Capstone Project 1

HOUSING PRICE PREDICTION

N BHARATH | Mentor – SIDDHARTH DIXIT

HOUSING PRICE PREDICTION

OVERVIEW

This project is to predict the housing prices in the given area considering various elements like whether the house contains car garage, swimming pool and how many bedrooms it contains and what is the dimensions of the building etc.

CLIENT

This project does not have a definitive client. But the analysis performed could be of use to anyone in the Real Estate Business (House Owners, Buyers, Tenets, etc).

If we ca make use of the data to find some insights from the past data and could possibly predict the future, many House Owners and Tenets could use this recommendation system on their own. Giving good recommendations directly entails one or many of the following:

1. Customers use the platform more frequently due to the quality and relevance of content shown to them.
2. Better User Experience. Customers do not relay on brokers and will search on their own according to their need and deed.

DATA

The data used in this project has been obtained from Kaggle and it is available in csv (Coma Separated File) format, the data set consists of 1460 instances of training data and 1460 of test data. Total number of attributes equals 81, of which 36 is quantitative, 43 categorical + Id and Sale Price.

DATA WRANGLING

This section describes the various data cleaning and data wrangling methods applied on the Movie datasets to make it more suitable for further analysis. The following sections are divided based on the procedures followed.

Cleaning

The dataset had a lot of features which had 0s for values it did not possess. These values were

converted to NaN. These NaN values are filled in two ways according to the type of the data

1. Categorical type was filled with the mode values.
2. Numerical type was filled with mean values.

Removing Unnecessary Features

This process was done in different ways

Number of missing values

If a particular feature has more that 50% of missing values in it then most of the times that particular feature will not play any significant role in learning the model. These features must be removed and here features like 'Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature' has more than 50% of missing values and has been removed

Significance

From co-relation we get to know features that are important for predicting the output and after performing co-relation features which has values greater than 0.7 are carried forward and remaining were dropped

Outliers

For any feature in the data whose Z-Value is > 3 those values are called as Outliers and Removing those values will help a lot in predicting the accurate output and here values which has Z-Value > 3 has been removed.

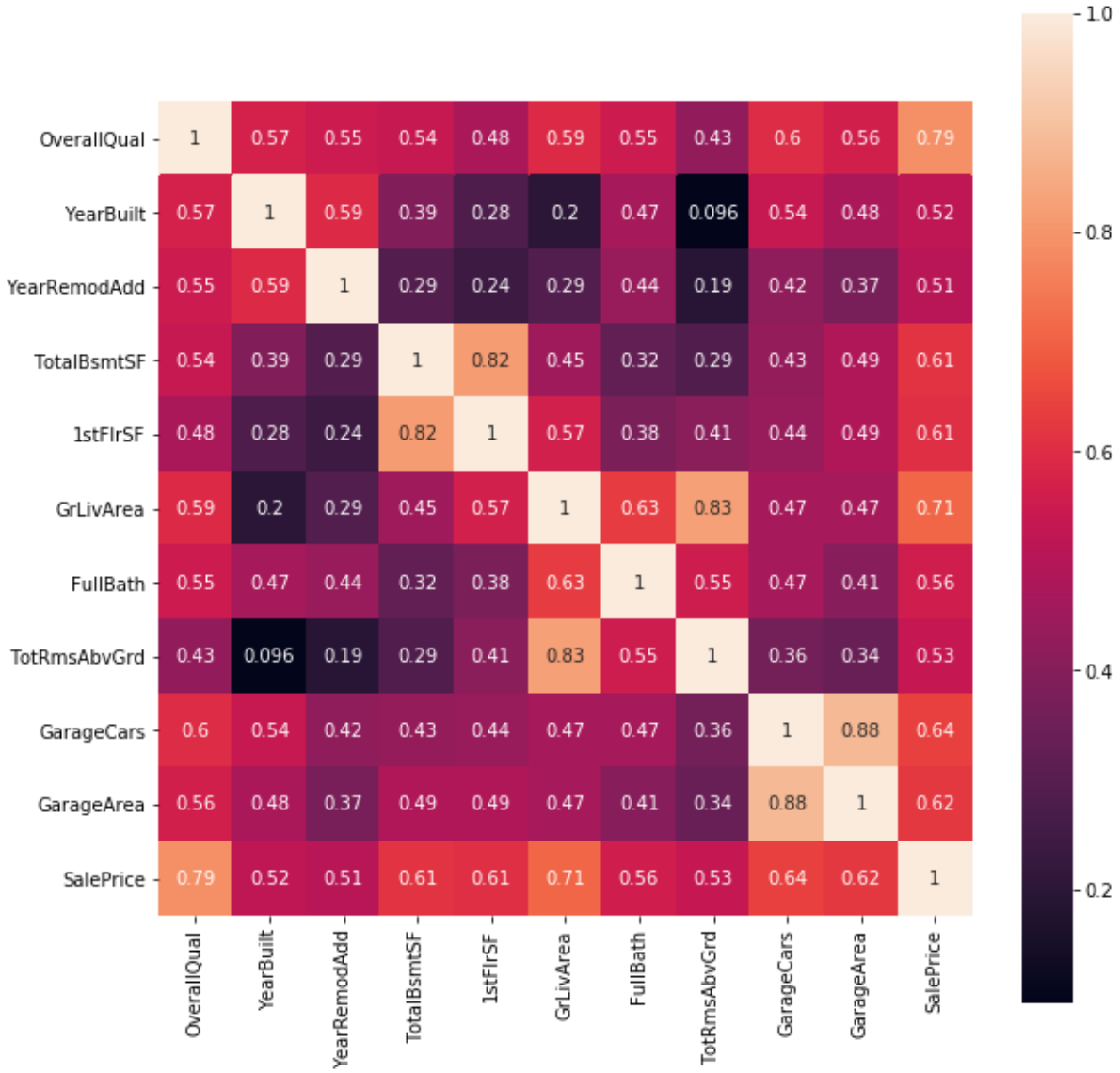
EDA (Exploratory Data Analysis)

