* Введение в оценку почему это важно и почему важно автоматизировать и смоделировать это
* Теоретические аспекты именно рынка недвижимости, потом рынка коммерческой недвижимости и затем сферы оценки
* Поиск и сбор данных, написание программ для сбора готовых данных
* Анализ собранных данных с использованием как питона, так и кьюгиса
* Модели
* Практическое применение

РАССМОТРЕТЬ МОДЕЛИ В ТЕОРИИ, ПИТОН И ГИТХАБ

В условиях современной экономической системы, сложившейся в Российской Федерации – роль рынка недвижимости очень велика. Только по прогнозным расчеты за 2024 год, операции по недвижимости составляют 9,2% от ВВП, почти аналогичный результат – 9,1% за 2023. В особенности очень важна коммерческая недвижимость – как показатель вовлеченности населения в образование бизнесов. Экономика коммерческой недвижимости вышла из пандемийного кризиса, и, несмотря на социально-экономическую нестабильность, сложившуюся в 2022 году - показывает динамичное развитие в последние годы, тому доказательство - увеличение объема инвестиций в 1,5 раза по всему коммерческому сектору, по торговым объектам недвижимости – в 2 раза.

Для успешного функционирования рынка такого масштаба и объема - очень важно иметь структурированную систему оценивания объектов недвижимости, которая будет функционировать внутри самого рынка

Ц

**Теоретические аспекты**

При изучении сферы коммерческой недвижимости, первоочередно необходимо определить используемое понятие не только коммерческой недвижимости, но и что подразумевается под недвижимостью и сферой недвижимости в экономике, какие нормативные акты регулируют данную сферу, на каких принципах построен коммерческий рынок.

Согласно Гражданскому кодексу Российской Федерации (ст. 130), к недвижимым вещам (недвижимое имущество, недвижимость) относятся земельные участки, участки недр и все, что прочно связано с землей, то есть объекты, перемещение которых без несоразмерного ущерба их назначению невозможно, в том числе здания, сооружения, объекты незавершенного строительства. Аналогично 130 статье Гражданского кодекса - Федеральный закон № 122-ФЗ от 21 июля 1997 года «О государственной регистрации прав на недвижимое имущество и сделок с ним» (ст. 1) определяет недвижимое имущество как земельные участки, участки недр и все объекты, которые связаны с землей так, что их перемещение без несоразмерного ущерба их назначению невозможно, в том числе здания, сооружения, жилые и нежилые помещения, предприятия как имущественные комплексы.

Исходя из указанного, упрощенным для восприятия термином недвижимости можно считать следующее – объекты (здания, сооружения, жилые и нежилые помещения, предприятия), перемещение которых невозможно без причинения ущерба объекту и его назначению или ограничено по тем же причинам, есть недвижимое имущество.

В таком случае, рынок недвижимости представляет собой систему взаимодействий различных уровней и механизмов, которые регулируют процессы создания, управления, передачи и использования объектов недвижимости для конкретных целей.

С учетом того, что недвижимое имущество делится на жилое и нежилое – недвижимость по своему назначению также делится на 2 вида: жилая и коммерческая. Ввиду того, что в данной работе проводится исследование и оценка коммерческой недвижимости, рассматривать тоже имеет смысл именно её.

Коммерческая недвижимость – это собственность, используемая исключительно для решения бизнес-ориентированных задач, то есть получения дохода, или для создания рабочего места, вместо жилого помещения. 8 Чаще всего коммерческая недвижимость сдается в аренду арендаторам для ведения приносящей доход деятельности. Можно выделить несколько признаков нежилого помещения:

* Юридический статус – «нежилое».
* Наличие отдельного входа, изолированного от входа в жилой дом.
* Нельзя зарегистрировать граждан, ведь оно не предназначено для проживания.
* Можно использовать для регистрации юридического адреса ООО.
* Повышенные требования к содержанию.

Разделяют три основных вида коммерческой недвижимости:

* Офисные объекты, коммерческие объекты, предназначенные для размещения предприятий, организаций и профессиональных служб. Могут быть как небольшими помещениями, так и крупными комплексами. По своей сути делятся на три категории, в зависимости от качества самого объекта, его местоположения, оснащения и арендной ставки: класс А, класс В, класс С.
* Торговые помещения или помещения свободного типа — это коммерческие объекты, предлагающие потребительские товары и услуги. Классифицируются в зависимости от их местоположения и состава арендаторов.
* Промышленные (складские) объекты представляют собой помещения, предназначенные для производства товаров, их хранения или для организации функционирования логистической системы. Разделяют на склады для массовых грузов, гибкие помещения, промышленные объекты, а также холодильные камеры.

Понимая устройство недвижимости, важно также определить принципы, используемые при ее оценке.

**ПРИНЦИПЫ ОЦЕНКИ НЕДВИЖИМОСТИ**

Оценкой недвижимости принято считать процесс определения рыночной или иной стоимости недвижимости и создание определенного отчета о процессе и результатах оценки, то есть официального документа, подтверждающего ликвидность объекта и объективность и независимость самой оценки. Сама оценка рыночной стоимости недвижимого имущества на территории Российской Федерации выполняется в соответствие с требованиями Федерального закона № 135-ФЗ от 29.07.98 г. «Об оценочной деятельности в Российской Федерации», Приказа МЭРТ РФ № 256 от 20.07.2007 г. «Об утверждении федерального стандарта оценки «Общие понятия оценки, подходы к оценке и требования по проведению оценки (ФСО № 1)», Приказа МЭРТ РФ № 255 от 20.07.2007 «Об утверждении федерального стандарта оценки «Цель оценки и виды стоимости (ФСО № 2)», Приказа МЭРТ РФ № 254 от 20.07.2007 «Об утверждении федерального стандарта оценки «Требования к отчету об оценке (ФСО № 3)», Международных стандартов МСО-2005, Европейских стандартов оценки ЕГАО, 2003.

С учетом изучения указанных выше документов – рыночной стоимостью объекта оценки принято считать наиболее вероятную цену, по которой данный объект может быть отчужден на открытом рынке и в условиях конкуренции, то есть обе стороны сделки действуют разумно, располагают всей необходимой информацией, а на стоимость объекта не влияют чрезвычайные обстоятельства, а именно:

* Одна из сторон сделки не обязана отчуждать объект оценки, а другая сторона не обязана принимать исполнение;
* Стороны сделки хорошо осведомлены о предмете сделки и действуют в своих интересах;
* Объект оценки представлен на открытый рынок в форме публичной оферты, типичной для аналогичных объектов оценки;
* Цена сделки представляет собой рыночно-ориентированное вознаграждение за объект оценки;
* Участие в сделке добровольное для обеих сторон;
* Платеж за объект оценки должен быть выражен в денежной форме.

Сама же оценка должна нести в себе не только объективность и независимость, но и структурность, системность, поэтапность в процессе. Поэтому есть определенный порядок оценки:

1. Заключение договора о проведении услуг об оценке объекта(ов);
2. Установление количественных и качественных характеристик объекта;
3. Анализ аналогичного объекту рынка;
4. Выбор методов оценки объекта и проведение соответствующих расчетов;
5. Обобщение результатов и составление отчетов по деятельности;
6. Подготовка и передача заказчику отчета об оценке.

Обращаясь к методам оценки, важно правильно и точно определять, каким методом лучше воспользоваться, в зависимости от поставленных задач, для этого необходимо безусловное понимание существующих методов и принципов пользования ими. Тем не менее, важно применять данные методы в совокупности, для более структурного подхода.

Доходный подход – совокупность методов оценки, основанных на определении потенциальных доходов от использования объекта оценки – аренды или перепродажи. Одним из основных методов расчета потенциального дохода (ПД) стоит считать следующую математическую модель:

где S – площадь, сдаваемая в аренду (м2);

Ca – арендная ставка (руб./м2).

Сравнительный подход – основной подход, основанный на сравнении объекта с аналогичным объектами на рынке. Аналоги определяются учитывая технические характеристики объектов, их геопозицию и планировку. Математическая модель оценки объекта недвижимости с использованием следующего подхода будет представлена следующим образом:

где k - количество аналогов,

– оценка рыночной стоимости объекта оценки,

– оценка стоимости объекта аналога,

– вклад i-го аналога в стоимость объекта,

Затратный подход – совокупность методов оценки, основанных на определении затрат, необходимых для воспроизводства или замещения объекта оценки с учетом износа или устаревания. Рассмотрим подход с точки зрения одного из вариантов математической модели:

где C – стоимость оцениваемого объекта (руб.);

С кв. – стоимость кв. метра (руб./м2);

S – площадь объекта (м2);

k1 – коэфф. разницы между объектом и аналогом;

k2 – коэфф. изменения стоимости строительства;

k3 – коэфф. прибыли застройщика;

k4 – коэфф. НДС.

При проведении оценки каждым из указанных методов необходимо соблюдать систему принципов оценки, ведь на данной системе строится сама сфера оказания данной услуги. Здесь имеет влияние и принципы, согласно которым работает рыночная среда недвижимости, и принципы отношения к объекту собственником.

Если изучать последние – становится понятно, что важными элементами являются полезность, рынок, а также долгосрочной доходности. Соответственно, принципом полезности стоит считать способность объекта в удовлетворении имеющихся потребностей, которые собственник будет требовать с данного объекта. При увеличении полезности – соразмерно увеличивается и оценочная стоимость. Принцип рынка указывает на возможность потенциального собственника выбрать объект с наименьшей стоимостью эквивалентной полезности. Принцип долгосрочной доходности подтверждает ожидания инвестора получить доходы, со вложенных в объект средств, в будущем. Следовательно сумма потенциальных доходов образовывает оценочную стоимость.

Изучая принципы рыночной среды, которые влияют на оценку объекта важно отметить следующие. Принцип эффективных вложений – вклад в объект эффективен в том случае, если он повышают рыночную стоимость объекта и не превышает полученный прирост в стоимости. Принцип спроса и предложения – цены реагируют в соответствие с законами о спросе и предложении, то есть стабильны в случае равновесия, наличие дефицита предложения повышает цену и превышение предложения над спросом соответственно понижает цену. Принцип конкуренции – фактор конкуренции уравнивает потенциальную доходность вложений. Принцип эффективного использования – Объект используется наиболее выгодным способом, учитывая альтернативные издержки, физические возможности, законодательство, максимальную возможную продуктивность.

НАПИСАТЬ ГЛАВУ ПРО АНАЛИЗ НЫНЕШНЕГО РЫНКА

Данные принципы и основы рынка недвижимости используются повсеместно уже продолжительно время, однако с развитием новых технологий, появлением машинного обучения, искусственного интеллекта и всецелым развитием рынка – важно учитывать не только базовую материальную часть, на которой строится оценка, но и как оценка развивается как наука, как сфера оказания услуг, какие методы используются для улучшения качества и точности оценки, исключая факторы человеческой ошибки, узкой выборочности данных и условной субъективности. Всецелом, важна информация о направлении, в котором движется данная наука, для использования соответствующих стандартов при проведении такого рода исследования. Поэтому важным элементом работы стоит считать исследование передовых, как зарубежных, так и отечественных, работ в данной сфере. Особенно в рамках использования машинного обучения для построения модели оценки стоимости коммерческого объекта недвижимости.

**АНАЛИЗ ИСТОЧНИКОВ**

Для изучения моделирования оценки недвижимости важно понимать:

* Насколько сильно отличается моделирование коммерческой недвижимости от жилой?
* Какие данные необходимы для моделирования стоимости коммерческой недвижимости?
* С какими моделями лучше всего подходить к оценке данным такого рода?
* Какие результаты по аналогичным моделям?

Исходя из изученных зарубежных источников на тему первого вопроса, в особенности на примере следующего источника [Joseph Gyourko UNDERSTANDING COMMERCIAL REAL ESTATE: JUST HOW DIFFERENT FROM HOUSING IS IT? // NBER WORKING PAPER SERIES. – 2009], на основе проведенных исследований, расчетов и выявлений закономерностей на исторических данных, следует заключение, что рынки жилой и коммерческой недвижимости зависят от общих фундаментальных факторов. С учетом изменений, происходящих в разных структурах экономики, в разных регионах и не смотря на факт разных продолжительностях бизнес-циклов в этих двух сферах рынка недвижимости – что жилая, что коммерческая недвижимость показывают очень схожие тенденции и темпы роста стоимости. Это объясняется достаточно сильной зависимостью от макропараметров экономики и положения, в котором находится рынок недвижимости в целом.

<https://irnr.ru/wp-content/uploads/Концепция/Ликвидность/210_2_2023.pdf> ИНФОРМ-ОЦЕНКА

Тем не менее, при рассмотрении ситуации на отечественном рынке недвижимости – аналогичного заключения сделать не получается. При рассмотрении аналитического исследования по рынку недвижимости на территории РФ, и в особенности в регионе с самой большой ликвидностью коммерческой недвижимости – Санкт-Петербурге, мы видим, что средняя экспозиция объектов на рынке коммерческой недвижимости с 2022 по 2023 увеличилась во всех секторах, на 20% в торговом секторе, 15% в сегменте офисной недвижимости и 10% в складской. Помимо этого, коммерческие сегменты рынка показывают падение спроса при высоком предложении, за исключением производственно-складской недвижимости, показывающие хорошие результаты с учетом развитие электронной коммерции и расширения логистических сетей.

В то же время, жилая недвижимость в Санкт-Петербурге показывает диаметрально противоположные результаты. Высокий спрос, низкое предложение, рост цен на недвижимость и сильный приток населения. При этом средняя продолжительность экспозиции на первичном рынке уменьшилась на 13% с января 2022 года по январь 2023. При этом, при сравнении самой продолжительности экспозиции – коммерческая выше в несколько раз.

При изучении передовых работ на рынке коммерческой недвижимости, затрагивающих основы оценки недвижимости и содержащие основную материальную часть, касающуюся данного сектора – можно сделать выводы об основных факторах, оказывающих влияние на стоимость объекта. [https://iksadyayinevi.com/wp-content/uploads/2021/11/MODERN-METHODS-APPROACH-IN-REAL-ESTATE-VALUATION.pdf]

Первостепенным фактором, влияющим на стоимость ультимативно считается геопозиция объекта, дистанция объекта от центра, доступ к транспорту и объектам общественного пользования (больницы, образовательные учреждения, торговые центры). Помимо этого, стоимость — это технические характеристики объекта, это также площадь и форма помещения, возраст объекта, срок экспозиции на рынке, качество материалов построенного здания, инфраструктура и доступ к парковочным местам и другие особенности, зависящие индивидуального от каждого покупателя и собственника. В моменте оценки также учитываются макроэкономические факторы.

Подходя однако к вопросу изучения современных подходов в оценке недвижимости – важно обратить внимание на все возможные варианты, их плюсы, минусы и особенности. Обратившись к зарубежной работе, рассматривающей результаты работ, в которых были изучены подходы к оценке недвижимости [MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW], можно заметить, что 11 из 24 лучших результатов в исследованиях показывали модели, основанные на алгоритме случайного леса. Алгоритм случайного леса, он же Random Forest – алгоритм машинного обучения, заключающийся в использовании совокупности деревьев решений. Данный алгоритм представляется очень гибким, универсальным, а также устойчивым к межфакторной корреляции, что вполне объясняет, почему он показывал лучшие результаты и в других исследованиях, проведенных на похожую тему.

Однако для определения наилучшего результата, будет важно учитывать все доступные модели, для выявления особенностей данных и понимания всех возможных подходов к данному вопросу, оставаясь при этом объективным.

