



МИНОБРНАУКИ РОССИИ  
федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ  
УНИВЕРСИТЕТ»  
(СПбГЭУ)

**ОТЧЕТ**  
**по Производственной преддипломной практике**

Наименование организации прохождения практической подготовки:

**ООО «Актив-Инвест»**  
(наименование организации)

Направление **38.03.01 «Экономика»**  
(шифр, наименование)

Направленность: **Математическое моделирование и анализ данных в  
экономике** (наименование)

Обучающийся **Дмитриев Александр Ростиславович**  
(Ф.И.О. полностью)

Группа **Э-2010**  
(номер группы)

Подпись

Руководитель  
по практической подготовке от СПбГЭУ  
**Заграновская Анна Васильевна, к.э.н., доцент, доц.**  
(Ф.И.О., ученая степень, должность, ученое звание)

\_\_\_\_\_  
(подпись руководителя)

Оценка по итогам защиты отчета \_\_\_\_\_

Санкт-Петербург  
2024 г.



МИНОБРНАУКИ РОССИИ  
федеральное государственное бюджетное образовательное  
учреждение  
высшего образования  
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ  
ЭКОНОМИЧЕСКИЙ  
УНИВЕРСИТЕТ»  
(СПБГЭУ)

СОГЛАСОВАНО:

Руководитель по практической подготовке от  
профильной организации  
(заполняется в случае прохождения  
практической подготовки в профильной  
организации)

УТВЕРЖДАЮ:

Заведующий кафедрой прикладной математики  
и экономико-математических методов

Директор ООО «АКТИВ-ИНВЕСТ»  
Елисеева Елена Сергеевна

Фридман Григорий Морицович



(подпись)

« » 20\_\_ г.

### Индивидуальное задание на производственную преддипломную практику

Обучающегося 4 курса  
(курс обучения)

Дмитриева Александра Ростиславовича  
(Ф.И.О. полностью)

Направление/специальность: 38.03.01 «Экономика»

Направленность: Математическое моделирование и анализ данных в экономике

Тема ВКР/НКР: Моделирование оценки объекта недвижимости  
(заполняется в случае прохождения преддипломной практики)

Наименование организации прохождения практической подготовки:  
ООО «АКТИВ-ИНВЕСТ»

Сроки практической подготовки с 18.04.2024 по 17.05.2024

Руководитель по практической подготовке от СПбГЭУ

Полякова Светлана Петровна, доцент кафедры ПМиЭММ

Совместный рабочий график  
с указанием видов работ, связанных с будущей профессиональной  
деятельностью

№ п/п	Перечень заданий, подлежащих разработке	Календарные сроки (даты выполнения)
1	Выявление и изучение особенностей функционирования современного российского рынка коммерческой недвижимости.	18.04.2024 - 22.04.2024
2	Сбор и анализ данных, необходимых для дальнейшего использования при построении моделей.	23.04.2024 – 01.05.2024
3	Согласование с руководителем практики от предприятия индивидуального задания на практику.	02.05.2024
4	Использование методов машинного обучения для оценки стоимости квадратного метра торгового объекта.	03.05.2024 - 09.05.2024
5	Анализ полученных результатов, внесение корректировок в данные для улучшения результатов.	10.05.2024 - 14.05.2024
6	Сделать выводы об итогах. Интерпретировать полученные результаты.	15.05.2024 - 16.05.2024
7	Обобщение материалов и подготовка отчёта и презентации по результатам практики.	17.05.2024

С заданием ознакомлен \_\_\_\_\_  
(подпись обучающегося)

Руководитель по практической подготовке от СПбГЭУ

\_\_\_\_\_  
(Подпись) Полякова С.П.  
(Расшифровка)

Руководитель по практической подготовке от профильной организации  
(заполняется в случае прохождения практической подготовки в профильной организации)

\_\_\_\_\_  
(Подпись) Елисеева Е. С.  
(Расшифровка)

Обучающийся прошел инструктаж по ознакомлению с требованиями охраны труда, техники безопасности, пожарной безопасности, а также с правилами внутреннего распорядка. Вводный инструктаж и инструктаж на рабочем месте пройдены с оформлением установленной документации.

Руководитель по практической подготовке от организации/профильной организации назначен приказом № \_\_\_\_\_ дата \_\_\_\_\_ и соответствует требованиям трудового законодательства Российской Федерации о допуске к педагогической деятельности.

\_\_\_\_\_  
(Ф.И.О. должность)  
(подпись)



## АКТИВ-ИНВЕСТ

Общество с ограниченной ответственностью «АКТИВ-ИНВЕСТ»

ИНН / КПП 2308016670 / 230801001

ОГРН 1022301221360

Юридический адрес: 350020, КРАСНОДАРСКИЙ КРАЙ, Г. КРАСНОДАР, УЛ.

РАШПИЛЕВСКАЯ, Д. 179, К. 1, ПОМЕЩ. 45

ОКПО 26577398

ОКАТО 03401364000

### ОТЗЫВ О ПРОХОЖДЕНИИ ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ (ПРЕДДИПЛОМНОЙ)

Обучающегося: Дмитриева Александра Ростиславовича

Направление/специальность: 38.03.01 «Экономика»

Направленность (профиль)/специализация: Математическое моделирование и анализ данных в экономике

В организации в период: с 18.04.2024 по 17.05.2024

#### Результаты работы

За время прохождения производственной практики в ООО «АКТИВ-ИНВЕСТ» студент 4 курса факультета экономики и финансов Дмитриев Александр Ростиславович выполнил задачи по сбору, обработке и анализу данных, после чего смоделировал оценку стоимости квадратного метра торгового объекта коммерческой недвижимости, а также сделал выводы о проделанной работе и ее результатах, учитывая также и внесенные корректировки.

#### Личные и деловые качества (компетенции)

В период прохождения практики, Дмитриев Александр Ростиславович показал себя как ответственного работника, заинтересованного в улучшении своих навыков, наряду с качественным подходом к выполнению поставленных задач.

#### Качество отчета, оценка

Отчет по проделанной работе свидетельствует о высоком уровне теоретической подготовки, способности корректно применять имеющиеся знания на практике. Выполнен в срок, в полном объеме, в соответствии с индивидуальным заданием и с учетом всех требований. Рекомендуемая оценка – «Отлично».

#### Рекомендации

Рекомендовано изучить возможность создания среды для практического использования моделей для оценки.

Подпись: \_\_\_\_\_

Должность: Директор ООО «АКТИВ-ИНВЕСТ»

ФИО: Елисеева Елена Сергеевна



Тел. +7 (861) 279-84-00

## ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ .....	6
1. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ИССЛЕДОВАНИЯ В РАМКАХ КОММЕРЧЕСКОЙ НЕДВИЖИМОСТИ .....	8
1.2. ПОСТАНОВКА ЗАДАЧИ И ОПРЕДЕЛЕНИЕ ИСТОЧНИКА ДАННЫХ.....	11
2. СБОР ДАННЫХ, ИХ АНАЛИЗ И ПОСТРОЕНИЕ МОДЕЛЕЙ .....	13
2.1 . СБОР ДАННЫХ.....	13
2.2. РАБОТА С КООРДИНАТНЫМИ ДАННЫМИ .....	16
2.3. ГЕОГРАФИЧЕСКИЙ АНАЛИЗ .....	18
2.4. ОБРАБОТКА ДАННЫХ И ПОДГОТОВКА К ПОСТРОЕНИЮ МОДЕЛЕЙ .....	20
2.5. ВИЗУАЛИЗАЦИЯ ОСОБЕННОСТЕЙ ДАННЫХ .....	21
2.6. ПОСТРОЕНИЕ МОДЕЛЕЙ.....	25
2.7. ПОДБОР НОВЫХ ПАРАМЕТРОВ .....	31
ЗАКЛЮЧЕНИЕ .....	39
СПИСОК ИСТОЧНИКОВ .....	40
ПРИЛОЖЕНИЕ А .....	42

## **ВВЕДЕНИЕ**

Выпускная квалификационная работа является показателем знаний и навыков студента, поэтому практическая часть является основной при подведении итогов работы. В настоящем отчете упор будет сделан на исследование оценки объектов на российском рынке коммерческой недвижимости, сбора данных, анализа и последующего выявления закономерностей для построения модели, способной корректно и структурно оценивать объект недвижимости, в зависимости от определенных параметров. В настоящем отчете будет представлен результат практической деятельности, которая войдет в ВКР.

Целями преддипломной практики являются:

- закрепление и углубление теоретических знаний студентов, полученных в процессе обучения на всех видах аудиторных и внеаудиторных занятий;
- закрепление навыков научно-исследовательской работы;
- приобретение самостоятельного опыта и овладение практическими навыками, передовыми методами труда в финансовых учреждениях, организациях и корпорациях;
- сбор, анализ и обобщение материалов для написания выпускной квалификационной работы;
- овладение методами и приемами сбора данных, их обработки, анализа и построения моделей.

Задачами преддипломной практики являются:

- Выявление и изучение особенностей функционирования современного российского рынка коммерческой недвижимости.
- Сбор и анализ данных, необходимых для дальнейшего использования при построении моделей.
- Согласование с руководителем практики от предприятия индивидуального задания на практику.

- Использование методов машинного обучения для оценки стоимости квадратного метра торгового объекта.
- Анализ полученных результатов, внесение корректировок в данные для улучшения результатов.
- Сделать выводы об итогах.
- Интерпретировать полученные результаты.
- Обобщение материалов и подготовка отчёта и презентации по результатам практики.

Объект исследования – ООО «АКТИВ-ИНВЕСТ». Предметом исследования является российский рынок коммерческой недвижимости.

