МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ

ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»



**Курсовая работа**

на тему «Построение модели линейной регрессии на данных о недвижимости»

Обучающегося 3 курса

очной формы обучения

направление подготовки:

09.03.01 Информатика и вычислительная техника

направленность (профиль):

Технологии разработки программного обеспечения

Волжанин Александр Павлович

Руководитель курсовой работы:

Доктор педагогических наук, профессор

Ильина Татьяна Сергеевна

Санкт-Петербург

2024

**СОДЕРЖАНИЕ**

ВВЕДНИЕ

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛИНЕЙНОЙ РЕГРЕССИИ

ОПИСАНИЕ ДАННЫХ  
ПОСТРОЕНИЕ МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ

ЗАКЛЮЧЕНИЕ

ЛИТЕРАТУРА

**ВВЕДЕНИЕ**

### Актуальность темы

Рынок недвижимости занимает значительное место в экономике любого региона, влияя на социально-экономическое развитие, инвестиционный климат и уровень жизни населения. В условиях динамичных изменений на рынке недвижимости, точная оценка стоимости объектов становится критически важной как для потенциальных покупателей и продавцов, так и для инвесторов и государственных органов. Эффективные методы прогнозирования цен на недвижимость позволяют принимать обоснованные решения, минимизировать риски и оптимизировать инвестиционные стратегии.

С развитием информационных технологий и доступностью больших объемов данных, появились возможности для применения современных статистических и машинно-обучающих методов анализа и прогнозирования цен на недвижимость. Одним из наиболее распространенных и эффективных методов является линейная регрессия, которая позволяет выявить и количественно оценить зависимость стоимости недвижимости от различных факторов.

**Цель работы:**  
Разработать и проанализировать модель линейной регрессии для предсказания стоимости квартиры на основе её характеристик, используя реальные данные о продаже недвижимости в Санкт-Петербурге.

1. Исследовать теоретические основы линейной регрессии и её применение в оценке недвижимости.
2. Провести предобработку и анализ исходных данных.
3. Построить модель линейной регрессии, учитывающую основные характеристики квартир.
4. Оценить точность и качество построенной модели.
5. Визуализировать результаты моделирования.
6. Предложить рекомендации по улучшению модели и дальнейшим направлениям исследования.

**Объект исследования:**  
Рынок жилой недвижимости Санкт-Петербурга с 2018 по 2021 год.

**Предмет исследования:**  
Зависимость стоимости квартир от их площади с использованием модели линейной регрессии.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛИНЕЙНОЙ РЕГРЕССИИ

**Понятие и классификация регрессионных моделей**

Регрессионный анализ является одним из фундаментальных методов статистического моделирования, используемым для изучения зависимости между зависимой переменной и одной или несколькими независимыми переменными. Основная цель регрессии — предсказание значений зависимой переменной на основе значений независимых переменных и понимание влияния последних на первую.

**Модель линейной регрессии**

Линейная регрессия предполагает, что зависимость между зависимой переменной Y и независимыми переменными X1,X2,…,Xp является линейной. Общая форма модели линейной регрессии записывается следующим образом:

Y = β₀ + β₁X₁ + β₂X₂ + … + βₚXₚ + ε

Где:

* Y - зависимая переменная
* X₁, X₂, …, Xₚ - независимые переменные
* β₀ - свободный член (интерцепт)
* β₁, β₂, …, βₚ - коэффициенты регрессии
* ε - случайная ошибка

**Основные предположения линейной регрессии**

Для корректности применения линейной регрессии необходимо соблюдение ряда предположений:

1. **Линейность**: Зависимость между зависимой и независимыми переменными должна быть линейной.
2. **Независимость ошибок**: Ошибки модели ε должны быть независимыми друг от друга.
3. **Гомоскедастичность**: Дисперсия ошибок должна быть постоянной для всех значений независимых переменных.
4. **Нормальность ошибок**: Ошибки модели ε должны распределяться нормально.
5. **Отсутствие мультиколлинеарности**: Независимые переменные не должны быть линейно зависимы друг от друга.

