



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

ОТЧЕТ
по Производственной практике
(по получению профессиональных умений
и опыта профессиональной деятельности)

Наименование организации прохождения практической подготовки:

ООО «Актив-Инвест»
(наименование организации)

Направление **38.03.01 «Экономика»**
(цифр, 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
Полякова Светлана Петровна, 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



(подпись)

20__ г.

Фридман Григорий Морицович

(подпись)

**Индивидуальное задание
на производственную практику
(по получению профессиональных умений
и опыта профессиональной деятельности)**

Обучающегося 4 курса Дмитриева Александра Ростиславовича
(курс обучения) (Ф.И.О. полностью)

Направление/специальность: 38.03.01 «Экономика»

Направленность: Математическое моделирование и анализ данных в
экономике

Тема ВКР/НКР: _____

(заполняется в случае прохождения преддипломной практики)

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

Сроки практической подготовки с 04.04.2024 по 17.04.2024

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	Ознакомление с правилами внутреннего распорядка на предприятии, прохождение инструктажа по технике безопасности и охране труда.	04.04.2024
2	Ознакомление с аналитическими задачами, решаемыми подразделением. Согласование с руководителем практики от предприятия индивидуального задания на практику	04.04.2024
3	Сбор фактического материала по рынку недвижимости.	05.04.2024
4	Обработка собранного материала.	06-13.04.2024
5	Анализ имеющихся данных.	13-15.04.2024
6	Сделать выводы о собранных данных и их возможном влиянии на результаты машинного обучения при оценке коммерческого объекта недвижимости. Описать выявленные проблемы.	15-16.04.2024
7	Обобщение материалов и подготовка отчёта по результатам практики	17.04.2024

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 (ПО ПОЛУЧЕНИЮ ПРОФЕССИОНАЛЬНЫХ УМЕНИЙ И ОПЫТА ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ)

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 04.04.2024 по 17.04.2024

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 фактического материала, его обработке, анализу, а также сделал выводы о собранных данных, об их потенциальном качестве в рамках оценки коммерческой недвижимости с помощью различных методов машинного обучения.

Личные и деловые качества (компетенции)

В период прохождения практики, Дмитриев Александр Ростиславович показал умение работать в системе с регулярной отчетностью о проделанной работе, продемонстрировал умение находить нестандартные решения задач, в добавок к дисциплинированному и ответственному подходу к деятельности, хорошим навыком коммуникации и другим аспектам работы как с коллективом, так и вне.

Качество отчета, оценка

Отчет по проделанной работе выполнен в срок, в полном объеме, в соответствии с изложенными стандартами и с учетом всех требований. Рекомендуемая оценка – «Отлично».

Рекомендации

Рекомендовано уделить внимание поиску дополнительных данных для улучшения потенциальных результатов при оценке коммерческой недвижимости с помощью машинного обучения.

Подпись: _____

Должность: Директор ООО «АКТИВ-ИНВЕСТ»

ФИО: Елисеева Елена Сергеевна



Тел. +7 (861) 279-84-00

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	6
1. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ И АНАЛИЗ ФИНАНСОВОЙ ОТЧЕТНОСТИ ООО «АКТИВ-ИНВЕСТ».....	8
1.1. Характеристики и основная деятельность ООО «Актив-Инвест»	8
1.2. Анализ финансовой отчетности ООО «Актив-Инвест»	8
2. ЛИЧНЫЙ ВКЛАД В ДЕЯТЕЛЬНОСТЬ КОМПАНИИ	11
2.1. Постановка задачи и определение источника данных	11
2.2. Сбор данных.....	12
2.3. Обработка данных	16
2.4. Работа с координатными данными	18
2.5. Географический анализ	20
2.6. Анализ значимости параметров	22
ЗАКЛЮЧЕНИЕ.....	26
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	27

ВВЕДЕНИЕ

Производственная практика была пройдена в компании ООО «Актив-Инвест», занимающейся оценкой недвижимости. Ранее был выявлен потенциал в применении машинного обучения в рамках модернизации процессов деятельности компании. Выбор системы подхода к применению машинного обучения первостепенно оказывает влияние на точность оценки объекта. Точность оценки имеет большую роль для всей деятельности компании, поэтому первостепенно важно достижение наилучшего результата.

