

HOUSING: PRICE PREDICTION

Submitted by:

PAVANKUMAR BOLLI

**ACKNOWLEDGMENT**

I would like to express my special thanks to flip robo technlogies for given me such a wonderfull of apportunity to explore and insite work on a data science projects.

And and also thanks to Datatraiend institute’s mentors to motivate and understand the concepts of Data science. I have got lot of help from their resourses and concept to complete these project.

Link that helps me to complete this project

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>

**INTRODUCTION**

* Business Problem Framing:-

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate

market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market

and there are various companies working in the domain. Data science comes as a very important tool to solve problems

in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and

focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling,

recommendation systems are some of the machine learning techniques used for achieving the business goals for housing

companies. Our problem is related to one such housing company.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses

data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same

purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file

below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model

using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest

in them or not.

* Business Problem Framing

Now days in the field of business’s. Real estate business have a major role and also have huge profit and revenue.

Now price of houses are depend’s on various aspects. Like depends on the cost of living, surrounding area, cleaness, Lot area in sqr fit, number of floors, type of neighbour’s are present and also depends on city etc.

Based on these feature’s price of house’s are decided. Before investing a money in this field we need to predict the price of house and there future scope into that area/location.

* Conceptual Background of the Domain Problem

The company has collected a data set from the sale of houses in Australia.

* Review of Literature

Type of houses

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

Motivation of this project is to determine which feature are affecting much to predict the sale price. Here, there are number of feature and motive is to extract a those feature’s which is more usefull for to predict the sale price of the house.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this sample data have soo many feature’s. Some of feature’s have heighly correlated to our target column i.e Sale price. Data set have both categorical and continues data. Training data set contains 1168 observations and have 80 features with one target column. Some of the continues column’s have skewness. Categorical features have two type ordinal and nominal data. Data set have null values also. Soo, i have handel the null values present in the data set. Handle the skewness columns and convert into normalized column by using Power transformer. I have used to encode ordinal and nominal data seperatly. Firstly i thought to use get dummies or one hot encoder for nominal data but we have soo many nominal feature’s get dummies will create new column to each lables present in respective nominal column. This may lead to heigh dimensional problem which will not perform tree based algorithm. So i have used target encode and ordinal encoder. Ordinal encoder is used for the ordinal columns.

* Data Sources and their formats

Format of the data set is csv. Pandas i used to read and access the data for understand the structue of the data. From the data set we can see that heigh number of data is in categorical in ordinal and some of nominal data. Null values also present in the categorical data. Only in one continues column have null values.

* Data Preprocessing Done

In this project i have used in preprocessing that is Power Transformer and Standard scaller. My obervation is that accuracy of the model varies before preprocessing steps and after preprocessing. Decision tree and or tree based algorith not much affect whether we apply preprocessing or not. But in athor models like Linear Regression have soo much influes of preprocessing step. These type of models need preprocessing like need to handlel outliers, skewness scallig into the similar range to entire data set so model get not biased on certain features while prediction.

* Data Inputs- Logic- Output Relationships

Aim of this project is to identify the best input’s. Based on this inputs sale price will predict accuratly. Here some of the inputs which are heigly affecting to predict the sale price like Overal quality of the house which is most important feature to predict the price od house. In this columns have 10 number which are ordinal. 10 means the quality of house is soo rich having well neigbours waulity of material used while building the house all the thing’s are consider based on this sale price will increase or decrease. Suppose the quality of houses rated out of 10 have only 3 or 4 this type of houses price have less comapred to rating have more than 7. Such that other features also acts over target column i.e sale price and help to predict the price of house.

* Hardware and Software Requirements and Tools Used

ListinTools and libraries that i have used to solve project:

* Software: Anaconda, Jupyter Notebook, Python3
* Libraries: Numpy, Pandas, Matplotlib, Seaborn,

Sklearn.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

While performing EDA we have got insite information from dataset to predict Sale price. Some of the columns are interdependent like garage area and garege can have number of cars space. Number of cars are directly correlated to area of garage. Multicorrelation will effect some of the models.

Most of the columns are ordinal type which giving the

grades, rating and ranks to the house’s which have directly correlation to our target column.

