МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

РЕГРЕССИОННЫЙ АНАЛИЗ РЫНКА ЖИЛЬЯ В ГОРОДЕ САРАТОВЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Алейникова Владислава Александровича

Научный руководитель
доцент, к.ф.-м.н., доцент _____ В. В. Новиков

Заведующий кафедрой
д.ф.-м.н., доцент _____ С. П. Сидоров

Саратов 2023

ВВЕДЕНИЕ

В современной экономике рыночные процессы играют важную роль в развитии отдельных компаний, секторов экономики и всей национальной экономики в целом. Одним из ключевых инструментов анализа рыночных процессов является регрессионный анализ, который позволяет определить взаимосвязь между различными переменными и прогнозировать их будущее поведение.

Актуальность темы. Одним из наиболее значимых секторов экономики является рынок жилья, который влияет на многие аспекты жизни общества, такие как уровень жизни, социальную мобильность и экономическую стабильность. В связи с непрерывным ростом числа населения и строительстве новой жилой недвижимости, в городах возникает проблема выгодной покупки и продажи недвижимости. Среди большого количества предложений становится сложно анализировать и выбирать лучшие варианты на рынке. В решении этой проблемы может помочь регрессионный анализ рынка жилья, что делает эту тему очень актуальной в современных реалиях.

Целью бакалаврской работы является изучение возможностей приложения регрессионного анализа к рынку жилой недвижимости, а также реализация программы, осуществляющей регрессионный анализ рынка жилья в городе Саратове.

Объект исследования – рынок жилой недвижимости города Саратова.

Предмет исследования – построение модели линейной регрессии для анализа рынка жилья и прогнозирования цен жилой недвижимости.

Для достижения поставленной цели необходимо выполнить следующие **задачи**:

1. Рассмотреть теоретические аспекты исследования рынка жилой недвижимости и регрессионного анализа.
2. Выбрать источник данных для исследования.
3. Реализовать программу для сбора данных.
4. Рассмотреть собранные данные и провести разведочный анализ.
5. Реализовать программу, осуществляющую регрессионный анализ рын-

ка жилья в городе Саратове.

6. Проанализировать полученные результаты.

Структура и содержание бакалаврской работы. Работа состоит из введения, двух разделов, заключения, списка использованных источников, содержащего 29 наименований, и двух приложений, содержащих код программ. Общий объем работы составляет 60 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В **первом** разделе приводится необходимая теория, которая будет использована в данной работе.

Рынок жилой недвижимости. Основная информация. Рынок жилой недвижимости представляет собой сектор экономики, где осуществляется купля-продажа и аренда жилых объектов, таких как квартиры, дома, дачи и другие жилые помещения. Этот рынок является важной частью экономической инфраструктуры каждого общества, поскольку обеспечивает людей жильем, создает рабочие места в строительстве и связанных отраслях, а также имеет существенное влияние на экономическую активность.

Необходимость прогнозирования и оценки цены жилья. Необходимость проведения оценки стоимости недвижимого имущества объясняется ее важностью в сферах экономики общества, ее ролью в обеспечении материального благополучия и качества жизни общества. Рынок жилья постоянно меняется под влиянием различных факторов. Оценка и прогнозирование стоимости жилья играют важную роль при финансовых решениях, инвестициях, обеспечении стабильности и государственном регулировании. Они способствуют эффективному функционированию рынка жилой недвижимости и устойчивому развитию этого сектора экономики.

Влияние характеристик жилья на формирование цены. При покупке или продаже жилья на рынке недвижимости множество факторов влияют на ее стоимость. Характеристики самой квартиры играют ключевую роль в формировании цены. Например, общая площадь квартиры существенно влияет на ее цену. Квартиры с большей площадью обычно имеют более высокую стоимость. Это связано с тем, что большая площадь предоставля-

ет больше места для проживания, а также дает больше возможностей для размещения мебели и создания комфортной обстановки.

Регрессионный анализ в прогнозировании цены жилья. Регрессионный анализ является одним из основных инструментов, используемых в экономике и финансовой аналитике для прогнозирования стоимости жилой недвижимости.

