

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
GRADUATION THESIS**

Прогнозирование ставки репетитора при помощи методов машинного обучения

Обучающийся / Student Антоненко Елизавета Павловна

Факультет/институт/кластер/ Faculty/Institute/Cluster высшая школа цифровой культуры
Группа/Group S42012

Направление подготовки/ Subject area 02.04.03 Математическое обеспечение и администрирование информационных систем

Образовательная программа / Educational program Аналитика данных 2021


Язык реализации ОП / Language of the educational program Русский

Статус ОП / Status of educational program

Квалификация/ Degree level Магистр

Руководитель ВКР/ Thesis supervisor Михайлова Елена Георгиевна, доцент, кандидат физико-математических наук, Университет ИТМО, высшая школа цифровой культуры, директор

Обучающийся/Student


Документ подписан	
Антоненко Елизавета Павловна	
29.05.2023	

(эл. подпись/ signature)

Антоненко
Елизавета
Павловна

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Михайлова Елена Георгиевна	
29.05.2023	

(эл. подпись/ signature)

Михайлова
Елена
Георгиевна

(Фамилия И.О./ name
and surname)

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**ЗАДАНИЕ НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ /
OBJECTIVES FOR A GRADUATION THESIS**

Обучающийся / Student Антоненко Елизавета Павловна

Факультет/институт/кластер/ Faculty/Institute/Cluster высшая школа цифровой культуры
Группа/Group S42012

Направление подготовки/ Subject area 02.04.03 Математическое обеспечение и администрирование информационных систем

Образовательная программа / Educational program Аналитика данных 2021

Язык реализации ОП / Language of the educational program Русский

Статус ОП / Status of educational program

Квалификация/ Degree level Магистр

Тема ВКР/ Thesis topic Прогнозирование ставки репетитора при помощи методов машинного обучения

Руководитель ВКР/ Thesis supervisor Михайлова Елена Георгиевна, доцент, кандидат физико-математических наук, Университет ИТМО, высшая школа цифровой культуры, директор

Основные вопросы, подлежащие разработке / Key issues to be analyzed

Техническое задание: Прогнозирование ставки репетитора по описанию и отзывам с помощью методов машинного обучения.

Цели и задачи работы:

1. Изучить теоретические основы предварительной обработки данных и построения регрессионных моделей.
2. Собрать необходимые данные и сформировать выборку.
3. Провести предварительную обработку данных.
4. Провести графический и факторный анализ.
5. Обучить несколько моделей.
6. Провести корректировку параметров.
7. Сравнить построенные модели и сделать вывод.

Данные для анализа взяты с сайта repetitors.info. Анализ проводится с помощью методов машинного обучения на языке Python.

Материалы, рекомендуемые для изучения:

1. Для анализа задачи и актуальности - статьи со статистическими данными о ставках репетиторов от компаний, работающих в сфере репетиторства
2. Для теоретических основ :
МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ - В. М. БУРЕ, Е. М. ПАРИЛИНА, А. А. СЕДАКОВ

Naked Statistics: Stripping the Dread from the Data - Charles Wheelan

3. Для работы с методом Python:

документации необходимых библиотек

Selenium with Python - Baiju Muthukadan

Дата выдачи задания / Assignment issued on: 24.04.2023

Срок представления готовой ВКР / Deadline for final edition of the thesis 24.05.2023

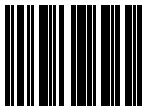
Характеристика темы ВКР / Description of thesis subject (topic)

Тема в области фундаментальных исследований / Subject of fundamental research: нет / not

Тема в области прикладных исследований / Subject of applied research: да / yes

СОГЛАСОВАНО / AGREED:

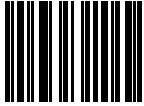
Руководитель ВКР/
Thesis supervisor

Документ подписан	
Михайлова Елена Георгиевна	
27.05.2023	

(эл. подпись)

Михайлова
Елена
Георгиевна

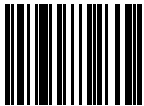
Задание принял к
исполнению/ Objectives
assumed BY

Документ подписан	
Антоненко Елизавета Павловна	
27.05.2023	

(эл. подпись)

Антоненко
Елизавета
Павловна

Руководитель ОП/ Head
of educational program

Документ подписан	
Михайлова Елена Георгиевна	
29.05.2023	

(эл. подпись)

Михайлова
Елена
Георгиевна

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
SUMMARY OF A GRADUATION THESIS**

Обучающийся / Student Антоненко Елизавета Павловна
Факультет/институт/кластер/ Faculty/Institute/Cluster высшая школа цифровой культуры
Группа/Group S42012
Направление подготовки/ Subject area 02.04.03 Математическое обеспечение и администрирование информационных систем
Образовательная программа / Educational program Аналитика данных 2021
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Магистр
Тема ВКР/ Thesis topic Прогнозирование ставки репетитора при помощи методов машинного обучения
Руководитель ВКР/ Thesis supervisor Михайлова Елена Георгиевна, доцент, кандидат физико-математических наук, Университет ИТМО, высшая школа цифровой культуры, директор

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
DESCRIPTION OF THE GRADUATION THESIS**

Цель исследования / Research goal

В данной статье исследуется проблема прогнозирования ставки репетитора на основе данных анкеты и других параметров. Целью исследования является определение наиболее важных признаков, влияющих на ставку, и сравнение различных методов машинного обучения для достижения наилучшего прогноза.

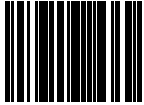
Задачи, решаемые в ВКР / Research tasks

Для исследования были поставлены следующие задачи: собрать необходимые данные, провести качественную предобработку данных, построить различные модели и провести анализ качества и сравнение моделей.

Краткая характеристика полученных результатов / Short summary of results/findings

Лучшего качества достигли модели градиентного бустинга. Наиболее значимыми параметрами оказались: город, количество отзывов, опыт репетитора, средняя оценка.

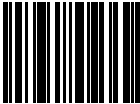
Обучающийся/Student

Документ подписан	
Антоненко Елизавета Павловна	
29.05.2023	

Антоненко
Елизавета
Павловна

Руководитель ВКР/
Thesis supervisor

(эл. подпись/ signature)

Документ подписан	
Михайлова Елена Георгиевна	
29.05.2023	

(эл. подпись/ signature)

(Фамилия И.О./ name
and surname)

Михайлова
Елена
Георгиевна

(Фамилия И.О./ name
and surname)

Содержание

Глава 1. Введение	2
1.1. Факторы	2
1.2. Актуальность	4
1.3. Задачи	5
Глава 2. Теоретические основы	6
2.1. Предварительная обработка данных	6
2.2. Предварительная обработка текста	11
2.3. Постановка задачи регрессии и описание моделей	14
2.4. Критерии качества модели	20
2.5. Способы подбора гиперпараметров	22
Глава 3. Список библиотек Python	26
Глава 4. Сбор и подготовка данных к анализу	28
4.1. Сбор данных	28
4.2. Предобработка данных	29
4.3. Предобработка текста	32
Глава 5. Анализ данных	34
5.1. Графический анализ	34
5.2. Факторный анализ	35
5.3. Обучение моделей	36
5.4. Корректировка параметров	42
Глава 6. Заключение	46
Список литературы	48

Глава 1. Введение

Репетитор – это педагог, дающий частные уроки на дому или дистанционно. Современные образовательные стандарты и высокие требования к уровню знаний в школе требуют массового привлечения квалифицированных педагогов для индивидуальных занятий. Поэтому услуги репетиторов пользуются огромной популярностью практически на всех этапах обучения [1].

Тема данной работы - прогнозирование ставки репетитора - актуальна в наше время, так как может помочь как самим репетиторам, так и ученикам. Для репетиторов правильная установка цены может привести к увеличению числа клиентов и увеличению дохода. Для учеников правильно выбранная цена может помочь найти подходящего репетитора в рамках своего бюджета. Кроме того, прогнозирование ставки может быть полезно для онлайн-платформ, которые предоставляют услуги репетиторов. Здесь анализ данных о ставках может помочь платформе оптимизировать ценообразование и увеличить доходность.

Важно также знание признаков, влияющих на ставку репетитора. Это позволяет определить, какие навыки, опыт и характеристики сильно влияют на стоимость его услуг. Это может помочь репетиторам сосредоточиться на развитии и совершенствовании ключевых навыков, что в конечном итоге повысит качество предоставляемых ими услуг. Зная, какие признаки важны для определения ставки репетитора, можно лучше распределить ресурсы, направленные на развитие и поддержку репетиторов. Это позволит инвестировать в наиболее важные аспекты и обеспечить максимальную эффективность использования ресурсов.

1.1 Факторы

Для прогнозирования необходимо выделить факторы, которые могут повлиять на предсказание ставки. После ознакомления со статьями компаний, работающих в сфере репетиторства ([2] - [10]), были выделены следующие факторы:

1. Опыт работы.

На стоимость занятий в первую очередь влияет опыт репетитора. Чем больше опыта и квалификации, тем выше может быть ставка за занятие.

2. Предмет.

Сложность предмета или высокий спрос также могут влиять на ставку репетитора. Английский язык и математика — предметы, по которым чаще всего ищут педагога. Спрос на них выше, а значит, и цена тоже.

3. Регион.

Цена услуг учителя зависит от города, в котором он преподает. В Москве и Санкт-Петербурге стоимость выше, чем в регионах. По данным Института образования ВШЭ, средняя ставка в час у репетиторов в Москве составляет 691 рубль, в регионах — 427 рублей [8].

4. Наличие и количество отзывов.

Часто клиенты читают отзывы о репетиторах, прежде чем сделать выбор. Если репетитор имеет достаточное количество положительных отзывов, то это говорит о нем, как о хорошем специалисте.

5. Средняя оценка на платформе.

Средняя оценка считается как среднее по всем оценкам, которые поставили репетитору и очень влияет на решение будущего клиента.

6. Образование.

Частные уроки дают учителя с разным уровнем образования. Это может быть студент — тогда цена урока будет низкой. А может быть преподаватель с дипломом хорошего вуза и профильными сертификатами. Согласно опросу на одной из платформ для репетиторов, самые высокооплачиваемые репетиторы — преподаватели вузов, самые низкооплачиваемые — студенты [9].

7. Формат занятия.

Онлайн-занятия могут быть дешевле, чем занятия вживую, так как не требуют дополнительных расходов на дорогу и аренду помещения.

8. Цель занятий.

Занятие может быть с целью заполнить пробелы по предмету или, наоборот, продвинуться вперед по программе, а могут содержать конкретную цель - подготовка к экзаменам или олимпиадам. В среднем по России цена за занятия по подготовке к ЕГЭ/ОГЭ увеличивается примерно на 30%, по сравнению с другими [10].

1.2 Актуальность

Изучив причины, влияющие на изменение ставки за занятия, рассмотрим, для кого полезно предсказывать эту ставку:

1. Начинающий репетитор.

Если репетитор только начинает свою деятельность и еще не имеет достаточного опыта и рекомендаций, то установка слишком высокой цены может отпугнуть потенциальных клиентов, а слишком низкой - обесценить его труд.

2. Репетор, у которого нет клиентов.

