Housing: Price Prediction

Submitted by:

Shravani Natakala

ACKNOWLEDGMENT

I would like to extend my sincere and heartfelt gratitude to DataTrained Academy who helped me in this and has always been very cooperative and without their guidance, cooperation and encouragement, the projects couldn't have been what it evolved to be.

Other sources used as a reference for this project are:

https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1

INTRODUCTION

Houses are one of the necessary needs of each and every person around the world and therefore housing and real estate market is one of the markets which highly contribute in world's economy. It is a very large market and therefore has various companies working in this 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. This project 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 company wants to figure out factors influencing the prices of houses. And also want to understand the pricing dynamics of new market so that they can manipulate the strategies and concentrate on the areas that will yield high returns.

In this project we have build a machine learning model predicting sale prices of house using regression algorithms. We have also performed Extensive EDA to gain insights of data. Also the data consists of missing values so we have made few assumptions to fill those null values.

ANALYTICAL FRAMING

Description of each variable

Target Variable:

SalePrice: Price of houses.

Features:

1. MSSubClass: (Categorical) type of dwelling involved in the sale.

2. MSZoning: (Categorical) general zoning classification of the sale.

3. LotFrontage: (Continuous) Linear feet of street connected to property

4. LotArea: (Continuous) Lot size in square feet

5. Street: (Categorical) Type of road access to property

6. Alley: (Categorical) Type of alley access to property

7. LotShape: (Categorical) General shape of property

8. LandContour: (Categorical) Flatness of the property

9. Utilities: (Categorical) Type of utilities available

10. LotConfig: (Categorical) Lot configuration

11. LandSlope: (Categorical) Slope of property

12. Neighborhood: (Categorical) Physical locations within Ames city limits

13. Condition1: (Categorical) Proximity to various conditions

14. Condition2: (Categorical) Proximity to various conditions (if more than one is present)

15. BldgType: (Categorical) Type of dwelling

16. HouseStyle: (Categorical) Style of dwelling

17. OverallQual: (Categorical) Rates the overall material and finish of the house

18. OverallCond: (Categorical) Rates the overall condition of the house

19. YearBuilt: (continous discrete) Original construction date

20. YearRemodAdd: (Continuous) Remodel date (same as construction date if no remodeling or additions)

21. RoofStyle: (Categorical) Type of roof

22. RoofMatl: (Categorical) Roof material

23. Exterior1st: (Categorical) Exterior covering on house

24. Exterior2nd: (Categorical) Exterior covering on house (if more than one material)

25. MasVnrType: (Categorical) Masonry veneer type

26. MasVnrArea: (Continuous) Masonry veneer area in square feet

27. ExterQual: (Categorical) Evaluates the quality of the material on the exterior

28. ExterCond: (Categorical) Evaluates the present condition of the material on the exterior

29. Foundation: (Categorical) Type of foundation

30. BsmtQual: (Categorical) Evaluates the height of the basement

31. BsmtCond: (Categorical) Evaluates the general condition of the basement

32. BsmtExposure: (Categorical) Refers to walkout or garden level walls

33. BsmtFinType1: (Categorical) Rating of basement finished area

34. BsmtFinSF1: (Continuous) Type 1 finished square feet

35. BsmtFinType2: (Categorical) Rating of basement finished area (if multiple types)

36. BsmtFinSF2: (Continuous) Type 2 finished square feet

