HOUSING PROJECT

SUBMITTED BY:-

RISHABH JOHRI

DATA SCIENCE-INTERN

ACKNOWLEDGEMENT:-

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a learning dataset and it has helped me to improve my analytical skills. And I want to express my gratitude to Ms. Gulshana Chaudhary , the

(SME of Flip Robo Technologies), she is the person who has helped me to get out of all the difficulties which I faced during the project. She also encouraged me a lot with her unconditional support I have ended up with this Project.

HOUSING PROJECT FINDINGS:-

1. 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 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 onchanging trends in housing sales and purchases. Predictive modeling, Market mix modeling, 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.

House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. House Price prediction, is important to drive Real Estate efficiency. As earlier, Houseprices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency. The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analysing previous market trends and price ranges, and also upcoming developments the future prices will be predicted as per cost of property depending on number of attributes considered.

Conceptual Background of the Domain Problem:

The real estate market is one of the most competitive in terms of pricing and same tends to vary significantly based on numerous factors; forecasting property price is an important module in decision making for both the buyers

and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies. 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?

Why is house price prediction important?

- House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improving the Real Estate efficiency.

There are three factors that influence the price of a house which includes:- Physicalconditions, concept and the Location. Hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore, in this project report we present various important features to use while predicting housing prices with good accuracy. While using features in a regression model, some feature engineering is required for more better prediction.

Review of Literature:-

The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is 6 attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming.
The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Therefore, in this project report, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model, some feature engineering is required for better prediction.

The primary aim of this report is to use these Machine Learning Techniques and convert them into Machine Learning models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, A seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house.

Motivation for the Problem Undertaken:

In this, we have 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. The relationship between house prices and the economy is an important motivating factor for predicting house.

2.Analytical Problem Framing:-

Mathematical / Analytical Modelling:-

This particular problem has two datasets one is train data-set and the other is test data-set. I have built model using train data-set and predicted Sale Price for test dataset. By looking into the target column, I came to know that the entries of Sale Price column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 80% null values and more than 85% zero values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. While checking the null values in the datasets I found many columns with Nan values and I replaced those nan values with suitable entries like mean for numerical columns and mode for categorical columns. To get better insight on the features I have used plotting like distribution plot, bar plot, reg plot and strip plot. With these plotting I was able to understand the relation between the features in a better way. Also, I found outliers and skewness in the dataset so I removed outliers using ZScore method and removed the skewness using Yeo-Johnson method. I have used all the regression models while building model, then tuned the best model and saved the best model. At last, I have predicted the sale price for test dataset using the saved model of train dataset.

Data Sources and their formats:-

The data collected and provided to me by internship company – Flip Robo Technologies in csv file format. The Company provided me two datasets one is train and other is test. I built model using train dataset and predicted Sale Price

using test dataset. The train dataset was having 1168 rows and 81 columns including target, and the test dataset was having 292 rows and 80 columns excluding target. In this particular dataset I have object, float and integer types of data and did analysis of both the datasets separately for effective and efficient process ahead.

Data PreProcessing :-

¬ As a first step I have imported all required libraries and I have imported both the datasets which were in csv format.

¬ Then I did all the statistical analysis like checking shape, nunique- for unique values, value counts, info etc.

While checking the info of the datasets I found some columns with more than 80% null values, so these columns will create skewness in datasets so I decided to drop those columns.

¬ Then while looking into the value counts I found some columns with more than 85% zero values this also creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 85% zero values.

¬ While checking for null values I found null values in most of the columns and I have used imputation method to replace those null values (mode for categorical column and mean for numerical columns).

¬ In Id and Utilities column the unique counts were 1168 and 1 respectively, which means all the entries in Id column are unique and ID is the identity number given for particular asset and all the entries in Utilities column were same so, these two column will not help us in model building. So, it is decided to drop these columns.

¬ Next as a part of feature extraction I converted all the year columns to there respective age. Thinking that age will help us more than year.

¬ And all these steps were performed to both train and test datasets separately.

Data Inputs , Logic and Output Relations:-

1. I have used box plot for each pair of categorical features that shows the relation with the median sale price for all the sub categories in each categorical feature.

2. And also for continuous numerical variables I have used reg plot to show the relationship between continuous numerical variable and target variable.

3. I found that there is a linear relationship between continuous numerical variable and Sale Price.

Hardware and Software Requirements and Tools Used:-

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware used:

- 1. Processor — core i5 2. RAM — 8 GB 3. HDD — 1TB.

Software Required: - 1. Anaconda.

Libraries Used:-

Import numpy as np

Import pandas as pd

Import seaborn as sns

Import matplotlib.pyplot as plt

Import warnings

warnings.filterwarnings(“ignore”)

Data Analysis and Visualization:-

Identification of Possible Problem Solving Approach (Methods):

Here, i have used imputation method to replace null values. To remove outliers I have used Zscore method. And to remove skewness, Yeo-Johnson method is used. To encode the categorical columns, I have used Ordinal Encoding. And in this, the correlation between dependent and independent features is checked. Also I have used standardization, then followed by model building with all regression algorithms.

Testing of Identified Algorithms:-

Here, Sale Price was my target and it is a continuous column so this problem is regression problem. And I have used all regression algorithms to build my model. By looking into the difference of r2 score and cross validation score, I found GradientBoostingRegressor as a best model with least difference between Cv and R2 Score, Also to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have go through cross validation. Below are the list of regression algorithms I have used in my project:-

- GradientBoostingRegressor.
- RandomForestRegressor.
- KNeighborsRegressor .
- DecisionTreeRegressor.
- ExtraTreesRegressor.

Key Metrics for success in solving Problem Under consideration:-

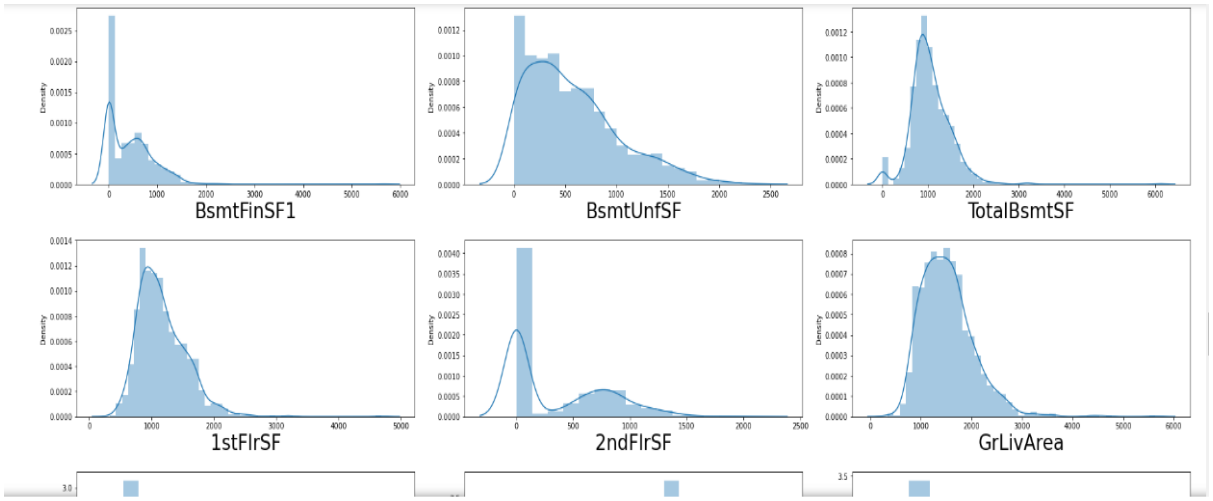
I have used the following metrics for evaluation:

