

Housing Price Prediction

Submitted by:

Moncy Kurien

**ACKNOWLEDGMENT**

The data was collected by the company named Surprise Housing from the Sale of Houses in Australia.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain Problem

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. Companies can increase their overall revenue, profit by smartly investing in the houses with the right price. Using Data science, these companies can predict the prices of the house and invest the right amount on the house and sell them at a higher price.

* Review of Literature

In this project, the different features of the houses are analysed and the features that contribute more towards the price of the house are identified and are used to predict the house price. For this purpose, the dataset from the sale of houses in Australia is used.

There were missing values in many features which are imputed. PowerTransformer is used to fix the skewness of the data.

For scaling, MinMaxScaler is used. The dataset contains about 33% of outliers. Since it’s such a huge number, these outliers could just be natural outliers which won’t be an outlier if we have more data. The following features are dropped for specific reasons:  
1. LotFrontage has only 34% correlation with the SalePrice(target variable). Hence, it is dropped.

2. The YearBulit and the GarageYrBlt are highly correlated to each other. Hence, GarageYrBlt is dropped.

3. 1stFlrSF and TotalBsmtSF are highly correlated to each other. Hence dropped 1stFlrSF.

4. GrLivArea and TotRmsAbvGr are highly correlated to each other. Hence, dropped TotRmsAbvGr.

5. GarageCars and GarageArea are highly correlated to each other. Hence, dropped GarageArea.

6. Since Id is unique to each house, it has been dropped.

OneHotEncoder is used for encoding the categorical data so that any new values in the feature in test data or any future data can be ignored.

The Final hyperparameter tuned model chosen is the XGBRegressor model.

The cross val score of the final model with the whole train data is: 0.8832426817573655

and the Variance is: 0.0007309462170151364

* Motivation for the Problem Undertaken

Real Estate is a very popular domain for investment for many companies as well as many individuals. This project will help the companies/individuals to make the right decisions in terms of buying a house in Australia.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Since the target value SalePrice is a continuous variable, the problem is considered as a regression problem. Although there are some variables with good correlation, the linear models except Ridge did not perform well. The Ensemble models such as Gradient Boosting and Extreme Gradient Boosting techniques did a very good job in predicting the Sale price.

* Data Sources and their formats

The dataset was collected from the sale of houses in Australia. Data contains 1460 entries each having 81 variables.

The Variable data types are as follows:

Data columns (total 81 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Id 1168 non-null int64