После изучения структуры корпоративных процессов, первоочередной задачей стал сбор данных, их анализ и дальнейшее моделирование.

Основополагающими целями практики стоит считать анализ, оценка и прогнозирование стоимости коммерческой торговой недвижимости. Это в том числе подразумевает под собой применение на практике методик сбора и анализа данных, работы с экономической интерпретацией, а также применение статистических и аналитических инструментов для работы над созданием надежных моделей прогнозирования.

Практика направлена на выполнение следующих задач, согласованных в индивидуальном задании:

1. Ознакомление с правилами внутреннего распорядка на предприятии, прохождение инструктажа по технике безопасности и охране труда.
2. Ознакомление с аналитическими задачами, решаемыми подразделением. Согласование с руководителем практики от предприятия индивидуального задания на практику.
3. Сбор фактического материала по рынку недвижимости.
4. Обработка собранного материала.

5. Анализ имеющихся данных.
6. Сделать выводы о собранных данных и их возможном влиянии на результаты машинного обучения при оценке коммерческого объекта недвижимости. Описать выявленные проблемы.
7. Обобщение материалов и подготовка отчёта по результатам практики

Объектом практики же является ООО «Актив-Инвест». Предметом практики стоит понимать процессы анализа и модернизации оценки рынка коммерческой торговой недвижимости, используя машинное обучение.

В ходе практики основное внимание уделяется изучению объектов торговой недвижимости, таким как помещения свободного типа, будущие магазины и помещения в коммерческих комплексах. Работа включает анализ различных параметров этих объектов, таких как координатное местоположение, характеристика здания и объекта, удаленность от определенных точек, размеры, текущее состояние рынка, что в итоге позволяет формировать базу данных для моделей оценки стоимости.

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

1. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ И АНАЛИЗ ФИНАНСОВОЙ ОТЧЕТНОСТИ ООО «АКТИВ-ИНВЕСТ»

1.1. Характеристики и основная деятельность ООО «Актив-Инвест»

ООО «Актив-Инвест», является одной из крупнейших компаний Краснодарского края, которая занимается предоставлением посреднических услуг при оценке недвижимого имущества (код по ОКВЭД2 68.31.5). Ежегодно, база объектов предприятия расширяется, и для регулярного проведения оценки недвижимости – необходимо учитывать своевременную информацию, запросы рынка, его особенности. В особенности очень важно участвовать в динамичном развитии методов оценки недвижимости, и в общем развитии рынка.

Помимо этого, численность ООО «Актив-Инвест» насчитывает 24 человека и числится в реестре малых предприятий. Также предприятие подвержено большому потоку клиентов, поступающему, как и из агентств недвижимости, с которыми работает ООО «Актив-Инвест», так и от физических лиц напрямую.

Что указывает на постоянное большое количество работы, связанной с оценкой и ограниченное имеющимися ресурсами. В связи с чем отделу оценки недвижимости и другого имущества была поставлена задача по применению более современного, автоматизированного и научного подхода к оценке.

Подробное описание проводимого исследования на тему оценки эффективности применения данного подхода приведено в следующей главе.

1.2. Анализ финансовой отчетности ООО «Актив-Инвест»

Для полноценного анализа деятельности ООО «Актив-Инвест», необходимо обратиться к финансовой отчетности компании и провести ее

анализ, даст более наглядное представление о результатах деятельности. Рассматриваемый период – два года. Но для лучшего понимания отчета, стоит учесть экономическую среду, в которой компания осуществляла свою деятельность за последние года.