В условиях современной экономики, роль рынка коммерческой недвижимости несомненна. Она обусловлена динамичным развитием рынка последние годы, только за 2023 год объем инвестиций вырос в 1,5 раза по всему коммерческому сектору недвижимости, по торговой – в 2. Эффективное управление активами, точная и объективная оценка и прогнозирование стоимости коммерческой недвижимости могут значительно повысить инвестиционную привлекательность и экономическую стабильность компаний и частных инвесторов, не говоря уже об улучшении качества оказываемых предприятиями услуг в рамках работы с недвижимостью.

Очень важно для проведения такого рода анализов осуществить доскональное изучение доступных источников данных, провести полноценный анализ сферы коммерческой недвижимости и принципов оценки объектов.

# **1. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ИССЛЕДОВАНИЯ В РАМКАХ КОММЕРЧЕСКОЙ НЕДВИЖИМОСТИ**

## **1.1. ОСОБЕННОСТИ ФУНКЦИОНИРОВАНИЯ СОВРЕМЕННОГО РОССИЙСКОГО РЫНКА КОММЕРЧЕСКОЙ НЕДВИЖИМОСТИ**

Если рассматривать рынок недвижимости как совокупность взаимодействий рынков разного уровня, которые определяют процессы руководства и использования объектов недвижимости, то первоочередно необходимо понимать на что разделен рынок недвижимости, на чем основывается ценообразование объектов, а также, какие методы позволяют рассматривать данную сферу с точки зрения статистики и с целью обнаружения взаимосвязей.

Рынок разделяется на жилую и нежилую (коммерческую) недвижимость. Коммерческая недвижимость – это собственность, используемая исключительно для решения бизнес-ориентированных задач, то есть получения дохода, или для создания рабочего места, вместо жилого помещения. Это обширная категория недвижимости, включающая три основных вида:

- Офисные объекты — это коммерческие объекты, предназначенные для размещения предприятий, организаций и профессиональных служб. Они могут быть как небольшими, рассчитанными на одного арендатора, так и крупными комплексами. Офисные здания в своей сути делятся на три категории в зависимости от качества, местоположения, оснащения и арендных ставок: класс А, класс В и класс С.
- Торговые помещения (помещения свободного типа) — это коммерческие объекты, предлагающие потребительские товары и услуги. Торговые площади часто классифицируются в зависимости от их местоположения и состава арендаторов.
- Промышленные (складские) объекты - к промышленным объектам относятся складские помещения, распределительные центры,



производственные мощности и складские помещения для гибких перевозок. Они предназначены для производства, хранения и доставки товаров.

Оценка же рыночной стоимости недвижимого имущества на территории российской федерации выполняется в соответствии 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 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.

Оценка рыночной стоимости недвижимого объекта должна производиться структурно, с определенной этапностью:

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 объекта – важно правильно и точно определять, каким методом и когда лучше воспользоваться, а для этого безусловно необходимо понимание какие методы существуют и используют. Тем не менее, важно применять и рассматривать данные методы в совокупности.

Доходный подход – совокупность методов оценки, основанных на определении потенциальных доходов от использования объекта оценки – аренды/перепродажи.

Сравнительный подход – методы оценки, основанные на сравнении объекта оценки с аналогами на рынке. Аналог определяется, как по техническим характеристикам объектам, так и по геопозиции и планировке.

Затратный подход – совокупность методов оценки, основанных на определении затрат, необходимых для воспроизводства или замещения объекта оценки с учетом износа или устаревания.

При изучении же самого рынка – можно наблюдать отсутствие строгой зависимости между различными параметрами самого объекта и его стоимостью. Зачастую крупным фактором ценообразования служит длительность открытого объявления о продаже или другие внешние факторы. Ключевые факторы объекта безусловно представляют из-себя основной набор значений, влияющий на стоимость, однако тенденции последних лет

показывают не только рост средней стоимости квадратного метра объекта, но и отсутствие единой системы оценки объектов. Во многом потому, что такую крупную сферу практически невозможно контролировать полностью, тем не менее за счет времени и внешних факторов изначально некорректная, то есть не соответствующая рыночной оценке, цена имеет закономерность снижаться к линии спроса в данном сегменте.

Во многом благодаря такому недостатку сферы – очень важно использовать научный подход при оценке объекта, исключая факторы человеческой ошибки и другие недостатки.

## **1.2. ПОСТАНОВКА ЗАДАЧИ И ОПРЕДЕЛЕНИЕ ИСТОЧНИКА ДАННЫХ**

Для анализа рынка и построения в дальнейшем модели – необходимо собрать основную выборку по недвижимости. В большинстве компаний, несмотря на наличие собственной базы объектов - регулярно практикуется сбор и использование открытых источников, в особенности для обновления базы данных и понимания функционирования рынка, нахождения закономерностей. Также, стоит отметить, в рамках работы будут использованы данные именно по Санкт-Петербургу и ближайшей части Ленинградской области, но это не отменяет возможностей использования данного подхода к любой местности, при наличии основных параметров.

Для основного источника данных был выбран Циан – как площадка, в сфере которой участвует больше всего агентств недвижимости, в которой находится наибольшее количество объявлений данного типа, а также как качественный сервис, руководство которого крайне подробно следит за корректностью указанных данных. Циан предоставляет к открытому доступу следующие данные со страницы объявлений:

- URL;
- Стоимость;
- Площадь;
- Стоимость за кв метр;

- Дом;
- Этаж;
- Назначение (офис, склад, торговое);
- Тип здания;
- Фиктивные переменные;
- Агентство;
- Расстояние до метро в минутах;

Используя эти и другие данные необходимо построить модель, способную производить оценку объекта – иными словами, рассчитывать стоимость квадратного метра данного объекта. Критически важным фактором в данном исследовании стоит считать координаты объектов, они критически важны будут как для обработки данных, создании новых переменных, так и для прогнозного значения модели.

## **2. СБОР ДАННЫХ, ИХ АНАЛИЗ И ПОСТРОЕНИЕ МОДЕЛЕЙ**

### **2.1. СБОР ДАННЫХ**

Для сбора данной информации придется обратиться к написанию программы-парсера. Парсер — это программа, сервис или скрипт, который собирает данные с указанных веб-ресурсов, обрабатывает их и выдает в нужном формате. Использование парсеров очень распространено в современном мире, во многом это один из основных способов получения информации из открытого доступа. Более того, в сфере недвижимости это цифровой фундамент для изучения рынка. Все современные агентства стараются постоянно следить за новыми данными и собирать максимум из них. Такой задачей и занимается один из отделов компании «Актив-Инвест».

После изучения 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. Проконсультировавшись со специалистами в «Актив-Инвест» – была проделана работа по написанию парсера специально для данного сайта, с целью собрать максимум возможной информации.

Для данной задачи были изучены основные библиотеки для парсинга на языке программирования python – “Beautiful4Soup” и “Selenium”. Используя основные API этих библиотек, были изучены возможности для сбора данных.

Выводом стало, что самым оптимальным и эффективным способом стоит считать использование библиотеки “Selenium”, во многом благодаря возможностям данной библиотеки более точно взаимодействовать с сайтами, используя имитацию нажатия кнопок, открытия ссылок и перехода от страницы к странице.

Для того, чтобы собрать все объекты, первоначально необходимо собрать все ссылки на них с предварительно настроенной фильтрацией сайта. Настроим среду для кода и импортируем необходимые библиотеки.

Был написан код для парсинга ссылок всех объектов в сфере коммерческой недвижимости в Санкт-Петербурге и Ленинградской области.

```
driver = webdriver.Chrome()
hrefs = []

try:
    for page in range(1,50):
        driver.get(url=f'https://spb.cian.ru/cat.php?deal_type=sale&engine_version=2&offer_type=offices&office_type%5B%5D=2&office_type%5B1%5D=4&office_type%5B2%5D=5&office_type%5B3%5D=7&office_type%5B4%5D=9&office_type%5B5%5D=12&p={page}&region=2&sort=creation_date_desc')
        time.sleep(10)
        print('main base loaded')
        # Постепенная загрузка страницы
        skrol = 0
        for i in range(15):
            skrol += 1000
            driver.execute_script(f'window.scrollTo(0, {skrol});')
            time.sleep(1)
        time.sleep(10)
        print(f'page {page} loaded')
        # Взятие href с каждой страницы
        for i in range(1, 34):
            try:
                item = driver.find_element(By.XPATH, f'/html/body/div[1]/div/div[5]/div[{i}]/div/div[2]/div/div[1]/div[1]/div[1]/h3/div/a').get_attribute('href')
                hrefs.append(item)
                time.sleep(1)
            except Exception as inner_ex:
                print(f'Error in inner loop: {inner_ex}')
                continue
    except Exception as ex:
        print(f'Error in outer loop: {ex}')

driver.close()
driver.quit()
```

Рисунок 1 – Парсер ссылок объявлений.

Данный код включает в себя цикл с искусственными задержками, для прохождения по каждому из 34 объявлений на всех 50 страницах сайта с определенными нами заранее условиями фильтрации. В цикле предусмотрено использование скрипта для вертикального перемещения по странице, с целью загрузки и обработки всех объявлений, взятие href-ссылок объектов, то есть url-ссылок опубликованных объявлений, учитывая структуру 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  $m^2$ ;
- Этаж;
- Свободно ли помещение;
- Город;
- Район;
- Округ;
- Улица;
- Номер дома;
- Ссылка на страницу продавца;
- Описание;
- Список ближайших станции метро (мин);
- Список ближайших выездов на шоссе (км).

Работа данного кода суммарно занимает приблизительно 7 часов с учетом обработки всех 1348 объявлений, полученных с работы предыдущего парсера. Такая длительность объясняется наличием задержек для загрузки 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 и не подразумевают какой-либо практической ценности.

