

**HOUSING: PRICE PREDICTION**

Submitted by:

AMIT PURI GOSWAMI

**ACKNOWLEDGMENT**

The project is done under the kind guidance of Mr. Kasif Khan. Some research papers and data is used from different online material available freely.

The data belongs to sale of houses in Australia and used by Surprise Housing to predict the property price.

**INTRODUCTION**

* The main problem is to predict the House Price and to find out the features highly affect the same.

The company Surprise Housing wish to step in the property market of Australia to make maximum profit by purchasing the housing property in less price and selling at a higher price.

This is a very common strategy any company make for business. Here comes the role of the Machine Learning to help the company identify the best rate (low ) of the property than the expected. By knowing this the company can purchase the property and later sale on at a higher price.

The problems is very similar to other problems in the real world, we may need to know the weather condition on the basis of some features , we may require to know the price of any other goods by using the machine learning model. The list goes on.

* Conceptual Background of the Domain Problem

We need to identify the most contributing variables which affect the SalePrice highly. Apart from this we would like to know the estimated SalePrice of the House Property.

* Review of Literature

It is researched that OverallQual is the highest affecting factor for the SalePrice other decisive

factors are :

YearBuilt 0.514408

TotRmsAbvGrd 0.528363

FullBath 0.554988

1stFlrSF 0.587642

TotalBsmtSF 0.595042

GarageArea 0.619000

GarageCars 0.628329

GrLivArea 0.707300

**OverallQual 0.789185**

* Motivation for the Problem Undertaken

There is a big motive behind this as if the model works well it will give us best prediction to get good profits from the purchase and sale of House Properties.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

There is no substantial multi collinearity found so no columns are deleted.

Outliers are found but the data loss is 100% so no outliers are removed.

* Data Sources and their formats

Data is coming from sale of houses in Australia.

* Data Preprocessing Done

Data is tested for null values, blank values and treated with suitable values. In string features it is replaced with ‘NoVal’ and in numeric feature it is replaced with mean value of the features.

* Data Inputs- Logic- Output Relationships

Input is always important for output. The value in the highly contributing factors drastically changes the output.

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

NA

* Hardware and Software Requirements and Tools Used

Basic Hardware are : i5 processor, 8GB RAM, other standard h/w

Softwares used : Jupyter, MS Excel, MS Word

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Standard process of data cleansing, feature selection, EDA, Outliers Removal, Skewness Removal, Handling Multi collinearity , dividing data in train and test, Standardization, finding best random state, model selection, finding best cv, Parameter finetuning, Finalization of the model and predicting.

* Testing of Identified Approaches (Algorithms)

