PLATA, Giannah Faith F.
BSCS 3B-IS

REPORT

- Data Preprocessing

The first step involved loading the dataset and conducting exploratory data analysis (EDA) to understand the relationships between various features and house prices. Missing values were checked, and since none were found, the next step was to normalize the features to ensure they were on a similar scale. Standardization was applied to the continuous variables, while categorical variables were encoded if necessary. This preparation ensured that the dataset was ready for modeling.

- Model Development

A multiple regression model was implemented using Python's Scikit-learn library. The dataset was split into training (70%) and testing (30%) sets to evaluate the model's performance. The model was trained on the training set, and feature selection was performed to identify the most significant predictors, which included size, number of bedrooms, age, and proximity to downtown.

- Model Evaluation

The model's performance was assessed using several metrics, including Mean Squared Error (MSE), R-squared, and Adjusted R-squared. The MSE indicated the average squared difference between predicted and actual prices, while R-squared provided insight into the proportion of variance explained by the model. Additionally, a plot of predicted prices against actual prices was created to visualize the model's accuracy, showing a strong correlation between the two.

- Model Improvement Attempts

To improve the model, various techniques were considered, such as polynomial regression to capture non-linear relationships and regularization methods like Ridge or Lasso regression to prevent overfitting. However, initial results indicated that the linear model performed adequately, so further complexity was deferred for future iterations.

- Challenges Faced

One of the main challenges encountered was ensuring that the normalization of features did not distort the relationships between them. This was addressed by carefully selecting the normalization method and validating the results through visualization. Additionally, feature selection was crucial, as including irrelevant features could lead to overfitting, which was mitigated by using correlation analysis and evaluating model performance iteratively.

- Visualizations and Plots

1. **Scatter Plots**: These illustrated the relationships between individual features and house prices, highlighting key correlations.
2. **Correlation Matrix**: This provided a comprehensive view of how features correlated with each other and with the target variable (house price).
3. **Predicted vs. Actual Prices Plot**: This plot demonstrated the model's accuracy, with points closely aligning along the identity line indicating good predictive performance.

Conclusion

The developed multiple regression model effectively predicts house prices based on key features such as size, number of bedrooms, age, and proximity to downtown. Its applicability in real-world scenarios is significant, particularly for real estate companies looking to estimate property values accurately. However, potential limitations include the model's reliance on linear relationships and its sensitivity to outliers, which could affect predictions in cases of extreme values. Future work could explore more complex models or additional features to enhance predictive accuracy further.