По итогам 2023 года объем инвестиций в коммерческую недвижимость России достиг рекордных значений, в 1,5 раза больше по сравнению с 2022 годом. При этом инвестиции в складскую, торговую и офисную недвижимость выросли более чем в два раза. И похожий рост прогнозируют на ближайшие годы.

Таблица 1 – Горизонтальный и вертикальный анализ бухгалтерского баланса ООО «АКТИВ-ИНВЕСТ».

Наименование показателя	2023 г.		2022 г.		Изменение, ±	Темп роста, %
	Тыс. руб.	уд.вес, %	Тыс. руб.	уд.вес, %		
Денежные средства и денежные эквиваленты	1 725	15,90	8 291	47,95	-6 566	-79,19%
Финансовые и другие оборотные активы (включая дебиторскую задолженность)	9 121	84,10	9 000	52,05	121	1,34%
БАЛАНС	10 846	100	17 291	100	-6 445	-37,27%
Капитал и резервы	9 695	89,39	15 760	91,15	-6 065	-38,48%
Краткосрочные заемные средства	-	-	-	-	-	-
Кредиторская задолженность	1 151	10,61	1 531	8,85	-380	-24,82%
БАЛАНС	10 846	100	17 291	100	-6 445	-37,27%

Стоит изучить доступную информацию по активам, что количество, а также удельный вес денежных средств и их эквивалентов уменьшился на 6 566 тыс. руб. или на 79,19%. Касательно финансовых и других оборотных активов – изменение положительное незначительное на 121 тыс. руб. ~1,34%.

В пассивах заметно значительное уменьшение капитала и резервов на 6 065 тыс. руб., или на 38,48 %. Существует кредиторская задолженность, которая уменьшилась на 37,27%, то есть на 6 445 тыс. руб.

В целом, ввиду отсутствия более подробных исторических данных – трудно сказать о характере данных изменений, тем не менее, учитывая структуру компании – можно предполагать, что кредиторская задолженность, а также часть изменений в активах обусловлена взаимодействием с системой дочерних и материнских предприятий.

Таблица 2 – Финансовые результаты ООО «АКТИВ-ИНВЕСТ».

Наименование показателя	2023	2022	Изменение, тыс. руб.	Темп роста, %
Выручка	48 148	47 475	673	1,42%
Расходы по обычной деятельности	-16 114	-18 686	2 572	-13,76%
Прочие доходы	3 713	989	2 724	275,43%
Прочие расходы	-919	-1 698	779	-45,88%
Налоги на прибыль(доходы)	-1 560	-1 440	-120	8,33%
Чистая прибыль(убыток)	33 268	26 640	6 628	24,88%

Изучая финансовые результаты, становится заметно значительное увеличение чистой прибыли, в основном за счет значительного снижения расходов по обычной деятельности (на 13,76%), а также серьезного снижения прочих расходов (на 45,88%). К тому же большое влияние оказало увеличено доходов, полученных не по основному роду деятельности – 2 724 тыс. руб. и 275,43%. Доходы за основной род деятельности при этом не показал значительного роста, всего 1,42%.

Трудно объяснить все закономерности ввиду отсутствия исторических и подробных данных об операциях, однако фактические результаты показывают эффекты качественного менеджмента доходов и расходов.

2. ЛИЧНЫЙ ВКЛАД В ДЕЯТЕЛЬНОСТЬ КОМПАНИИ

2.1. Постановка задачи и определение источника данных

Для анализа рынка и построения в дальнейшем модели – необходимо собрать основную базу данных недвижимости. В организации, как правило, используется собственная база данных, однако в рамках конфиденциальности собственности компании возникает необходимость использовать данные с открытых источников, тем не менее, в самой компании регулярно практикуется сбор и использование открытых источников, в особенности для обновления базы данных и понимания функционирования рынка. 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.

На нынешнем рынке есть несколько крупных открытых баз недвижимости. В первую очередь это такие сайты как «Авито», «Циан», «Яндекс Недвижимость». Однако есть и другие, достаточно подробные информационные источники, однако они принадлежат определенным агентствам недвижимости, то есть подразумевают узконаправленность имеющихся данных и не предоставляют необходимый для данного рода работы объем и спектр объектов.