Filling null values have major role in this dataset beacuse dataset have lots of null values in both categorical and continues data. Based on the type of column i have repalce the null values to some other values. I have handle null values in continuos column after encoding entire data set into numberical then after based on all the columns null values are imported bye usinf KNN Imputer.

* Testing of Identified Approaches (Algorithms)

I have used some of the model of machine learning like Linear regression, KNN, Decision Tree Regressor, SVR, SGD Regressor. I have used pipeline so each model can work and give there ouputs. It used to select best random state to split the data to resepective model and aslo find best cv(cross validation). Each model have differect random state and cv. I have tested by applying this pipeline before scalling, after scalling the data. And also before feature selection and after feature selection. How models are giving there output aslo observed that after scalling the data linear regression model’s accuracy has increase heigh. To handle the problem of overfitting and underfitting the model. Each model will give there out put after tunning the model.

* Run and Evaluate selected models

In this project i got good accuracies from Linear regression and decision tree regression algorithms.

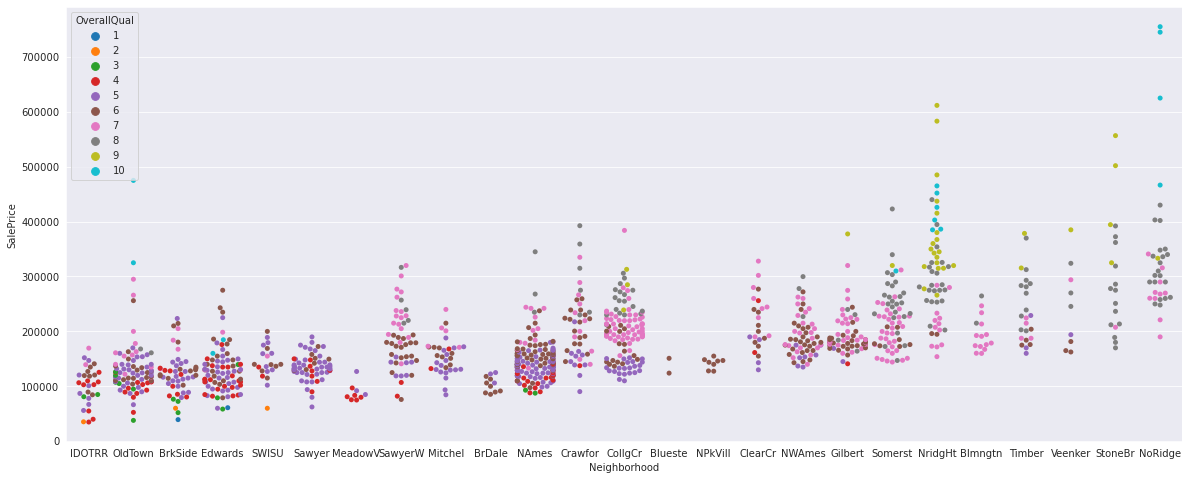
Linear Regression is used in suppervised regression model. Linear regression is used to fit the best fit line on the basis of that line sale prices would predicted. Best fit line is a line where the difference between actual and predicted values is less that is called best fit line.   
Linear regression contain two major values based on this values line would drawn i.e intercept and the slope of the line. Best slope would be calculated by using the gradient decent method with the small learning rate.

