Report

Over view of the Machine Problem, Predicting House Prices with Multiple Regression. Using multiple regression approaches, we developed a predictive model to estimate property prices. The intention was to create a model that would be helpful for real estate agents, who could utilize it in alongside their existing knowledge to estimate the price of new homes based on characteristics like the number of bedrooms, age, and location in relation to downtown, among other things. I reached at that technique through exploratory data analysis, which identified the crucial relationships between these variables and housing prices. Then, a comprehensive pre-processing step was executed to handle any missing data and standardize the characteristics for similarity. Then, a multiple linear regression model was created and its performance metrics were described, allowing to assess the model's accuracy using metrics like Mean Squared Error (MSE) and R-squared. It had several difficulties, such as inconsistent column names and missing data, but these were all resolved? an attempt to ensure the model's quality.

Characteristic Statistics

Size (sqft): This one is universal because homes naturally exist in a wide variety of shapes and sizes.
Bedrooms: Compared to house size, this characteristic typically has less detail and a smaller range.
Age: The construction dates of houses can influence the difference in cost.
Miles to Downtown: The closer the location, the more likely it is that the value will equal the price
Price (Output): The training data's output that ought to be correlated with the other attributes.

Data exploration

Exploratory data analysis aims to identify the structure of data and the correlations between parts of it. The correlation matrix revealed an important beneficial relationship between house size and price; other characteristics, such as bedrooms and age, were not well correlated or had stronger or opposite relationships. Scatter plots revealed expected price patterns: larger houses are generally more expensive, while proximity to downtown had little positive correlation with price. The feature-based data analysis assisted in identifying learn patterns and detecting potential data problems.

Data preprocessing

The data preprocessing part involved the ways to deal with missing values and normalizing features. The missing values were managed by calculating the gaps and filling them with mean values while the feature scaling was done using normalization in the form of StandardScaler. I think that it is a necessary part of the dataset preparation due to the fact that it allows getting access to all features on the same level and scale. There is no need for any categorical encoding since there were no categorical features in the dataset, which consisted only of numerical features.

Model Development

The dataset's characteristics were used to create a linear regression model with multiple variables for predicting housing prices. A 70% training set and 30% test set were used to divide the data. The original feature selection included all of the options available. Further feature selection can enhance the model's accuracy. Preprocessing was critical since characteristics were standardized and splitting was carefully done.

Model Evaluation

One way to determine how well the model performs is to evaluate the difference between the actual and predicted prices. In this scenario, it could be evaluated with Mean Squared Error and R-squared. MSE shows the average squared error between the predicted and actual prices. R-squared shows how well the regression line explains the variation of prices. It is also possible to visually evaluate the accuracy of the model if the actual prices and predicted prices are shown on the same graph.

Challenges

A number of challenges were faced, including inconsistencies in column names and an absence of data. Differences in named columns were resolved by checking the dataset and then changing the names properly. The mean values computed from the remaining valid data for the individual record were utilized to fill the empty slots. These techniques ensured that no biases were created throughout the data beforehand procedures and that the dataset's quality was not affected.

Conclusion

The linear regression model accurately predicts house values based on factors such as size, number of bedrooms, age, and distance from downtown. While it provides helpful information for real estate professionals, its usefulness is limited by its basic characteristics and assumptions about linear relationships. Improving accuracy in the future could involve bringing together characteristics and getting into more advanced models to identify complex patterns in the data.