МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное общеобразовательное учреждение высшего образования

«Липецкий государственный педагогический университет имени П.П. Семенова-Тян-Шанского»

кафедра математики и физики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Бондаренко Вадима Олесьевича

Направление подготовки 01.03.02 Прикладная математика и информатика

Направленность: Прикладная математика и информатика в экономике

Тема: Применение алгоритмов машинного обучения для анализа рынка недвижимости

Утвержден
приказом ректора университета
№ 622 от 10 ноября 2023 г.

Руководитель: канд. физ.-мат. наук, доцент Калитвин В.А. _____ 10 июня 2024 г.	Студент гр. 03-ОФО-ЕМТ-ПМИ 2020 Бондаренко В.О. _____ 10 июня 2024 г.

ДОПУСКАЕТСЯ К ЗАЩИТЕ в ГЭК

заведующий кафедрой
канд. пед. наук., доцент _____ Жигаленко С.Г.

Липецк 2024 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 3
ГЛАВА I. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА РЫНКА НЕДВИЖИМОСТИ. 5
1.1 Типы машинного обучения 6
1.1.1 Обучение с учителем 7
1.1.2 Обучение без учителя 9
1.2 Методы машинного обучения 11
1.2.1 Метод k-средних 11
1.2.2 Ансамблевые методы регрессии: случайные леса и градиентный бустинг 12
1.3 Характеристика рынка недвижимости г. Липецка 15
ГЛАВА 2. ПОСТРОЕНИЕ МОДЕЛЕЙ ДЛЯ АНАЛИЗА РЫНКА НЕДВИЖИМОСТИ 18
2.1 Выбор инструментов 18
2.2 Сбор и обработка данных 20
2.3 Кластерный анализ 23
2.5 Регрессионный анализ 34
2.5.1 Линейная регрессия 36
2.5.2 Случайный лес 39
2.5.3 Градиентный бустинг 44
ЗАКЛЮЧЕНИЕ 50
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 51

ВВЕДЕНИЕ

В настоящее время применение методов машинного обучения играет важную роль в исследованиях, включая анализ рынка недвижимости. Эта технология обеспечивает возможность более глубокого и детального изучения динамического и сложного рынка недвижимости. В условиях быстро меняющейся экономической среды и усиленной конкуренции на рынке недвижимости применение методов машинного обучения приобретает стратегическое значение. Алгоритмы машинного обучения способны эффективно обрабатывать большие объемы данных, выявлять скрытые закономерности и предсказывать тенденции. В контексте рынка недвижимости это открывает новые перспективы для точного прогнозирования цен, выявления потенциальных инвестиционных возможностей и оптимизации стратегий управления недвижимостью.

Объект исследования – применение методов машинного обучения в анализе рынка недвижимости.

Предмет исследования – анализ и разработка алгоритмов машинного обучения, которые могут быть эффективно использованы для прогнозирования цен на недвижимость и кластеризации по стоимости и расположению объектов.

Цель работы: изучение методов машинного обучения и их применение для анализа вторичного рынка недвижимости в г. Липецк.

Для достижения поставленной цели необходимо решить следующие задачи:

1) изучить теоретические аспекты машинного обучения;
2) собрать и обработать данные;
3) реализовать и оценить модели;
4) сформулировать выводы по результатам исследования.

Методологическая основа:

1) Изучение литературы по методам машинного обучения и их применению в анализе рынка недвижимости.

2) Сбор данных о вторичном рынке недвижимости в г. Липецк из базы данных, предварительная обработка данных.
3) Разработка моделей машинного обучения для прогнозирования цен на недвижимость и кластеризации объектов по стоимости и расположению.
4) Оценка производительности моделей с использованием метрик точности.

Структура дипломной работы: введение, основная часть, заключение, список использованных источников.

ГЛАВА I. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА РЫНКА НЕДВИЖИМОСТИ

В нынешнем мире, машинное обучение, как практическое применение и наука об алгоритмах, которое своей компьютерной мощью может превратить большие объёмы данных в знания, благодаря библиотекам с открытом кодом, которые были разработаны в последние годы [16].

В главе будут раскрыты следующе темы:

- Типы машинного обучения;
- Методы машинного обучения;
- Характеристика рынка недвижимости г. Липецка [16].

Современные технологии предоставляют нам обширный ресурс - обильный поток структурированных и неструктурированных данных. В данном контексте машинное обучение, выступает как ключевая область развития, возникшая во второй половине двадцатого века в рамках искусственного интеллекта. Оно представляет собой сферу, где самообучающие алгоритмы позволяют извлекать знания из данных с целью формирования прогнозов.

В отличие от традиционных методов, требующих ручного вывода правил и создания моделей на основе анализа объемных данных, машинное обучение предлагает более эффективный подход для сбора знаний из данных. Этот процесс постепенно повышает эффективность прогнозирующих моделей и обеспечивает принятие решений, опираясь на данные.

В наше время важность машинного обучения выходит за пределы компьютерных наук и активно влияет на повседневную жизнь. Машинное обучение становится неотъемлемой частью нашего опыта, обеспечивая более эффективные решения и прогнозы, основанные на обширном объеме доступной информации [16].

1.1 Типы машинного обучения

В данном разделе мы рассмотрим два типа машинного обучения: обучение с учителем (supervised learning), обучение без учителя (unsupervised learning). На (рис. 1), показаны фундаментальные отличия между двумя типами машинного обучения [15].

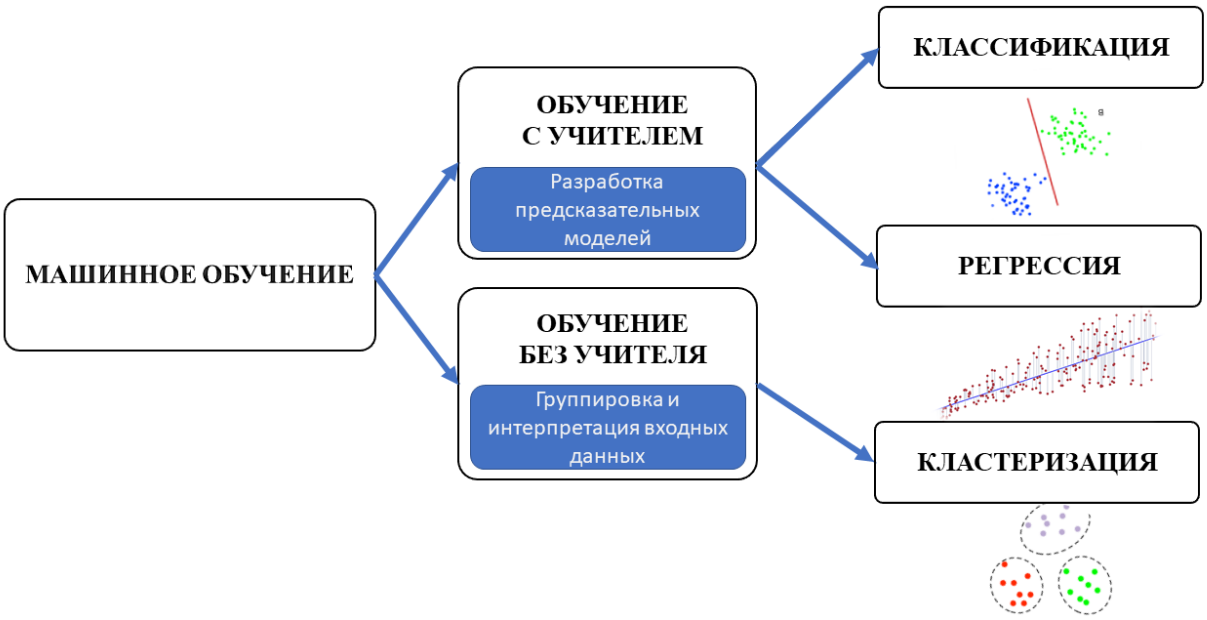

Рисунок 1 – Типы машинного обучения

На (рис. 1), продемонстрированы 2 типа машинного обучения, есть ещё третий тип, так называемый: «обучение с подкреплением (reinforcement learning)», то есть предметная область, в которой может применяться данный подход, является доменом, где модель взаимодействует с окружающей средой, принимая решения и получая обратную связь на основе своих действий. В этом разделе мы поговорим про два типа машинного обучения, а именно обучение с учителем и без учителя, поэтому обучение с подкреплением не будет использоваться, поскольку в данном контексте нам не требуется модель, взаимодействующую с окружающей средой и принимающую решения на основе обратной связи [16].

1.1.1 Обучение с учителем

В обучении с учителем два основных подхода - регрессия и классификация. Регрессия применяется, когда цель состоит в прогнозировании непрерывной переменной, в то время как классификация используется для определения принадлежности объекта к заданным классам.

Главная цель обучения с учителем заключается в настройке модели на помеченных обучающих данных, что позволяет модели делать прогнозы на новых, ранее не встречавшихся данных. Термин "с учителем" относится к наличию набора обучающих примеров, где желаемые выходные данные уже известны.

На (рис. 2), представлена типичная последовательность действий с учителем, где помеченные метки обучающих данных передаются алгоритму машинного обучения для подгонки к прогнозирующей модели, которая может вырабатывать прогнозы на новых непомеченных входных данных.

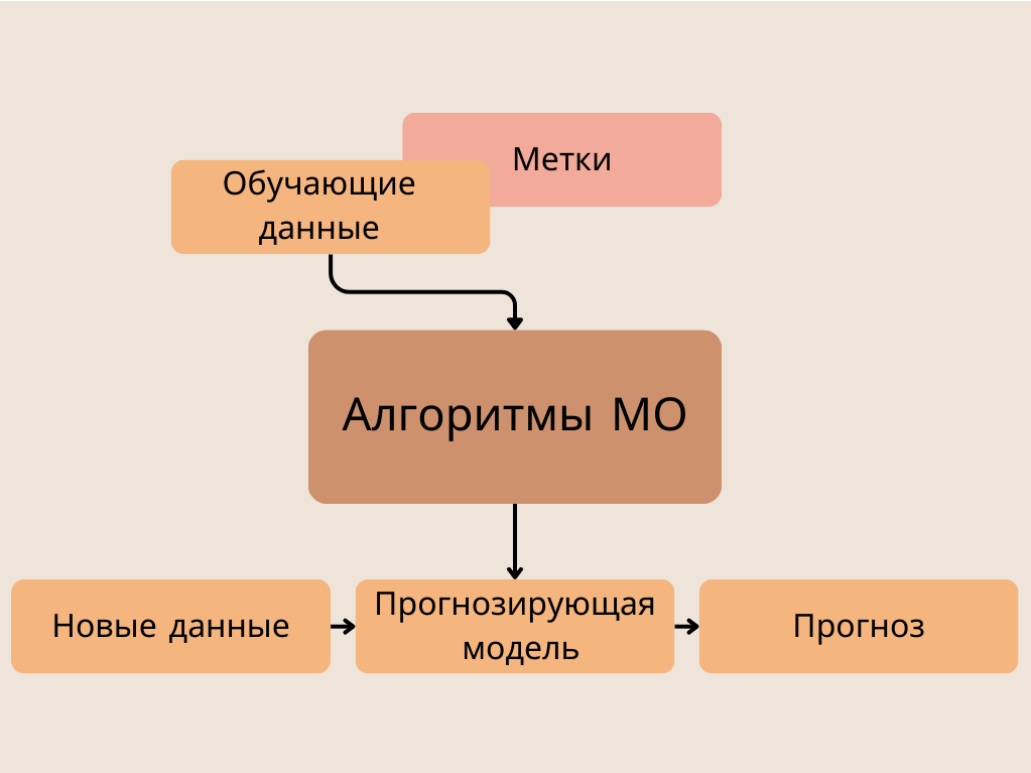

Рисунок 2 – Схема действий при обучении с учителем

Обучение с учителем включает в себя задачи классификации и регрессии. В задаче классификации, где рассматриваются метки дискретных классов, модель стремится присвоить объектам конкретные категории, как показано на (рис. 3). На (рис. 3) продемонстрирована концепция задачи двоичной классификации, где набор данных оказывается двумерным, то есть с каждым образцом ассоциированы два значения: x_1 и x_2.

В прочем, представление классов в наборе меток не обязательно должно быть ограничено двоичным характером. Модель, прошедшая обучение по методу учителя, может присваивать новому неотмеченному экземпляру любую метку класса, присутствующую в обучающем наборе данных.

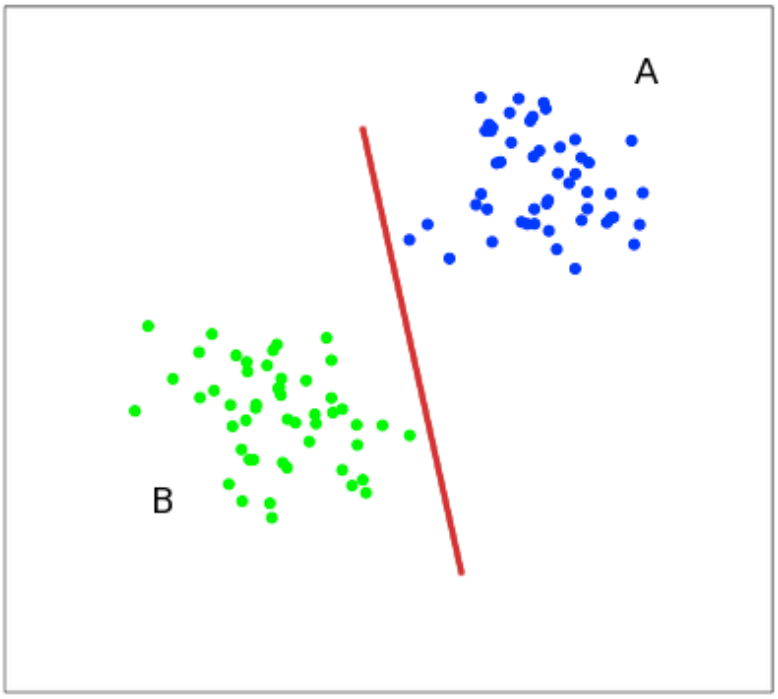

Рисунок 3 – Концепция задачи двоичной классификации

В регрессии, другой подкатегории обучения с учителем, модель настраивается для предсказания непрерывных значений, обеспечивая более точные прогнозы для числовых выходных данных, как показано на (рис. 4). На (рис. 4) продемонстрирована концепция линейной регрессии. Для переменной x и целевой переменной y мы проводим подгонку прямой линии к этим данным с целью минимизации расстояния, обычно используемого в виде среднеквадратичного отклонения, между точками данных и

аппроксимированной линией. После этого мы можем использовать определенные на основе имеющихся данных значения свободного члена и наклона для прогнозирования целевой переменной новых данных [16].

