

**Housing Price Prediction**

Submitted by:

**Yuvarani N**

**ACKNOWLEDGMENT**

The first and almost main source which helped me to work on this project is my passion towards Data Science and understandings about the data science prototypes and machine learning models which I gained from the DataTrained live classes provided by Shankar sir. I would like to thank Shankar sir for providing the clear-cut information about each topic in a very easy and understandable manner.

The experience which I gained from working on practice and evaluation projects from Data trained also helped me a lot. I also used to refer Kaggle very often for discussion about the projects and getting to understand and description from different perspectives about the data set. FlipRoboTechnologies is offering me the real-time internship projects to get the real industry experience in the data science world. I would like to thank Flip Robo for offering such projects to gain and improve our knowledge towards data science world.

**INTRODUCTION**

* Business Problem Framing

Houses are one of the necessary need 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. 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 below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

* Conceptual Background of the Domain Problem

Real estate business is growing very well in recent days. Based on my understanding, when the properties are new and it is located in the residential areas with all the public utilities, then the cost of the property will be significantly higher. So the most deciding factors will be the year it is built and the area it is located and the lot area size and the utilities available at the property.

* Motivation for the Problem Undertaken

The Motivation part I felt here is, this project will help to 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

In this problem we have a target column which is sales price of the property, it is continous data. Hence we need to approach this problem with regression model.

`I have also noticed that few features with outliers and data skewness, so I have reduced them using PowerTransformer as much as possible, and then I have built various regression algorithms and based on the various metrics found that XGBRegressor provides the best accuracy up to 89%.

* Data Sources and their formats

The data was provided to me by Flip Robo and the file was in a excel format. There were 1168 rows of data and 80 different features in test data and 292 records in test data. In order to predict the sale price of the property, I have built the regression model.

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

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

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

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

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

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

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

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

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

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

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

* Data Pre-processing Done

1. Checked for null values, and removed the features which are having more than 50% of null values. Handled the remaining null values with respective mean or mode values.
2. Dropped the unnecessary columns such as (Id)
3. Applied PowerTransformer to reduce the skewness.
4. Checked for multi-collinearity issue, and found few features had multi-collinearity issue and removed those features.

* Data Inputs- Logic- Output Relationships

1. Plotted distribution plot , to find the data distribution is normal or skewed.
2. Plotted scatter plot to find the relationship between the continous data features and the target variable saleprice.
3. Plotted bar plot to find the relationship between the categorical feature and the target variable.

* Hardware and Software Requirements and Tools Used

The is the bare minimum required to run this project

Hardware:

• Process: Intel core i5 and above

• RAM: 4GB and above

• SSD: 250GB and above

Software:

• Anaconda (Jupyter Notebook)

Libraries:

**• import pandas as pd**: pandas is a popular Python-based data analysis toolkit which can be imported using import pandas as pd.

It presents a diverse range of utilities, ranging from parsing multiple file formatsto converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.

**• import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

• **import seaborn as sns**: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas’ data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

• Import matplotlib.pyplot as plt: matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure,creates a plotting area in a figure, plotssome linesin a plotting area, decorates the plot with labels, etc.

* + from sklearn.linear\_model import LinearRegression
  + from sklearn.ensemble import RandomForestRegressor,GradientBoostingRegressor,AdaBoostRegressor
  + from sklearn.svm import SVR
  + from sklearn.neighbors import KNeighborsRegressor
  + from sklearn.tree import DecisionTreeRegressor
  + from sklearn.metrics import r2\_score,mean\_absolute\_error,mean\_squared\_error
  + from sklearn.model\_selection import train\_test\_split,cross\_val\_score
  + import xgboost
  + from xgboost import XGBRegressor

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Found that few features with outliers and removed them to improve the model accuracy.

* Testing of Identified Approaches (Algorithms)

Since our target variable is a continous column ,I have used below regression algorithms and used ‘R2\_score’, ‘cross\_val\_score’ to calculate the model accuracy and find the best model out of the below models.

