FLIP ROBO

House Price Prediction

Submitted by:

Gargi Saha Samanta

ACKNOWLEDGMENT

I would like to express my gratitude towards FlipRobo Technologies for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons and my mentor(Ms. Shristi Maan) for giving me such attention and time as and whenever required.

INTRODUCTION

- ## Business Problem Statement
  - ➢ 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 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
  - ➢ 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.

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

- ## Motivation for the Problem Undertaken
  - ➢ Having lived in India for so many years if there is one thing that I had been taking for granted, it's that housing and rental prices continue to rise. Housing prices have recovered remarkably well, especially in major housing markets.

- So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.
- So we have to predict the pricing as per customers requirement and needs.

Analytical Problem Framing

- ## Dataset Representation:

```
# Load the dataset
df_train=pd.read_csv('house_train.csv')
df_test=pd.read_csv('house_test.csv')
```

```
# View the train data
df_train.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl

Observation:

1. Seeing the data we have to build a model which can be used to predict the SalePrice.
2. The data seems to be a combination of both numerical and categorical features.

So clearly it is a regression problem.

- ## Data Sources and their formats & inferences

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
RMResidential 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
SevSevere 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)
HipHip	

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)
MnMimimum 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
SevSeverely 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 Preprocessing**

Change the categorical values into numerical

```
categoricals
len(categoricals)
```

```
43
```

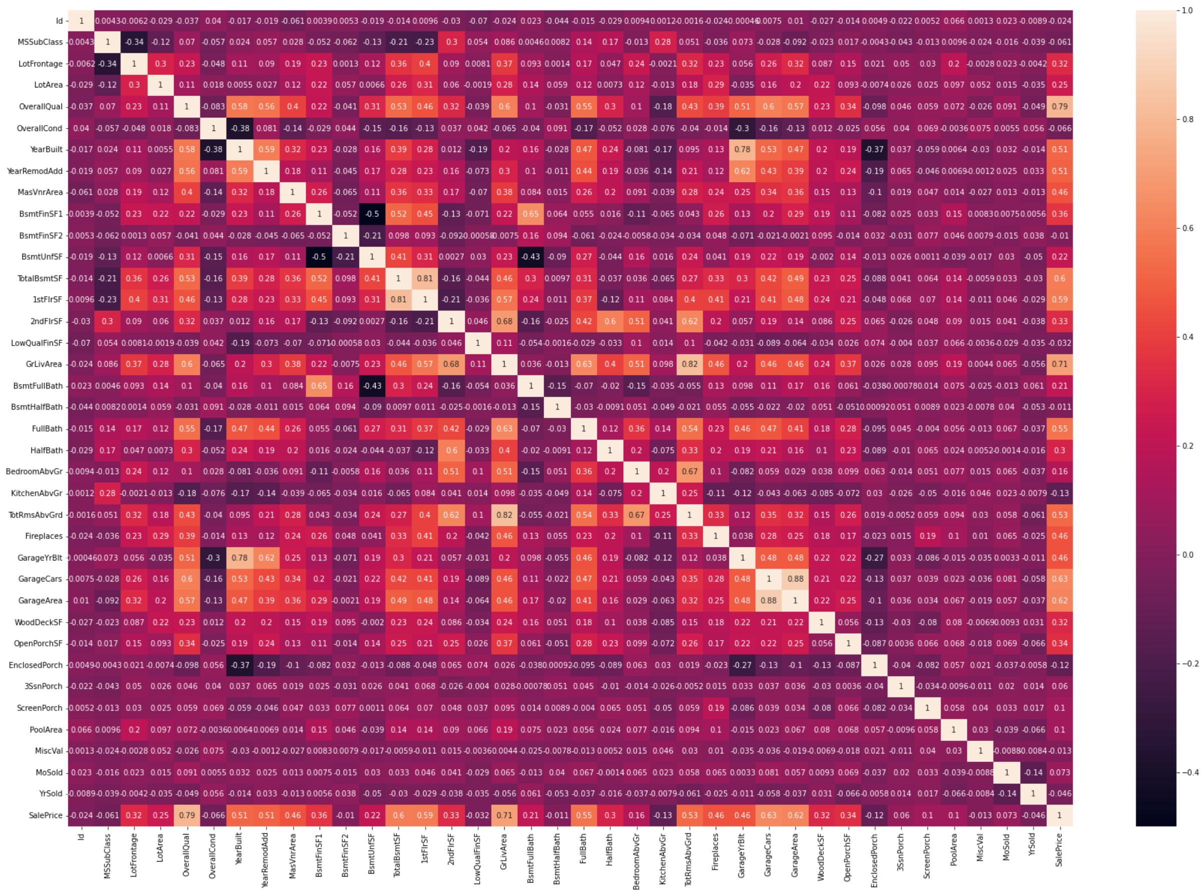
```
from sklearn.preprocessing import LabelEncoder
encoder=LabelEncoder()
```