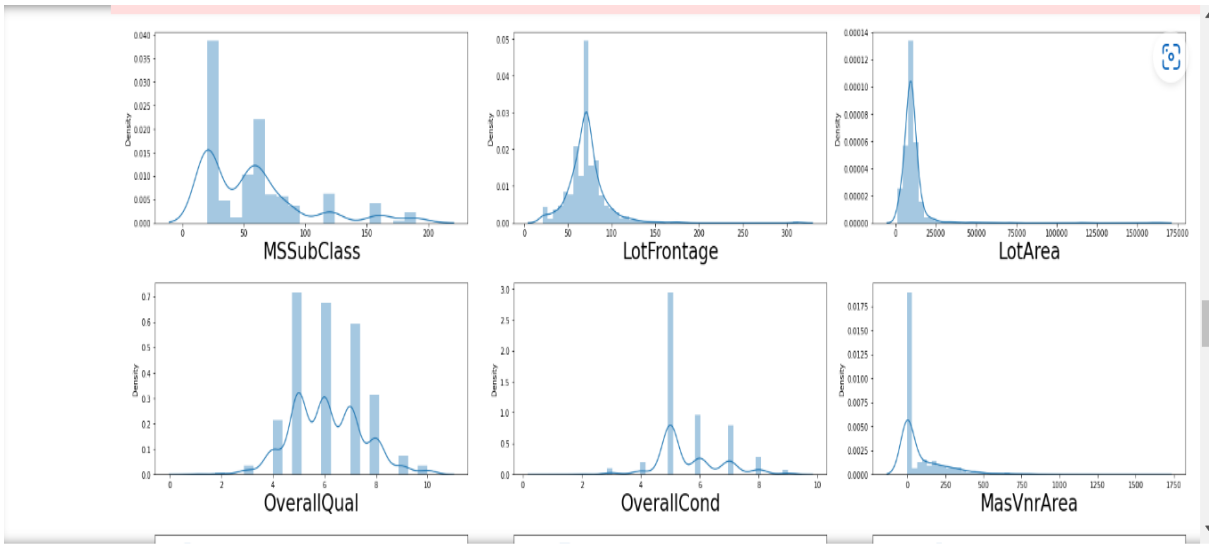
• I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.

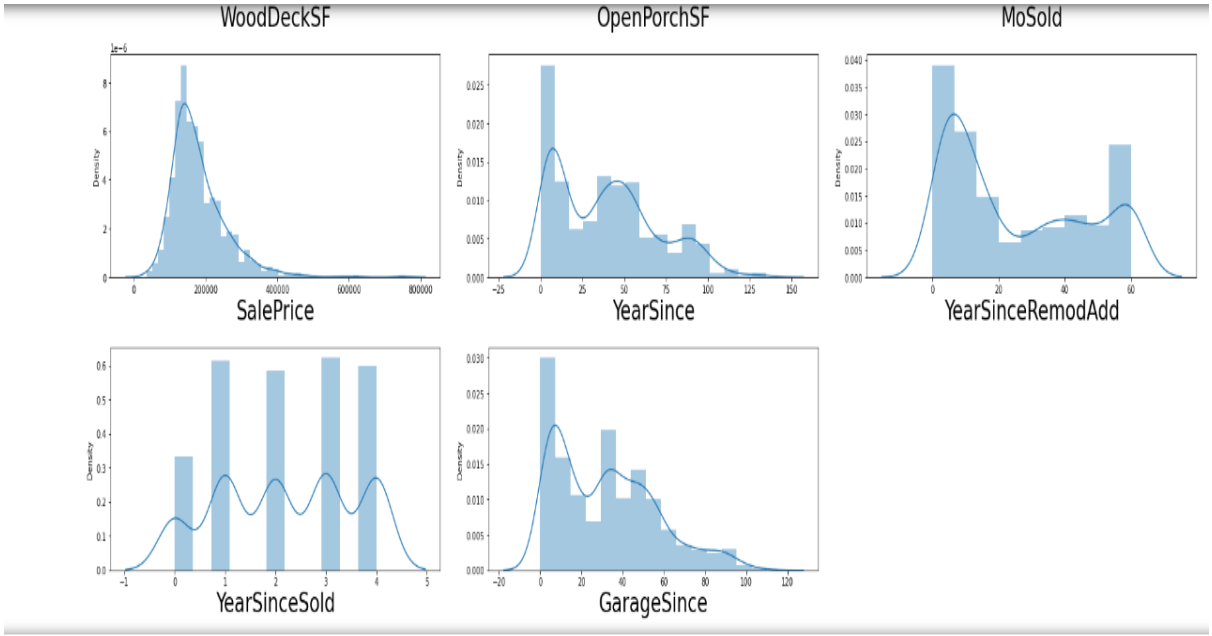
• I have used root mean square deviation which is one of the most commonly used measures for evaluating the quality of predictions. • I have used r2 score which tells us how accurate our model is.