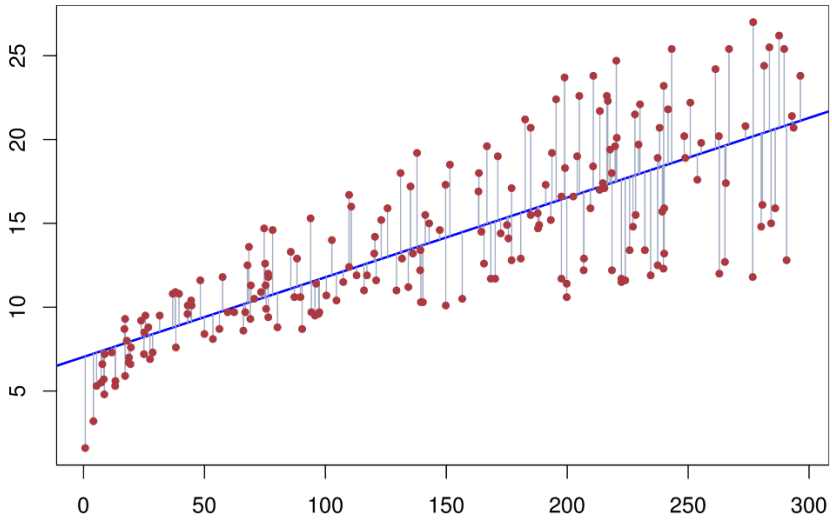

Рисунок 4 – Концепция линейной регрессии

1.1.2 Обучение без учителя

В процессе обучения с учителем мы заранее знаем правильные ответы, когда обучаем модель. В случае обучения с подкреплением мы определяем, как оценивать действия агента через награды. Однако, при обучении без учителя, мы оперируем непомеченными данными или данными неизвестной структуры. Применение методов обучения без учителя позволяет исследовать структуру данных, извлекая значимую информацию, не зависящую от известной целевой переменной или функции награды.

В контексте обучения без учителя, обычно выделяют два ключевых метода, активно применяемых: кластеризация и снижение размерности. Суть кластеризации заключается в исследовании данных с целью выявления схожих между собой объектов и их группировке в кластеры, как показано на (рис. 5). Этот метод помогает выделить внутренние закономерности, отражающие структуру данных. Кластеризацию иногда называют классификацией без учителя, так как оно очень великолепным способом,

справляется с структурированием информации и выведения значимых взаимосвязей из данных.

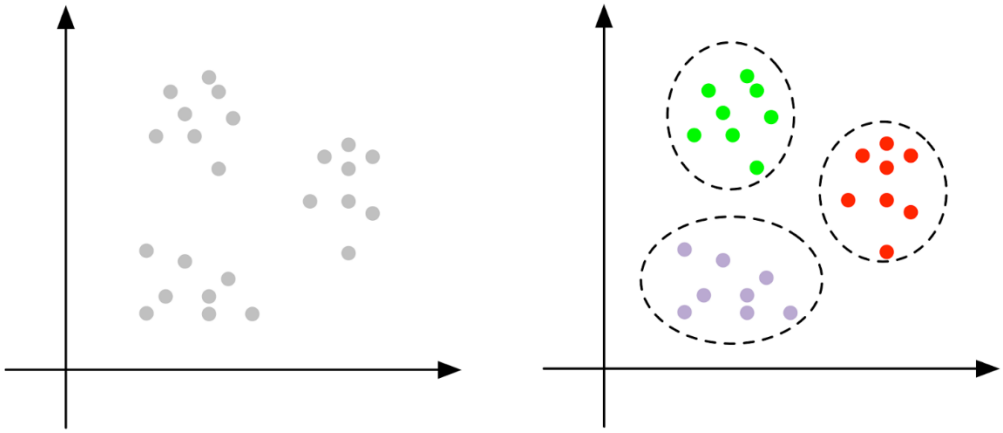

Рисунок 5 – Пример использовании кластеризации

Ещё одной подобластью обучения без учителя, является понижение размерности для сжатия данных, данный способ фокусируется на уменьшении числа признаков в наборе данных, сохраняя при этом существенные характеристики, как показано на (рис. 6). На (рис. 6), приведён пример применения нелинейного понижения размерности для сжатия данных трёхмерного пространства признаков в новое двумерное пространство признаков. Таким образом, данный процесс, позволяет устранить избыточность информации и сосредоточить своё внимание на ключевых аспектах данных, что особенно полезно при работе с большим объёмом информации.

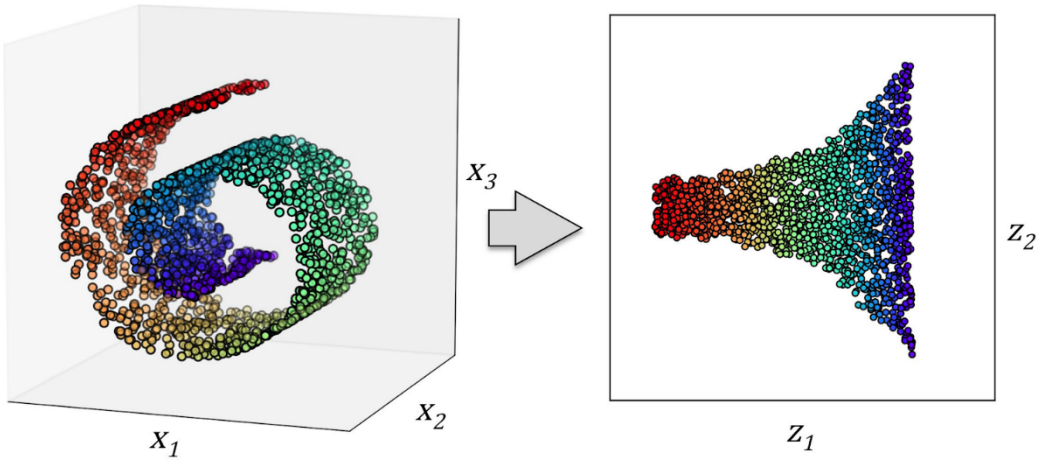

Рисунок 6 – Пример использования понижения размерности

Оба таких подхода, играют важную роль в обучении без учителя, предоставляя инструменты для анализа и извлечения структуру из непомеченных данных [12].

1.2 Методы машинного обучения

После того, как мы более подробно ознакомились с двумя типами машинного обучения, стоит важным рассмотреть их методы, которые мы планируем использовать в рамках нашего исследования. Существует большое количество методов машинного обучения, но в данном случае мы рассмотрим только несколько, которые хорошо подходит для анализа рынка недвижимости, а именно для задачи кластеризации в обучении без учителя, мы подробно разберём метод k-средних (k-means), также уделим внимание обучению с учителем, а именно для задач регрессии разберём такие модели, как линейная регрессия (linear regression), и два ансамблевых метода: случайные леса (Random Forests) и градиентный бустинг (Gradient Boosting).

1.2.1 Метод k-средних

Самый известный и универсальный метод для задачи кластеризации представляет собой метод k-средних. Он успешно масштабируется для обработки больших объёмов данных и применяется в различных областях. Суть этого метода, заключается в минимизации критерия, известного как инерция или сумма квадратов внутри кластера. Этот алгоритм делит набор N образцы X в K не пересекающие кластеры C, каждый из которых описывается средним μ_j, образцов в кластере, средними значениями называют центроидами кластера. Алгоритм K-средних направлен на выбор центроидов, минимимизирующих инерцию или критерий суммы квадратов внутри кластера:

$$\sum_{i=0}^{n} \min_{\mu \in C} (\|x_i - \mu_j\|^2)$$

Инерцию можно определить как меру того, насколько кластеры внутренне связаны. В общих чертах алгоритм k-средних состоит 3 шагов:

1) Выбор k начальных центроидов (k — это количество кластеров).
2) Присвоение точек к ближайшим центроидам. Каждая точка данных присваивается к ближайшему центроиду в пространстве признаков. Это создает k кластеров, где каждый кластер представлен центроидом и точками, которые ему присвоены.
3) Обновление центроидов, пересчитываются центроиды для каждого кластера, опираясь на средние значения признаков точек внутри кластера. Этот процесс продолжается до тех пор, пока центроиды не стабилизируются и не изменятся менее чем на некоторый заранее заданный порог.

Алгоритм k-средних стремится минимизировать суммарное квадратичное отклонение между точками кластера и их соответствующими центроидами. Как результат, после нескольких итераций центроиды становятся представительными для средних значений внутри каждого кластера, а точки принадлежат кластерам на основе близости к центроидам [8].

1.2.2 Ансамблевые методы регрессии: случайные леса и градиентный бустинг

Ансамблевые методы, такие как случайные леса и градиентный бустинг, представляют собой мощные инструменты в области машинного обучения. Эти методы основаны на идее объединения нескольких слабых моделей для создания сильной и устойчивой прогностической силы. В данном разделе мы рассмотрим два важных ансамблевых подхода: случайные леса, которые строятся на основе решающих деревьев, и градиентный бустинг, который

последовательно улучшает модель, фокусируясь на ошибках предыдущих итераций. Детальное погружение в эти методы позволит понять их принципы работы и применение в различных задачах машинного обучения [3].

Метод регрессии случайный лес (Random Forests), это универсальный метод машинного обучения, который объединяет прогнозы нескольких деревьев решений, чтобы уменьшить переобучение и повысить точность. Концепция случайного леса, представлена на (рис. 7).

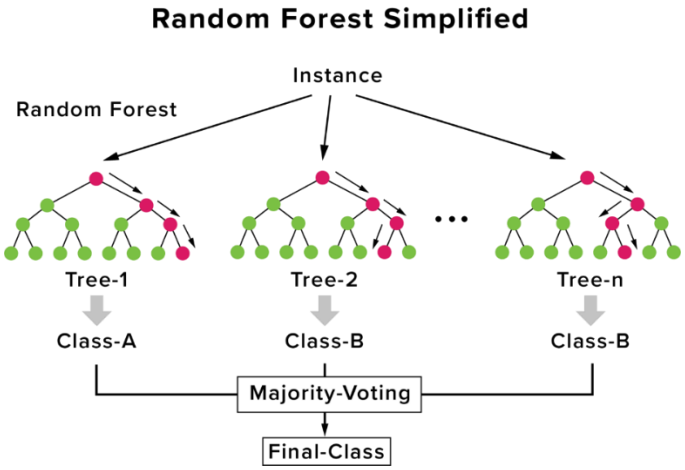

Рисунок 7 – Концепция упрощённого алгоритма случайный лес

Работа алгоритма состоит из четырёх этапов:

1) Создания случайных выборок: из исходного набора данных случайным образом выбираются подмножества данных с возвращением, то есть элементы могут быть выбраны несколько раз, в то время как другие могут быть пропущены. Эти случайные подвыборки используются для обучения каждого дерева в лесу.
2) Построение деревьев решений: для каждой случайной подвыборки строится отдельное дерево решений. На каждом узле данных происходит разделение на основе определенного признака и порогового значения. Процесс строится до достижения критерия остановки, такого как заданная глубина дерева или минимальное количество элементов в узле.
3) Голосование: каждое дерево в случайном лесу предсказывает числовое значение для каждого элемента тестового набора данных. На этапе

голосования прогнозы каждого дерева "голосуют" за числовые значения, формируя множество прогнозов для каждого элемента данных.

4) Выбор окончательного результата: окончательное предсказание определяется выбором прогноза с наибольшим средним значением (или медианой). Этот шаг помогает сгладить возможные переобучения, возникающие при использовании отдельных деревьев в задачах регрессии.

Алгоритм случайного леса обладает характеристиками стабильности и надежности, что делает его эффективным средством для анализа различных видов информации. Путем комбинирования выводов из различных деревьев, случайный лес способствует улучшению точности прогнозов и снижению риска переобучения [18].

Еще один метод ансамблевой регрессии – это градиентный бустинг (Gradient Boosting). Подобно своему предшественнику, случайному лесу, этот метод сильно зависит от использования алгоритма деревьев решений.

Процесс обучения алгоритма градиентного бустинга начинается с создания отдельного листа, основываясь на значениях выходного набора данных. Этот лист представляет собой начальное предположение о выходных значениях набора данных. По мере выполнения алгоритм строит новые деревья решений, основываясь на значениях ошибок, вычисленных по результатам предыдущих деревьев. Например, в случае непрерывных целевых переменных начальным предположением для алгоритма градиентного бустинга может быть среднее значение целевой переменной (выхода). Концепция алгоритма градиентного бустинга, представлена на (рис. 8).

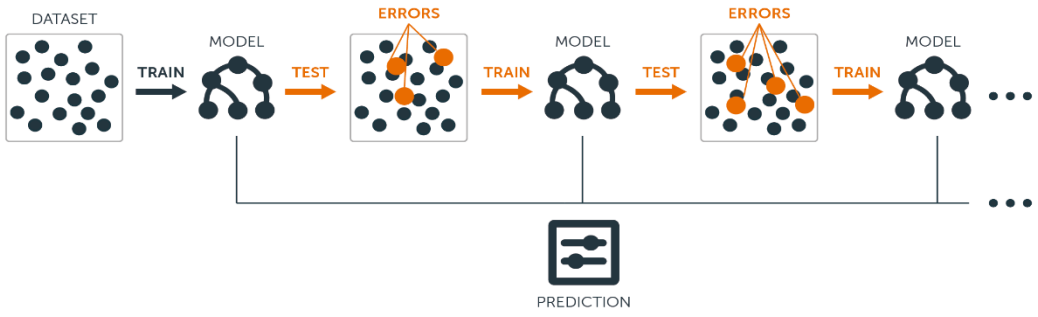

Рисунок 8 – Концепция алгоритма бустинга

Как показано в (рис. 8), основная идея алгоритма градиентного бустинга заключается в пошаговом обучении слабых моделей, часто представленных в виде деревьев решений, с последующей компенсацией их ошибок.

Чаще всего, помимо стандартной библиотеки GradientBoosting для реализации алгоритма градиентного бустинга используется более мощная и эффективная альтернатива в виде XGBoost, представляющей собой регуляризованную версию градиентного бустинга [14].