• LinearRegression

• RandomForestRegressor

• GradientBoostingRegressor

• AdaBoostRegressor

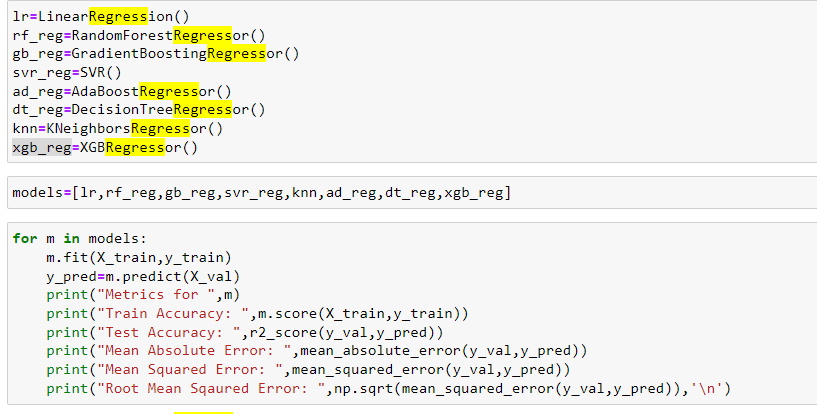
• KNeighborsRegressor

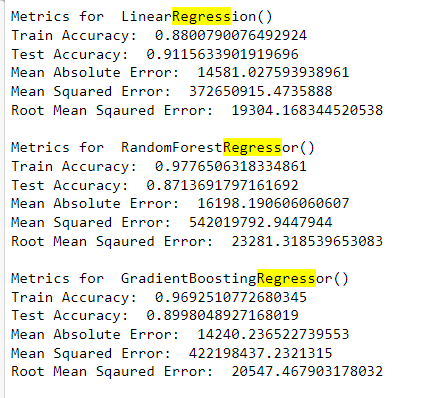
• DecisionTreeRegressor

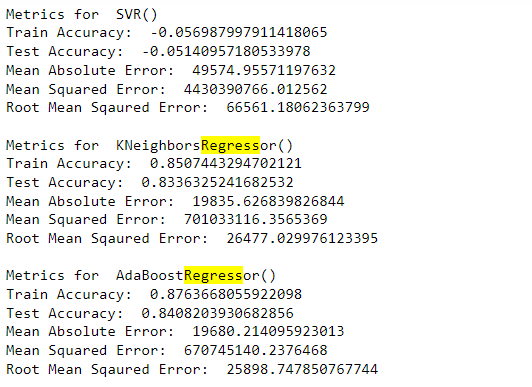
• SVR

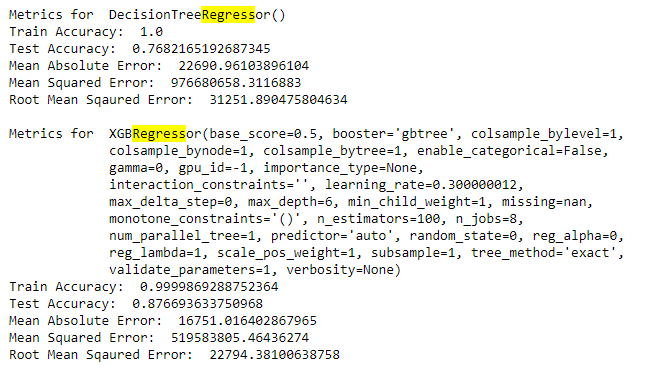
• XGBRegressor

* Run and Evaluate selected models

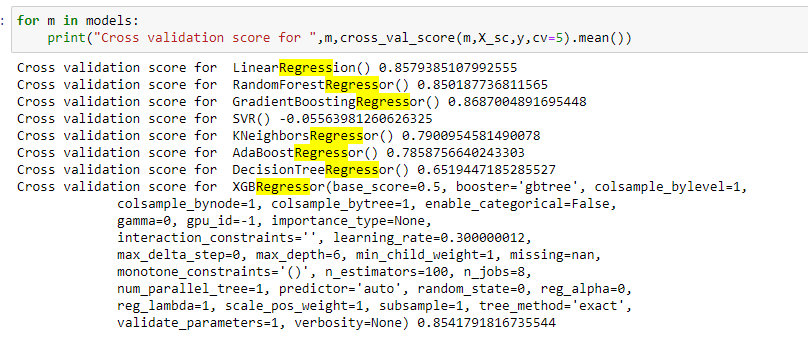




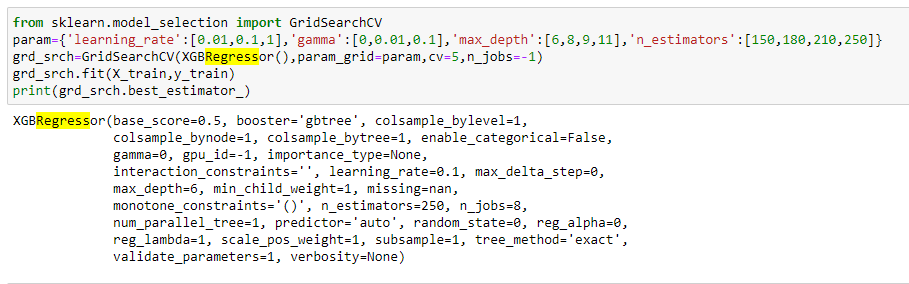


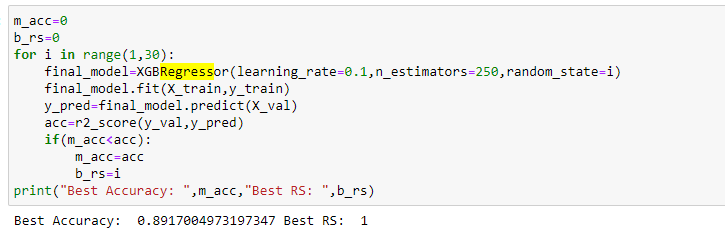


**Cross validation check:**



**Hyper Parameter Tuning:**





**Based on the model accuracy and cross validation XGBRegressor is having less difference. Selected XGBRegressor as the final model to our dataset.Hyper