**СБОР ДАННЫХ**

Для сбора данных такого рода не будет достаточным усилием обратиться к собранным базам данных, доступным в открытом доступе. Данные такого рода представляют собой достаточно узкую выборку, рамки которой были определены исключительно самим автором собранной выборки. Ввиду невозможности оценить корректность и актуальность данных такого рода – необходимо обратиться к рынку в настоящем моменте, с целью собрать все реальные объекты с нынешними ценами. Помимо этого, важно, чтобы эти данные были в открытом доступе и сам процесс сбора был автоматизирован и доступен к использованию в любое необходимое время, в особенности из-за необходимости большого количества самих наблюдений, которые в данном случае будут представлять собой объектами недвижимости. Единственным из современных доступных способов сбора данных при таких условиях является парсинг.

Парсинг – это процесс автоматизированного сбора данных из электронного ресурса, с помощью кода, программы или сервиса. Для написания такого рода программы необходимо обратиться к языку программирования Python и к библиотекам, доступным в нем – в частности, к основным библиотекам, позволяющим эффективно осуществлять парсинг – «Selenium» и «BeautifulSoup». Изучив API данных библиотек, стали понятны принципы, по которым будет необходимо собирать данные, учитывая эту информацию – первостепенной задачей становился сам процесс выбора основного источника данных, который предложит наибольшее количество доступных и необходимых параметров, даст необходимое количество наблюдений, а также будет представлять собой качественный набор данных – полный, актуальный и корректный.

В процессе исследования основных источников на тему парсеров, их типов, принципов использования и основных библиотек на языке программирования python возникла необходимость также и в изучении основ написания сайтов – html, для понимания структуры страниц и данных на этих страницах.

После изучения современных электронных площадок для объявлений по коммерческой недвижимости в Санкт-Петербурге, в частности по торговому сектору недвижимости, наилучшее качество данных, вместе с набором доступных параметров и пригодным для парсинга устройства сайта, демонстрировал сайт объявлений Циан. Помимо этого, Циан предоставляет значительное количество объявлений, подходящих под определенные нами параметры.

Для того, чтобы собрать все объекты, первоначально необходимо собрать все ссылки на них с предварительно настроенной фильтрацией сайта. Был написан код для парсинга URL-ссылок всех объектов в сфере коммерческой недвижимости в Санкт-Петербурге.

Данный код включает в себя цикл с искусственными задержками, для прохождения по каждому из 34 объявлений на всех 50 страницах сайта с определенным нами заранее условиями фильтрации. В цикле предусмотрено использование скрипта для вертикального перемещения по странице, с целью загрузки и обработки всех объявления, взятие ссылок опубликованных объявлений, учитывая структуру html-кода сайта «Циан».

Также было предусмотрено возникновение ошибок, почему после возникновения ошибки сам процесс парсинга не прекращается. Каждая ссылка объявления, при этом, записывается в список и затем сохраняется.

После успешного парсинга ссылок на объекты, возникает надобность парсинга самих объявлений, а для данной работы необходимо провести доскональное изучение кода страницы объявлений чтобы ответить на основные вопросы при такого рода парсинге.

Какие объекты размещены? Какой путь они имеют? В каком формате те или иные данные записываются? Имеет ли значение в каком порядке эти данные собираются? Обратившись к структуре кода одной из страниц, была написана программа-парсер [приложение А].

В данном коде происходят действия сбора данных по циклу всех объектов. При открытии каждого объявления, происходит задержка в 10 секунд для полной загрузки страницы и, соответственно, для корректного функционирования парсера. На каждой странице происходит обращение к следующим объектам с извлечением текста и дальнейшим сохранением его, вместе с ссылкой на данное объявление в соответствующем с кодом порядке:

* Название объявления;
* Цена объекта;
* Цена за квадрат;
* Площадь м^2;
* Этаж;
* Максимальный этаж здания
* Свободно ли помещение;
* Город;
* Район;
* Округ;
* Улица;
* Номер дома;
* Ссылка на страницу продавца;
* Описание.

Работа данного кода суммарно заняла приблизительно 7 часов с учетом обработки всех 1348 объявлений, полученных с работы предыдущего парсера. Такая длительность объясняется наличием задержек для прогрузки страниц и дополнительного времени для сбора информации с каждой страницы.

Также стоит отметить, что все эти данные нуждаются в тщательной обработке, ведь все они имеют на данный момент текстовый или даже списочный формат, а некоторые и вовсе могут не подразумевать какой-либо практической ценности. Однако, прежде чем обратиться к обработке данных, необходимо собрать координаты для этих данных.

**РАБОТА С КООРДИНАТНЫМИ ДАННЫМИ**

Для обработки адресов и перевода их в систему координат было решено обратиться к геокодеру от компании «Яндекс». Был изучен API геокодера и способы его применения. Первоначально стоит обратиться к термину.

Геокодер — это программа, которая расшифровывает адрес текстового формата в географические координаты согласно открытой базе данных того или иного источника.

После ознакомления с особенностями геокодера – был создан ключ, для использования его внутри среды кода на языке программирования python. Реализована функция для отправления запросов в виде адреса и получения результата в формате координат. Имелись определенные ограничения для бесплатного использования геокодера в рамках научной работы, одним из таких ограничений, оказавших влияние на обработку данных, был лимит на кол-во запросов в сутки – 1000. Также были учтены возможные возникновения ошибок или факт ненайденных координат.

Данная функция была использована в цикле для получения координат по всему списку данных. Так как данных больше 1000, добавление координат было в несколько этапов. Вместе со вторым этапом также были собраны координаты станций метрополитена в Санкт-Петербурге. На момент отправления запросов – их число составляло 72.

Для получения максимума пользы из информации по метро, были рассчитаны расстояния от объектов до всех станций метро, чтобы потом была возможность определить количество станций метро возле объектов. Для этого пришлось обратиться к формуле расчета расстояния между двумя точками на координатной плоскости Земли, ввиду технической невозможности расчета времени пути пешим ходом, на машине и на общественном транспорте. Большинство отечественных сервисов, предлагающих такие функции, не были в открытом доступе, а зарубежные сервисы, хоть и ультимативно предлагали такой доступ – показывали некорректные значения, ввиду недостатка данных, а также были недоступны для использования из Российской Федерации.

Где:

L — искомое расстояние между пунктами 1 и 2;

q — расстояние в одном градусе, константа ~ 111.2 км;

φ₁, φ₂ — географические широты пунктов 1 и 2;

λ₁, λ₂ — географические долготы пунктов 1 и 2.

Используя эту формулу, были произведены расчеты расстояний между каждым объектом и каждой станцией метро. Помимо этого, предварительно была найдена база данных по Санкт-Петербургу, включающая основные объекты, по типу зданий, парковок, остановок общественного транспорта, дорог и других данных, не представлявших особого научного интереса для поставленных в данной работе задач.

Для поставленной задачи – прогнозирования стоимости квадратного метра торгового помещения – наличие парковок является достаточно важным фактором. Это подтверждается и с точки зрения рассуждения, и при обращении к источникам, изучающим факторы влияния на стоимость объектов. Данные по координатам парковок были введены, а затем обработаны с учетом расстояния до объектов. Был добавлен столбец, содержащий количество парковок в радиусе одного километра от каждого из объектов.

К имеющимся данным было решено добавить разделение объектов на районы Санкт-Петербурга, используя фиктивные переменные. При этом стоит отметить, что в выборке не оказалось данных по следующим районам: Колпинский, Курортный, Кронштадтский и Петродворцовый. В связи с чем, данные факторы были убраны из наших данных.

Для выполнения этой задачи был использован вышеуказанный набор геопространственных данных по Санкт-Петербургу, разделяющий город на существующие 18 районов. Из имеющегося набора координат в рамках системы координат – были собраны названия районов в отдельный список. Данные были конвертированы в формат полигональных данных и затем, с помощью функции для проверки, в какой из районов входит объект и последующего обновления столбцов соответствующих районов, данные были обработаны и приведены в надлежащий вид.

**ОБРАБОТКА ДАННЫХ**

Обработка первоначальных, так называемых «грязных» данных, является одной из важнейших частей работы с данными. Очень важно не потерять данные, обнаружить все ошибки и аномалии, ведь для построения дальнейших моделей нужно иметь подготовленные данные, не вызывающие, как и логических, так и технических конфликтов.

В первую очередь было необходимо определить столбцы, которые могли быть полезны для нас и нести смысл, который позволят, или более подробно изучить объект, или окажутся значимыми факторами для построения модели. Также, стоит отметить, что отсутствует необходимость обработки параметров, которые были добавлены отдельно от основной выборки, ведь произведенные расчеты были точны.

Одним из важных факторов является площадь самого объекта. Такие данные нам нужны в числовом формате, а именно – в десятичном, ведь площадь может быть не целым числом. Для обработки данного столбца были использованы функции для замены лишних пробелов и очищения символов единиц измерения площади и любых других символов возле необходимых нам данных. Используя указанные функции, был очищен столбец от лишних пробелов и была произведена замена запятых на точки, после чего столбец был переведен из текстовых значений в десятичные числа. Похожие действия были произведены со столбцами полной цены объекты и цены квадратного метра в нем.

Столбец с признаком, отвечающим за свободность объекта на данный момент – были заменены соответствующие записи на 0 в случае, если объект свободен, и 1, если он занят. При подробном изучении было найдено определенное количество объектов, которые содержали не точные данные касательно занятости объекта, поэтому они предварительно были помечены как занятые.

Согласно сайту «Циан», в объявлении указывался этаж самого объекта, а также максимальный этаж здания, в котором объект находился. Для корректного вывода информации и сохранении этажа и максимального этажа здания – была написана функция [приложение Б] для вывода этих данных и созданы два отдельных столбца для каждого из признаков.

Наличие подробного адреса – один из важных фактов выбора именно площадки «Циан», ведь с помощью точного адреса есть возможность преобразовать такого рода данные в координаты и затем использовать их для геоспатического анализа и построения моделей. Для этого необходимо было объединить все столбцы с городом, районом, округом, улицей и домом в одну строку, приведя их при этом в формат, по которому можно было бы найти этот объект на картах.

Учитывая необходимость работы с координатами, было решено собрать координаты станций Санкт-Петербургского Метрополитена и затем рассчитать расстояние до объектов недвижимости, а затем сам признак - количество станций метро в радиусе километра.

Тем не менее, предварительно было необходимо обработать аномальные значения по всем факторам, разобраться с пропусками и предварительно оценить качество собранных данных.

При исследовании количества пропусков в данных согласно столбцам, был получен следующий результат:

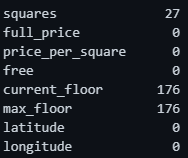


Рисунок № - количество пропущенных значений по столбцам

Ввиду некорректности и ошибочности данных в строках без площади, которые, скорее всего, были вызваны ошибками в оформлении объявления – было решено избавиться от 27 строк с пропущенными значениями площади объекта. Кроме того, эти строки совпадали с пропусками и по этажам объектов, поэтому после исключения 27 строк, лишь 149 показывали пропуски по этажам.

Изучив особенности данных, а именно средние, медианные и модальные значения по данным столбцам, было решено заполнить пропуски по максимальному этажу с помощью метода ближайших соседей.

Был произведен анализ аномальности распределения данных по всем имеющимся параметрам. Рассматривая распределение наблюдений по площади, получаем следующую картину:

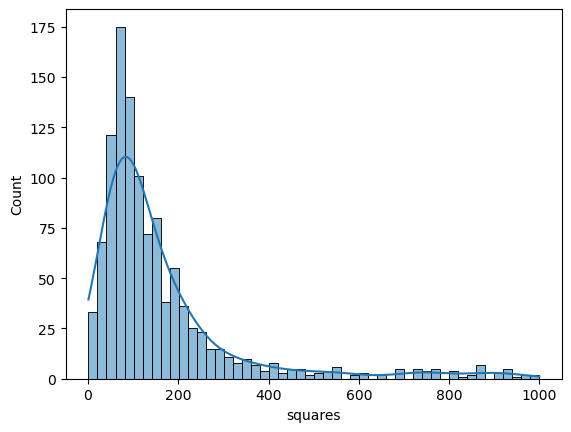


Рисунок № – Распределение наблюдений по площади объекта (м^2).

Здесь заметна сильная правосторонняя асимметрия, кроме того, объекты торговой недвижимости с площадью более чем в 500 квадратных метров крайне редко встречаются на рынке, о чем также говорит и исследование https://irnr.ru/wp-content/uploads/Концепция/Цены\_и\_ставки/Аналитический\_обзор11.pdf от ООО «Информ-Оценка», что предложение на рынке продажи торговой недвижимости, с площадью, включающей или превышающей 500 метров квадратных - менее 10%. Исходя из указанных исследований, а также отсутствия в выборке достаточного количества наблюдений, превышающих 300 метров квадратных, было решено ограничить выборку данными с площадью, не превышающей 300 метров квадратных. После чего, было получено следующее распределение:

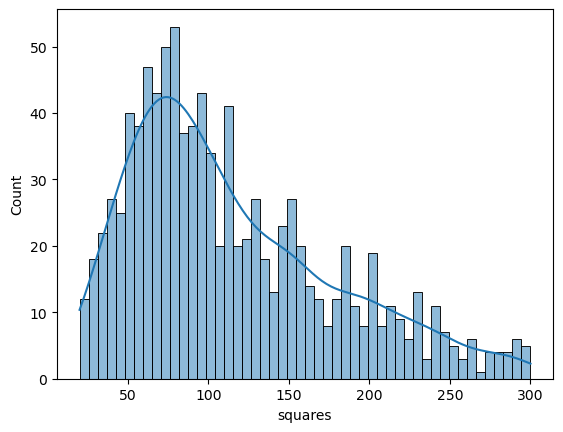


Рисунок № – Скорректированное распределение наблюдений по площади объекта (м^2).

После чего был исследован целевой параметр настоящего исследования – стоимость квадратного метра. Согласно необработанным данным, распределение данного параметра будет иметь следующий вид:

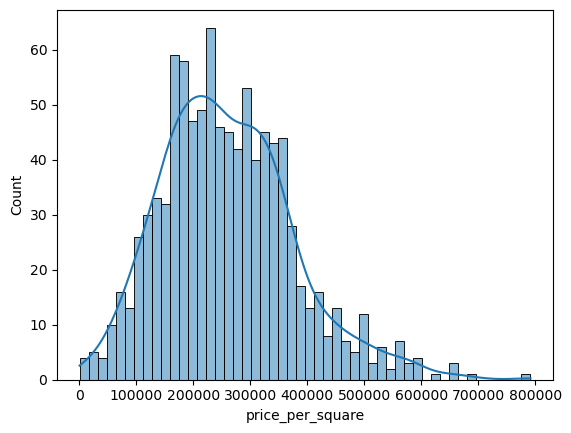


Рисунок № – Распределение наблюдений по стоимости квадратного метра (руб/м^2).

Данная гистограмма также показывает небольшую правостороннюю асимметрию, тем не менее, предельно заметно, после какой стоимости квадратного метра количество объектов значительно уменьшается. Кроме этого, согласно указанному ранее исследованию, в половине сделок цена покупки в рублях за квадратный метр исчислялась измерялась в рамках от 125 тысяч рублей за квадратный метр, до 250 тысяч рублей за квадратный метр. Было принято решение ограничить выборку в следующем диапазоне: от 40 000 рублей за квадратный метр до 450 000 рублей за квадратный метр. Распределение после изменений можно рассмотреть на следующем рисунке:

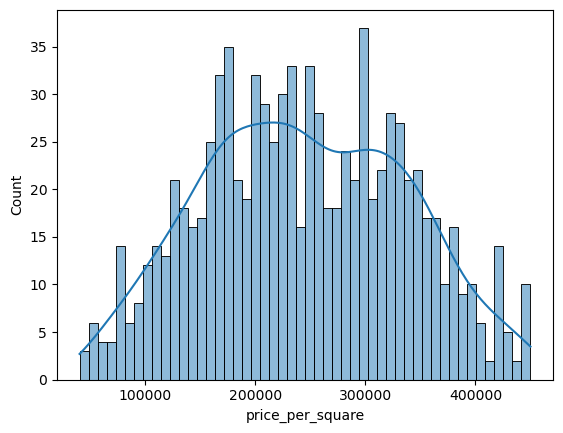


Рисунок № – Скорректированное распределение наблюдений по стоимости квадратного метра (руб/м^2).