Таким образом, градиентный бустинг нацелен на пошаговое улучшение прогнозов, учитывая допущенные предыдущими моделями ошибки и фокусируясь на областях, где эти ошибки были наиболее существенными.

1.3 Характеристика рынка недвижимости г. Липецка

Российский рынок недвижимости, включая город Липецк, подвержен влиянию общих законов экономики, но в то же время, его развитие определяется спецификой страны и отражает общий вектор перехода от стихийного к уравновешенному и цивилизованному состоянию. Состояние рынка недвижимости несет в себе индикаторы национальной экономики, оказывая влияние на бизнес, социально-демографические задачи и стратегические направления развития страны.

Эксперты отмечают, что региональные параметры рынка недвижимости формируются под влиянием изменений в экономической, социальной и экологической сферах. Важно отметить, что социальное благополучие и качество жизни в регионе тесно связаны с развитием рынка недвижимости, включая доступность жилья, его стоимость и качество.

В 2023 году в Липецкой области был отмечен повышенный спрос на квартиры, сопровождаемый значительным ростом цен. Жители региона, опасаясь обесценивания своих сбережений, активно вкладывали в недвижимость (рис. 9). Согласно официальным данным Минстроя, средняя рыночная стоимость жилья в IV квартале 2023 года почти достигла 83 тысяч

рублей за 1 кв. м, что существенно превышает уровень в IV квартале 2022 года (75 тысяч рублей за 1 кв. м) [3]. Однако следует отметить, что данные Минстроя представляют собой усредненные значения, не всегда точно отражающие реальную ситуацию на рынке. Не смотря на существующее множество факторов, участвующих в процессе ценообразования недвижимости, большинство из них малопригодно для решения задач прогнозирования.

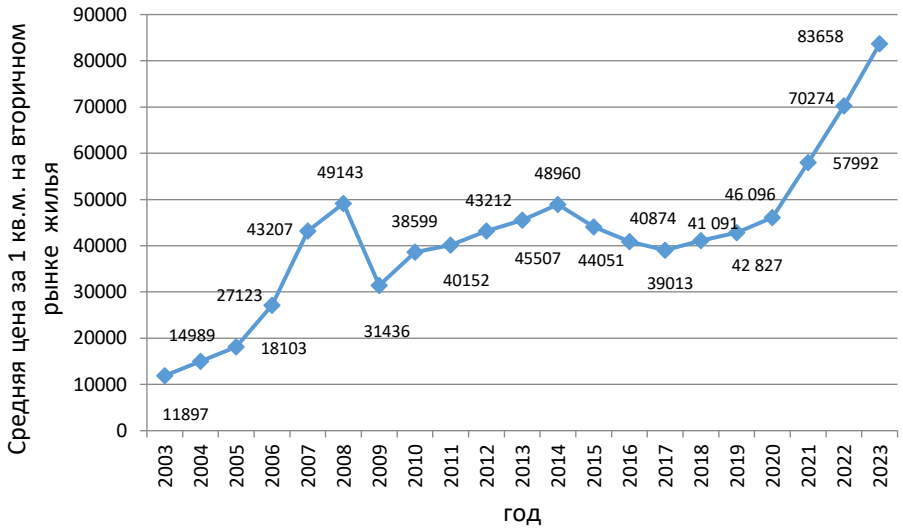

Рисунок 9 – Ценовая динамика на вторичном рынке жилья в Липецке по данным Минстроя

Нами был построен график динамики средних цен вторичного рынка недвижимости за период 2020-2023 гг. (рис. 10), в котором наблюдается такая же тенденция как и по оффициальным данным Минстроя.

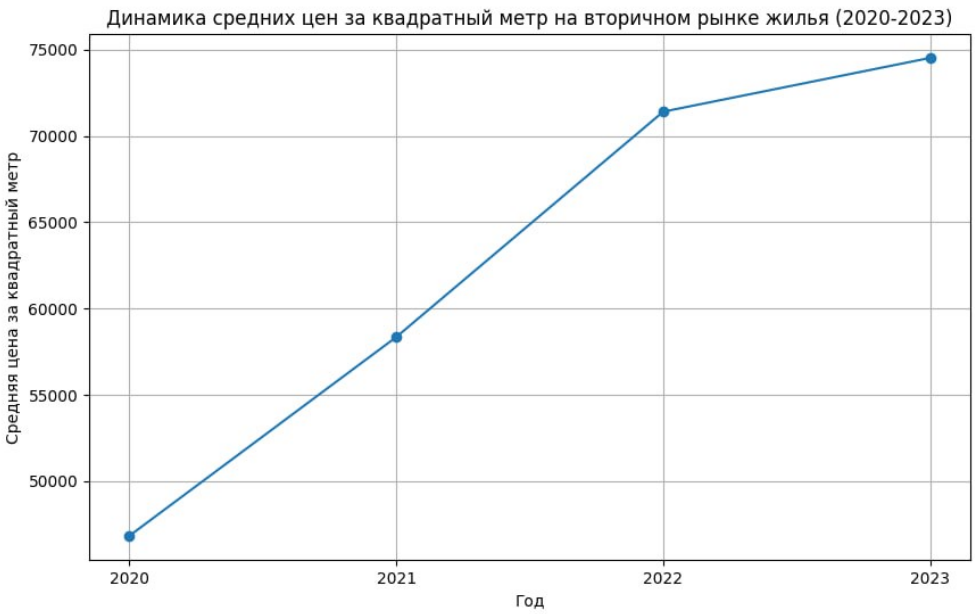

Рисунок 10 – Ценовая динамика на вторичном рынке жилья в Липецке по нашим данным

Использование методов машинного обучения в анализе вторичного рынка недвижимости позволяет более точно и детально изучать динамику изменений цен, предсказывать тенденции и выявлять факторы, влияющие на спрос и предложение. Это открывает новые возможности для эффективного управления недвижимостью, принятия обоснованных решений в условиях динамичного рынка и повышения прозрачности сделок.

Таким образом, внедрение методов машинного обучения в анализ рынка недвижимости Липецка предоставляет инструментарий для более глубокого понимания его характеристик, определения реальных динамик и повышения эффективности управления недвижимым имуществом в регионе.

ГЛАВА 2. ПОСТРОЕНИЕ МОДЕЛЕЙ ДЛЯ АНАЛИЗА РЫНКА НЕДВИЖИМОСТИ

Машинное обучение, включает в себя в первую очередь постановку задачи, то есть выбор входных данных и используемых методов машинного обучения. Следующим шагом и самым главным, является сбор данных, в нашем случае, это характеристики недвижимости, учитывая широкий спектр данных, важно оценить их актуальность, доступность и достоверность. После выбора данных нужно их подготовить, для этого необходимо избавиться от выбросов и пропусков, а также дубликатов, провести стандартизацию числовых переменных и кодирование категориальных данных, после этого разделить данные на обучающую и тестовую выборку для последующего обучения и оценки модели. Для выбора модели нужно отталкиваться от характера данных и поставленной цели. Обучение и оценка модели предполагает правильной настройки параметров модели и использований метрик эффективности для определения точности модели. Последним и цикличным шагом для достижения хороших результатов, является анализ сильных и слабых сторон, а также понимание насколько модель, соответствует поставленным задачам и требованиям, в случае необходимости внести корректировки в модель, изменить параметры, добавить новые данные или изменить метод машинного обучения для улучшения её производительности.

2.1 Выбор инструментов

В мире, обладающем большим количеством данных, всё больше приобретает популярность язык программирования «Python». Это очень высокоуровневый и интерпретируемый язык сценариев, он прост и понятен в синтаксисе, а порог вхождения для изучения, меньше, чем у других языков программирования, чем обуславливается динамичным развитием рынка и привлечением новой аудиторией, как показано на (рис. 11).

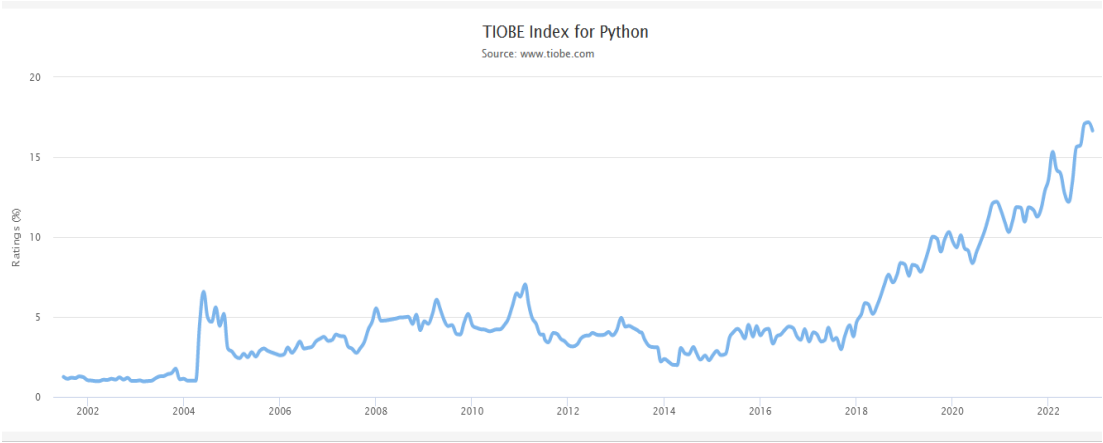

Рисунок 11 – Тенденция популярности языка программирования «Python»

Так же, «Python», обладает большой документацией и мощными библиотеками для работы с данными, что позволяет нам в данной работе воспользоваться следующими библиотеками:

- Pandas (полезная библиотека в анализе уже структурированных данных);
- Matplotlib (данная библиотека хорошо пригодится для визуализации данных двумерных и трёхмерных графиков);
- Seaborn (это библиотека основанная на Matplotlib и предназначена для создания статистических графиков);
- Scikit-learn (это одна из популярных библиотек для машинного обучения);
- SQLite3 (это библиотека предназначена для построения запросов в БД SQLite);

Для хранения данных в нашей работе, мы будем использовать СУБД SQLite, а в качестве инструмента для манипуляций над этими данными, мы воспользуемся инструментом SQLiteStudio (3.4.4), из-за высокой производительности и быстрого подключения для небольших проектов.

На основании выше сказанного, для проведения обработки данных и построение модели станет язык программирования «Python (3.11)», а средой выполнения «Visual Studio Code (1.86.2)» [2].

2.2 Сбор и обработка данных

Самый трудоёмкий процесс для проведения анализа, это сбор и подготовка данных. В нашей работе будут использованы данные, предоставленные Липецкой организацией «АН Квадрат» из их базы данных по продажам за временной период 2020 – 2023 гг.

Учитывая, что основной целью работы является анализ вторичного рынка жилья г. Липецка, была проведена предварительная обработка данных, было проделано следующее:

- Удалены пустые столбцы;
- Удалены объекты, не представляющие ценность (аренда, комнаты, участки, дома, объекты из других областей и городов);
- Объекты первичного рынка были также удалены;
- Была произведена очистка строк, содержащих наибольшее количество пропусков
- Дубликатов не оказалось

На следующем этапе была проведена визуализация данных на географической карте. Для этого были собраны координаты границ округов города Липецка с помощью сервиса карт OpenStreetMap. С помощью языка программирования python и библиотеки request был написан код для построения запроса к этому сервису для получения координат границ округов Липецка и сохранения их в файл json. После чего с помощью библиотек folium, json и sqlite3 была написана функция для построения маркеров на карте, результат проделанной работы представлен на (рис. 12).

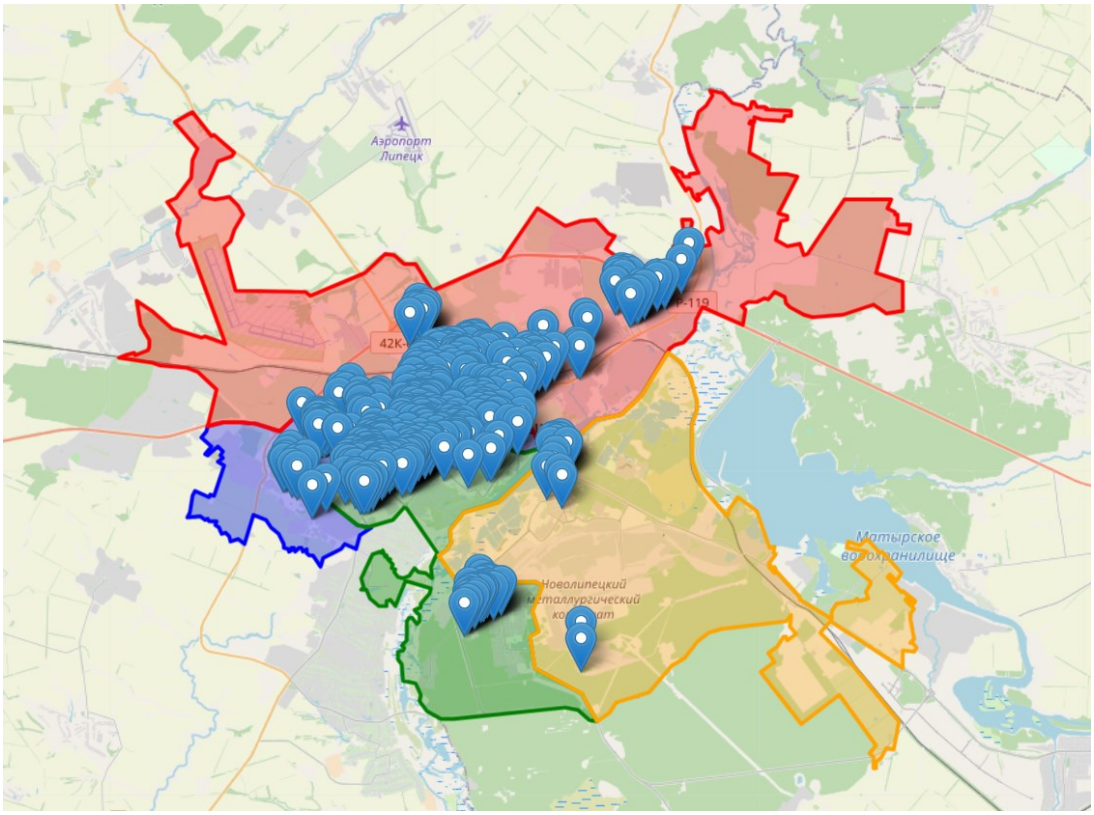

Рисунок 12 – Визуализация маркеров на карте г. Липецка

Визуализация объектов на карте города Липецка (рис. 12), нужна была, чтобы удостовериться, что объекты рынка недвижимости не выходят за границы районов г. Липецка. Это было необходимо, чтобы построить более эффективную модель из-за особенностей объектов и их цен за пределами города.