В случае прогнозирования стоимости жилья, зависимая переменная представляет собой цену квартиры или дома, а независимые переменные включают различные характеристики недвижимости, такие как площадь, количество комнат и другие факторы, которые могут влиять на ее стоимость.

Математические основы регрессионного анализа. Рассмотрим многомерную регрессионную модель (multiple regression model), или модель множественной регрессии:

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, n$$

или

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots n, \tag{1}$$

где x_{tp} - значения регрессора x_p в наблюдении t, а $x_{t1} = 1, t = 1, \ldots, n$. С учетом этого замечания не будем далее различать модели вида (1) со свободным членом или без свободного члена.

Гипотезы, лежащие в основе модели множественной регрессии, являются естественным обобщением модели парной регрессии:

1. $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_{t_1} t = 1, \ldots, n-$ спецификация модели.
2. $x_{t1}, \ldots, x_{tk}-$ детерминированные величины. Векторы $x_s = (x_{1s}, \ldots, x_{ns})'$, $s = 1, \ldots, k$ линейно независимы в R^n.

3a. $\mathrm{E}\varepsilon_t = 0, \mathrm{E}\left(\varepsilon_t^2\right) = \mathrm{V}\left(\varepsilon_t\right) = \sigma^2-$ не зависит от t.

3b. $\mathrm{E}\left(\varepsilon_t \varepsilon_s\right) = 0$ при $t \neq s-$ статистическая независимость (некоррелированность) ошибок для разных наблюдений. Часто добавляется следующее условие.

3c. Ошибки $\varepsilon_t, t = 1, \ldots, n$ имеют совместное нормальное распределение: $\varepsilon_t \sim N\left(0, \sigma^2\right)$.

В этом случае модель называется нормальной линейной регрессионной

(classical normal linear regression model).

Гипотезы, лежащие в основе множественной регрессии, удобно записать в матричной форме, которая главным образом и будет использоваться в дальнейшем. Пустъ y обозначает $n \times 1$ матрицу (вектор-столбец) $(y_1, \ldots, y_n)'$, $\beta = (\beta_1, \ldots, \beta_k)' - k \times 1$ вектор коэффициентов; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' - n \times 1$ вектор ошибок;

$X = \begin{bmatrix} x_{11} & \ldots & x_{1k} \\ \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{nk} \end{bmatrix} - n \times k$ матрицу объясняющих переменных.

Столбцами матрицы X являются $n \times 1$ векторы регрессоров $x_s = (x_{1s}, \ldots, x_{ns})', s = 1, \ldots, k$. Условия $1 - 3$ в матричной записи выглядят следующим образом:

1. $y = X\beta + \varepsilon$ — спецификация модели;
2. X — детерминированная матрица, имеет максимальный ранг k;

3a,b. $\mathrm{E}(\varepsilon) = 0; \mathrm{V}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon') = \sigma^2 I_n$

дополнительное условие:

3c. $\varepsilon \sim N\left(0, \sigma^2 I_n\right)$, т. е. ε — нормально распределенный случайный вектор со средним 0 и матрицей ковариаций $\sigma^2 I_n$ (нормальная линейная регрессионная модель).

Во **втором** разделе будет описан процесс регрессионного анализа рынка жилья в городе Саратове, с постановкой задачи, сбором необходимой информации, описанием работы программы и выводом.

Поставленная задача звучит следующим образом: необходимо собрать данные по рынку жилой недвижимости (в частности – по квартирам) города Саратов, и на основании этих данных при помощи регрессионного анализа отобрать наилучшие варианты по соотношению цены к качеству. Наилучшие варианты будут выбираться исходя из предсказанных цен на квартиры. Будет вычисляться так называемый «коэффициент недооценённости» жилья, который представляет собой отношение предсказанной цены к реальной цене жилья для каждой квартиры из собранных данных.

Сбор данных. Источником данных для исследования рынка квартир послужила популярная интернет-площадка для размещения объявлений – сайт «Avito».

Для того чтобы собрать большое количество данных с этого сайта необходимо было реализовать программу для парсинга. Парсер — это программа для сбора и систематизации информации, размещенной на различных сайтах. Парсер работает следующим образом: он анализирует страницу на наличие контента, соответствующего заранее заданным параметрам, а потом извлекает его, превратив в систематизированные данные.