37. BsmtUnfSF: (Continuous) Unfinished square feet of basement area

38. TotalBsmtSF: (Continuous) Total square feet of basement area

39. Heating: (Categorical) Type of heating

40. HeatingQC: (Categorical) Heating quality and condition

41. **CentralAir:** (Categorical) Central air conditioning

42. Electrical: (Categorical) Electrical system

43. 1stFlrSF: (Continuous) First Floor square feet

44. 2ndFlrSF: (Continuous) Second floor square feet

45. LowQualFinSF: (Continuous) Low quality finished square feet (all floors)

46. GrLivArea: (Continuous) Above grade (ground) living area square feet

47. BsmtFullBath: (Continuous discrete) Basement full bathrooms

48. BsmtHalfBath: (Continuous discrete) Basement half bathrooms

49. FullBath: (Continuous discrete) Full bathrooms above grade

50. HalfBath: (Continuous discrete) Half baths above grade

51. Bedroom: (Continuous discrete) Bedrooms above grade (does NOT include basement bedrooms)

52. Kitchen: (discrete) Kitchens above grade

53. KitchenQual: (Categorical) Kitchen quality

54. TotRmsAbvGrd: (Continuous discrete) Total rooms above grade (does not include bathrooms)

55. Functional: (Categorical) Home functionality (Assume typical unless deductions are warranted)

56. Fireplaces: (Continuous discrete) Number of fireplaces

57. FireplaceQu: (Categorical) Fireplace quality

58. GarageType: (Categorical) Garage location

59. GarageYrBlt: (Continuous discrete) Year garage was built

60. GarageFinish: (Categorical) Interior finish of the garage

61. GarageCars: (Continuous discrete) Size of garage in car capacity

62. GarageArea: (Continuous) Size of garage in square feet

63. GarageQual: (Categorical) Garage quality

64. GarageCond: (Categorical) Garage condition

65. PavedDrive: (Categorical) Paved driveway

66. WoodDeckSF: (Continuous) Wood deck area in square feet

67. OpenPorchSF: (Continuous) Open porch area in square feet

68. EnclosedPorch: (Continuous) Enclosed porch area in square feet

69. 3SsnPorch: (Continuous) Three season porch area in square feet

70. ScreenPorch: (Continuous) Screen porch area in square feet

71. PoolArea: (Continuous) Pool area in square feet

72. PoolQC: (Categorical) Pool quality

73. Fence: (Categorical) Fence quality

74. MiscFeature: (Categorical) Miscellaneous feature not covered in other categories

75. MiscVal: (Continuous discrete) $Value of miscellaneous feature

76. MoSold: (Continuous discrete) Month Sold (MM)

77. YrSold: (Continuous discrete) Year Sold (YYYY)

78. SaleType: (Categorical) Type of sale

79. SaleCondition: (Categorical) Condition of sale

80. Id : Identity number of each house.

The data provides has both object and numeric data types column. It also has null values.

Hardware and Software Requirements and Tools Used

- Laptop (Project done in jupyter notebook)
- scikit-learn
- matplotlib
- pandas
- numpy

Steps Done:

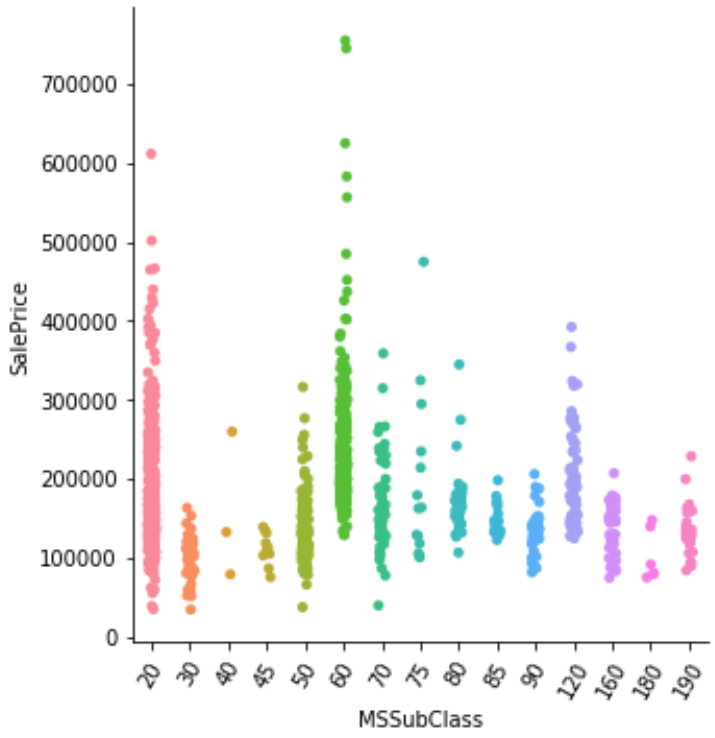
- Data Analytics done by deeply looking into statistical description of data and visualizing distribution of each feature.
- Bivariate analysis also done to determine relationship of each feature with target variable.
- Since data consists of null values the conclusions made from EDA are used to fill those missing data.
- Data pre-processing involves filling all the missing data with appropriate values and encoding all object type variables.
- Data cleaning done by removing certain features which have too little impact on target variable and also have outliers. Further dealt with skewness using various transformation techniques.
- Building of model and using various metrics to determine best model.
- Tuning model to make it more efficient.
- Further performing all the date processing in test data and predicting prices using model build.