Parameter tuning helped to increase the accuracy from 87% to 89%**

* Key Metrics for success in solving problem under consideration

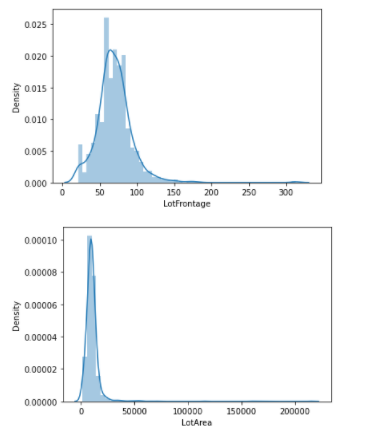
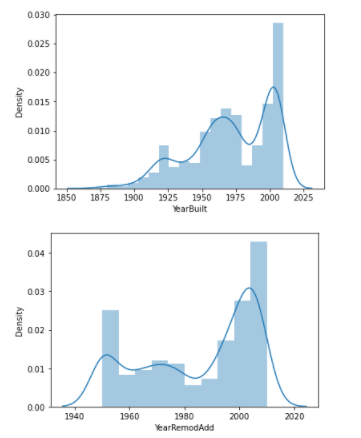
I have used the following metrics for evaluation:

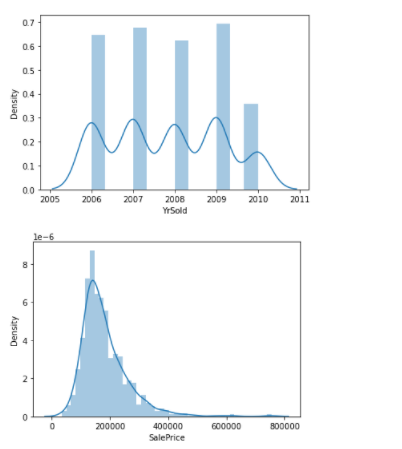
**• R2 score:** R-squared is a statistical measure that represents the goodness of fit of a regression model. **The ideal value for r-square is 1.**

• **Cross\_val\_score**: To run cross-validation on multiple metrics and also to return train scores, fit times and score times. Get predictions from each split of cross-validation for diagnostic purposes. Make a scorer from a performance metric or loss function.

