**Boston Housing Report**

**Boston Housing Dataset Analysis**

**1. Dataset and Its Features**

The Boston Housing dataset consists of various features related to housing in the Boston area. The features are as follows:

* **CRIM**: Per capita crime rate by town.
* **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft.
* **INDUS**: Proportion of non-retail business acres per town.
* **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
* **NOX**: Nitric oxide 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 Black residents by town.
* **LSTAT**: Percentage of lower status of the population.
* **MEDV**: Median value of owner-occupied homes in $1000s.

**2. Data Preprocessing Steps**

Data preprocessing included the following steps:

* **Log Transformation**: Applied log transformation to skewed features to reduce skewness and make the data more normally distributed.
* **Outlier Handling**: Identified and capped outliers in features to the upper and lower bounds based on the interquartile range (IQR).

**3. Model Training and Evaluation**

The dataset is split into training and testing sets. Six different models are trained on these split dataset, and the model is evaluated based on the following metrics:

* **Mean Absolute Error (MAE)**
* **Mean Squared Error (MSE)**
* **R-squared (R²)**
* **Adjusted R-squared**

Here are the results from the model evaluation:

Train score : 0.8938931512384102

Test score : 0.8516679363182201

r2 score: 0.8516679363182201

Adjusted\_r2\_score: 0.8480574195551748

MAE: 2.2252221298767836

MSE: 7.257138801879341

**4. Interpretation of the Model's Performance and Coefficients**

AdaBoost does not provide coefficients and intercepts as linear regression does. Instead, it provides feature importances, which indicate the relative importance of each feature in making predictions. The higher the feature importance, the more significant the feature is in predicting the target variable (housing prices).

Linear Regression coefficients and intercepts

Feature Coefficient

0 CRIM -1.113643

1 ZN 0.210993

2 INDUS -0.584474

3 CHAS 0.404992

4 NOX -1.386081

5 RM 1.713043

6 AGE -0.121816

7 DIS -2.647300

8 TAX -0.063869

9 PTRATIO -1.176908

10 B 0.272759

11 LSTAT -4.082870

Intercept: 21.848754127602746

AdaBoost Feature importance

Feature Importance

11 LSTAT 0.396239

5 RM 0.150850

7 DIS 0.112488

9 PTRATIO 0.086496

4 NOX 0.072783

0 CRIM 0.070417

8 TAX 0.058101

2 INDUS 0.036627

6 AGE 0.015773

10 B 0.000203

1 ZN 0.000022

3 CHAS 0.000000

**5. Challenges Faced**

* **Data Skewness**: Some features had significant skewness which required log transformation.
* **Outliers**: Identifying and handling outliers was necessary to improve model performance.

By focusing on these aspects, the model's performance was enhanced, resulting in better predictions of housing prices in the Boston area.

* All The Visualization graph are added to the Notebook
* AdaBoost Actual vs Predicted graph



**Conclusion**

This brief report summarizes the key steps and findings from the analysis of the Boston Housing dataset. The preprocessing steps, model training, and evaluation metrics provided insights into the relationships between various features and housing prices, as well as the challenges encountered during the task.