Conclusions from EDA

By Univariate analysis we observed the patterns in data to determine how to treat null values. We observed that most of the missing data is because of missing of that feature from house.

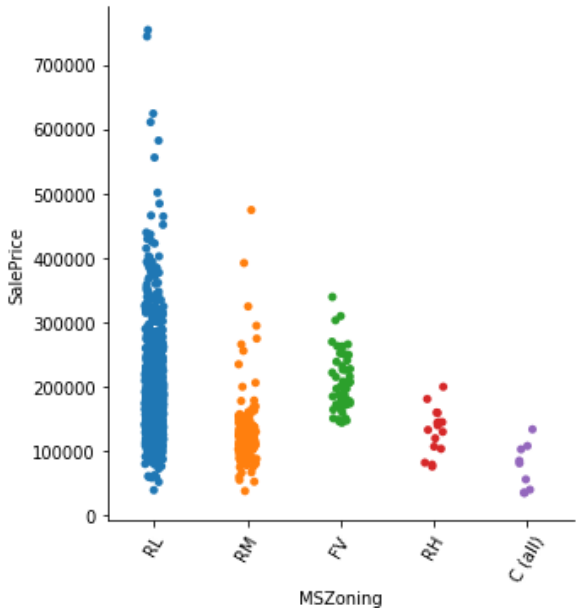
Bivariate Analysis of each variables has been done in order to look closely at how each feature is impacting sales. Below are some features where we observed direct relationship with “SalePrice”.

MSSubClass



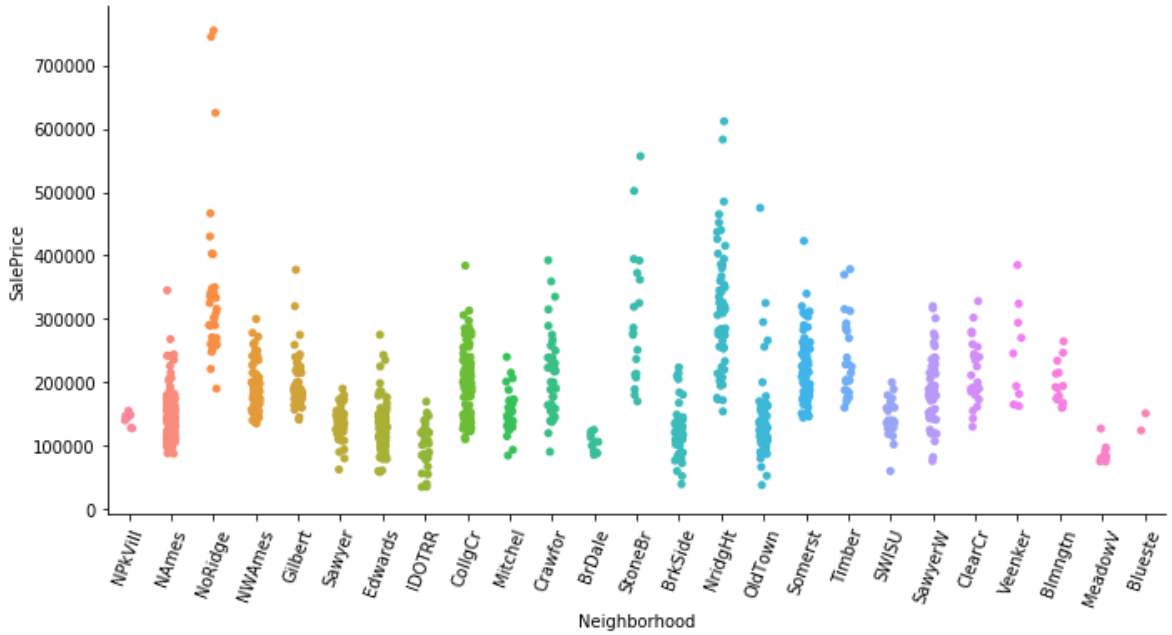
We observe that one storey newer style and 2 storey newer styles may have higher prices.