1 MSSubClass 1168 non-null int64

2 MSZoning 1168 non-null object

3 LotFrontage 954 non-null float64

4 LotArea 1168 non-null int64

5 Street 1168 non-null object

6 Alley 77 non-null object

7 LotShape 1168 non-null object

8 LandContour 1168 non-null object

9 Utilities 1168 non-null object

10 LotConfig 1168 non-null object

11 LandSlope 1168 non-null object

12 Neighborhood 1168 non-null object

13 Condition1 1168 non-null object

14 Condition2 1168 non-null object

15 BldgType 1168 non-null object

16 HouseStyle 1168 non-null object

17 OverallQual 1168 non-null int64

18 OverallCond 1168 non-null int64

19 YearBuilt 1168 non-null int64

20 YearRemodAdd 1168 non-null int64

21 RoofStyle 1168 non-null object

22 RoofMatl 1168 non-null object

23 Exterior1st 1168 non-null object

24 Exterior2nd 1168 non-null object

25 MasVnrType 1161 non-null object

26 MasVnrArea 1161 non-null float64

27 ExterQual 1168 non-null object

28 ExterCond 1168 non-null object

29 Foundation 1168 non-null object

30 BsmtQual 1138 non-null object

31 BsmtCond 1138 non-null object

32 BsmtExposure 1137 non-null object

33 BsmtFinType1 1138 non-null object

34 BsmtFinSF1 1168 non-null int64

35 BsmtFinType2 1137 non-null object

36 BsmtFinSF2 1168 non-null int64

37 BsmtUnfSF 1168 non-null int64

38 TotalBsmtSF 1168 non-null int64

39 Heating 1168 non-null object

40 HeatingQC 1168 non-null object

41 CentralAir 1168 non-null object

42 Electrical 1168 non-null object

43 1stFlrSF 1168 non-null int64

44 2ndFlrSF 1168 non-null int64

45 LowQualFinSF 1168 non-null int64

46 GrLivArea 1168 non-null int64

47 BsmtFullBath 1168 non-null int64

48 BsmtHalfBath 1168 non-null int64

49 FullBath 1168 non-null int64

50 HalfBath 1168 non-null int64

51 BedroomAbvGr 1168 non-null int64

52 KitchenAbvGr 1168 non-null int64

53 KitchenQual 1168 non-null object

54 TotRmsAbvGrd 1168 non-null int64

55 Functional 1168 non-null object

56 Fireplaces 1168 non-null int64

57 FireplaceQu 617 non-null object

58 GarageType 1104 non-null object

59 GarageYrBlt 1104 non-null float64

60 GarageFinish 1104 non-null object

61 GarageCars 1168 non-null int64

62 GarageArea 1168 non-null int64

63 GarageQual 1104 non-null object

64 GarageCond 1104 non-null object

65 PavedDrive 1168 non-null object

66 WoodDeckSF 1168 non-null int64

67 OpenPorchSF 1168 non-null int64

68 EnclosedPorch 1168 non-null int64

69 3SsnPorch 1168 non-null int64

70 ScreenPorch 1168 non-null int64

71 PoolArea 1168 non-null int64

72 PoolQC 7 non-null object

73 Fence 237 non-null object

74 MiscFeature 44 non-null object

75 MiscVal 1168 non-null int64

76 MoSold 1168 non-null int64

77 YrSold 1168 non-null int64

78 SaleType 1168 non-null object

79 SaleCondition 1168 non-null object

80 SalePrice 1168 non-null int64

dtypes: float64(3), int64(35), object(43)

memory usage: 739.2+ KB

* Data Preprocessing Done

The missing values in the categorical features are assumed to missing because that feature is not available for a house. The missing MasVnrArea is set to 0 because the values were missing for houses that did not have Masonary Veneer.

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

Assumed that the missing categories are because the features were not available for the houses.

* Hardware and Software Requirements and Tools Used

1. Google Colab

2. SKLEARN

3. MATPLOTLIB

4. SEABORN

5. PANDAS

6. NUMPY

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

This problem can be solved using Regression models. The R2 scores and Mean Squared Errors of different Base models and the ensemble models are compared, and the final best fit is used as the final model. The Final hyperparameter tuned model chosen is XGBRegressor model.

* Testing of Identified Approaches (Algorithms)

The following models were tried to solve this problem:

1. Lasso

2. Ridge

3. SVR

4. DecisionTreeRegressor

5. KNeighborsRegressor

6. XGBRegressor

7. RandomForestRegressor

8. GradientBoostingRegressor and

9. AdaBoostRegressor

* Run and evaluate selected models

Model: Lasso

R2 score: -0.00916039323870992

Variance: 0.0001389141606020204

==================================================

Model: Ridge

R2 score: 0.8253377629095887

Variance: 0.007238669191463325

==================================================

Model: SVR

R2 score: 0.7375989377657173

Variance: 0.0060217148913749245

==================================================

Model: DecisionTreeRegressor

R2 score: 0.6526893652219987

Variance: 0.015108446011103113

==================================================

Model: KNeighborsRegressor

R2 score: 0.7132321605153435

Variance: 0.0056167405609302765

==================================================

Model: XGBRegressor

R2 score: 0.8554513892311743

Variance: 0.00368889368828731

==================================================

Model: RandomForestRegressor

R2 score: 0.8341334147620156

Variance: 0.0042206029168412994

==================================================

Model: GradientBoostingRegressor

R2 score: 0.8578791144380302

Variance: 0.0037000220230117564

==================================================

Model: AdaBoostRegressor

R2 score: 0.7765540528658017

Variance: 0.004527487045823097

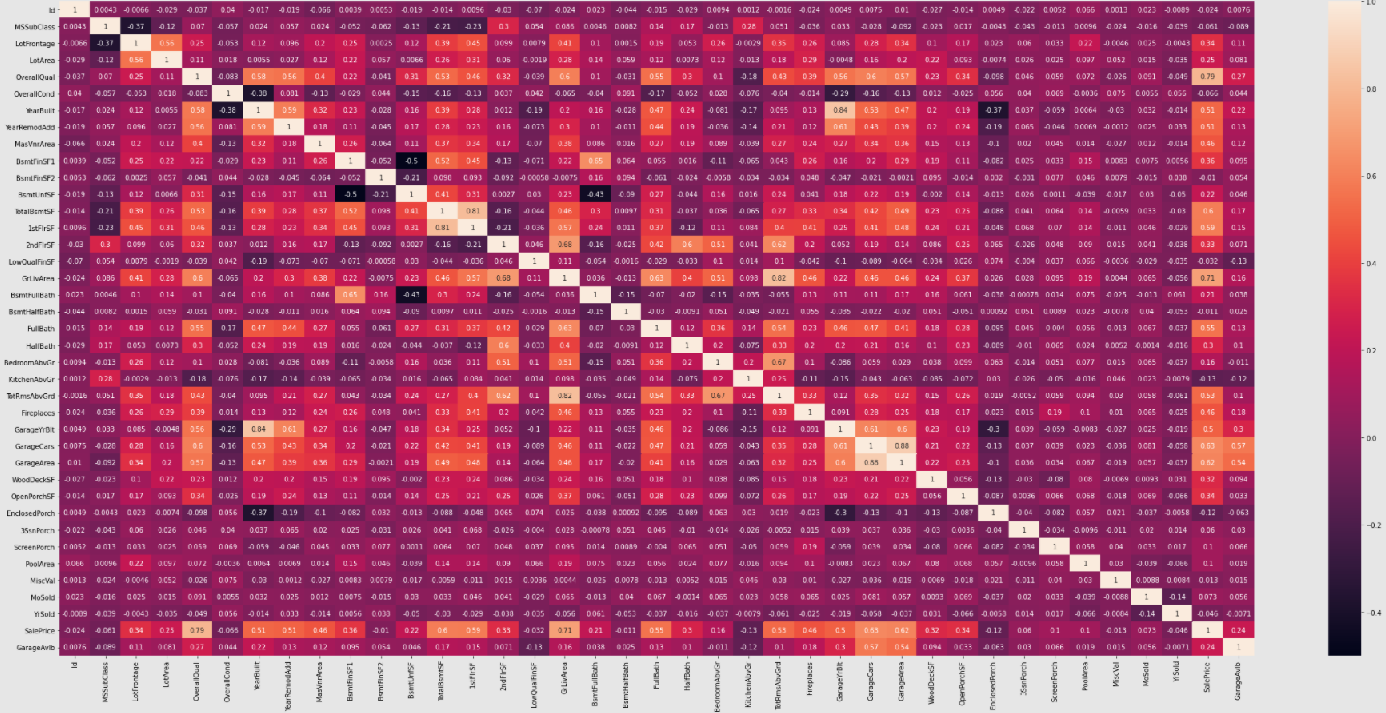
==================================================

* Key Metrics for success in solving problem under consideration.

