**Boston House Price Prediction using Machine learning**

By

Mohammed Suraj Shaik (U00846116)

**Introduction**

House price analysis will be beneficial to real estate sectors where important features related to the house can be analyzed with the help of machine learning algorithms. This will be beneficial to corporate sectors and real estate agencies where they could fix and decide the price of a house based on important attributes such as number of bedrooms, lot area, parking space, etc. Machine learning is capable to extract important patterns from the data where the house price can be predicted based on different features.

The project is focused on building house price prediction model with the help of machine learning algorithms where regression techniques will be used to predict the house price based on different features of the data.

The model will be helpful in real estate sectors and other construction companies where the value of the house can be predicted based on the features related to it and it will help in maximizing the profits of particular business which are concerned with housing sectors.

**Solution**

* **Data Preparation**

Data preparation includes treating of outliers and imputation of the missing values if present in the data. Outliers affect the accuracy of the models and also the correlation of features with the house price will be analyzed in this phase. R programming will be used for the entire house prediction analysis

* **Model Building**

Model building includes implementation of machine learning models in predicting the outcome where the data is splitted into training and validation. Regression algorithms such as Ridge Regression, lasso Regression and Multiple Linear Regression will be applied to predicting the house price.

* **Model Evaluation**

For evaluation of the model, different performance metrics such as mean squared error, root mean square error and R square score will be implemented to evaluate the errors generated during the prediction.

* **Model tuning**

Model tuning includes testing of different parameters that are involved in each machine learning models where the most effective parameters will be included in the final prediction with the help of model tuning techniques.

**About the Data:**

The data is collected from Kaggle which is known as Boston housing data.

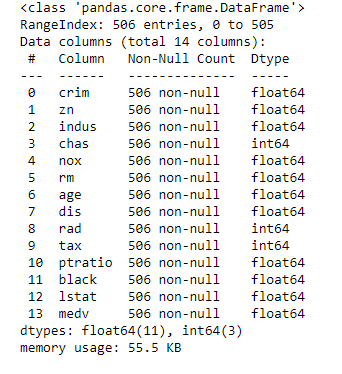
The data contain 506 attributes along with 14 columns where the target variable contains the median price of Boston houses and is continuous in nature. All the features are numeric in nature and the description of all the features are given below.

The Boston data frame has 506 rows and 14 columns.

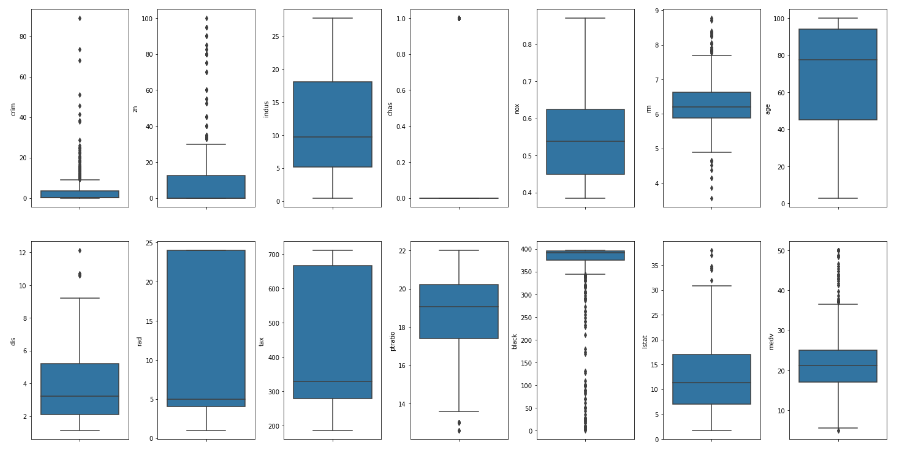
1. CRIM - per capita crime rate by town.
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. NOX - nitrogen oxides concentration (parts per 10 million).
6. RM - average number of rooms per dwelling.
7. AGE - proportion of owner-occupied units built prior to 1940.
8. DIS - weighted mean of distances to five Boston employment centres.
9. RAD - index of accessibility to radial highways.
10. TAX - full-value property-tax rate per \$10,000.
11. PTRATIO - pupil-teacher ratio by town.
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
13. LSTAT - lower status of the population (percent).
14. MEDV- median value of owner-occupied homes in \$1000s.

**Empirical Experiments**

* **Data Exploration**



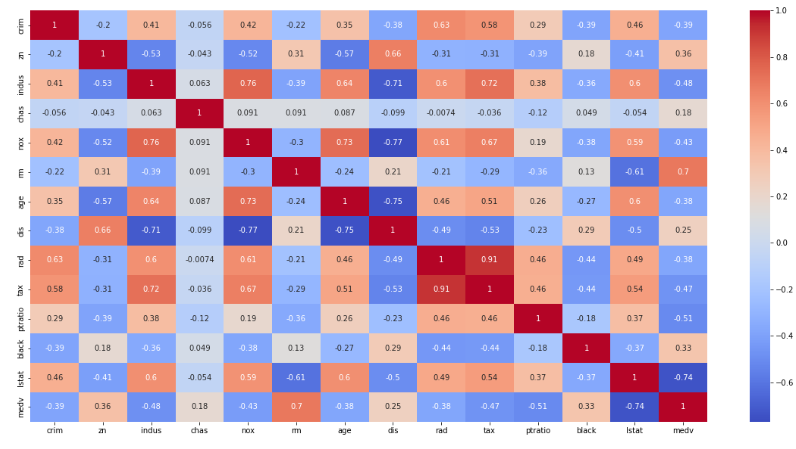
There are 14 columns in the database do not contain any missing values and all the columns are numeric in nature. None of the values are categorical and the output of the value in the median value of the house.