* Visualizations

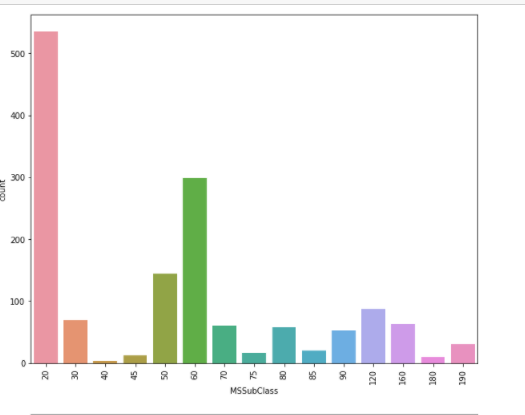
**Univariate analysis:**

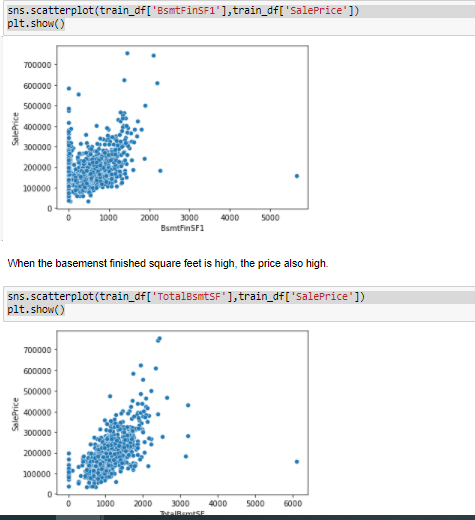
 



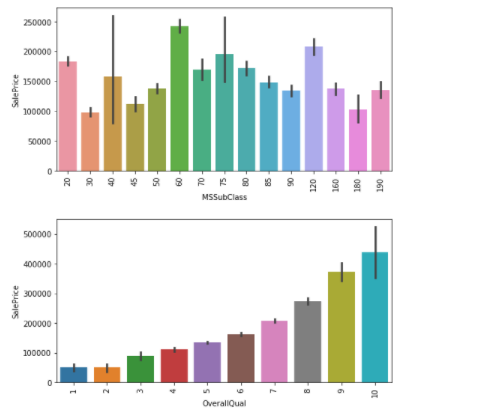
Based on the above distribution plot diagrams it is clear that almost all the features are having skewed data.

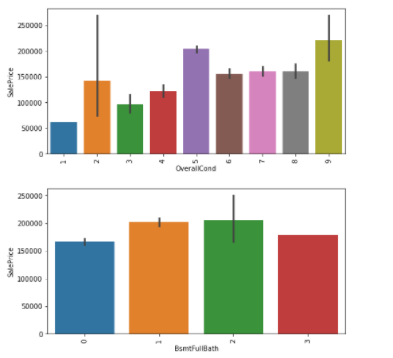
**Bivariate analysis:**

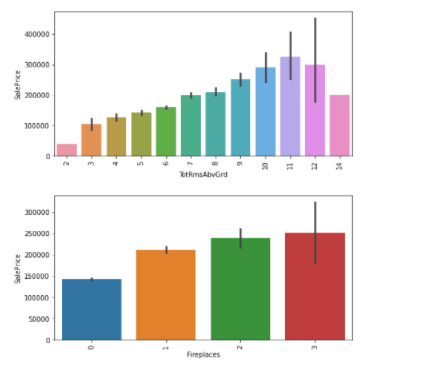


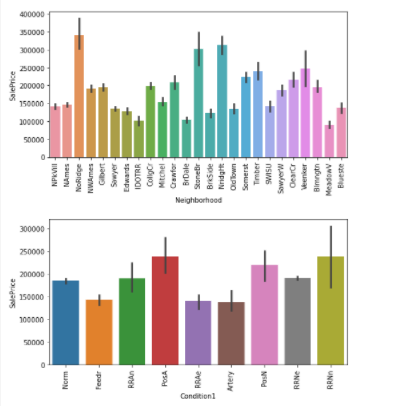


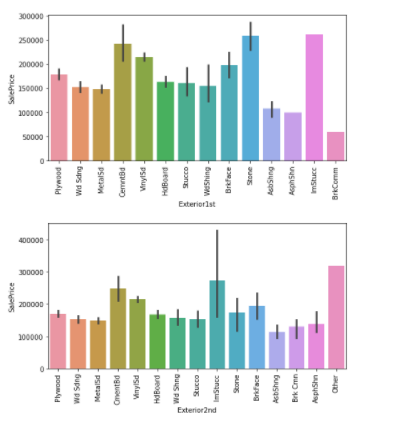


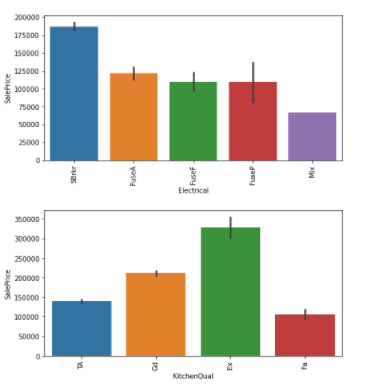


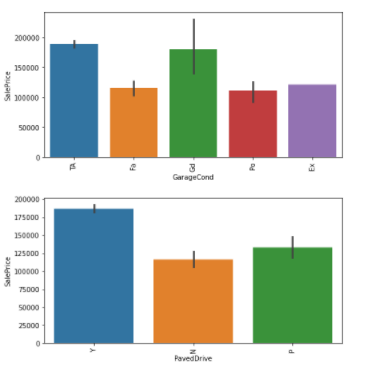


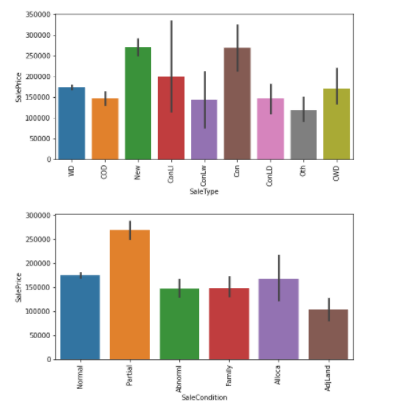












Observations from above plots:

1. There are majority of the properties are belongs to 1-STORY 1946 & NEWER ALL STYLES and followed by 2-STORY 1946 & NEWER

2. The quality of overall material and finish of the house is mostly average and followed by above average.

3. Majority of the properties are having average condition.

4. Most of the properites does not have full bathroom in basement and the second higest is with one bathroom at basement.