Изучив площадь и целевой признак, признаками, которые осталось рассмотреть оставались являлись этажность и координаты самого торгового объекта.

Особенности признака этажа, на котором находится сам объект, показывали, что использование данного признака будет не совсем корректно, практически все объекты расположены на первом этаже, о чем нам говорит и распределение данного параметра на следующем рисунке.

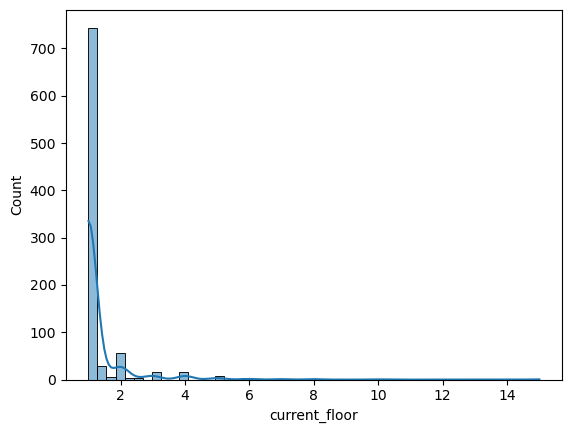


Рисунок № – Распределение наблюдений по этажу.

Такой признак не привнесет никакого логического смысла в исследование, кроме того, он не сможет показывать связь с целевым признаком, учитывая сильную асимметрию в распределении данных.

Аналогичная ситуация возникла и с признаком занятости объекта, в нем лишь несколько десятков объявлений представлялись занятыми, почему было принято решение исключить из исследования и этот параметр.

Тем не менее, мы имеем параметр по максимальному этажу здания, что в принципе подразумевает под собой и учет возраста здания, его качества и других особенностей, учитывая, что структура этажей зданий в Санкт-Петербурге довольно последовательна. Здесь и большое количество зданий в центре города по 4-5 этажей. Большое количество зданий, построенных в советские годы, с 5 и 9 этажами. Более новые же здания не так ограничены в количестве этажей, однако тоже имеют свои закономерности.

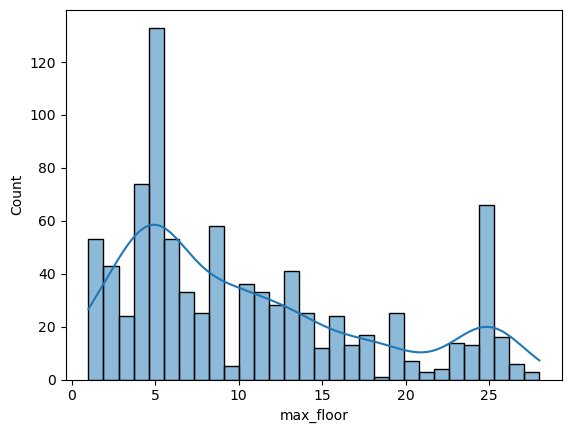


Рисунок № – Распределение наблюдений по максимальному этажу.

Распределение, представленное на рисунке, демонстрирует, что значительно большая часть объявлений находится в пятиэтажных зданиях. Далее идут четырехэтажные, двадцати-пятиэтажные, девяти и одноэтажные здания. В целом, на данном этапе уже можно предположить, что данный параметр будет представлять большой интерес в дальнейшем исследовании.

Переходя к координатным признакам, важно понимать, что использование простых гистограмм для понимания распределения объектов будет не совсем информативно и в некоей степени не корректно. Поэтому стоит обратиться именно к координатной плоскости.

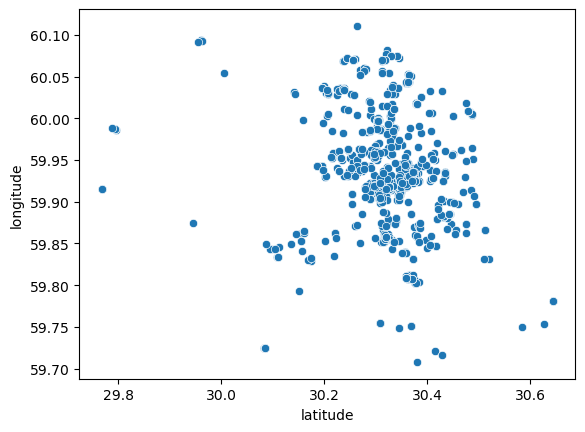


Рисунок № – Распределение наблюдений согласно координатам.

Учитывая большую связь текущей выборки с координатными данными, важно обращаться к данной выборке соответственно, с геоспатическим подходом к анализу. Тем не менее, для исследования, такого количества наблюдений будет недостаточно, как и признаков, почему стала ясна необходимость в расширении выборки и привлечения большего количества признаков.

**аренда**

Был изучен аналогичный по фильтрам раздел торговых объектов на площадке Циан, по аренде торговых объектов, откуда также был произведен сбор данных с помощью парсинга. Стоит указать, что предварительно были откорректированы названия некоторых столбцов-признаков и URL-ссылка в парсере на подборку коммерческой недвижимости.

Рассмотрим признаки, которые собираются с арендной выборки:

* Название объявления;
* Стоимость аренды объекта в месяц (руб.);
* Стоимость аренды за квадрат в год (руб.);
* Площадь (м^2);
* Этаж;
* Свободно ли помещение;
* Город;
* Район;
* Округ;
* Улица;
* Номер дома;
* Ссылка на страницу продавца;
* Описание.

При соединении двух наборов данных такого рода – было необходимо учитывать не только совпадающие признаки, но и признак, информация о котором нам первостепенна важна – стоимость квадратного метра при покупке недвижимости. У новой выборки данных такой информации нет, однако есть признак, который потенциально имеет большое влияние на возможность моделировать стоимость квадратного метра, но отсутствует в исходном наборе данных – стоимости аренды объекта в месяц.

При изучении возможности исправления возникшей ситуации с данными, было вновь уделено особое внимание возможности заполнения пропусков данных с помощью KNN - метода ближайших соседей. Наборы данных по покупке и аренде недвижимости необходимо было привести к одному виду, в особенности было важно иметь одни и те же зависимые переменные, чтобы корректно и максимально точно заполнить пропуски с помощью учета влияния имеющихся зависимых переменных.

Поэтому собранные признаки для основного набора данных были аналогичным образом сформированы в арендном наборе данных, учитывая географические координаты, количество парковок, отношение к тому или иному району, количество станций метро в радиусе 500 м. Важной частью стояла также и предобработка данных, позволяющая исключить выбросы и аномалии, которые были аналогичным первому набору данных образом исключены.

После чего было произведено применение метода заполнения пропусков с помощью метода ближайших соседей, которое позволило бы нам иметь признак, демонстрирующий стоимость аренды в месяц у аналогично объекта, ровно также, как и целевой признак, который показывал наряду с фактической стоимостью квадратного метра - стоимость квадратного метра у объекта аналогичного тому, который имеет параметры данного объекта.

Имеющийся признак стоимости аренды наблюдения в месяц был переведен в признак стоимости аренды квадратного метра, с помощью следующей формулы:

После чего, было изучено распределение добавленного параметра:

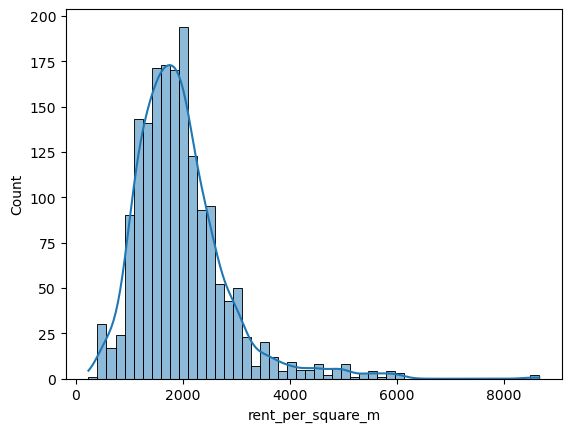


Рисунок № – Распределение наблюдений по стоимости аренды квадратного метра.

Возникает нежеланная ситуация, а именно - недостаток данных со стоимостью аренды квадратного метра, превышающей 4000 рублей за квадратный метр. В данном случае был изучено источник, изучающий данный вопрос (<https://irnr.ru/wp-content/uploads/Концепция/Цены_и_ставки/Аналитический_обзор11.pdf>). С учетом указанных фактов, было решено исключить из выборки наблюдение, в которых стоимость аренды квадратного метра превышает 4000 рублей и недостает до 500 рублей. На выходе, после внесения корректировок в выборку, имеем следующее распределение:

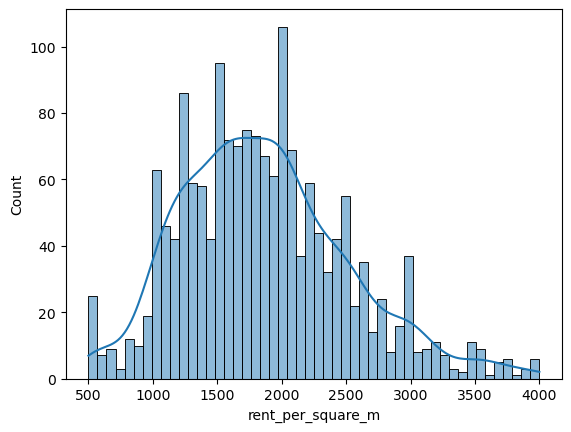


Рисунок № – Скорректированное распределение наблюдений по стоимости аренды квадратного метра.

Одним из важных факторов, которые также стоит учесть при обработке данных – удаленность от центра города, ведь некоторые объекты находятся слишком далеко и будут влиять на общий результат.

Для этого необходимо иметь отчетную точку, которая будет определять центр города. Для корректного использования, стоит изучить распределение данных на карте и учесть, где в Санкт-Петербурге будет самый дорогой квадратный метр. Это можно осуществить с помощью изучения координат объектов, принадлежащих к самым дорогим квадратным метрам. Ограничив выборку от 75% квантиля и выше, получаем, что медианное значение координат следующее: 59.936589 градусов восточной долготы и 30.312419 градусов северной широты, что в переводе на адрес будет Невский проспект, дом 3. Что находится прямо возле Дворцовой площади и Зимнего дворца.

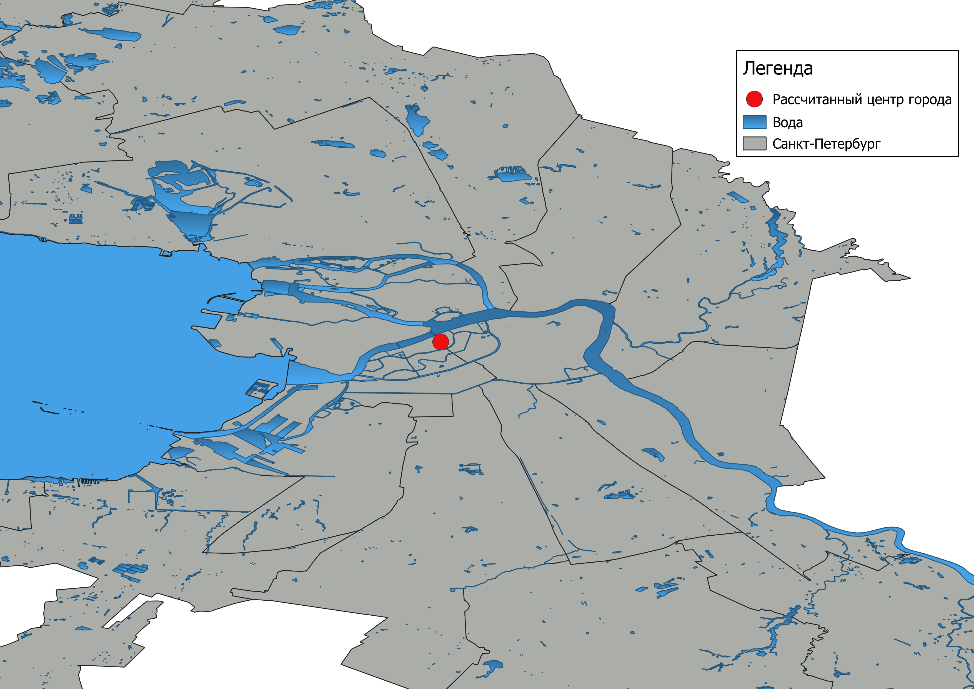


Рисунок № – Рассчитанный центр города.

После чего, необходимо ввести признак, показывающий географическое расстояние от рассчитанного центра города до указанного объекта. Используя циклы и указанную ранее формулу расчета географического расстояние между двумя объектами, получаем новый столбец с данными. Признак такого рода позволит не только корректно оценивать влияние отдаленности от центра на стоимость объекта, но и обработать данные, чрезмерно удаленные от центра Санкт-Петербурга, где они не будут представлять нам интереса, ввиду нахождения немного в другой категории. Рассмотрим распределение данных по рассчитанному параметру на следующем рисунке.

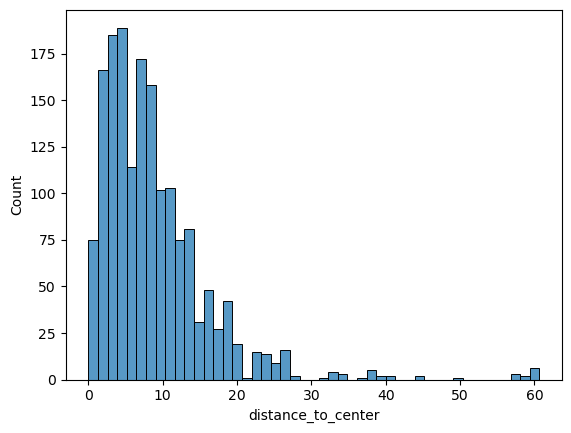


Рисунок № – Распределение по рассчитанному расстоянию от центра города.

Получив такой результат, то есть сильную правостороннюю асимметрию, было решено ограничить выборку наблюдениями, находящимися в радиусе 20 км от точки центра.

Для добавления признаков, связанных с учетом района, в котором находится коммерческий объект, необходимо было воспользоваться собранной базой данных, содержащей информацию о границах районов в формате полигональных данных. Преодобработав эти данные, выборка наблюдений была проверена, на вхождение, по точке координат, в те или иные районы. После чего, была произведена проверка, в каких районах объектов не найдено, или найдено слишком мало, чтобы использовать этот район в исследовании. Рассмотрим на следующем рисунке количество объектов в каждом из районов Санкт-Петербурга:

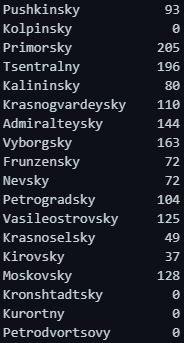


Рисунок № – Количество объектов по районам Санкт-Петербурга.

Общая сумма по районам совпала с количество объектов в выборке. Единственными районами, которые не будут входить в дальнейшее исследование, стоит считать: Колпинский, Кронштадтский, Курортный и Петродворцовый.

Была рассчитана средняя цена квадратного метра в каждом из районов. Помимо этого, учитывая распределение парковок по всему городу, был рассчитан показатель населения в радиусе 1 км от объекта, согласно расчету количества населения на 1 парковку. После произведения расчетов, в среднем, на 1 парковку приходилось примерно 689 человек. Данный показатель был умножен на количество парковок в радиусе 1 километра для каждого наблюдения и введен как признак ввиду того, что использование просто популяции по району не было бы демонстративно и эффективно, районы различаются по размерам, когда как такого родам параметр показывает примерную населенность зоны вокруг объекта.