Чаще всего отказы приходят из-за предложения слишком высокой цены. Если стоимость услуг репетитора значительно выше, чем у конкурентов, то это может стать причиной, по которой клиенты выберут другого репетитора. Кроме того, если репетитор ставит слишком высокую цену, но при этом не предлагает соответствующего качества услуг, то это также может негативно сказаться на его репутации и привести к уменьшению количества учеников.

3. Ученики и их родители.

Оценка диапазона ставок по запросу клиента может помочь сделать более обоснованный выбор и выбрать репетитора, который соответствует финансовым возможностям и ожиданиям.

4. Платформы, которые предоставляют услуги репетиторства.

Анализируя ставки различных репетиторов, можно определить среднюю цену на занятие и тенденции в изменении цен, что может помочь улучшить качество услуг и повысить их доступность. Также можно использовать эти данные для привлечения и удержания как учеников, так и репетиторов.

1.3 Задачи

Целью данной работы является построение одной или нескольких моделей, предсказывающих ставку репетитора по входящим факторам. Для достижения цели были сформулированы следующие задачи:

1. Изучить теоретические основы предварительной обработки данных и построения регрессионных моделей.
2. Собрать необходимые данные и сформировать выборку.
3. Провести предварительную обработку данных.
4. Провести графический и факторный анализ.
5. Обучить несколько моделей.
6. Провести корректировку параметров.
7. Сравнить построенные модели и сделать вывод.

Работа с данными будет проведена, используя методы языка Python [30].

Глава 2. Теоретические основы

Основными источниками для изучения предварительного и регрессионного анализа были следующие статьи и методические пособия: [11] - [21]

2.1 Предварительная обработка данных

Методы предобработки данных - это техники обработки данных, которые выполняются до обучения модели. Они позволяют улучшить качество модели и избежать ошибок, связанных с качеством входных данных.

1. Удаление дубликатов.
2. Обработка пропущенных значений.
3. Кодирование категориальных признаков.
4. Удаление выбросов.
5. Выбор признаков.

Рассмотрим этапы подробнее:

1. Удаление дубликатов.

В большинстве случаев дубликаты рассматриваются как негативный фактор, и в процессе очистки данных от них стремятся избавиться. Потому что дублирующие записи не несут полезной информации и могут только пагубно повлиять на обучение модели.

2. Обработка пропущенных значений.

Пропущенные значения могут появляться по разным причинам: ошибки при вводе данных, отсутствие данных, пропуски в измерениях и т.д. Это важный этап предобработки данных, который заключается в заполнении или удалении пропущенных значений в наборе данных.

Есть несколько способов обработки пропущенных значений:

(a) Удаление строк или столбцов с пропущенными значениями

Этот метод используется, если пропущенные значения занимают небольшую долю от общего количества данных. Удаление пропущенных значений может привести к потере информации, поэтому он должен использоваться с осторожностью.

(b) Заполнение пропущенных значений.

Данный метод используется, если пропущенные значения занимают значительную долю от общего количества данных. Заполнение может происходить с помощью разных методов, например: заполнение медианой, модой (наиболее часто встречаемое значение), интерполяция (метод в основном используется во временных рядах и заключается в вычислении пропущенных значений на основе соседних)

.

Важно выбрать метод обработки пропущенных значений, который соответствует характеру данных и цели анализа. Неправильная обработка пропущенных значений может привести к искажению результатов анализа и появлению ошибок в моделях машинного обучения.

3. Кодирование категориальных признаков.

Категориальные признаки - это переменные, которые принимают значения из заданного набора категорий. Некоторые модели машинного обучения могут работать только с числовыми данными, поэтому категориальные признаки должны быть преобразованы в числовые признаки.

Самый популярный метод для кодирования - One-hot encoding. При этом методе каждая категория преобразуется в бинарный вектор, где для каждой категории устанавливается только одно значение 1, а все остальные равны 0. Этот метод хорошо работает в случае, когда категориальный признак имеет небольшое число уникальных значений.

Существует еще один метод - Label encoding, в котором каждая кате-

гория преобразуется в уникальное числовое значение. Этот метод может быть эффективен в случае, если категориальный признак имеет большое число уникальных значений, но при этом он может вызывать проблемы, когда модель начинает интерпретировать кодирование как упорядоченный ранг признаков.

4. **Обработка выбросов.**

В данных могут присутствовать значения, являющиеся выбросами. Выбросы - экстремальные значения во входных данных, которые находятся далеко за пределами других наблюдений. Они могут возникать из-за ошибок измерения, несоответствия формату или неправильного ввода данных. Своими значениями они сильно путают модель, влияя на ее обучение. Обработка выбросов позволяет улучшить качество работы модели.

Для того, чтобы определить, является ли значение выбросом, пользуются характеристикой выборки, называемой межквартильным размахом:

$$IQR = Q_3 - Q_1,$$

где Q_1 - первая квартиль - это такое число, что ровно 25% выборки меньше него, а Q_3 - третья квартиль — число, меньше которого ровно 75% всех значений признака.

Принято за выбросы считать значения, которые выходят за 1.5 межквартильных размаха, то есть не принадлежат отрезку:

$$[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$$

Существует несколько методов обработки выбросов:

- (а) Удаление выбросов.

В этом методе выбросы просто удаляются из набора данных. Данный метод может быть эффективным, если количество выбросов невелико. Однако при удалении выбросов может быть потеряно много информации.

(b) Замена выбросов.

В этом методе выбросы заменяются на другие значения, например, на среднее или медиану. Этот метод может быть эффективным, если выбросов немного, и они не слишком сильно отличаются от остальных значений. С другой стороны замена выбросов на среднее значение может привести к смещению данных, что может негативно сказаться на точности модели.

(c) Использование статистических методов.

Этот метод основан на анализе статистических свойств данных, таких как среднее значение и дисперсия. Например, можно использовать правило трех сигм, при котором выбросы определяются как значения, которые находятся за пределами трех сигм от среднего значения. Метод может быть эффективным, если распределение данных приближается к нормальному распределению.

5. Отбор признаков и снижение размерности.

Этот этап заключается в выборе наиболее значимых признаков, которые влияют на целевую переменную и исключении ненужных признаков или преобразовании исходных признаков в новые, более компактные компоненты, с максимальным сохранением информации о данных.

Существует много методов для отбора параметров и снижения размерности, познакомимся с самыми распространенными:

(a) Метод корреляционного отбора признаков.

Один из методов отбора факторов базируется на анализе матрицы парных коэффициентов корреляции.

Корреляция измеряет степень связи между двумя явлениями. Коэффициент корреляции представляет собой число в диапазоне от -1 до 1 и им выражается связь между двумя переменными. Чем ближе корреляция к 1 или -1 , тем сильнее связь между переменными. Нулевая корреляция говорит об отсутствии значимой связи между двумя переменными.

Корреляционная матрица представляет собой квадратную таблицу, в которой каждый элемент является коэффициентом корреляции между двумя признаками. Считается, что две переменные явно коллинеарны, если парный коэффициент корреляции превышает значение 0.7 . В таком случае одна из них исключается из модели. Предпочтение отдается тому фактору, который достаточно тесно связан с результативным фактором, но имеет при этом наименьшую тесноту связи с другими объясняющими факторами.

(b) Метод главных компонент.

Метод используется для преобразования исходных признаков в новое пространство, которое содержит меньшее количество переменных (главных компонент), при этом сохраняется наибольшую дисперсию данных.

Основная идея метода главных компонент заключается в том, что исходные признаки коррелируют друг с другом, и некоторые из них могут быть выражены через линейные комбинации других. Эти линейные комбинации называются главными компонентами и представляют собой новые признаки, каждый из которых является линейной комбинацией исходных.

(c) Метод отбора признаков на основе важности.

Это один из способов оценки важности признаков. Существует несколько способов оценки feature importance, один из наиболее распространенных — это использование алгоритмов, которые на основе значений признаков и целевой переменной автоматически определяют, насколько каждый признак влияет на предска-

зание.

Например, одним из популярных алгоритмов является RandomForest, который может рассчитывать важность каждого признака на основе их использования при построении деревьев решений. Для этого алгоритм переставляет значения каждого признака случайным образом и смотрит, как это влияет на точность модели. Если признак является важным, то перестановка его значений должна существенно снизить точность модели.

2.2 Предварительная обработка текста

Предварительная обработка текста - это процесс преобразования текстовых данных в структурированный формат, понятный для алгоритмов машинного обучения. Предобработка состоит из различных этапов, которые могут отличаться в зависимости от задачи и реализации:

1. Токенизация.

Чаще всего используется простое разделение текста на слова (токены), но могут быть и другие методы, например, разбиение на символы или n-граммы.

2. Приведение к нижнему регистру.

Данный этап позволяет избежать дублирования признаков, если слово встречается в разных регистрах.

3. Удаление специальных символов.

При обработке текста специальные символы (например, знаки препинания или символы математических операций) могут искажать смысл текста и тем самым затруднять работу алгоритмов обработки текста.

4. Удаление стоп-слов.

Стоп-слова - это слова, которые часто встречаются в текстах, но не несут смысловой нагрузки и не влияют на результаты анализа текста. Это могут быть предлоги, союзы, местоимения и т.д.

Удаление стоп-слов может быть полезным для улучшения качества анализа текста, так как они уменьшают шум и повышают точность выделения ключевых слов. Кроме того, процесс может ускорить обработку текстов, так как это уменьшает объем данных.

5. Лемматизация.

Лемматизация - приведение всех слов к их базовой форме (лемме). Например, слова "книга", "книги" и "книгу" будут приведены к форме "книга". Лемматизация позволяет уменьшить размерность пространства признаков и избежать дублирования.

6. Стемминг.

Это процесс нахождения основы слова (stem) путём удаления его окончаний. Например, слова "книга", "книги" и "книгу" будут приведены к форме "книг". Стемминг также позволяет уменьшить размерность пространства признаков и устранить шум.

7. Векторизация.

Это этап преобразования текста в числовой формат - вектор, понятный для алгоритмов машинного обучения. Обычно для векторизации текста используются следующие методы:

(a) Мешок слов (Bag-Of-Words)

Простейший способ векторизации заключается в заполнении вектора частотами появлений всех уникальных слов словаря в каждом документе. Словарь, в данном случае, представляет собой список уникальных слов. Результатом является матрица, где каждый столбец соответствует уникальному слову, а каждая строка - отдельному тексту.

У метода есть 2 значительных недостатка: игнорирование контекста и большая размерность пространства признаков. Но также метод имеет и преимущества над другими: не требует сложных вычислений, поэтому он может использоваться для обра-

ботки больших объемов данных и быстрого прототипирования моделей.

(b) *TF-IDF* (Term Frequency - Inverse Document Frequency)

Главная идея этого подхода состоит в том, что основной смысл документа закодирован в более редких словах. В этом методе каждое слово получает вес, который зависит от того, как часто оно встречается в конкретном тексте, а также от того, насколько оно уникально для всего корпуса текстов. Реализуется он благодаря двум параметрам: *TF* и *IDF*.

TF - частота слова в документе. Вычисляется как отношение числа вхождений некоторого слова к общему числу слов документа.

$$TF(t, d) = \frac{n_t}{\sum_k n_k},$$

где t - слово, d - документ, n_t - частота слова в документе, n_k - уникальное слово в документе.