Программу для парсинга было решено реализовать при помощи языка программирования C# (C Sharp).

Полный код программы для парсинга приведен в придожении А бакалаврской работы.

Для начала были подключены необходимые пространства имен для использования классов, которые понадобятся при парсинге сайта. Выбранные в работе пространства имен позволяют осуществлять: поддержку глобализации и локализации в приложении, взаимодействие с серверами, осуществление работы с сетевыми ресурсами, работу с различными кодировками текста, считывание HTML страниц и работу с CSV файлами.

После подключения необходимых пространств имен следует создание нового пустого списка объектов User, создание потоков и объектов для записи данных в файл CSV и экземпляра HtmlParser, который будет использоваться для разбора HTML-кода. Так же был создан объект HttpClient для отправки HTTP-запросов и добавление заголовка User-Agent для имитации работы веббраузера.

Далее была прописана запись заголовков столбцов в файл CSV. В объявлениях квартир были выбраны следующие поля с характеристиками:

— Количество комнат (room_no)
— Общая площадь квартиры (size_t)
— Этаж (floor)
— Количество этажей в доме (floors)
— Цена (price)
— Цена за квадратный метр (price_sqm)
— Площадь кухни (size_k)
— Жилая площадь квартиры (size_l)
— Наличие и тип балкона (balc)

— Тип ремонта в квартире (decoration)
— Тип санузла (bath)
— Адрес (adress)

Был создан класс User с набором свойств, которые представляют различные данные, извлекаемые из HTML-кода и сохраняемые в CSV-файл. Это данные каждой из выбранных характеристик квартиры.

Сам сбор данных со страницы с объявлением происходит в работе цикла, который проходит все страницы и собирает информацию по характеристикам каждой из квартир из заданных полей.

Отправляется GET-запрос на сайт с объявлениями. Полученный HTML-код считывается и с помощью HtmlParser выполняется разбор HTML-кода и создается структурированное представление документа. Из считанного HTML-кода извлекаются необходимые элемент представляющую информацию о характеристиках квартиры. Создается новый объект User с соответствующими считанными элементами, и значения этих элементов, содержащие данные о конкретной характеристике квартир, записываются в CSV-файл.

В результате работы программы было получено 2000 объявлений о продаже квартир в городе Саратове. Все эти данные были сохраненны в файл maindata.csv.

Разведочный анализ данных. Для дальнейшей работы с полученными данными необходимо было провести разведочный анализ, а именно:

1. Проверить данные на пропуски и при их наличии заполнить их;
2. Проверить данные на наличие отклонений и аномалий (выбросов) в распределении значений признаков;
3. Закодировать номинальные признаки.

Разведочный анализ данных проводился на языке программирования Python в бесплатной интерактивной облачной среде для работы с кодом – Google Colab.

Сначала были подключены все необходимые библиотеки для проведения разведочного анализа. После этого был подключен файл и считаны с него все данные.

В ходе исследования данных на пропуски были найдены пустые значения. Был проведен анализ этих пропусков и они все были заполнены.

Далее следовала проверка на наличие отклонений и аномалий в распределении значений признаков. Она происходили при помощи построения и анализа диаграмм размаха. Исходя из анализа построенных графиков аномалий и отклонений в распределении значений признаков не было обнаружено.

После этого было необходимо закодировать номинальные признаки, которые будут участвовать в построении модели регрессии. Таких признаков было три – это «balc» (тип балкона), «decoration» (тип ремонта) и «bath» (тип ванной комнаты).

Каждый из столбцов анализировался и их текстовые значения расположенные по возрастанию качества соответствующей характеристики заменялись численными значениями.

Построение регрессионной модели. Построение модели так же происходило на языке программирования Python. Полный код программы, реализующей регрессионный анализ рынка жилья представлен в приложении Б бакалаврской работы.

Перед построением модели регрессии были визуализированы зависимости характеристик квартир. В том числе вычилсялась корреляция признаков, строился график тепловой карты и парный график зависимости признаков. В ходе анализа построенных графиков были выделены основные зависимости между различными признаками и сделаны выводы о их значимости.

Сначала в среду разработки был подключен файл, содержащий все данные с объявлениями, с заполненными пропусками и закодированными номинальными признаками.