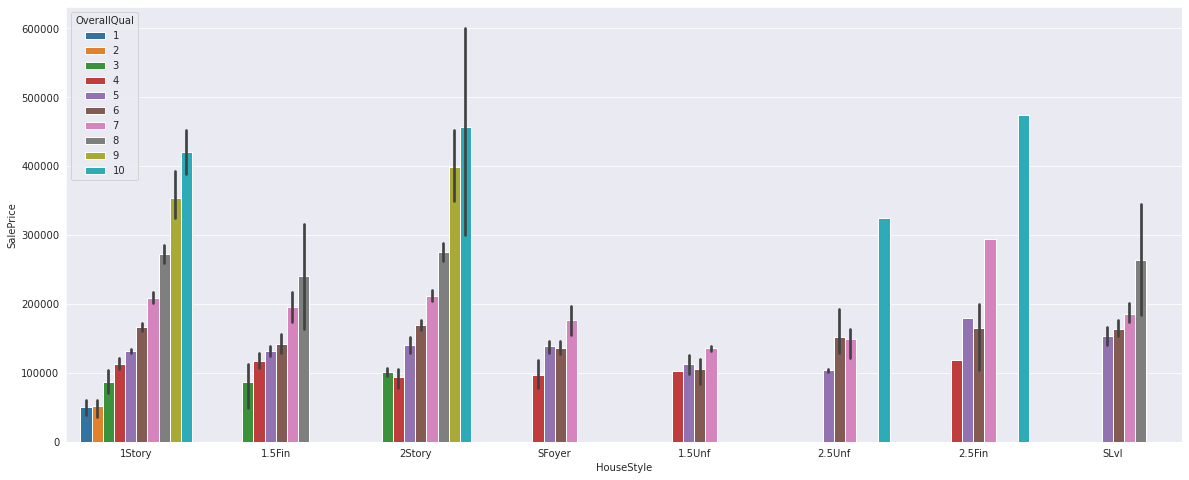
Decision tree is a tree based algorithm it cab be used on both classification and regression problems.

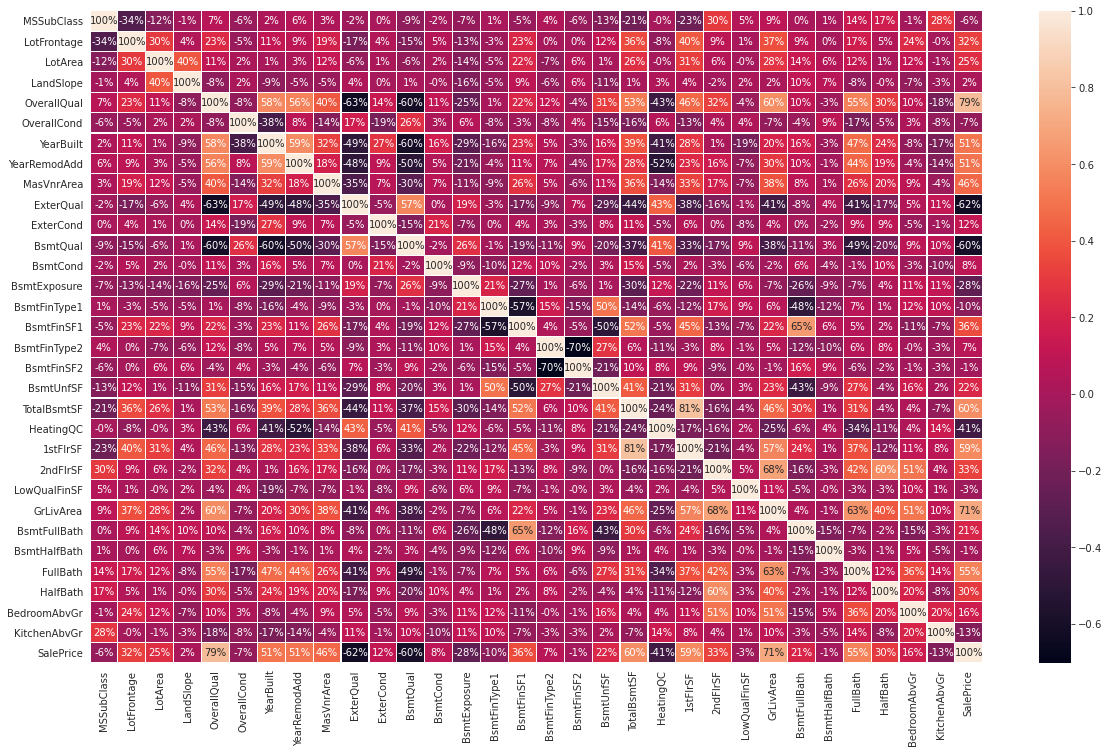
Now we have a regression problem so i used decision tree regressor. In this algorithm first it is calculte the impurity of each column and check the information gain. Which gives heigh information gain that columns beacomes a root node then their child would be their features. Then again calcualte the entrophy or gini and information gain like this tree will build.