IDF - обратная частота документа - инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$IDF(t, D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|},$$

где D - корпус документов d , $|D|$ - мощность корпуса, а в знаменателе считается количество документов, в которых есть слово t . Логарифмирование используется для сглаживания весов слов, что позволяет сделать веса более равномерными и избежать ситуации, когда редкие слова получают слишком высокий вес, а часто встречающиеся слова - низкий.

Итоговая формула вычисления веса (значимости) слова t в документе d из корпуса D :

$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D),$$

Таким образом большой вес в $TF-IDF$ получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

По сравнению с мешком слов, у данного метода есть большое преимущество - возможность оценить важность слова в контексте. Но этот метод, как и предыдущий, не учитывает порядок слов, что может привести к потере информации, связанной с контекстом.

(с) *Word2Vec*

Данный метод преобразует каждое слово в вектор фиксированной длины.

Word2Vec имеет две основные архитектуры: *CBOW* (Continuous Bag-of-Words) и *skip-gram*. Архитектура *CBOW* пытается предсказать текущее слово на основе контекста, тогда как архитектура *Skip-gram* пытается предсказать контекст на основе текущего слова. В обоих случаях алгоритм обучается на большом корпусе текста, чтобы определить взаимосвязи между словами. Этот метод учитывает не только частоту слов, но и контекст их использования, что позволяет получить более информативные векторы.

2.3 Постановка задачи регрессии и описание моделей

В работе решается задача регрессии.

Задача регрессии - это задача машинного обучения, в которой требуется построить модель, предсказывающая непрерывный числовой выход на основе входных параметров.

Методов, решающих задачу регрессии, очень много. В список часто используемых точно входят:

1. Линейная регрессия.
2. Регрессия на основе деревьев решений.

3. Метод опорных векторов (SVM).
4. Градиентный бустинг.

Рассмотрим каждый метод отдельно:

1. Линейная регрессия.

Это метод, который используется для определения линейной зависимости между независимыми переменными и зависимой переменной, путем построения линейной функции, вида:

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m + \varepsilon,$$

где y — зависимая переменная (результативный признак), x_1, x_2, \dots, x_m — независимые переменные (факторы), a_0, a_1, \dots, a_m — параметры модели, которые определяют коэффициенты наклона прямой, а ε — случайная ошибка.

Задача линейной регрессии — определить наилучшие значения параметров модели a_0, a_1, \dots, a_m таким образом, чтобы минимизировать сумму квадратов ошибок. Это достигается путем использования метода наименьших квадратов, который минимизирует сумму квадратов расстояний между прогнозируемыми значениями y_i и истинными значениями \hat{y}_i :

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min,$$

где n — количество наблюдений.

Преимущества линейной регрессии — простота и легкость в использовании, быстрое время обучения модели, интерпретируемость результатов, хорошие результаты, когда связь между переменными линейная. В недостатки модели можно записать: чувствительность к выбросам и мультиколлинеарности (высокой корреляции между независимыми переменными).

В Python модель линейной регрессии реализована в библиотеке `scikit-learn` в классе **LinearRegression**. В работе также рассматриваются методы **Lasso** (Least Absolute Shrinkage and Selection Operator) и **Ridge** (Regularized Least Squares Regression) - это методы регуляризации линейной регрессии, которые используются для уменьшения переобучения модели. В обычной линейной регрессии минимизируется функция среднеквадратической ошибки, которая может привести к переобучению модели. Lasso и Ridge добавляют штрафы на коэффициенты модели, чтобы ограничить их значения. Разница между Lasso и Ridge заключается в том, какой тип штрафа используется.

В Lasso используется L_1 -регуляризация, которая добавляет сумму абсолютных значений коэффициентов. В итоге целевая функция выглядит следующим образом:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^m |a_j|$$

В Ridge используется L_2 -регуляризация, которая добавляет сумму квадратов коэффициентов

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^m a_j^2$$

Обе регуляризации имеют параметр λ , который контролирует силу регуляризации. Чем больше λ , тем сильнее регуляризация, и модель более склонна к уменьшению весов. При этом, если λ слишком велико, то модель может потерять способность обобщаться на новые данные. Поэтому настройка этого параметра является важной частью обучения модели. Можно рассмотреть как близкие к нулю значения - 0.001, так и большие значения - 1000.

2. Регрессия на основе деревьев решений.

Метод основывается на построении древовидной структуры, в которой каждый узел представляет тест на одном из признаков, а каждый лист соответствует возможному значению этого признака.

Процесс построения начинается с корневого узла, который содержит все доступные образцы данных. Затем на каждом шаге модель разделяет образцы на две подгруппы в зависимости от значения одного из признаков. Разделение происходит таким образом, чтобы минимизировать среднеквадратичную ошибку предсказаний в каждой подгруппе. Процесс разделения продолжается до тех пор, пока не будет достигнут критерий остановки.

Модели, основанные на рассматриваемом методе, имеют ряд преимуществ: они легко интерпретируемы и могут помочь понять важность каждого признака в принятии решения, могут обрабатывать как категориальные, так и числовые признаки, а так же устойчивы к выбросам. Из недостатков метода можно выделить следующие: деревья решений часто склонны к переобучению, не могут эффективно обрабатывать мультиколлинеарные признаки.

Для предотвращения переобучения деревьев решений, существуют несколько подходов:

(а) Ранняя остановка.

В данном подходе на каждой итерации проверяется изменение ошибки. Если ошибка перестает значительно улучшаться или начинает увеличиваться, разделение прекращается и происходит остановка построения дерева. Преимущество метода — сокращение временных затрат на обучение. Главный недостаток — ранняя остановка может негативно сказаться на точности дерева.

(б) Ограничение глубины дерева.

Алгоритм останавливается после достижения установленного числа разбиений в ветвях.

(с) Задание минимального количества образцов в листе.

Устанавливается минимальное количество образцов, которое должно находиться в каждом листе. Если количество образцов в листе становится меньше этого значения, то дальнейшее разбиение прекращается.

(d) Ограничение количества листьев.

Алгоритм останавливается после достижения установленного числа листьев в дереве.

(e) Задание минимального количества образцов для разделения.

Устанавливается минимальное количество образцов, которое должно находиться в узле, чтобы разделение этого узла было допустимым. Если количество образцов меньше или равно заданому значению, то разделение узла не происходит, и узел становится листом дерева.

(f) Ограничение времени выполнения.

Если время превышает установленное значение, процесс прерывается.

В работе рассматриваются модели **DecisionTreeRegression** - классическая модель регрессии, основанная на методе дерева решений, и **RandomForestRegression** - расширение идеи деревьев решений, где несколько деревьев объединяются для улучшения качества предсказаний. Случайный лес строит несколько деревьев решений, используя случайные подвыборки данных и случайные подмножества признаков. Каждое дерево обучается независимо, и для каждого разделения узла выбирается случайное подмножество признаков. При предсказании модель усредняет результаты всех деревьев. Такая техника называется Бэггинг.

3. Градиентный бустинг.

Бустинг - это техника построения ансамблей, в которой предсказатели построены не независимо (как в случае со случайным лесом),

а последовательно. Идея заключается в том, что следующая модель будет учиться на ошибках предыдущей.

Градиентный бустинг использует метод градиентного спуска для оптимизации функции потерь. Он строит новую модель, минимизируя градиент функции потерь, что позволяет учитывать ошибки предыдущих моделей.

У градиентного бустинга много преимуществ: он способен моделировать сложные нелинейные зависимости в данных, хорошо справляется с выбросами и шумом, способен работать с различными типами данных, включая числовые, категориальные и текстовые. Но так же, как и у любого алгоритма, есть свои недостатки: обучение и применение градиентного бустинга может быть времязатратным процессом, градиентный бустинг может быть склонен к переобучению, особенно при недостаточном количестве данных или неоптимальной настройке параметров.

В работе рассматриваются следующие модели : **CatBoostRegressor**, **LightGBMRegression**.

У каждой из них есть преимущества, которые будут полезны для задачи:

- (a) CatBoostRegressor - эффективно обрабатывает разреженные данные, что является преимуществом в задачах с большими и разреженными наборами данных, например, в области рекомендательных систем или анализа текста.
- (b) LightGBMRegression - применяет стратегию Leaf-Wise роста деревьев, которая отличается от традиционной стратегии Level-Wise. Вместо деления узла на всех уровнях одновременно, она делает разделение наиболее информативных узлов, что может привести к более компактным и быстрорастущим деревьям. Также модель автоматически обрабатывает категориальные признаки, применяя метод кодирования, что упрощает использование таких признаков в модели.

2.4 Критерии качества модели

Проверка качества оцененной регрессионной модели может проводиться по следующим критериям:

1. Среднеквадратическая ошибка (Mean Squared Error)

MSE является одним из наиболее распространенных критериев качества для регрессии. Он вычисляется как среднее значение квадратов разностей между предсказанными значениями модели и истинными значениями целевой переменной:

$$MSE = \frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

2. Среднеквадратическое отклонение (Root Mean Squared Error)

RMSE является квадратным корнем из MSE:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Большой плюс этой функции в том, что она имеет ту же размерность, что и целевая переменная, что облегчает интерпретацию показателя ошибки.

3. Средняя абсолютная ошибка (Mean Absolute Error)

MAE - это среднее абсолютных отклонений прогнозных значений от реальных значений.

$$MAE = \frac{1}{N} \sum_{i=1}^n |y_i - \hat{y}_i|$$

MAE не учитывает величину отклонения и может быть более подходящей метрикой, если важно избежать больших ошибок.

4. Средняя ошибка аппроксимации (Mean Absolute Percentage Error)

MAPE измеряет среднюю абсолютную процентную ошибку между прогнозируемыми значениями модели и фактическими значениями целевой переменной:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100\%$$

Ошибка аппроксимации не более 8–12% свидетельствует о хорошем качестве модели.

5. Коэффициент детерминации (R-squared)

R^2 - это коэффициент детерминации, который есть мера объясненной дисперсии модели относительно общей дисперсии целевой переменной:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2},$$

где \bar{y}_i - среднее фактических значений.

В формуле числитель дроби представляет собой сумму квадратов остатков модели, а знаменатель - общую сумму квадратов отклонений фактических значений от их среднего.

Таким образом R^2 показывает, как хорошо модель подходит для данных. В отличие от предыдущих критериев, для которых качество определяет близость значения к нулю, R^2 чаще всего принимает значение от 0 до 1, где 1 означает, что модель абсолютно точна, а 0 - что модель не предсказывает ничего лучше, чем простое среднее.

В редких случаях R^2 может быть отрицательным, если модель хуже чем простая модель, которая всегда предсказывает среднее значение.

Это происходит, когда модель сильно недообучена и не может описать данные.

2.5 Способы подбора гиперпараметров

Гиперпараметры модели - это параметры, которые определяют саму структуру модели или способ ее обучения. Они не могут быть автоматически определены в процессе обучения модели, а должны быть явно заданы пользователем перед обучением. Гиперпараметры влияют на поведение модели и ее способность обобщать данные.

Цель состоит в том, чтобы найти комбинацию гиперпараметров, которая обеспечивает наилучшую производительность модели на проверочном наборе данных.

Выбор оптимальных значений гиперпараметров модели является важной задачей при разработке и настройке моделей машинного обучения. Это может быть выполнено с использованием различных методов. В данной работе будут рассмотрены некоторые из них:

1. Сеточный поиск (Grid Search)

Это метод перебора комбинаций гиперпараметров модели, путем создания сетки всех возможных комбинаций их значений.