После предобработки данных, категориальные признаки были преобразованы в числовой формат. В данном случае мы использовали библиотеку scikit-learn, модуль Label Encoding с целью подготовки данных для использования в алгоритмах машинного обучения. Мы решили использовать этот метод, из-за его популярности и экономии пространства.

После предобработки данных можно представить новый формат таблицы, имеющий правильный тип данных, отсутствие пропусков и дубликатов, результат можно увидеть в таблице 1.

Таблица 1 – Итоговой вариант признаков

Числовые признаки	
Название столбца	Тип данных
Год	INT
Месяц	INT
День	INT
Цена	INT
Количество комнат	INT
Общая площадь	FLOAT
Жилая площадь	FLOAT
Площадь кухни	FLOAT
Этаж	INT
Этажность	INT
Долгота	FLOAT
Широта	FLOAT
Категориальные признаки	
Название столбца	Тип данных
Материал дома	INT
Тип ремонта	INT
Район	INT
Тип санузла (совмещённый/раздельный)	INT
Тип комнат (изолированные/смежные)	INT
Тип водоснабжения	INT
Балкон/лоджия/ничего	INT

2.3 Кластерный анализ

Основная цель кластерного анализа заключается в разделении объектов на группы или кластеры таким образом, чтобы внутри каждой группы наблюдения были более похожи друг на друга, чем на объекты из других кластеров. Поэтому одной из основных задач нашего исследования является формирование кластеров на основе стоимости квартир с применением метода k-средних (KMeans) для более глубокого понимания процессов ценообразования в городе Липецке.

Для начала мы планируем написать функцию на Python, которая позволит нам географически визуализировать ценообразование недвижимости в Липецке, используя средние значения цен (рис. 13). Это позволит нам лучше понять пространственную структуру данных и подготовить основу для последующего кластерного анализа и интерпретации результатов.

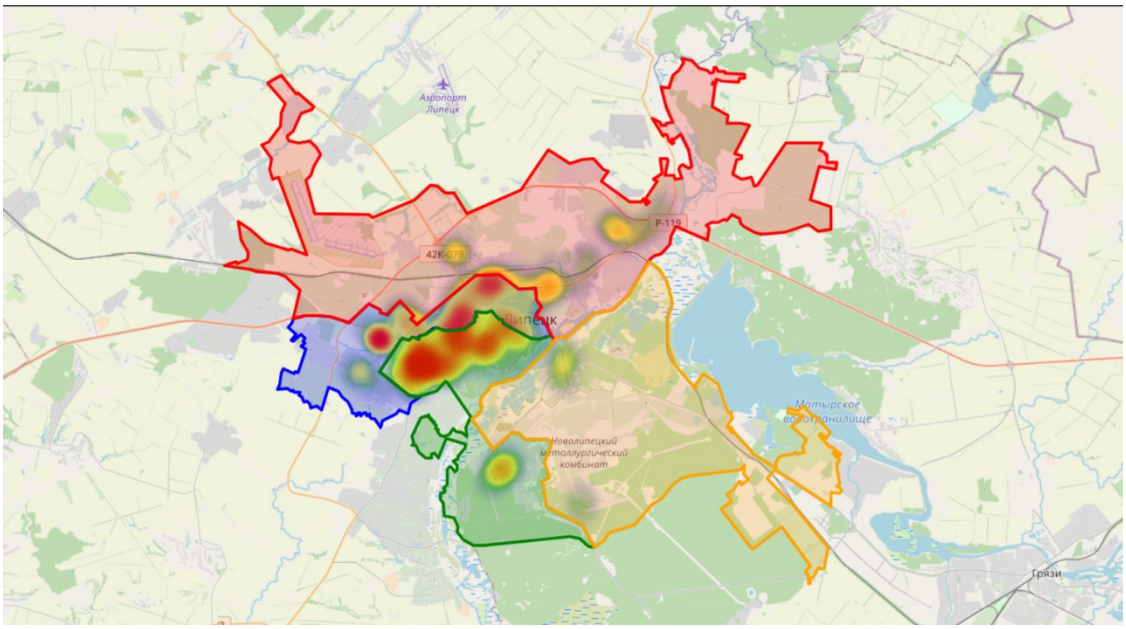

Рисунок 13 – Тепловая карта цен г. Липецка

На (рис. 13) отчетливо видно, что более дорогое жилье представлено в Советском и Октябрьском районах города Липецка, в то время как наименее дорогое жилье располагается в Правобережном и Левобережном районах. Этот анализ географического распределения цен становится отправной точкой для последующего кластерного анализа. Мы планируем использовать метод k-

средних (KMeans) для выделения кластеров, однако перед этим необходимо определить оптимальное количество кластеров с помощью метода локтя.

Суть в этом методе, графическое определение оптимального числа кластеров в методе k-средних (KMeans). Идея заключается в том, чтобы найти точку на графике, где увеличение числа кластеров приводит к заметному уменьшению внутрикластерного расстояния или суммы квадратов расстояний между точками и центроидами их кластера.

Для обеспечения корректной работы метода k-средних (KMeans), который чувствителен к масштабу данных, мы проведем стандартизацию данных. Этот процесс заключается в приведении всех числовых переменных к общему масштабу путем вычитания среднего значения и деления на стандартное отклонение. Такая нормализация данных особенно важна, когда признаки имеют различные диапазоны значений или измеряются в разных единицах. В нашем случае, поскольку цены на жилье в каждом году могут значительно отличаться, использование метода StandardScaler для стандартизации этих переменных обеспечивает удобство анализа. Это позволяет привести признаки к стандартным единицам, где среднее значение становится равным 0, а стандартное отклонение – 1, что в свою очередь облегчает работу с данными, имеющими различные диапазоны и значения.

В качестве функции потерь будет выступать сумма квадратов внутрикластерных расстояний (WCSS):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \min \left(\left\| x_i^{(j)} - c_j \right\| \right)^2, \text{ где:}$$

k - количество кластеров,

n - количество точек данных,

$x_i^{(j)}$ - i-я точка данных внутри j-го кластера,

c_j- центроид для j-го кластера.

После определения функции потерь, с помощью Python нами был построен график с использованием метода локтя, представленный на (рис. 14).

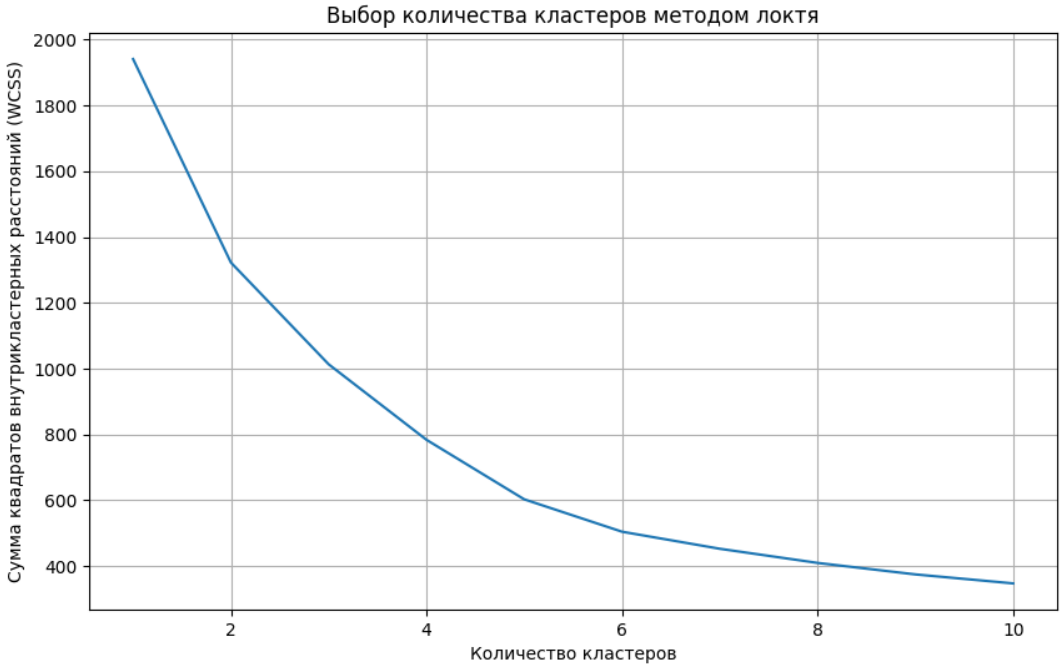

Рисунок 14 – Выбор количества кластеров методом локтя

Как видно на (рис. 14), когда мы перешли от пяти до шести кластеров, ошибка перестала существенно уменьшаться. Это объясняет, что выбор количества кластеров равен пяти.

Нами был создан объект класса модели с использованием пяти кластеров, а также были проведены эксперименты с различными гипермаркетами (параметрами) модели, представленные на (рис. 15).

```
24  # Определение числа кластеров
25  num_clusters = 5
26  kmeans = KMeans(
27      n_clusters=num_clusters,
28      init='k-means++',
29      max_iter=300,
30      n_init=20,
31      random_state=42
32  )
33  kmeans.fit(scaled_features)
```

Рисунок 15 – Гипермаркеты модели k-средних

Как видно на (рис. 15), параметры модели k-средних имеют свои особенности:

- n_clusters: это количество кластеров, на которые мы хотим разбить наши наблюдения, в данном случае было определено пять, 4 и 6 тоже проверялись, но результаты получились некорректными.
- init: определяет, как мы выберем первоначальное расположение (инициализацию) центроидов. Есть два варианта, выбрать центроиды случайно init = random или выбрать их так, чтобы центроиды с самого начала располагались максимально далеко друг от друга init = k-means++; второй вариант оптимальнее, был выбран он и проверен метод инициализации init = random, результаты были неизменны.
- n_init: сколько раз алгоритм будет инициализирован, т.е. сколько раз будут выбраны центроиды до начала оптимизации; на выходе будет выбран тот вариант, где ошибка была минимальна, было выбрано разное кол-во, результаты неизменны.
- max_iter: максимальное количество итераций алгоритма после первоначального выбора центроидов, было выбрано разное кол-во, результаты оказались неизменны.
- random_state: воспроизводимость результата.

Данные параметры прошли ряд изменений, которые особой роли в улучшение не сыграли, но и не показали себя хуже. Для оценки качества модели кластерного анализа мы использовали две ключевые метрики: коэффициент силуэта и индекс Калински-Харабаша.

Коэффициент силуэта, имеющий значение 0,375, отражает степень схожести объектов как внутри кластера, так и между кластерами. Этот коэффициент принимает значения от -1 до 1, где высокие значения указывают на хорошую кластеризацию, близкие к 0 – на перекрывание кластеров, а значения, приближающиеся к 1, свидетельствуют о четком разделении кластеров. Оценка в 0,375 обычно считается довольно хорошей, но важно учитывать контекст и сравнивать результаты с альтернативными подходами.

Индекс Калински-Харабаша, равный 347,1, также является важной метрикой, оценивающей компактность и разделение кластеров. Большие значения этого индекса обычно указывают на более четкое разделение между кластерами и более компактные кластеры. Учитывая небольшой объем данных, значение 347,1 кажется довольно высоким.

Исходя из результатов метрики, сложно сделать выводы о правильности модели, так как это в основном зависит от контекста задачи и структуры данных. Для лучшего понимания, стоит визуализировать полученные результаты кластерного анализа. Для этого мы воспользуемся библиотекой Folium, чтобы продемонстрировать результаты на карте OpenStreetMap, представленная на (рис. 16).

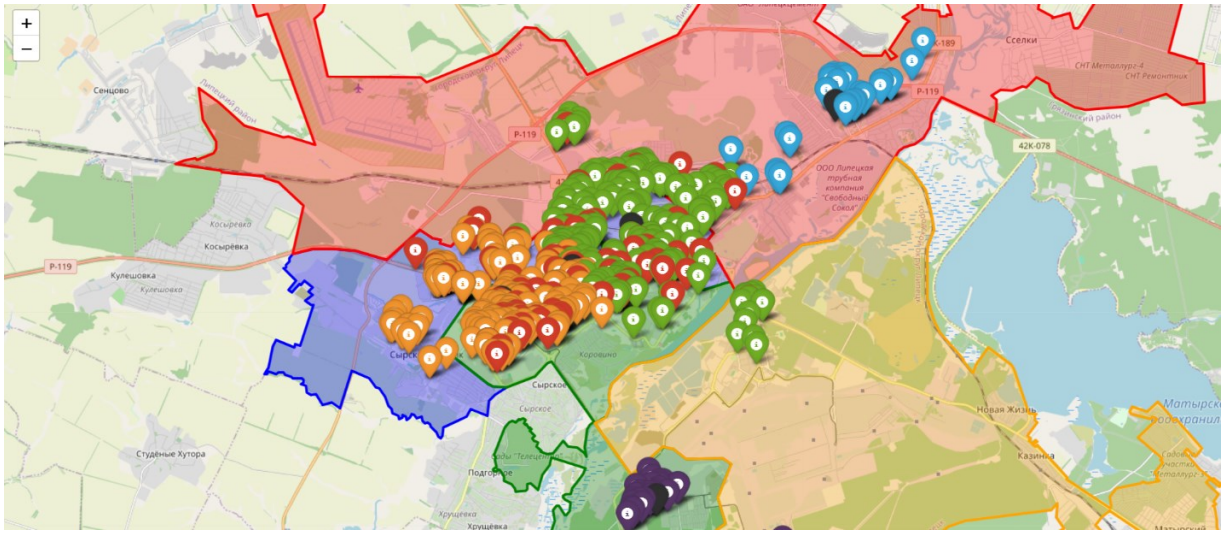

Рисунок 16 – Визуализация кластеров

На (рис. 16) можно увидеть, что кластеризация соответствует ожиданиям и логике, что считается положительным аспектом, так как правильность модели зависит от того, насколько успешно она отражает структуру выбранных данных и решает конкретную задачу. В нашем случае задача по формированию кластеров по стоимости в каждом районе выполнена, каждый кластер имеет следующие группы цен:

- Кластер 0 (зелёный кластер) имеет группу стоимости объектов (244) от 700.000 до 3.790.000;

- Кластер 1 (оранжевый кластер) имеет группу стоимости объектов (249) от 1.155.000 до 4.450.000;
- Кластер 2 (синий кластер) имеет группу стоимости объектов (35) от 700.000 до 4.550.000;
- Кластер 3 (красный кластер) имеет группу стоимости объектов (87) от 3.600.000 до 10.900.000;
- Кластер 4 (фиолетовый кластер) имеет группу стоимости объектов (32) от 600.000 до 5.300.000;

Из предоставленных групп кластеров по стоимости объектов и их визуализация (рис. 16) говорит о том, что кластеры 0, 1, 2 и 4 были правильно разбиты, а вот кластер 3 говорит о дорогом сегменте объектов, это могут быть как и недавно появившиеся новостройки, так и дома чуть выше среднего сегмента рынка. Центроиды групп кластеров (рис. 16), представлены в виде чёрных маркеров.