Сами же фиктивные переменные районов были исключены из выборки, как и расстояние до станций метро, то есть были оставлены только производные признаки из данных подборок. Это связано с большим значением межфакторной корреляции указанных признаков между собой, а также нестабильной связью всех признаков к целевому признаку. Кроме того, в выборке присутствуют признаки в виде координат каждого объекта, что поможет учесть географическое расположение.

После проведения данных этапов предобработки и увеличения выборки – мы имеем 1578 строк и 15 признаков. Рассмотрим признаки и их обозначение в используемой среде для программирования на языке python.

* 'price\_per\_square' – стоимость квадратного метра (руб./м2), целевой признак;
* 'rent\_per\_month' – стоимость аренды в месяц аналогичного объекта (руб.);
* 'squares' – площадь объекта (м2);
* 'max\_floor' – максимальный этаж здания объекта;
* 'latitude' – (северная) широта (°);
* 'longitude' – (восточная) долгота (°);
* 'metro\_250m' – количество станций метров в радиусе 250 метров от объекта (шт.);
* 'metro\_500m' – количество станций метров в радиусе 500 метров от объекта (шт.);
* 'metro\_1km' – количество станций метров в радиусе километра от объекта (шт.);
* 'parkings\_1km' – количество парковок в радиусе километра от объекта (шт.);
* 'rent\_per\_square\_m' - стоимость аренды квадратного метра в месяц аналогичного объекта (руб./м2);
* 'distance\_to\_center' – географическое расстояние до рассчитанного центра город (км);
* 'population' – население согласно району (человек);
* 'mean\_distr\_pps' – средняя стоимость квадратного метра согласно району (руб./м2);
* 'pop\_per\_park' – рассчитанное население в радиусе 1 км от каждого объекта (человек).

Однако важно понимать, что возможно не все из них окажутся достаточно значимыми, чтобы использоваться в построении моделей. Все собранные признаки, при этом, несут весомую роль в процессе анализа рынка коммерческой торговой недвижимости как и в целом, так и в рамках данной научной работы.

**АНАЛИЗ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ**

**ГЕОГРАФИЧЕСКИЙ АНАЛИЗ**

С учетом имеющихся географических данных, с помощью свободной кроссплатформенной геоинформационной системы – QGIS, предназначенной для создания, редактирования, визуализации, анализа и публикации геопространственной информации, важно изучить их на координатной системе с наложенной на нее картой Санкт-Петербурга, размеченной по имеющимся границам районов города, для проведения ознакомительного анализа.

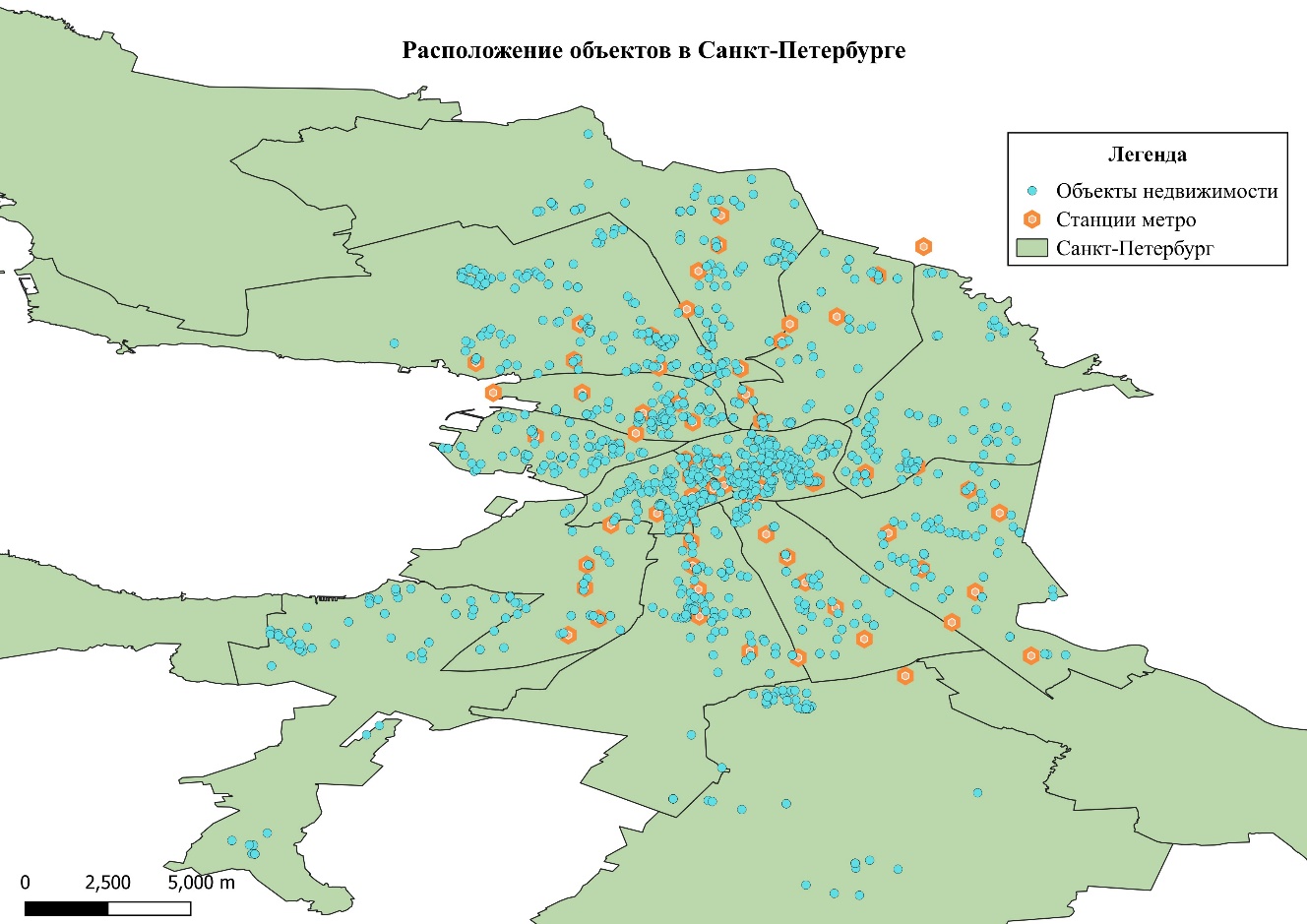


Рисунок № – Расположение объектов в Санкт-Петербурге.

В соответствии с рисунком №, на данной карте расположен Санкт-Петербург, разделенный на административные деления – районы. Синим цветом на данном рисунке обозначены объекты, которые находятся в собранных данных, оранжевым шестиугольником – станции метро. Указанную информацию можно получить из указанной легенды как к этому рисунку, так и последующим. Помимо этого, наличие шкалы масштаба позволяет понимать и интерпретировать данные инфографики. Одной из заметных зависимостей указанного рисунка становится факт скопления большинства объектов у станций метро, что может оказаться существенным фактором в дальнейшей работе.

При более подробном рассмотрении – на рисунке № следует обратить внимание на отображение всех зданий и парковок, существовавших в Санкт-Петербурге на начало 2023 года, когда база географических данных была собрана, данное отображение показывает высокий уровень сопоставляемости полученных с помощью геокодера значений координат, с реальными зданиями.

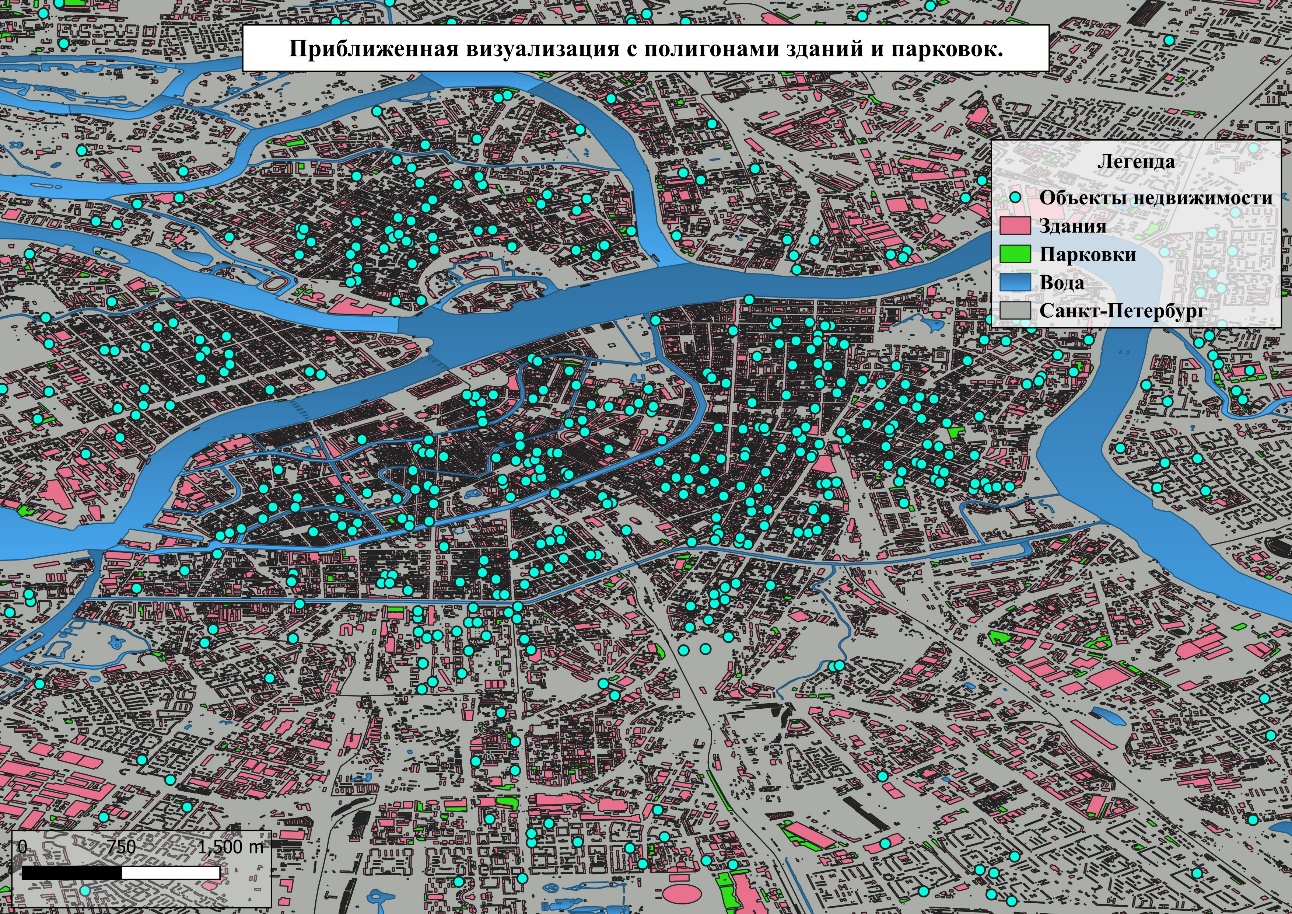


Рисунок № – Приближенная визуализация с полигонами зданий и парковок.

Предварительно можем отметить, что основная доля коммерческих объектов из нашей выборки располагается в Центральном и Адмиралтейском районах, в остальных районах распределение примерно одинаковое между собой и склонно к образованию в кластеры, не считая, конечно же, удаленных от центра районов города.

С учетом предварительного выявления группировки объектов в определенных частях города по геопозиции, было решено оценить уровень кластеризации объектов, рассчитывая размер кластера как радиус в 500 метров, что видно на следующем рисунке.

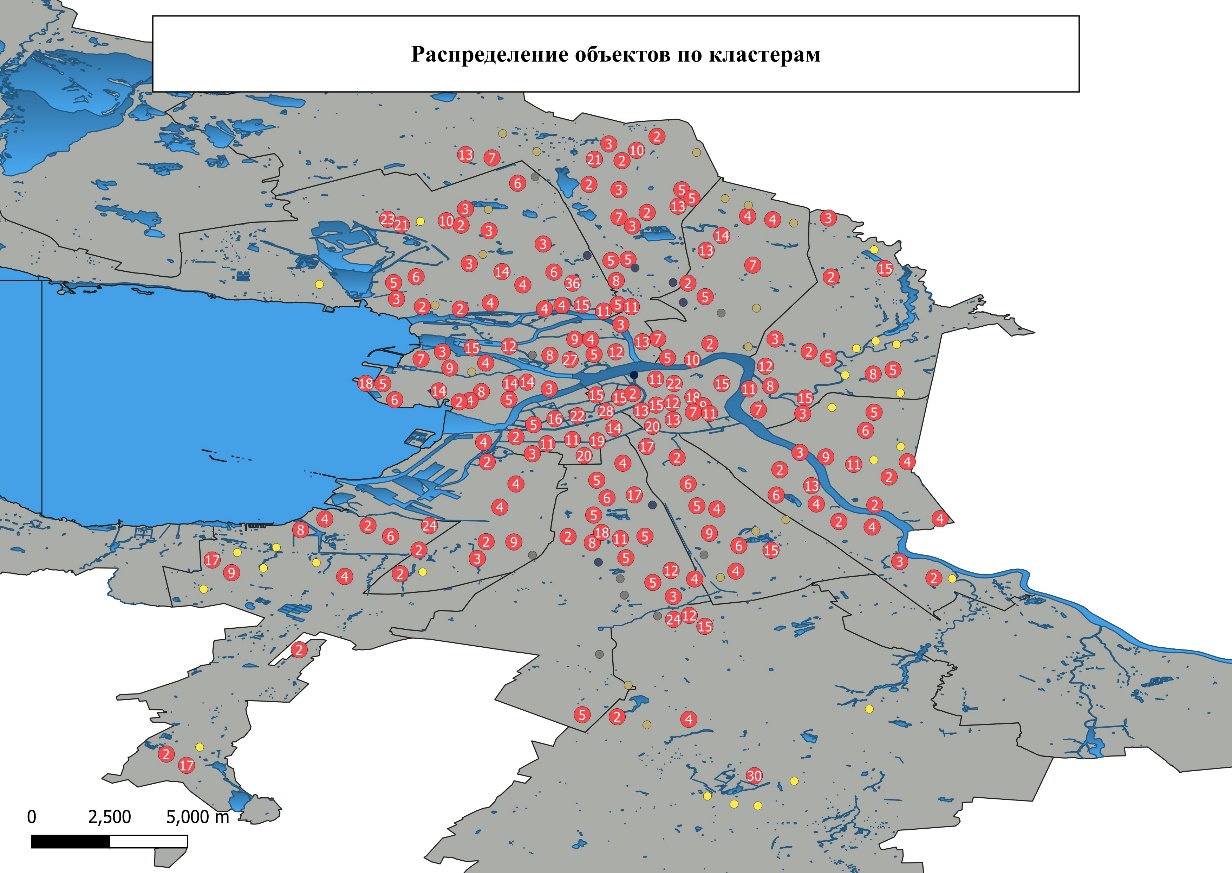


Рисунок № - Распределение объектов по кластерам.

Действительно, большая часть объектов располагается в определенных кластерах, совместно с другими объектами. Можно сделать заключение, что существенная часть находится в бизнес-центрах, новых жилых комплексах, а также в общедоступных частях города Санкт-Петербург, где скапливаются торговые помещения.

Помимо этого, критически важным стоял вопрос изучения распределения и особенностей самих данных вне рамок геоспатического анализа и координатной плоскости. Учитывая проведенный анализ распределения при обработке данных, возникала необходимость повторного изучения, ввиду обновления с помощью присоединения новых данных. Рассмотрим распределение некоторых факторов, с учетом внесенных изменений.