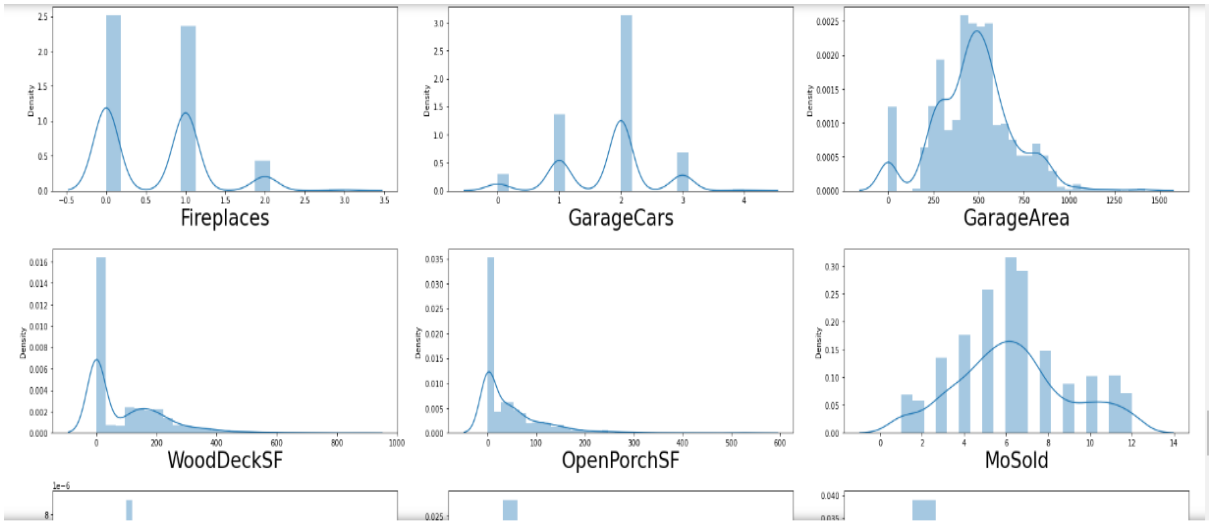
VISUALISATIONS:-

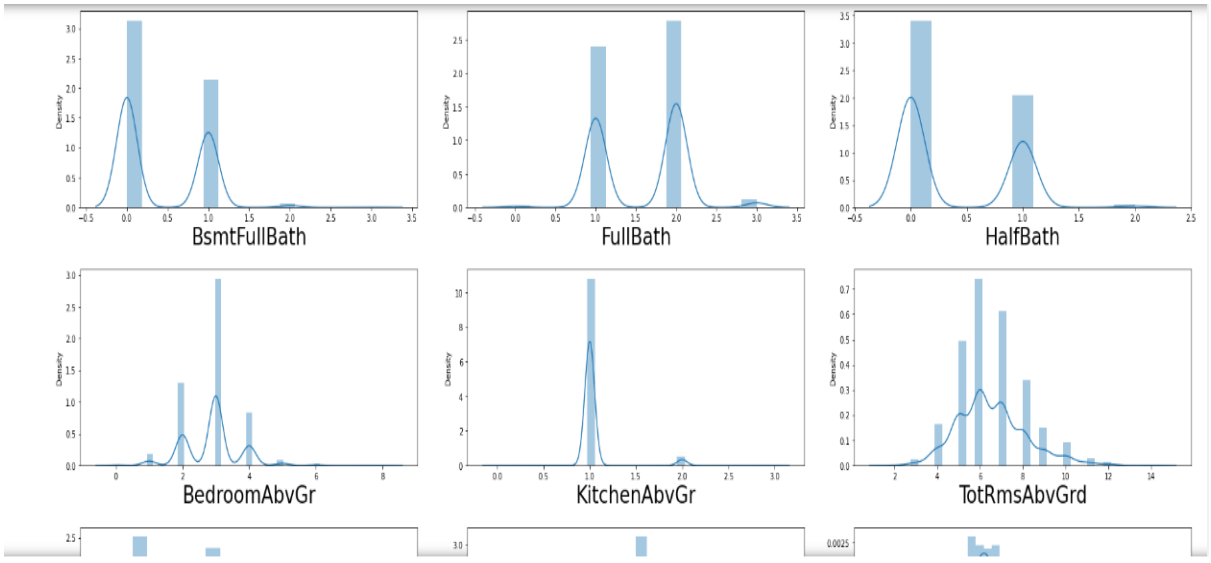
Visualisation of Numerical data using Dist plot method:-