This step helps us to find out important features of the data, and relation between dependent and independent variables of the data. It is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a data set. EDA is performed in order to define and refine the selection of feature variables that will be used for machine learning.

EDA of this project is started with finding important features of the data and it is done by using Heat map

Heat Map

The heatmap is the best way to get a quick overview of correlated features



The above heat map shows that

1 OverallQual', 'GrLivArea' and 'TotalBsmtSF' are strongly correlated with 'SalePrice'.

2 'GarageCars' and 'GarageArea' are strongly correlated variables.

3 'TotalBsmtSF' and '1stFloor' seem to be correlated with each other.

4 'TotRmsAbvGrd' and 'GrLivArea' also seem to correlated with each other

Pair Plot

Although we already know some of the main figures, this pair plot gives us a reasonable overview insight about the correlated features

1. One interesting observation is between 'TotalBsmtSF' and 'GrLiveArea'. In this figure we can see the dots drawing a linear line, which almost acts like a border. It totally makes sense that the majority of the dots stay below that line. Basement areas can be equal to the above ground living area, but it is not expected a basement area bigger than the above ground living area.

2. One more interesting observation is between 'SalePrice' and 'YearBuilt'. In the bottom of the 'dots cloud', we see what almost appears to be an exponential function. We can also see this same tendency in the upper limit of the 'dots cloud'
3. Last observation is that prices are increasing faster now with respect to previous years.

REGRESSION: PREDICTING HOUSE PRICES

Predicting Movie Revenues is an extremely popular problem in Machine Learning which has created a huge amount of literature.

Feature Engineering

1. Converting all categorical features into Booleans using One Hot Encoding.
2. Removing feature variables that are co-related to each other like 'GarageCars' and 'GarageArea', 'TotalBsmtSF' and '1stFloor', and 'TotRmsAbvGrd' and 'GrLivArea' are strongly co-related to each other, so one of these features has to be removed.
3. Converting features which are not normally distributed into normal distribution using logarithmic transformation such as ‘SalePrice’, 'GrLivArea', ‘TotalBsmtSF’.