## 2.2. РАБОТА С КООРДИНАТНЫМИ ДАННЫМИ

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

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

После ознакомления с особенностями геокодера — был создан ключ API для использования его внутри среды кода на языке программирования python и написана функция для отправления запросов в виде адреса и получения результата в формате координат. Также имеются 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.

```
import requests

def get_coordinates(address):
    api_key = "ce2524a7-073a-41e9-8f8d-47caa1182bdc"
    url = f"https://geocode-maps.yandex.ru/1.x/?apikey={api_key}&geocode={address}&format=json"
    response = requests.get(url)
    data = response.json()
    if response.status_code == 200:
        found = int(data['response']['GeoObjectCollection']['metaDataProperty']['GeocoderResponseMetaData']['found'])
        if found > 0:
            coords_str = data['response']['GeoObjectCollection']['featureMember'][0]['GeoObject']['Point']['pos']
            coordinates = tuple(map(float, coords_str.split()))
            return coordinates
        else:
            print("Объекты по заданному адресу не найдены.")
            return None
    else:
        print("Ошибка при получении координат.")
        return None
```

Рисунок 2 — Код получения координат с помощью API геокодера.

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 2 этапа. Вместе со вторым этапом также были собраны координаты

открытых и ремонтирующихся станций метрополитена в Санкт-Петербурге. На момент отправления запросов – их число составляло 72.

Динамику и статистику запросов также можно рассмотреть в отдельном блоке сайта геокодера.

Для получения максимума пользы из информации по метро, рассчитаем к основным данным расстояние до всех станций метро, чтобы потом мы могли определить, какие станции оставить, а какие убрать. К тому же рассчитаем количество станций метро на определенных расстояниях от объекта. Для этого обратимся к формуле расчета расстояния между двумя точками на координатной плоскости Земли:

$$L = q * \arccos(\sin\varphi_1 * \sin\varphi_2 + \cos\varphi_1 * \cos\varphi_2 * \cos|\lambda_1 - \lambda_2|)$$

Где:

$L$  — искомое расстояние между пунктами 1 и 2;

$q$  — расстояние в одном градусе, константа  $\sim 111.2$  км;

$\varphi_1, \varphi_2$  — географические широты пунктов 1 и 2;

$\lambda_1, \lambda_2$  — географические долготы пунктов 1 и 2.

Используя эту формулу, рассчитываем расстояние между каждым объектом и каждой станцией метро, используя циклы. Однако стоит учитывать возможное возникновение погрешностей и неточностей.

Помимо этого, предварительно была найдена платная база данных по Санкт-Петербургу, включающая основные объекты, по типу зданий, парковок, ж/д путей, дорог и так далее.

Для поставленной задачи – прогнозирования стоимости квадратного метра торгового помещения – наличия 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.

### 2.3. ГЕОГРАФИЧЕСКИЙ АНАЛИЗ

С учетом имеющихся географических данных, с помощью свободной кроссплатформенной геоинформационной системы – QGIS, предназначенной для создания, редактирования, визуализации, анализа и публикации геопространственной информации, выведем их на координатную систему с наложенной на нее картой Санкт-Петербурга для проведения ознакомительного анализа.

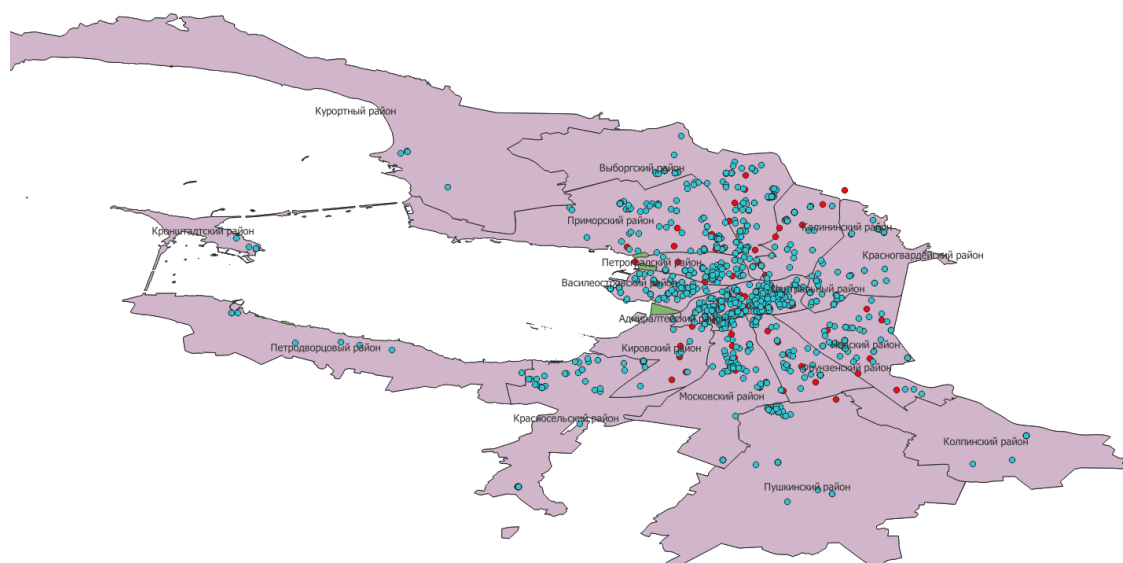


Рисунок 3 – Визуализация объектов в QGIS.

На данной карте расположен Санкт-Петербург, разделенный на административные деления – районы. Синим цветом на данном рисунке обозначены объекты, которые находятся в собранных данных, красным – станции метро.

При рассмотрении более подробно – можно также обнаружить, что на карте обозначены все здания и парковки, существовавшие в Санкт-Петербурге на 2022 год, это поможет нам проверить совпадение значений координат из данных с реальными зданиями.



Рисунок 3 – 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, не считая, конечно же, удаленных от центра районов города.

Стоит также рассмотреть вывод горячей карты, с учетом цены квадратного метра, чем ярче скопление точек на карте – тем дороже там квадратный метр. Видим примерно аналогичные результаты, основное скопление объектов с высокой стоимостью квадратного метра находится также в центральной части города:



Рисунок 4 – Горячая карта объектов по стоимости квадратного метра.

## **2.4. ОБРАБОТКА ДАННЫХ И ПОДГОТОВКА К ПОСТРОЕНИЮ МОДЕЛЕЙ**

Для дальнейшей работы нам все-таки необходимо снова обратиться к методам обработки данных, однако уже не первичной – необходимо обработать выбросы, аномальные значения и пропуски в столбцах данных, которые могут иметь значимость для нас как исследователей.

При изучении объявлений и столбцов с данными были обнаружены ошибки в оформлении информации об объектах, в связи с чем цена за квадратный метр была ниже, чем само количество квадратных метров – в связи с чем есть необходимость в удалении указанных данных из нашей выборки. Также, учитывая тенденции рынка и его средние значения, а также консультацию с агентами по недвижимости – было решено не учитывать в выборке торговые объекты с площадью больше 200 метров квадратных. Они представляют собой несколько иную подкатегорию, нежели изучаемую в



данной работе, и имеют очень низкую востребованность ввиду своей дороговизны, в связи с чем гораздо дольше находятся на рынке и корректируют свою цену для привлечения инвесторов или покупателей.

Помимо этого, при определении целевого признака – цены за квадратный метр, возникла надобность в исключении полной стоимости объекта, она бы не позволяла рассматривать построение моделей без учета сильной межфакторной корреляции, по этой же причине данный фактор и не был выбран как целевой.

Использование всех данных по станциям метро не показало наличие какой-то серьезной зависимости и создавало лишь большие мультиколлинеарные проблемы, поэтому было решено исключить данные параметры.

Целевой признак также нуждается в обработке, ведь помимо обработки аномальных значений, стоит задача максимальной нормализации данных – для улучшения результатов будущих моделей.

## **2.5. ВИЗУАЛИЗАЦИЯ ОСОБЕННОСТЕЙ ДАННЫХ**

При изучении целевого признака, обратим внимание в первую очередь на распределение параметра:

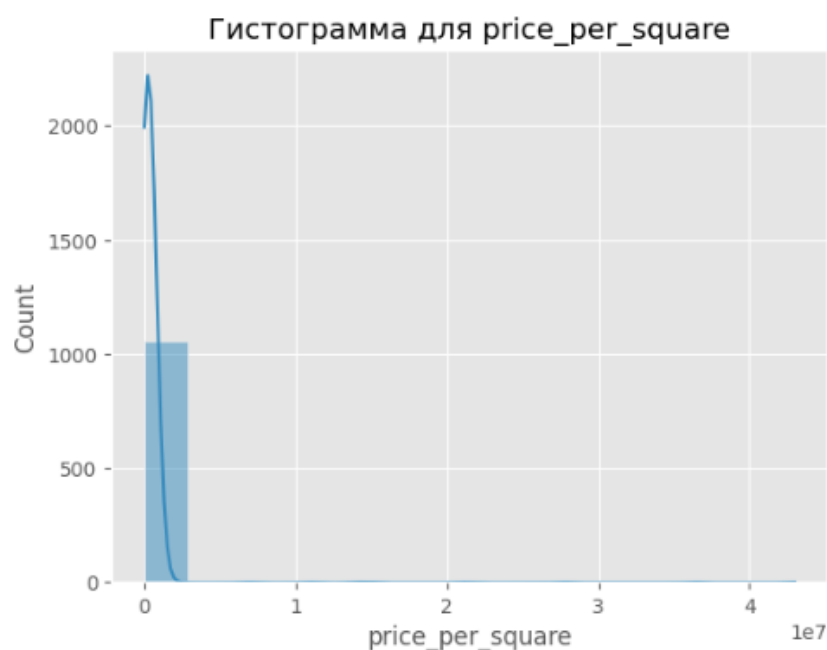


Рисунок 5 – Распределение целевого признака.