MSZoning:



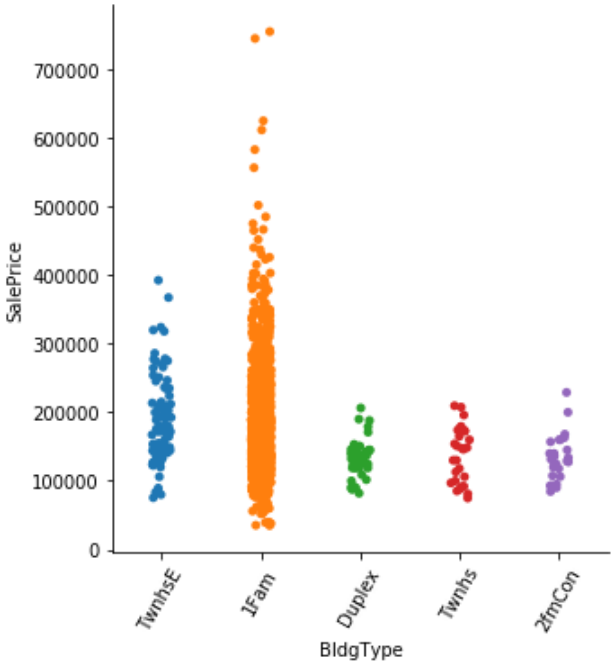
Location of houses in residential spaces with low or medium density impacts the price of house.

Neighborhood:



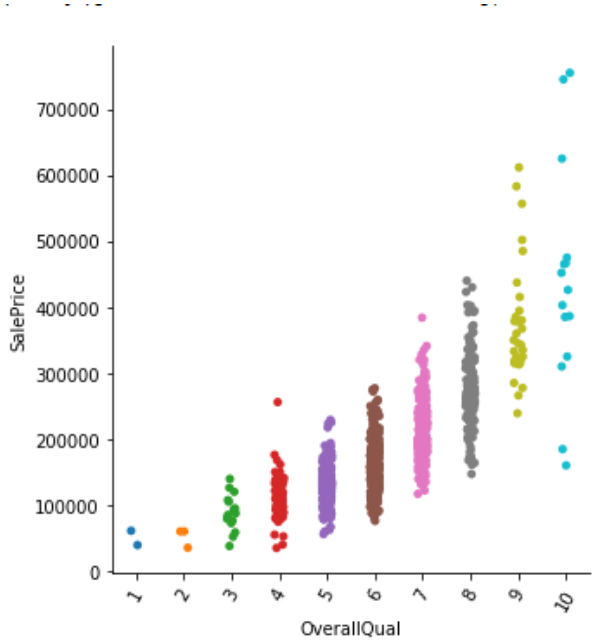
Locality of house also impacts the sale price. Here we observe houses located in northridge area are of higher value.

BldgType:

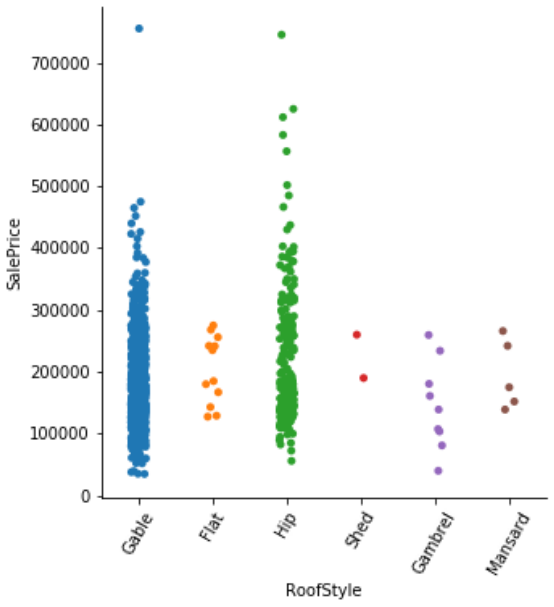


Single family detached type of dwelling is most preferred by people and thus may lead to higher prices.

OverallQual:

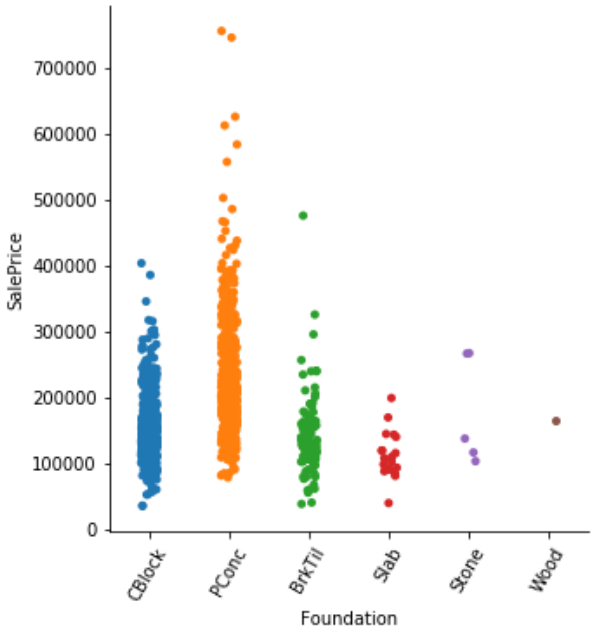


RoofStyle:



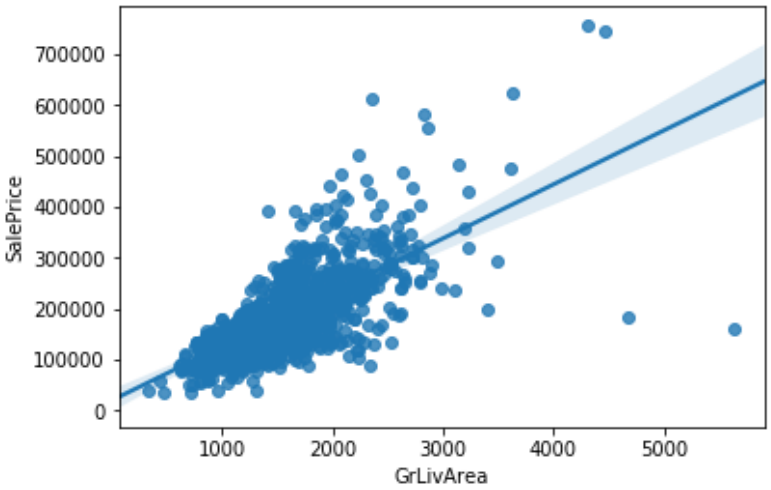
Roof style also impacts the sale price. Here, Hip and Gable are giving higher prices.

Foundation:



Prices maybe high for concrete material used in houses.

GrLivArea:

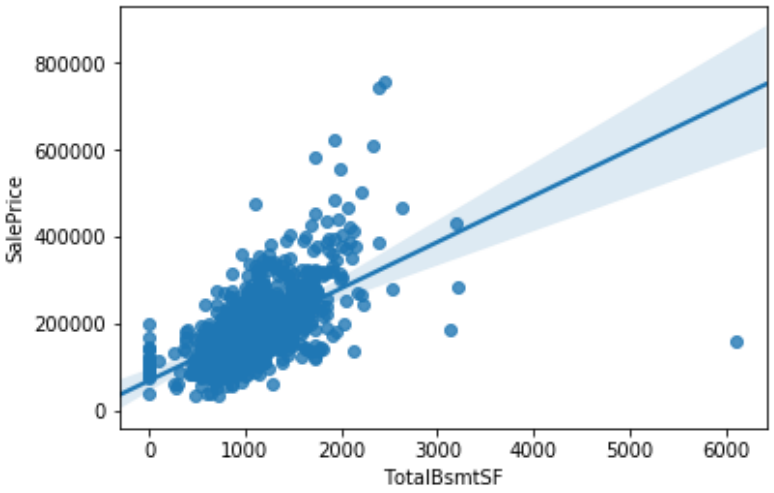


With increase in Living Area sale price also increases.

GarageCars :

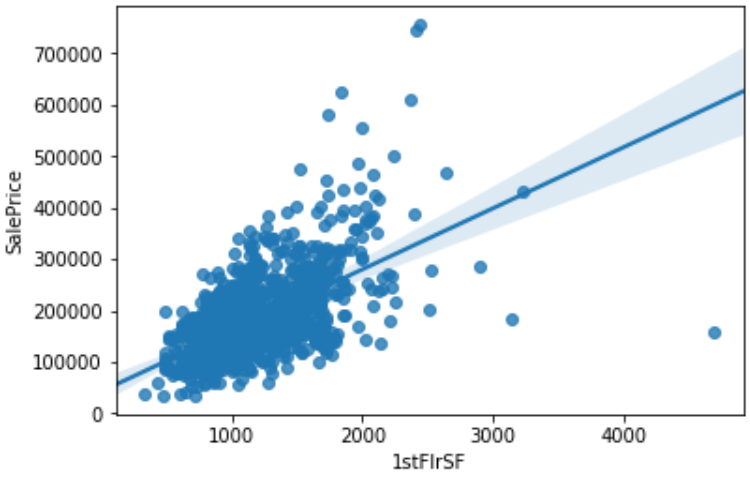
Increase in number of cars indirectly depends on area and thus this feature positively effects sale price.