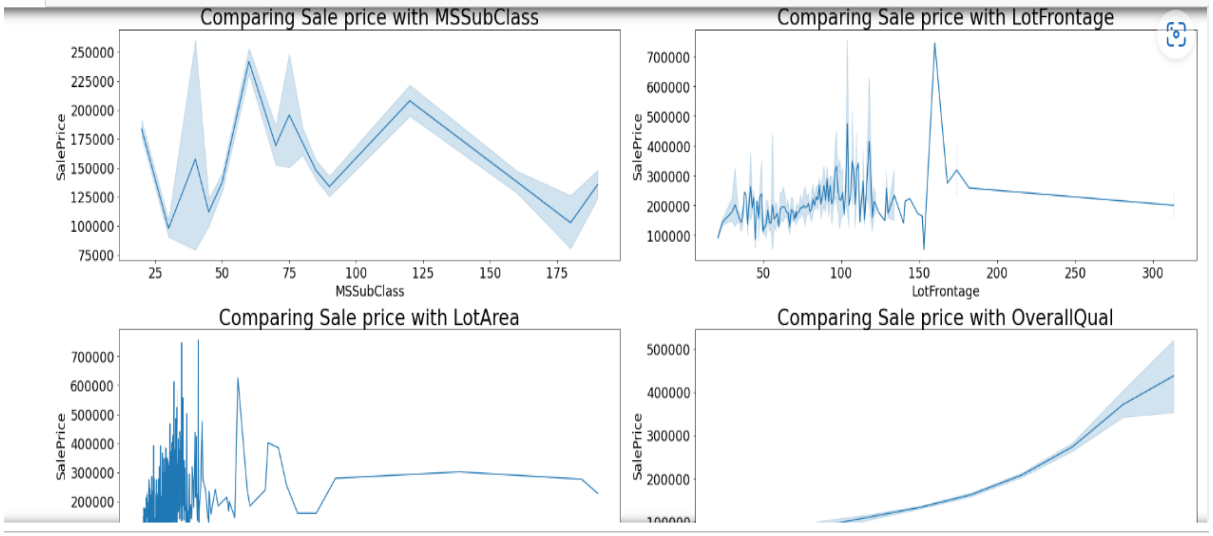


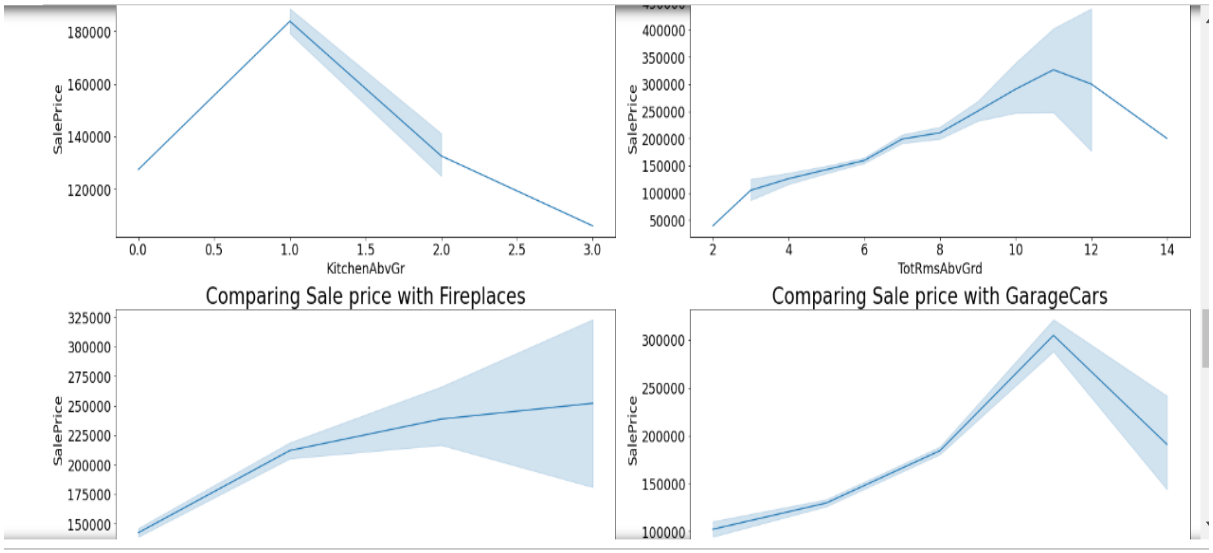

WoodDeckSF
OpenPorchSF
MoSold
SalePrice
YearSince
YearSinceRemodAdd
YearSinceSold
GarageSince

