HOUSE PRICE ANALYSIS AND PREDICTION

Submitted by: Sana Qayyum

INTRODUCTION

In recent years, the real estate market has experienced significant fluctuations, making accurate house price predictions crucial for buyers, sellers, and investors. This project aims to develop a predictive model for house prices using various features, including property type, location, and area size

OBJECTIVES

- Examine the provided dataset for house listings to understand its structure, features, and potential issues such as missing values and outliers.
- To Clean and Preprocess the Data:
- Identify and handle missing values and outliers to ensure data quality.
- Transform categorical features into numerical values using appropriate encoding techniques.
- To Perform Exploratory Data Analysis (EDA):
- Visualize the distribution of house prices and other relevant features to identify trends and patterns.
- Investigate potential relationships between features and the target variable (house prices) through visualizations such as scatter plots and box plots.
- To Engineer Relevant Features:
- Create new features that may enhance the predictive power of the model, such as price per square foot and transformations of existing features.
- To Develop Predictive Models:
- Train and evaluate multiple machine learning models, ito predict house prices based on the available features.
- To Evaluate Model Performance:
- Assess the accuracy and effectiveness of each model using performance metrics such as RMSE, R-squared, and visualizations of predicted vs. actual prices.

DATA EXPLORATION AND CLEANING

The initial step in preparing the dataset involved addressing various data quality issues, including missing values, inconsistencies, and outliers.

- Missing Values:

Analyzed each feature for missing values and determined appropriate strategies for handling them.

- Outlier Detection and Handling:

Identified outliers using the Interquartile Range (IQR) method. Values beyond 1.5 times the IQR from the first and third quartiles were flagged as outliers. These were removed to prevent them from skewing the analysis and model performance.

- Data Type Conversion:

Ensured that all features were of the correct data type. For example, categorical features were converted from object types to categorical types, facilitating easier processing during encoding.

DATA CLEANING

The Dataset consists of 20 columns.To ensure the dataset was well-suited for analysis and modeling, several preprocessing steps were performed. This involved removing redundant or irrelevant columns that were not necessary for the predictive model. Specifically, removing these columns, the dataset was streamlined to focus on the more relevant features, allowing for more efficient and accurate model development.

Handling Outliers in the Dataset

During data analysis, I discovered some irregularities when inspecting the unique values for the baths and bedrooms columns. To ensure the dataset's quality and remove unrealistic or erroneous data points, I applied the following conditions to filter out outliers:

- Bathrooms greater than Bedrooms: It is uncommon for a property to have more bathrooms than bedrooms. Therefore, I removed all instances where the number of bathrooms exceeded the number of bedrooms.
- Bedrooms and Bathrooms Constraints: Upon further analysis, I found that certain properties had unrealistic values for bedrooms and bathrooms:
- Properties with 0 bedrooms or 0 bathrooms were considered invalid.
- Properties with more than 10 bathrooms or more than 12 bedrooms were deemed extreme outliers.
- These conditions helped to clean the data, ensuring that the remaining values for bathrooms and bedrooms were realistic and within a plausible range, improving the overall quality and reliability of the dataset.

DATA EXPLORATION

After cleaning the data, exploratory data analysis (EDA) was conducted to gain insights into the dataset and understand relationships between features and the target variable (house prices):

- Descriptive Statistics:
  - Generated summary statistics for numerical features, including count, mean, median, minimum, and maximum values. This provided a basic understanding of the distribution of house prices and other features.
- Distribution Analysis:
  - Plotted histograms and box plots to visualize the distribution of house prices. The initial analysis revealed a right-skewed distribution, which prompted the application of log transformation to normalize the data.
- Correlation Analysis:
  - Computed the correlation matrix to examine relationships between numerical features. Visualizations such as heatmaps were used to identify strong correlations that could indicate potential predictors for the house price.
- Visualizations:
  - Created scatter plots and bar plots to explore relationships between individual features and house prices. For example, scatter plots of bedrooms and bathrooms against prices helped visualize the impact of these features on the target variable.

FEATURE ENGINEERING:

One of the crucial aspects of property data is Area Size, which was provided in different units—Marla and Kanal. To ensure uniformity in the dataset and facilitate accurate modeling, I standardized the area size by converting it all to square feet (sqft). This involved the following steps:

1. Conversion Based on Area Type:

   - If the area type was Marla, I converted it to square feet by multiplying the area size by 272.25 (since 1 Marla equals 272.25 sqft).
   - If the area type was Kanal, I used a conversion factor of 5445 (since 1 Kanal equals 5445 sqft).
   - For other area types or already existing square feet values, no transformation was applied.
   - After converting the area sizes to a common unit, the original columns (Area Type and Area Size) became redundant and were removed from the dataset.
   - Following the standardization of area sizes, I calculated the price per square foot for each property by dividing the total price by the area size in square feet, which provided a consistent metric to compare properties of different sizes

ANALYSIS:

Average Property Prices by City

To analyze the variation in property prices across different cities, I plotted the average price for each city using the following bar plot

- Lahore: As evident from the bar plot, Lahore stands out as having the highest average property prices. This could be attributed to rapid urbanization, high demand for residential and commercial spaces, and the presence of premium housing developments.
- Karachi: Following Lahore, Karachi shows the second-highest average property prices. Karachi, being a major economic hub, sees high property prices, especially in commercial zones and upscale residential areas.
- Islamabad: Islamabad comes third in terms of property prices. As the capital city, Islamabad has relatively high prices, driven by well-planned residential areas, government offices, and embassies.

Ave**rage Property Prices by Property Type**

To understand how property prices vary across different property types, I plotted the average price for each property category using the following bar plot.

- Farmhouses: Farmhouses are shown to have the highest average price. This could be because farmhouses are typically located in larger plots of land and often offer luxurious amenities, making them more expensive compared to other property types.
- Houses: Houses follow farmhouses in terms of price. Houses, particularly in urban areas, often come with higher prices due to increasing demand for residential spaces and limited availability of land.
- Flats/Apartments: Flats have the lowest average price among the property types. Flats generally have smaller spaces compared to houses or farmhouses, and their prices reflect this.

Log Transformation:

In this project, the prices were right-skewed, meaning a small number of properties had very high prices, causing the data to deviate from a normal distribution.

The log transformation was applied to the target variable (price per square foot) to reduce the impact of outliers and bring the distribution closer to normality. This helps linear models perform better, as they generally assume normally

MODELS

In this project, several machine learning models were employed to predict house prices based on the available features. Each model was evaluated for its performance, and insights were drawn from their results.

1. Linear Regression

Overview: Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (house price) and one or more independent variables (features). It assumes a linear relationship between the variables.
Implementation:

- The model was trained on the training dataset using features such as property type, city, number of bedrooms, bathrooms, and price per square foot.
- Performance was evaluated using metrics like Root Mean Squared Error (RMSE) and R-squared.

Results: The linear regression model provided a baseline for comparison with more complex models. Its simplicity and interpretability made it a good starting point.

2**. Random Forest**
Overview: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction (for regression) of the individual trees. It is effective at capturing complex relationships and handling overfitting.
Implementation:

- The model was trained using the same features as the linear regression model.
- Hyperparameters were tuned using cross-validation to optimize model performance.

Results: The Random Forest model demonstrated improved performance over the linear regression model, showing a lower RMSE and higher R-squared value. Its ability to handle non-linear relationships made it a suitable choice for this dataset.

Model Evaluation
Each model's performance was evaluated based on RMSE and R-squared values on the validation and test datasets