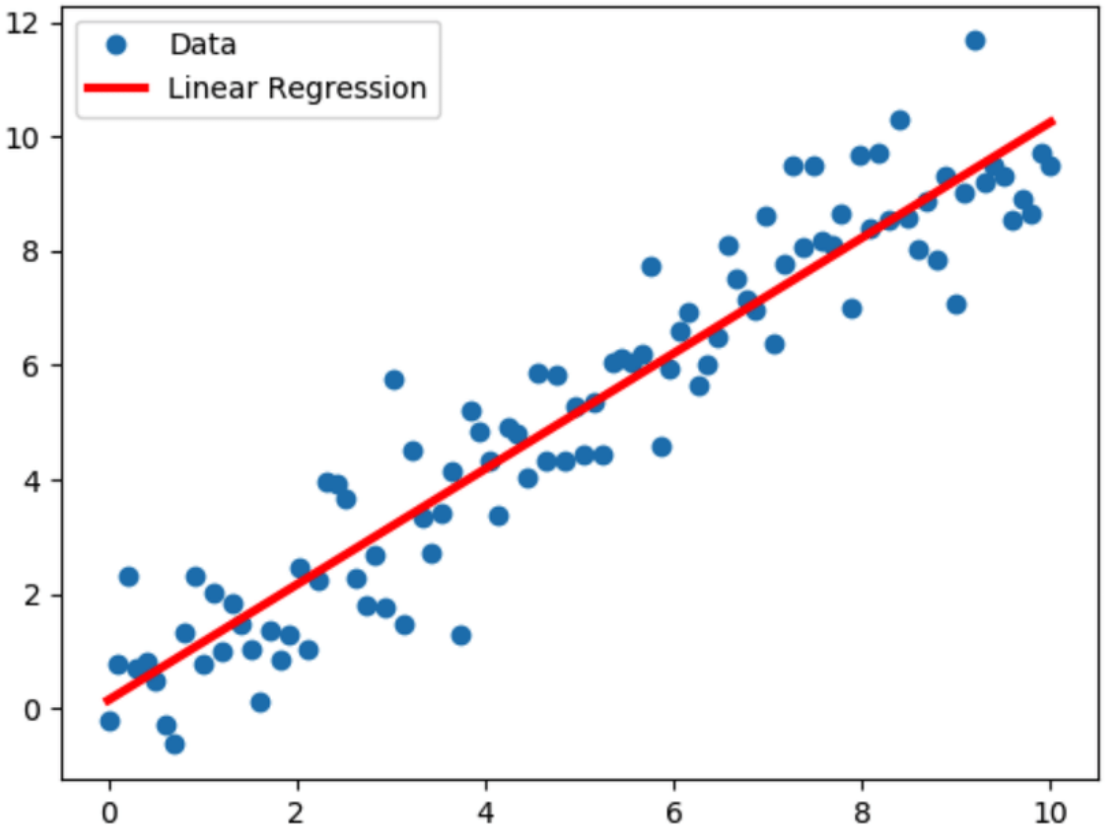
MODEL

Here to solve this problem I have used different regression models.

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable.

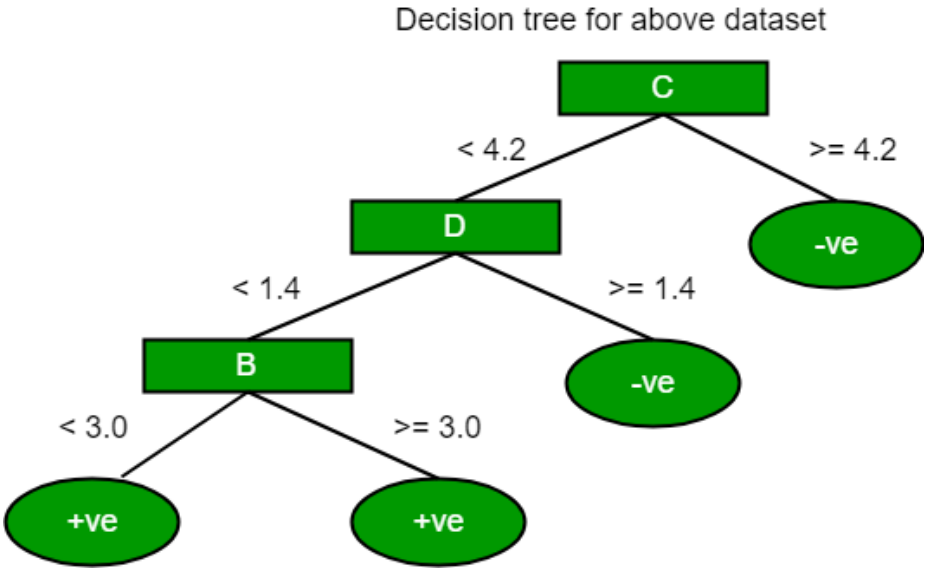
A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept