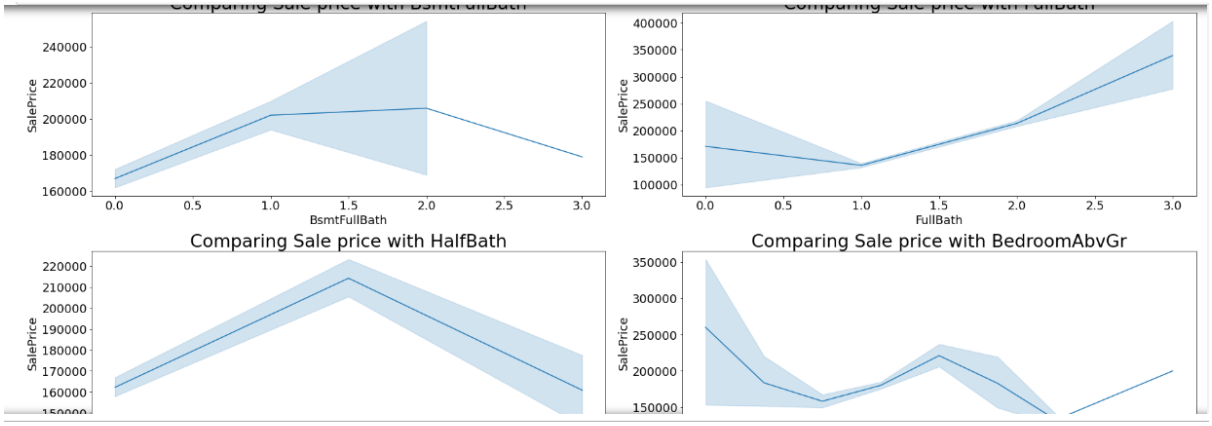

Fireplaces
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
MoSold


BsmtFullBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd

COMPARING NUMERICAL DATA WITH TARGET COLUMN:-





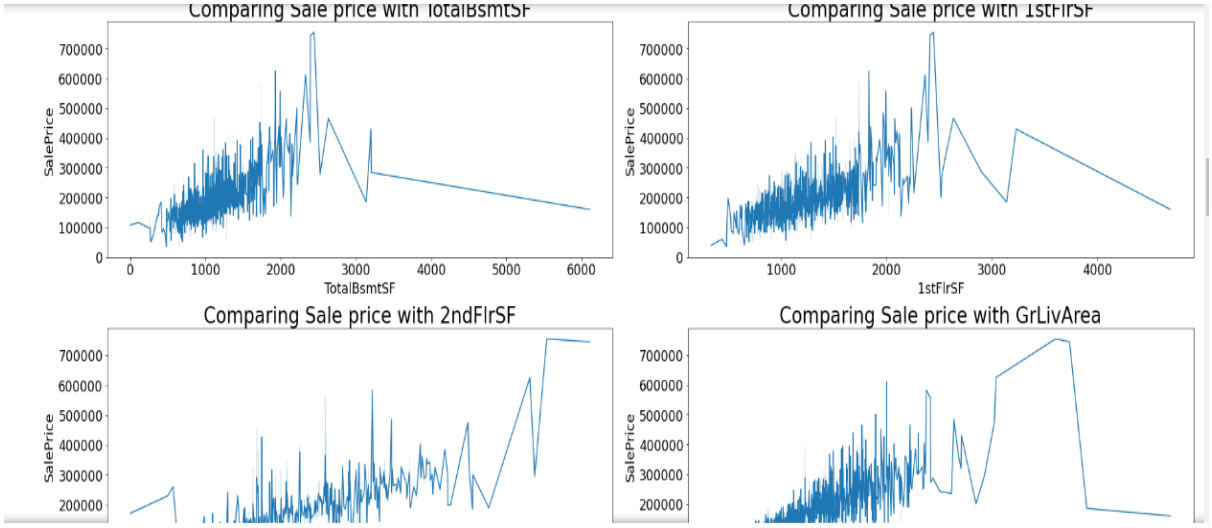


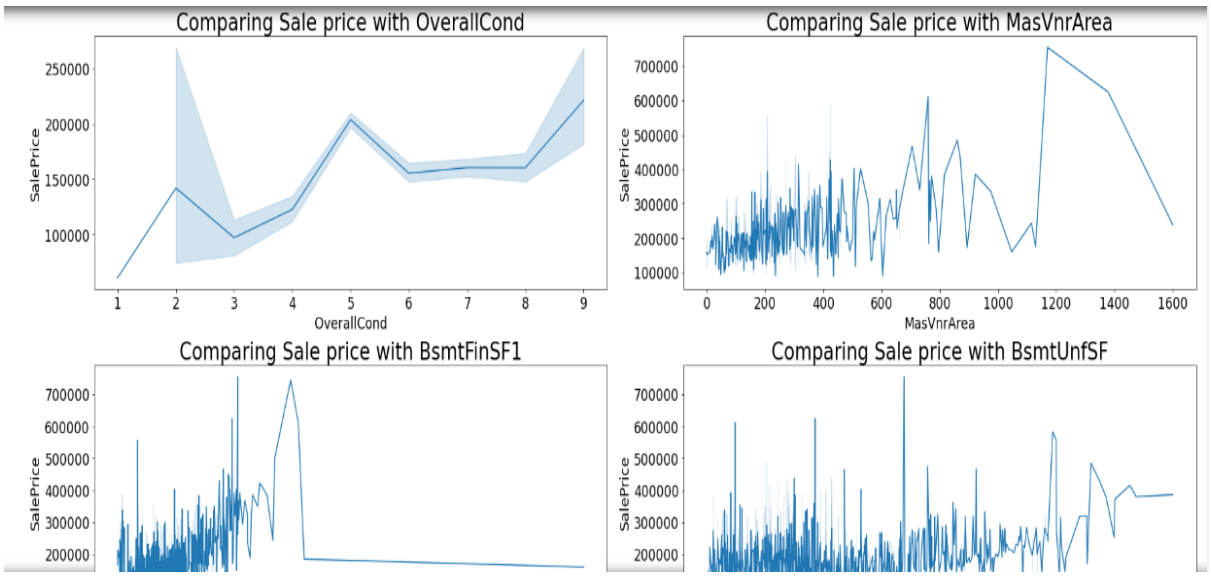
1.In the Graph, the sales is good and SalePrice is also high 1-STORY 1946 & NEWER ALL STYLES(20) and 2-STORY 1946 & NEWER(60) types of dwelling(MSSuubClass).

2.In the Graph, graph in Linear feet of street connected to property(LotFrontage) most of sales are happening below 135 feet but best SalePrice is between 160 to 180 feet.

3.In 3rd Graph, Lot size in square feet(LotArea) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.

4.In 4th Graph, Rates the overall material and finish of the house(OverallQual) is increasing linearly sales is also increasing And SalePrice is also increasing





Observations:-

1.In the Graph, the sales is good and Sale Price is also high 1-STORY 1946 & NEWER ALL STYLES(20) and 2-STORY 1946 & NEWER(60) types of dwelling(MS Sub Class).

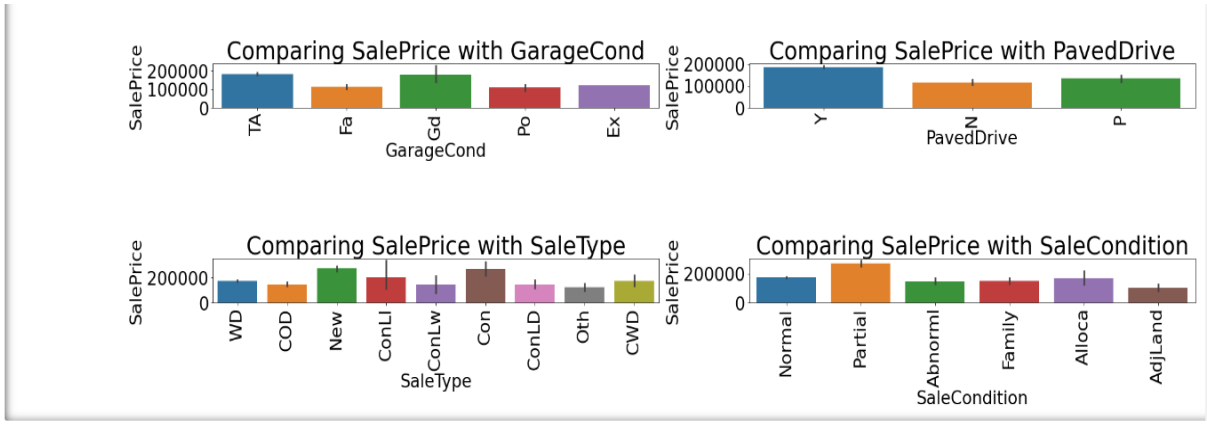
2.In the Graph, graph in Linear feet of street connected to property(LotFrontage) most of sales are happening below 135 feet but best SalePrice is between 160 to 180 feet.

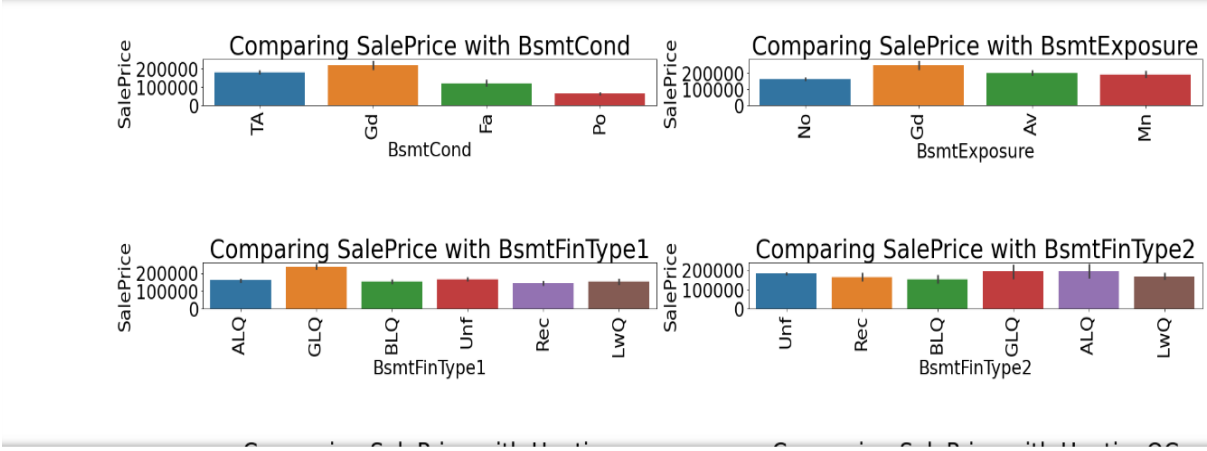
3.In 3rd Graph, Lot size in square feet(LotArea) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.

