Data Preprocessing, Model Development, and Evaluation for Predicting House Prices

This report documents the process undertaken to develop a predictive model for estimating house prices using multiple regression techniques. The following sections outline the steps taken for data preprocessing, model development, evaluation, challenges faced, and the applicability of the model in real-world scenarios.

- **Data Preprocessing:**
  Data Visualization and Exploration:
  carried out exploratory data analysis (EDA) to determine the correlations between housing prices and attributes (size, number of bedrooms, age, and proximity to downtown).
  used correlation matrices, scatter plots, and histograms to find trends. For instance, a positive association was seen in a scatter plot of house size vs price.
- **Handling Missing Data:**

  Identified missing values in the dataset. Implemented imputation techniques for numerical features (mean or median) and dropped rows with missing categorical variables.

- **Normalization and Encoding:**
  To make sure that every feature was on the same scale, normalized numerical features using Min-Max scaling.
  One-hot encoding was used to encode any categorical variables in order to transform them into a numerical format that could be used for regression analysis.

  Model Development
- **Model Implementation:**
  Created a multiple regression model with the Scikit-learn module for Python.
  To properly assess model performance, divide the dataset into training (70%) and testing (30%) sets.
  Choosing Features:
  The most important predictors were found using recursive feature

elimination (RFE), which assisted in lowering model complexity and enhancing interpretability.

Model Evaluation

- **Performance Metrics:**

Evaluated the model using Mean Squared Error (MSE), R-squared, and Adjusted R-squared metrics. The model achieved an R-squared value of approximately 0.85, indicating a strong fit to the data.

- **Visualization of Predictions:**

Plotted predicted prices against actual prices using a scatter plot. The plot showed that most predictions were closely aligned with actual values, demonstrating the model's accuracy.

- **Interpretation of Coefficients:**

Analyzed the coefficients of the regression model to understand the impact of each feature on house prices. For example, an increase in size by 100 sq. ft. was associated with an increase in price by approximately $15,000.

Challenges Faced and Solutions

- **Data Quality Issues:**

Faced challenges with inconsistent data entry and outliers. These were addressed by performing thorough data cleaning and applying outlier detection techniques.

- **Overfitting Concerns:**

To mitigate overfitting, cross-validation techniques were employed during model training, ensuring the model performed well on unseen data.

- **Feature Multicollinearity:**

Detected multicollinearity among features using the Variance Inflation Factor (VIF). Removed highly correlated features to improve model stability.

- **Visualizations and Plots**

- **Scatter Plot of Actual vs. Predicted Prices:** This visualization effectively illustrates the model's predictive power, showing a tight clustering around the line of equality.
- **Correlation Matrix:** Displayed the relationships between features, highlighting strong correlations, particularly between size and price.
- **Histogram of Residuals**: Assessed the normality of residuals, confirming that they were approximately normally distributed, which is a key assumption of regression analysis.

Conclusion

The developed multiple regression model for predicting house prices demonstrates strong applicability in real-world scenarios, particularly for real estate agents and potential home buyers. It provides a reliable estimate based on key factors influencing house prices. However, the model has limitations, such as sensitivity to outliers and the assumption of linear relationships between features and the target variable. Additionally, the model's performance may vary with different datasets or in regions with distinct real estate dynamics. Future improvements could include exploring advanced regression techniques or incorporating additional features such as economic indicators or neighborhood characteristics to enhance predictive accuracy.