Первым признаком для изучения, следует обозначить целевой – стоимость квадратного метра. Следующий рисунок демонстрирует обновленное распределение.

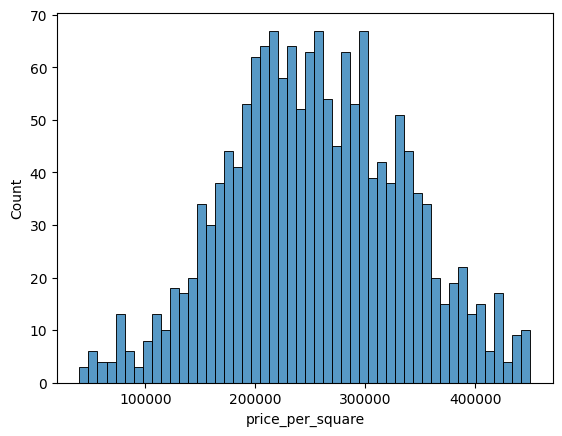


Рисунок № - Распределение обновленного целевого признака.

По сравнению с предыдущим рисунком, демонстрирующим распределение целевого признака, заметно значительное изменение, особенно в заполненности данных, позволяющих утверждать о нормализации распределения целевого параметра. Это говорит об эффективности примененного объединения двух выборок и корректности обеих частей данных.

Изучив распределение стоимости квадратного метра, была необходимость понимания аналогичного распределения с учетом географической интерпретации. Воспользовавшись возможностью распределения объектов по стоимости квадратного метра на группы, был образован график градационного распределения объектов на рисунке №, представленный далее.

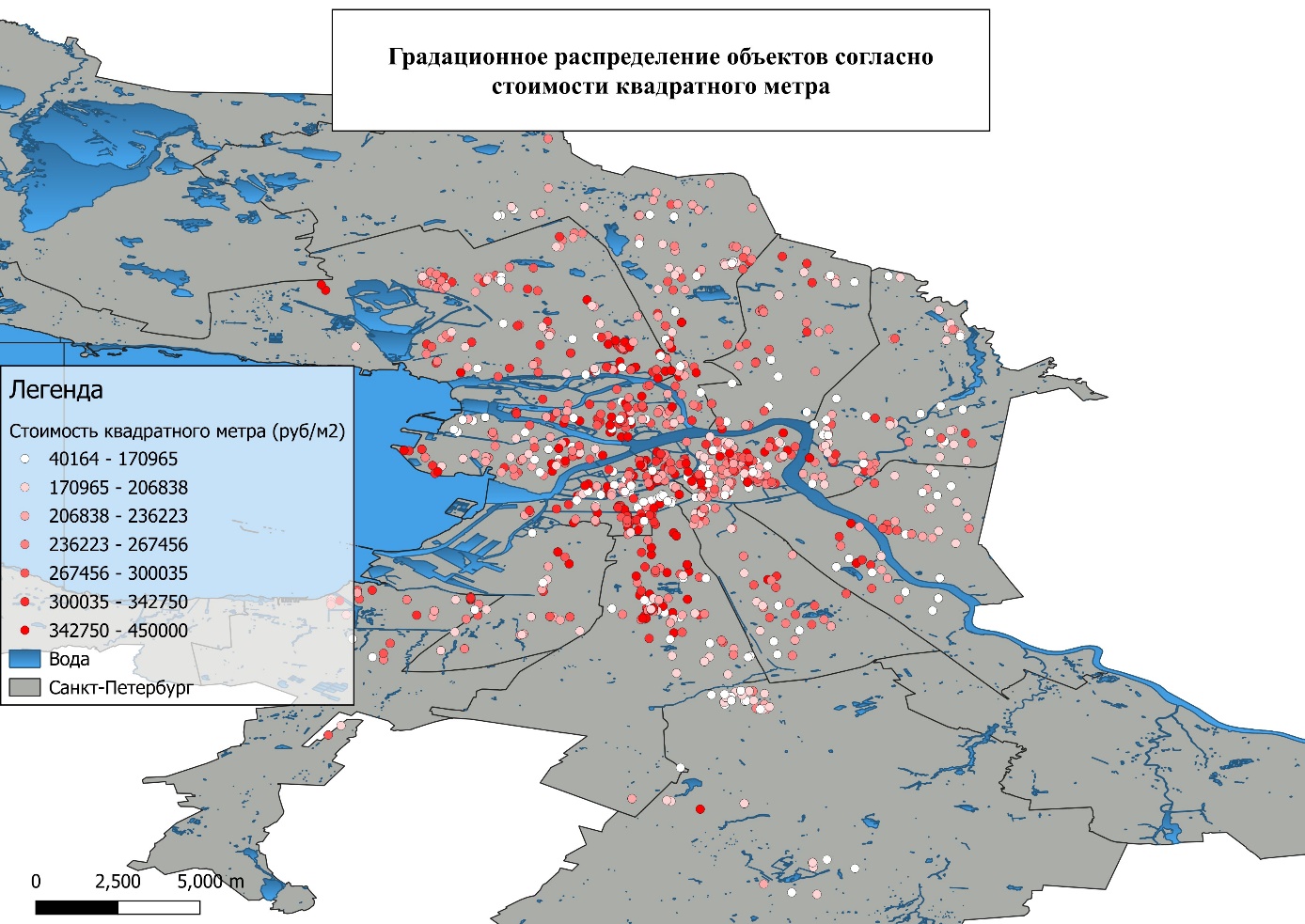


Рисунок - Градационное распределение объектов согласно стоимости квадратного метра.

В первую очередь стоит отметить, что распределение объектов с разной стоимостью квадратного метра достаточно равномерное – в каждом из районов заметны объекты по всей шкале. Однако заметны определенные тенденции, в особенности, что несмотря на достаточно высокую стоимость около центра, по мере удаления от него – трудно заметить уменьшение стоимости квадратного метра. Необходимо рассмотреть данную зависимость более точно, помимо этого – учесть распределение стоимости квадратного метра по районам.

Для выяснения характера зависимости между расстоянием до центра и стоимостью квадратного метра, выведем точечный график на следующем рисунке.

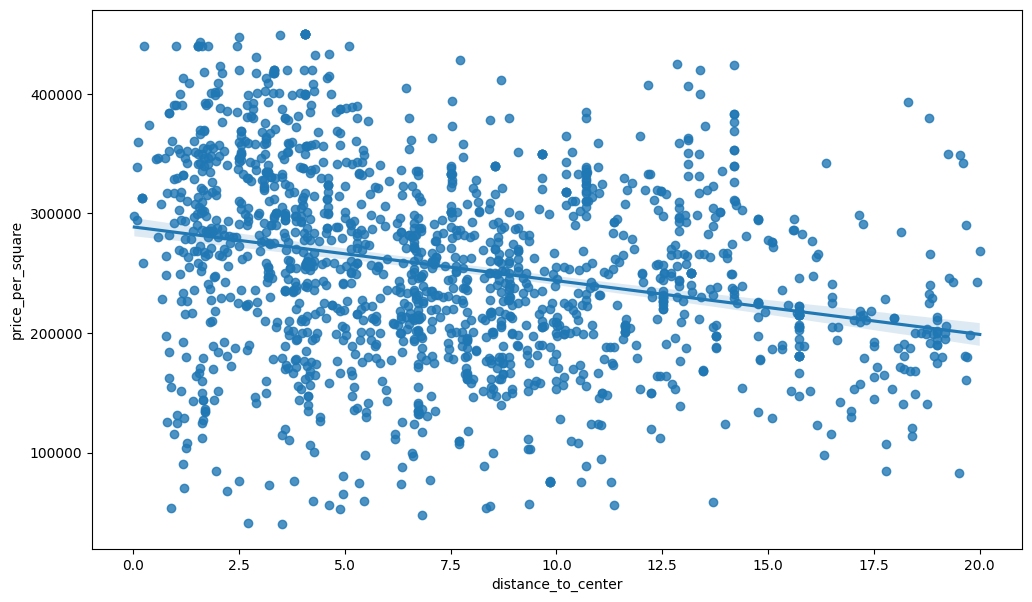


Рисунок – Точечный график зависимости стоимости квадратного метра по мере удаленности от рассчитанного центра.

Исходя из графика, можем утверждать, что по мере отдаления от центра – встречается не только значительно меньшее количество объектов, но и снижается стоимость квадратного метра. Тем не менее, мы все еще имеем наблюдения, которые не показывают снижение в стоимости, даже при значительном расстоянии от центра, что, впрочем, никак не опровергает сделанное выше заключение. Логично предположить, что в неких районах, отдаленных от центра, стоимость квадратного метра не падает, из-за уникальных особенностей каждой части города.

Используя вычисленные данные по средним ценам в каждом из районов, безусловной частью анализа становился процесс исследования влияния района города на стоимость квадратного метра торгового объекта. Результат чего, можно видеть на следующем рисунке.

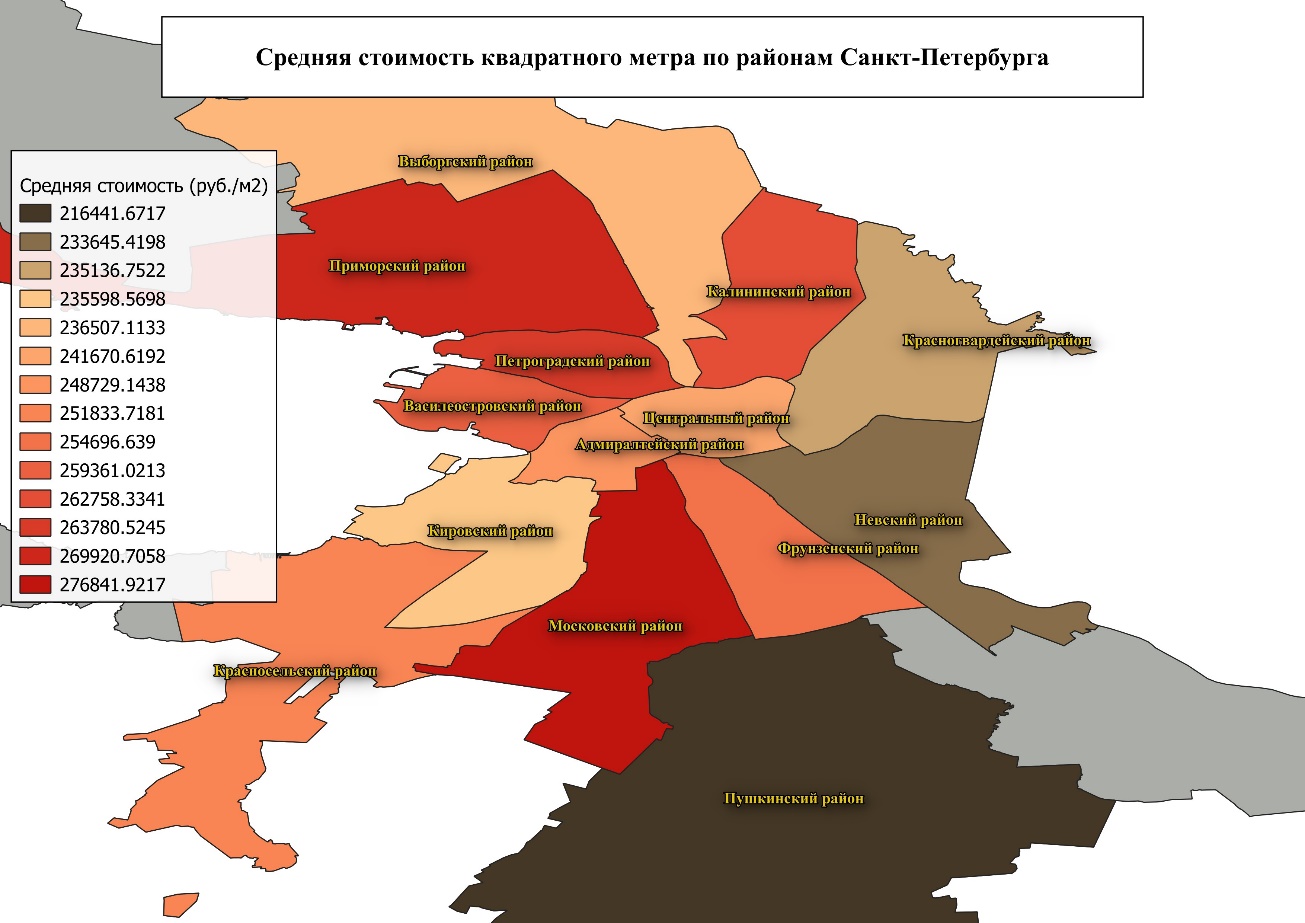


Рисунок - Средняя стоимость квадратного метра по районам Санкт-Петербурга.

Согласно рисунку №, видим, что, в среднем, самый дорогой квадратный метр будет в Московском и Приморском районах, это 276 841 и 254 696 рублей за квадратный метр соответственно. Это обуславливается большой концентрацией новых зданий и сумме нежилой площади в каждом из районов. Тем не менее, выделяющимся результатом можно считать относительно невысокую стоимость квадратного метра в Адмиралтейском и Центральном районах. В данном случае важно учитывать средний возраст зданий районов. Последние же представляют концентрацию исторических архитектурных памятников, которые отличаются более низким качеством помещений и небольшими площадями, в сравнении с новыми помещениями в районах, отдаленных от центра. С отрывом самые низкие показатели средней стоимости квадратного метра в Пушкинском районе – 216 441 рублей за квадратный метр.

Кроме того, нельзя не учитывать факт отсутствия в выборке данного исследования объектов недвижимости со стоимостью квадратного метра, превышающую 450 000 рублей и площадью, превышающей 300 метров квадратных.

Площадь, как один из важнейших факторов при оценке недвижимости, имеет достаточно спорную зависимость со стоимостью квадратного метра. Один из принципов зависимости между площадью и стоимостью квадратного метра - стоимость квадратного метра снижается пропорционально росту площади объекта. Важно проверить и подтвердить данный принцип на используемых данных. Выведем точечный график линейной зависимости между указанными параметрами.

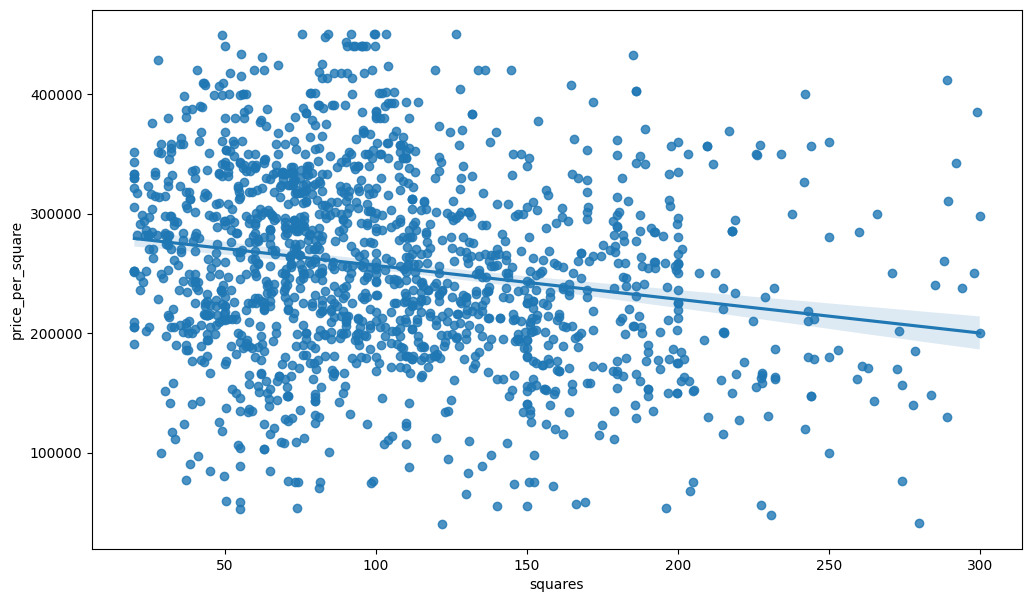


Рисунок – Точечный график зависимости стоимости квадратного метра по мере увеличения площади объекта.