```
#  Converting all categorical variables using label encoding rather than using One hot encoder to

for feature in categoricals:
    df_train[feature]=encoder.fit_transform(df_train[feature].astype(str))

for feature in categoricals_test:
    df_test[feature]=encoder.fit_transform(df_test[feature].astype(str))
```

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

➢ **Correlation :**

```
plt.figure(figsize=(30,20))
sns.heatmap(df_train.corr(),annot=True,robust=False)
```



Observation:
1. We can see that there is a positive correlation between the almost all the features and the SalePrice.
2. Few of the features are showing strong relation but few are not.

- Assumption for the problem:

➢ So clearly it is a Regression problem.

- Hardware and Software Requirements and Tools Used

Software Used:

- Jupyter Notebook
- Ms-Paint

- MS-PowerPoint
- MS-Word

Hardware used:

- Laptop
- Good internet connectivity

Model/s Development and Evaluation

- **Testing of Identified Approaches (Algorithms)**
- **Techniques:**

- LinearRegression()
- DecisionTreeRegressor()
- KNeighborsRegressor()
- RandomForestRegressor()
- GradientBoostingRegressor()

- Running the selected Models:

Techniques:

```
# Split the train dataset:

X_train,X_test,Y_train,Y_test=train_test_split(x_scaled,Y,test_size=0.2,random_state=42)
```

```
# creating objects for given
Linear=LinearRegression()
DecisionTree=DecisionTreeRegressor()
knn=KNeighborsRegressor()
Random=RandomForestRegressor()
gbr=GradientBoostingRegressor()

```

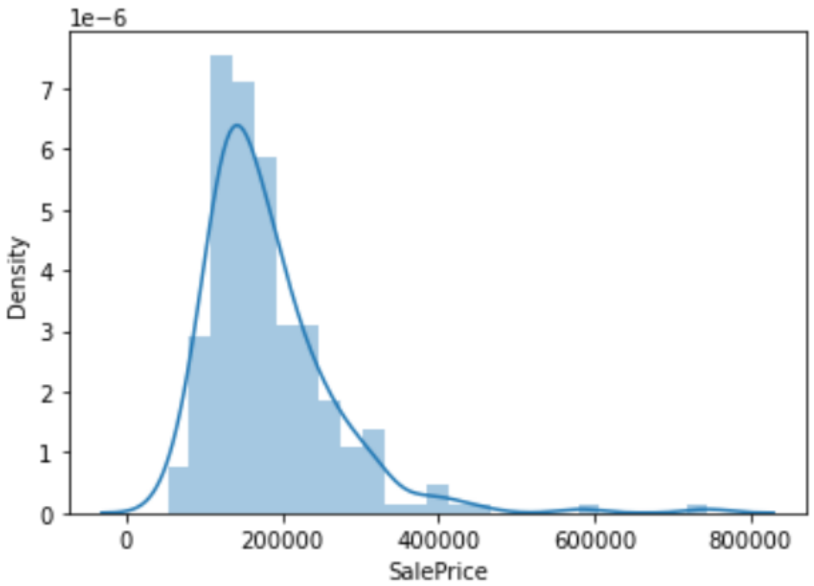
```
# For fitting the model

alg=[Linear,DecisionTree,knn,Random,gbr]
acc_models={}
for model in alg:
    model.fit(X_train,Y_train)
    Y_pred=model.predict(X_test)
    acc_models[model]=round(r2_score(Y_test,Y_pred)*100,1)
    print("Model Name:",model)
    print('Accuracy ::',{round(r2_score(Y_test,Y_pred)*100,1)})
    print('Mean Absolute Error(MAE) is::',{mean_absolute_error(Y_test,Y_pred)})
    print('Mean Squared Error(MSE) ::',{mean_squared_error(Y_test,Y_pred)})
    print('Root Mean Squared Error is ::',{np.sqrt(mean_squared_error(Y_test,Y_pred))})
    print("-------------------------------------------------------")
```