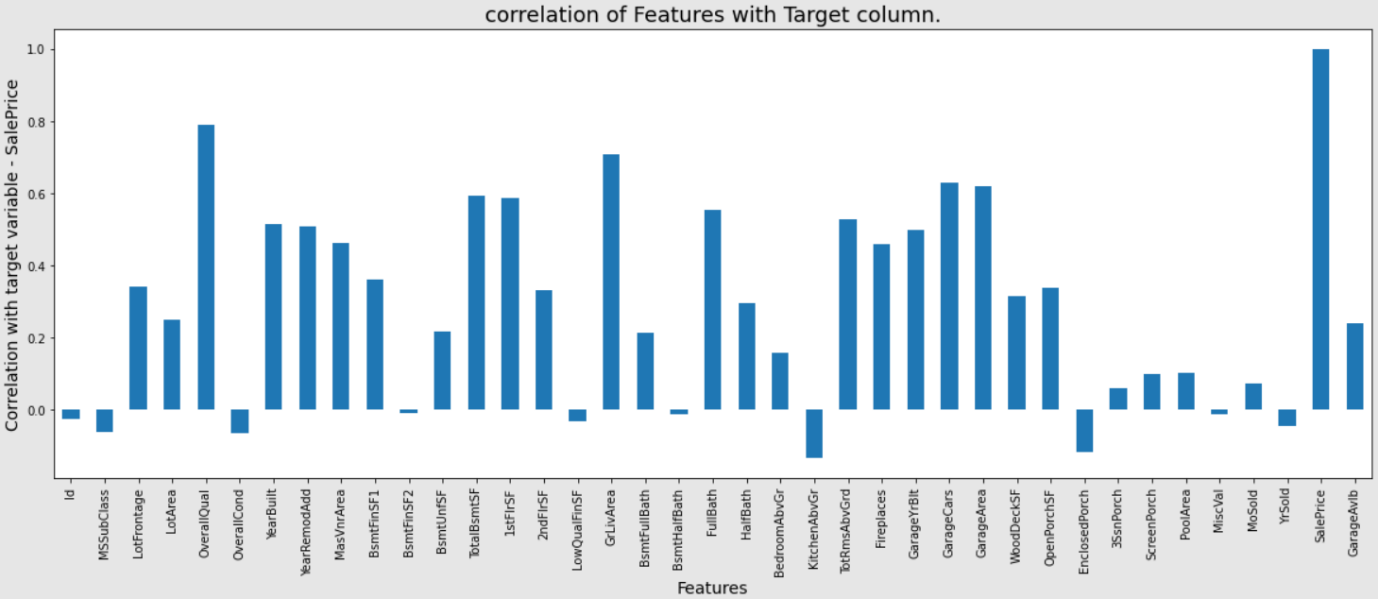
Since this is a regression problem, I used the R2 score and Mean Square Error as the Key metrics to find the best fitting model.

* Visualizations

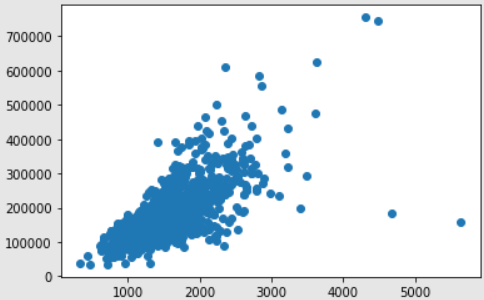
1. **Correlation of features:**



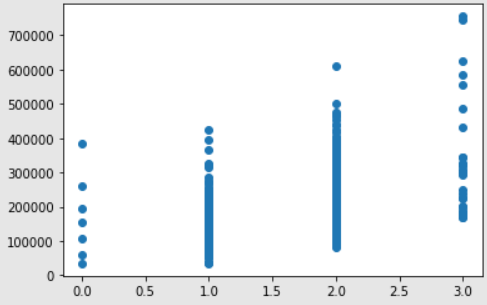
1. **Correlation of features with target:**