TotalBsmtSF :



with increase in Area of basement sale price also increases.

1stFlrSF:



with increase in Area of first floor sale price also increases.

Next, we moved to check for correlation and observed very less multicollinearity between variables. Also the maximum correlation was observed with overall quality of house which was 78.9%.

Further we observe outliers in these columns : LotFrontage, LotArea, MasVnrArea, BsmtFinSF2, GarageArea, GrLivArea, WoodDeckSF, OpenPorch, EnclosedPorch, ScreenPorch and 3SsnPorch.

Data pre-processing and cleaning

Data preprocessing here involves dealing with null values and encoding object type variables to make it machine learning model friendly. For null values we observed that values are missing because of feature not present in house therefore we have to make null as category and thus we make an assumption that if data is missing it means that feature is missing in houses.

Since ‘utilities’ had only one value and ‘Id’ is identity number thus not making any effect our target variable so we will delete them. Now further looking at correlation of each feature and outliers contained in data we decide to remove certain features.

"BsmtHalfBath","EnclosedPorch","3SsnPorch","ScreenPorch","PoolArea","MasVnrType", "BsmtFinSF2" and "MiscVal".

Next I have removed skewness in data using square root transformation.

MODEL BUILDING AND EVALUATION

Algorithms used are:

- Linear Regression (Model regularization using Lasso CV)
- Decision Tree Regression
- KNeighbor Regression
- Random Forest Regression

Since range of target variable is too high we are getting a high value of mse therefore we will only look for R2 score and CV Score to determine our best model.

Linear Regression: R2 Score = 84%
CV Score = 67.49%

Lasso CV: R2 Score = 84.36%
CV Score = 68.62%

KNeighbor Regression : R2 Score =75.73%
CV Score = 70.23%

Decision Tree Regression: R2 Score = 75.99%
CV Score = 74.17%

Random Forest Regression: R2 Score = 85.94%
CV Score = 83.80%

Model Evaluation:
Difference between R2 Score and CV Score for Linear Regression is: 16.55%
Difference between R2 Score and CV Score for KNN Model is: 5.49%
Difference between R2 Score and CV Score for Lasso CV is: 15.74%
Difference between R2 Score and CV Score for Decision Tree Regressor is: 1.81%
Difference between R2 Score and CV Score for Random Forest Regressor is: 2.13%

By looking at accuracy score and CV score we choose Random forest regression model as our best model and now we will move forward to tune it.

Hyper Parametric Tuning

After parametric tuning of random forest model using Grid Search CV we obtain following results:

RandomForest regression:

Accuracy = 88.16%
Root Mean Squared Error= 25450.76284324928
Mean Absolute Error= 17743.409360730595

In the last step we have loaded the test data and performed all the processing and cleaning done in training set to predict the prices of house and saved that predicted prices in a csv file.

Conclusions

Key Findings and Conclusions of the Study

Features strongly influencing the Sale Price of houses are:

Overall Quality, Ground Living Area, Garage Cars, Garage Area, Total Basement Area, 1st Floor Area, Full Bathrooms, Total Rooms above Ground, Year Built, Year Remodeled (if d one), Masonry Veneer Area, Fireplaces, Garage Type, Garage Finished or not, Kitchen Qu ality, Basement Quality and Exterior Material Quality

We have deleted a few features which somehow where not contributing to our model. They are: Basement Half Bathroom, Enclosed Porch Area, 3 Season Porch Area, Screen Porch Area, Pool Area, Masonry Veneer Type, Basement Finished Area Type 2, Utilities, Id and Miscellaneous feature Value

The model build gives an accuracy of 88.16% after hyper parametric tuning.

Learning Outcomes of the Study in respect of Data Science

This project helped me to learn how to deal with missing data and do visualizations to gain insights on data. It helped me to gain conclusions from graphs. Also it helped me in exploring multiple algorithms and metrics to get the best output.

Limitations of this work and Scope for Future Work

Since the data keeps changing we cannot fully rely on this project in the distant future we need to update it with updating in data. Also this project is done with limited sources more algorithms were not explored. And there was no information on null values so we made an assumption that missing data means missing feature.