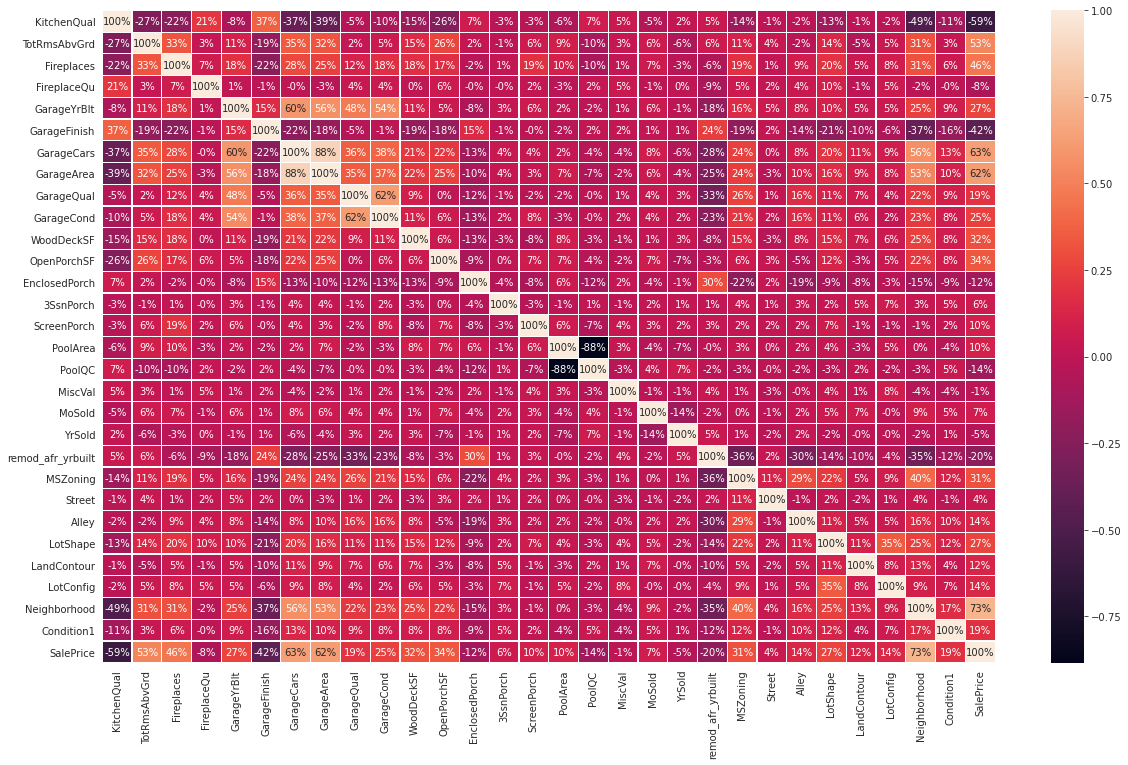
Problem of this algorithm having heigh chances of overfittig the model which may lead heigh variance in testing. To prevent this problem we have parameter’s like depth of the tree. If the tree will read each and every point’s then it may have overfitting problem. If we hyper tune this algorithm problem could be solved.