Для каждого гиперпараметра необходимо указать несколько значений и Grid Search перебирает всевозможные комбинации, чтобы определить оптимальные значения. Он оценивает модель для каждой комбинации гиперпараметров с использованием кросс-валидации и выбирает лучшую комбинацию на основе заданной метрики оценки.

Grid Search является простым и понятным подходом к подбору гиперпараметров, который не требует сложной настройки или дополнительных алгоритмов. Однако, при большом пространстве поиска и большом количестве гиперпараметров данный метод может быть очень затратным по вычислительным ресурсам и времени.

2. Случайный поиск (Random Search)

В отличие от Grid Search, где перебираются все возможные комбинации, Random Search выбирает случайные комбинации гиперпараметров из заданного пространства.

Данный метод позволяет более эффективно исследовать большие пространства гиперпараметров, особенно если некоторые из них имеют меньшую важность.

3. Байесовская оптимизация (Bayesian Optimization)

Это метод оптимизации гиперпараметров модели, который использует байесовский подход. Вместо перебора всех возможных комбинаций гиперпараметров, Bayesian Optimization строит модель для улучшения целевой функции (например, значение метрики качества) и принимает решения о следующих наборах гиперпараметров, исходя из предыдущих наблюдений.

Bayesian Optimization обновляет свое априорное знание о функции потерь после каждого наблюдения, что позволяет адаптироваться к предыдущим результатам и сосредоточиться на более перспективных областях пространства гиперпараметров. Также метод может эффективно работать с шумными и нелинейными функциями потерь. Но, в сравнении с простыми методами поиска, Bayesian Optimization требует больше вычислительных ресурсов.

В данной работе будут настраиваться гиперпараметры для следующих моделей:

1. DecisionTreeRegression

Имеет следующие гиперпараметры:

- criterion

Критерий, используемый для измерения качества разделения. Возможные значения: "mse" (среднеквадратичная ошибка) или "mae" (средняя абсолютная ошибка).

- max_depth

Максимальная глубина дерева. Ограничение глубины помогает предотвратить переобучение модели.

- `min_samples_split`

Минимальное количество образцов, необходимых для разделения узла.

- `min_samples_leaf`

Минимальное количество образцов, необходимых для образования листового узла. Если число образцов в узле меньше указанного значения, то создание нового узла будет прекращено.

- `max_features`

Количество признаков, рассматриваемых при поиске наилучшего разделения. Можно указать число признаков или процент от общего числа признаков.

2. RandomForestRegression

Так как RandomForestRegression основывается на комбинировании множества деревьев решений, все гиперпараметры, рассмотренные выше, можно подбирать и для данной модели.

Дополнительно можно подобрать `n_estimators` - это количество деревьев в лесу. Чем больше деревьев, тем более стабильная и точная модель может быть построена, но с увеличением числа деревьев увеличивается время обучения.

3. CatBoostRegressor

У CatBoostRegressor также есть несколько гиперпараметров, которые можно настроить для оптимизации модели:

- `learning_rate`

Скорость обучения модели. Определяет вклад каждого дерева в итоговое предсказание. Более низкое значение обычно требует большего количества деревьев для достижения хороших результатов.

- depth

Максимальная глубина дерева. Определяет количество уровней разделения в каждом дереве. Более глубокие деревья могут захватывать более сложные взаимосвязи в данных, но могут также привести к переобучению.

- l2_leaf_reg

Коэффициент регуляризации L2 для листьев дерева. Используется для контроля сложности модели и предотвращения переобучения.

- iterations

Количество итераций (деревьев), которые следует построить. Большее количество итераций может улучшить точность модели, но также увеличивает время обучения.

Глава 3. Список библиотек Python

Список библиотек Python [22] - [30], которые были использованы для реализации всех этапов:

- Для чтения и предобработки: Numpy, Pandas, RE.

Numpy - библиотека для выполнения операций с массивами и математических вычислений.

Pandas - библиотека для работы с данными, предоставляющая высокоуровневые структуры данных и инструменты для обработки, фильтрации, анализа и визуализации данных.

RE (Regular Expressions) предоставляет функционал для работы с регулярными выражениями для поиска, извлечения и модификации подстрок на основе определенных шаблонов.

- Для предобработки текста: NLTK, Rymorphy2.

NLTK - библиотека для обработки естественного языка.

Rymorphy2 - библиотека для морфологического анализа русских слов. Она позволяет выполнять различные операции с русскими словами, такие как лемматизация, получение грамматической информации, склонение и прочие морфологические преобразования.

- Для графического и корреляционного анализа: Seaborn, Matplotlib, Wordcloud.

Matplotlib - это библиотека визуализации данных в Python. Она предоставляет широкий спектр возможностей для создания различных типов графиков, диаграмм и визуальных представлений данных.

Seaborn - это так же библиотека визуализации данных на основе Matplotlib. Seaborn является мощным инструментом для создания привлекательных и информативных графиков, и она широко используется для визуализации данных, анализа и исследования.

Wordcloud - Библиотека, которая предоставляет функциональность для создания облака слов на основе текстовых данных.

- Для векторизации текстовых данных: `Collections`.

Collections - это модуль, предоставляющий альтернативные структуры данных и расширенные версии встроенных контейнеров данных, таких как списки, кортежи, словари и множества. Он предоставляет эффективные реализации контейнеров с дополнительными возможностями и функциональностью.

- Для обучения моделей: `Scikit-learn`, `CatBoost`, `LightGBM`.

Scikit-learn - библиотека машинного обучения с широким спектром алгоритмов и инструментов для классификации, регрессии, кластеризации, обработки текстов, извлечения признаков и многого другого.

CatBoost - это библиотека машинного обучения, разработанная компанией Yandex. Библиотека предоставляет реализацию алгоритма градиентного бустинга.

LightGBM - это библиотека машинного обучения, которая предоставляет реализацию алгоритма градиентного бустинга над деревьями. С помощью этой библиотеки будем задавать и обучать модель `LightGBMRegression`.

Глава 4. Сбор и подготовка данных к анализу

4.1 Сбор данных

Данные были собраны с сайта repetitors.info. Это сервис по подбору репетиторов с большой базой преподавателей (более 300 тысяч анкет). Для выгрузки данных был написан код на языке Python, где использовалась программная библиотека для управления браузерами - Selenium WebDriver [31] - [32]. Отдельно собирались анкеты из городов: Москва, Санкт-Петербург, Краснодар, Казань, Екатеринбург, Нижний Новгород, Ростов-на-Дону. Также для сбора использовалась фильтрация с сайта по предмету и опыту работы и эти значения записывались в отдельные факторы.

По итогу было собран датасет, размером 190284 записей, содержащий столбцы:

1. **Link** - ссылка на анкету
2. **Name** - ФИО репетитора
3. **Description** - полное описание анкеты
4. **Mark** - средняя оценка по отзывам и количество отзывов
5. **Subject** - предметы
6. **Experience** - опыт репетитора

	Link	Name	Description	Mark	Subject	Experience
0	https://msk.repetitors.info/repetitor/?p=Levac...	Левачев Сергей Михайлович	Выбрать Левачев Сергей Михайлович Репетитор по...	оценка 4,91 70 отзывов	химия	Преподаватель вуза
1	https://msk.repetitors.info/repetitor/?p=Tsyts...	Цыцарев Леонид Владимирович	Выбрать Цыцарев Леонид Владимирович Проводит д...	оценка 4,97 37 отзывов	математика	Небольшой опыт
2	https://msk.repetitors.info/repetitor/?p=TlebzuMP	Тлебзу Маргарита Петровна	Выбрать Тлебзу Маргарита Петровна Предметы: ан...	оценка 5,00 7 отзывов	английский язык	Средний опыт
3	https://msk.repetitors.info/repetitor/?p=KarasSR	Карась Светлана Ростиславовна	Выбрать Карась Светлана Ростиславовна Проводит...	оценка 4,88 16 отзывов	английский язык	Средний опыт
4	https://spb.repetitors.info/repetitor/?p=LapinaMA	Лапина Мария Андреевна	Выбрать Лапина Мария Андреевна Репетитор по ма...	оценка 5,00 4 отзыва	математика	Небольшой опыт

4.2 Предобработка данных

С собранными данными были проведены следующие этапы предварительной обработки:

1. Удалены дубликаты.
2. Удалены пропущенные значения.
3. Сформирован столбец city.

Анкетам из городов Санкт-Петербург и Москва присвоено значение Москва, остальным – Регион. Гистограмма распределения городов показана на рис. 1.

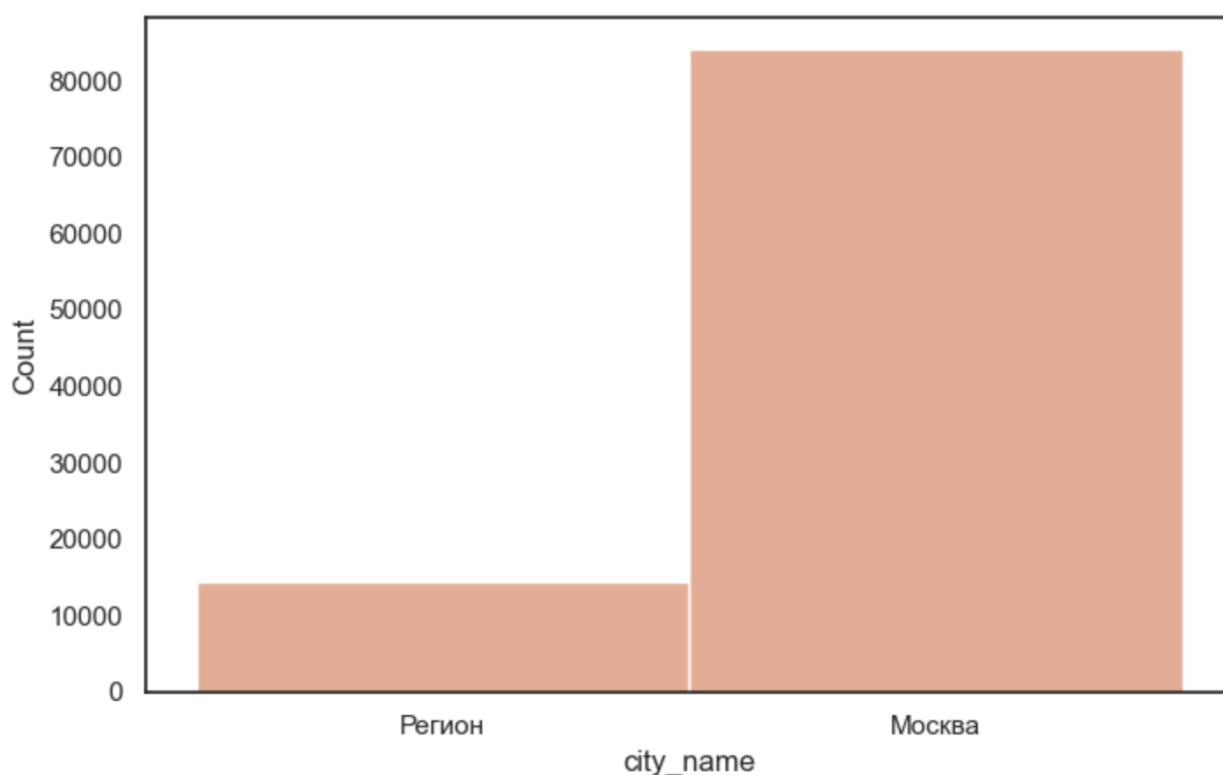


Рис. 1: Гистограмма распределения параметра city.