Algorithms used are

1. LinearRegression
2. DecisionTreeRegressor
3. RandomForestRegressor
4. Ridge
5. SVR
6. KNeighborsRegressor

* Run and Evaluate selected models

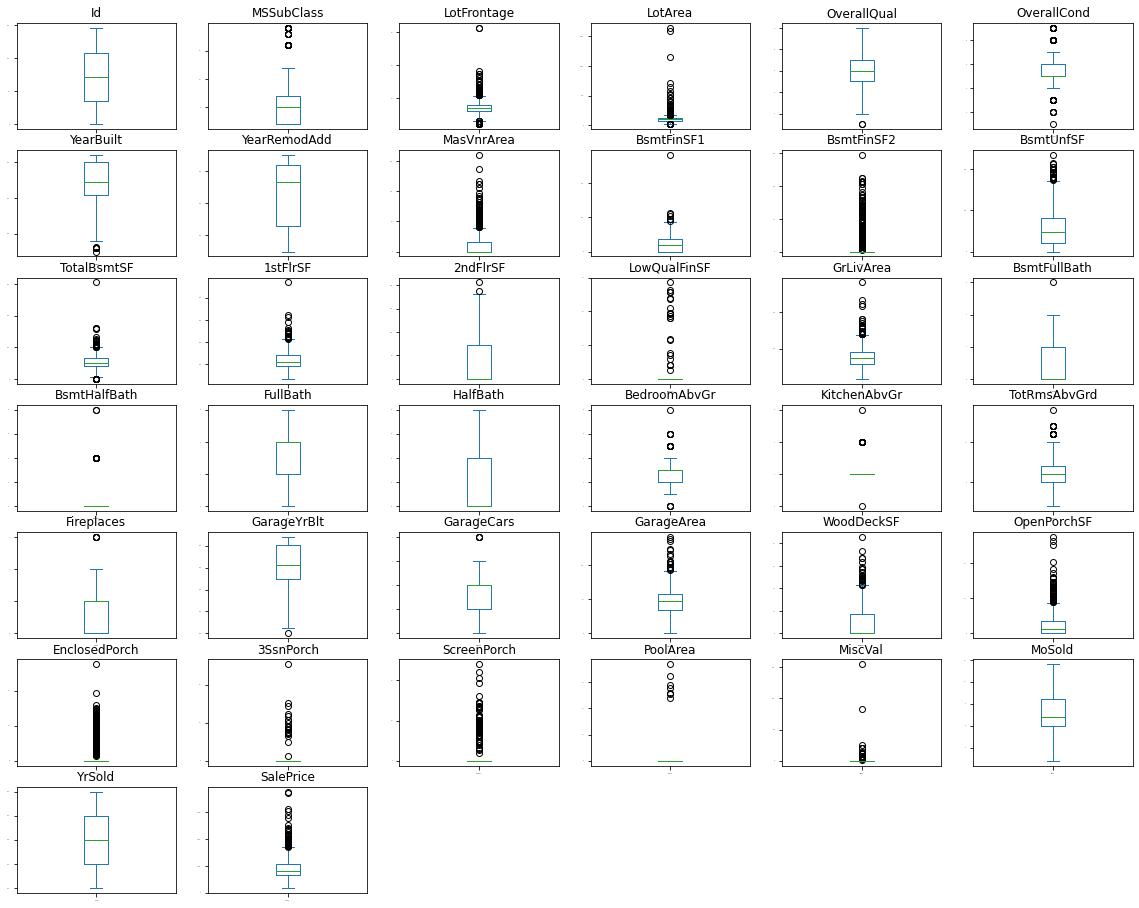
After running and evaluating RandomForestRegressor is found to be the best model.