Видим, что предположение корректно, теперь, учитывая данную особенность интерпретации – необходимо понимать распределение значений площади объектов по координатной плоскости имеющегося геопространства. На следующем рисунке можно заметить, как изменяются значения площади объектов в зависимости от местонахождения:



Рисунок – Градационное распределение по площади торгового объекта в Санкт-Петербурге.

На рисунке № заметно, что объекты, с площадью 130 метров квадратных и выше, по большей части отдалены от центральной части города, но тем не менее присутствуют в немалом количестве в Адмиралтейском и Центральном районе. Однако определенно большая часть торговых помещений в данных районах имеет площадь ниже 130 метров квадратных. Кроме того, можно сделать утверждение, что достаточно значительная часть помещений города в этом диапазоне площади сосредоточена в центральной части.

Ещё одним весомым фактором при оценке стоимости коммерческой недвижимости принято считать доступ к парковкам возле объекта. Это во многом применимо к офисной и промышленной недвижимости, однако и для торгового сектора имеет значимое влияние на целевой признак. С учетом рассчитанного признака количества парковок в радиусе 1 км от объекта, имеем следующую инфографику:

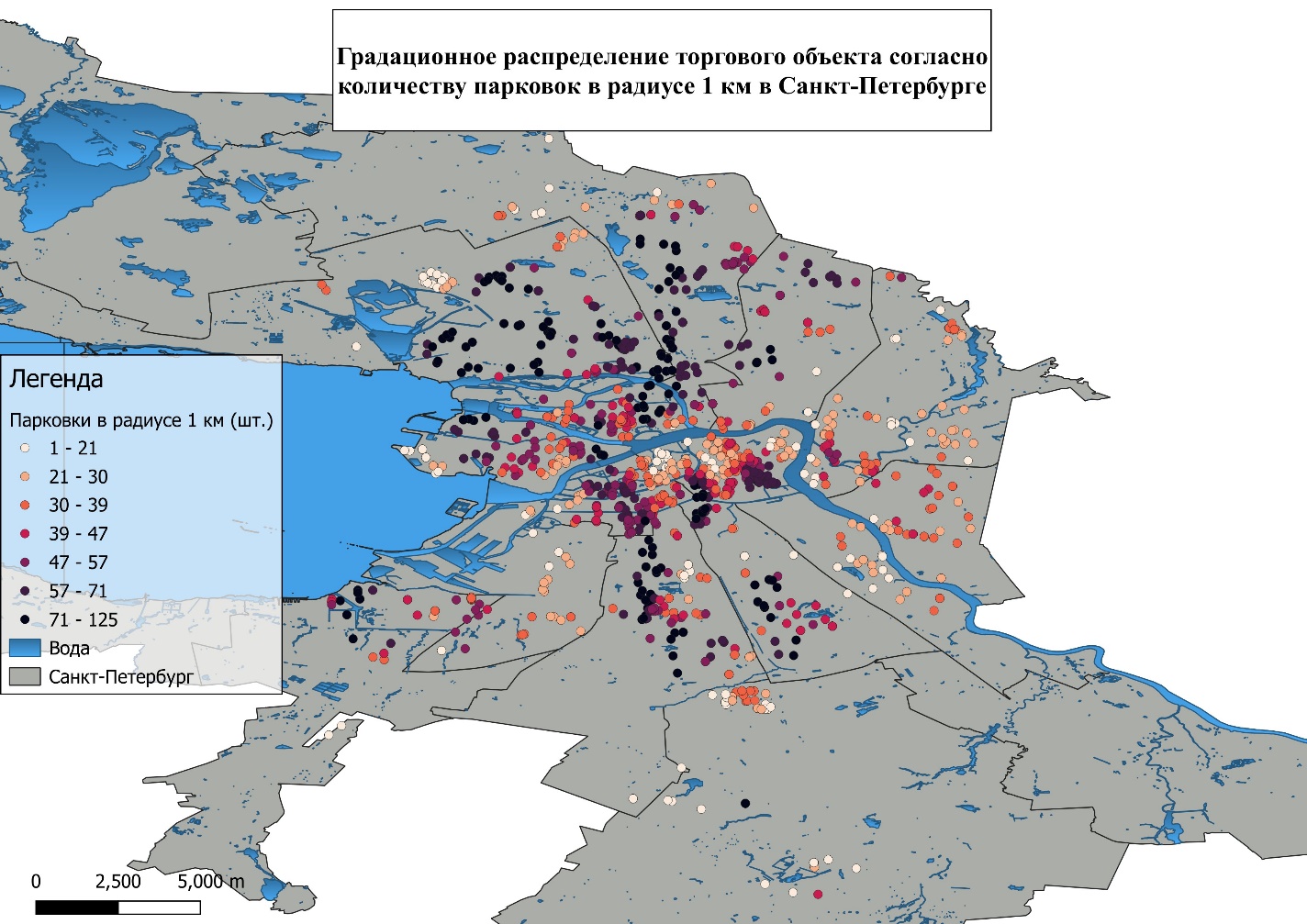


Рисунок – Градационное распределение торгового объекта согласно количеству парковок в радиусе 1 км в Санкт-Петербурге.

Можно заметить, что в количество парковок в радиусе км от объекта стремится к уменьшению, при приближении к центральной части города. Также небольшое количество парковок возле торговых объектов показывают Невский, Кировский и Красногвардейский районы. При этом, согласно количеству парковок, можно считать наиболее благоприятными для владельца автомобиля – Московский, Приморский и Выборгский. Во многом это связано с доступностью использования пространства под парковки в новых районах, не во вред городского планирования. Однако, как сильно связаны с собой целевой признак и признак, учитывающий количество парковок в радиусе 1 километра от объекта. Следующая точечная диаграмма демонстрирует эту связь.

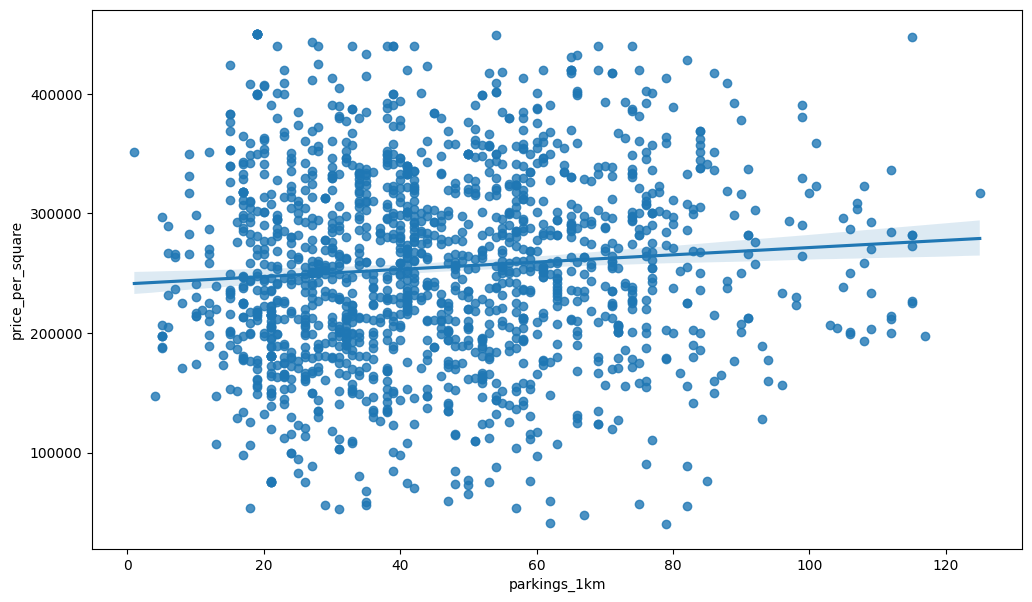


Рисунок – Точечный график зависимости стоимости квадратного метра по мере увеличения количества парковок в радиусе 1 км от объекта.

Подтверждаем наличие связи, однако отмечаем малое количество объектов, окруженных больших количеством парковок и небольшое различие в цене, вплоть до 80 парковок в радиусе километра.

Стоит обратить внимание и на наличие сгруппированных выбросов – в выборке присутствуют объекты, находящиеся в новостройках или бизнес-центрах, где они образуют большую плотность объектов в одной локации, однако создают прецедент, что возле них небольшое количество или полное отсутствие парковок. Данная проблема может возникнуть из-за отсутствия парковок в настоящем моменте, но ультимативно все-таки показывает реальное положение и не приносит проблему неполноты данных.

Важно изучить также и разделение на этажности зданий, в которых находятся объекты. Город, ввиду своей структуры, разделен на кластеры зданий одного типа, важно же учитывать как это выглядит в общем географическом формате распределения по городу.



Рисунок – Градационное распределение максимального этажа здания объекта.

Максимальный этаж здания, в котором находится объект, представляет собой не просто этажность, но по своей сути учитывает возраст объекта, инфраструктуру, которая будет расположена вокруг, но и, как мы видим, географическое расположение объекта. Зависимость цены от местоположения объекта не просто логична, последовательна, но и визуально-подтверждена, учитывая закономерности, обнаруженные с помощью проведенного геоспатического анализа.

Изучая такой показатель как стоимость квадратного метра, критически важной задачей становится понимание, как эта стоимость, в совокупности с площадью влияет на итоговый результат, полную стоимость продажи объекта. Данный параметр не несет пользы для оценки самого объекта недвижимости, но представляет большую роль в рамках анализа такого крупного геопространства, как Санкт-Петербург. Поэтому имело смысл обозначить распределение полной стоимости объекта по городу, после чего сравнить визуализацию с представленными ранее факторами и оценить структуру связей параметров в рамках города со своими уникальными особенностями.

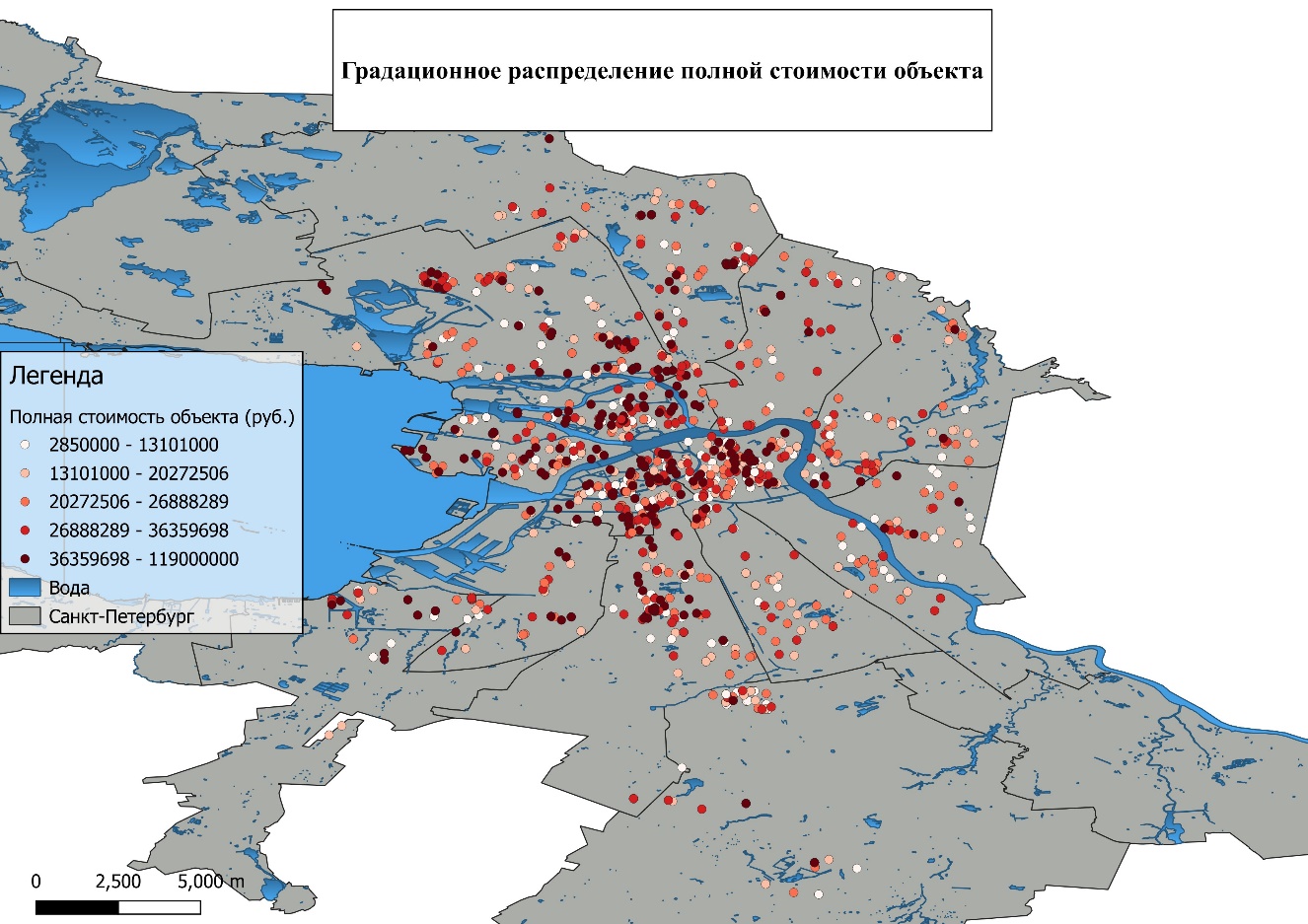


Рисунок – Градационное распределение полной стоимости объекта.

С учетом градационного распределения полной стоимости объекта, указанного на рисунке, можем заметить, что самые дорогостоящие объекты расположены в центральной части Санкт-Петербурга, а также немного удаленные на север и юг от рассчитанного центра города. Восточная часть города же показывает значительное уменьшение в стоимости, несмотря на подобное остальной части города распределения максимального этажа здания, и площади объекта. Признаком, который отличает данные части города можно считать количество парковок, предположительно интерпретируя его как качество инфраструктуры в этой части города. Возможный недостаток качественных условий для торгового бизнеса негативно влияет на стоимость как и квадратного метра объекта, так и его полную стоимость.

Проведя полный и корректный общий анализ собранных данных, с учетом географических координат, важно оценить влияние зависимых признаков на целевой признак. Корреляционный анализ поможет понять, какие признаки стоит включать в дальнейшее исследование и построение моделей для оценки стоимости квадратного метра коммерческого торгового объекта недвижимости, а какие будет необходимо исключить.

**Корреляционный анализ**

Корреляционный анализ – это один из статистических методов изучения тесноты связи между теми или иными признаками. Он является фундаментом для построения точных и корректных моделей.

В первую очередь, для определения метода оценки корреляции между признаками, поймем какое распределение у имеющихся данных.

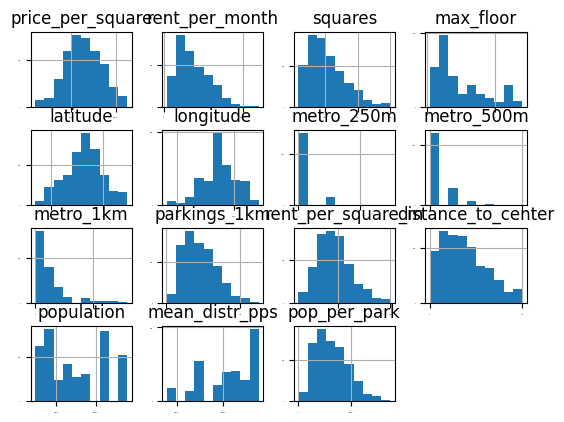


Рисунок – Распределение всех признаков в выборке.