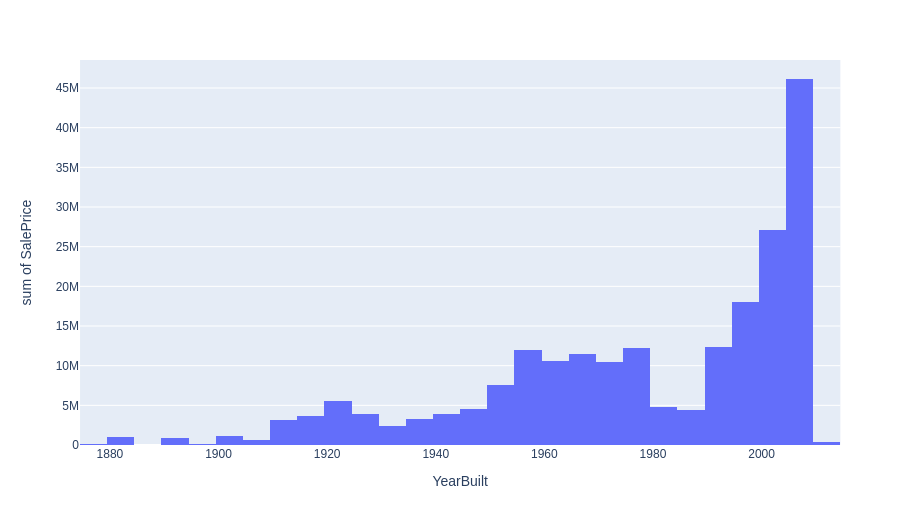
* Visualizations

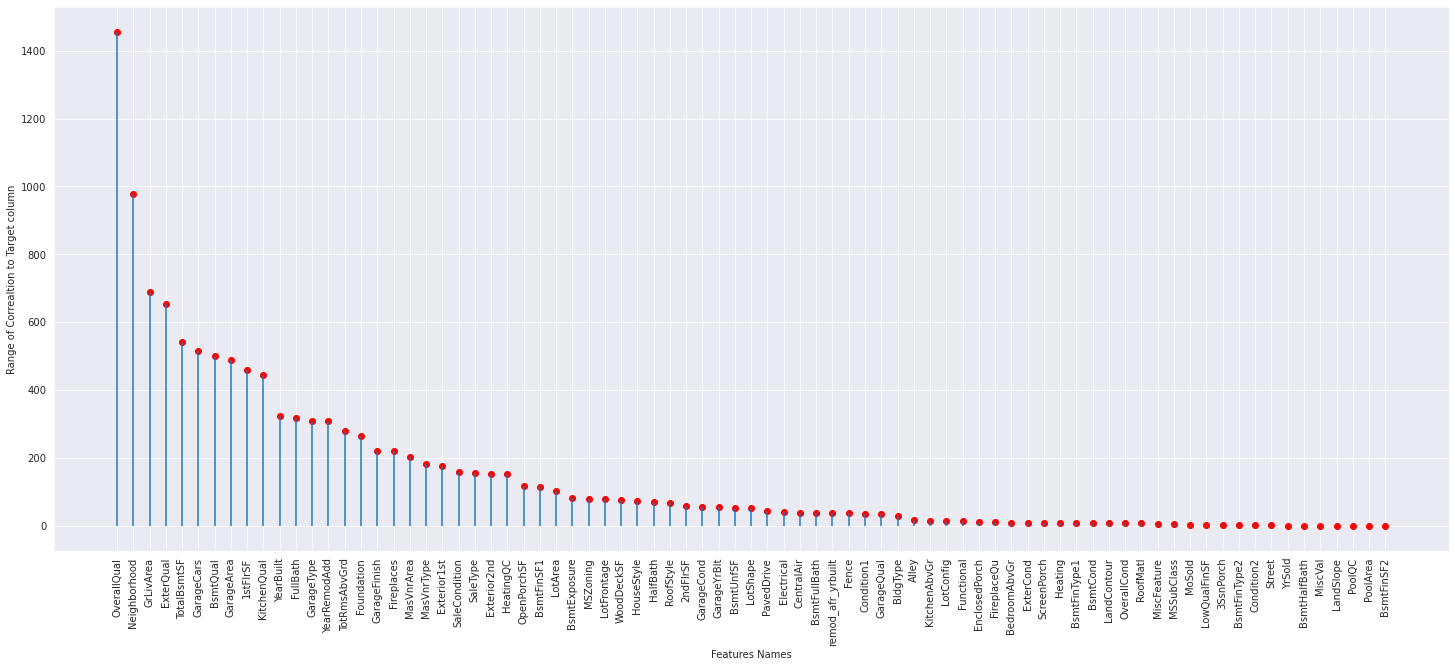












* Interpretation of the Results

To see the accuracy of each model i have used r2 score. R2 score is the score how good slope has genreate for the best fit line. Linear regression and decision tree regression is giving the r2 score near to 88.4 %.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function and Target encoder and ordinary encoder. The main challenging is that to run all model by hypertuning each model and fetching each model best random state and best cv once and get the best model name as a result.