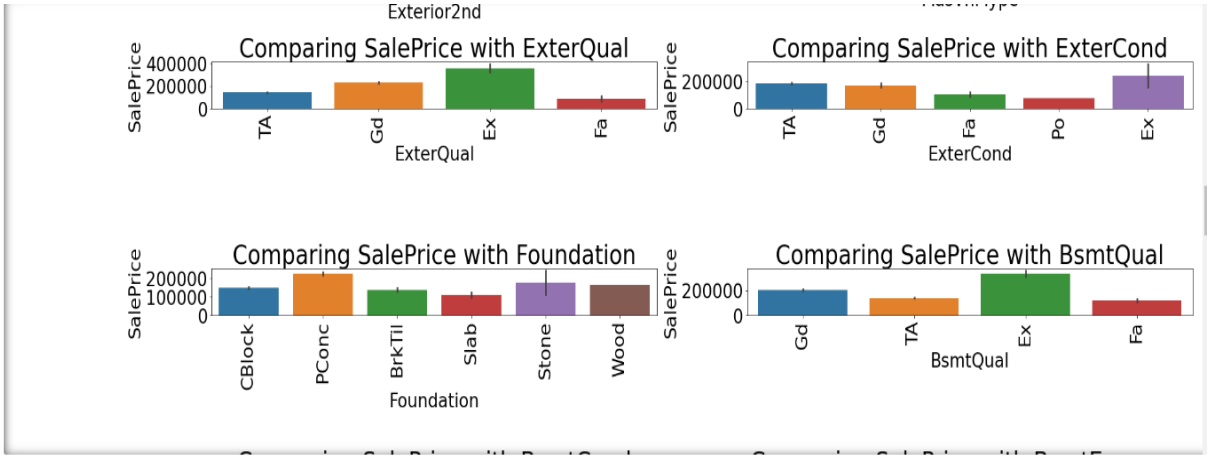
4.In 4th Graph, Rates the overall material and finish of the house(OverallQual) is increasing linearly sales is also increasing And SalePrice is also increasing linearly.

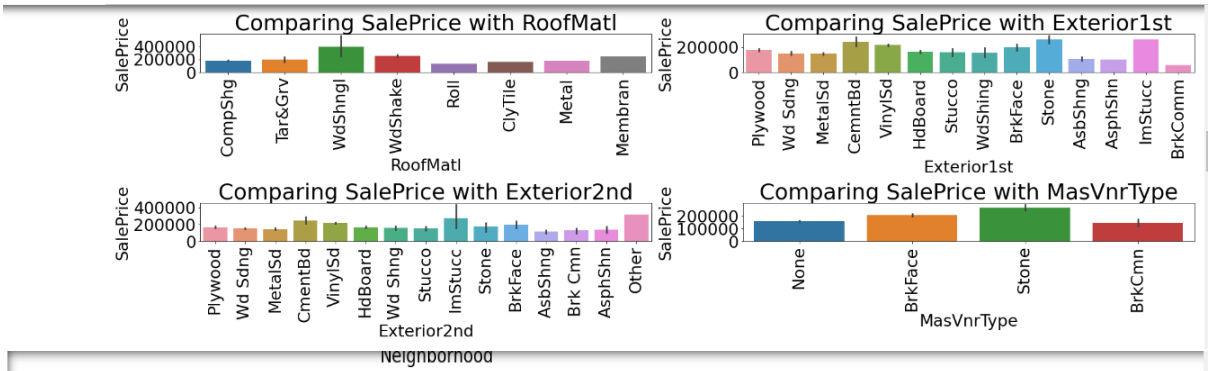
5.In 5th Graph, (Average) overall condition of the house(OverallCond) the sales is high and SalePrice is also high.

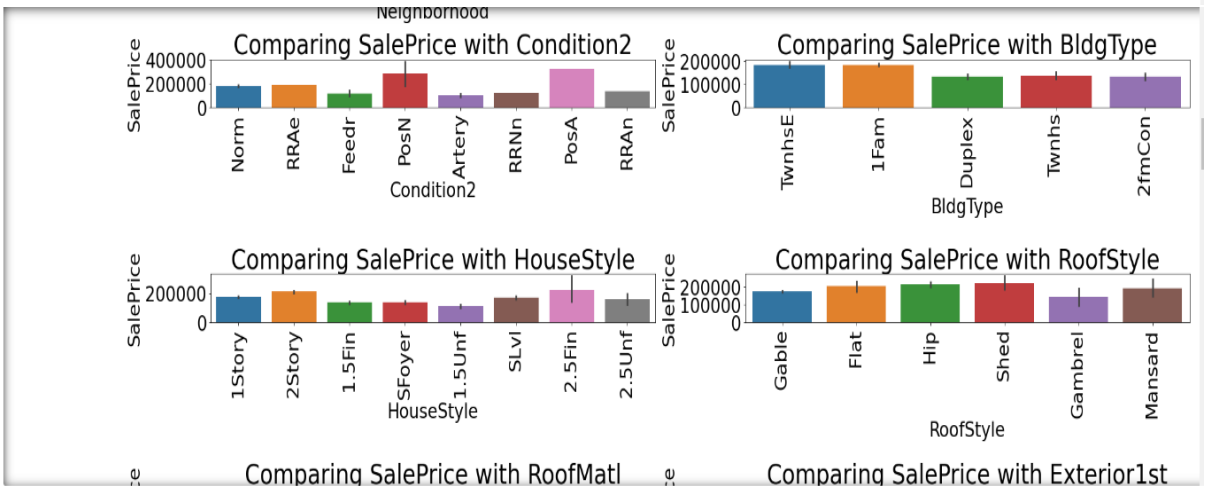
Visualisation of Categorical Features with Target Column:-