Также, следует отметить различие количества объектов в каждом кластере, это может быть обусловлено неоднородностью рынка, плотностью застроек, особенностями районов и многими другими факторами, визуализация кластеров по количеству наблюдений, представлено на (рис. 17).

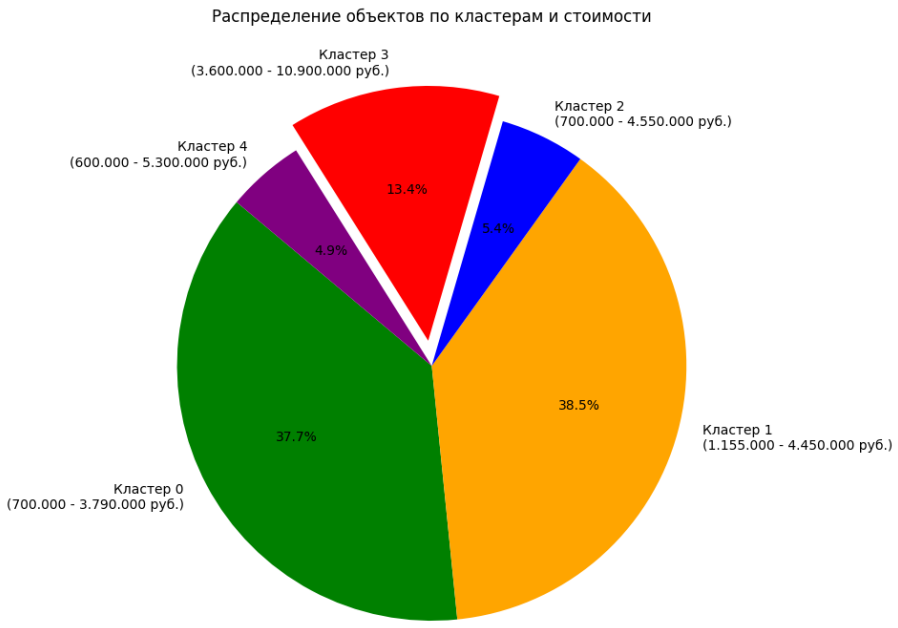

Рисунок 17 – Количество наблюдений в каждом кластере

Для лучшего понимания взаимосвязей между факторами в каждом кластере была создана тепловая карта корреляций. Этот визуальный метод представления матрицы корреляций позволяет быстро и наглядно оценить силу и направление связи между различными переменными. Мы воспользовались функцией corr() из библиотеки pandas для расчёта корреляционной матрицы между всеми парами признаков. Для расчёта корреляционной матрицы между всеми парами признаков для цены был использован коэффициент Пирсона. Затем мы использовали функцию heatmap() из библиотеки seaborn для визуализации тепловой карты корреляций для каждого кластера, представленные на (рис. 18-22). Кроме того, была сделана общая визуализация влияния признаков на цену недвижимости, представленная на (рис. 23) [4].