С учетом такого результата, обратим внимание на anomalously высокие и критически низкие показатели стоимости квадратного метра. Стоит при этом учесть, что такого рода данные по большей части были ошибочно определены в исходном источнике, однако даже если некоторые из них не ошибочные – такие данные во всяком исходе не пригодны для нашего исследования, так как попадают кардинально в другое направление коммерческой сферы.

Поэтому мы обрабатываем данные, исключая объекты со стоимостью квадратного метра менее 30 000 рублей и более 450 000 рублей. При этом также убираем данные ниже 5% квантиля и выше 95%, чтобы даже с учетом исключенных данных – устранить разброс значений, нормализовать распределение. Выведем гистограмму и получим следующий результат:

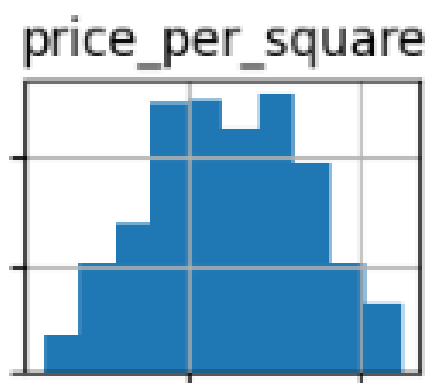


Рисунок 6 – Распределение обработанного целевого признака.

Выведем подобную статистику для каждого оставшегося параметра в наших данных:

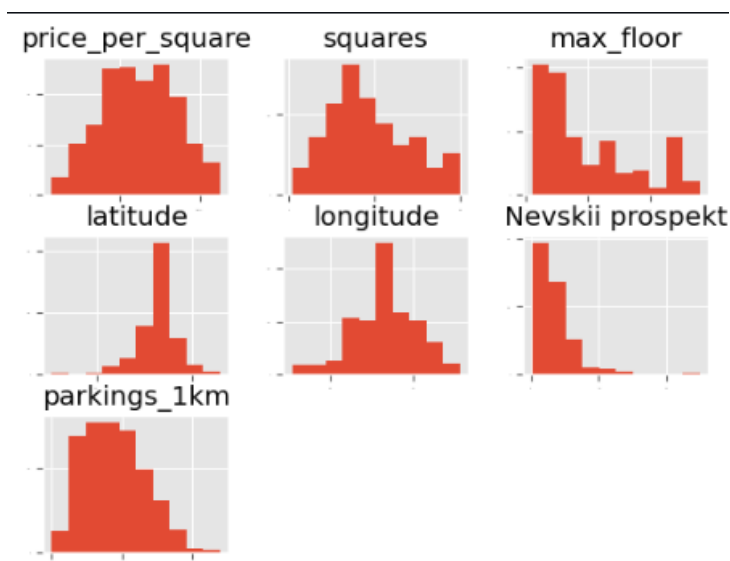


Рисунок 7 – Распределение используемых параметров.

Видим, что целевой признак нормализован, в то время как некоторые зависимые имеют определенную асимметрию. Рассмотрим количество наблюдений, объект в которых находится на разных расстояниях от центра, определим с какого расстояния, можно считать, что объекты нерелевантны или не настолько востребованы, чтобы оставлять наличие выбросов.

Путем практического исследования было решено остановиться на 30 км от центра – таким образом мы не исключаем все объявления за чертой города, однако не учитываем чрезмерно удаленные, позволяя нам больше отталкиваться от цен города. Смотрим на изменения, а также на графики плотности данных параметров:

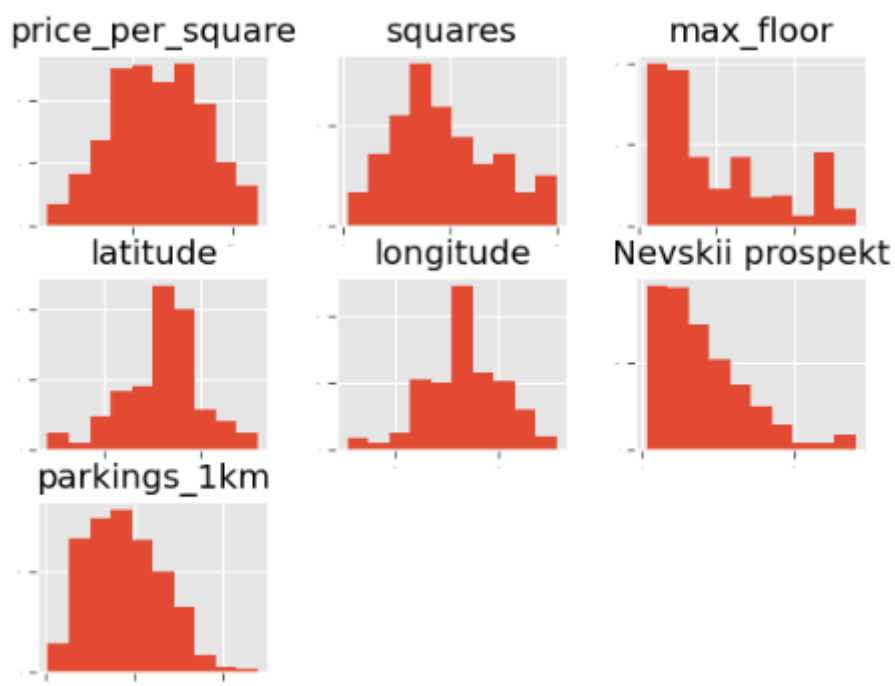


Рисунок 8 – Распределение обработанных параметров.

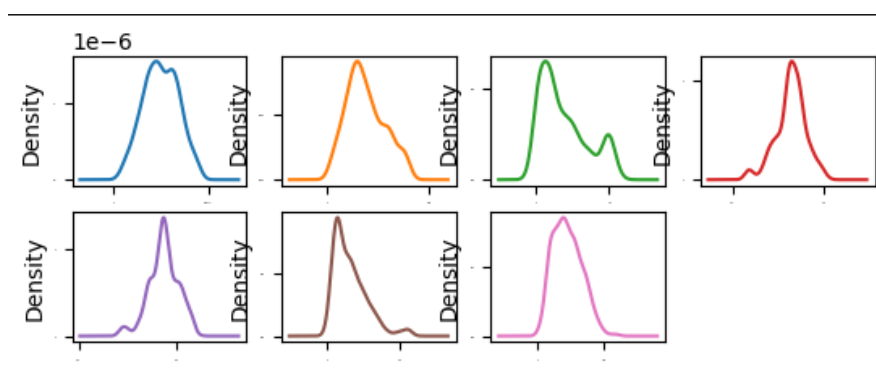


Рисунок 9 – Плотность обработанных параметров.

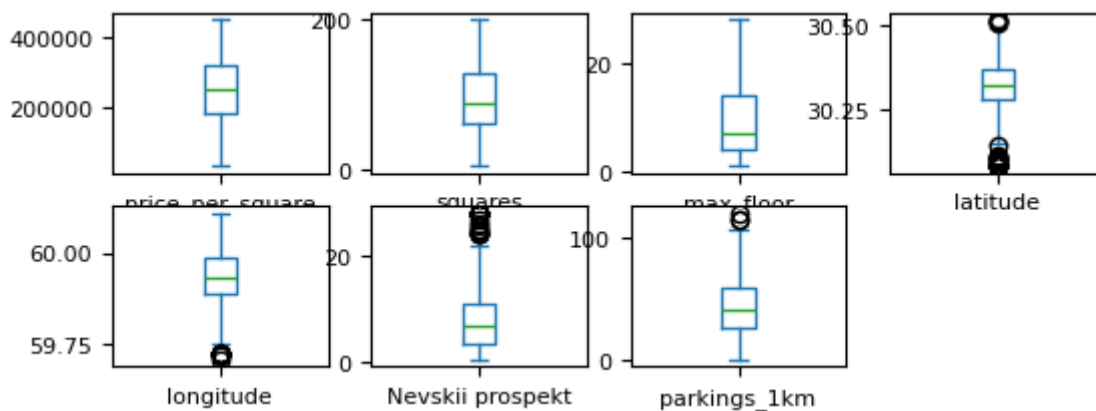


Рисунок 10 – Вохплот`ы обработанных параметров.

Не менее важно рассмотреть корреляционную зависимость между оставшимися параметрами. С учетом недостатка данных и основной зависимости стоимости объекта от координат, результаты приемлемы, но безусловно должен подниматься вопрос о нахождении большего количества параметров, которые могут оказать существенное влияние на целевой параметр – стоимость квадратного метра.

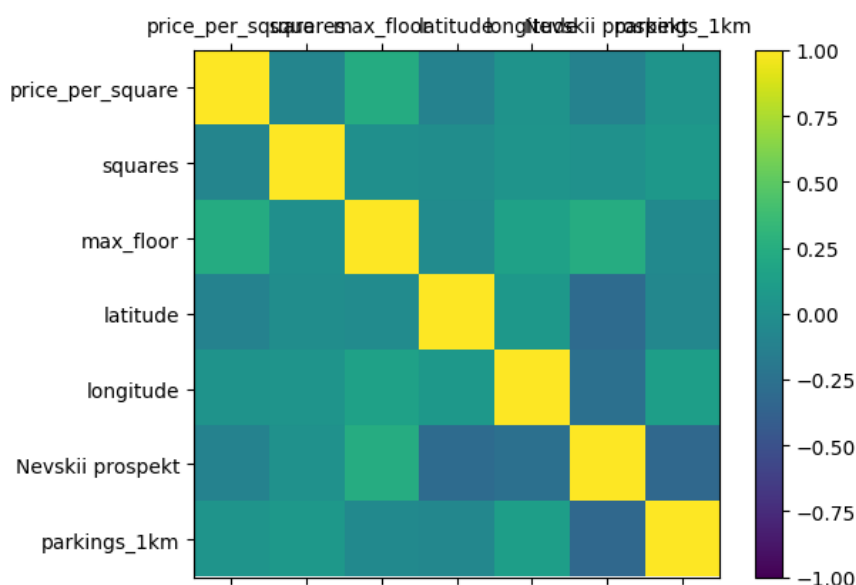


Рисунок 11 – Корреляция обработанных параметров.