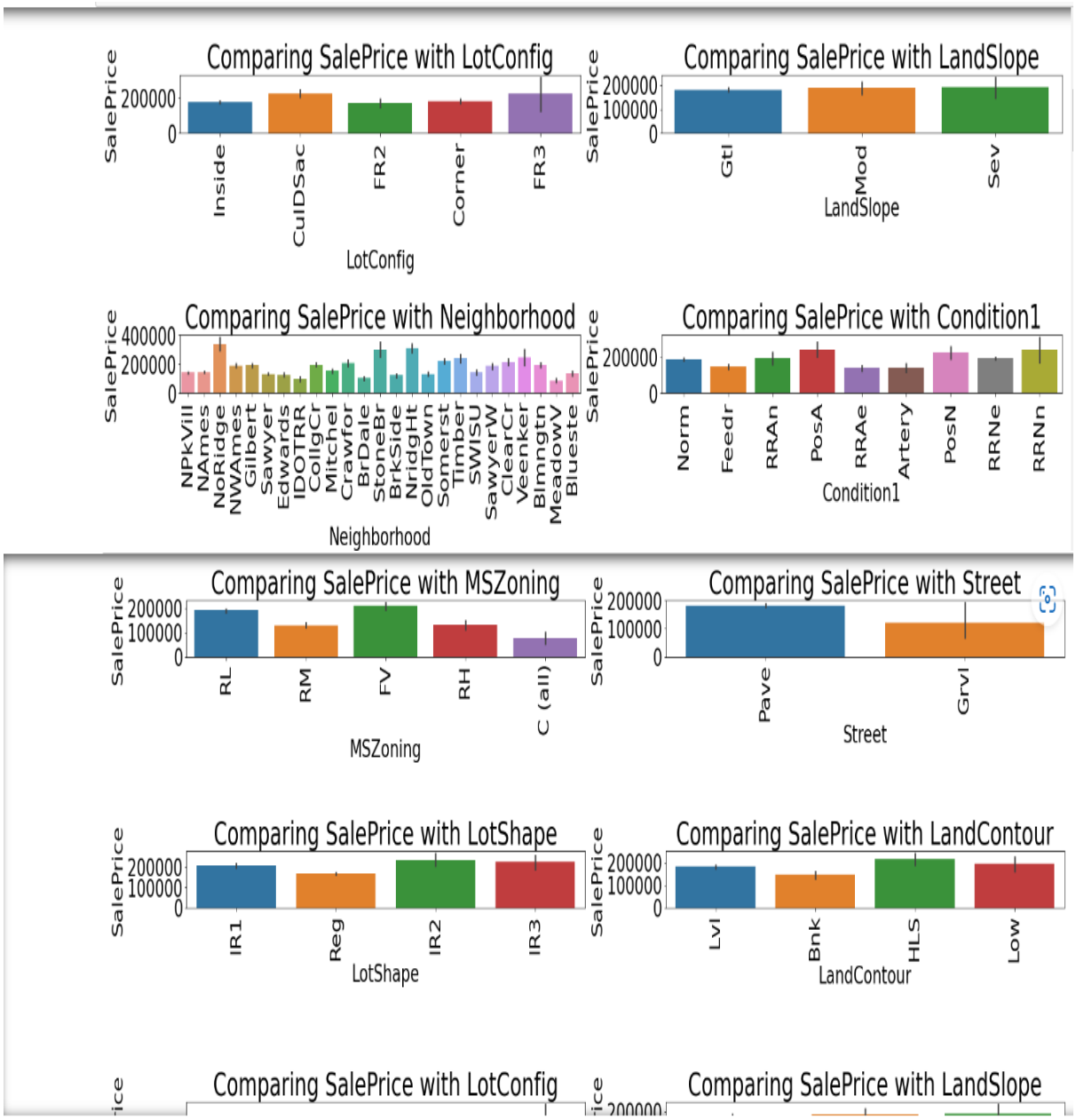



Exterior2nd
Comparing SalePrice with ExterQual
SalePrice
400000
200000
0
TA
Gd
Ex
Fa
ExterQual
Comparing SalePrice with ExterCond
SalePrice
200000
0
TA
Gd
Fa
Po
Ex
ExterCond
Comparing SalePrice with Foundation
SalePrice
200000
100000
0
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation
Comparing SalePrice with BsmtQual
SalePrice
200000
0
Gd
TA
Ex
Fa
BsmtQual


Comparing SalePrice with RoofMatl
SalePrice
400000
200000
0
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
RoofMatl
Comparing SalePrice with Exterior1st
SalePrice
200000
0
Plywood
Wd Sdng
MetalSd
CemntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
Exterior1st
Comparing SalePrice with Exterior2nd
SalePrice
400000
200000
0
Plywood
Wd Sdng
MetalSd
CmentBd
VinylSd
HdBoard
Wd Shng
Stucco
ImStucc
Stone
BrkFace
AsbShng
Brk Cmn
AsphShn
Other
Exterior2nd
Comparing SalePrice with MasVnrType
SalePrice
200000
100000
0
None
BrkFace
Stone
BrkCmn
MasVnrType


Neighborhood
Comparing SalePrice with Condition2
SalePrice
400000
200000
0
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn
Condition2
Comparing SalePrice with BldgType
SalePrice
200000
100000
0
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType
Comparing SalePrice with HouseStyle
SalePrice
200000
0
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
HouseStyle
Comparing SalePrice with RoofStyle
SalePrice
200000
100000
0
Gable
Flat
Hip
Shed
Gambrel
Mansard
RoofStyle

Comparing SalePrice with RoofMatl
Comparing SalePrice with Exterior1st



HERE ARE SOME MODELS WHICH I HAVE TESTED ON:-

```
Best Accuracy Score is :  0.9064585047615917 on Random_State :  78
```

In [171]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.30, random_state=78)
```

As Target Variable is Numerical, it is a Regression Problem:-

In [173]:
```
# For importing all necessary libraries for model building:-
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import ExtraTreesRegressor, GradientBoostingRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neighbors import KNeighborsRegressor as KNN
from sklearn.metrics import classification_report
from sklearn import metrics
from sklearn.model_selection import cross_val_score
```

In [175]:
```
# for checking R2 score with 1st Algorithm:-
Gbc=GradientBoostingRegressor()
Gbc.fit(x_train,y_train)
pred=Gbc.predict(x_test)
R2_Score=r2_score(y_test,pred)
print('r2_score: ',R2_Score)
print('mean_squared_error: ',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error: ',metrics.mean_absolute_error(y_test,pred))

# cross validation score
score= cross_val_score(Gbc, x, y, cv=10).mean()
print("Cross Validation Score: ", score)
```

```
r2_score:  0.9121445608743504
mean_squared_error:  352343741.0319501
mean_absolute_error:  14086.414357089254
Cross Validation Score:  0.8770830334229022
```

In [176]:
```
# For checking with 2nd algorithm:-
rfr=RandomForestRegressor()
rfr.fit(x_train,y_train)
pred=rfr.predict(x_test)
R2_Score=r2_score(y_test,pred)
print('r2_score: ',R2_Score)
print('mean_squared_error: ',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error: ',metrics.mean_absolute_error(y_test,pred))

# cross validation score
score= cross_val_score(rfr, x, y, cv=10).mean()
print("Cross Validation Score: ", score)
```

```
r2_score:  0.8988790957317421
mean_squared_error:  405544813.85557866
mean_absolute_error:  14472.405471698112
Cross Validation Score:  0.8583234780855118
```

In [177]:
```
# For checking with 3rd alogirthm:-
knn=KNN()
knn.fit(x_train,y_train)
pred=knn.predict(x_test)
R2_Score=r2_score(y_test,pred)
print('r2_score: ',R2_Score)
print('mean_squared_error: ',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error: ',metrics.mean_absolute_error(y_test,pred))

# cross validation score
score= cross_val_score(knn, x, y, cv=10).mean()
print("Cross Validation Score: ", score)
```

```
r2_score:  0.8513402380466143
mean_squared_error:  596199133.358868
mean_absolute_error:  18232.7213836478
Cross Validation Score:  0.7923976719328912
```

In [178]:
```
# For checking with 4th algorithm:-
DT=DecisionTreeRegressor()
DT.fit(x_train,y_train)
pred=DT.predict(x_test)
R2_Score=r2_score(y_test,pred)
print('r2_score: ',R2_Score)
print('mean_squared_error: ',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error: ',metrics.mean_absolute_error(y_test,pred))