**Методы оценки параметров модели**

Наиболее распространённым методом оценки параметров линейной регрессии является метод наименьших квадратов (МНК). Он заключается в минимизации суммы квадратов отклонений фактических значений зависимой переменной от предсказанных моделью:

Minimize ∑ᵢ₌₁ⁿ (Yᵢ − Ŷᵢ)²

Где Ŷᵢ​=β0​+β1​Xi1​+β2​Xi2​+…+βp​Xip​.

**Критерии качества модели**

Для оценки качества построенной модели линейной регрессии используются различные метрики:

* **Коэффициент детерминации (R²)**: показывает долю объяснённой вариации зависимой переменной моделью.
* **Средняя абсолютная ошибка (MAE)**: Среднее значение абсолютных отклонений предсказанных значений от фактических.
* **Корень из средней квадратичной ошибки (RMSE)**: Квадратный корень из средней квадратичной ошибки, более чувствителен к большим отклонениям.

### ОПИСАНИЕ ДАННЫХ

### Источник и характер данных

Для построения модели использованы реальные данные о продаже квартир в Санкт-Петербурге, охватывающие период с 2018 по 2021 годы. Набор данных содержит 547 706 наблюдений и 13 признаков, предоставляющих информацию о различных характеристиках сделок с недвижимостью.

### Структура и содержание набора данных

Основные переменные набора данных включают:

* price — стоимость квартиры (в рублях).
* date, time — дата и время сделки.
* geo\_lat, geo\_lon — географические координаты местоположения квартиры.
* rooms — количество комнат.
* area — общая площадь квартиры (в квадратных метрах).
* building\_type — тип здания (панельный, кирпичный и др.).
* region — район расположения квартиры.
* kitchen\_area — площадь кухни.
* Дополнительные признаки, такие как floor (этаж), year\_built (год постройки здания), heating\_type (тип отопления) и другие.

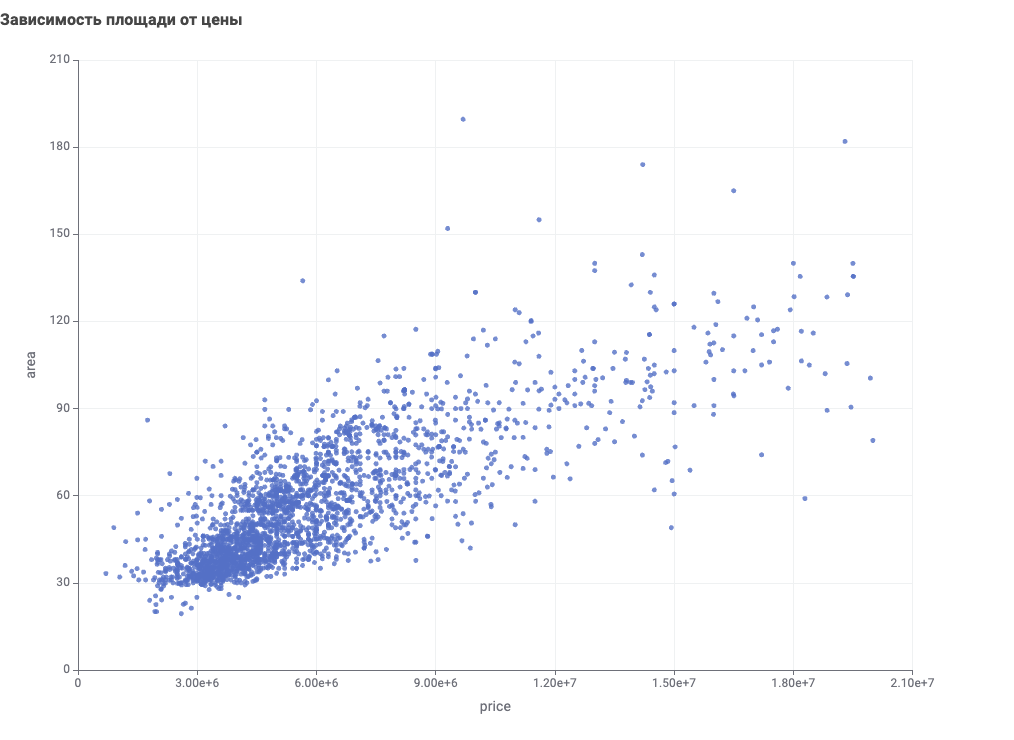
Мы будем использовать для создания модели линейной регрессии только цену и площадь.