```
acc_models
```

```
{LinearRegression(): 71.0,
 DecisionTreeRegressor(): 56.7,
 KNeighborsRegressor(): 62.1,
 RandomForestRegressor(): 78.2,
 GradientBoostingRegressor(): 81.0}
```

```
sns.distplot((Y_test))
```

```
<AxesSubplot:xlabel='SalePrice', ylabel='Density'>
```

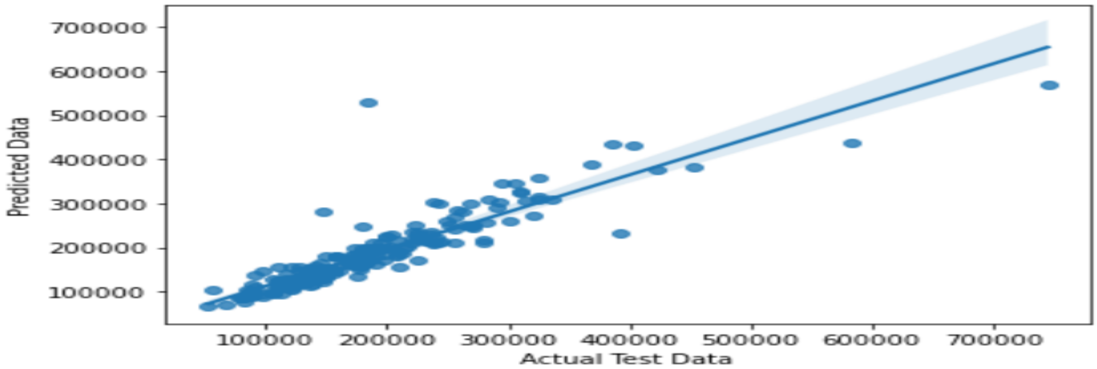


Observations:

Gradient Boosting showing best result

Graph After fitting the model:

```
GBR=GradientBoostingRegressor()
GBR.fit(X_train,Y_train)
Y_pred=GBR.predict(X_test)
sns.regplot(Y_test,Y_pred)
plt.xlabel("Actual Test Data")
plt.ylabel("Predicted Data")
plt.tight_layout()
```



Key Metrics for success in solving problem under consideration

Hyper Tuning the Gradient Boosting Algorithm:

```
X_train, X_test, Y_train, Y_test = train_test_split(x_scaled, Y, test_size=0.20, random_state=72)

from sklearn.model_selection import GridSearchCV

param_grid = {"min_samples_leaf" : [1,2,3],
              "min_samples_split" : [2,3,4],
              "n_estimators" : [100,200],
              "learning_rate" : [0.1,0.2]}
grid_search = GridSearchCV(gbr, param_grid=param_grid)
grid_search.fit(X_train, Y_train)
grid_search.best_params_
```

```
{'learning_rate': 0.2,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 200}
```

```
best_model = GradientBoostingRegressor(learning_rate=0.2,min_samples_split=3,min_samples_leaf=1,n_
X_train, X_test, Y_train, Y_test = train_test_split(x_scaled,Y, test_size=0.20, random_state=72)
best_model.fit(X_train, Y_train)
Y_pred = best_model.predict(X_test)
r2_score(Y_test, Y_pred)
```

```
0.916385624451561
```

```python
df_test['SalePrice'] = best_model.predict(df_test)
```

```python
df_test[['Id', 'SalePrice']].to_csv('House_Price_submission.csv', index=False)
```

```python
submission = pd.read_csv(r'House_Price_submission.csv')
submission.shape
```