Из указанных источников необходимо выбрать один основной, соответствующий основным параметрам:

- Полнота данных – наличие максимума информации об объекте;
- Целостность и корректность данных;
- Отсутствие ложных данных;
- Наличие критически важных параметров: адрес, квадратура, стоимость.

Стоит рассмотреть конкретные примеры объектов на данных площадках. В таком случае будет возможность изучить возможный набор данных, который можно будет собрать с каждого объекта.

На данном рисунке видно пример объекта с сайта «Яндекс Недвижимость». Если изучить код страницы, а также визуальную составляющую – со страницы объявления можно извлечь следующую информацию:

С учетом проведенного анализа, включая изучение корректности данных – было решено, что основным источником для данных стоит использовать Циан. Яндекс и Авито не так подробно изучают объявления и оставляют большое количество ложных объявлений, которые размещают некоторые агентства недвижимости для привлечения клиентов.

2.2. 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=28&office_type%5B%5D=4&office_type%5B%5D=5&office_type%5B%5D=7&office_type%5B%5D=9&office_type%5B%5D=12&p={page}&region=2&sort=creation_date_desc')
        time.sleep(10)
        print('main base loaded')
        # Постепенная прогрузка 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')
        # Валидация href с каждой страницы
        for i in range(1, 35):
            try:
                item = driver.find_element(By.XPATH, f'/html/body/div[1]/div/div[5]/div[1]/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 определенным нами заранее условиями фильтрации. В цикле предусмотрено использование скрипта для вертикального перемещения по странице, с целью загрузки и обработки всех объявления, взятие href-ссылок объектов, то есть ugl-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.

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.

При изучении написанного парсера, необходимо обратить внимание на структуру внутри цикла. Для каждого из объектов используются операторы “try”, в случае если на какой-либо из страниц не окажется данного объекта или возникнет проблема с его интерпретацией библиотекой selenium. Такая система используется для всех объектов.

Однако ввиду особой структуры объектов, указывающих информацию о расположенных вблизи метро и выездов на шоссе, необходимо было сделать некоторые изменения, а именно:

```
# Метро
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 метро, 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
```

Рисунок 2 – Обработка списков метро в рамках парсера.

Поиск элемента, в котором содержатся данные о станциях метро (шоссе), запуск цикла для каждого из элементов в так называемом «списке», затем вывод названия станции метро в отдельную переменную и затем поиск элемента, указывающего на время (км) до метро (шоссе), после чего такой элемент добавляется в список станций метро (выездов на шоссе) у этого объекта.

В случае если у объекта не указаны станции метро (шоссе) – значения столбца остаются пустыми.

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.

2.3. Обработка данных

Обработка первоначальных, так называемых «грязных» данных, является одной из важнейших частей Data-деятельности. 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 необходимо определить столбцы, которые могут быть полезны для нас и нести смысл, который позволит или более подробно изучить объект или окажется значимым фактором для построения модели.

Одним из важных факторов является площадь самого объекта. Такие данные нам нужны в числовом формате, а именно – float, ведь площадь может быть не целым числом. Для обработки данного столбца используем функции `replace` и `strip` и дополнительно библиотеку `re` для очищения символов единиц измерения площади и любых других символов возле. Используя указанные функции очистим столбец от лишних пробелов и заменим запятые на точки, после чего сможем перевести столбец из `str`-значения во `float`. Первоначальная обработка с площадью закончена. Похожие действия проведем и со столбцами полной цены и цены

квадратного метра, только вместо единиц измерения площади будут символы рубля.

Обращаясь к следующему столбцу «free» - видим, что свободен ли объект описывается словами – заменяем 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 совсем точные данные касательно занятости объекта, поэтому предварительно были помечены как занятые.