Изучая рисунок №, становится заметно, что целевой признак имеет нормальное распределение, как и координаты самого объекта – долгота и широта. Однако в остальных параметрах видим наличие правосторонней асимметрии, это касается стоимости аренды аналогичного объекта в месяц, площади объекта, максимального этажа здания, дистанции до центра. Некоторые параметры имеют чрезвычайно большое количество нулевых значений – это параметры, связанные с количеством метро в разном радиусе от объекта недвижимости. Распределение населения по районам достаточно сложно интерпретировать, основываясь на нормализации данных. Объектов слишком много, чтобы такие фиксированные значения населения для данных из района показывали нормальное распределение. Однако, с учетом наличия признака примерного рассчитанного населения в радиусе 1 км от объекта, мы имеем более закономерный результат, хоть и с небольшой правосторонней асимметрией. Практически аналогичная ситуация возникает и со средним значением стоимости квадратного метра по районам. Еще одним фактором, приблизившемся к нормализации распределения стоит считать стоимость аренды квадратного метра аналогичного объекта в месяц.

С учетом наличия факторов, которые никоим образом не подразумевают под собой нормальное распределение, важно обратиться к коэффициенту корреляции Кендалла, используя возможности в библиотеке pandas на языке программирования python. Выведем горячую карту корреляции между признаками нашей выборки на следующем рисунке.

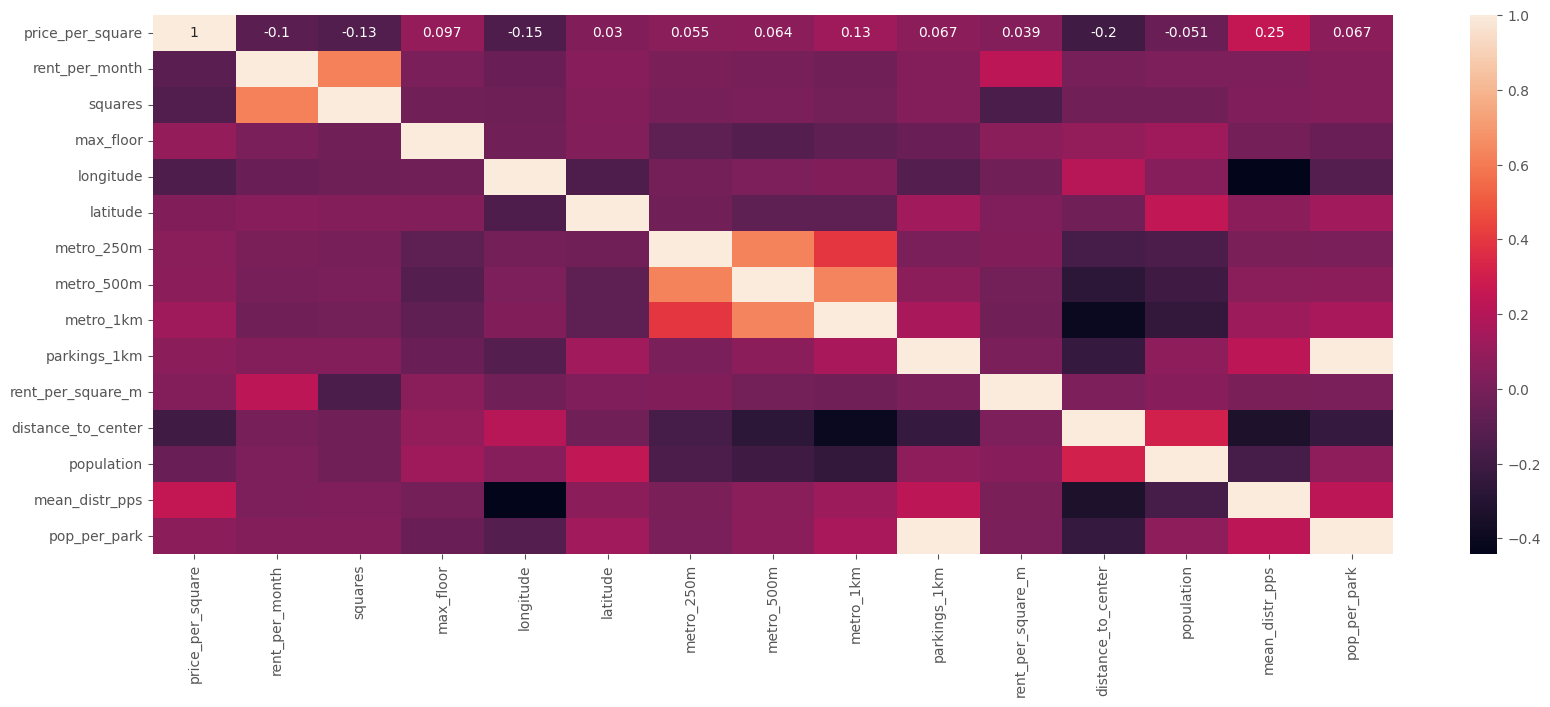


Рисунок – Горячая карта корреляции признаков.

Исходя из рисунка №, можем отметить отсутствие заметной связи между какими-либо признаками и целевым признаком стоимости квадратного метра. Тем не менее, это не означает, что в совокупности, данные параметры не могут смоделировать точный результат квадратного метра. Рынок недвижимости достаточно сложен и не структурирован. Поэтому даже при наличии параметров, которые являются главнообразующими в рамках стоимости недвижимости, возможно отсутствие строго-определенной связи между признаками. Главной задачей, при этом, является отбор параметров, которые могут негативно повлиять на точность или последовательность моделирования данных.

Одной из серьезного рода проблем можно считать наличие сильной межфакторной корреляции между признаками, демонстрирующими количество станций метров в определенных радиусах от объекта. При этом, самым значимым параметром среди указанных трех является количество станций метро в радиусе 1 км от объекта. Сравнив значения связи между признаками, и оценив значимость данных признаков - исключим из выборки 2 признака по метро, с радиусом в 250 и 500 метров.

Помимо этого, сильную межфакторную корреляцию показывают признаки площади объекта и стоимости аренды аналогичного объекта за месяц, согласно шкале, указанной на горячей карте на рисунке №, значение межфакторной корреляции составляет примерно 0.5, заметная теснота связи, превышающая при этом тесноту связи со стоимостью квадратного метра. Учитывая, что теснота связи площади с целевым признаком выше – был исключен признак стоимости аренды аналогичного объекта.

Заполнить логическое присутствие признака, связанного с уровнем потенциального дохода с объекта может признак стоимости аренды квадратного метра. Если однако обратить на величину корреляционного отношения, увидим, что связь не является даже слабой. Помимо этого, фактор коррелирует с площадью объекта. К тому же, отсутствие связи между целевым и указанным признаком не представляет другого решения, кроме как исключения из выборки указанного фактора. Обратим внимание на другие признаки, корреляционная связь целевого признака с которыми также низка.

Одним из таких признаков является количество парковок в радиусе 1 километра. Согласно корреляционному отношению, полученному с последнего рисунка и вывода корреляционной статистики, равному 0.067 – связь практически отсутствует. Однако, как было изучено нами в предыдущей главе, связь между параметрами действительно есть, в том числе и учитывая распределение по карте. Исключение данного признака не несет в себе последовательности исследования, хоть безусловно и имеет в своем логическом основании причины. Такая же ситуация образуется и с признаком максимального этажа, несмотря на корреляционное отношение, равное 0.097 – признак значим и действительно указывает не только на этажность здания, но и подразумевает его возраст и предрасположенность к тому или иному кластеру в определенной части города, что было обсуждено также в предыдущей главе.

Подобная же ситуация возникает и с параметром населения в радиусе 1 километра от объекта, однако ввиду того, что данный признак был рассчитан с помощью признака количества парковок – мы имеем очень сильную прямую тесноту связи между двумя зависимыми признаками, почему, при равенстве корреляционного отношения данных параметров к целевому признаку – мы исключим производный признак, оставив количество парковок в радиусе 1 километра от объекта.

Указывая на недостаток связи населения в определенном радиусе от объекта, нельзя не учитывать и недостаток связи между признаком с данными по населению в каждом из районов и целевым признаком. Признак создает большое количество связей, теснота которых превышает тесноту данного признака с целевым. Признак населения, оформленный на основе фиктивных переменных районов исключен и не несет научного интереса, по крайней мере, в имеющемся, в данном случае, формате.

Особенным случаем стоит назвать связь целевого признака с координатами. Исходя из полученного статистического вывода, долгота объекта имеет хоть и слабую, но значимую связь, в то время как широта объекта показывает, что данный признак не значим для ценообразования. С точки зрения логики и сохранения принципа использования данных в том наборе данных, без которого они не будут представлять из себя весомую роль, то есть нарушая полноту данных, исключение одного из факторов не позволительно. В таком случае существует лишь два возможных решения событий, а это либо исключение обоих признаков, либо использование обоих. Учитывая критическую роль координат в данном исследовании и в ценообразовании квадрата недвижимости, было решено оставить в выборке оба этих показателя. Тем не менее, важно понимать причины такого рода проблемы. С учетом проведенного географического анализа, в особенности рисунка №, можно указать на следующие причины:

* Объекты всех градаций цены присутствуют в любой части города, при движении по оси ординат и отсутствуют при движении по оси абсцисс.
* Сконцентрированность объектов в западной части города.
* Низкое качество объектов в восточной части города.
* Рассредоточенность объектов по вертикали.

Из этих причин, можно сделать вывод, что в рамках общего подхода к образованию стоимости из координат, оба признака влияют на стоимость квадратного метра. Однако в рамках используемой выборки, то есть полного рынка Санкт-Петербурга, широта действительно не значима как признак. Тем не менее, целесообразно оставить оба признака, но при этом учитывать имеющийся дисбаланс в распределении объектов в Санкт-Петербурге.

Исходя из проведенного корреляционного анализа и процесса определения только необходимых данных – мы имеем следующие признаки, роль в ценообразовании стоимости квадратного метра коммерческого торгового объекта в Санкт-Петербурге которых действительно важна:

* Площадь объекта (м2);
* Максимальный этаж здания объекта;
* Количество станций метро в радиусе 1 км (шт.);
* Координаты: долгота и широта (°);
* Количество парковок в радиусе 1 км (шт.);
* Географическая удаленность от центра (км);
* Средняя стоимость квадратного метра в районе, в котором находится объект (руб./м2).

Исходя из полученных результатов и будет строиться дальнейшее построение моделей для оценки целевого признака.

**Машинное обучение**

Отталкиваясь от изученных источников, мы знаем, что наиболее оптимальным методом машинного обучения, скорее всего, будет являться метод случайного леса.

Тем не менее, для всеобъемлющего и полного исследования – важно изучить все основные методы и выбрать наиболее точный для использования в моделировании оценки торговой недвижимости, а именно – стоимости квадратного метра в Санкт-Петербурге. Ключевым показателем результатов моделирования будет являться среднее абсолютное процентное отклонение, также известное, как MAPE. Но важно учитывать и такие расчетные статистические показатели, как коэффициент корреляции, среднее абсолютное отклонение и среднее квадратичное отклонение. Использование всех данных показателей в совокупности позволит максимально точно и объективно оценивать возникающие результаты в процессе моделирования нашего целевого показателя.

Помимо этого, есть несколько принципов, которые мы обязаны учитывать при построении моделей. Одним из них является использования кросс-валидации [почитать про cross\_val\_score] для возможности оценки моделей на незнакомых для нее данных, что позволит нам избежать переобучения модели.

Напишем код для выбора наилучшего отношения тестовой выборки к тренировочной.

Помимо этого, важно рассмотреть использование не только стандартных имеющихся данных, но и данных стандартизированных. Соответственно, полное исследование даст нам определенный спектр для оценки результатов и понимания, как лучше понимать эти данные.

Выбрав используемый нами набор данных, с указанными в предыдущей главе оставшимися параметрами, мы разделяем выборку на 2 части – целевой признак и зависимые признаки. После чего мы используем кросс-валидацию, определим размер тестовой и тренировочной выборки. На данный момент важно определить основное отношение этих двух выборок.

На данный момент отношение тренировочной к тестовой выборке будет следующим – 8:2. Это позволит моделям обучиться на большом количестве наблюдений и при этом даст значительное количество наблюдений для проверки результатов. Помимо этого, перед распределением, которое как и построение моделей, так и вся работа с данными производится на языке программирования python, важно задать семя случайного распределения. Чтобы даже случайность имела определенную последовательность, которую мы могли бы произвести повторно, для оценки результатов достаточно точно и корректно.

Про kfold.

Как уже было указано, основным показателем для оценки моделей будет являться среднее абсолютное процентное отклонение, далее обозначаемое как MAPE, ведь основной целью данной работы является возможность моделирования корректной стоимости объекта коммерческой недвижимости. Построим несколько основных моделей, а именно:

* Линейная регрессия;
* Лассо;
* Эластичная сеть;
* Метод К-ближайших соседей;
* Решающие деревья;
* Метод опорных векторов.

Рассмотрим результаты каждого из метода, опираясь на указанный ранее показатель, то есть выведем среднее абсолютное процентное отклонение для каждого из метода а также стандартное отклонение от данного среднего. Линейная регрессия показала точность в 0.294483, то есть модель в среднем ошибается на 29,4483% (+- 0.2066%). Лассо же показала практически аналогичный результат, данная модель ошибается в среднем в 29,4480% (+- 0.2061%). Эластичная сеть имела следующие результаты: 29,4512% (+- 0.1439%). Метод K-ближайших соседей показал среднюю ошибку на 30.7437% с разбросом в (+- 0.0029%). Решающие деревья показал ошибку в 31.4379% (+- 0.3307%). А метод опорных векторов в 33.5565% (+- 0.6456%). Подробную визуализацию можно рассмотреть на следующем рисунке, демонстрирующим сравнение указанных алгоритмов.

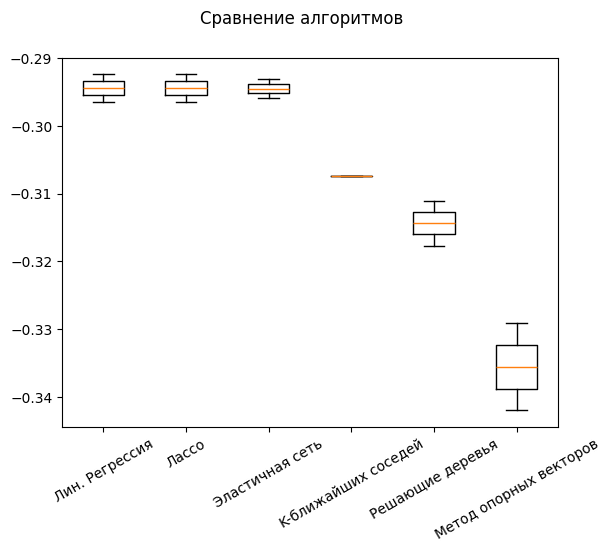


Рисунок – Сравнение алгоритмов согласно MAPE.

