FLIP ROBO

House Price Prediction

Submitted by:

PREETI SINGH

ACKNOWLEDGMENT

I would like to thank the Gulshana Chaudhary my SME in Internship from Flip-Robo Technology for providing me with Train and Test Dataset , Sample Documentation and Data Description without which working in this project would have been very challenging . I would also like to mention I received help and resource from various websites like Scikit Learn, Kaggle, TowardsDataScience and Medium etc.

INTRODUCTION

- **Business Problem Framing**

  Houses are one of the necessary needs 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.

- **Conceptual Background of the Domain Problem**

  Observation Based on the above data, we can drop the following columns - LotFrontage - Alley - FireplaceQu - PoolQC - Fence - MiscFeature - Id (dropping this not because of count, irrelevant) - MoSold (dropping this not because of count, irrelevant) - Street (dropping this not because of count, irrelevant) - Utilities (dropping this not because of count, irrelevant) Review of Literature

  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**

  It is required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

  Data contains 1460 entries each having 81 variables. Data contains Null values. We need to treat them using the domain knowledge and our own understanding. Extensive EDA has to be performed to gain relationships of important variable and price. Data contains numerical as well as categorical variable. We need to handle them accordingly. We have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters. We need to find important features which affect the price positively or negatively. Two datasets are being provided to you (test.csv, train.csv). We will train on train.csv dataset and predict on test.csv file.

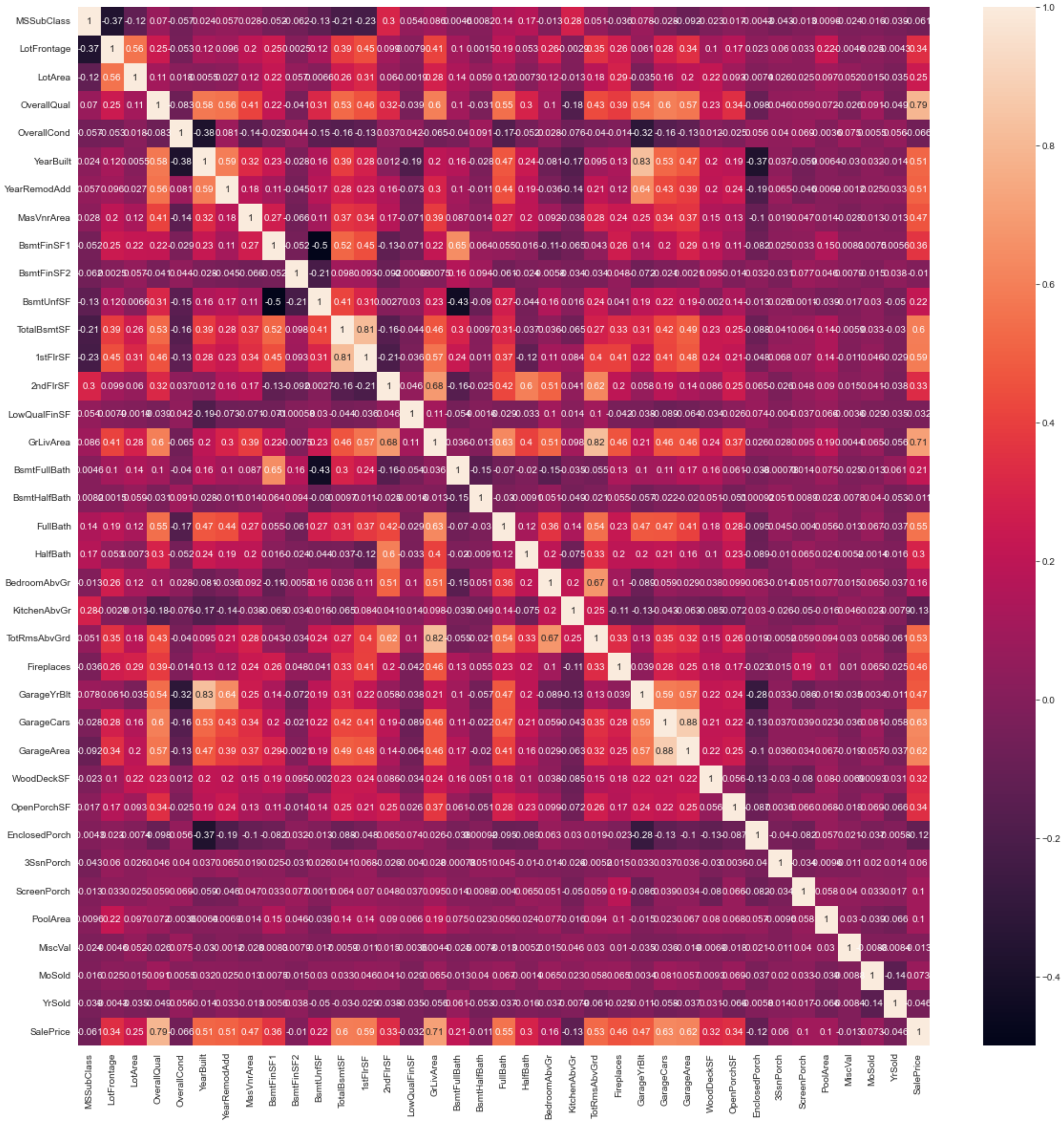
- **Data Sources and their formats**

  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.

