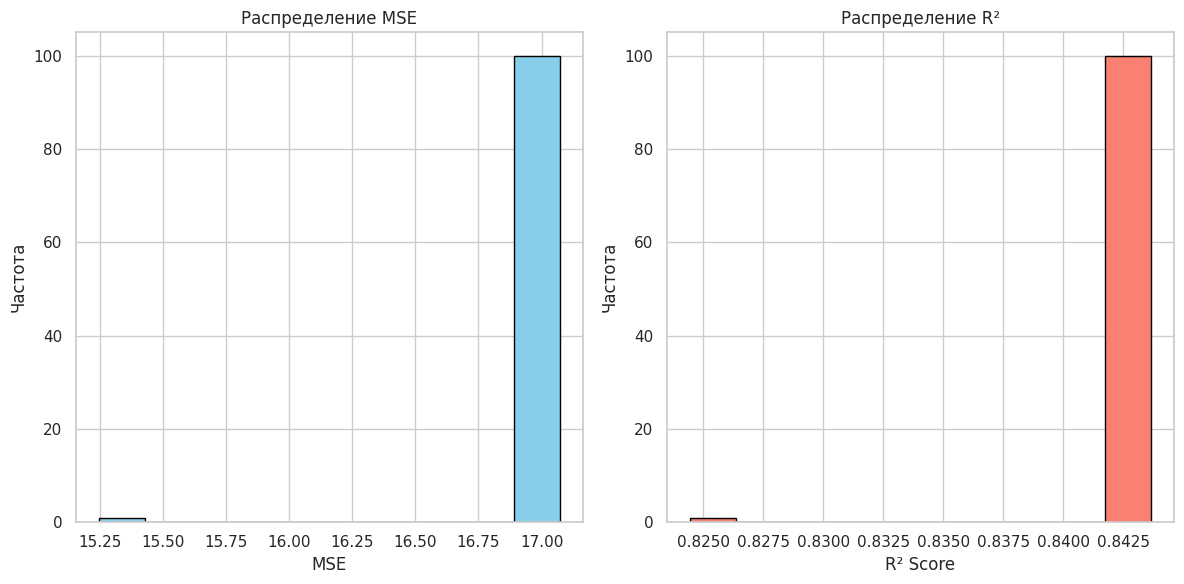
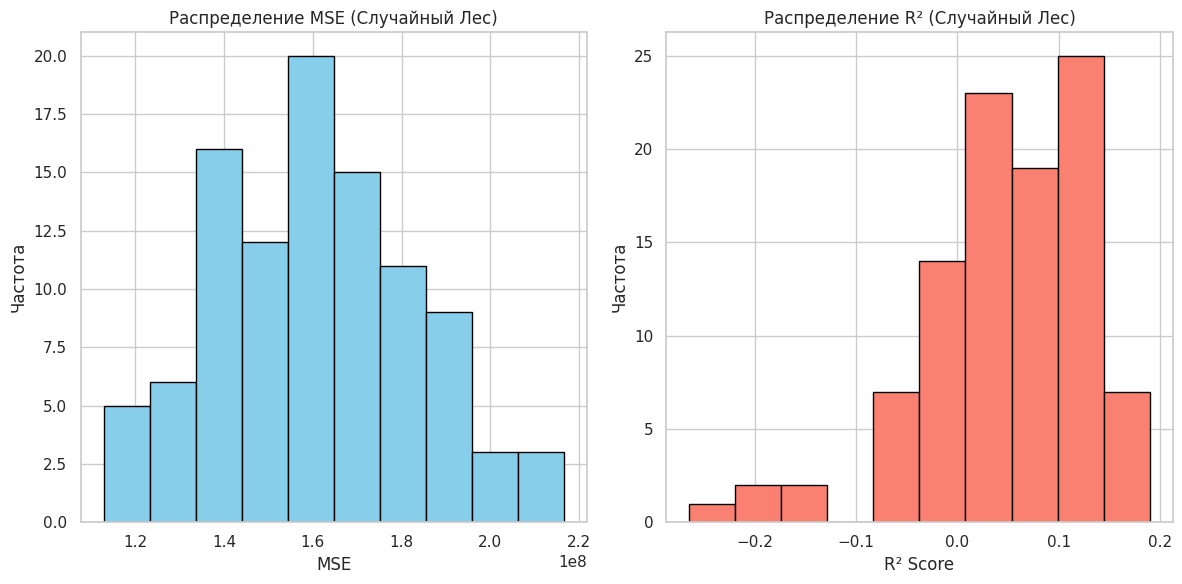
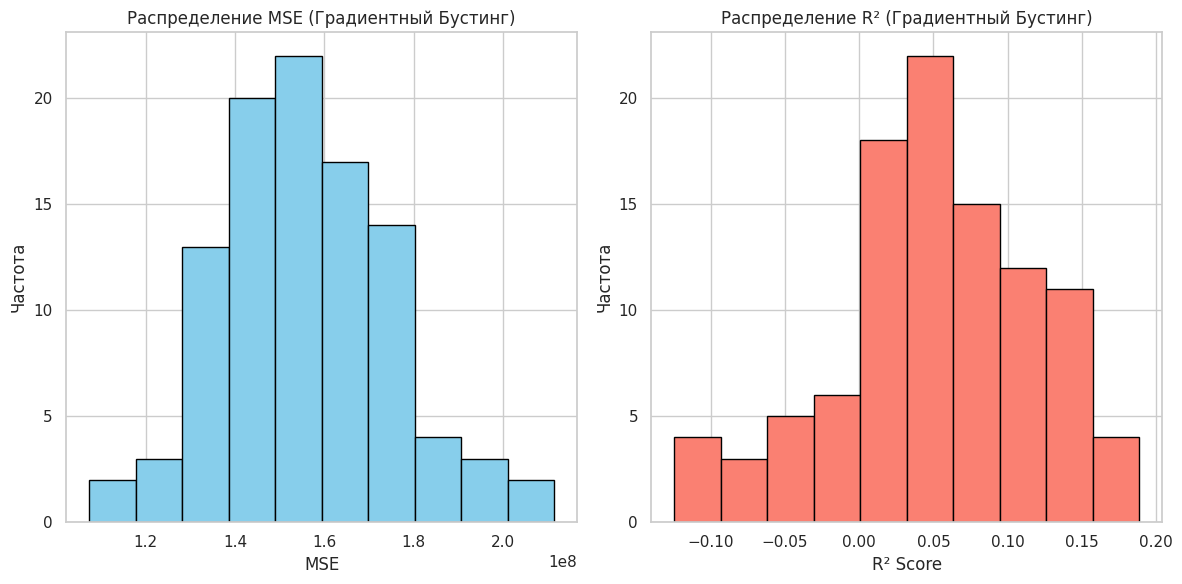