После использования указанных методов разработки выводим количество строк с данными объектов, которые мы будем использовать для построения моделей. Их число - 801 объект.

## 2.6. ПОСТРОЕНИЕ МОДЕЛЕЙ

При изучении источников, рассматривавших самые эффективные способы прогнозирования цен на коммерческую недвижимость – основной вектор направления стоял к случайному лесу или методу ближайших соседей, во многом также потому, что данные методы в некоей мере нивелируют межфакторную корреляцию, что является достаточно серьезным вопросом в данной выборке - однако для структурного понимания данных необходимо рассмотреть всевозможные варианты.

Первоначально – разделим данные на тренировочную и тестовую выборки. В данном случае было выбрано распределение в 50%, а основным параметром для сравнения результатов модели будем считать MAPE - среднее абсолютное процентное отклонение прогнозируемого значения от реального. Напишем код, для вывода данной статистики по основным регрессионным методам на обычных данных с использованием кросс-валидации с заранее определенными параметрам. Выведем результаты в виде boxplot`а:

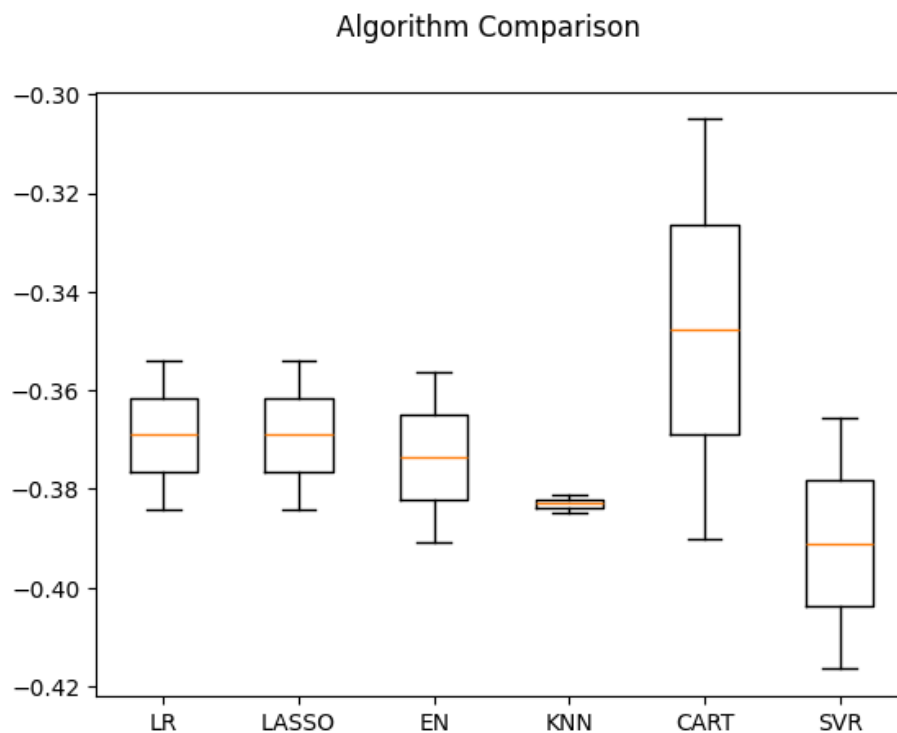


Рисунок 12 – Сравнение вывода MAPE регрессионных алгоритмов на нестандартизированных данных.

Видим, что в рамках прогнозирования на нестандартизированных данных примерно одинаково себя показывают данные методы, рассмотрим среднее значения:

LR: -0.369073;

LASSO: -0.369071;

EN: -0.373562;

KNN: -0.382985;

CART: -0.350150;

SVR: -0.390972.



Видим, что в среднем самый лучший (минимальный) результат показывает регрессор деревьев решений (CART), однако имеет достаточно большой разброс, поэтому в данном случае, предпочтительней было бы использовать метод ближайших соседей (KNN), ввиду его стабильности в показаниях. Теперь стандартизируем данные и применим их для построения этих же моделей:

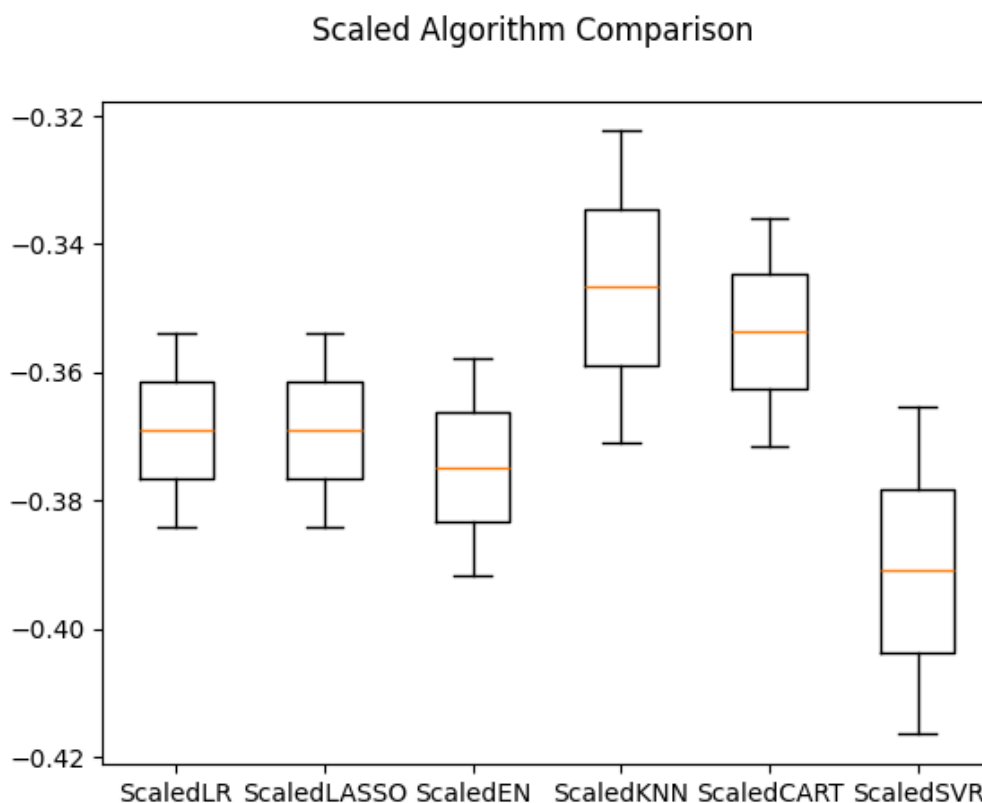


Рисунок 13 – Сравнение вывода MAPE регрессионных алгоритмов на стандартизированных данных.

Видим, что результаты не сильно улучшились по сравнению с не масштабированными данными, скорее даже ухудшились. Рассмотрим значения повнимательнее:

ScaledLR: -0.369073;

ScaledLASSO: -0.369073;

ScaledEN: -0.374864;

ScaledKNN: -0.346791;

ScaledCART: -0.353721;

ScaledSVR: -0.390965.

Все еще наиболее предпочтительным стоит считать метод ближайших соседей (ScaledKNN), хоть при стандартизированных данных стандартное отклонение значительно увеличилось по сравнению с прошлыми результатами. Линейная регрессия, L1 и комбинация L1&L2 не показали изменений по сравнению с предыдущими расчетами.

Перейдем к рассмотрению ансамблевых методов, а именно – Адабустинг, Градиентный бустинг, Случайный лес и Extra-Trees.

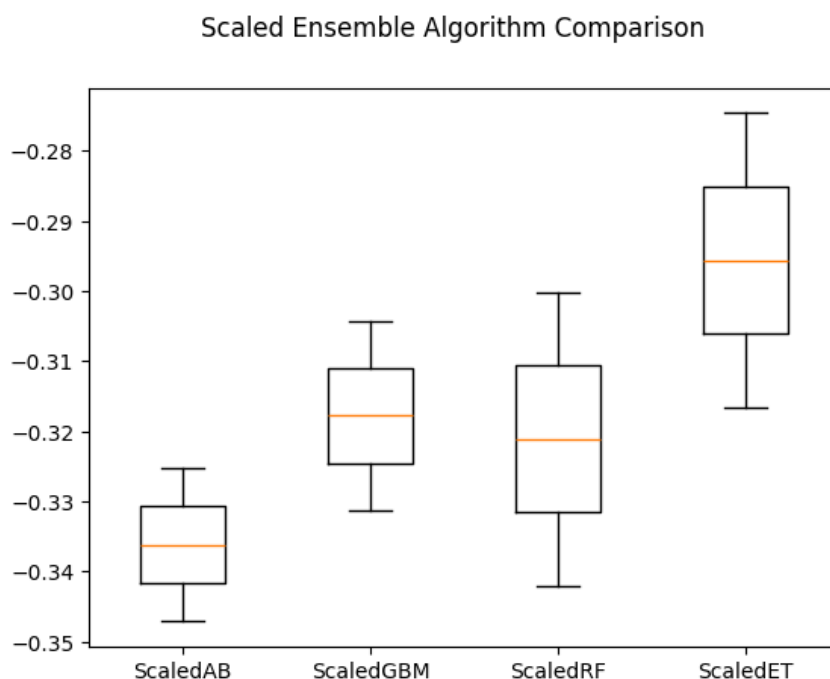


Рисунок 14 – Сравнение вывода MAPE ансамблевых алгоритмов на стандартизированных данных.

Лучший результат показывает Extra-Trees Regressor, со средним значением MAPE в 0.304085 и стандартной девиацией в 0.028873. Рассматривая другие параметры, видим –

ScaledAB: -0.343236 (0.019689);

ScaledGBM: -0.318066 (0.015337);

ScaledRF: -0.313031 (0.015420).

