Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Name: Nixon E. Coronado
Year & Section: BSCS – 2A

Introduction

Predicting house prices is a crucial task in the real estate industry, as it helps buyers, sellers, and investors make informed decisions. In this report, I will explore a dataset of house prices and develop a multiple linear regression model to predict house prices based on various features. I will also analyze the relationships between the features and the target variable, as well as evaluate the model's performance using various metrics.

Exploratory Data Analysis (EDA)

I began by loading the house prices dataset using pd.read_csv('datasets_house_prices.csv'). To gain a better understanding of the data, I explored the dataset using data.info(), data.describe(), and data.isnull().sum(). This allowed me to understand the structure, content, and quality of the data, ensuring that there were no missing values.

Next, I identified the feature columns ('Size (sqft)', 'Bedrooms', 'Age', and 'Proximity to Downtown (miles)') and the target variable ('Price'). To prepare the data for modeling, I performed feature scaling using StandardScaler() to standardize the feature values, ensuring that all features contribute equally to the distance calculations.

To visualize the relationships between the features and the target variable, I created scatter plots. The Size (sqft) vs Price plot showed a strong positive linear relationship, suggesting that larger houses tend to have higher prices. The Bedrooms vs Price plot did not reveal a clear pattern, indicating a weak relationship between the number of bedrooms and price. The Age vs Price and Proximity to Downtown (miles) vs Price plots showed random distributions, suggesting no noticeable relationships between these features and price.

I also generated histograms to understand the distributions of the features. The Size (sqft) histogram showed a fairly uniform distribution with slight peaks around 3500 and 2500 sq. ft. The Bedrooms histogram revealed that most properties have 1, 3, or 5 bedrooms. The Age histogram had peaks at around 20 years and 50-60 years, with a relatively flat distribution across other ranges. The Proximity to Downtown (miles) histogram showed a somewhat uniform distribution with slight peaks around 10 and 25 miles.

Finally, I calculated the correlation matrix to identify the relationships between features and the target variable. The Size (sqft) and Price had a perfect positive correlation (1.00), while Bedrooms and Price had a very weak negative correlation (-0.05). Age and

Price had almost no correlation (0.01), and Proximity to Downtown (miles) and Price had a very weak negative correlation (-0.03).

Model Development and Evaluation

After preprocessing the data, I split it into training and test sets using train_test_split() with a test size of 0.3 (30% for testing). I created a linear regression model using LinearRegression() and performed feature selection using Recursive Feature Elimination (RFE) to identify the most significant predictors. RFE was fit on the training data using rfe.fit(X_train, y_train), and the selected features were 'Size (sqft)', 'Bedrooms', and 'Proximity to Downtown (miles)'.

I then trained the model on the selected features using model.fit(X_train_selected, y_train) and evaluated its performance on the test set using model.score(X_test_selected, y_test), which yielded an R-squared value of 0.9982. To further assess the model's performance, I calculated the Mean Squared Error (MSE), R-squared, and Adjusted R-squared. The MSE was 128679770.26, the R-squared was 0.9982, and the Adjusted R-squared was 0.9982.

I interpreted the model coefficients to understand the significance of each feature:

- Size (sqft): 278664.85 (positive, significant)
- Bedrooms: 7157.32 (positive, significant)
- Proximity to Downtown (miles): -8515.08 (negative, significant)

These coefficients suggest that larger houses with more bedrooms and closer proximity to downtown tend to have higher prices.

To visualize the model's accuracy, I plotted the actual vs. predicted prices. The points clustered closely around the reference line, indicating good predictive performance.

Challenges and Limitations

While the dataset did not have any missing values, handling missing data can be a challenge in real-world scenarios. Additionally, there may be opportunities to create new features or transform existing ones to improve the model's performance. The high R-squared and Adjusted R-squared values suggest that the model may be overfitting to the training data, and techniques like cross-validation and regularization could help address this issue. Finally, the model's performance on the test set is excellent, but it may not generalize well to new, unseen data, and further validation on additional datasets would be necessary to assess its real-world applicability.

Conclusion

The multiple linear regression model developed using the selected features (Size (sqft), Bedrooms, and Proximity to Downtown (miles)) demonstrates strong predictive power, with an R-squared value of 0.9982 on the test set. However, it's important to note that the model's performance may be influenced by the specific characteristics of the dataset used for training and testing, and additional factors such as neighborhood quality, school districts, and market conditions may also play a role in determining house prices.