5.Majority of the properites does not have half bathrooms at basement.

6. Most of the properties are having 2 bathrooms above the basement and follwed by few properties with 1 bathroom.

7. Mostly there is no half bathrooms above basement and some properties with 1 half bathrooms.

8. Majority of the properites are having 3 bedrooms above the grade.

9. 1 kitchen is available for most of the properties.

10. Most of the properties are having 6 rooms above the basement and followed by properties with 7 rooms.

11. Most of the properties are having no fire place and followed by 1 fire place.

12. Most of the properites are with grage size of 2 cars.

13. Most of the properties are from the Residential Low Density zone.

14. Most of the properites are having paved street to the property.

15. Most of the properties are in regular shape

16. Majority of the properties are in near flat same level land counter

17. Almost the proerties are having all public utilities available at the property.

18. Majority of the properties are belongs to indie lot.

19. Majority of the properties are having gentle slope.

20. Most of the properites are Single-family Detached type of dwelling.

21. Most of the properties are belongs to 1 story house style

22. Majority of the properites are having Gable style roofs.

23. Mostly the roof material is Standard (Composite) Shingle.

24. Mostly the exterior covering is Vinyl Siding.

25. Most of the properties are with average quality of the material on the exterior.

26. Most of the properties are with average exterior condition.

27. Mostly the properties are having foundation type as Poured Contrete or Cinder Block.

28. Most of the time, the height of basement is average and good.

29. Most of the times the basement condition is Average.

30. Mostly the type of heating is Gas forced warm air furnace.

31. Most of the times Heating quality and condition is excellent.

32. Majority of the properties are having central Air conditioning.

33. Mostly the properties are having Standard Circuit Breakers & Romex electrical system.

34. Most of the times, the kitchen quality is average.

35. Mostly the Home functionality is typical.

36. Mostly the Garage is attached from home.

37. Most of the times the Interior finish of the garage is unfinished.

38. Mostly the garage quality is average.

39. Mostly the garage condistion is average.

40. Mostly the drive way is paved.

41. Mostly the type of sale is Warranty Deed - Conventional.

42. Mostly the sale condition is Normal.

* Interpretation of the Results

Observations:

1. When the type of property is 2-STORY 1946 & NEWER the price is higher and when the type is 1-STORY 1945 & OLDER the price is low.

2. When the overall material and finish of the house is excellent, the price is high.

3. When the garage size is big, the price of the property is high.

4. The price of the property is higher in the zones of Floating Village Residential and Residential Low Density.

5. Almost all the properties are having all public utilities available.

6. When the lot configuration is Cul-de-sac and Frontage on 3 sides of property, the the price of the property is high.

7. When the type of house is Frontage on 3 sides of property and Two story, then the price is more.

8. When the roof material is Wood Shingles then the price is more.

9. When the Exterior covering is Stone and Imitation Stucco the price of the house is higher.

10. When the Masonry veneer type is Stone, the price of the property is high.

11. When the quality of the material on the exterior is excellent, the price is higher.

12. When the present condition of the material on the exterior is excellent, the price of the property is high.

13. When the type of the foundation is Poured Contrete then the price is higher.

14. When the height of the basement is excellent, the price of the property is high.

15. When the basement finished area is Good Living Quarters, the price of the property is higher.

16. When the garage type is Built-In (Garage part of house - typically has room above garage), the price is higher.

17. When the garage quality is excellent, the price is higher.

18. When the sale type is Contract 15% Down payment regular terms and Home just constructed and sold then the price is higher.

19. When the sale condition is partial(Home was not completed when last assessed (associated with New Homes)), the price is higher.

We able to successfully create a model with 89% of accuracy score.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project, I have used machine learning algorithms to predict the sales price of the property. I have mentioned the step-by-step process that I have done to get the best accuracy score of my model. After performing all the steps, I have found that “XGBRegressor ” fits well for our model and we able to build model with best accuracy score of 89%.

* Learning Outcomes of the Study in respect of Data Science

Visualization is such a powerful tool, which helps us in interpreting the dataset in a very easy manner.

This project helped me in breaking a task into sub-tasks and working on each part individually to easily complete the problem without any hassle.