REPORT

SUMMARIZING THE DATA AND ITS PROCESSING

1. The dataset and its features:

- **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 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 centres
- **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 -** % lower status of the population
- **MEDV -** Median value of owner-occupied homes in $1000's

2. Data preprocessing steps:

- **DATA TYPE -** Changing the data type of the columns from object to float for analysing the features and performing operations.
- **HANDLE OUTLIERS -** Checking for the outliers columns and capping the outliers for most of them.
- **CHECKING FOR CONSTANT COLUMNS –** Checking the dataset if any constant column is present so that I can remove them.
- **CORRELATION –** Checking the correlation of the columns and removing the related columns.
- **SKEWNESS –** Checking the skewness of the columns.
- **TRANSFORM –** Transforming the columns so that the model performs better on them.
- **TRAIN-TEST –** Splitting the data into Training and Test set.
- **SCALING** – Scaling the data using Standard Scaler.

3. Model training and evaluation results:

- **LINEAR REGRESSION –** Training Linear Regression model which gave a r2 score of 0.63.
- **SVR –** Training Support Vector Regressor on the data gave us r2-score of 0.73.
- **XGBOOST –** Applying XgBoost on the training data gave an adjusted r2-score of 0.73 on test data.
- **RANDOM FOREST REGRESSOR –** Random Forest on applying on the data performed a result with adjusted r2-score of 0.75.
- **HYPER-PARAMETER TUNING –** Applying Grid Search CV on the data for Random Forest gave a result of adjusted r2-score 0.73 which is less than the normal Random Forest result without tunning. So, neglect the tunning.

4. Interpretation of the model's performance and coefficients:

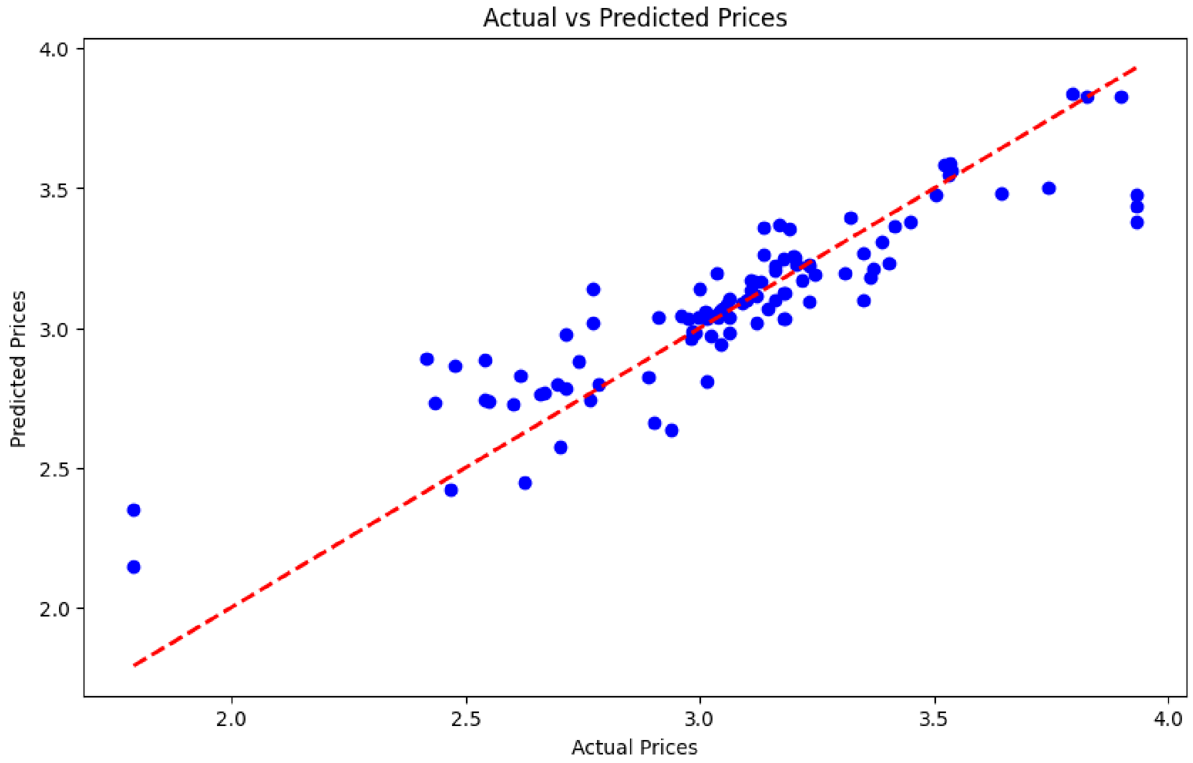
- Hence, I concluded Random Forest as my final model whose adjusted r2-score was 0.75.
- I have checked for important features using Random Forest through which I came across the fact that some features were not participating much in output generation.
- After removing those columns and again training the model, the adj. r2-score increased to 0.77

5. Any challenges faced during the task:

- The only challenge I faced during the task is that the result of Hyper Parameter Tuning degraded the performance of the model.

ALL THE VISUALIZATION AND GRAPHS ARE ADDED TO THE NOTEBOOK.

FINAL ACTUAL VS PREDICTED PRICE GRAPH:



Ankit Raj
SIC-**21BCSB32**
Silicon Institute of Technology