Далее основываясь на информации, полученной при визуализации зависимостей, был сделан вывод, что цена квартиры линейно зависит от её параметров. Также была выдвинута гипотеза о том, что среднестатистическая цена квартиры с некоторым набором параметров является показателем её качества. В этом случае можно вывести функцию зависимости качества квартиры от её параметров. Имея такую функцию, можно посчитать, какие из имеющихся на рынке квартир переоценены, а какие недооценены. Именно недооцененные квартиры и необходимо найти.

Для построения линейной модели использовалась библиотека scikit-learn, и её модули.

Перед построением модели необходимо было разделить исходные данные на тестовую и тренировочную части. В качестве зависимой переменной, которую необходимо предсказать, была выбрана цена квартиры. В качестве независимых переменных были выбраны все остальные столбцы таблицы с данными кроме столбца price_sqm поскольку он напрямую зависит от цены и общей площади квартиры.

Далее создается объект модели линейной регрессии с помощью конструктора LinearRegression(). Этот объект будет использоваться для обучения модели и предсказания цены квартир. Модель адаптируется к обучающим данным и вычисляет параметры регрессии. На этом этапе модель «обучается» на обучающих данных путем нахождения оптимальных коэффициентов для линейной комбинации признаков, которая наилучшим образом объясняет целевую переменную. В процессе обучения модель анализирует обучающие данные и пытается построить прямую линию регрессии, которая наиболее близко соответствует имеющимся данным и будет использоваться для предсказания целевых значений на новых данных.

После построения модели рассматривались её коэффициенты, они позволяют сделать выводы о значимости признаков модели.

По результатам работы модели регрессии необходимо было убедиться в её качестве, то есть оценить построенную модель регрессии. Для этого были подсчитаны среднеквадратическая ошибка (RMSE) и коэффициент детерминации (R^2), которые являются важной частью регрессионного анализа. Значения вычисленных оценок представлены ниже.

```
RMSE: 9260.1675072945467
R^2: 0.9599682108365556
```

Для построенной модели, значение RMSE приблизительно равно 9260, это означает, что средняя разность предсказанной цены и фактической цены составляет 9260 рублей. Значение R^2 для построенной модели приблизительно равно 0.96%. В данном случае значение оценки R^2 является очень хорошей, что позволяет сделать выводы о хорошем качестве построенной модели регрессии.

После рассмотрения оценок, результаты работы модели сохранялись в

отдельную переменную, так же как и цены квартир из выборки. Это необходимо для дальнейших расчетов и выбора наиболее выгодных предложений на рынке квартир. Так же используя эти переменные был построен график, представленный на рисунке 1. Точки на грифике, которые находятся над красной линией, представляют квартиры, цены на которые оказались завышены по сравнению с предсказанными моделью регрессии ценами квартир. Это означает, что эти квартиры считаются переоцененными. С другой стороны, точки, которые находятся под красной линией, представляют собой недооцененные квартиры. Чем дальше точка находится от линии, тем сильнее она недооценена или переоценена. Таким образом, из анализа графика был сделан вывод, что существует довольно много квартир, цены на которые оказались ниже предсказанной стоимости. Это может представлять интерес для покупателей, которые ищут недвижимость по более выгодной цене.

