

HOUSE PRICE PREDICTION

Submitted by:

SURJEET SINGH

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to **FlipRobo** who gave me the golden opportunity to do this wonderful project on the topic **House Price Prediction** project. The data used in this project was provided by the **FlipRobo** itself.

**INTRODUCTION**

* Business Problem Framing

As the population of the world is increasing therefore, house for living is a necessity. The real estate became a big market where several companies are working. The challenging part in this domain is to correctly place the price of a house based on the facilities provided in the house. Here the data science comes into picture. We will build a predictive model which will help the real estate companies to place the correct price of the house.

* Review of Literature

Along with the facilities inside the house, the price of a house also depends upon the current market condition. For example, during COVID-19 pandemic, the market was on downfall due to fewer sales of flats. The price of the house decrease if there is very noisy roads in the vicinity. The price of the house increases if the utility are nearby for example Gym, grocery etc. The following link was helpful for me to understand the problem in a better way. <https://www.investopedia.com/articles/mortgages-real-estate/08/fair-price-on-home.asp>

**Analytical Problem Framing**

* Data Sources and their formats

The data was provided by the FlipRobo for this project. The dataset has 1168 rows and 81 features. The dataset contains mixed type of features, such as numerical features, categorical features, and discrete features in the categorical features.

* Data Preprocessing Done

There are some of the features having missing values more than 90%. We have removed those features having missing values more than 1000. For rest of the features we have fill missing values with median. We have not used mean here because some of the houses has huge values which can affect imputation. There are different types of variables such as numerical variables, categorical variables etc. We have used label encoding in the case of categorical variables and minmax scalar for scaling the data.

* Data Inputs- Logic- Output Relationships

Since this is a regression problem, the output of the model must be a continuous number. And there are several variables which have values of different range which will be deciding factor for a machine learning model to predict the price of a house.

* State the set of assumptions (if any) related to the problem under consideration

In this case we have not removed the outliers. As we have assumed that some of the houses which are luxurious have very high price. The model should predict the price of such houses with accuracy otherwise it will be a loss for the company as big values return big profits.

* Hardware and Software Requirements and Tools Used

All the codes are written in python 3. Jupyter notebook used to create all visualizations, training, and testing of the machine learning models. Device used in this project : hp laptop with intel i5 process and 4 GB RAM.

**Model/s Development and Evaluation**

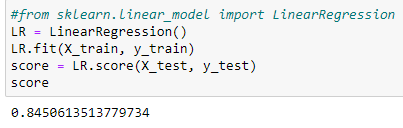
* Identification of possible problem-solving approaches (methods)

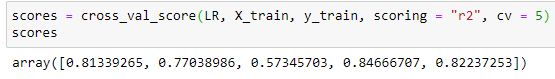
It is a regression problem and we will use linear regression, decision tree, random forest regression and gradient boost regression algorithms to solve this problem.

* Testing of Identified Approaches (Algorithms)

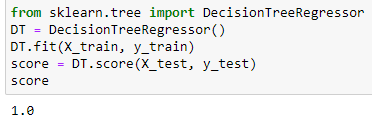
We have applied four different machine learning models to solve this problem. We have also performed the cross validation for each model to ensure the accuracy of the model.

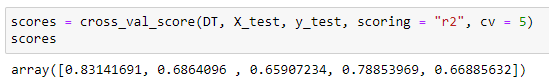
* Run and Evaluate selected models



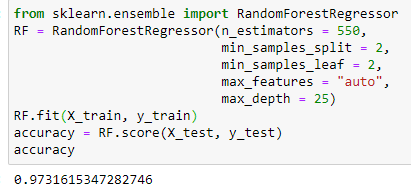


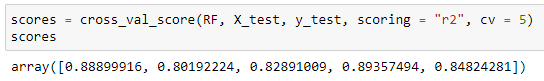
The linear regression model has performed well and returns a good score of 85%.





The decision tree model have preformed super on the test data. But in the cross validation, its performance reduced and it has fluctuations in score also.





