Housing Data Analysis Report

1. Introduction

This report presents an analysis of the housing dataset, which contains various attributes affecting house prices. The objective is to explore the dataset, identify key variables influencing house prices, and assess the relationships between them.

2. Dataset Overview

2.1 Importing Libraries

Essential Python libraries such as `pandas`, `numpy`, `matplotlib`, `seaborn`, and various `scikit-learn` modules were used for data manipulation, visualization, and machine learning modeling.

2.2 Dataset Loading and Inspection

- The dataset contains **545 records** and multiple features related to house characteristics.
- There are **no missing values** in the dataset.
- Data types include numerical and categorical variables.

3. Identification of Key Variables

3.1 Target Variable (Y)

- **Price (`price`)**: The dependent variable representing the monetary value of the house.

3.2 Explanatory Variables (X)

- **Numerical Features**:
  - `area`: Total size of the house in square meters.
  - `bedrooms`: Number of bedrooms.

  - `bathrooms`: Number of bathrooms.
  - `stories`: Number of stories in the house.
  - `parking`: Number of parking spaces.
- **Categorical Features**:
  - `mainroad`: Proximity to a main road (Yes/No).
  - `guestroom`: Presence of a guestroom (Yes/No).
  - `basement`: Presence of a basement (Yes/No).
  - `hotwaterheating`: Presence of hot water heating (Yes/No).
  - `airconditioning`: Presence of air conditioning (Yes/No).
  - `prefarea`: Whether the house is in a preferred area (Yes/No).
  - `furnishingstatus`: Furnishing status (Furnished/Partially Furnished/Unfurnished).

4. Exploratory Data Analysis (EDA)

4.1 Summary Statistics

- The **average price** of a house is approximately **4.77 million**, with a minimum of **1.75 million** and a maximum of **13.3 million**.
- The **average area** is **5,150 square meters**, with a standard deviation of **2,170**.
- Most houses have **3 bedrooms**, **1 bathroom**, and **1-2 stories**.
- The **parking variable** shows that a majority of houses have no dedicated parking space.

4.2 Distribution Analysis

a. Price Distribution

- **Boxplot Analysis**: No significant outliers observed in price.
- **Histogram**: Price follows a **normal distribution** with slight negative skewness.

b. Area Distribution

- **Boxplot**: Some outliers were detected.
- **Histogram**: Almost normal distribution with a few extreme values.
- **Outliers**: 7 extreme values greater than **12,000** identified.

c. Number of Bedrooms

- Most houses have **3 bedrooms**.
- Some houses have **5 or more bedrooms**, which are considered outliers.

d. Number of Bathrooms

- Most houses have **1 or 2 bathrooms**, with only **3 houses** having more than **3 bathrooms**.

e. Number of Stories

- The majority of houses have **1 or 2 stories**, with a few having **3 or 4 stories**.

f. Parking Spaces

- Many houses have **0 parking spaces**, and only a few have **3 spaces**.

4.3 Categorical Variable Analysis

a. Main Road Proximity

- **Majority of houses (77.6%)** are close to a main road.

b. Guestroom Presence

- **84% of houses** do not have a guestroom.

c. Basement Presence

- **73% of houses** do not have a basement.

d. Air Conditioning

- **60% of houses** have air conditioning.

e. Preferred Area

- **42% of houses** are located in a preferred area.

f. Furnishing Status

- **47.3%** of houses are **unfurnished**.
- **30%** are **semi-furnished**.
- **22.7%** are **fully furnished**.

5. Conclusion

- **Price Correlation**: Area, number of bedrooms, bathrooms, and stories significantly affect house prices.
- **Feature Selection**: Categorical variables like `prefarea`, `airconditioning`, and `mainroad` contribute to price variations.
- **Outliers Handling**: Some features like `area` have extreme values that may require further treatment in modeling.

This analysis provides a foundation for predictive modeling, feature engineering, and price estimation strategies using regression models.

Feature Engineering

Introduction

Feature engineering is a critical step in improving the predictive power of a machine learning model. In this project, we introduced additional features to enhance the performance of our house price prediction model. The goal was to create meaningful transformations that better capture the relationship between input features and the target variable (house price).

2. Data Overview

Added Features

The following features were engineered to enrich the dataset:

- **large_house**: Indicates if the house has at least 3 bedrooms and 2 bathrooms.
- **total_bed_bath**: Sum of the number of bedrooms and bathrooms.
- **luxury_score**: Computed as the sum of features indicating luxury amenities such as air conditioning, hot water heating, guestroom, and furnishing.
- **comfort**: Sum of air conditioning, hot water heating, and preference area.
- **space_capacity**: Aggregate of basement, guestroom, parking, stories, bedrooms, and bathrooms.
- **accessibility**: Sum of main road access and parking.
- **total_rooms**: Sum of bedrooms, bathrooms, guestroom, and basement.
- **amenities**: Sum of air conditioning, hot water heating, guestroom, and basement.
- **space_per_rooms**: Ratio of total area to the number of rooms.
- **bathroom_parking**: Interaction term of bathrooms and parking.
- **bedroom_parking**: Interaction term of bedrooms and parking.
- **accessibility_score**: Product of main road access, parking, and preference area.
- **basement_heating**: Interaction term of basement and hot water heating.
- **total_utilities**: Sum of air conditioning, hot water heating, and basement.