### Предобработка данных

Для обеспечения качества и корректности анализа данных были проведены следующие этапы предобработки:

1. **Очистка данных:**
   1. Удаление дубликатов записей.
   2. Исправление или удаление некорректных значений
2. **Фильтрация данных:**
   1. **Стоимость квартиры:** от 300 000 руб. до 20 000 000 руб. — для исключения нерелевантных или ошибочных записей.
   2. **Количество комнат:** больше 0 — исключение квартир без комнат.
   3. **Площадь квартиры:** больше 0 и не превышает 190 кв.м. — для исключения аномально больших или нулевых значений.
3. **Обработка пропущенных значений:**
   1. Пропущенные значения заменены на соответствующие статистические показатели (например, среднее значение для числовых переменных).
   2. В случае значительного количества пропусков по некоторым признакам, эти признаки были исключены из дальнейшего анализа.

График после фильтрации аномальных данных:



ПОСТРОЕНИЕ МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ

**Обучение модели линейной регрессии:**

Модель обучена с использованием узла **Linear Regression Learner**, где зависимой переменной была выбрана цена квартиры (price), а независимой — площадь квартиры (area).

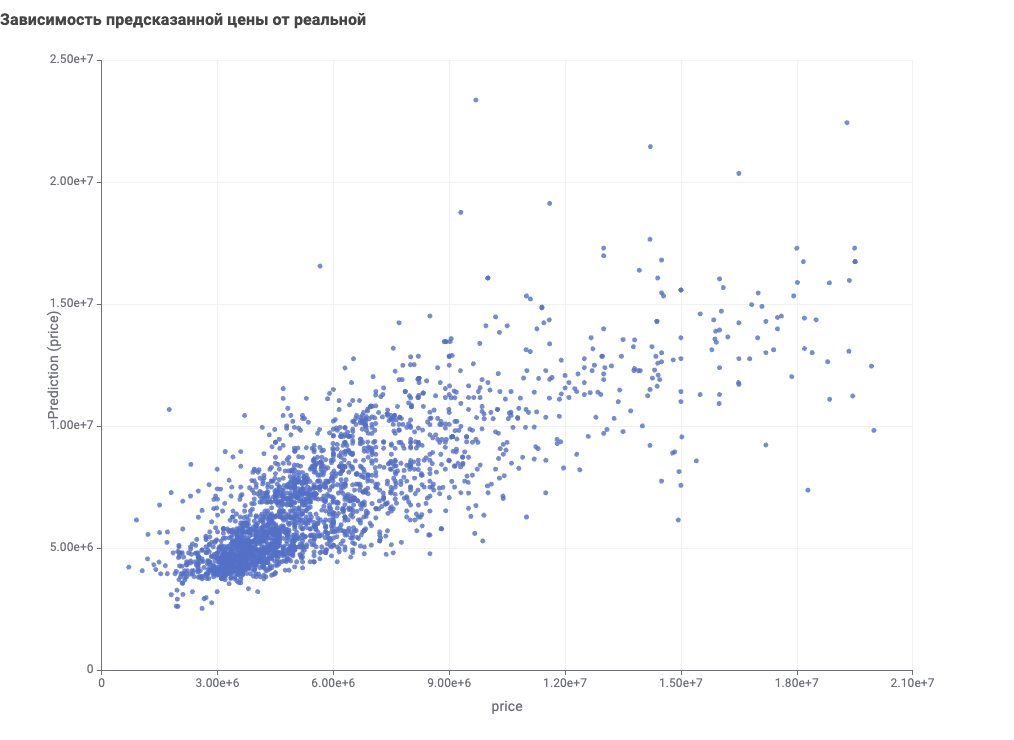
**Прогнозирование:**

Узел **Regression Predictor** применён для получения предсказанных значений стоимости квартир на основе обученной модели.