Следующей задачей стояло обработка значений этажа. 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 – напишем функцию для вывода этих данных и создадим два отдельных столбца для них.

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.

При попытках обработки столбцов с данными по шоссе возникли проблемы вопроса интерпретации возможного формата обработанных данных. Представленные в виде текста станции метро не позволяли бы учесть их как фиктивные переменные, к тому же был важен фактор дистанции до метро, а циан обозначал его в минутах и мог указывать это время либо как на транспорте, либо пешим ходом, что добавляло проблемы в интерпретации. Такие же проблемы возникли и с шоссе, однако ввиду

большого количества уникальных выездов на шоссе – решение такого рода проблем было найдено только для метро.

Учитывая работу с координатами, было решено собрать координаты станций Санкт-Петербургского Метрополитена и затем рассчитать расстояние до объектов недвижимости, а также количество станций метро в радиусе нескольких километров.

2.4. 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"][0]["GeoObjectCollection"]["metaDataProperty"]["GeocoderResponseMetaData"]["found"])
        if found > 0:
            coords_str = data["response"][0]["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 3 – 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.

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 ~ 111.2 км;

φ_1, φ_2 — географические широты пунктов 1 и 2;

λ_1, λ_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.

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

Для выполнения этой задачи был также использован платный набор геопространственных данных по Санкт-Петербургу, разделяющий город на существующие 18 районов. Из имеющегося набора полигонов в рамках системы координат – были собраны названия районов в отдельный список. После чего были импортированы библиотека `shapely`, позволяющая работать с полигональными и координатными данными. С помощью метода `loads` данные были конвертированы в формат полигональных данных и затем, была написана функция для проверки, в какой из районов входит объект и последующего обновления столбцов соответствующих районов.

После этого также убираем столбец с центральным районом, для корректного использования фиктивных переменных.

2.5. 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.

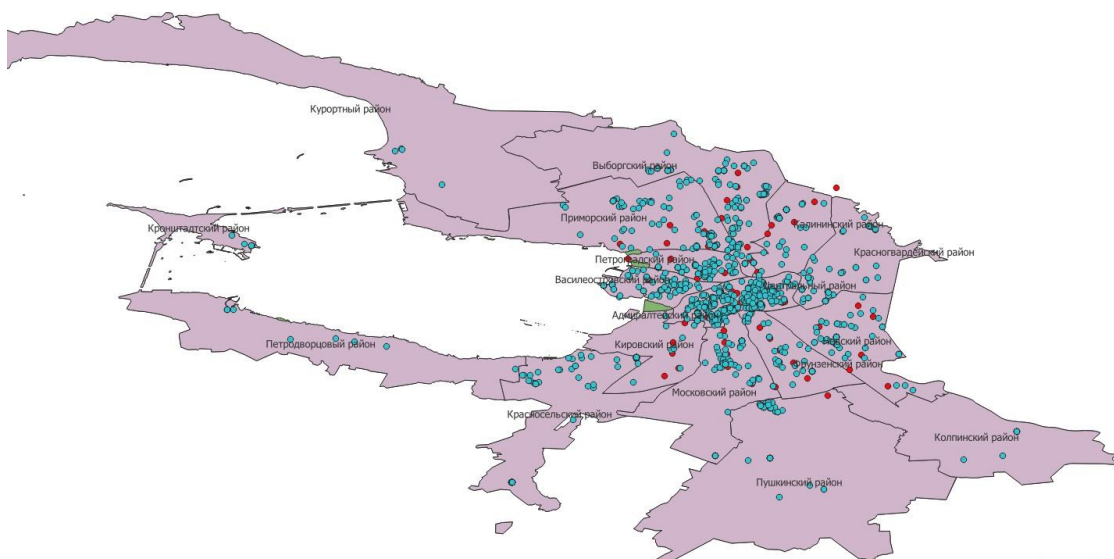


Рисунок 4 – Визуализация объектов в QGIS.

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.