4. Преобразован столбец subject.

Предметы были объединены в 4 группы:

- (а) Естественно-научные: математика, информатика, физика, программирование, химия, биология

- (b) Гуманитарные: география, история, русский язык, экономика, обществознание, литература
- (c) Языки: английский, немецкий, французский, китайский, испанский, итальянский, японский
- (d) Начальные классы: начальная школа, подготовка к школе, музыка

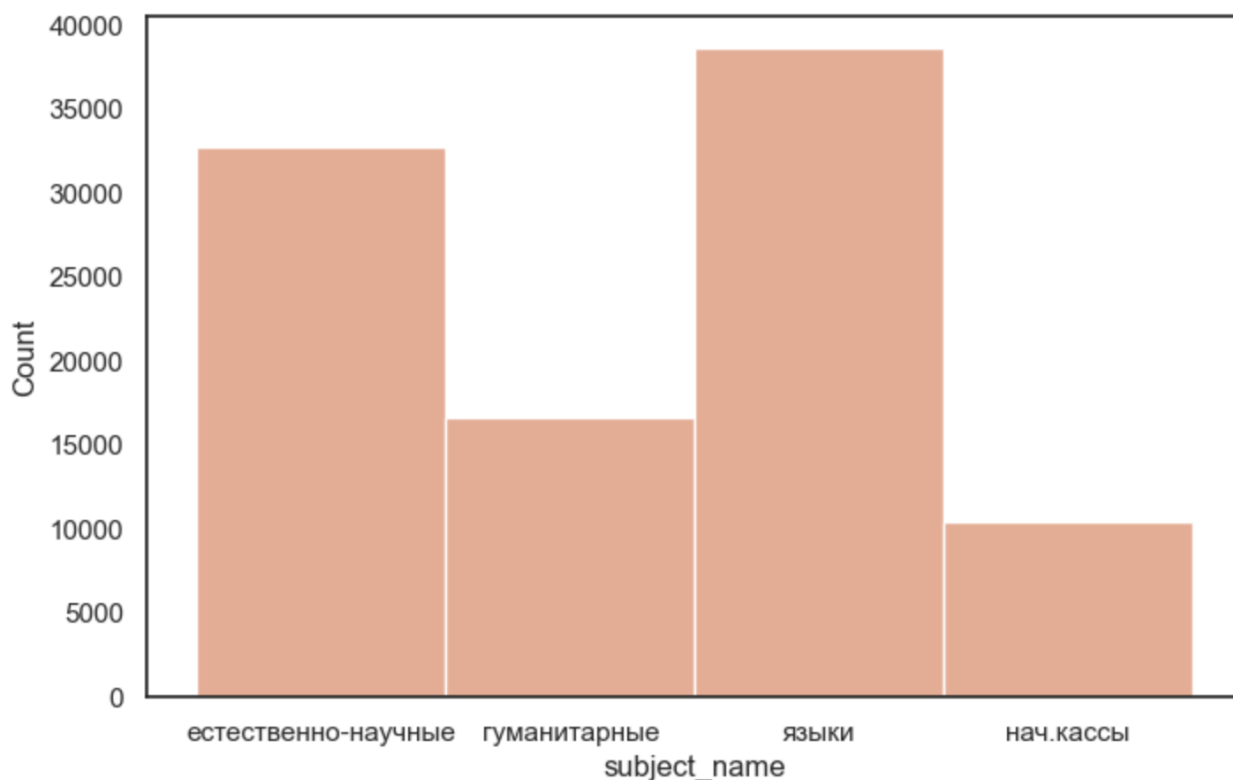


Рис. 2: Гистограмма распределения параметра subject.

5. Преобразован столбец experience.

Опыт был разбит на 3 группы:

- (a) Небольшой опыт
- (b) Средний опыт
- (c) Большой опыт: серьёзный опыт, преподаватель вуза, репетитор-эксперт, профессор, школьный учитель, преподаватель курсов.

Гистограмма распределения опыта показана на рис. [3](#).

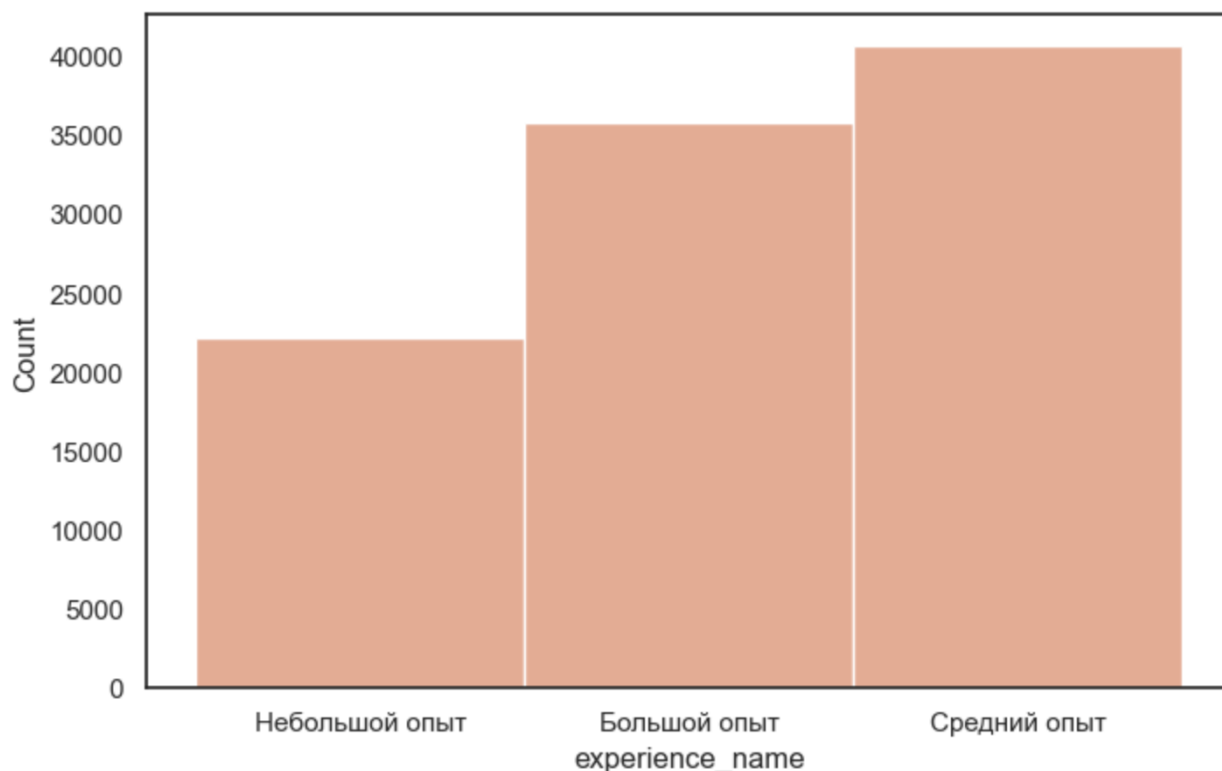


Рис. 3: Гистограмма распределения параметра experience.

6. Сформированы столбцы mark и review_count.

С помощью регулярных выражений для каждой анкеты в столбце mark было найдено количество отзывов и значение записано в отдельный столбец review_count, а в mark была записана только средняя оценка по отзывам. Если отзывы и, соответственно, оценки отсутствуют, им присваивались нулевые значения.

	link	name	description	mark	subject	experience	city	subject_name	experience_name	city_name	review_count
0	https://kzn.repetitors.info/repetitor?p=Ahmet...	Мулекова Ильвира Фаниловна	Выбрать Мулекова Ильвира Фаниловна Репетитор п...	5.0	3	3	1	языки	Большой опыт	Регион	4
1	https://spb.repetitors.info/repetitor?p=UdalovPP	Удалов Павел Павлович	Выбрать Удалов Павел Павлович Предметы: матема...	0.0	1	2	2	естественно-научные	Средний опыт	Москва	0

7. Столбцы city, subject, experience приведены в категориальную форму.

8. Сформирован результирующий признак price_avg.

Были оставлены только анкеты, в которых стоимость измеряется в «руб./ч.». Из столбца description, с помощью регулярных выражений,

был найден блок с ценами за занятия, в нем найдены и усреднены цены для каждого репетитора.

9. Удалены выбросы.

Проведена обработка выбросов для результирующего фактора `price_avg`. Построены боксплоты до и после удаления выбросов:

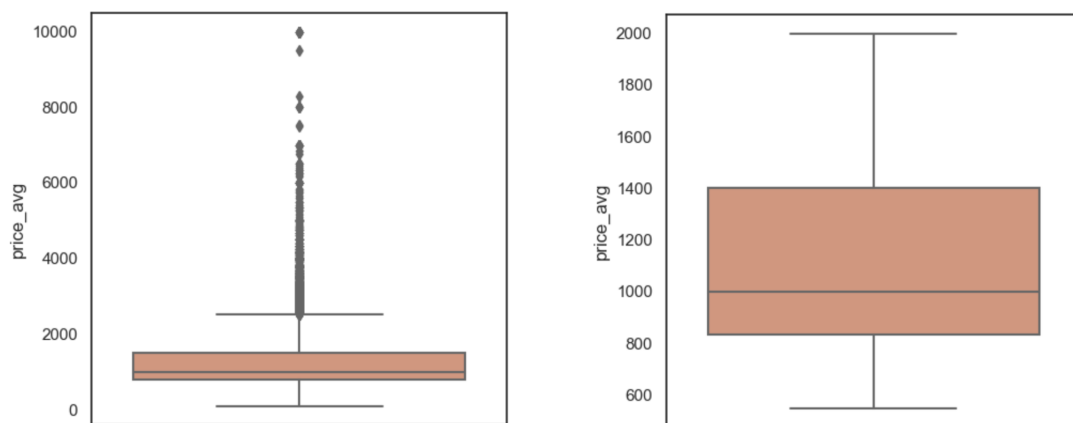


Рис. 4: Боксплот `price_avg` до и после удаления выбросов.

4.3 Предобработка текста

В рассматриваемом датасете после предобработки данных остался только один текстовый признак – `description`. В работе были проведены следующие этапы:

1. Токенизация
2. Приведение к нижнему регистру
3. Удаление знаков пунктуации, спец-символов, цифр
4. Удаление стоп-слов из готового русского словаря стоп-слов библиотеки `nltk`
5. Лемматизация
6. Стемминг

Пример работы кода:

До предобработки	После предобработки
Карначук Ирина Юрьевна Репетитор по русскому языку. Образование: • Воронежский государственный университет, филология, 2015 г. Опыт: • Опыт репетиторства — 13 лет. Учитель высшей категории. Действующий эксперт предметной комиссии ЕГЭ и ОГЭ по русскому языку. Знаю, как научить, чтобы выпускник получил желаемые баллы.	карначук ирина юриевич русский образование воронежский государственный университет филология опыт опыт репетиторство высокий категория действовать эксперт предметный комиссия егэ огэ русский знать научить выпускник получить желаемый балл

Далее был составлен словарь уникальных слов. Длина словаря - 66 тысяч. Среди самых частовстречающихся были такие слова, как: опыт, образование, английский, выезд, школа, дистанционный, государственный, математика, педагогический, егэ.