- **Data Pre-processing Done**

  Storing null values in train, then printing columns with more than 0 null values Impute missing values Dropping columns which has around 50 percent missing values Selecting categorical features and encoded get dummies of input features Scaling input and test data using MinMaxScaler module.

- **Data Inputs- Logic- Output Relationships**

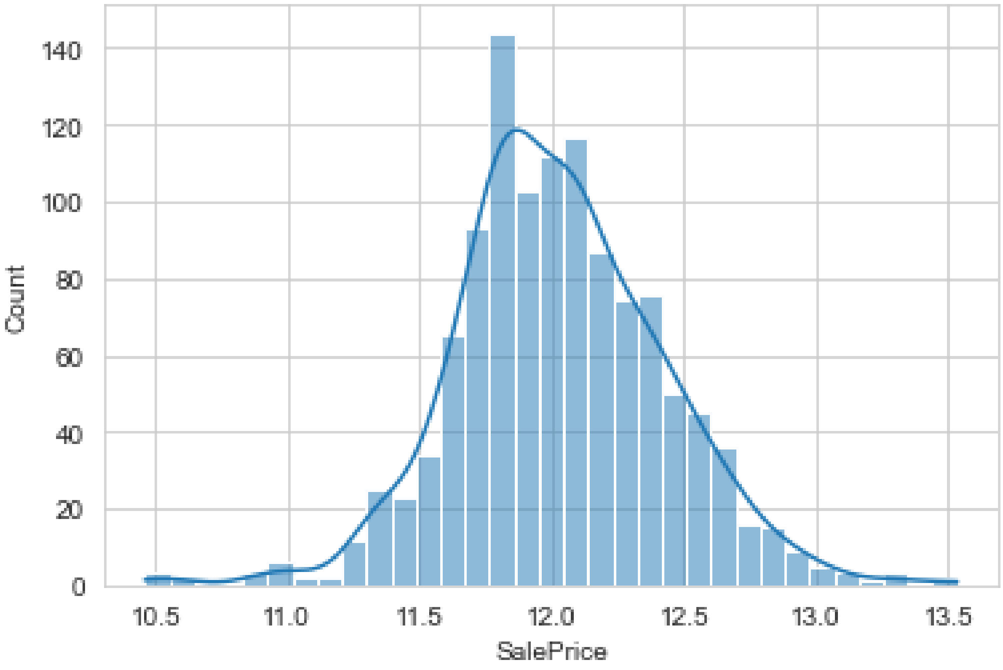


- **Hardware and Software Requirements and Tools Used**

The General Hardware used for this project is: - 8 GB RAM 512GB SSD Intel i5 processor The Software and tools used for this project is: - Python (Jupyter Notebook) Scikit Learn Various tools: - Pandas, Matplotlib, NumPy, Seaborn etc.

Model/s Development and Evaluation

Following is the distribution plot of MarketValue which is our Target.