```
(292, 2)
```

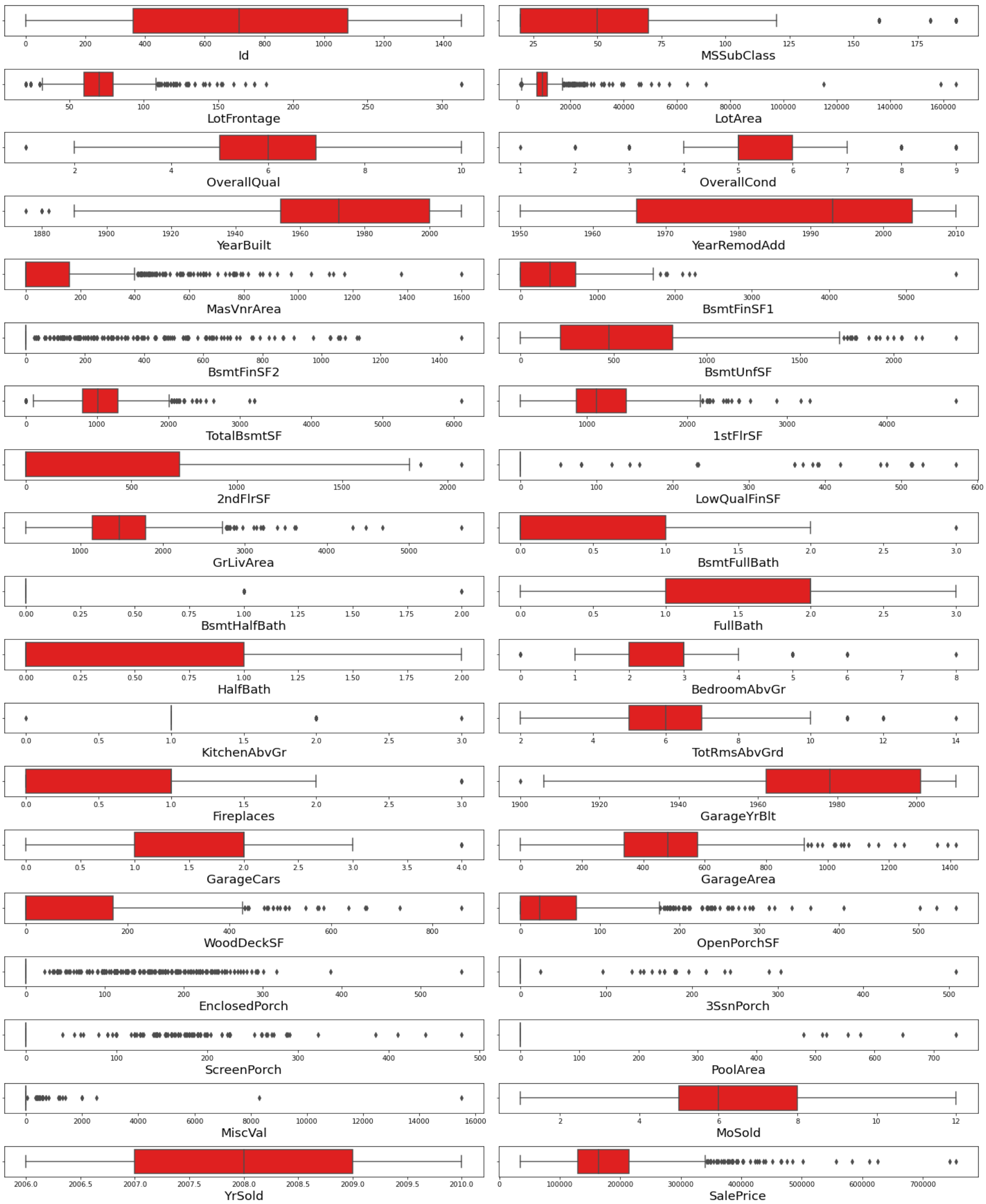
```python
submission
```

1	1018	250764.159506
2	929	237982.485986
3	1148	188367.514023
4	1227	208346.650108
5	650	79911.192588
6	1453	130145.871060
7	152	284122.890173
8	427	239908.663531
9	776	155651.652826

- ## Visualizations:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

Univariate Analysis of Numerical variables:

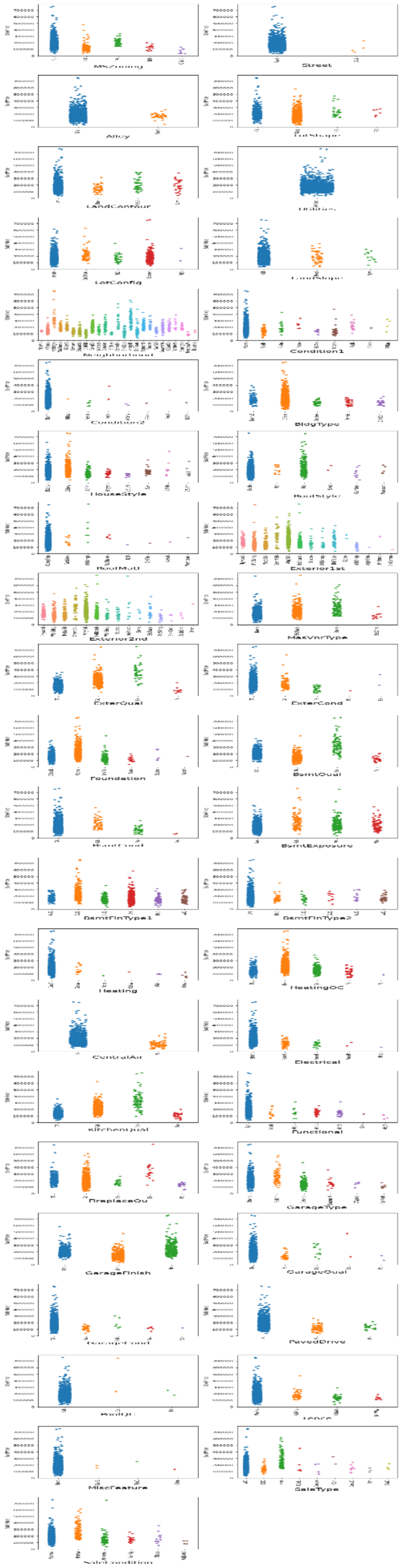


OBSERVATION:

1. Some features such as Id, YearRemodAdd,BsmFullBath,FullBath,HalfBathFirePlace,MoSold,YrSold are not having outliers.

2. Rest of the features are more or less having outliers .

Bivariate plottings:



• Interpretation of the Results:

1. FV is highest in price followed by RL and RH.

2. Streets having Pave and Alley having Grvl is having high Price.

3. LotShape of IR2 is high in Price.

4. LandContour with HLS ,LotConfig with FR3,LandSlope woth Sev are having higher prices than the other subcategories.

5. Condition 1 withRRNn nad PosA have high price.

6. Condition2 with PonA follewd by PosN are having prices.

7. BldgType of Twnhse,HouseStyle of 2.5Unf,RoofStyle of Shed, RoofMatl of Wdshngl, Exterior1st of

stone and Imstucc are high prices whereas Exterior2nd with other and Imstucc have high price.

8. MasVnrType with stone, ExterQual with Ex,ExterCond withEx,Foundation with Pconc, BsmtQual with ex

BmstCond with Gd,BmstExposer with Gd,BsmtFinType1 with GQL,BsmtFinType2 with GQL andAQL are

high in Price.

9. Heating with GasA ,HeatingQc with Ex,CentralAir with Yes Electrical with SBrkr, KitchenQual with

Ex ,Funtional with Typ,FireplaceQu with Ex, GarageType with BuiltIn, GarageFinish with Fin has

high Price.

10. GarageQual with Ex,GarageCond with Gd,PavesDrive with Y,PoolQc with Ex ,Fence with MnPrv nad

GdPrv are high in Price.

11. SaleType of con and new ,SaleCondition with Partial are having highest SalePrice.

CONCLUSION

So, our study showed that…….

Gradient Boosting Regression displayed the best performance for this Dataset and can be used for deploying purposes.

- ## Learning Outcomes of the Study in respect of Data Science

- Our customers requirements are our highest priority so the project was built to satisfy their needs so the project works well and there is no customer churn

- We should maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.
- So we have to predict the pricing as per customers requirement and needs.

- ## Limitations of this work and Scope for Future Work
- 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.
- But still customers are always comparing the prices hence we should keep on updating our project to meet their necessity.

Thank You