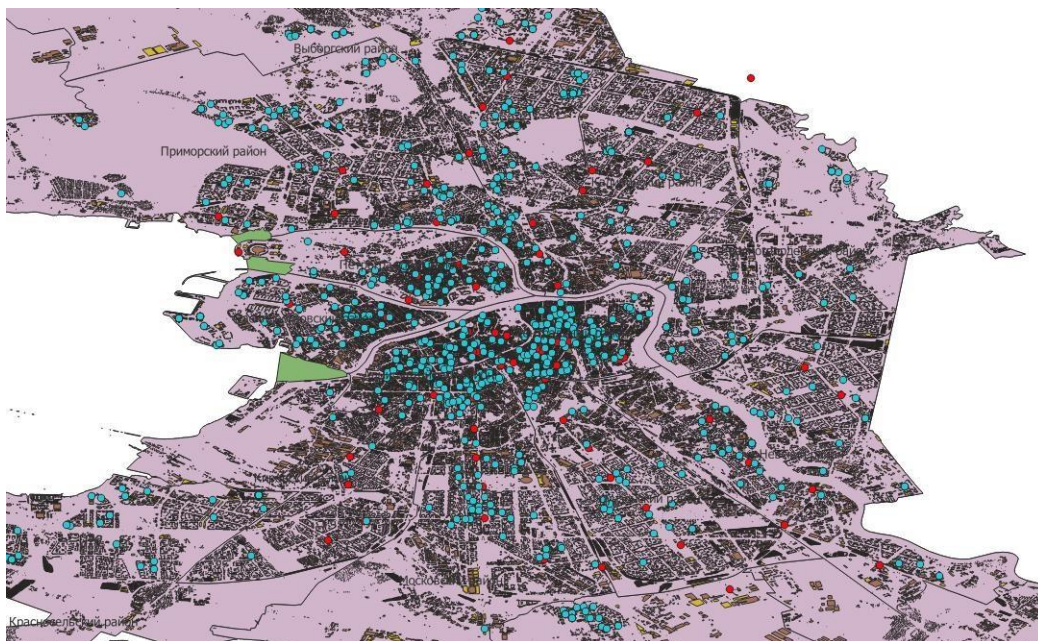


Рисунок 5 – Приближенная визуализация с полигонами зданий и парковок.

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.

При этом стоит отметить, что хоть QGIS критически важен в плане визуального и общего анализа и, соответственно, понимания данных – работа с таким ресурсом требует большого количества часов для изучения и использования на практике всех возможных методов, которые позволяют работать с данными, поэтому к более подробному изучению данной системы было решено вернуться позднее.

2.6. Анализ значимости параметров

Теперь же возникает немного другого рода проблема – чрезмерное количество факторов в нашем наборе данных, которые при этом создают большую межфакторную корреляцию – дистанции до станций метро. Выведем корреляционную матрицу частей столбцов нашего Dataframe.

