BOSTON HOUSING PRICE PREDICTION MODEL

Introduction

Using the provided dataset, to create a machine learning model to forecast the price of a home.

Before moving on to the model construction, where the data is trained and tested, the given dataset has undergone data pre-processing and a thorough exploratory data analysis.

Data Description

The Boston Housing Dataset is derived from data gathered by the U.S. Census Service about housing in the Boston, Massachusetts, area. The columns of the dataset are described 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 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

Approach

Before beginning to develop a model, it is necessary to pre-process the data and perform the necessary visualisations in order to better comprehend the data.

Building a regression model is preferable for this housing price prediction model in order to improve prediction and accuracy.

Visualization

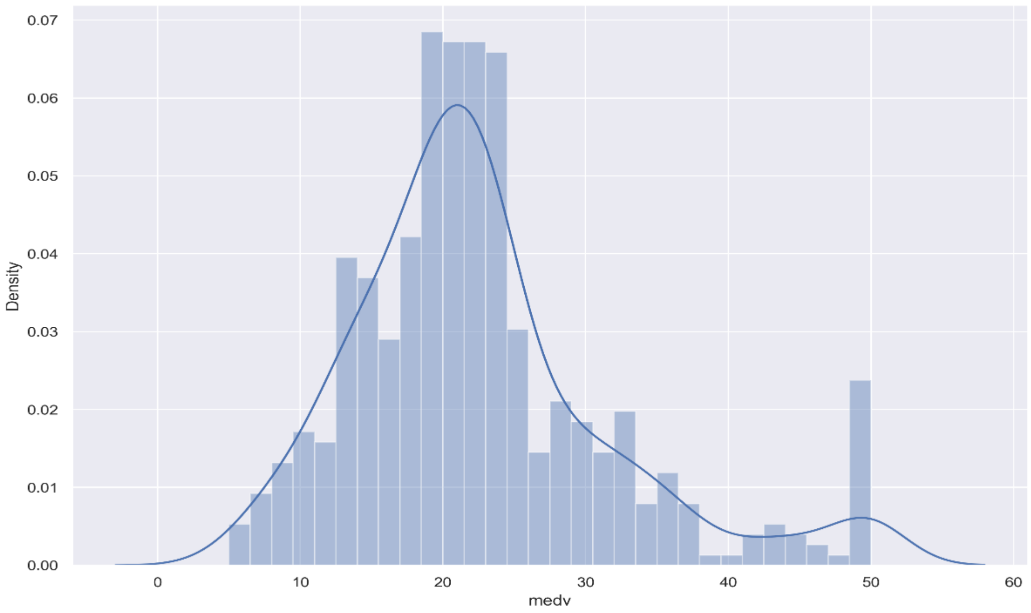

Figure 1 - "MEDV" Distribution

The above chart shows the distribution of the “MEDV” column, which is the target variable. The distribution seem to a normal distribution and it is a regression model.

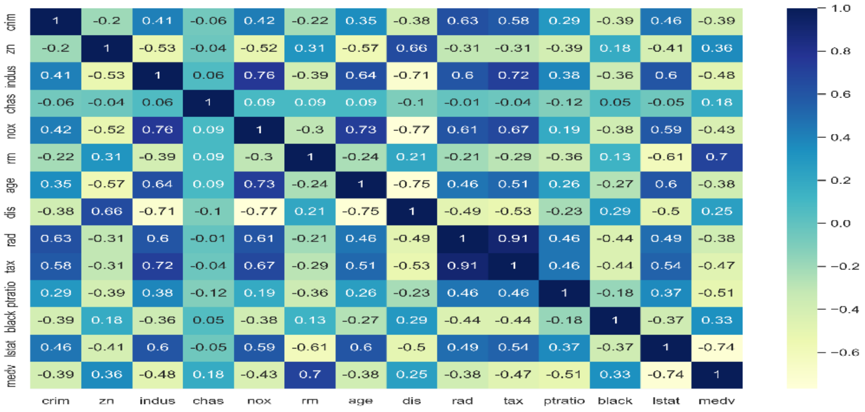

Figure 2 Heatmap

- From the above heatmap we can see that **Lstat and RM** are strongly correlated with **MEDV**
- **TAX** and **RAD** are strongly correlated together, so we don't include them in feature selection to avoid multicollinearity

Algorithm

As mentioned earlier, it is a machine learning algorithm based on supervised learning is linear regression. It executes a regression operation. Regression uses independent variables to model a goal prediction value. It is mostly used to determine how variables and forecasting relate to one another. Regression models vary according to the number of independent variables they use and the type of relationship they take into account between the dependent and independent variables.

Evaluation

R Square – The model is evaluated using the r square value.

The r square value for the **training model** is **0.6391406783112359**

The **RMSE** value is **5.505471273611083**

The r square value for the **testing model** is **0.6305371137570698**

The **RMSE** value is **5.640085632223884**
The r square value of both training and testing model is around 65% and therefore the model is a good fit.

When assessing the effectiveness of a regression-based machine learning model, one of the most crucial metrics is the R2 score. Also known as the coefficient of determination, it is pronounced as R squared. To measure the variance in the predictions that the dataset can explain.

Result

The developed prediction model has a decent r score and so it is a good fit model.

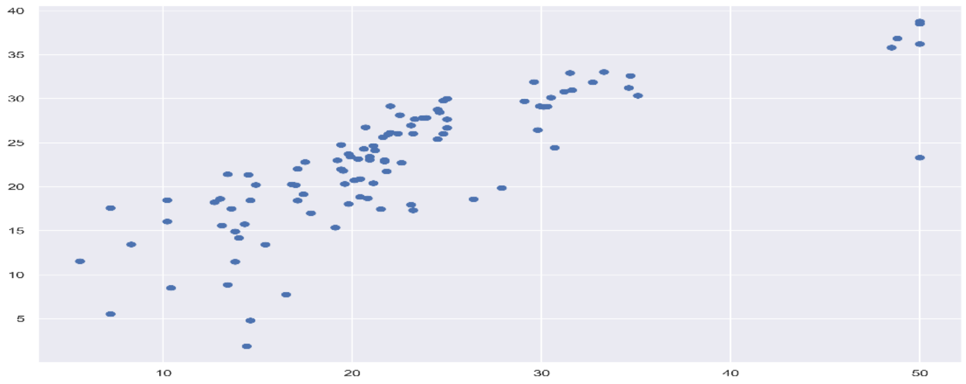

Figure 3 Scatterplot

The above scatterplot shows the testing and prediction values. Ideally it should be a straight line for a good model.

Conclusion

The Boston House Price Prediction model has been developed using the provided dataset which is collected from the US. The developed linear regression model proved to be good fit with a decent r squared value which can used for future predictions of the house price at Boston.

Future Work

The model's efficiency can be improved by adding more data. With the help of more data, the accuracy of prediction will be increased significantly.