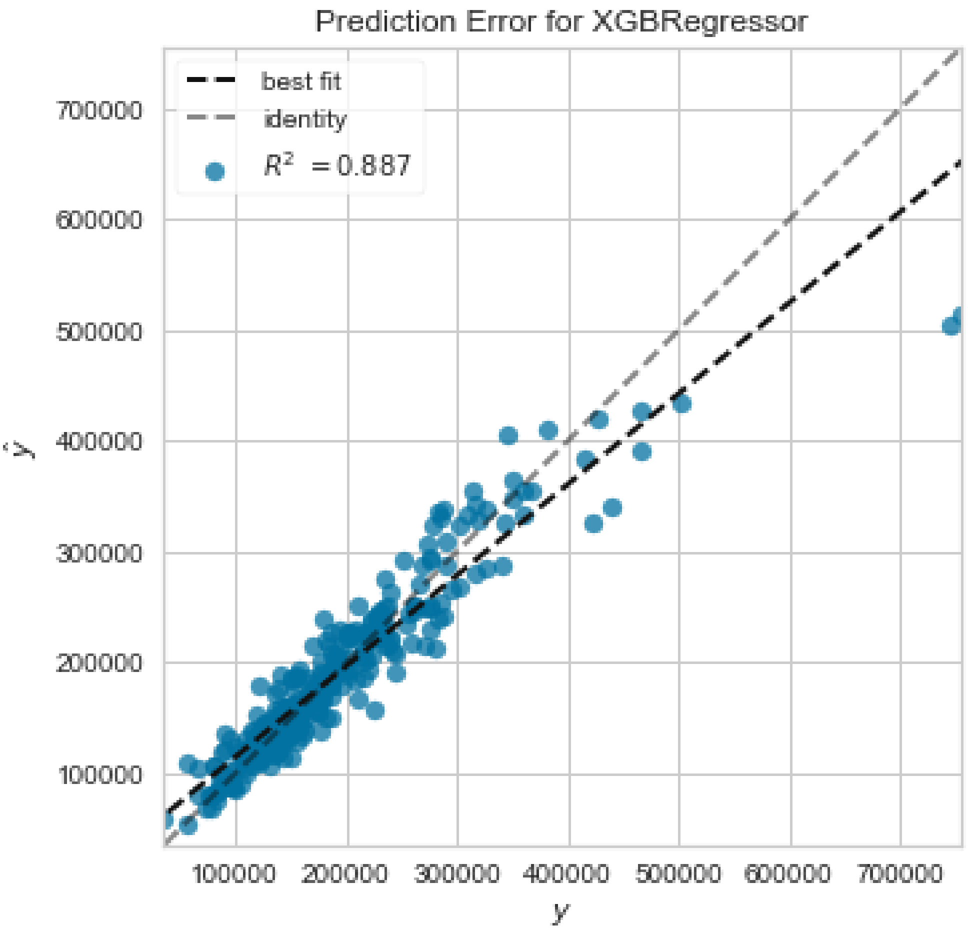


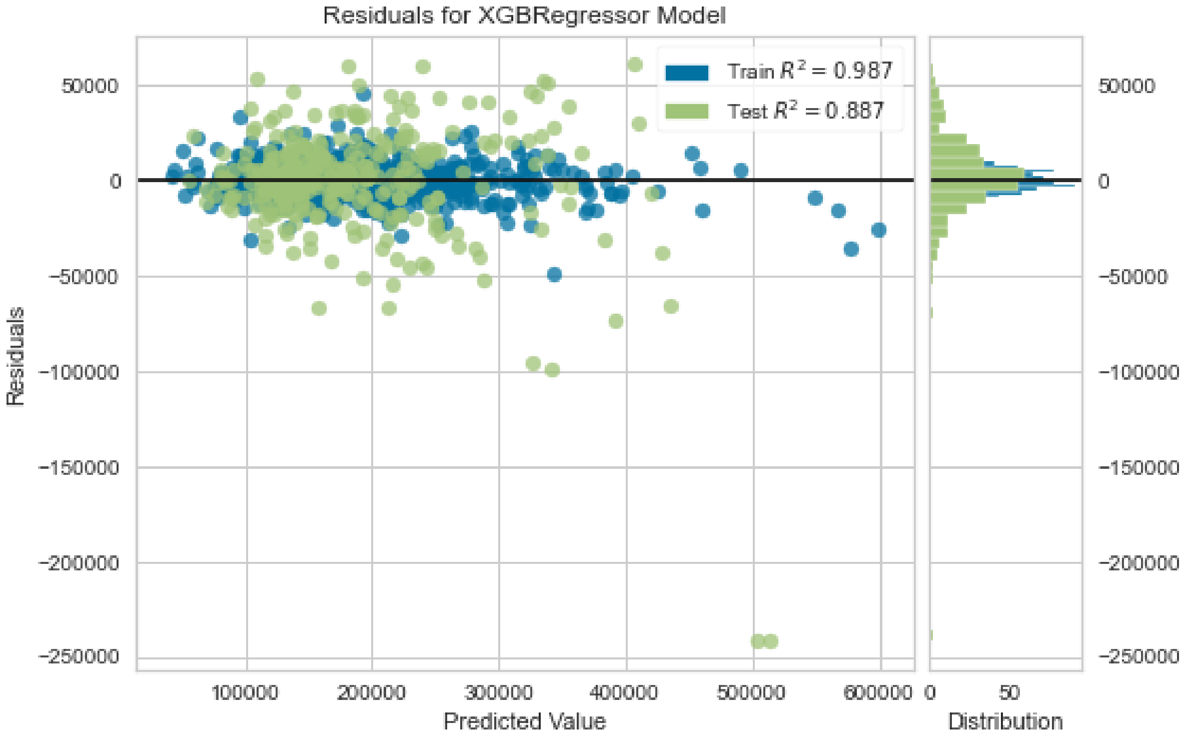
The Different Models which are used are: -

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. DecisionTree Regressor
5. Support Vector Regressor
6. KNeighbors Regressor
7. Random Forest Regressor
8. XGBoost Regressor
9. Elastinet Regressor
10. SGD Regressor
11. Bagging Regressor
12. Adaboost Regressor
13. Gradient Regressor

CONCLUSION

I used various different types of models and used R2 score, RSME score and Cross Validation Score to determine which model is best. After training and testing all the models I have came to this conclusion that XGBoost Regressor is giving the best performance with 88% test R2 score, 98% of training R2 score, 0.82 cross validation score, RMSE score 32674.21346224513.




Residuals for XGBRegressor Model
Train $R^2 = 0.987$
Test $R^2 = 0.887$
Residuals
50000
0
−50000
−100000
−150000
−200000
−250000
100000
200000
300000
400000
500000
600000
Predicted Value
0
50
Distribution