

Содержание

1. Цель работы, выходная переменная и признаки	3
2. Предварительная обработка данных	3
3. Разделение набора данных	3
4. Модели машинного обучения	4
5. Прогнозные значения	4
6. График регрессии, прогнозные значения и уравнение регрессии	4
7. Оценка точности моделей	5
Заключение	5

1. Цель работы, выходная переменная и признаки

Цель работы: Разработать модели машинного обучения для прогнозирования арендной платы за жилье на основе предоставленных данных.

Выходная переменная (целевая):

- Rent: (арендная плата).

Признаки набора данных:

- BHK: Количество спален, холлов и кухонь (числовой).
- Size: Площадь жилья (числовой).
- Floor: Этаж (категориальный).
- Area Type: Тип площади (категориальный).
- Area Locality: Район (категориальный).
- City: Город (категориальный).
- Furnishing Status: Статус мебелировки (категориальный).
- Tenant Preferred: Предпочтительный арендатор (категориальный).
- Bathroom: Количество ванных комнат (числовой).
- Point of Contact: Контакт (категориальный).

2. Предварительная обработка данных

Данные загружены из файла House_Rent_Dataset.csv.

- Пропущенные значения: В данных нет пропусков (проверено с помощью `isnull().sum()`).
- Исключены: Posted On, Floor, Area Locality (из-за большого количества уникальных значений).
- Кодирование символьных данных: Категориальные признаки (Area Type, City, Furnishing Status, Tenant Preferred, Point of Contact) закодированы с помощью OneHotEncoder.

Предобработчик: ColumnTransformer с OneHotEncoder для категориальных и remainder='passthrough' для числовых.

3. Разделение набора данных

Набор данных разделен на обучающую (80%) и тестовую (20%) выборки с помощью `train_test_split` (random_state не указан явно, но подразумевается воспроизводимость).

X: Все признаки кроме Rent, Posted On, Area Locality, Floor. y: Rent.

4. Модели машинного обучения

Созданные модели:

- Линейная регрессия: `LinearRegression`.
- Полиномиальная регрессия: `PolynomialFeatures(degree=2) + LinearRegression`.
- Ридж-регрессия: `Ridge(alpha=1.0 по умолчанию)`.

Модели объединены в пайплайны с предобработчиком.

5. Прогнозные значения

Прогнозы вычислены с помощью `predict` на тестовой выборке для каждой модели.

6. График регрессии, прогнозные значения и уравнение регрессии

График: Построен scatter plot для Size vs Rent с данными, тестовыми значениями и прогнозами линейной регрессии (`matplotlib`).

Уравнения регрессии (коэффициенты извлечены из моделей):

- Линейная:

$$\begin{aligned} \text{Rent} = & -18486.370 \\ & +5553.960 \cdot \text{Area Type_Built Area} - 1275.123 \cdot \text{Area Type_Carpet Area} \\ & -4278.837 \cdot \text{Area Type_Super Area} - 4874.722 \cdot \text{City_Bangalore} \\ & -12703.227 \cdot \text{City_Chennai} + 2349.352 \cdot \text{City_Delhi} \\ & -21727.989 \cdot \text{City_Hyderabad} - 7842.364 \cdot \text{City_Kolkata} + 44798.950 \cdot \text{City_Mumbai} \\ & +5569.976 \cdot \text{Furnishing Status_Furnished} - 3045.807 \cdot \text{Furnishing Status_Semi-Furnished} \\ & -2524.169 \cdot \text{Furnishing Status_Unfurnished} + 4079.037 \cdot \text{Tenant Preferred_Bachelors} \\ & +3307.130 \cdot \text{Tenant Preferred_Bachelors/Family} - 7386.167 \cdot \text{Tenant Preferred_Family} \\ & -5005.848 \cdot \text{Point of Contact_Contact Agent} + 19397.413 \cdot \text{Point of Contact_Contact Builder} \\ & -14391.565 \cdot \text{Point of Contact_Contact Owner} + 3669.642 \cdot \text{BHK} \\ & +36.602 \cdot \text{Size} + 11502.331 \cdot \text{Bathroom} \end{aligned}$$

7. Оценка точности моделей

Модель	R^2	RMSE	MAE	MSE
Линейная регрессия	0.5205	43717.0552	22232.6300	1911180916.2022
Полиномиальная регрессия	0.7958	28528.0921	14062.6489	813852036.6120
Ридж-регрессия	0.5205	43716.6342	22222.5073	1911144103.7330

Полиномиальная модель показывает наилучшие результаты (высокий R^2 , низкие ошибки).

Заключение

В работе реализован регрессионный анализ для прогнозирования арендной платы. Полиномиальная регрессия степени 2 оказалась наиболее точной. Рекомендуется дальнейшая оптимизация (например, подбор гиперпараметров) для улучшения моделей.