С учетом указанных результатов стоит присмотреться именно к Extra Trees, учитывая его значение MAPE. Стоит отметить и градиентный бустинг, у данного метода относительно небольшой разброс и его приемлимое значение

процента абсолютного отклонения. Применим метод GridSearch для поиска наилучших параметров для Extra-Trees модели. Согласно выводу сети – лучший результат MAPE достигается при 150 базовых моделях на тестовых данных и равен 0.292196. После обучения модели с такими параметрами на тренировочной выборке и применении модели на тестовой получаем следующие результаты. Среднее абсолютное отклонение равняется ~56660 рубля, что в процентном отклонении равняется 35.3%.

Для понимания качества результата – обратимся к более подробному рассмотрению модели – влияние параметров. Ввиду невозможности интерпретировать ансамблевые методы настолько же точно, как регрессионные – обратимся к методам SHAP. Это метод интерпретации моделей машинного обучения, который предоставляет объяснения для индивидуальных прогнозов. Он основан на концепции кооперативных игр и использует значения из теории игр для определения вклада каждого признака в прогноз модели. Сами SHAP-значения — это оценки вклада каждого признака в прогноз модели для конкретного наблюдения данных. SHAP-значения представляют собой разницу между средним прогнозом модели и прогнозом модели для данного наблюдения при условии определенного набора признаков.

Рассмотрим предварительно важность признаков, используемых в модели.

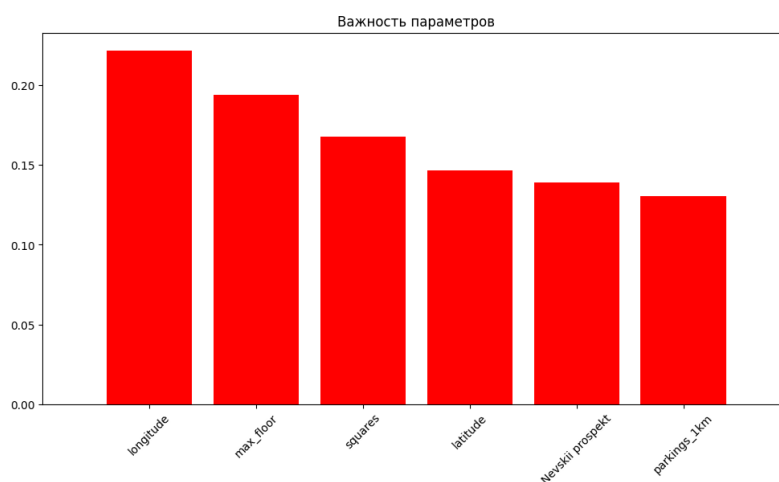


Рисунок 15 – Сравнение важности параметров для модели ETR.

Переходя к интерпретации роли параметров методами SHAP – стоит отметить основные принципы – на данном графике представлены по оси Y – ранжированные по важности параметры, важность отдельных наблюдений можно также заметить благодаря цветовой разметке, красные наблюдения оказались более важными, более влиятельными для модели, чем синие. По оси X же располагаются SHAP-значения, которые, как уже было сказано отображают вклад каждого признака, отрицательный или положительный, в прогноз модели относительно среднего прогноза. Распределение наблюдений по оси X позволяет оценить как в среднем каждый из параметров влияет на прогнозное значение. Помимо этого – данные на графике стандартизованы

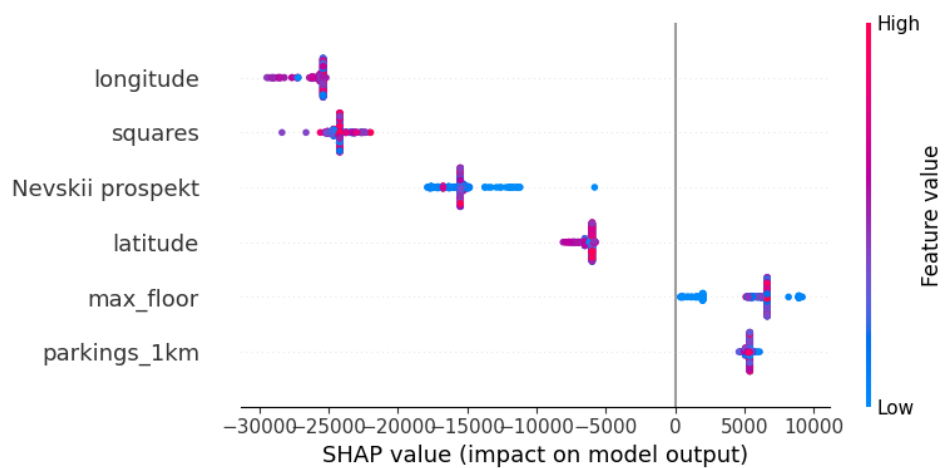


Рисунок 16 – Сравнение влияние параметров для модели ETR согласно SHAP.

Если рассматривать средние значение, получится выявить более структурную интерпретацию:

- squares -24237.017969;
- max\_floor 5622.295619;
- latitude -6181.290438;
- longitude -25707.471252;
- Nevskii prospekt -15534.455458;
- parkings\_1km 5290.673031.

Согласно модели, получается, что в среднем площадь объекта недвижимости снижает стоимость квадратного метра данного объекта на 24 тысячи 237 рублей; в среднем, максимальный этаж в здании объекта повышает стоимость его квадратного метра на 5 тысяч 622 рубля; в среднем, координаты широты объекта понижают стоимость квадратного метра объекта на 6 тысяч 181 рубль; в среднем, координаты долготы объекта понижают стоимость объекта на 25 тысяч 707 рублей; в среднем, расстояние объекта от центра Санкт-Петербурга понижают стоимость его квадратного метра на 15 тысяч 534 рубля; в среднем, количество парковок в радиусе одного километра от объекта, повышает стоимость квадратного метра объекта на 5 тысяч 290 рублей.

Такие результаты не являются предпочтительными – слишком большое отклонение между прогнозными значениями и фактическими. Для решения подробного рода проблемы стоит отталкиваться от поиска новых факторов и добавления их для повышения значимости. Более того, стоит рассмотреть вопрос расширения выборки, ввиду достаточного малого количества наблюдений, что тоже негативно сказывается на обучении модели. Также, стоит обратить внимание на использование данных без стандартизации – при сравнении результатов случайного леса с и без стандартизации – второй вариант показал себя результативнее.

## **2.7. ПОДБОР НОВЫХ ПАРАМЕТРОВ**

Первой частью плана по улучшению результатов стоит считать привлечение новых данных и нового фактора к ним. Решением стала выборка данных с Циана по аренде аналогичных торговых помещений. Был отредактирован использованный ранее парсер, после чего все данные были собраны и приведены в аналогичный основному набору данных вид. Также были собраны координаты данных объектов, отредактировано расстояние от станции метро Невский проспект и парковки в радиусе 1 километра.

После чего два набора данных были объединены, однако возникла другого рода задача – заполнить пропущенные значения в обеих выборках.

Ведь арендные строки содержат информацию о стоимости аренды в месяц, когда стоимостные – о стоимости квадратного метра объекта.

```
kndf = pd.merge(ordf, ardf, how='outer')
✓ 0.0s

kndf.isna().sum()
✓ 0.0s
```

price_per_square	1293
squares	0
max_floor	98
latitude	0
longitude	0
Nevskii prospekt	0
parkings_1km	0
rent_per_month	801
clas	0

dtype: int64

Рисунок 17 – Пропущенные значения в каждом из столбцов.

На данном выводе можно заметить указанную особенность. Впрочем, такого рода несоответствие не означает наличие проблемы, скорее напротив – смешивание такого набора данных позволит нам расширить нашу выборку почти в два раза и добавить достаточно весомый признак – стоимость аренды помещения в месяц.

Решить проблему заполнения пропусков возможно с помощью использования метода ближайших соседей. Алгоритм, основываясь на значениях остальных параметров рассчитает, сколько бы стоил квадратный метр в данном арендном помещении и, соответственно, сколько бы стоила аренда в данном помещении.

```
from sklearn.impute import KNNImputer

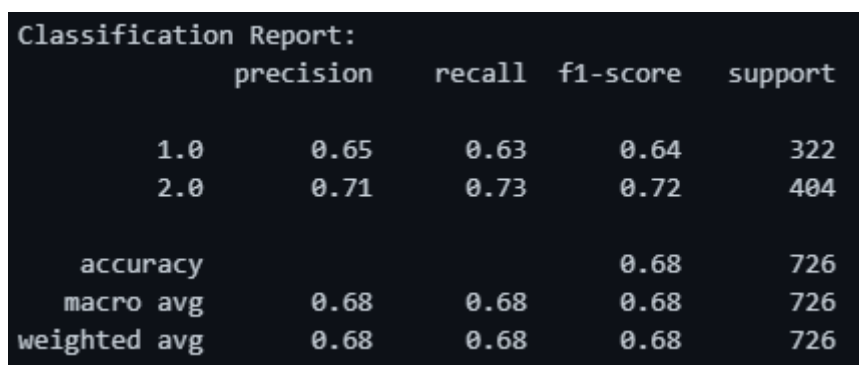
imputer = KNNImputer()
✓ 0.6s

imputer = KNNImputer()
|
filled_data = imputer.fit_transform(kndf)
✓ 0.1s
```



Рисунок 18 – Применение алгоритма метода ближайших соседей для заполнения пропусков в данных.

