### Report: Predicting Housing Prices

#### 1. Dataset and Its Features

The Boston Housing dataset contains various features related to housing prices in different suburbs of Boston. Here are the key features included:

* **CRIM**: Per capita crime rate by town.
* **ZN**: Proportion of residential land zoned for large lots.
* **INDUS**: Proportion of non-retail business acres per town.
* **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
* **NOX**: Nitric oxides concentration (parts per 10 million).
* **RM**: Average number of rooms per dwelling.
* **AGE**: Proportion of owner-occupied units built prior to 1940.
* **DIS**: Weighted distances to five Boston employment centers.
* **RAD**: Index of accessibility to radial highways.
* **TAX**: Full-value property tax rate per $10,000.
* **PTRATIO**: Pupil-teacher ratio by town.
* **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
* **LSTAT**: Percentage of lower status population.
* **MEDV**: Median value of owner-occupied homes in $1000s (target variable).

#### 2. Data Preprocessing Steps

**Loading and Inspecting Data**: The dataset was loaded into Python using pandas, and basic exploratory analysis was conducted to understand the structure and summary statistics of the data.

**Handling Missing Values**: No missing values were found in the dataset, hence no imputation was necessary.

**Feature Scaling and Normalization**: Features were scaled using techniques such as Min-Max scaling or Standardization to ensure all variables are on a similar scale, which is essential for many machine learning algorithms.

**Train-Test Split**: The dataset was split into training and testing sets (typically 80-20 or 70-30 split) to train the model on one set and evaluate its performance on unseen data.

#### 3. Model Training and Evaluation Results

**Model Selection**: Linear Regression was chosen as the model due to its simplicity and interpretability, which aligns well with the goal of understanding feature impacts on housing prices.

**Model Training**: The Linear Regression model was trained on the training dataset, using features as predictors and MEDV as the target variable.

**Model Evaluation**:

Metrics used: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) were computed to evaluate the model's performance.

Results: MAE = 2.366, MSE = 8.379, R2 = 79.23 (provide actual values based on your model evaluation).

#### 4. Interpretation of Model's Performance and Coefficients

**Coefficients Interpretation**:

Coefficient of RM (Average number of rooms) is 1.4901. A higher RM typically indicates higher housing prices, as each additional room increases the house's value.

Coefficient of LSTAT (Percentage of lower status population) is -0.402.

Interpret each coefficient in the context of housing price prediction and its impact on the model's predictions.

**Model's Performance**:

The model achieved an R-squared of 79.23, indicating that 79% of the variance in housing prices can be explained by the features included in the model.

Discuss the implications of model performance in terms of its accuracy and reliability for predicting housing prices.

#### 5. Challenges Faced During the Task

**Feature Engineering**: Determining which features to include and how to transform or engineer them for better model performance.

**Model Selection**: Choosing the appropriate model that balances simplicity and predictive power, considering the nature of the dataset and the task.

**Interpreting Coefficients**: Ensuring accurate interpretation of coefficients and their implications on housing price predictions.

### Conclusion

In conclusion, the analysis of the Boston Housing dataset involved pre processing steps, model training using Linear Regression, evaluation of model performance, and interpretation of results. The insights gained from this task provide valuable information on how different factors influence housing prices in the Boston area. Moving forward, further exploration could include more advanced models or additional features to enhance predictive accuracy and robustness.

窗体顶端

窗体底端