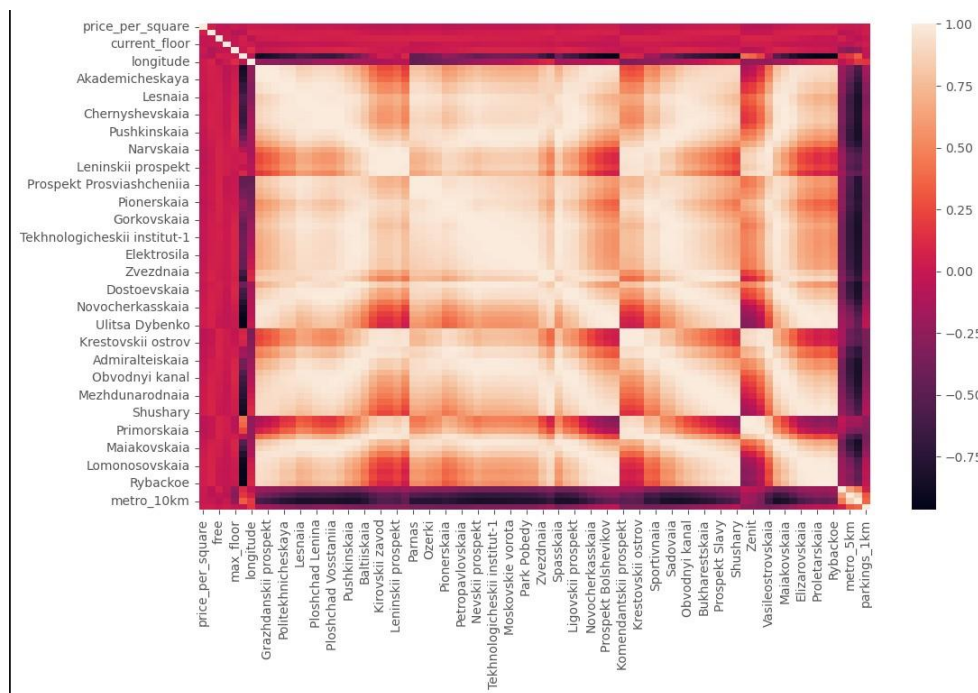


Рисунок 6 – Корреляционная матрица части параметров.

Было решено избавиться от большей части признаков на данном этапе модели и в случае необходимости – все же учесть некоторые станции. Также была изучена столбчатая диаграмма, показывающая корреляционную связь с целевым признаком:

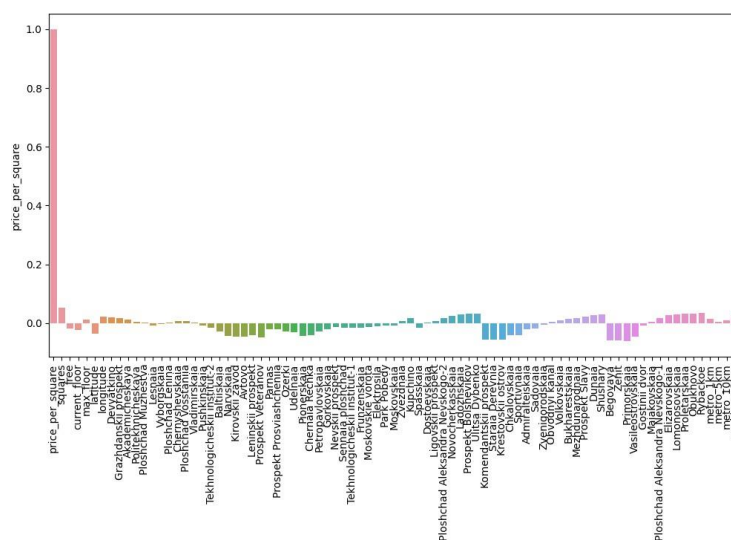


Рисунок 7 – Диаграмма корреляции параметров по метро.

По ней, во многом, можно сделать заключение о большой роли отрицательного влияния конечных станций на стоимость квадратного

метра. Было испробовано несколько вариантов включения определенного количества факторов станций, в особенности использование только конечных станций метрополитена и Невского проспекта, однако во всех вариантах они вызывают корреляцию между собой и не оказывают значимого влияния на все параметры прогнозирования целевого параметра.

Поэтому было решено на данный момент оставить только станцию метро - Невский проспект, ведь большинство объектов расположены в центре или вокруг центра, а данная станция во многом и представляет собой центр города, что позволит нам также в дальнейшем интерпретировать данный параметр как влияние расстояния от центра Санкт-Петербурга.

Рассмотрим наличие влияния фиктивных параметров между собой и с целевым параметром, а также проверим значимость параметров с учетом также и основных факторов в рамках поставленной задачи.