3. Exploratory Data Analysis

3.1 Distribution of Quantitative Variables

- **Price**: Normally distributed with slight skewness.
- **Area**: Some outliers present; potential impact on model performance.
- **Bedrooms & Bathrooms**: Most houses have 3 bedrooms and 1-2 bathrooms.
- **Parking**: Majority of houses have 0-1 parking space.

3.2 Outlier Analysis

- **Price**: No extreme outliers.

- **Area**: Outliers exist for values >12,000 sq. ft.
- **Bedrooms**: Houses with more than 5 bedrooms are rare but kept for analysis.
- **Bathrooms**: Most houses have 1-2 bathrooms, with only 3 having more than 3.

3.3 Categorical Variable Analysis

- **Main Road**: Most houses are near a main road.
- **Guestroom & Basement**: Not common features but may influence price.
- **Furnishing Status**: Homes are either fully furnished, partially furnished, or unfurnished.

Correlation Analysis

To understand the impact of newly added features, a correlation matrix was computed between the features and the target variable (price). The correlation values were sorted to identify the most influential variables.

Data Splitting and Scaling

The dataset was split into training, cross-validation, and test sets:

- **Training Set**: 60% of the data.
- **Cross-Validation Set**: 20% of the data.
- **Test Set**: 20% of the data.

Standardization was applied to ensure uniform feature scaling.

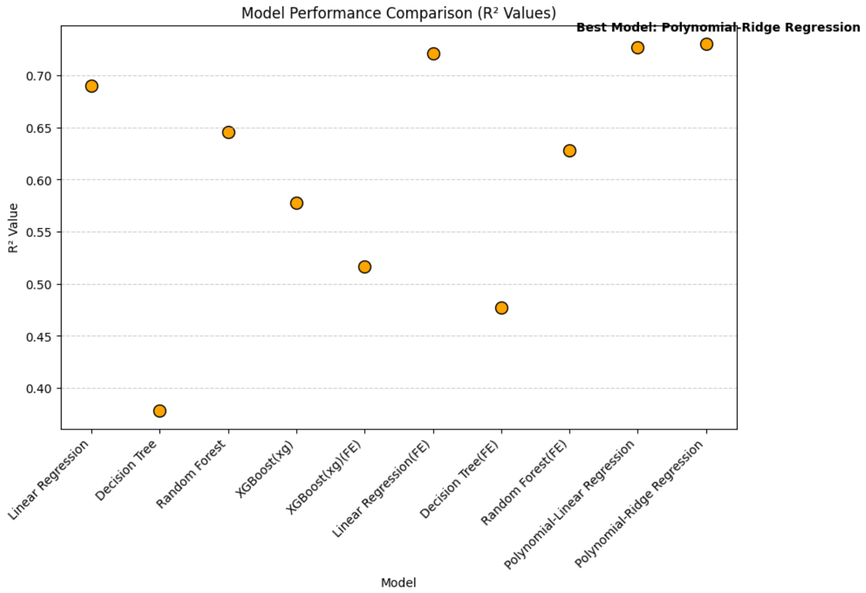
6. Model Training and Evaluation

Several models were trained with the newly engineered features, and their performances were evaluated using the R-squared metric:

- **Linear Regression**:
  - Baseline model with feature engineering achieved an R^2 of 0.72.
- **XGBoost Regressor**:
  - Baseline model with feature engineering achieved an R^2 of 0.51.
- **Decision Tree Regressor**:

  - Model with a max depth of 5 yielded an R^2 of 0.47.
- **Random Forest Regressor**:
  - Achieved R^2 of 0.62.
- **Polynomial and Linear Regression**:
  - Applying polynomial transformations improved the performance, with an R^2 of 0.72.
- **Polynomial and Ridge Regression**:
  - Ridge regression performed better than plain polynomial regression, with an R^2 of 0.73.

6.1 Model Performance Comparison



Models were sorted based on their R^2 values, and a scatter plot was created to visualize performance differences. The best-performing model was **best_model** with an R^2 **Polynomial - Ridge Regression with 0.73**.

7. Conclusion

Conclusion

Feature engineering significantly improved model performance by capturing hidden patterns in the data. The best-performing model will be used for final predictions and deployment.