Для определения наиболее лучшего подхода, необходимо рассмотреть также использование данных алгоритмов при стандартизации данных. Воспользовавшись функцией StandardScaler, существующая выборка была стандартизирована, после чего построены аналогичные модели и выведены результаты в виде среднего значения MAPE и его стандартного отклонения. Рассмотрим и оценим данные результаты в формате названия метода, его показателя среднего абсолютного процентного отклонения и стандартного отклонения:

* Стандартизированная Лин. регрессия: 29.4483% (0.002066)
* Стандартизированная регуляризация Лассо: 29.4483% (0.002066)
* Стандартизированная Эластичная сеть: 29.9320% (0.005157)
* Стандартизированный метод К-ближайших соседей: 28.7242% (0.000674)
* Стандартизированные решающие деревья: 32.0327% (0.011057)
* Стандартизированный метод опорных векторов: 33.5535% (0.006461)

Обратимся также и к реализованной визуализации на следующем рисунке для более наглядного восприятия и оценки:

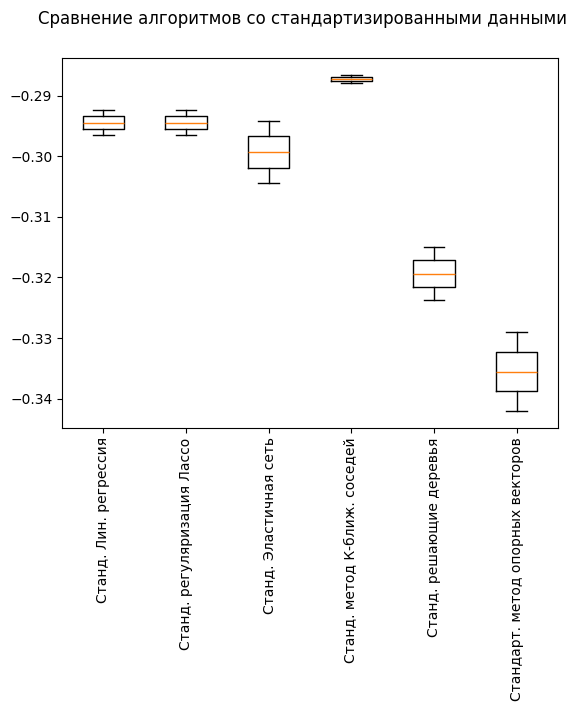


Рисунок – Сравнение стандартизированных алгоритмов согласно MAPE.

На данный момент видим, что особого различия между моделями со стандартными и стандартизированными данными практически нет. Модель в среднем ошибается на 29% на тестовых данных, что достаточно негативный результат проделанной работы. Заметно радикальное улучшение метода ближайших соседей, что более чем закономерно, учитывая стандартизацию данных, остальные же алгоритмы не показали существенного изменения при стандартизации данных. Однако, крайне рано делать существенные выводы. Гораздо важнее обратиться к ансамблевым методам, способным заметно улучшить полученные результаты.

В первую очередь рассмотрим алгоритмы на стандартизированных данных, в их числе будет:

* Ада бустинг;
* Градиентный бустинг;
* Случайный лес;
* Ансамбль дополнительных деревьев.

Рассмотрим сравнение на следующем рисунке:

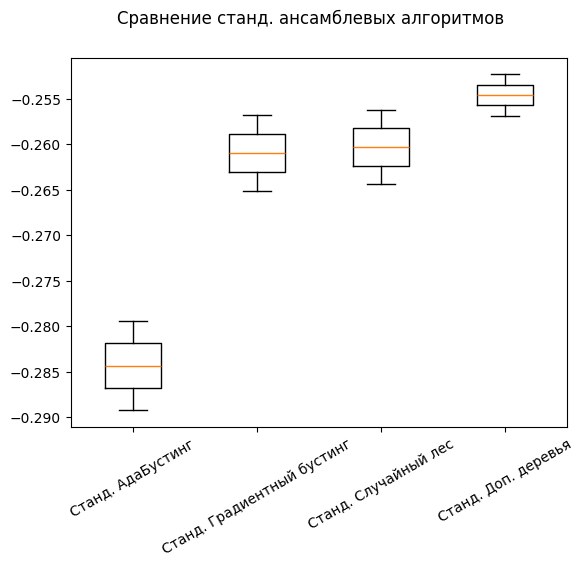


Рисунок – Сравнение стандартизированных ансамблевых алгоритмов согласно MAPE.

Основываясь на рисунке №, можем заметить значительные изменения по сравнению с предыдущими результатами. Наиболее предпочитаемыми моделями из ныне рассмотренных, стоит считать ансамбль дополнительных деревьев со средним абсолютным процентным отклонение равным в среднем 25.4619% со стандартным отклонением в 0.2279%, случайный лес со средним абсолютным процентным отклонение равным в среднем 26.0320% со стандартным отклонением в 0.4099% и градиентный бустинг со средним абсолютным процентным отклонение равным в среднем 26.0989% со стандартным отклонением в 0.4148%. Адабустинг на стандартизированных данных же не показал значительного улучшения результата по сравнению с обычными алгоритмами, рассмотренными ранее. Рассмотрим результаты построения моделей на данных, не подвергнутых стандартизации.

Согласно полученной статистике, имеем следующие показатели среднего значения и стандартного отклонения среднего значения у среднего абсолютного процентного отклонения:

* Адабустинг: MAPE = 28.5546% (0.008087);
* Градиентный бустинг: MAPE = 26.0886% (0.003800);
* Случайный лес: MAPE = 24.8374% (0.003626);
* Дополнительные деревья: MAPE = 24.4669% (0.003769).

Рассмотрим данные статистические показатели более подробно с помощью полученной визуализации на следующем рисунке:

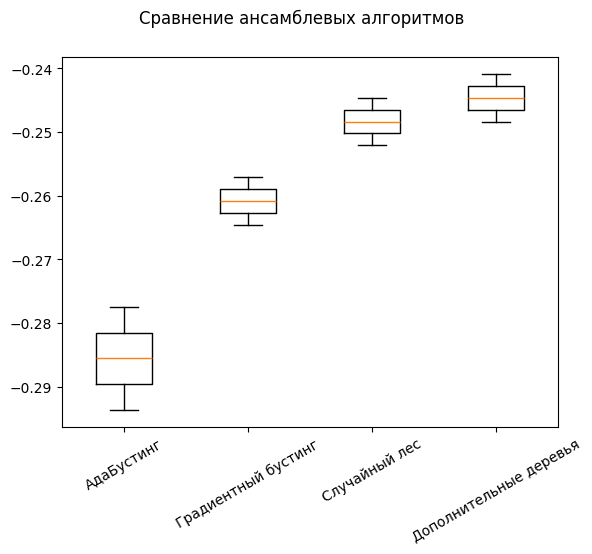


Рисунок – Сравнение ансамблевых алгоритмов согласно MAPE.

По сравнению с реализацией данных алгоритмов на стандартизированных данных, мы видим заметное улучшение в моделировании оценки стоимости квадратного метра с помощью случайного леса и дополнительных деревьев. Что не удивительно, ведь данные алгоритмы гораздо эффективнее работают на не стандартизированных данных. Учитывая, что два данных алгоритма показывают наилучшие результаты, стоит изучить при каких параметрах данные алгоритмы покажут наилучший результат. Рассмотрим данный вопрос с помощью GridSearch. Параметры, значение которых мы планируем отбирать:

Количество деревьев в ансамбле - чем больше деревьев, тем более стабильным и точным будет предсказание модели, но также увеличится время обучения и предсказания. Рассматриваем следующие значения: 50, 100, 150, 200, 250, 300, 350, 400.

Максимальная глубина каждого дерева: ограничивает количество уровней в дереве, регулируя степень переобучения данных. Изучаем следующие возможные значения: без ограничений, 10, 20, 30, 40, 50.

Минимальное количество образцов, необходимое для создания одного листа: данный параметр помогает контролировать рост дерева. Проверяем оптимальное количество среди следующих значений: 1, 2, 4.

Использование или не использование возвращения случайной выборки с возвращением.

В первую очередь запустим поиск по сетке для ансамбля дополнительных деревьев. Рассчитав значение среднего абсолютного процентного отклонения для 288 комбинаций параметров, мы получаем значение САПО равное 24.3562%, при параметрах соответственно указанным ранее: 400, 1, 20, не использование возвращения случайной выборки с возвращением.

Применим данные рассчитанные параметры для построения итоговой модели ансамбля дополнительных деревьев. Обучим модель на тренировочных данных, которые составляют 80% всего набора данных. Набор показателей, которые критически важно рассмотреть при оценке результата модели, включает в себя: коэффициент детерминации, для понимания как расчетные параметры объясняют изменение изучаемого параметра, среднее квадратическое отклонение, далее указываемое как СКО, для понимания разброса значений относительно среднего значения, среднее абсолютное отклонение, далее упоминаемое как САО, для оценки отклонения отдельных значений от среднего значения и среднее абсолютное процентное отклонение, далее упоминаемое как САПО, для упрощенной интерпретации предыдущего показателя.

На тренировочных данных мы имеем следующие результаты:

* Коэффициент детерминации = 99.97 (%)
* СКО = 1198.1776
* САО = 442.7246 (руб.)
* САПО = 0.2185 (%)

Исходя из указанных показателей, модель ошибается на тренировочных данных в среднем на две десятых процента. Безусловно, данные показатели скорее говорят о негативном качестве модели. Логичным заключением стоит считать, что ансамбль дополнительных деревьев переобучен и предназначен для данных другого объема. Тем не менее, важно рассмотреть, как модель покажет себя на тестовой выборке.

* Коэффициент детерминации = 40.99 (%)
* СКО = 59850.0712
* САО = 43456.8520 (руб.)
* САПО = 21.91 (%)

Исходя из полученных статистических показателей, расчетные параметры объясняют девиацию изучаемого параметра, стоимости квадратного метра, на 40.99%, модель в среднем ошибается на 21.91%, что равняется 43456.8520 в эквиваленте рублей. Промежуточные итоги говорят о том, что собранные параметры не позволяют оценить недвижимость с высокой точностью. Об этом говорит как и невозможность объяснить изменения целевого признака, так и статистически достаточно высокая ошибка. Рассмотрим результаты подбора наилучших параметров для алгоритма случайного леса.

Рассчитав значение среднего абсолютного процентного отклонения для 288 комбинаций параметров на тренировочной выборке, мы получаем значение САПО равное 24.6256%, при параметрах соответственно указанным ранее: 100, 1, не ограниченное, использование возвращения случайной выборки с возвращением.

На тренировочных данных мы имеем следующие результаты:

* Коэффициент детерминации = 91.74 (%)
* СКО = 23103.3346
* САО = 16675.8290 (руб.)
* САПО = 8.6309 (%)

Исходя из указанных показателей, модель ошибается на тренировочных данных в среднем на восемь целых шесть десятых процента. Такой результат гораздо более последователен и разумен, по сравнению с предыдущим. Логичным предварительным заключением стоит считать, что случайный лес подходит данной выборке значительно больше, чем ансамбль дополнительных деревьев, так как он не допускает переобучения. Тем не менее, важно рассмотреть, как модель покажет себя на тестовой выборке.

* Коэффициент детерминации = 44.97 (%)
* СКО = 57794.9583
* САО = 43021.9394 (руб.)
* САПО = 21.49 (%)

Исходя из полученных результатов алгоритма случайного леса на тестовой выборке мы замечаем улучшение, по сравнению с ансамблем дополнительных деревьев, что только подтверждает сделанное нами заключение. Основной и итоговой моделью данного исследования стоит считать модель случайного леса, расчетные параметры которой объясняют девиацию изучаемого параметра на 44.97 %, при этом допуская ошибку модели в среднем равную 21.49 %, эквивалентную 43021.94 рублям.

После окончания подбора оптимального алгоритма и создания модели, важно понимать, как мы получаем данные значения. Встает вопрос интерпретации параметров и их влияния на моделированное значение целевого параметра.

**Интерпретация параметров**

После создания оптимальной модели для оценки стоимости квадратного метра, интерпретации статистических показателей, важно понимать, как данная модель функционирует. В отличии от линейной регрессии, интерпретация параметров и их влияния на результат модели в ансамблевом алгоритме, а именно в случайном лесе, не настолько прямолинейна. Она требует подхода более подробного, хоть и не подразумевает такого строгого определения влияния тех или иных параметров, как в случае с линейной регрессией.

Одним из способов понимания, на чем модель строит свою оценку – использование показателя важности признаков. Важность каждого признака показывает, насколько полезен этот признак для модели при предсказании целевой переменной. Важность признака рассчитывается на основе того, как часто и с каким выигрышем в точности модель использует этот признак при делении данных на узлы в деревьях. Чем больше значение, тем важнее этот признак для модели. Рассмотрим вывод важности признаков для построенной модели случайного леса для оценки стоимости квадратного метра коммерческого торгового объекта недвижимости на следующем рисунке.

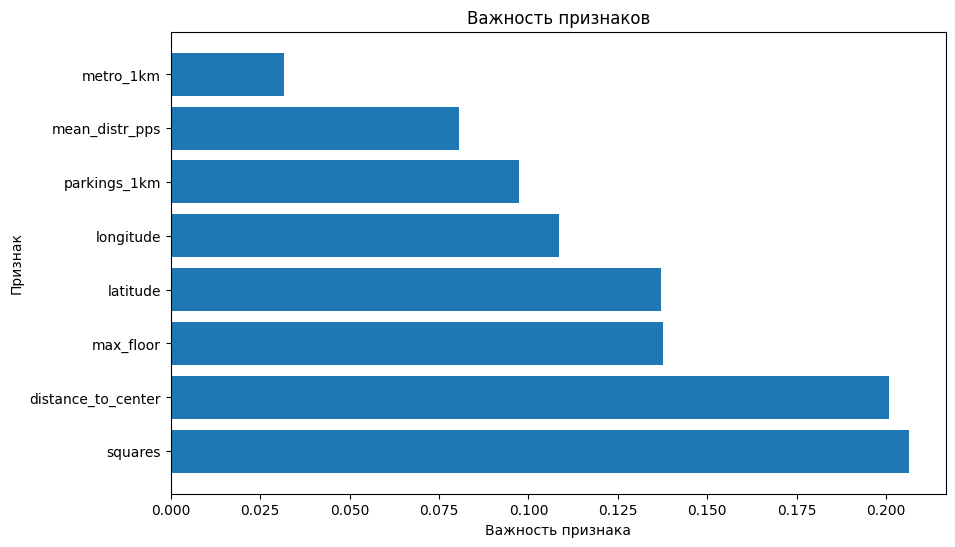


Рисунок – Важность признаков для модели.

Для более точного рассмотрения важности, стоит рассмотреть более подробные значения важности, но тем не менее, можем заметить, что основными признаками, влияющими на оценку модели являются площадь объекта и географическое расстояние до рассчитанного центра города Санкт-Петербурга. Наименее важным же признаком является количество станций метро в радиусе километра от объекта, что могло возникнуть ввиду большого количества наблюдений, в которых данный признак равняется нулю, снижая количество использований данного признака в процессе построения модели. Указанное в рисунке № влияние признаков надлежит более подробному изучению, стоит также отметить, что сумма влияния всех признаков равняется 100%. Площадь имеет влияние, равное 20.6374 %, расстояние до рассчитанного центра равняется 20.0837 %, максимальный этаж составляет 13.7506% влияния, широта влияет на 13.7156 % из 100 %, долгота же на 10.8559 %, количество парковок в радиусе 1 километра на 9.7259 %, средняя цена квадратного метра на 8.0524 %, и влияние количества станций метро в радиусе 1 километра равняется 3.1784 % от общей суммы.

Тем не менее, это также может говорить нам об отсутствии возможности достижения такой точности на нынешнем рынке коммерческой торговой недвижимости в Санкт-Петербурге. Рынок претерпел множество структурных изменений ввиду неожиданных изменений макроэкономических параметров, геополитических причин и социального напряжения. С учетом такого рода факторов – нельзя сказать, что данное исследование не показало возможность использовать модель для корректной оценки на рынке. Большой преградой в данном исследовании стало в том числе ограниченная доступность критически важных данных для оценки недвижимости. Тем не менее, учитывая проделанную работу и имеющиеся обстоятельства, которые оказывают отрицательное влияние на исследование подобного рода можно заявить о достаточно хороших результатах, позволяющих определить рамки стоимости квадратного метра для каждого торгового объекта при знании его основных параметров.