<https://github.com/Learner2020-J/KC-House-Price-Prediction>

**INTRODUCTION**

Applications of Machine Learning in Various Fields have helped a lot. Nowadays it is used in almost every customer service industry, increasing business profit, forecasting of weather. In similar way we can use it for prediction of House Prices for different Customers.

We will use the technique of Linear Regression to predict. The idea of regression is pretty simple: given enough data, you can observe the relationship between your target parameter (the output) and other parameters (the input), and then apply this relationship function to real observed data.

We are using machine learning as it is helpful to get the insights out of data. As, the price of house depends on lots of variables and factors which are many times not taken into consideration while purchasing the house.

Using different variables one can find out the relationships among different variables which help to determine the price of house. Lots of Real Estates Company Try to look for Potential buyers who are looking for lots of factors Prices, Square foot area of House, No of Floors, Bedrooms, Bathrooms and a lot of other variables. To get proper Customer and appropriate price we are performing this Machine Learning Study using the Linear Regression Technique.

**EXPLORATORY DATA ANALYSIS**

Here, we do the Exploratory Data Analysis of the House Price Data set in Jupyter Notebook. We perform different operations on the data set of house price and try to find out different insights. We find out the kind of variables present like Categorical and Numerical. Then does the Data contains any kind of Missing Values, Null Variables, and Duplicated Values etc. Then we try to find out the Statistical quantities related with the variables and different relationship among each other.

We also look for outliers in the data set. As lots of time this outliers values doesn’t help to get the proper statistical distribution.

We then perform Uni-variant and Bi-Variant Analysis of different parameters. Then plot different variables in different kind of plots and try to seek the relationships among each other. Then we try to find out the correlation among different variables. We then see whether the variables are normal distributed or not. The distribution helps us to know whether the variables are normal distributed or any kind of skewness is present. Relationship among different variables and different plot helps us to get the more knowledge.

Then we separate the categorical and numerical variables and perform different data operation on the datasets. After this we create correlation-matrix and plot the Heat Map to find out the relationship among each other and with other variables.

**ANOMALY DETECTION**

Anomaly Detection is the process of identifying unexpected items or events in data sets, which differ from the normal. In datasets pre-processing often there are certain data where some data might contain anomaly whose removal helps in getting a better accuracy and statistical insights.

Anomalies are also referred as Noise, outliers, Deviations, Exceptions. In this house price data sets there are some anomaly and some variables which have been excluded while performing the EDA. In the model we need to consider only those variables which are useful for the proper price predictions.

Like id, date, zip-Code, yr\_renvoated, yr\_built, lat, long etc.

**MODEL BUILDING**

After doing the EDA and Anomaly Detection model building phases is taken into consideration. The model building we use different Python Libraries for the model building operations. Especially Sckit-learn, numpy, pandas, seaborn, matplotlib, Scipy etc.

We then plot the variables in the normal distribution and then we do the antilog transformation. After that we find out the different scores of MAE, MSE, RMSE, R2 etc. Then we use find out the MAE, MSE, RMSE, and R2 from the train data.

Then we perform the same operation on test data and validation data and find out the MAE, MSE, RMSE, R2 etc.

**CONCLUSION**

Then, we find out the Train results, Valid Results & Test Results

Train Results

R squared 0.8765613103211521

Mean Absolute Error: 628927.862437973

Mean Squared Error: 1288742623779436.2

Root Mean Squared Error: 35899061.600262426

Validation Results

R squared 0.8726594724919401

Mean Absolute Error: 74084.07380079047

Mean Squared Error: 16007433073.085873

Root Mean Squared Error: 126520.48479628061

Valid Accuracy 0.8748868587230378

Mean Absolute Error: 75679.50116903569

Mean Squared Error: 16629365581.845093

Root Mean Squared Error: 128954.89747134497