Глава 5. Анализ данных

5.1 Графический анализ

Рассмотрим графики изменения результирующего признака от факторов.

Видно, что город влияет на стоимость достаточно значимо (рис.5(a)). С изменением опыта цены меняются не сильно, но тенденция видна (рис.5(b)).

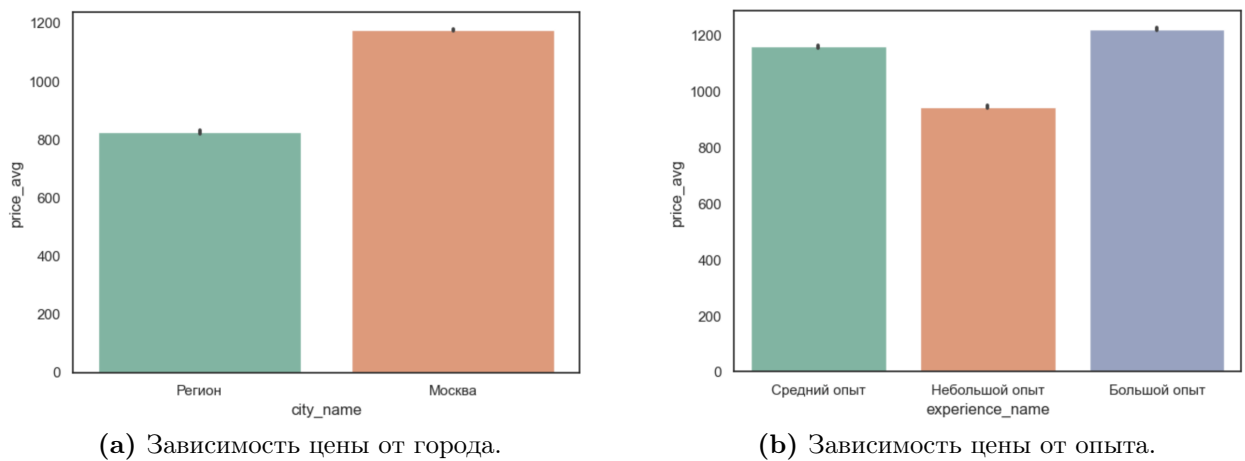


Рис. 5

Как видно на рис.6(a), предмет репетитора практически не влияет на стоимость занятий. Если рассматривать график отношения стоимости от количества отзывов, то здесь все достаточно логично распределено – чем больше отзывов, тем выше средняя стоимость занятий (рис.6(b)).

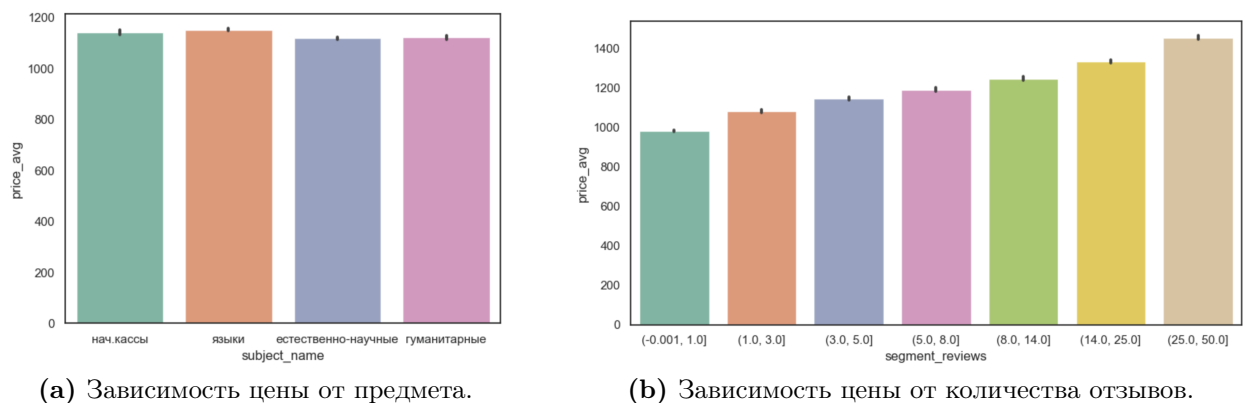


Рис. 6

Интересное распределение показала средняя оценка репетитора - до

последнего сегмента цена логично увеличивается с увеличением оценки, но последний столбец выделяется (рис. 6(a)).

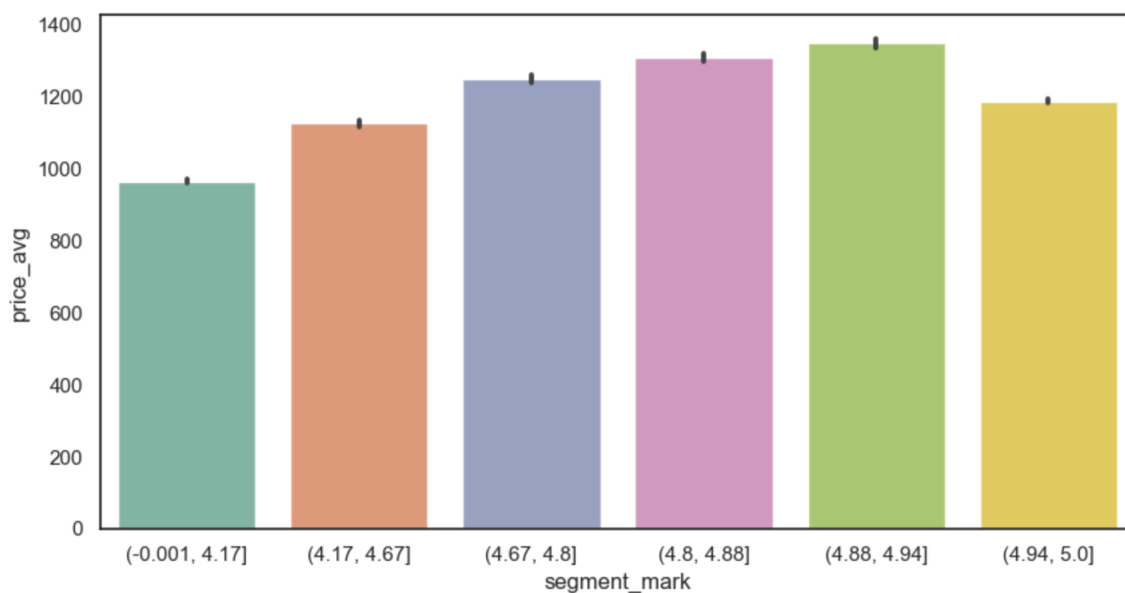


Рис. 7: Зависимость цены от средней оценки.

Это возникает потому, что существует большое количество начинающих репетиторов, у которых немного отзывов, но все они отличные. Цена занятий небольшая, потому что опыта у репетитора пока недостаточно.

5.2 Факторный анализ

Построена корреляционная матрица (рис. 8), по результатам которой был сделан вывод, что есть признаки, такие как предметы (math, hum, language, elementary), которые очень незначительно влияют на изменение стоимости.

Поэтому в работе будет рассматриваться датасеты, включающие эти признаки и не включающие.

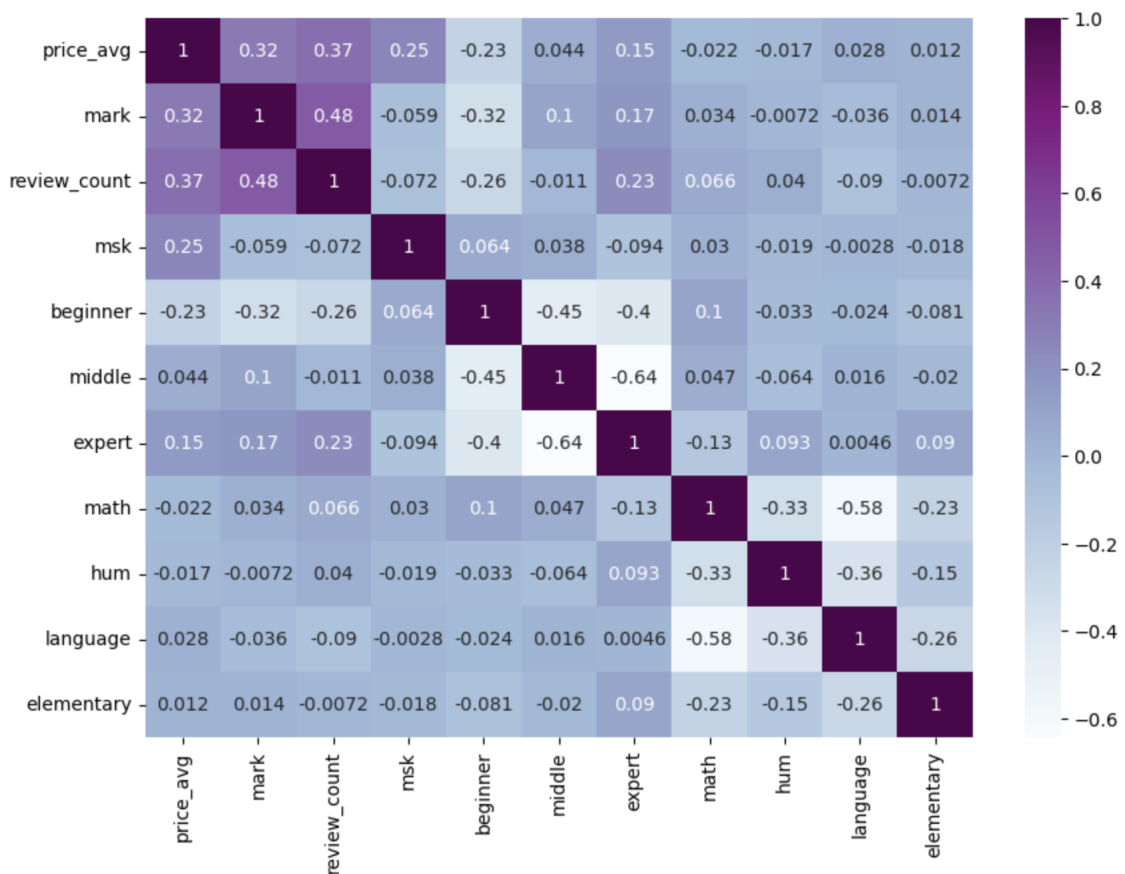


Рис. 8: Корреляционная матрица.

5.3 Обучение моделей

В работе будут анализироваться следующие модели:

1. LinearRegression
2. Ridge
3. Lasso
4. DecisionTreeRegressor
5. RandomForestRegressor
6. LGBMRegressor
7. CatBoostRegressor

Рассмотрим датасет со всеми входными параметрами. Для векторизации текста используем мешок слов, используя функцию `CountVectorizer()`

библиотеки `sklearn`. Чтобы ускорить обучение и предотвратить переобучение моделей, сократим пространство признаков, используя дополнительный параметр `max_features`. `max_features = N` оставляет топ-N наиболее часто встречающихся слов и векторизует только эти слова, игнорируя менее частые. Зададим `max_features = 3000` (в рамках данного этапа были также рассмотрены значения параметра 2000 и 4000, но для обоих значений модели показали результаты хуже, что может говорить о недообучении для случая 2000 параметров и возможном переобучении для 4000).