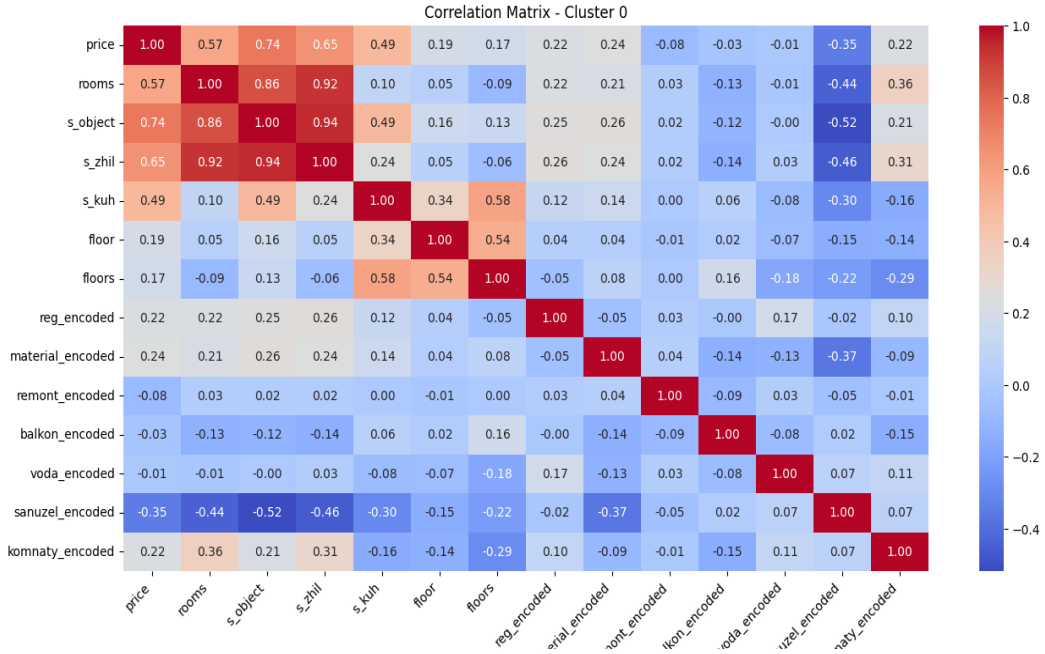

Рисунок 18 – Визуализация корреляции между признаками и ценой для кластера 0

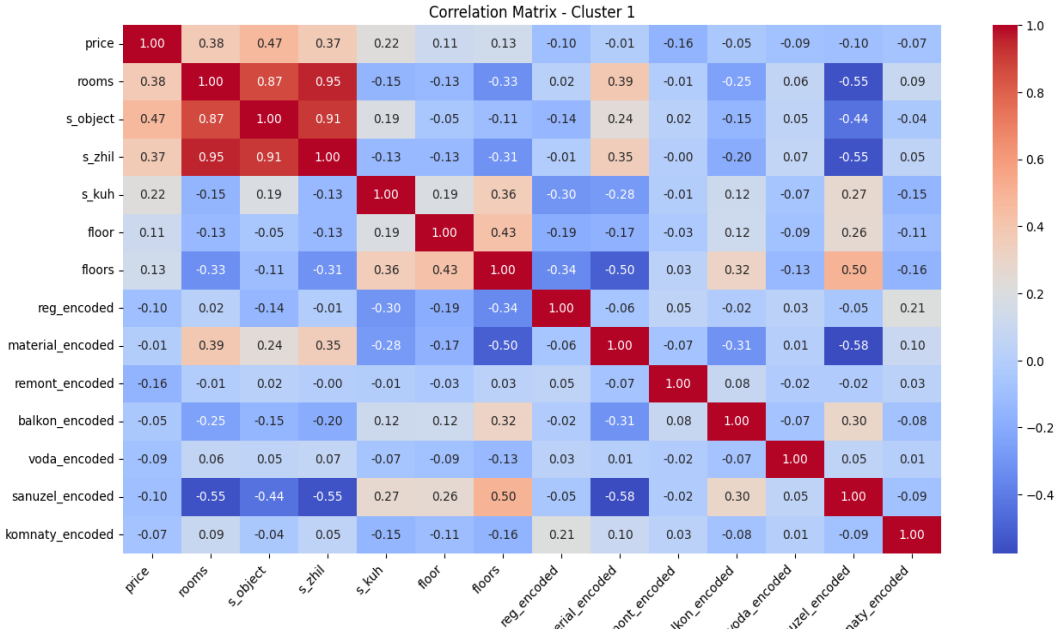

Рисунок 19 – Визуализация корреляции между признаками и ценой для кластера 1

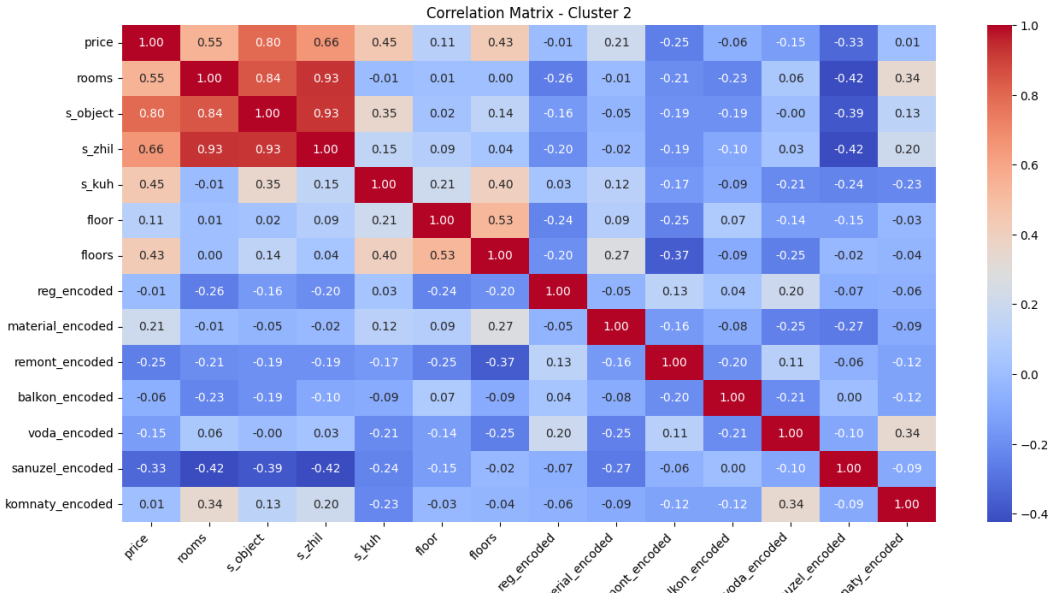

Рисунок 20 – Визуализация корреляции между признаками и ценой для кластера 2

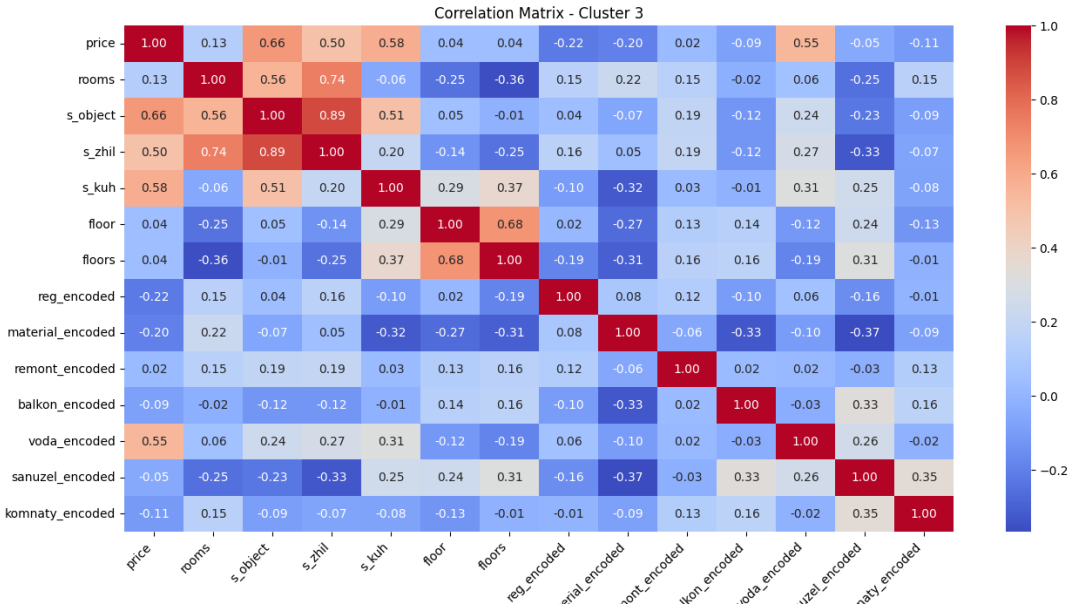

Рисунок 21 – Визуализация корреляции между признаками и ценой для кластера 3

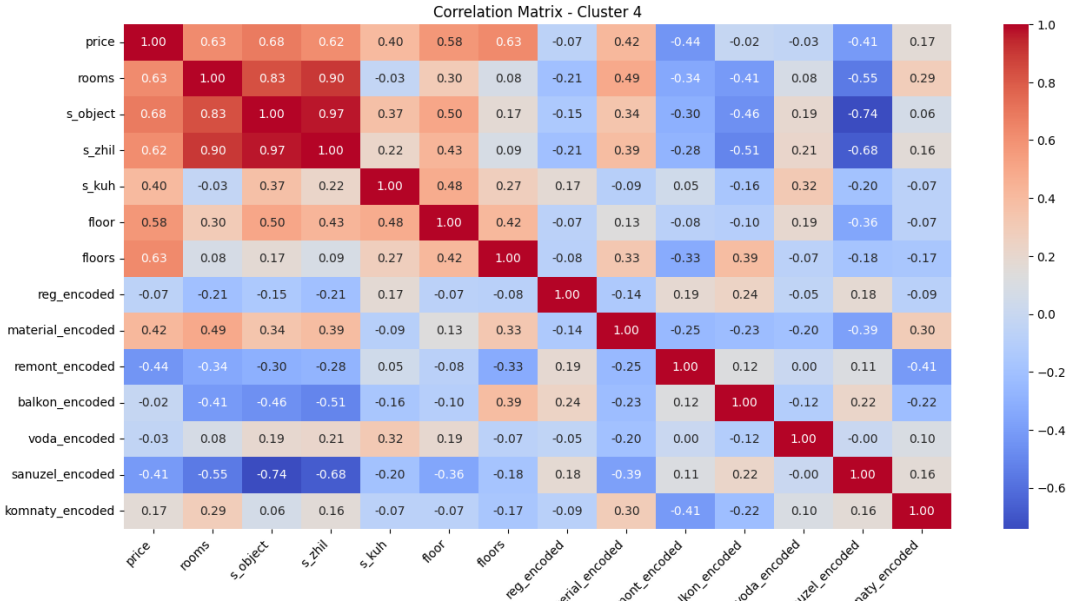

Рисунок 22 – Визуализация корреляции между признаками и ценой для кластера 4

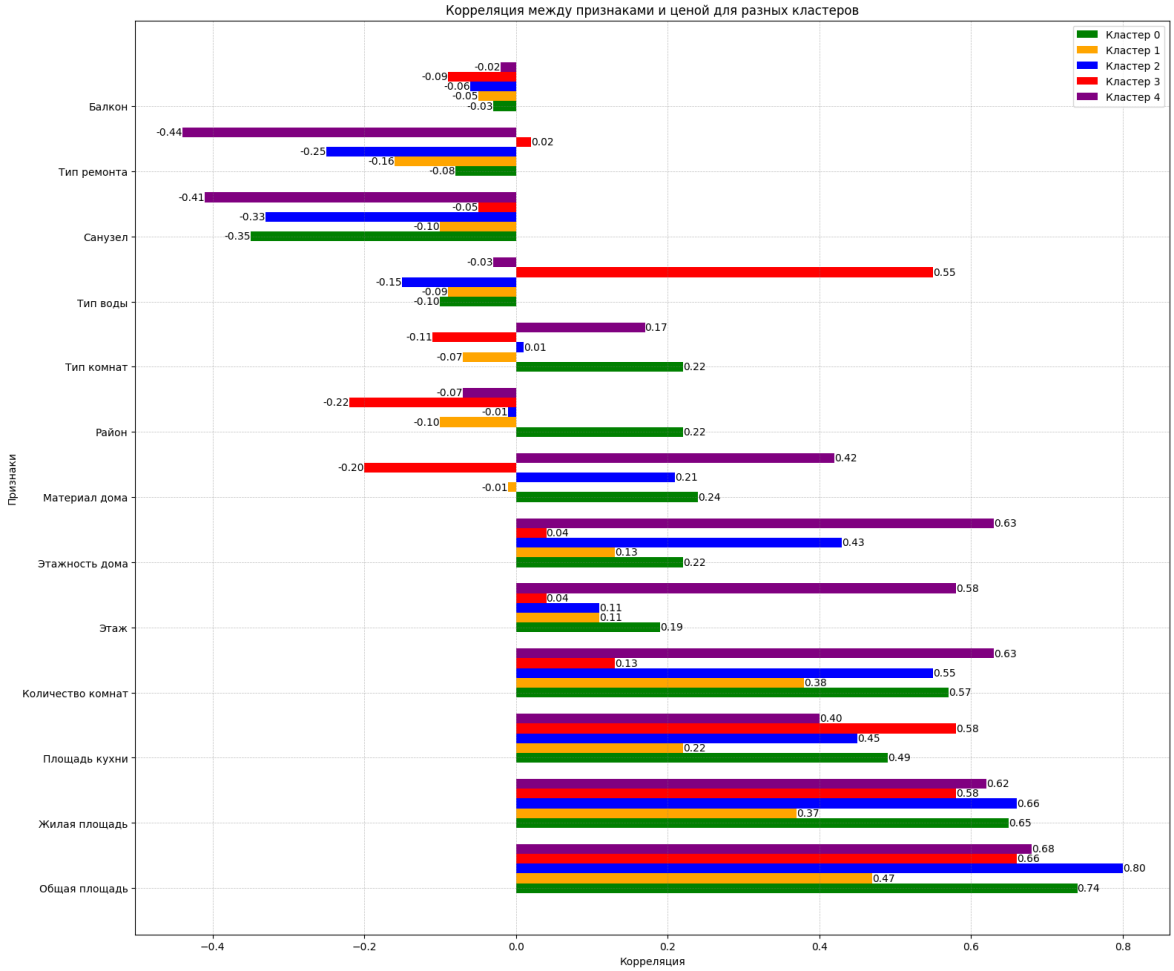

Рисунок 23 – Визуализация корреляции между признаками и ценой для разных кластеров

Исходя из (рис. 23) и (рис. 16) можно сделать совместный выводы по кластерному и корреляционного анализу для каждого кластера:

1. Кластеры 0, 1 и 2 демонстрируют высокую положительную корреляцию с ценой для таких признаков, как общая площадь, жилая площадь, площадь кухни и количество комнат. Это означает, что увеличение данных характеристик сопровождается ростом цен на жильё. В кластере 3, помимо этих признаков, также наблюдается заметная положительная корреляция с ценой для признака тип воды. Это можно объяснить тем, что данный сегмент рынка является более дорогим и предпочтительным для жизни, например, наличием индивидуального отопления в виде газовых котлов многоквартирного дома, что может повлиять на повышение цен на жилье в этом кластере. В кластере 4, кроме характеристик общей площади, жилой

площади и количества комнат, также отмечается положительная корреляция с ценой для признаков этажа и этажности дома. Это можно объяснить тем, что данный кластер расположен в более старом районе города (ЛТЗ), который не имеет больших этажностей, что делает квартиры на первых и последних этажах менее привлекательными из-за особенностей состояния дома и давней застройки, поэтому разница в цене на квартиру между этажами может сильно различаться.

2. Обратная корреляционная связь между типом ремонта и ценой наблюдается в кластерах 0, 1, 2 и 4. В этих кластерах можно выделить тенденцию, при которой наблюдается некоторое влияние на снижение цены в кластерах 0 и 1, а также более заметное влияние в кластерах 2 (Сокол) и 4 (ЛТЗ). Это может быть обусловлено тем, что в более старых районах города, таких как ЛТЗ и Сокол, плохое состояние ремонта квартиры оказывает более существенное влияние на цену жилья, чем в кластерах 0 и 1 (Центр или Новые районы). В кластерах 0 (частично Советский, частично Правобережный район), 2 (Сокол) и 4 (ЛТЗ) также наблюдается умеренная обратная связь цены жилья с наличием санузла. Это может быть объяснено тем, что в таких районах часто встречаются дома с планировкой жилья, где санузел является совмещенным, что, в свою очередь, может отражаться на снижении цены жилья.

Таким образом, анализ кластеров и корреляционных связей позволяет более глубоко понять факторы, влияющие на цены жилья в различных районах. Эти результаты могут быть полезны для определения стратегий ценообразования, а также для выявления потенциальных факторов, которые могут оказывать влияние на спрос и предложение на рынке недвижимости. Дальнейшие исследования могут включать анализ других характеристик жилья и их влияния на цену, а также изучение долгосрочных тенденций в развитии рынка недвижимости.

2.5 Регрессионный анализ

Каждый объект недвижимости обладает уникальным набором свойств, влияющих на его цену на рынке. Наша задача заключается в использовании данных о квартирах, где для каждой фиксируются характеристики объекта и соответствующая рыночная цена. Мы стремимся разработать модель, которая, основываясь на этих данных, предоставит нам инструмент для прогнозирования наиболее вероятной рыночной цены нового объекта на основе его характеристик. Важно отметить, что мы учитываем все факторы, воздействующие на стоимость объекта, включая технические параметры и его местоположение.

Для последующего проведения регрессионного анализа будет использован традиционный метод машинного обучения, такой как линейная регрессия и ансамблевые методы: случайный лес и градиентный бустинг.

Исходный набор данных был разбит на две части с пропорцией 70% на обучающую и 30% на тестовую выборки. Первая часть использовалась для обучения моделей, в то время как вторая – для оценки точности их прогнозов.

Также, перед обучением модели проведём логарифмическое преобразование к ценам (рис. 24), так как разброс значений целевой переменной варьируется в зависимости от времени (2020-2023г.), использование логарифмического преобразования может помочь сгладить сильное правостороннее распределение и сделать его более нормальным.

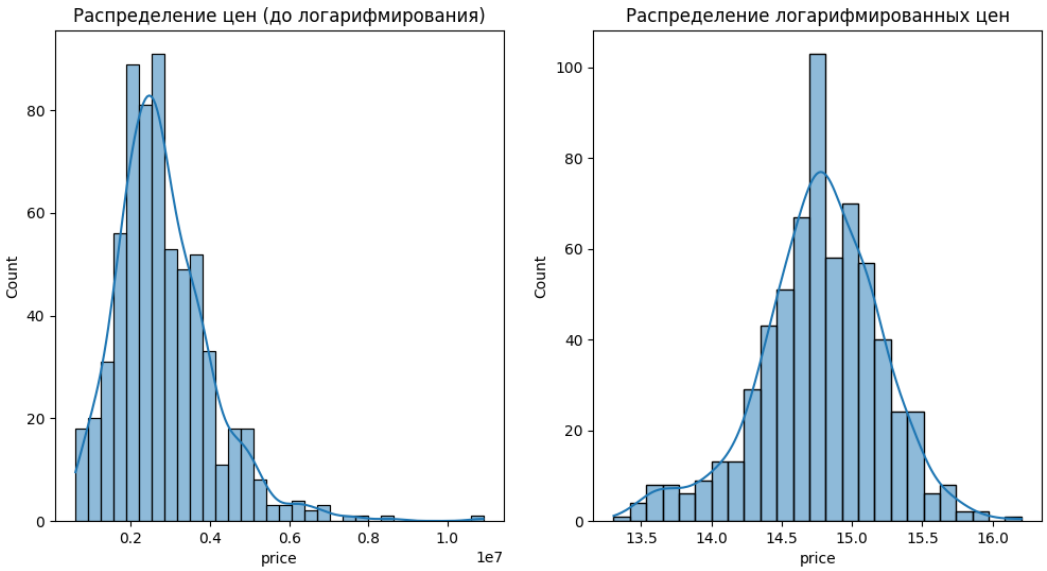

Рисунок 24 – Распределение Гаусса

На (рис. 24) видно, что распределение цены до логарифмирования скошено вправо, так как оно имеет «хвост» в правой части распределения, а распределение после логарифмирования не имеет перекоса и может считаться приемлемым результатом. Для независимых переменных X (числовые и категориальные признаки), будет произведена стандартизация, чтобы уравновесить влияние различных признаков на модель.

Для оценки эффективности, мы проведём вычисление различных метрик точности для каждого результата. В частности:

1) Коэффициент детерминации (R^2) представляет собой долю объясненной моделью дисперсии. Чем ближе значение R^2 к 1, тем лучше соответствие модели данным:

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

n – количество объектов в выборке;

2) Средняя абсолютная процентная ошибка (MAPE) показывает среднюю величину ошибки модели в процентном отношении:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100;$$

n – количество объектов в выборке;

3) Медианная абсолютная процентная ошибка (MedAPE) представляет собой среднее значение среди всех упорядоченных процентных ошибок.

$$\mathrm{MedAPE} = \mathrm{med}\left\{\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right\} \times 100;$$

n – количество объектов в выборке;

4) Среднеквадратическая ошибка модели (MSE), измеряет среднее значение квадратов разностей между предсказанными и фактическими значениями целевой переменной. Она штрафует за большие ошибки и может быть полезна, если важны точные значения предсказаний:

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

n – количество объектов в выборке;

Выбранные метрики были вычислены как для набора данных, используемого в процессе обучения, так и для выделенной тестовой выборки. Стоит отметить, что тестовая выборка полностью отсутствовала в процессе обучения модели, что гарантирует независимость оценок точности. Характеристики точности были рассчитаны путем сопоставления прогнозных значений рыночной стоимости с реальными ценами предложений для тех же объектов, что отражает реальную точность предсказаний. [10]

2.5.1 Линейная регрессия

Для построения модели линейной регрессии, мы воспользуемся модулем LinearRegression, для этого будет написана функция для обучения модели, где Y – зависимая переменная (цена), а X – независимые переменные (числовые и категориальные признаки). Результаты были получены в виде точечного графика и линией тренда (рис. 25).

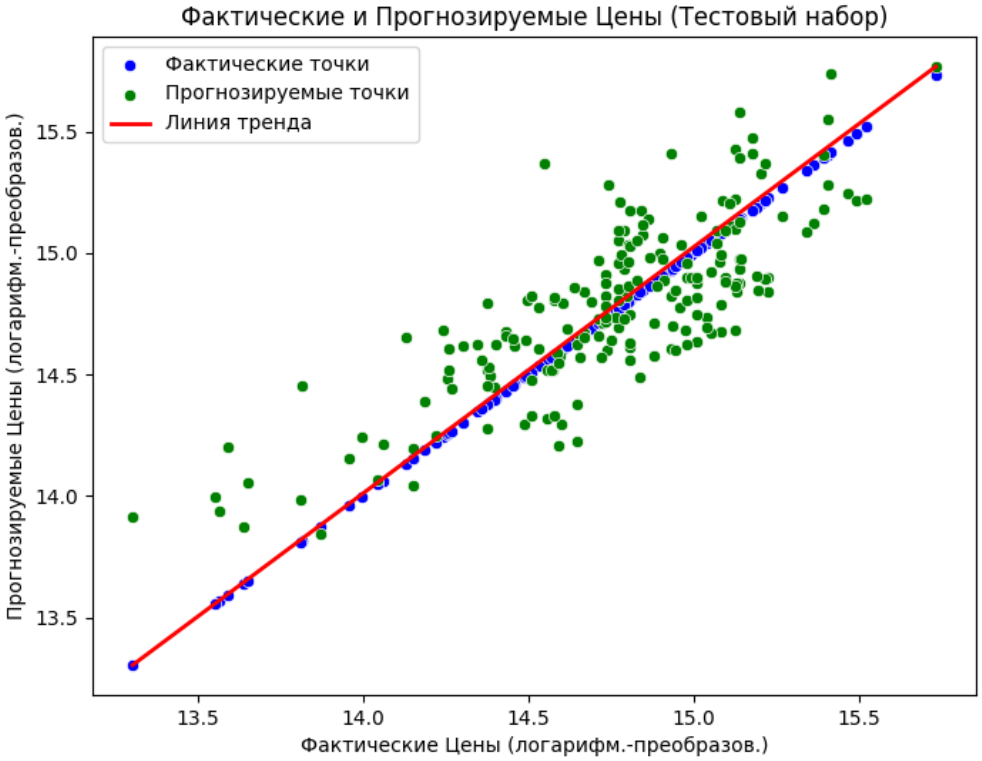

Рисунок 25 – Точечный график Линейной Регрессии

Как видно на (рис. 25), расхождение точек прогнозируемой цены от линии тренда может указывать на погрешности модели в определенных областях. Для лучшего понимания воспользуемся метриками эффективности в таблице 2.