Для этого мы напишем код, создав отдельный объект – dataframe, в котором с помощью библиотеки `scipy` используем функцию `pearsonr`, позволяющую нам вывести p-value для всех используемых факторов:

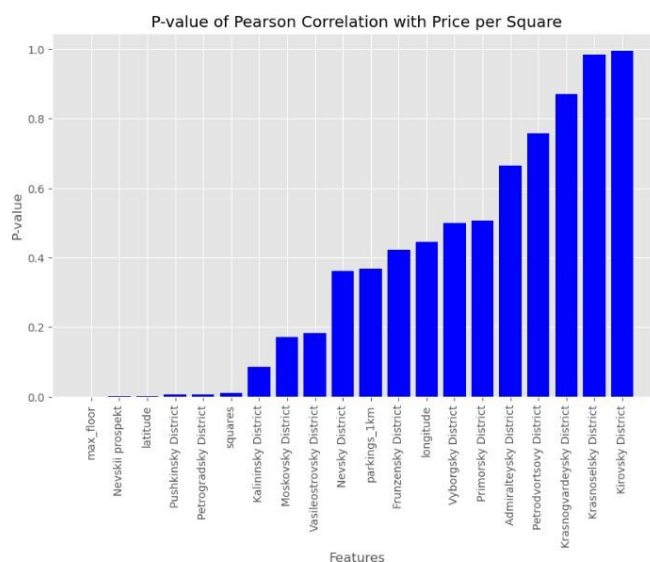


Рисунок 8 – Диаграмма значимости параметров и фиктивных переменных по районам.

Анализируя выявленные результаты, можно в целом утверждать, что никакой статистически различимой значимости параметров – фиктивных переменных районов Санкт-Петербурга – не может быть в рамках данной задачи. Возможно, причина лежит в необходимости большего объема равномерных данных – больше помещений в каждом из районов, для обнаружения серьезной зависимости в рамках разных районов.

Учитывая имеющиеся результаты проведенного анализа данных, стоит учесть только самые необходимые параметры, которые показывают наличие статистического смысла в виде зависимости с целевым показателем и наличие логического смысла в рамках данного бизнеса. На данный момент таковыми являются:

- price_per_square (Целевой признак, стоимость кв. метра);
- squares (Площадь объекта);
- max_floor (Максимальный этаж здания);
- longitude (Долгота);
- latitude (Широта);
- Nevskii prospekt (Удаленность от центра);
- parkings_1km (Количество парковок в радиусе 1 км).

ЗАКЛЮЧЕНИЕ

Полученные данные предварительно говорят о недостаточности используемых данных в рамках решения вопроса создания модели прогнозирования. Решение данной проблемы предполагает поиск, и использование более критически важных данных, осуществление более трудоемкой работы с их обработкой и подразумевает использование в дальнейшем новейших методов машинного обучения, способных нивелировать недостатки выборки и улучшить результаты прогнозов.

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 ООО «АКТИВ-ИНВЕСТ».

В дальнейшем планируется критическое увеличение данных, то есть добавление новых параметров, реализация модели прогнозирования и создание функционирующей среды для возможности практического использования.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

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 (дата обращения: 02.04.2024).
3. О. В. Мезенцева Оценка коммерческой недвижимости. - Екатеринбург: УрФУ, 2011. - 115 с (дата обращения: 28.03.2024).
4. Beautiful Soup Documentation // beautiful-soup URL: <https://beautiful-soup.readthedocs.io/en/latest/> (дата обращения: 20.03.2024).
5. 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
6. git API // git URL: <https://git-scm.com/docs/api-credentials> (дата обращения: 01.04.2024).
7. QGIS geo-analysis // docs.scipy.org URL: <https://qgis3-10-geoanalysis-un.readthedocs.io/en/latest/> (дата обращения: 13.04.2024).
8. SciPy API // docs.scipy.org URL: <https://docs.scipy.org/doc/scipy/reference/index.html> (дата обращения: 10.04.2024).
9. 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
(дата обращения: 05.04.2024).

10. WebDriver API // selenium-python URL: <https://selenium-python.readthedocs.io/api.html> (дата обращения: 20.03.2024).