**Результаты оценки:**

1. Коэффициент детерминации (R2R^2R2) составил 0,594, что указывает на то, что модель объясняет 59,4% изменчивости данных.
2. Средняя абсолютная ошибка (MAE): 1,550,151 руб.
3. Корень из средней квадратичной ошибки (RMSE): 2,207,532 руб.

Построен график зависимости предсказанных значений стоимости квартир от их фактических значений:



Большинство точек расположены около диагонали, что свидетельствует о приемлемом качестве модели.

ЗАКЛЮЧЕНИЕ

В ходе работы была построена модель линейной регрессии для предсказания стоимости квартир в Санкт-Петербурге.

Основные результаты и выводы:

**Эффективность модели:**

1. Модель объясняет 59,4% изменчивости данных (R2=0,594R^2 = 0,594R2=0,594).
2. Средняя абсолютная ошибка (MAE) составляет 1,550,151 руб., что указывает на значительное отклонение предсказаний для некоторых объектов.

**Основной фактор влияния:**

Площадь квартиры является ключевым параметром, определяющим её стоимость. Это подтверждено значимостью коэффициентов регрессии.

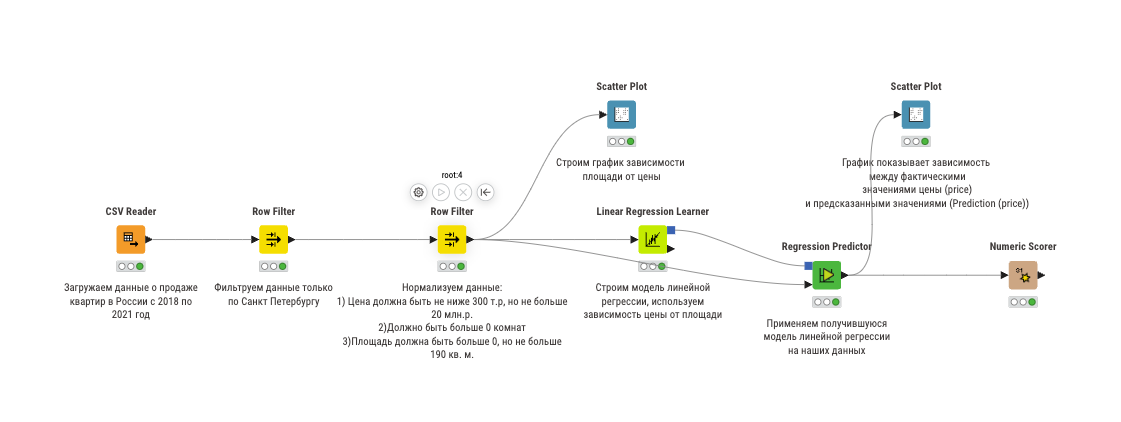
**Ограничения модели:**

Высокий уровень ошибки (RMSE = 2,207,532 руб.) может быть связан с отсутствием других значимых факторов, таких как расположение, тип здания, инфраструктура района.

**Рекомендации по улучшению:**

1. Добавление дополнительных признаков (например, координаты, тип здания).
2. Использование более сложных моделей машинного обучения (например, Random Forest или Gradient Boosted Trees).
3. Проведение детального анализа выбросов для улучшения качества данных.

Таким образом, разработанная модель представляет собой базовый инструмент для оценки стоимости квартир, который может быть доработан для повышения точности прогнозирования.

Полная картина узлов выглядит следующим образом:

ЛИТЕРАТУРА

Daniilak, M. Russia Real Estate 2018–2021 [электронный ресурс]. – Режим доступа: <https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-20182021/data> (дата обращения: 12.12.2024).

Обзор Knime Analytics Platform — open source системы для анализа данных [электронный ресурс]. – Хабр. – Режим доступа: <https://habr.com/ru/articles/320500/> (дата обращения: 12.12.2024).

KNIME. LinReg2LearnerNodeFactory2 [электронный ресурс]. – Режим доступа: <https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.regression.linear2.learner.LinReg2LearnerNodeFactory2> (дата обращения: 14.12.2024).