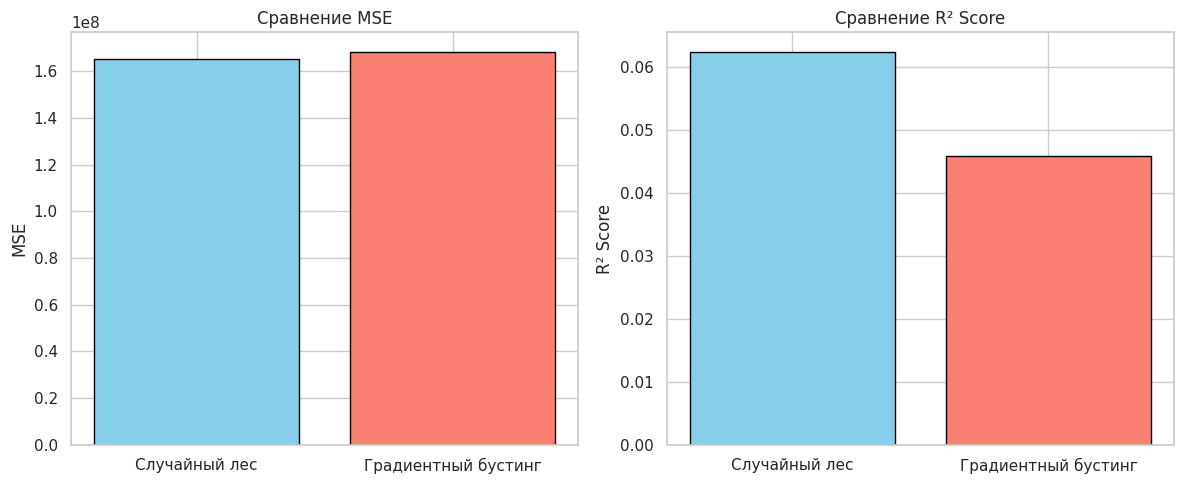
Рисунок 2. 100 итераций обучения модели

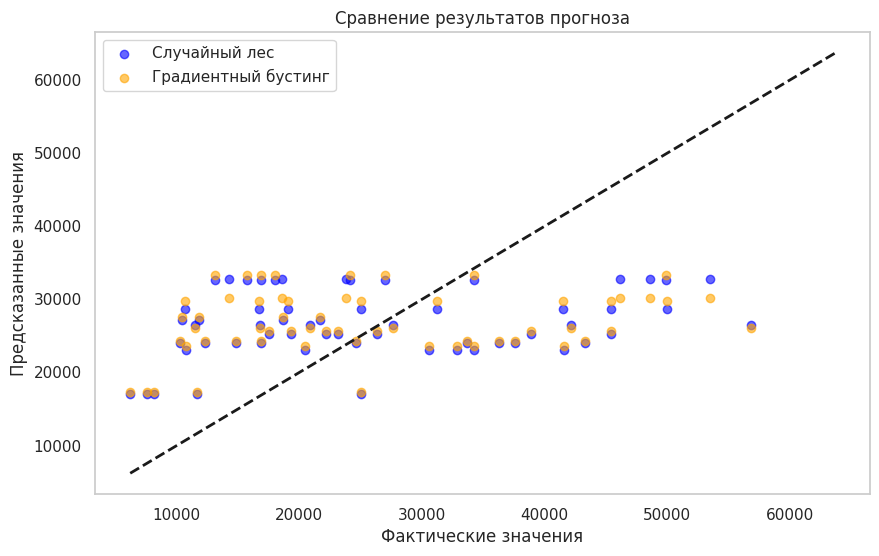
**3.3.2 Использование ансамблевых методов**

Для выполнения всех поставленных задач, было принято решение использовать ансамблевые методы, а именно случайный лес. Этот метод позволяет снизить риск переобучения посредством снижения чувствительности модели к шуму.

Рисунок 3. Случайный лес

Второй ансамблевый метод, который был применен в данной работе - это Градиентный бустинг. Этот метод часто используется для прогнозирования цен и финансового анализа.Рисунок 4. Градиентный бустинг.

Рисунок 4 – Сравнение эффективности использования двух методов.

Рисунок 5. Сравнение прогнозов.

После обучения модели можно сравнить значения, что бы убедится в том, что модель увеличила точность анализа и прогнозирования, и является работоспособной. Для этого используется график, который показывает фактические значения и прогнозы. Для оценки используются метрики MSE и R2.

Средняя MSE: 165370253.91

R² Score: 0.06

# Заключение

В ходе прохождения курсовой поставленные задачи по анализу данных, настройке среды программирования, выбора библиотек, поиску датасета, применению ансамблевых методов были выполнены. Была построена модель для анализа данных, было проведено прогнозирование изменения цен на образовательные услуги на основе данных, собранных в течение прошлых лет. Применение ансамблевых методов показало увеличение точности модели и улучшение прогнозирования. В ходе выполнения работы были получены важные теоретические и практические знания, были получены знания об анализе данных, их обработке прогнозу и проведении. Были получены навыки построения интеллектуальной системы и работы с ней.

Репозиторий: (https://github.com/LilBorrow/Kursovaya.git)