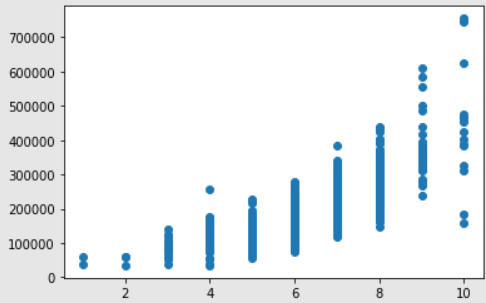
1. **GrLivArea vs SalePrice:**



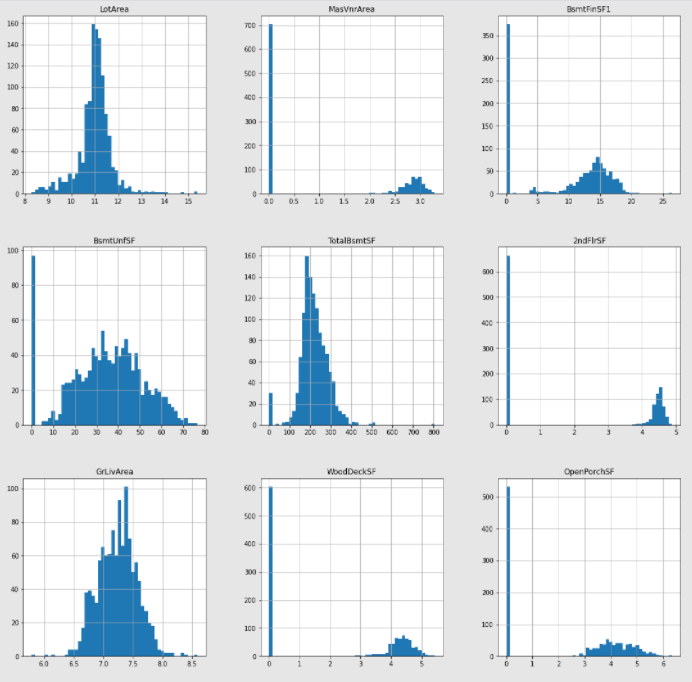
1. **FullBath vs SalePrice:**



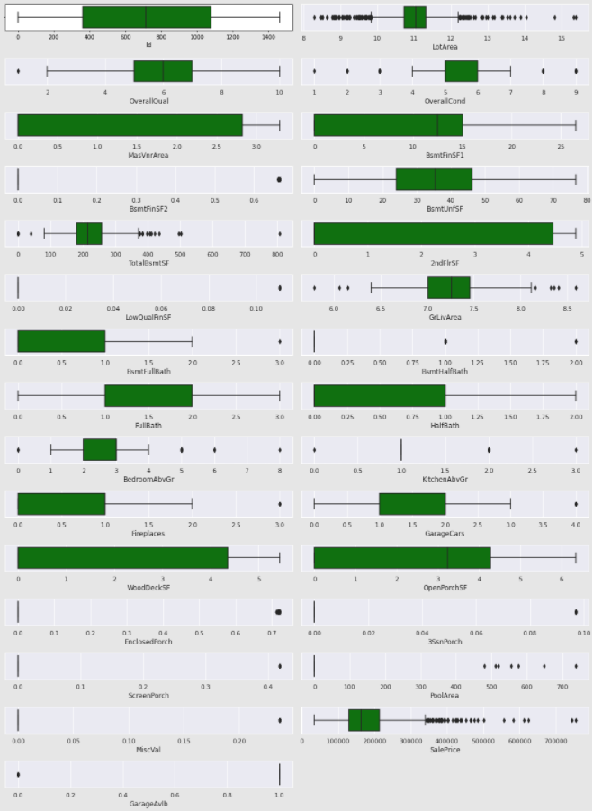
1. **OverallQual vs SalePrice:**



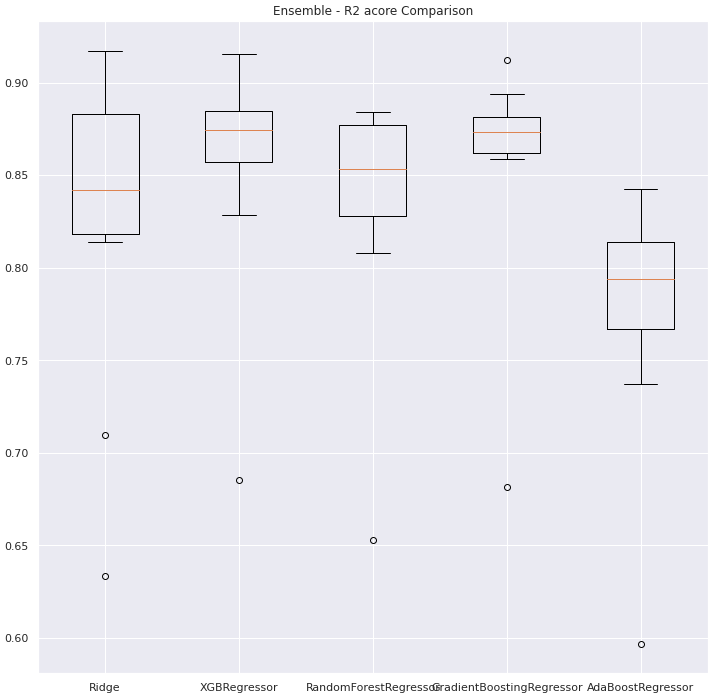
1. **Distribution of numerical features after transformation:**



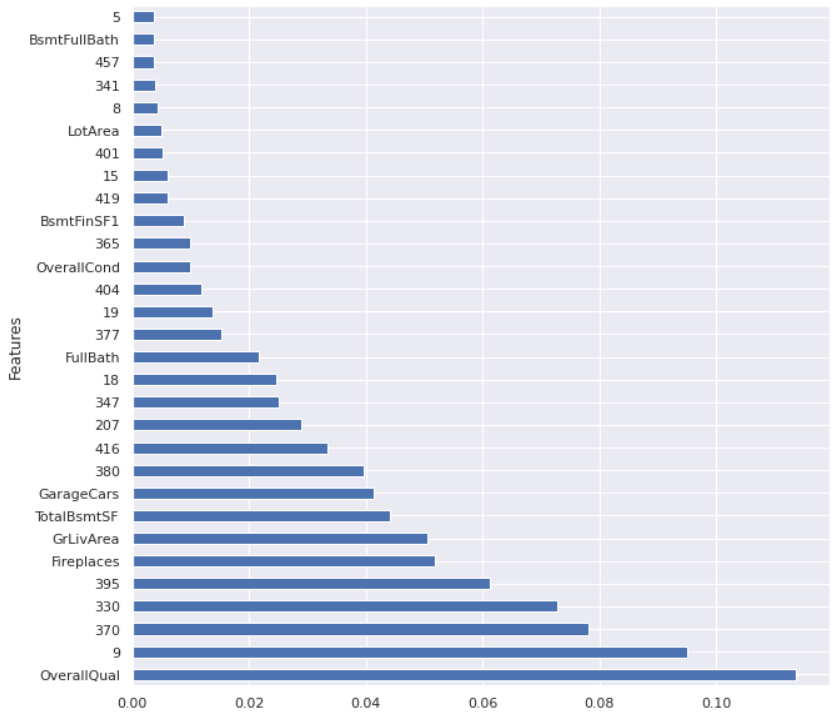
1. **Outliers:**



1. **Model comparison:**



1. **Feature Importance considered by XGBRegressor model:**



* Interpretation of the Results

1. There was multi-correlation between some features
2. The OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath are some of the highly correlated features with the target variable.
3. Ridge, XGBRegressor and GradientBoostingRegressor models showed very good performance.
4. OverallQual, FirePlaces, GrLivArea, TotalBsmtSF, GarageCars, FullBath, OverallCond, BsmtFinSF1, LotArea, BsmtFullBath and some of the encoded features are having influence in predicting the Saleprice while using XGBRegressor model as the Final model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath are some of the highly correlated features with the target variable.

OverallQual, FirePlaces, GrLivArea, TotalBsmtSF, GarageCars, FullBath, OverallCond, BsmtFinSF1, LotArea, BsmtFullBathare some of the most important features that influence the Sale Price of a house in Australia.

* Learning Outcomes of the Study in respect of Data Science

Changing the hyperparameter ranges based on the results will further improve the model performance.

* Limitations of this work and Scope for Future Work

There are only few data for training. The model will perform better by training with more data in future.