# cross validation score
score= cross_val_score(DT, x, y, cv=10).mean()
print("Cross Validation Score: ", score)
```

```
r2_score:  0.5931162442968136
mean_squared_error:  1631804997.8050315
mean_absolute_error:  27459.459119496856
Cross Validation Score:  0.7062252267865875
```

In [179]:
```
# For checking with 5th algorithm:-
etr=ExtraTreesRegressor()
etr.fit(x_train,y_train)
pred=etr.predict(x_test)
R2_Score=r2_score(y_test,pred)
print('r2_score: ',R2_Score)
print('mean_squared_error: ',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error: ',metrics.mean_absolute_error(y_test,pred))

# cross validation score
score= cross_val_score(etr, x, y, cv=10).mean()
print("Cross Validation Score: ", score)
```

```
r2_score:  0.9047487512888225
mean_squared_error:  382004593.4875163
mean_absolute_error:  14288.037987421383
Cross Validation Score:  0.8571092195675926
```

Here, the least difference between R2 Score and CV Score is in GradientBoostingRegressor, so this is the Final and best model for Hyper parameter tuning.

NOW THE HYPER PARAMETER TUNING DONE USING GRID SEARCH CV:-

Hyper Parameter Tuning using GridSearch CV:-

In [181]:
```
# For importing required libraries for Hyperparameter Tuning:-
from sklearn.model_selection import GridSearchCV
```

In [189]:
```
Parameters={"n_estimators":[100], "criterion":["squared_error"], "min_samples_split":[2],"min_samples_leaf":[1]}

Gb=GradientBoostingRegressor()
clf=GridSearchCV(Gb,Parameters,cv=5)
clf.fit(x_train,y_train)

print(clf.best_params_)
```

```
{'criterion': 'squared_error', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
```

In [190]:
```
Gb=GradientBoostingRegressor(criterion='squared_error',min_samples_split=2,min_samples_leaf=1,n_estimators=100)
Gb.fit(x_train,y_train)
Gb.score(x_train,y_train)
prediction=Gb.predict(x_test)
gbs=r2_score(y_test,prediction)
print("r2_score",gbs*100)

gbs_score=cross_val_score(Gb,x,y,cv=5)
gbc=gbs_score.mean()
print("Cross_Val_Score",gbc*100)
```

```
r2_score 91.41648470981778
Cross_Val_Score 86.84308279012463
```

Here, R2 score has been increased to 91.4% after doing Hyper parameter tuning. Now saving the model

SAVING THE MODEL AND LOADING FOR GETTING PREDICTIONS:-

Saving Model:-

In [192]:
```
# Here i am saving the best model:-
import pickle
Name='HousingProject'
pickle.dump(Gb,open(Name,'wb'))
```

Loading Saved Model and Predicting:-

In [193]:
```
# For loading the saved model
Loaded=pickle.load(open('HousingProject','rb'))
Predictions=Loaded.predict(x_test)
Predictions
```

Out[193]:
```
array([173469.89757969,  97273.56512385,  99178.50736733, 112042.27505788,
       114756.97728064, 154053.29036201, 234062.74486887, 188240.05584471,
       274816.2701618 , 154372.47929655, 144303.61113665, 165827.51322481,
       186991.39065812, 186470.49095746, 136977.01213427, 218338.35367927,
       141785.44701456,  92343.08741332, 215935.03614045, 274670.74803251,
       117699.80143621, 182257.96871946, 129776.52980049, 112427.56351052,
       165314.16012103, 367636.34163662, 175452.56278905, 154004.08507701,
       146683.77815181, 201466.24507574, 118422.12388206, 118668.04904585,
       342382.25604641, 138753.67874292, 180700.56222796, 261812.17593677,
       235104.39319122, 113673.49435841, 159959.32309002, 177641.76479206,
       137204.57633621, 375968.45891854, 114686.69207752, 105388.06658789,
       268267.12657236, 106143.69923305, 206653.79863512, 111903.66701686,
       119796.27090333, 221520.88520253, 283503.72816252, 172070.45100758,
       248008.22246201, 145187.71423177, 184198.77365671, 144747.39585985,
       167073.59132753, 115554.33450489, 308421.62420393, 159396.99261323,
       244811.04851223, 117173.98413039, 176850.7901214 , 125810.00518121,
       150602.91864289, 157366.2508658 , 355111.33300914, 302502.27253773,
       156038.96343826, 134578.96567703, 129486.30250032, 155069.31340511,
```

Interpretation of the Results:-

Here, the datasets were having null values and zero entries in maximum columns so we have to be careful while going through the statistical analysis of the datasets and proper plotting for proper type of features will help us to get better insights on the data. I found maximum numerical continuous columns were in linear relationship with target column.

I notice a huge amount of outliers and skewness in the data so we have chosen proper analytical methods to deal with the outliers and skewness. If this outliers and skewness are ignored, we might have lose our model accuracy which would have predicted wrong/ less accuracy.

Then we performed scaling on both train and test dataset so that our model don't get biased.

We used multiple models while building model using train dataset to get the best model out of it.

Then we used multiple metrics like MAE, MSE and R2_score which will help us to decide the best model

I found Gradient BoostingRegressor as the best model with 91.2% R2_score. Also, we managed to improve the accuracy of the best model by running hyper parameter tunning to 91.4%.

At the end, I have predicted the Sale Price for test dataset using saved model of train dataset. I was able to get the predictions near to actual values.

Conclusion:-

In this project , we have used machine learning algorithms to predict the house prices. We have mentioned the step by step procedure to analyse the dataset and finding the correlation between the features. Here, we can select the features which are not correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a csv file was generated consisting of predicted house prices. Hence, we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then the Data Frame is also saved for predicting Sale prices of test dataset.

Learning Outcomes of the Study:-

We found that the dataset was interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphs also, it has made me to understand what data is trying to say. Data cleaning is one of the most important steps for removing missing values and to replace null values with its respective mean, median or mode.

This study is an exploratory attempt to use five machine learning algorithms in estimating housing prices, and then compare their results.

To conclude, the application of Machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.

Limitations are:

1. First demerit is the data leakage when we merge both train and test datasets.

2. Then when there are more number of outliers and skewness in dataset, these two will reduce our model accuracy.

3. Then we have tried our best to deal with outliers, skewness, null values and zero values. So, then it looks good that we have achieved a accuracy of 91.2% even after dealing all these drawbacks.

4. This study did not cover all regression algorithms however, it was focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones which was also good for learning.

Thanks