From the data exploration we can see that the features such as ZN and the features such as CRIM contain outliers in most of the cases and the distribution of the data is skewed. Also, the median value of the house which is the output variable contain outliers where the values are above 40

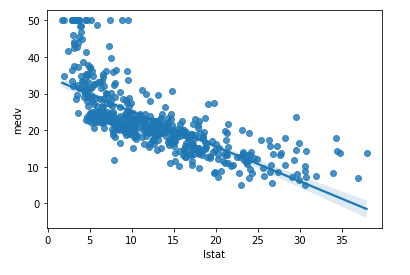
.

**Feature Correlation:**

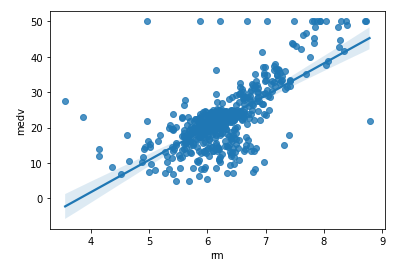


* The features such as DIS and INDUS, DIS and NOX and DIS and age has high negative correlation.
* Also, the features such as NM and LSTAT, output and LSTAT have high negative correlation.
* The feature such as CRIM and tax, Indus and tax, rad and tax have high positive correlation.
* The features such as NM without output variable MEDV also have good positive correlation.

**Feature Visualization:**



The plot indicates that increase of median value of the houses decreases the lowest status of the population.



The median value of the house and the average room per dwellings show a positive correlation as the median value of the house is higher in case of higher average number of rooms per dwelling.

**Model Results:**

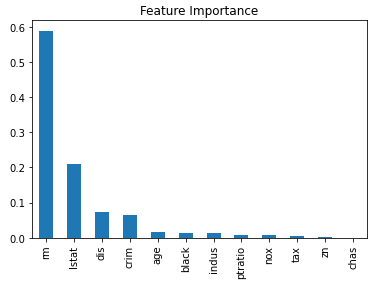
Various machine learning algorithms are used to predict the house price and the cross-validation score is calculated along with the mean square error. The values of different types of algorithms used are given below

|  |  |  |
| --- | --- | --- |
| Model | MSE | Cross validation score |
| Lasso Regression | 72.26 | 96.03 |
| Ridge Regression | 26.37 | 36.41 |
| Elastic Net Regression | 72.26 | 96.03 |
| Ridge CV | 23.56 | 32.43 |
| Decision Tree Regressor | 21.81 | 39.20 |

From the above comparison of the models, the best cross validation score given by Lasso and Elastic net regression. Although both the regression are based on regularization parameters, they performed better compared to other regression techniques in predicting the house price.

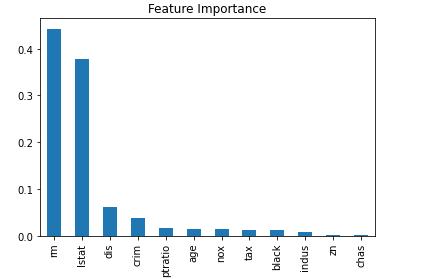
**Feature Importance:**

1. **Decision Trees Regressor**



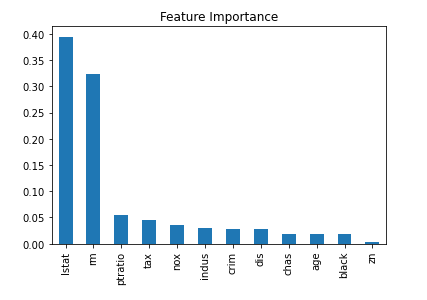
As per the decision tree, the features such as nm, lstat, dis, crim are the highest important features in predicting house price.

1. **Random Forest Regressor**



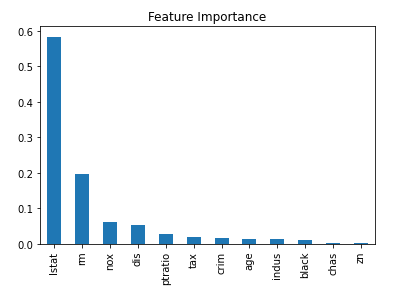
The random forest gives similar feature importance compared to the decision tree where the features such as nm, lstat, dis and crim are the most important features in predicting the house price.

1. **Extra Trees Regressor**



As per the extra trees regression the feature importance indicates features like lstat, nm, ptratio, tax, nox are the important features in predicting house price. Nm and lstat are the common important features found in decision tree, random forest and extra trees regression.

1. **XGBoost Regressor**



The features such as lstat, nm, nox and dis are the important features in predicting house price as per the XGBoost regression model.

**Conclusion:**

Lasso and Elastic Net Regression are found to be the most suitable algorithms to predict the house price and according to feature importance tests, most of the algorithms gave important features such as lstat, nm, nox, dis and crim are the most important features responsible to predict the house price. The outcome of the variables contain outliers which should be removed in future work. All the data size is small, outlier removal techniques might reduce the size of the data that will also reduce the importance of features in predicting the house price.