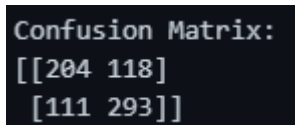
После использования данного алгоритма – очень важно оценить результат – насколько точно значения двух факторов были подобраны. При использовании классификатора ближайших соседей – был получен следующий результат:



Classification Report:				
	precision	recall	f1-score	support
1.0	0.65	0.63	0.64	322
2.0	0.71	0.73	0.72	404
accuracy			0.68	726
macro avg	0.68	0.68	0.68	726
weighted avg	0.68	0.68	0.68	726

Рисунок 19 – Результаты модели классификации.

В том числе матрица ошибок:



Confusion Matrix:	
[[204	118]
[111	293]]

Рисунок 20 – Матрица ошибок.

Основываясь на данных результатах, мы можем заметить, что большая часть прогнозов была предсказана верно, однако общий результат точности классификатора не показывает высоких результатов. Стоит отметить, что это и является ультимативной целью – это может свидетельствовать о том, что данные, заполненные алгоритмом, хорошо вписываются в логику данных и во внутренние закономерности выборки. Поэтому такой результат более чем подходит для решения поставленных задач.

После выявления проверки качества добавленных данных стоит проверить – насколько новые данные улучшают результаты моделей? Учитывая, что лучшие результаты на имеющихся данных показывали модели случайных лесов – стоит обратиться именно к ним для измерения изменений при расширении нашей выборки.

Была проведение кросс-валидация выборки в отношении 40% тестовых данных на 60% тренировочных. Коэффициент корреляции на тренировочных данных показал результат в 0.9116, а значения среднего квадратичного отклонения, среднего абсолютного отклонения и среднего абсолютного процентного отклонения составили 21300.727 , 15720.566 и 8.1 % соответственно. Достаточно хороший результат, хоть и на тренировочных данных. Тем не менее, хоть тестовая точность не показала подобных результатов – мы имеем значительное улучшение в точности модели по сравнению с предыдущими результатами. И если предыдущие результаты на тестовой модели показывали MAPE равную 0.292196, то сейчас мы имеем следующие результаты:

- $R^2$ : 0.31419075965671517
- MSE (СКО): 57228.42007749025
- MAE (CAO): 42625.42939557917
- MAPE (САПО): 0.210919059980457.

То есть качество модели улучшилось на 8 процентных пунктов. Если рассматривать результаты ExtraTreesRegressor – получаем наилучшие результаты на тренировочных данных:

- $R^2$ : 0.9996022455780086
- MSE (СКО): 1429.1161138446105
- MAE (CAO): 66.71441689640166
- MAPE (САПО): 0.00024171700035545643

Однако на тестовых данных результаты практически не изменились по сравнению с результатами обычного случайного леса. В первую очередь это стоит связывать с особенностями ExtraTrees модели – возможно такого количества данных недостаточно для корректного использования данной модели, ведь она попросту переобучается на этих данных и запоминает результат, вместо того чтобы независимо определять стоимость квадратного

метра. Перейдем к рассмотрению важности параметров, насколько сильно результаты модели зависят от определенных факторов:

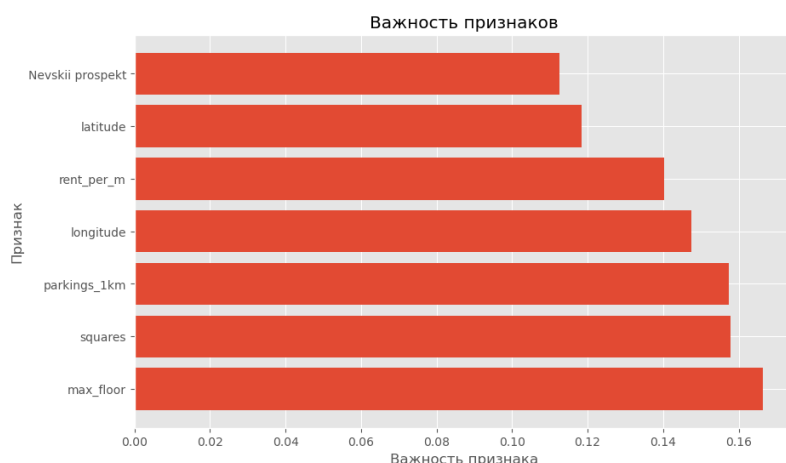


Рисунок 21 – Важность параметров для модели.

Видим, что введенный на данном этапе показатель стоимости аналогичного арендного помещения демонстрирует хорошие показатели влияния на прогноз модели. Рассмотрим Shar-график модели на нестандартизированных данных:

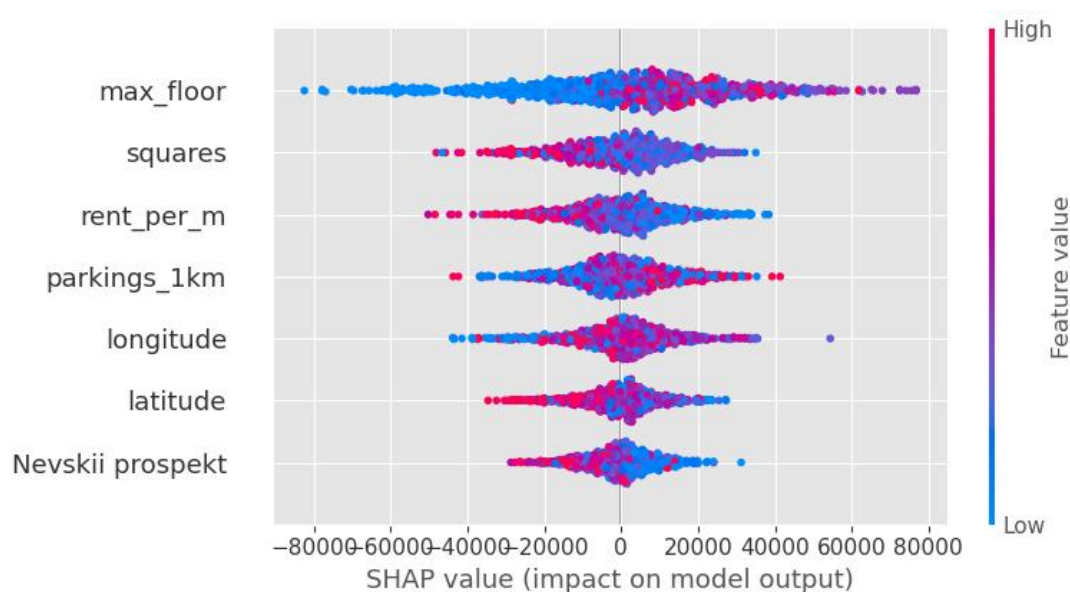


Рисунок 22 – Сравнение влияние параметров для второй модели согласно SHAP.

Выведем средние shar-значения по факторам.

- squares (площадь объекта): -249.2390
- max\_floor (максимальный этаж): 1276.5556
- latitude (широта): -303.7297

- longitude (долгота): 212.5159
- Nevskii prospekt (расстояние в км от Невского проспекта): -404.6131
- parkings\_1km (кол-во парковок в радиусе 1 км): -331.0429
- rent\_per\_m (стоимость аренды в месяц): -200.4467

Основываясь на графике, а также на численных значениях – стоит обратить внимание подробно на интерпретацию, какие значения действительно важны, то есть влиятельны для модели.

Рассмотрим интерпретацию на примере максимального этажа – значения, которые положительно влияют на стоимость квадратного метра – оказывают больше влияние на качество модели. Можем логически предположить, что на данном отрезке подразумеваются здания с высоким максимальным этажом – более современные здания.

После изучения результатов данной модели – было решено рассмотреть использование фиктивных переменных районов, однако добавить параметры, связанные с ними и убрать строки из незначимых для модели районов.

После проделывания указанных процессов, а также исключения незначимых параметров – мы имеем следующие столбцы районов: Admiralteysky, Vasileostrovsky, Vyborgsky, Kalininsky, Kirovsky, Krasnogvardeysky, Krasnoselsky, Moskovsky, Nevsky, Petrogradsky, Primorsky, Pushkinsky, Frunzensky, Tsentralny. По данным районам мы рассчитаем среднюю стоимость квадратного метра объектов и введем как еще один фактор.

Кроме того, было решено использовать информацию о населении Санкт-Петербурга как замена предполагаемому трафику возле объекта, ввиду технических трудностей получения такого рода данных. Использование населения по районам Санкт-Петербурга не дало желаемых результатов. Однако, было решено обратить внимание на распределение данных по парковкам по Санкт-Петербургу.



Рисунок 23 – Распределение парковок в Санкт-Петербурге.

После ознакомления с распределением – было решено учесть отношение населения к количеству парковок, после чего произвести расчет предполагаемого населения в области 1 км от объекта согласно количеству парковок в радиусе 1 км.

Данные факторы показали себя значимыми в рамках связи с целевым параметром, хоть и не показывали критичного увеличения в точности по итогу. При построении моделей – ExtraTrees оказались более предпочтительными в выборе из-за своих результатов.

Тренировочные данные:

- $R^2$ : 0.9999724534927307
- MSE: 379.6192780356386
- MAE: 22.914326854011826
- MAPE: 8.537753381060444e-05

Тестовые данные:

- $R^2$ : 0.39389013772492154
- MSE: 2758436128.179891
- MAE: 40105.37614942529
- MAPE: 0.1884151868585516

Остальные попытки внести новые параметры не показали изменение результата в нужную сторону, поэтому было решено на данный момент отталкиваться от данного набора данных и данной модели.

## ЗАКЛЮЧЕНИЕ

Полученные результаты говорят о важности использования точных и подробных данных для построения такого рода модели. Более того, модель такого рода должна иметь достаточно большую выборку. Однако критически-важным фактом необходимо считать наличие выборки, которая изначально размечена корректно. Ввиду использования данных с открытого рынка – мы имеем несправедливо переоцененные наблюдения, несправедливо недооцененные, ложные или некорректные наблюдения, которые не позволяют установить точные и статистически высоко-значимые связи между основными параметрами. Альтернативным решением в данной ситуации можно считать поиск и учет большего количества наблюдений, тем не менее, в рамках рынка, где достаточно часто встречается некорректная оценка – сам прогноз модели с полученной в результате ошибкой – входит в некие рамки разброса настоящего значения. К тому же, основным регулятором рынка в этом плане, стоит считать сам рынок – он корректирует изначальные оценки под запросы рынка.