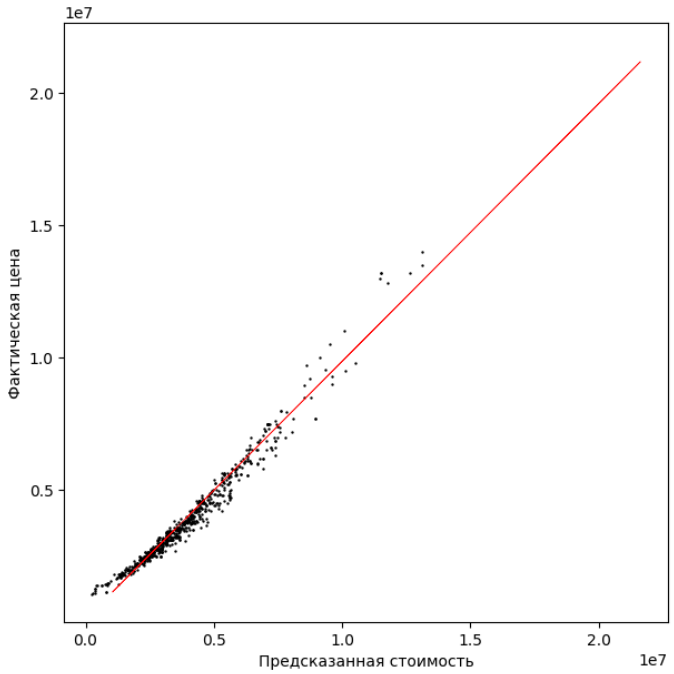

Рисунок 1 – Диаграмма зависимости предсказанной цены от актуальной

После того, когда были найдены квартиры, которые недооценены, были подсчитаны их «коэффициенты недооценённости». Был создан датафрейм, содержащий предсказанную цену и фактическую цену для каждой квартиры, а так же её коэффициент недооценённости. Полученный датафрейм был

отсортирован по убыванию этого коэффициента и выведено на экран первые 15 элементов полученного датафрейма. Результат работы программы представлен на рисунке 2.

	predicted_cost	actual_price	coefficient
578	3.499811e+06	2711500	1.290729
724	4.694871e+06	3650000	1.286266
47	3.876475e+06	3070000	1.262695
558	5.299871e+06	4200000	1.261874
250	4.491799e+06	3600000	1.247722
707	4.727447e+06	3800000	1.244065
590	5.549581e+06	4500000	1.233240
146	3.998643e+06	3300000	1.211710
378	5.203101e+06	4299000	1.210305
189	5.622359e+06	4650000	1.209109
482	4.684187e+06	3899000	1.201382
401	5.640959e+06	4699000	1.200459
207	5.628054e+06	4700000	1.197458
126	5.628054e+06	4700000	1.197458
432	4.166624e+06	3490000	1.193875

Рисунок 2 – Таблица с данными о квартирах с отсортированными по убыванию коэффициентами недооценённости

Анализ полученных результатов. В результате работы программы пользователю выдается таблица с данными, в которых содержится фактическая и предсказанная цена квартиры, а так же вычисленный для неё коэффициент недооценённости. Квартиры в этой таблицы расположены в порядке убывания коэффициента недооценённости. То есть первая квартира в таблице является самым выгодным вариантом для покупки, если не учитывать другие предпочтения пользователя. Пользователь может выбрать интересующие его характеристики и получить такую же таблицу с данными для квартир, удовлетворяющих его предпочтениям. Например, можно выбрать необходимое количество комнат, или тип ремонта в квартире и вывести данные с коэффициентом недооценённости для квартир с заданными требованиями.

Исходя из результатов работы программы коэффициент недооценённости первой квартиры в списке равен 1.290729, что означает, что первая квартира в списке недооценена примерно на 29%, вторая на 28.6% и так далее. Такой процент недооценённости является довольно ощутимым и делает покупку такой квартиры выгодной, на фоне характеристик и цен других квартир.

ЗАКЛЮЧЕНИЕ

В данной работе был рассмотрен регрессионый анализа рынка жилья в городе Саратове.

Была достигнута главная цель работы - изучение возможностей приложения регрессионного анализа к рынку жилой недвижимости, а также реализация программы, осуществляющей регрессионный анализ рынка жилья в городе Саратове.

В результате данной работы все поставленные задачи были выполнены, а именно:

1. рассмотрены теоретические аспекты исследования рынка жилой недвижимости и регрессионного анализа;
2. выбран источник данных для исследования;
3. реализована программа для сбора данных;
4. проведён разведочный анализ данных;
5. реализована программа для регрессионного анализа рынка жилья города Саратова;
6. были проанализированы полученные результаты.

Результаты работы можно использовать, чтобы понять динамику рынка недвижимости, определить потенциальные возможности для инвестиций и принять информированные решения при покупке недвижимости.

По итогам исследования можно с уверенностью сказать об эффективности метода регрессионно анализа рынка жилья. Такой метод экономит большое количество времени и предоставляет довольно обширную и наглядную информацию по многим параметрам рынка жилой недвижимости, что позволяет принимать обоснованное решение о покупке или инвестировании.