By using Linear Regression model accuracy, I have got is **89.1**

Decision Tree Regression

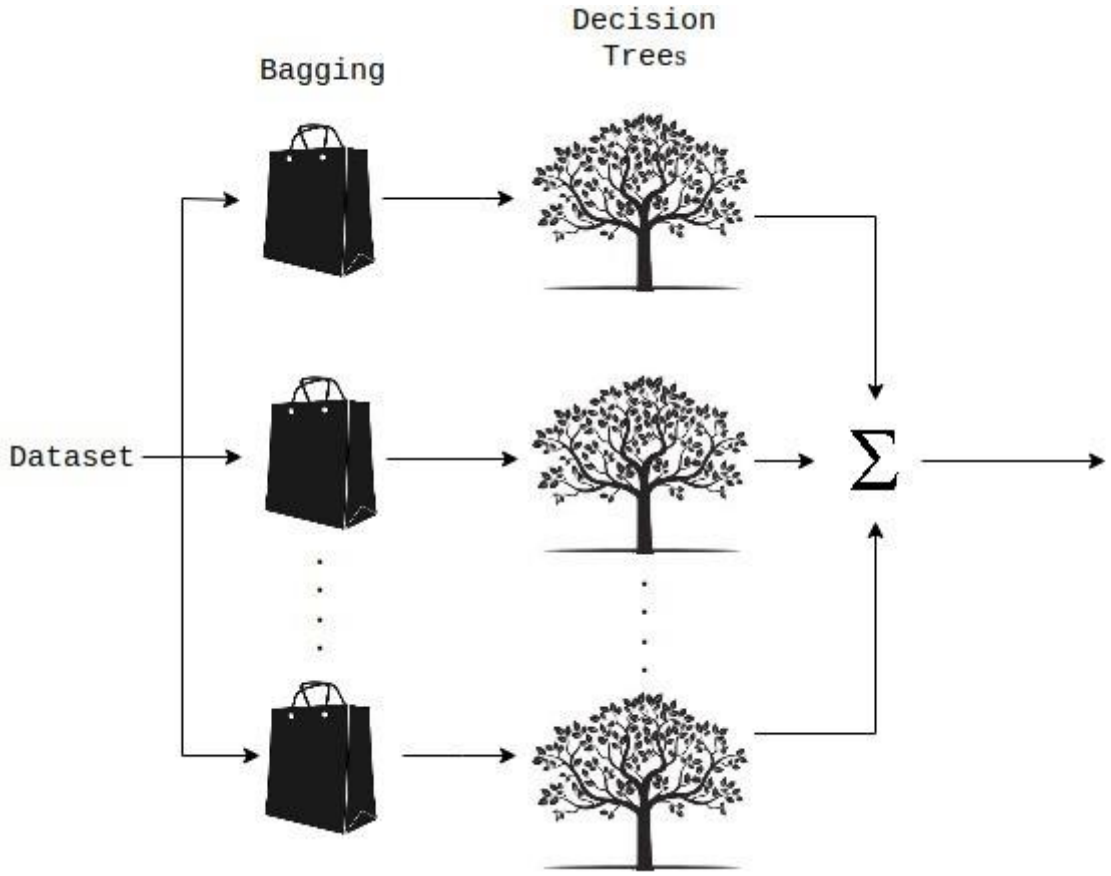
Decision tree algorithm is classification algorithm under supervised machine learning and it is simple to understand and use in data. The idea of Decision tree is to split the big data(root) into smaller(leaves)



By using Decision Tree Regression model accuracy, I have got is **82.9**

Random Forest

Random forest is collection of trees(forest) and it builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used for both classification and regression problems. Example: Suppose we have a bowl of 100 unique numbers from 0 to 99. We want to select a random sample of numbers from the bowl. If we put the number back in the bowl, it may be selected more than once.



By using Random Forest model accuracy, I have got is **88.2**

CONCLUSION

This report highlighted the processes of data wrangling, inferential statistics, data visualization, feature engineering and predictive modelling performed on the Housing Dataset. All the results and insights gained as part of these processes were also highlighted. With these insights, Linear Regression, Decision tree and Random Forest models were built to predict House prices with a Score of **0.891, 0.829** and **0.882** respectively.