В качестве метрик оценок точности будем рассматривать MAE, RMSE и R^2 . MAE и RMSE рассматриваются, чтобы значение можно было интерпретировать в рублях. Разница между этими двумя метриками покажет, есть ли в данных выбросы или ошибки, так как RMSE чувствителен к выбросам, в отличие от MAE. Дополнительно рассмотрим значение R^2 , чтобы посмотреть, сколько процентов общей вариации результативного признака объясняет модель.

Результаты работы моделей представлены в Таблице 1.

Таблица 1: Таблица результатов работы моделей

	RMSE	MAE	R^2
LinearRegression	297.0	231.2	0.4
Ridge	296.9	231.2	0.4
Lasso	296.5	232.0	0.4
DecisionTreeRegressor	305.6	238.7	0.3
RandomForestRegressor	302.4	236.4	0.3
LGBMRegressor	284.8	221.3	0.4
CatBoostRegressor	284.0	220.2	0.4

По результатам видно, что значение R^2 для всех построенных моделей достаточно низкое, зато для каждой модели значение MAE около 230. Это показывает, что в среднем прогнозы модели отклоняются от истинных значений на 230 рублей.

Лучше всего обучились модели, основанные на градиентном бустинге. Самая лучшая модель по всем критериям - CatBoostRegressor. Рассмотрим ее подробнее.

Отношение RMSE к MAE равно примерно 1.29, что говорит о том, что RMSE незначительно выше, чем MAE. Значит в датасете нет явных выбросов или больших ошибок. Проанализируем распределение ошибок (рис.9).

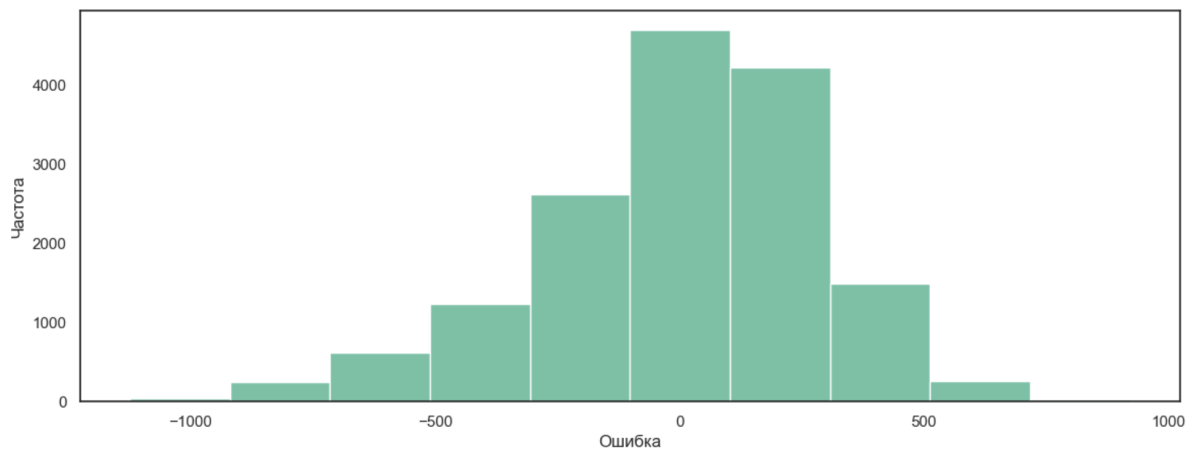


Рис. 9: График распределения ошибок.

Видно ярковыраженное нормальное распределение остатков, что говорит об отсутствии ошибок и выбросов в данных.

График предсказанных и истинных значений отображается на рис.10.

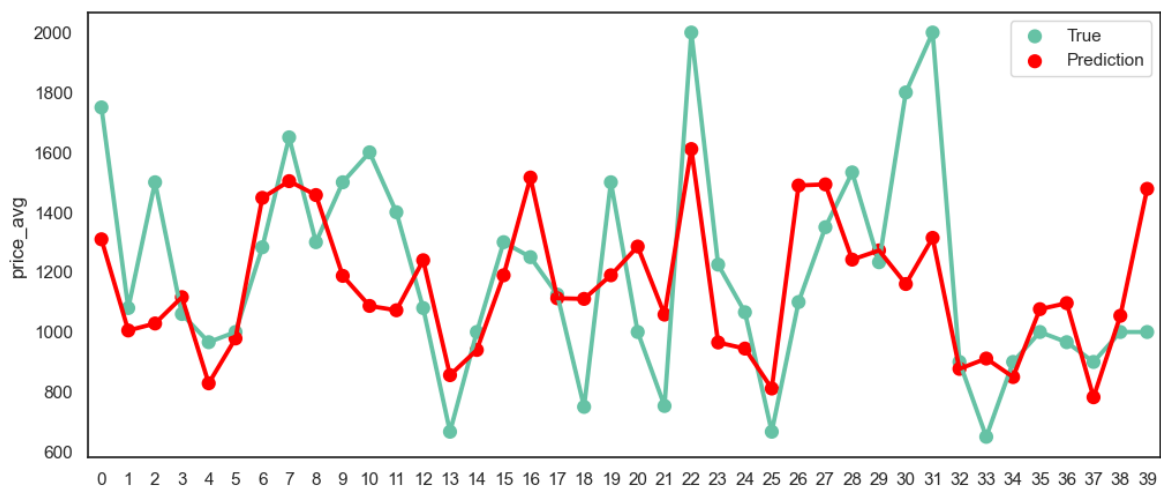


Рис. 10: График предсказанных и истинных значений.

По графику видно, что предсказание модели в большинстве случаев повторяет динамику истинных значений. В некоторых моментах модель точно определяет ставку.

Рассмотрим примеры (Таблица 2), разность между предсказываемым и реальным значением по модулю больше 500. Из таблицы видно, что, если

опыт большой (expert), оценка высокая и много отзывов, то модель корректирует в большую сторону. Если же опыт начинающий (beginner) или нет оценок вовсе - модель предсказывает оценку ниже.

Таблица 2: Таблица корректировки

Описание		
Проводит дистанционные занятия. ЮФУ, физико-математический факультет, специальность – учитель математики и информатики (2003 г.). Преподаватель техникума, опыт работы – 16 лет. Частный репетитор с 2000 года. Сертификация по математике пройдена.	Оценка Отзывы Опыт Предмет Город Истина Предсказ. значение	4.79 50 expert math region 560 1108
Образование: • МПГУ, факультет иностранных языков, специальность – преподаватель английского языка, 1999–2004 гг. Опыт: • Репетиторство — 17 лет. • Преподаватель английского языка в МГУ им. М.В. Ломоносова с 2008 года — 11 лет. • Педагогический стаж – 8 лет (МГУ им. М.В. Ломоносова, МГГУ им. М.А. Шолохова). Готовлю к ОГЭ и ЕГЭ и др. экзаменам. Минимум 2 раза в год езжу в Великобританию. Имею сертификат преподавателя бизнес английского.	Оценка Отзывы Опыт Предмет Город Истина Предсказ. значение	5.00 10 expert language msk 1200 1709
Образование: • Окончила Московский городской педагогический университет, Институт математики, информатики и естественных наук, направление подготовки – педагогическое образование с двумя профилями подготовки: биология, иностранный язык. Быстро нахожу контакт с детьми, настраиваю их на работу. Ответственно отношусь к работе с детьми! Даже если ребенку не очень нравится предмет, успеваемость держим на «4».	Оценка Отзывы Опыт Предмет Город Истина Предсказ. значение	4.50 2.0 beginner math msk 1700 825
Образование: РАМ им. Гнесиных, фортепианный факультет, специальность – фортепиано, 2 курс. Астраханский музыкальный колледж им. М.П. Мусоргского, отделение фортепиано, специальность – фортепиано. Стипендиат правительства Астраханской области. Опыт репетиторства – с 2013 года.	Оценка Отзывы Опыт Предмет Город Истина Предсказ. значение	0.00 0 beginner elementary msk 1500 866

Найдем значения важности каждого признака с помощью функции `get_feature_importance()`. Сформирована таблица первых важных факторов - Таблица 6. Здесь слова, заключенные в кавычки - это слова из описания репетитора.

Видно, что очень сильно влияют такие признаки, как: город, количе-

Таблица 3: Таблица важности признаков

Признак	Важность
msk	15.11
review_count	11.21
beginner	7.48
mark	7.29
"дистанционный"	4.26
expert	2.72
"достижение"	1.44
"мгу"	1.40
"выезд"	1.18
"курс"	1.11
опыт	0.91
"носитель"	0.88
"егэ"	0.64

ство отзывов, опыт, оценка, а также некоторые слова в описании: дистанционный, достижение, мгу, егэ и др. Также можно убедиться в том, что предметы имеют слабое влияние на предсказание стоимости занятия.

Теперь рассмотрим векторизацию TF-IDF функцией. Также воспользуемся параметром `max_features = 3000`. Результаты работы моделей можно посмотреть в Таблице 4.

Таблица 4: Таблица результатов работы моделей

	RMSE	MAE	R^2
LinearRegression	294.4	231.7	0.4
Ridge	292.3	230.2	0.4
Lasso	312.5	246.8	0.3
DecisionTreeRegressor	304.1	239.8	0.3
RandomForestRegressor	299.1	235.8	0.3
LGBMRegressor	284.2	222.8	0.4
CatBoostRegressor	283.7	222.5	0.4

Результаты оказались немного хуже, чем в случае простой векторизации мешком слов. Это может быть из-за того, что для слов, которые имели важную значимость при обучении, были посчитаны слишком ма-

ленькие значения весов и они не вошли в первые 3000 признаков. В данной же работе не нужно обращать внимание на важность слова в контексте всей коллекции документов. Достаточно выделять наличие слов, которые могут повлиять на ставку.

График предсказанных и истинных значений можно посмотреть на рис. 11. По графику видно, что значения предсказываются хуже.

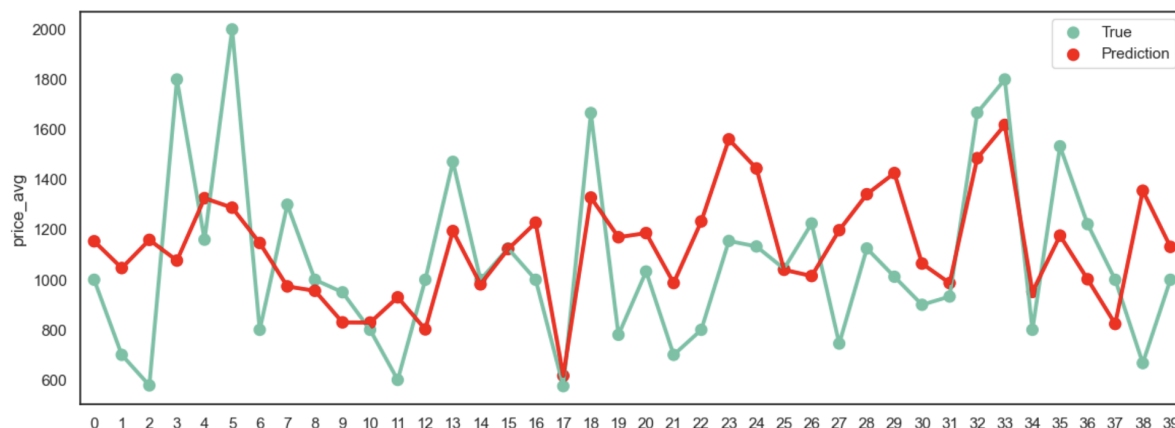


Рис. 11: График предсказанных и истинных значений.