# Список использованной литературы :

1. Григоров, А. А. Предпосылки применения искусственного интеллекта в справочных правовых системах / А. А. Григоров // Вестник науки. – 2024. – № 1
2. Дзялошинский, И. М. Искусственный интеллект: гуманитарная перспектива / И. М. Дзялошинский // Вестник Новосибирского государственного университета. Серия: История, филология. – 2022. – Т. 21, № 6. – С. 20-29.
3. Долженко, Е. Р. Искусственный интеллект и нейронные сети в управлении персоналом / Е. Р. Долженко // Управление развитием персонала. – 2023. – № 3. – С. 170-178.
4. Искусственный интеллект и трансформация менеджмента / Л. Ф. Никулин, В. В. Великороссов, С. А. Филин, А. Б. Ланчаков // Экономический анализ: теория и практика. – 2023. – Т. 22, № 3. – С. 556-573.
5. Использование возможностей искусственного интеллекта в рекламе / Д. А Шевченко, Е. М. Крюкова, В. В. Зеленов, В. В. Галстян // Практический маркетинг. – 2024. – № 1. – С. 60-64.
6. Карякин, В. В. Гибридные интеллектуальные системы как симбиоз

естественного и искусственного интеллектов / В. В. Карякин // Россия: тенденции и перспективы развития. – 2022. – № 17-1. – С. 652-655.

1. Рассел, С. Совместимость. Как контролировать искусственный интеллект практическое руководство / С. Рассел. – Москва : Альпина нон-фикшн, 2021. – 438 с.
2. Сергеев, Л. И. Цифровая экономика : учебник для вузов / Л. И. Сергеев, Д. Л. Сергеев, А. Л. Юданова ; под ред. Л. И. Сергеева. – 2-е изд., перераб. и доп. – Москва : Юрайт, 2024. – 437 с. – (Высшее образование).
3. Человек и системы искусственного интеллекта / В. А. Лекторский, С. Н. Васильев, В. Л. Макаров [и др.]. – Санкт-Петербург : Общество с ограниченной ответственностью "Издательство "Юридический

центр", 2022. – 328 с. – ISBN 978-5-94201-835-1.