* Key Metrics for success in solving problem under consideration

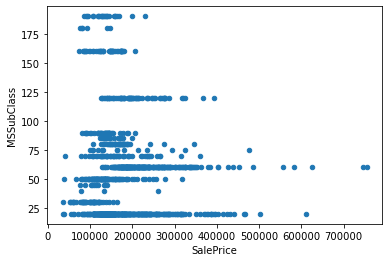
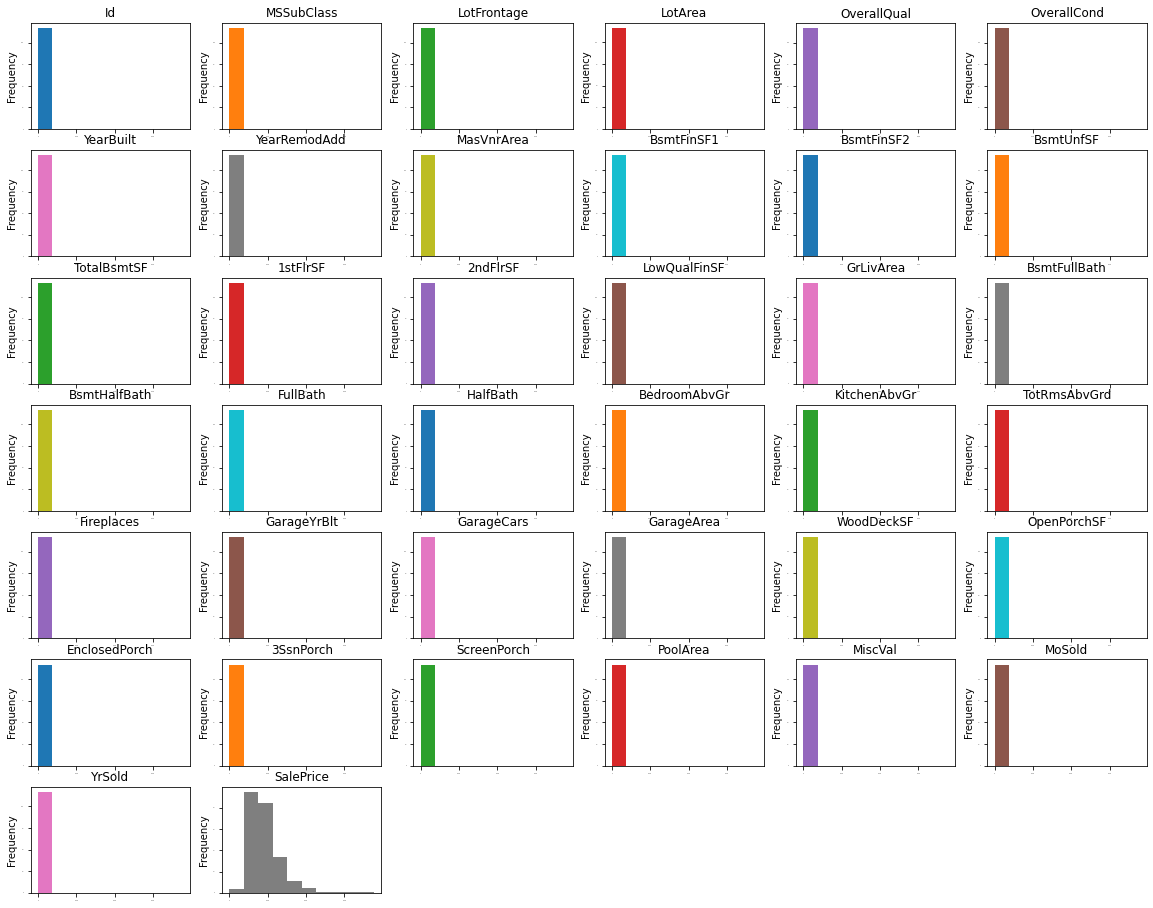
Best feature selection

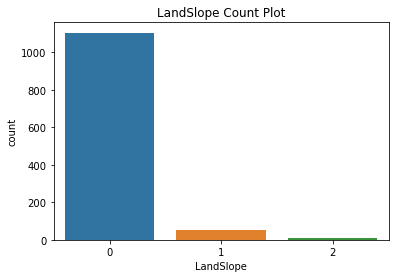
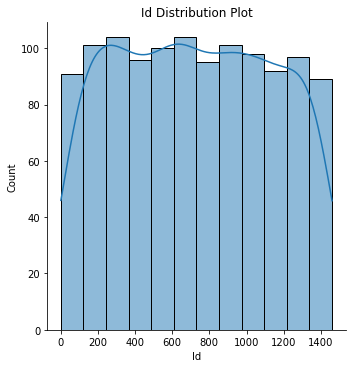
* Visualizations

1. Scatterplot
2. CountPlot
3. Boxplot



It shows that outliers are present.

Shows that Highest SalePrices is at 60 value of MSSubClass.



Landscope 0 is very common in all the property and its count is highest.

* Interpretation of the Results

**CONCLUSION**

* Key Findings and Conclusions of the Study

Highly contribution factor for SalePrice are

GarageCars 0.628329

GrLivArea 0.707300

**OverallQual 0.789185**

Highly negative contribution factor for SalePrice are

ExterQual -0.624820

BsmtQual -0.601307

KitchenQual -0.592468

* Learning Outcomes of the Study in respect of Data Science

Visualization helps us a lot in understanding the problem and identifying the work which will be suitable for model building.

* Limitations of this work and Scope for Future Work

No. of columns are 81 which is very high for a standard computer to handle and if we get high performing hardware little more accuracy can be obtained by getting best parameters.