Таблица 2 – Метрики эффективности для модели линейной регрессии

Линейная регрессия		
	Обучающая выборка	Тестовая выборка
MSE	0.062	0.059
R-squared	0.68	0.64
MAPE	20.72	20.74
MedAPE	17.22	18.3

Как видно в таблице 2, линейная регрессия, применяемая к данному набору данных, демонстрирует умеренную хорошую способность предсказывать цены на квартиры. Например, результаты на обучающем наборе данных, так как значение коэффициента детерминации составляет 0.68,

что говорит о том, что около 68% вариации зависимой переменной объясняется моделью. Однако на тестовом наборе данных модель не так хорошо справляется, так как значение коэффициента детерминации равно 0.64. Также среднеквадратическая ошибка на тестовом наборе данных составляет 0.059306, что относительно высоко, особенно учитывая, что ошибка на обучающем наборе данных равна 0.062765.

Результаты модели линейной регрессии, указывают на переобучении, это означает, что модель лучше подстроилась под обучающие данные и не обобщает свои знания на новые данные. Это может быть вызвано множеством факторов, таких как недостаточность количества данных, выбор неподходящих признаков, или использование модели слишком сложной для данной задачи, что может потребовать дополнительной настройки или использования более сложных методов регуляризации.

Для понимания значимости признаков, была построена таблица 3, в которой было выполнено преобразование значимости признаков в проценты для лучшей наглядности.

Таблица 3 – Значимость признаков для целевой переменной

Признаки	**Значимость (%)**
Кол-во комнат	0,86
Общая площадь	45,46
Жилая площадь	5,33
Площадь кухни	6,42
Этаж	1,43
Этажность	14,78
Широта	2,32
Долгота	1,38
Район	5,37
Материал дома	4,20
Ремонт	6,36
Наличие балкона	0,78
Вода	0,26
Санузел	0,74
Комнаты	4,30

Как видно из таблицы 3, наиболее значимыми признаками, являются общая площадь (45.46%), жилая площадь (5,33%) этажность (14.78%), площадь кухни (6.42%), материал дома (4,20%), ремонт (6,36%), комнаты (4,3%) и этаж (1,43%) эти технические признаки сильнее всего влияют на предсказание цен на недвижимость. Из географических признаков, стоит отметить важность района (5,37%), широта (2.32%) и долгота (1.38%), это может указывать на то, что местоположение объекта может влиять на цену. Существуют признаки с невысокой значимостью, такие как вода (0.26%), наличие балкона (0.78%), количество комнат (0,86%) и санузел (0.74%), эти признаки оказывают относительно слабое влияние на модель и скорее всего стоит рассмотреть возможность исключения признаков с низкой значимостью.

В целом, линейная регрессия, представляет собой эффективный инструмент для моделирования зависимости между различными параметрами и ценами на недвижимость, но для достижения более высокой точности, требуется рассмотреть нелинейную модель, которая также поможет уловить нелинейные связи. Следующий анализ будет проведён с использованием ансамблевого метода случайный лес.

2.5.2 Случайный лес

Для построения модели случайный лес, мы воспользуемся модулем RandomForestRegressor, для этого будет написана функция для обучения модели, где Y – зависимая переменная (цена), а X – независимые переменные (числовые и категориальные признаки).

В построении алгоритма случайного леса, мы будем использовать гипермаркеты (параметры), которые можно точно настроить для достижения большей точности модели машинного обучения, модели представлены на (рис. 26).

```
# Параметры для модели случайного леса
param_grid = {
 'max_depth': [90, 100, 105, 110, None],
 'max_features': ["sqrt", "log2", None],
 'min_samples_leaf': [1, 3, 5, 9],
 'min_samples_split': [2, 3, 4, 5, 10],
 'n_estimators': [100, 105, 110]
}
```

Рисунок 26 – Параметры модели Случайный лес

Как показано на (рис. 26) выбор параметров для модели случайного леса, осуществляется с целью достижения оптимальной производительности модели. Каждый параметр имеет свое влияние на процесс обучения и формирование модели. Ниже рассмотрим, почему были выбраны конкретные значения для каждого параметра:

1. Гипермаркет (max_depth): данный параметр определяет максимальную глубину деревьев в лесу. Увеличение глубины деревьев может привести к переобучению модели. В нашем случае были рассмотрены значения 90, 100, 105, 110 и None. Выбор глубины в пределах этого диапазона позволяет модели быть достаточно сложной для улавливания сложных закономерностей, но при этом избежать переобучения.

2. Гипермаркет (max_features): этот параметр определяет количество признаков, которые необходимо учитывать при каждом разделении. В вашем случае были рассмотрены значения sqrt (квадратный корень числа признаков), log2 (логарифм от числа признаков) и None (использование всех признаков). Ограничение количества признаков помогает уменьшить корреляцию между деревьями и улучшить обобщающую способность модели.

3. Гипермаркеты (min_samples_leaf и min_samples_split): эти параметры определяют минимальное количество образцов, необходимых для разделения узла и минимальное количество образцов, которые должны находиться в листовом узле. Они помогают предотвратить переобучение, за счет установки

минимального количества образцов, необходимых для того, чтобы разделение было считаться значимым.

4. Гипермаркет (n_estimators): данный параметр определяет количество деревьев в лесу. Большее количество деревьев может улучшить качество предсказания, но может также привести к увеличению времени обучения. Были рассмотрены значения 100, 105 и 110.

Выбор оптимальных значений этих параметров осуществляется через GridSearchCV, который перебирает различные комбинации параметров и оценивает их производительность с помощью кросс-валидации. Таким образом, выбор конкретных значений для каждого параметра основан на балансе между точностью модели и ее способностью к обобщению на новые данные.

По результатам GridSearchCV, были получены лучшие параметры: (max_depth: 100, max_features: sqrt, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 100). Эти параметры позволяют модели эффективно улавливать сложные закономерности в данных и обобщать их на новые случаи.

Также, были получены метрики эффективности модели, как на обучающей выборки, так и на тестовой, представленные в таблице 4.

Таблица 4 – Метрики эффективности для модели случайный лес

Случайный лес		
	Обучающая выборка	Тестовая выборка
MSE	0.01	0.06
R-squared	0.92	0.67
MAPE	9.87	19.73
MedAPE	8.09	16.58

В таблице 4 представлены результаты оценки модели на обучающей и тестовой выборках. На обучающей выборке модель достигла впечатляющих результатов: MSE - 0.01, R-квадрат - 0.92, MAPE - 9.87%, и MedAPE - 8.09%. Однако на тестовой выборке результаты были немного скромнее: MSE - 0.06, R-квадрат - 0.67, MAPE - 19.73%, и MedAPE - 16.58%.

Каждое дерево решений в модели случайного леса имеет высокую дисперсию, но при их параллельном объединении результирующая дисперсия становится невелика. Модель случайного леса более гибка и лучше адаптируется к сложным закономерностям в данных, чем линейная регрессия, но это может привести к переобучению.

Для модели случайного леса важно оценивать ее производительность на обеих выборках. Метрики на обучающей выборке помогают понять, как хорошо модель подгоняется под данные, а метрики на тестовой выборке демонстрируют ее способность к обобщению на новые данные.

Линейная регрессия обычно имеет меньшую сложность и меньшую склонность к переобучению, что позволяет оценивать ее производительность только на тестовой выборке. Однако модель случайного леса продемонстрировала хорошие результаты на обеих выборках, свидетельствуя о ее способности эффективно предсказывать целевую переменную (цену), представленную на (рис. 27), на основе предоставленных признаков.

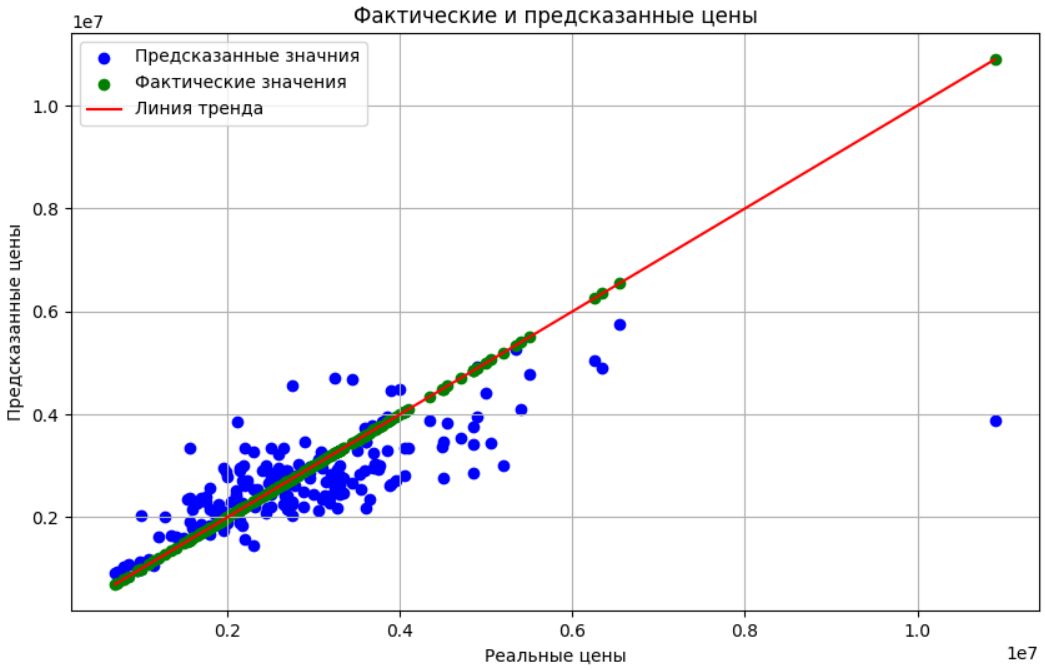

Рисунок 27 – Точечный график Случайного леса

Для понимания значимости признаков, была построена таблица 5, в которой было выполнено преобразование значимости признаков в проценты для лучшей наглядности.

Таблица 5 – Значимость признаков для целевой переменной

Признаки	**Значимость (%)**
Кол-во комнат	5,75
Общая площадь	27,36
Жилая площадь	17,11
Площадь кухни	16,30
Этаж	3,67
Этажность	8,77
Широта	4,33
Долгота	5,13
Район	2,14
Материал дома	1,50
Ремонт	2,12
Наличие балкона	1,54
Вода	0,34
Санузел	3,64
Комнаты	0,31

Как видно из таблицы 5, наиболее значимыми признаками для предсказания цен на недвижимость с использованием модели случайного леса оказались следующие характеристики: общая площадь (27.36%), жилая площадь (17.11%), площадь кухни (16.30%), количество комнат (5.75%) и этажность (8.77%). Эти факторы играют важную роль в формировании цены объекта недвижимости и сильно влияют на прогнозируемый результат.

Из географических признаков наибольшее влияние оказывают широта (4.33%) и долгота (5.13%). Это свидетельствует о значимости местоположения объекта недвижимости для его ценообразования.

Следует отметить, что признаки, такие как материал дома (1.50%), ремонт (2.12%), наличие балкона (1.54%), вода (0.34%) и санузел (3.64%), оказались менее значимыми для модели. Однако они все равно могут вносить свой вклад в предсказание цены на недвижимость.

В целом, результаты показывают, что модель случайного леса успешно выявила важные признаки, влияющие на цену недвижимости, следующей моделью станет Градиентный бустинг.

2.5.3 Градиентный бустинг

Для построения модели градиентный бустинг, мы воспользуемся модулем XGBRegressor, для этого будет написана функция для обучения модели, где Y – зависимая переменная (цена), а X – независимые переменные (числовые и категориальные признаки).

В построении алгоритма градиентный бустинг, мы будем использовать гипермаркеты (параметры), которые можно точно настроить для достижения большей точности модели машинного обучения, модели представлены на (рис. 28).

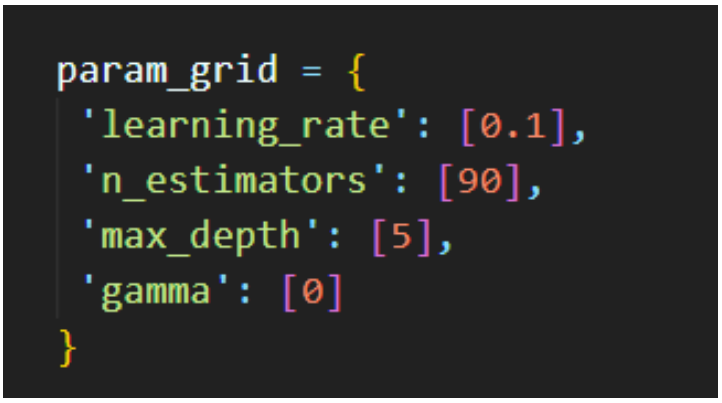

Рисунок 28 – Параметры модели Градиентный бустинг

Настройка параметров для модели градиентного бустинга (рис. 28) играет ключевую роль в оптимизации ее производительности. Давайте рассмотрим, почему мы выбрали определенные значения для каждого параметра:

1. Learning Rate (Скорость обучения): Этот параметр контролирует, насколько каждое новое дерево учитывает ошибки предыдущих деревьев. Малые значения означают более осторожное обучение, в то время как большие могут привести к переобучению. Мы выбрали значения (0.01, 0.05,

0.1, 0.2), чтобы оценить, насколько быстро или медленно модель должна учиться.

2. N Estimators (Количество деревьев): Этот параметр определяет, сколько деревьев будет добавлено к модели. Большее число деревьев может улучшить производительность, но также увеличить время обучения. Мы выбрали значения (80, 90, 100), чтобы найти оптимальное количество деревьев.

3. Max Depth (Максимальная глубина): Этот параметр определяет, насколько глубокими будут деревья в ансамбле. Слишком глубокие деревья могут привести к переобучению, а слишком мелкие — к недообучению. Мы выбрали значения (5, 6, 7), чтобы найти оптимальную глубину.