Так как почти в каждом описании уже прописаны предметы и они, соответственно, учитываются при векторизации и обучении, то признаки, составленные во время преобработки данных, могут только помешать и возможно ухудшить результаты. Попробуем убрать 4 фактора, отвечающие за предметы: math, hum, language, elementary.

Результаты работы моделей можно посмотреть в Таблице 5.

Таблица 5: Таблица результатов работы моделей

	RMSE	MAE	R^2
LinearRegression	298.03	231.94	0.4
Ridge	297.81	231.80	0.4
Lasso	298.00	233.00	0.4
DecisionTreeRegressor	307.71	240.14	0.3
RandomForestRegressor	303.17	237.03	0.3
LGBMRegressor	286.66	222.91	0.4
CatBoostRegressor	286.19	222.06	0.4

Значения немного ухудшились, что говорит о том, что столицы все же вносят какой-то вклад в предсказание.

5.4 Корректировка параметров

На данном этапе будем корректировать параметры для некоторых моделей, с помощью метода случайного поиска. Этот метод называется `RandomizedSearchCV` и он реализован в библиотеке `sklearn`.

Начнем с моделей `DecisionTreeRegressor` и `RandomForestRegressor`. Эти модели строились в прошлой главе хуже всего, но деревья решений склонны переобучаться, поэтому для них особенно важно подбирать гиперпараметры.

Построим сетку гиперпараметров для **`DecisionTreeRegressor`** и запустим поиск лучшей четверки параметров по сетке.

```
1  params = {
2      'max_depth': np.arange(2, 30),
3      'min_samples_split': np.arange(2, 10),
4      'min_samples_leaf': np.arange(1, 10),
5      'max_features': [1.0, 'sqrt', 'log2', None]
6  }
```

Лучшие гиперпараметры и результаты работы лучшей модели представлены ниже.

```
1  best_params = {
2      'min_samples_split': 6,
3      'min_samples_leaf': 9,
4      'max_features': None,
5      'max_depth': 8,
6  }
7
8  MAE = 236.79
```

Как видно, MAE выше, чем у построенных ранее моделей деревьев решений, но все еще ниже MAE лучшей модели.

Сформирована таблица первых важных факторов - Таблица [6](#). Здесь слова, заключенные в кавычки - это слова из описания репетитора.

Таблица 6: Таблица важности признаков

Признак	Важность
review_count	0.43
msk	0.20
beginner	0.13
mark	0.07
"дистанционный"	0.07
expert	0.04
middle	0.01

По результатам `DecisionTreeRegressor`, самым важным фактором, влияющим на ставку, является количество отзывов.

Найдем лучшие гиперпараметры для **`RandomForestRegressor`**. Построим сетку гиперпараметров и запустим поиск лучшей четверки параметров по сетке.

```
1  params = {
2      'n_estimators': randint(100, 1000),
3      'max_depth': randint(5, 50),
4      'max_features': [1.0, 'sqrt', 'log2', None],
5      'min_samples_split': randint(2, 10),
6      'min_samples_leaf': randint(1, 10)
7  }
```

Лучшие гиперпараметры и результаты работы лучшей модели представлены ниже.

```
1  best_params = {
2      'n_estimators': 700,
3      'max_depth': 13,
4      'max_features': 'log2',
5      'min_samples_split': 5,
6      'min_samples_leaf': 2
7  }
8  MAE = 235.12
```


MAE так же выше, чем у построенных ранее моделей, но все еще ниже лучшей модели.

Найдем лучшие гиперпараметры для модели **CatBoostRegressor**. Построим сетку гиперпараметров и запустим поиск лучшей четверки параметров по сетке.

```
1  params = {
2      'learning_rate': [0.05, 0.1, 0.3],
3      'depth': [3, 5, 7],
4      'l2_leaf_reg': uniform(0, 10),
5      'iterations': [100, 200, 300]
6  }
```

Лучшие гиперпараметры и результаты работы лучшей модели представлены ниже.

```
1  best_params = {
2      'depth': 7,
3      'iterations': 300,
4      'l2_leaf_reg': 4.9,
5      'learning_rate': 0.3
6  }
7
8  MAE = 218.0
```

Заметно улучшить модель не получилось. Но все же это лучший результат MAE, который был достигнут в работе.

График предсказанных и истинных значений отображается на рис. [12](#).

Список важных параметров совпадает с Таблицей [7](#), но значения важности изменились: более значимости стали первые 3 признака (город, количество отзывов, начальный опыт), все остальные признаки имеют значение важности немного ниже.

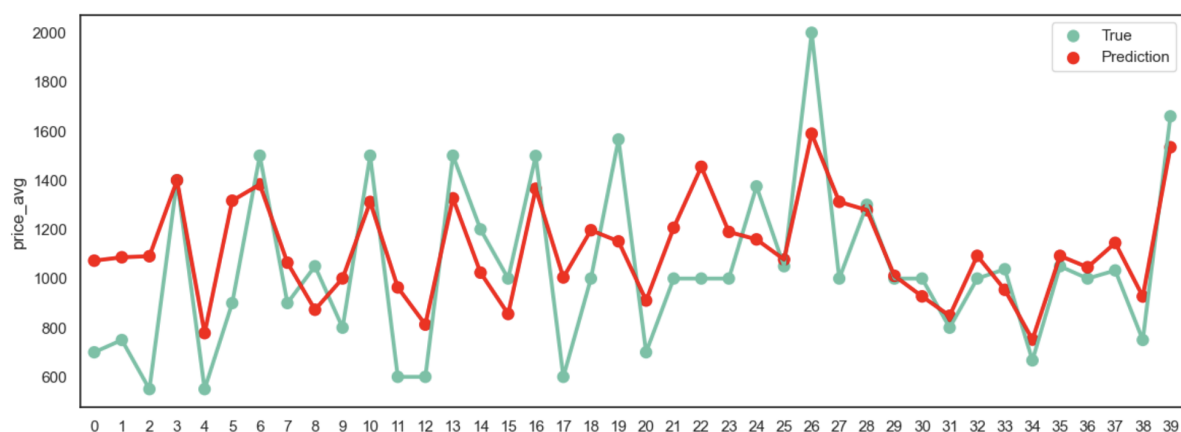


Рис. 12: График предсказанных и истинных значений.

Таблица 7: Таблица важности признаков

Признак	Важность
msk	16.01
review_count	12.61
beginner	8.11
mark	6.36
"дистанционный"	2.88
expert	2.07
"достижение"	1.32
"мгу"	1.28
"курс"	1.08
middle	1.03
"выезд"	0.91
"носитель"	0.88
"егэ"	0.68

В рамках данной работы были также построены модели и проанализированы результаты для данных без текстового параметра `description`, а также был рассмотрен случай уменьшения размерности данных методом главных компонент, но для обоих подходов результаты моделей показывали значения хуже.

Глава 6. Заключение

Для статистического анализа ставки репетитора были проделаны следующие задачи:

1. Изучен необходимый материал для сбора, предобработки и анализа данных.
2. Собраны необходимые данные и сформирована выборка.
3. Проведена предварительная обработка данных.
4. Проведены графический и факторный анализы.
5. Обучены модели: LinearRegression, Ridge, Lasso, DecisionTreeRegressor, RandomForestRegressor, LGBMRegressor, CatBoostRegressor.
6. Проведен этап подбора параметров для моделей: DecisionTreeRegressor, RandomForestRegressor, CatBoostRegressor.

Были рассмотрены методы векторизации: "Мешок слов" и TF-IDF. Результаты лучше показал первый подход.

По результатам анализа во всех случаях лучшей моделью оказалась модель градиентного бустинга **CatBoostRegressor**. Лучшие параметры и результаты модели представлены ниже. По результатам модели был сделан вывод о том, что в среднем модель ошибается в предсказании на 220 рублей, что показывает очень неплохой результат. Самыми важными параметрами модель посчитала:

1. Город, в котором репетитор преподает.
2. Количество у отзывов репетитора.
3. Опыт репетитора.
4. Средняя оценка.
5. Наличие таких слов в описании, как: дистанционный, достижение, мгу, курс, выезд, носитель, егэ.

Предмет не сыграл большой значимости в прогнозировании. Все остальные пункты из раздела "Факторы" подтвердили свою важность в обучении.

Список литературы

- [1] «Профессия Репетитор в МГУ»
- [2] Онлайн школа Skysmart «Сколько стоят занятия с репетитором»
- [3] «Как репетитору не продешевить с ценой урока»
- [4] Марианна Любарова (серфис Profi.ru) «Сколько стоят занятия с репетитором и от чего зависит цена?»
- [5] Регина Аюпова (сервис Skillbox) «Школьные репетиторы поднимают цену за свои услуги»
- [6] Данила Доманин (сервис Repit) «Как репетитору не продешевить с ценой урока»
- [7] Блог Учи.Дома «Стоимость занятий с репетитором»
- [8] Институт образования (ВШЭ) «Не школой единой – преподавание на онлайн-платформах и репетиторство стали полноценной частью профессии учителя»
- [9] Екатерина Шамаева (Тинькофф журнал) «Какие расценки у московских репетиторов»
- [10] Риа новости «Исследование показало, сколько зарабатывают репетиторы по языкам в России»
- [11] «Data Preprocessing in Data Mining»
- [12] «Предобработка данных (Data Preprocessing)»
- [13] Charles Wheelan «Naked Statistics: Stripping the Dread from the Data»
- [14] Шевляков Артём «Поиск выбросов и аномалий»
- [15] Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda «Applied Text Analysis with Python»

- [16] «Мультиклассовая классификация текста. Дисбаланс тренировочных данных и их генерация. Особенности взвешивания TF-IDF»
- [17] В. М. БУРЕ, Е. М. ПАРИЛИНА, А. А. СЕДАКОВ «МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ в R и Excel»
- [18] David M. Diez «An Introduction to Linear Regression Analysis»
- [19] «Быстрый градиентный бустинг с CatBoost»
- [20] «Как разработать ансамбль Light Gradient Boosted Machine (LightGBM)»
- [21] Roi Polanitzer «The Minimum Mean Absolute Error (MAE) Challenge»
- [22] «NumPy documentation»
- [23] Wes McKinney «pandas: powerful Python data analysis toolkit»
- [24] «NLTK documentation»
- [25] «Seaborn documentation»
- [26] «Matplotlib documentation»
- [27] «Scikit-learn documentation»
- [28] «CatBoost documentation»
- [29] «LightGBM documentation»
- [30] «Python documentation»
- [31] Baiju Muthukadan «Selenium with Python»
- [32] Гарри Персиваль «Python. Разработка на основе тестирования»