**Введение**

**Обзор проекта**

Проект посвящен разработке и оценке модели линейной регрессии для анализа рынка недвижимости. Основная цель - определить ключевые факторы, влияющие на цены на жилье, и оценить способность модели объяснять вариабельность этих цен.

**Общий подход к решению**

Сосредоточимся на создании и оценке модели линейной регрессии, которая будет анализировать различные характеристики жилой недвижимости, такие как площадь, количество спален, ванных комнат, качество и состояние строения, для предсказания их конечной цены. Это позволит оценить, какие факторы оказывают наибольшее влияние на стоимость недвижимости.

**Обзор данных**

**Источник и описание данных**

Исходные данные для нашей модели включают детализированный набор информации о различных объектах недвижимости. Данные содержат характеристики, такие как площадь, количество спален и ванных комнат, оценка качества и состояние строения, а также цену.

В датафрейме содержится информация об объектах недвижимости. Присутствуют как числовые (площадь, количество спален, ванных комнат, цена), так и категориальные (оценка качества, состояние строения) признаки.

**Предварительная обработка данных**

Перед началом анализа данные прошли этап предварительной обработки, включающий визуализацию и анализ распределения переменных.

# Импорт библиотек для анализа и визуализации данных

import matplotlib.pyplot as plt

import seaborn as sns

# Описательная статистика

desc\_stats = df.describe()

# Корреляционная матрица

corr\_matrix = df.corr()

# Визуализация корреляционной матрицы

plt.figure(figsize=(12, 10))

sns.heatmap(corr\_matrix, annot=True, fmt=".2f", cmap='coolwarm')

plt.title("Корреляционная матрица признаков")

plt.show()

desc\_stats, corr\_matrix['price'].sort\_values(ascending=False)

**Методология**

Описание алгоритма линейной регрессии:

Алгоритм линейной регрессии используется для моделирования и анализа взаимосвязей между зависимой и независимыми переменными. Он применяется для предсказания цены на основе различных характеристик недвижимости.

# Создание и обучение модели линейной регрессии

linear\_model = LinearRegression()

linear\_model.fit(X\_train, y\_train)

# Оценка модели на тестовой выборке

y\_pred = linear\_model.predict(X\_test)

# Рассчитываем метрики для оценки качества модели

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

# Вывод результатов

mse, r2

**Реализация алгоритма линейной регрессии:**

**Разделение данных на выборки:** Данные были разделены на обучающую и тестовую выборки.

# Выбираем предикторы и целевую переменную

X = df\_cleaned.drop('price', axis=1)

y = df\_cleaned['price']

# Разделение данных на обучающую (80%) и тестовую (20%) выборки

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Вывод размера обучающей и тестовой выборок

X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape

**Определение R²:**

R² является статистической мерой, отражающей долю вариации зависимой переменной (цены на недвижимость), которая объясняется независимыми переменными модели (такими как площадь, количество спален, ванных комнат, качество строения и др.).

Значение R² варьируется от 0 до 1, где близость к 1 указывает на то, что модель хорошо объясняет вариацию зависимой переменной.

**Использование R² в проекте:**

В проекте R² используется для оценки того, насколько хорошо модель линейной регрессии предсказывает цены на недвижимость.

При высоком значении R² можно сделать вывод, что выбранные признаки эффективно влияют на ценообразование недвижимости, а модель адекватно отражает реальные зависимости на рынке.

**Почему выбрана именно R²:**

**Интерпретируемость:** R² легко интерпретировать, так как он показывает долю вариации в целевой переменной, объясненную моделью. Это делает его понятным для объяснения результатов неспециалистам.

**Общепринятость:** R² является стандартной метрикой в регрессионном анализе, что делает её результаты сопоставимыми с другими исследованиями в этой области.

**Полезность в оценке модели:** Эта метрика позволяет оценить, насколько хорошо модель работает на данных, отличных от тех, на которых она обучалась, что является важным аспектом в проверке обобщающей способности модели.

**Обсуждение результатов**

**Интерпретация R²:**

Значение R², равное 0.701, указывает на то, что приблизительно 70.1% вариативности цены на жилье (зависимая переменная) может быть объяснено выбранными признаками (независимые переменные) в модели. Это довольно высокий уровень объяснительной способности, что говорит о том, что модель довольно эффективно улавливает взаимосвязи в данных.

**Эффективность модели:**

Поскольку значение R² близко к 1, можно сделать вывод, что модель линейной регрессии довольно хорошо подходит для данных. Она успешно учитывает значительную часть факторов, влияющих на цену недвижимости.

**Аспекты для улучшения:**

Несмотря на высокий R², всегда есть возможность для улучшения модели. Например, можно исследовать добавление или исключение некоторых признаков, использовать другие трансформации данных или даже применить более сложные модели, такие как ансамблевые методы или нейросети, для сравнения результатов.

**Заключение**

Модель линейной регрессии, разработанная в рамках проекта, показала достаточно высокую эффективность в предсказании цен на жилье, что подтверждается коэффициентом детерминации R², равным 0.701. Это означает, что модель успешно объясняет около 70.1% вариации цен на основе выбранных характеристик недвижимости.

**Список литературы и ресурсов**

В данном разделе представлены ключевые библиотеки и ресурсы, которые были использованы в проекте для разработки и анализа модели Random Forest:

**Pandas:** Основная библиотека Python для анализа и манипуляции данными. Использовалась для загрузки, очистки и предварительной обработки данных о ноутбуках.

Документация: <https://pandas.pydata.org/>

**NumPy:** Библиотека для поддержки больших многомерных массивов и матриц, включая математические функции для работы с этими массивами.

Документация: <https://numpy.org/doc/>

**Matplotlib:** Библиотека для создания статических, анимированных и интерактивных визуализаций в Python.

Документация: <https://matplotlib.org/stable/index.html>

**Seaborn:** Библиотека визуализации данных Python на основе matplotlib, предоставляющая более высокоуровневый интерфейс для создания привлекательных и информативных графиков статистических данных.

Документация: <https://seaborn.pydata.org/>

**Scikit-learn (sklearn):** Одна из основных библиотек машинного обучения в Python, использованная для реализации алгоритма R2, а также для предварительной обработки данных, разделения данных на обучающую и тестовую выборки и оценки эффективности модели.

Документация:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2\_score.html