4. Gamma (Гамма): Этот параметр указывает, как много должна уменьшиться потеря, чтобы произвести дополнительное разделение в узле дерева. Мы выбрали значения (0, 0.1), чтобы настроить чувствительность к потерям при разделении узлов.

Выбор значений параметров модели осуществляется с целью создания разнообразных комбинаций, чтобы оценить их влияние на производительность. Через метод GridSearchCV мы перебираем различные комбинации и оцениваем их производительность с помощью кросс-валидации. Таким образом, конкретные значения выбираются на основе баланса между точностью модели и ее способностью к обобщению на новые данные.

После применения GridSearchCV были определены лучшие параметры: gamma: 0, learning_rate: 0.1, max_depth: 5, n_estimators: 90. Эти параметры обеспечивают эффективное улавливание сложных закономерностей в данных и обобщение их на новые случаи.

Также, были получены метрики эффективности модели, как на обучающей выборки, так и на тестовой, представленные в таблице 6.

Таблица 6 – Метрики эффективности для модели градиентный бустинг

Градиентный бустинг		
	Обучающая выборка	Тестовая выборка
MSE	0.006	0.05
R-squared	0.97	0.70
MAPE	5.99	19.08
MedAPE	4,41	15.5

Как показано в таблице 6, модель градиентного бустинга показала высокую эффективность как на обучающей, так и на тестовой выборках. На обучающей выборке значение среднеквадратичной ошибки (MSE) составило 0.006, что указывает на низкую ошибку модели при предсказании значений на этом наборе данных. Коэффициент детерминации (R-squared) на обучающей выборке составил 0.97, что говорит о том, что модель объясняет 97% дисперсии зависимой переменной. Средняя абсолютная процентная ошибка (MAPE) и медианная абсолютная процентная ошибка (MedAPE) на обучающей выборке также оказались низкими: 5.99% и 4.41% соответственно.

На тестовой выборке модель также показала хорошие результаты, но с небольшим ухудшением по сравнению с обучающей выборкой. Значение MSE на тестовой выборке составило 0.05, R-squared - 0.70, MAPE - 19.08%, и MedAPE - 15.5%. Это указывает на способность модели к обобщению на новые данные, хотя наблюдается некоторое увеличение ошибок по сравнению с обучающей выборкой.

Таким образом, модель градиентный бустинг продемонстрировала хорошие результаты как на обучающей, так и на тестовой выборках, что свидетельствует о ее способности эффективно аппроксимировать и предсказывать целевую переменную (цену), представленной на (рис. 29) на основе предоставленных признаков.

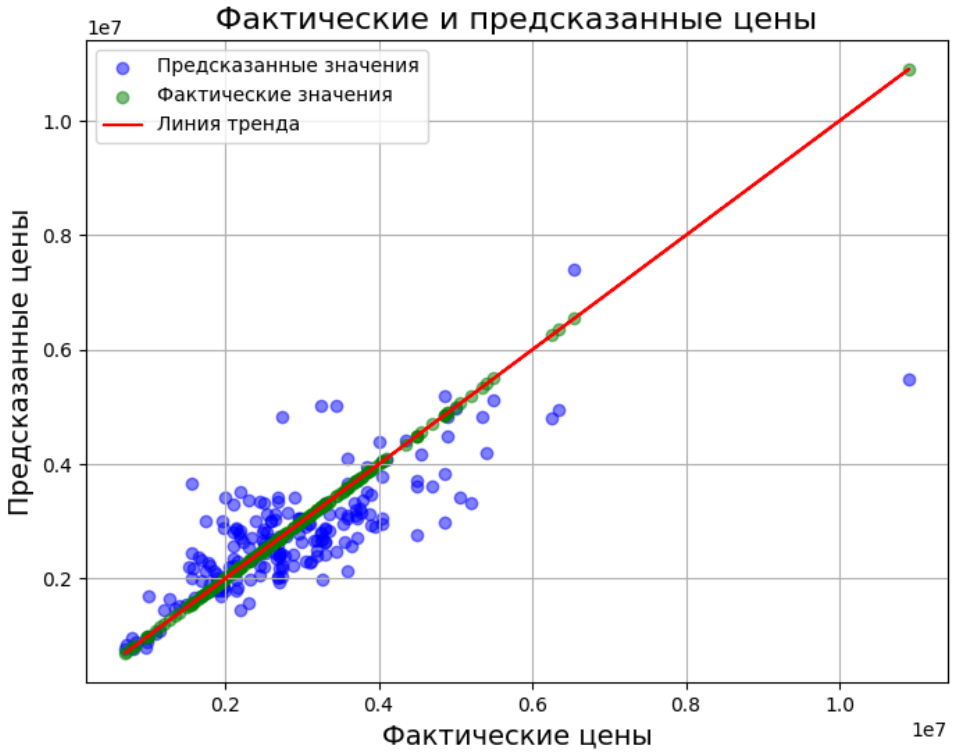

Рисунок 29 – Точечный график Градиентного бустинга

Для понимания значимости признаков, была построена таблица 7, в которой было выполнено преобразование значимости признаков в проценты для лучшей наглядности.

Таблица 7 – Значимость признаков для целевой переменной

Признаки	**Значимость (%)**
Кол-во комнат	0,67
Общая площадь	43,4
Жилая площадь	2,98
Площадь кухни	4,1
Этаж	2,91
Этажность	11,35
Широта	3,3
Долгота	3,63
Район	3,42
Материал дома	2,72
Ремонт	4,14
Наличие балкона	2,65
Вода	6,75
Санузел	3,45
Комнаты	3,64

Как видно из таблицы 7, наиболее значимыми признаками для предсказания цен на недвижимость с использованием модели случайного леса оказались следующие характеристики:

- Общая площадь (43.4%): данный признак оказался наиболее важным и составил почти половину от общей значимости всех признаков. Это подтверждает влияние общей площади объекта недвижимости на его стоимость, что логично, учитывая, что общая площадь является ключевым критерием при покупке жилья.

- Этажность (11.35%): этот признак также оказался важным, хотя и с меньшей значимостью по сравнению с общей площадью. Это указывает на то, что этажность здания также имеет существенное влияние на ценообразование в рынке недвижимости.

- Жилая площадь (2.98%), Площадь кухни (4.1%), Этаж (2.91%), Широта (3.3%), Долгота (3.63%), Район (3.42%): оставшиеся признаки также оказались важными, но с меньшей значимостью по сравнению с общей площадью и этажностью. Они все вносят свой вклад в прогнозирование цен на недвижимость и являются важными факторами при принятии решений о покупке или продаже недвижимости.

Из географических признаков наибольшее влияние оказали широта и долгота, что указывает на значимость местоположения объекта недвижимости для его стоимости.

Следует отметить, что признаки, такие как материал дома, ремонт, наличие балкона, вода и санузел, оказались менее значимыми для модели, но все же вносят свой вклад в прогнозирование цен на недвижимость.

Таким образом можно сделать выводы по 3 моделям регрессии. Из таблицы 8 видно, что метод градиентного бустинга демонстрирует наилучшую производительность по всем метрикам на тестовой выборке. Он превосходит как линейную регрессию, так и случайный лес, обеспечивая наивысший коэффициент детерминации (R^2) и наименьшие значения среднеквадратичной ошибки (MSE), средней абсолютной процентной ошибки

(MAPE) и медианной абсолютной процентной ошибки (MedAPE). Это указывает на то, что модель градиентного бустинга лучше всего подходит для предсказания цен вторичного рынка недвижимости с использованием машинного обучения в данном контексте.

Таблица 8 – Результаты моделей регрессии

Метод	R^2		MSE		MAPE		MedAPE	
	Train	Test	Train	Test	Train	Test	Train	Test
Linear Regression	0.68	0.64	0.062	0.059	20.72	20.74	17.22	18.3
Gradient Boosting	0.97	0.7	0.006	0.05	5.99	19.08	4.41	15.5
Random Forest	0.92	0.67	0.01	0.06	9.87	19.73	8.09	16.58

ЗАКЛЮЧЕНИЕ

В рамках данной работы были рассмотрены основные типы и методы машинного обучения, а также характеристика рынка недвижимости г. Липецка. Проведён кластерный и регрессионный анализ, основанные на методах машинного обучения.

Организацией АН «Квадрат», была предоставлена информация по объектам из их базы данных за период 2020-2023 гг.

В ходе проведения работы было выявлено, что применение методов машинного обучения в анализе вторичного рынка недвижимости г. Липецка, позволило выделить группы квартир по их стоимости и расположению с помощью метода кластерного анализа: k-средних, а также значительно повысить точность прогнозов цен с помощью регрессионных моделей: линейная регрессия, случайный лес и градиентный бустинг.

Кроме того, дополнительно корреляционный анализ позволил выделить ключевые признаки, влияющие на цены в каждом кластере, как положительно, так и негативно.

Таким образом, данная работа показала, что внедрение методов машинного обучения становится неотъемлемой частью современного анализа рынка недвижимости. Данное исследование не только раскрывает потенциал этих технологий, но и формирует основу для их практического применения, что в конечном итоге способствует улучшению качества управления недвижимостью и повышению конкурентоспособности участников рынка.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. **Nils Kok.** Big Data in real estate from manual appraisal to automated valuation / Nils Kok, Eija-Leena Koponen, Carmen Adriana Martinez-Barbosa // The journal of portfolio management. – 2017. – № 43(6). – p. 202-211.

2. Анализ данных с помощью Python / [Электронный ресурс] // GitHub: [сайт]. – URL: https://github.com/calistus-igwilo/House-sale-price-prediction-using-python/blob/master/Housing%20Price%20Prediction.md (дата обращения: 04.02.2024).

3. Ансамблевые методы / [Электронный ресурс] // scikit-learn: [сайт]. – URL: https://scikit-learn.ru/1-11-ensemble-methods/ (дата обращения: 14.03.2024).

4. **Астапов, Р. Л.** Автоматизация подбора параметров машинного обучения и обучение модели машинного обучения / Р. Л. Астапов, Р. М. Мухамадеева // Национальный исследовательский ядерный университет «МИФИ». – 2021. – № 5-2 (73). – С. 34-37.

5. **Баймуратов, И. Р.** Методы автоматизации машинного обучения: учебное пособие / И. Р. Баймуратов. – Санкт-Петербург: НИУ ИТМО, 2020. – 40 с.

6. **Елистратова, Е.** Градиентный бустинг / Е. Елистратова, К. Лунёв [Электронный ресурс] // Яндекс образование: [сайт]. – URL: https://education.yandex.ru/handbook/ml/article/gradientnyj-busting#spisok-literatury (дата обращения: 05.03.2024).

7. **Бородкин, К. В**. Исследование эффективности методов машинного обучения в задачах прогнозирования / К. В. Бородкин, Р. Р. Отырба, А. С. Пешков, М. А. Дрюченко // Сборник студенческих научных работ факультета компьютерных наук ВГУ. Министерство Науки и Высшего Образования РФ Федеральное Государственное Бюджетное Образовательное Учреждение Высшего Образования «Воронежский Государственный Университет». Том Выпуск 14 Часть 2. – Воронеж: Издательский дом ВГУ, 2020. – С. 28-36.

8. Кластеризация / [Электронный ресурс] // scikit-learn: [сайт]. – URL: https://scikit-learn.ru/clustering/ (дата обращения: 10.03.2024).

9. **Кугаевских, А. В.** Классические методы машинного обучения: учебное пособие / А. В. Кугаевских, Д. И. Муромцев, О. В. Кирсанова. – Санкт-Петербург: НИУ ИТМО, 2022. – 53 с.

10. **Лейфер, Л. А.** Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир / Л.А. Лейфер, Е.В. Чёрная // КиберЛенинка. – 2020. – № 3. – С. 32-42.

11. **Митина, О. А.** Технологии и инструментарий машинного обучения: учебное пособие / О. А. Митина, В. В. Жаров. – Москва: РТУ МИРЭА, 2023. – 76 с. – ISBN 978-5-7339-1758-0.

12. Обучение без учителя: 4 метода кластеризации данных на Python / [Электронный ресурс] // proglib: [сайт]. – URL: https://proglib.io/p/unsupervised-ml-with-python (дата обращения: 05.02.2024).

13. **Полищук, О. А.** Машинное обучение как современный подход к анализу данных / О. А. Полищук, А. Д. Мартыничева, П. Д. Егоров // Юго-Западный государственный университет. – 2022. – № 2. – С. 64-73.

14. Регрессия с повышением градиента / [Электронный ресурс] // scikit-learn: [сайт]. – URL: https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html#plot-training-deviance (дата обращения: 12.03.2024).

15. **Саксонов, П. В.** Обзор методов машинного обучения / П. В. Саксонов, А. А. Бауман // Современные тенденции и инновации в науке и производстве: Материалы XII Международной научно-практической конференции, Междуреченск, 26 апреля 2023 года / Редколлегия: Т. Н. Гвоздкова (отв. редактор), С. О. Марков [и др.]. – Междуреченск: Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2023. – С. 4441-4446.

16. **Себастьян, Р.** Python и машинное обучение: учебник и практикум / Р. Себастьян, В. Мирджалили. – 3-е изд. Пер. с англ. – СПб.: ООО «Диалектика», 2020. – 848 с. – ISBN: 978-5-907203-57-0.

17. **Сохина, С. А.** Машинное обучение. Методы машинного обучения / С. А. Сохина, С. А. Немченко // Современная наука в условиях модернизационных процессов: проблемы, реалии, перспективы: Сборник научных статей по материалам V Международной научно-практической конференции (Уфа, 30 апреля 2021 г.). – Уфа: Общество с ограниченной ответственностью "Научно-издательский центр «Вестник науки»", 2021. – С. 165-168.

18. **Страхов, Е.** Деревья решений и случайный лес / Е. Страхов [Электронный ресурс] // kaggle: [сайт]. – URL: https://www.kaggle.com/code/emstrakhov/decision-trees-and-random-forest (дата обращения: 03.03.2024).

19. **Штукатуров, С.** Реализация и разбор алгоритма «случайный лес» на Python / С. Штукатуров [Электронный ресурс] // Tproger: [сайт]. – URL: https://tproger.ru/translations/python-random-forest-implementation (дата обращения: 03.03.2024).

20. **Эрик, К.** Прогнозирование цен на жилье с помощью машинного обучения / К. Эрик [Электронный ресурс] // kaggle: [сайт]. – URL: https://www.kaggle.com/code/erick5/predicting-house-prices-with-machine-learning/notebook (дата обращения: 07.02.2024).