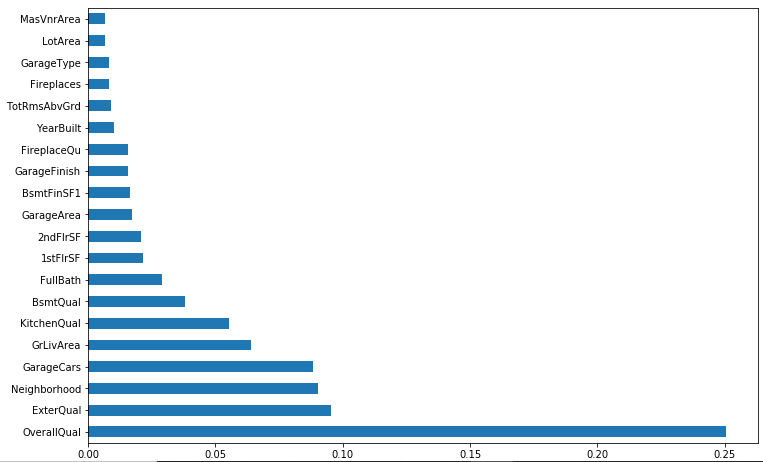
The score of random forest is also good. It has good accuracy for cross validation also. Random forest is also maintaining the accuracy in the cross validations.

* Key Metrics for success in solving problem under consideration

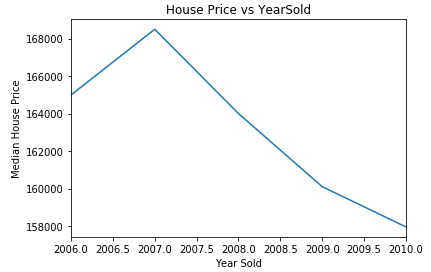
It is a regression problem. The predicted values by the machine learning models are continuous values. Therefore, RMSE is a good choice to evaluate the model.

* Visualizations

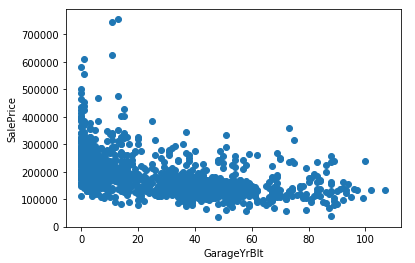
We have plotted so many graphs in this project. Since there are 81 variables so it is difficult to put every type of plot for every variable here. Therefore, we will show some of the graphs here.



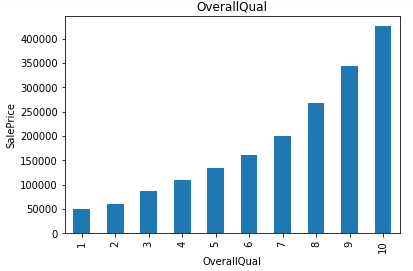
This graph is showing the feature importance of top 20 features. The most important feature for house price prediction is overall quality of the house.



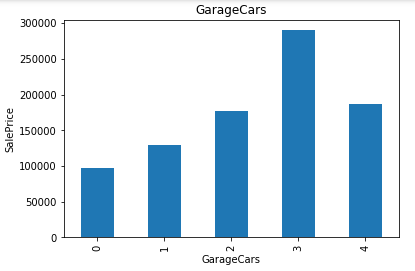
It is clear from the above graph that the price of the house is decreasing with time. It is intuitive that if the house gets older, its price will reduce.



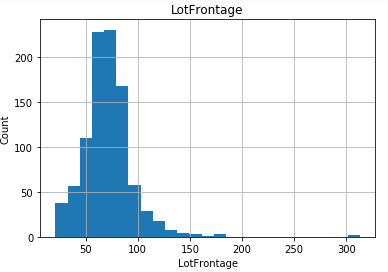
The above graph is on logarithmic scale between sales price and garage built year. As the age of the garage is increasing, the price of the house is degreasing. The similar behaviour have been observed for some other features such as remodificaiton of house, year of built etc.



There is a very beautiful exponential growth of sale price with the overall quality of the house. Also, it is the most important factor in deciding the price of the house.



The average price of the house is increasing as the number of garages increases. There is an exception for 4 garages.



The data distribution for most of the continuous variables is normal but has some skewness in it. The skewness can be removed by using log normal transformations.

We have also plotted the correlation graph but here it is difficult to fit because it has 81 features in it. That graph is shown in the jupyter notebook.

* Interpretation of the Results

Most of the continuous variables are normal but have some skewness in it. That skewness has been removed using the log normal transformation. The age of the house, affect the price of the house. As the house gets older, the price of the house decreases. The remodifications done in the house also reduce the price with time. The number of garages affects the price of the house positively. The most important factor in price prediction is overall quality of the house. There is an exponential growth in the average price of the house as overall quality of the house increases.

**CONCLUSION**

* Limitations of this work and Scope for Future Work

The number of instance in this project is low. We can increase the data to increase the accuracy. There 80 variables in the dataset. As we know that there is curse of dimensionality, we can use PCA to reduce the dimension of data. We can also remove the highly correlated features to reduce it.