Тем не менее, проделанная работа оказала значимое воздействие на возможности применения практических способностей в рамках планирования деятельности, исследования темы и литературы узкой направленности, поисков данных, их обработки, исследования и дальнейшего построения моделей как для изучения рынка, так и для усовершенствования устоявшихся подходов к оценке недвижимости.

В дальнейшем планируется имплементация такого рода модели в среду для практического использования.

## СПИСОК ИСТОЧНИКОВ

1. Астраханцева Ирина Александровна, Смирнова Наталья Владимировна Оценка стоимости коммерческой недвижимости на основе моделей машинного обучения // Научные труды ВЭО России. - 2022. - №237
2. Геокодер // Яндекс URL: [https://yandex.ru/maps-api/products/geocoder-api/?utm\\_source=yandex&utm\\_medium=cpc&utm\\_campaign=fmt\\_textcmp\\_brand\\_searchprod\\_api-mapsgeo\\_ru&utm\\_content=cid\\_99950950\\_gbid=5325032826\\_aid\\_15308233919\\_kid=48232840543\\_adp\\_no\\_pos\\_premium1\\_src\\_search\\_none\\_dvc\\_desktop\\_rtg\\_48232840543&utm\\_term=геокодер%20яндекс%20api&yclid=11291778458353663999](https://yandex.ru/maps-api/products/geocoder-api/?utm_source=yandex&utm_medium=cpc&utm_campaign=fmt_textcmp_brand_searchprod_api-mapsgeo_ru&utm_content=cid_99950950_gbid=5325032826_aid_15308233919_kid=48232840543_adp_no_pos_premium1_src_search_none_dvc_desktop_rtg_48232840543&utm_term=геокодер%20яндекс%20api&yclid=11291778458353663999) (дата обращения: 22.04.2024).
3. Головин Н.А., Азанова И.А., Балабаев В.Д., Гладких К.Д. Теоретические аспекты оценки коммерческой недвижимости // Economy and Business: Theory and Practices. - 2022. - №94
4. Д.С. Костяев, А.В. Доценко ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ ДЛЯ АВТОМАТИЗАЦИИ ОЦЕНКИ ЖИЛОЙ НЕДВИЖИМОСТИ // 2017
5. Захаров А.В., Харламов А.В. Техника геокодирования в построении географически взвешенных регрессионных моделей при массовой оценке в условиях неопределенности и неоднородности исходных данных // Имущественные отношения в РФ. - 2010. - №106
6. Катенев К.А., Лукинов В.Л. Методы автоматического анализа и извлечения данных с веб-страниц с помощью языка программирования Python // 2023
7. Киселева Елизавета Александровна Использование Selenium для парсинга отчетов в формате HTML // Приамурский государственный университет им. Шолом-Алейхема. – 2020
8. Лебеденко В.С., Абалтусова Е.И., Самсонова П.В., Ткаченко А.В. Методы автоматизированной оценке коммерческой недвижимости // Economy and Business: Theory and Practice. - 2022. - №93



9. Медведева Н.И., Бровкин А.В. Оценка рыночной стоимости объектов коммерческой недвижимости с использованием доходного подхода // Экономика: вчера, сегодня, завтра. 2018. Том 8. № 9А. С. 79-91
10. Мухаметзянова Е.В. ПРИМЕНЕНИЕ ПОДХОДОВ К ОЦЕНКЕ СТОИМОСТИ КОММЕРЧЕСКИХ ОБЪЕКТОВ НЕДВИЖИМОГО ИМУЩЕСТВА // Научное измерение. - 2023. - С. 18-25.
11. Носов В.В., Цыпин А.П. ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЦЕНЫ ОДНОКОМНАТНОЙ КВАРТИРЫ МЕТОДОМ ГЕОГРАФИЧЕСКИ ВЗВЕШЕННОЙ РЕГРЕССИИ // 2015
12. О. В. Мезенцева Оценка коммерческой недвижимости. - Екатеринбург: УрФУ, 2011. - 115 с (дата обращения: 28.04.2024).
13. Сурков Ф.А. Нейросетевые методы анализа данных в оценке недвижимости // 2016
14. Beautiful Soup Documentation // beautiful-soup URL: <https://beautiful-soup.readthedocs.io/en/latest/> (дата обращения: 20.04.2024).
15. Dr. Sweta R. Kumar, Swati Bhatt, Hasan Phudinawala Predicting House Price with Deep learning: A comparative study of Machine Learning Models // International Journal for Multidisciplinary Research. – 2023
16. git API // git URL: <https://git-scm.com/docs/api-credentials> (дата обращения: 21.04.2024).
17. QGIS geo-analysis // docs.scipy.org URL: <https://qgis3-10-geoanalysis-un.readthedocs.io/en/latest/> (дата обращения: 21.04.2024).
18. SciPy API // docs.scipy.org URL: <https://docs.scipy.org/doc/scipy/reference/index.html> (дата обращения: 10.05.2024).
19. Unleashing the Potential of Random Forest Regression: A Python Implementation Guide with Hyperparameter Tuning. // DEV URL: [https://dev.to/newbie\\_coder/unleashing-the-potential-of-random-forest-regression-a-python-implementation-guide-with-hyperparameter-tuning-167o](https://dev.to/newbie_coder/unleashing-the-potential-of-random-forest-regression-a-python-implementation-guide-with-hyperparameter-tuning-167o) (дата обращения: 05.05.2024).

## ПРИЛОЖЕНИЕ А

```
driver = webdriver.Chrome()
```

```
data_spec_all = {}
```

```
try:
```

```
    for k in range(655,1349):
```

```
#        Инициализация селениума
```

```
    driver.get(url=hrefs[k])
```

```
    time.sleep(10)
```

```
#        Название товара
```

```
    try:
```

```
        title = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[1]/div/h1').text
```

```
    except Exception as title_exception:
```

```
        title = None
```

```
#        Цены товара
```

```
    try:
```

```
        price =
driver.find_element(By.XPATH, "/html/body/div[2]/div/div[2]/div[3]/div/div[1]/div
[1]/div[3]/div/div[1]/span").text
```

```
    except Exception as price_exception:
```

```
        price = None
```

```
    try:
```

```
        square_price =
driver.find_element(By.XPATH, "/html/body/div[2]/div/div[2]/div[3]/div/div[1]/div
[3]/div/div/div[1]/span[2]").text
```

```
    except Exception as sq_price_exception:
```

```
        square_price = None
```

```
    try:
```

```

        squares =
driver.find_element(By.XPATH,"/html/body/div[2]/div/div[2]/div[2]/div[3]/div[1]/
div[2]/span[2]").text

    except Exception as sq_exception:

        squares = None


    try:

        floor =
driver.find_element(By.XPATH,"/html/body/div[2]/div/div[2]/div[2]/div[3]/div[2]/
div[2]/span[2]").text

    except Exception as floor_exception:

        floor = None


    try:

        free =
driver.find_element(By.XPATH,"/html/body/div[2]/div/div[2]/div[2]/div[3]/div[3]/
div[2]/span[2]").text

    except Exception as free_exception:

        free = None


    try:

        city = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/div/div/a[1]').te
xt

    except Exception as city_exception:

        city = None


    try:

        district = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/div/div/a[2]').te
xt

    except Exception as distr_exception:

        district = None


    try:

```

```

        okrug = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/div/div/a[3]').text

    except Exception as okrug_exception:

        okrug = None

    try:

        street = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/div/div/a[4]').text

    except Exception as street_exception:

        street = None

    try:

        house_number = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/div/div/a[5]').text

    except Exception as house_exception:

        house_number = None

#     Url нποδαφεια

    try:

        seller = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[3]/div/div[3]/div/div/div[1]/div/div[2]/div/div/div
[1]/a').get_attribute('href')

    except Exception as seller_exception:

        seller = None

    try:

        description = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/div[5]/div/div/div/div/span').text

    except Exception as description_exception:

        description = None

#     Mempo

    try:

```

```

sub_near = []

sub_list = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/ul[1]')

# Обработка данных списка, если он найден

for li in sub_list.find_elements(By.TAG_NAME, 'li'):
    sub = li.find_element(By.TAG_NAME, 'a').text
    time_value = li.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/ul[1]/li[1]/span')
    .text

    sub_near.append([sub,time_value])

except Exception as sub_exception:
    sub_near = None

# Итоже

try:
    highway_near = []

    way_list = driver.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/ul[2]')

    # Обработка данных списка, если он найден

    for li in way_list.find_elements(By.TAG_NAME, 'li'):
        way = li.find_element(By.TAG_NAME, 'a').text
        dist_value = li.find_element(By.XPATH,
'/html/body/div[2]/div/div[2]/div[2]/section/div/div/div[2]/address/ul[2]/li[2]/span')
        .text

        highway_near.append([way,dist_value])

    except Exception as highway_exception:
        highway_near = None

# Добавляем данные с url товара

```

```

        new_data = {'url': hrefs[k], 'seller_url': seller, 'title': title, 'squares': squares,
                    'full_price': price,
                    'price_per_square': square_price, 'city': city, 'district': district, 'okrug':
                    okrug, 'street': street, 'house_number': house_number,
                    'description':
                    description, 'sub_near': sub_near, 'highway_near': highway_near, 'floor': floor, 'free': free}

        cian = cian.append(new_data, ignore_index=True)

        time.sleep(5)
except Exception as ex:
        print(ex)
finally:
        driver.close()
        driver.quit()

```