**PROJECT PROPOSAL**

COMP 7118 Data Mining Spring 2022

Mohammed suraj shaik U00846116

Contact Email : mshaik1@memphis.edu

**Boston House Price Prediction using Machine learning**

**What is the problem/application?**

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.

**What methods will be tested or implemented?**

* **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.

**What data sets will be used**

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. TARGET - median value of owner-occupied homes in \$1000s.

**Deliverables**

* **Background Research**

This part will contain the background research of house price and how it impacts in real life problems. Also the aims, objective and the research question will be properly explained.

* **Literature Survey**

This will contain existing works on house price prediction analysis using machine learning approaches. The literature gap will be studied and evaluated.

* **Methods**

This will contain the methods and procedures applied in the project. All the machine learning algorithms implemented in R and will be explained in detail.

* **Results and analysis**

This will contain all the results including the error, top features, and other analysis achieved from the code. All the models will be implemented in R studio where R programming language will be applied.