FLIP ROBO

HOUSING PROJECT

Submitted by:

SHRUTI GAMNE

ACKNOWLEDGMENT

I take great pleasure to thank and acknowledgment the permission and allowance by Flip Robo Technologies and their inspiration provided. I extend whole hearted thanks to them under whom I worked and learned a lot and for enlightening me with their knowledge and experience to grow with the corporate working. Their guidance at every stage of the project enabled me to successfully complete this project which otherwise would not have been possible without their consent encouragement and motivation, without the support it was not possible for me to complete the report with fullest endeavour. I would like to extend my thanks to all my SME who supported me in carryout my operation successfully and generously and provided me vital information regarding my project objective.

INTRODUCTION

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. There are three factors that influence the price of a house which include physical conditions, concept and location.

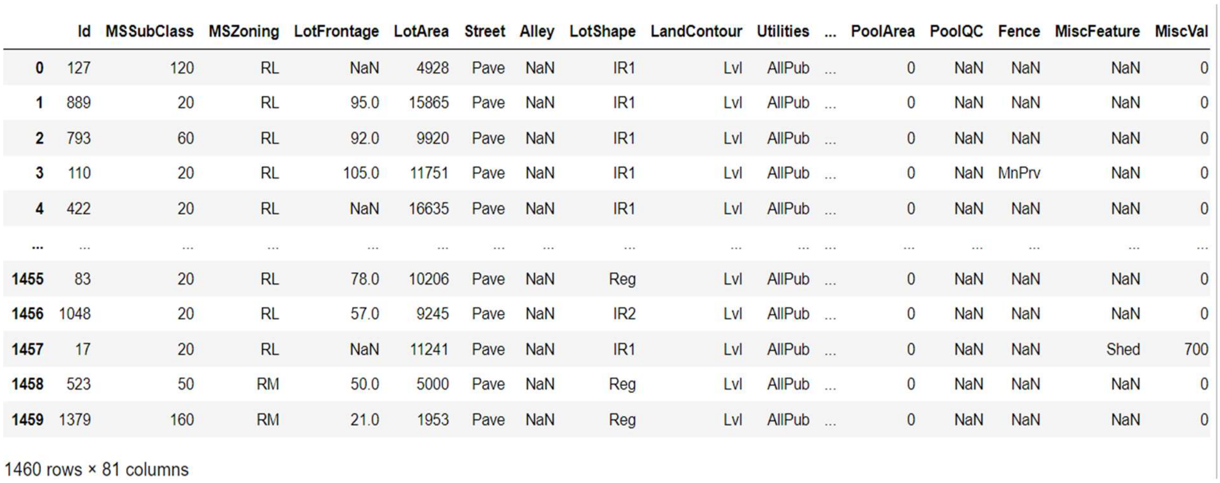
House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyse historical property transactions to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.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.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 8:2 for Training and Testing purpose respectively..

- Data format



	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1456	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1457	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	Shed	700
1458	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1459	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0

1460 rows × 81 columns

- Data Pre-processing Done

  Firstly I dropped the columns containing the NaN values in large numbers, then filled NaN values with help of fillna() with median wherever necessary ,then used Label Encoder to encode all the string values/categorical data.

- Data Inputs- Logic- Output Relationships

The description for the 81 features is given below:

MSSubClass: Identifies the type of dwelling involved in the sale. More the number of storey more will be the Sale Price

MSZoning: Identifies the general zoning classification of the sale. more will be the Sale Price

LotFrontage: More Linear feet of street connected to property. more will be the Sale Price

LotArea: Greater the Lot size in square feet Greater will be sale price.

Street: if Paved road has access to property Greater will be sale price.

LotShape: larger shape of property, Greater will be sale price.

LandContour: More Flatness of the property Greater will be sale price.

Utilities: All public Utilities available, Greater will be sale price.

LotConfig: Lot configuration

LandSlope: More the Slope of property, lesser will be the sale price

Neighborhood: Highly expensive area Greater will be sale price.

Condition1: Proximity to various conditions, more the amenities higher will be sale price

Condition2: Proximity to various conditions

BldgType: larger the type of dwelling Greater will be sale price.

HouseStyle: More number of storey more will be sale price.

OverallQual: Excellent overall material and finish of the house, Greater will be sale price.

OverallCond: Excellent overall condition of the house Greater will be sale price

YearBuilt: Original construction date, older the house lesser will be sale price

YearRemodAdd: Newer the Remodel date Greater will be sale price

RoofStyle: Costlie the Type of roof Greater will be sale price

RoofMatl: Costlier Roof material used then Greater will be sale price

Exterior1st: more the Exterior covering on house Greater will be sale price.

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

MasVnrType: more Masonry used Greater will be sale price

MasVnrArea: Greater Masonry veneer area in square feet Greater will be sale price

ExterQual: Poor the quality of the material on the exterior lesser will be the sale price

ExterCond: average present condition of the material on the exterior average will be the sale price

Foundation: Costlier Type of foundation Greater will be sale price

BsmtQual: More the height of the basement Greater will be sale price

BsmtCond: Poor - Severe cracking, settling, or wetness of the basement lesser will be the sale price

BsmtExposure: Good Exposure to walkout Higher will be price

BsmtFinType1: Unfinshed basement area lesser will be the sale price

BsmtUnfSF: Unfinished square feet of basement area lesser will be the sale price

TotalBsmtSF: more the square feet of basement area greater will be the sale price

Heating: costlier the heating greater will be the sale price

HeatingQC: Poor Heating quality and condition lesser will be the sale price

CentralAir: if has Central air conditioning greater will be the sale price

Electrical: excellent Electrical system provided greater will be the sale price

1stFlrSF: more the First Floor square feet greater will be the sale price

2ndFlrSF: more the Second floor square feet greater will be the sale price

LowQualFinSF: Low quality finished square feet (all floors) lesser will be sale price

GrLivArea: more Above grade (ground) living area square feet greater will be the sale price

BsmtFullBath: more Basement full bathrooms greater will be the sale price

BsmtHalfBath: more Basement half bathrooms greater will be the sale price

FullBath: more Full bathrooms above grade greater will be the sale price

HalfBath: Half baths above grade greater will be the sale price

Bedroom: Bedrooms above grade greater will be the sale price

Kitchen: Kitchens above grade greater will be the sale price

KitchenQual: Excellent Kitchen quality greater will be the sale price

TotRmsAbvGrd: Total rooms above grade greater will be the sale price

Functional: Severely Damaged Home lesser will be the sale price

Fireplaces: more Number of fireplaces greater will be the sale price

FireplaceQu: Poor Fireplace quality lesser will be the sale price

GarageType: More than one type of garage greater will be the sale price

GarageYrBlt: older the garage was built lesser will be the sale price

GarageFinish: Unfinished Interior of the garage lesser will be the sale price

GarageCars: more Size of garage in car capacity more sale price

GarageArea: more Size of garage in square feet more sale price

GarageQual: Poor Garage quality lesser will be the sale price

GarageCond: poor Garage condition lesser will be the sale price

PavedDrive: if dirt/gravel present in driveway lesser will be the sale price

WoodDeckSF: more Wood deck area in square feet greater will be the sale price

OpenPorchSF: larger Open porch area in square feet greater will be the sale price

EnclosedPorch: more Enclosed porch area in square feet greater will be the sale price

3SsnPorch: if has Three season porch area in square feet greater will be the sale price

ScreenPorch: larger Screen porch area in square feet greater will be the sale price

PoolArea: larger Pool area in square feet greater will be the sale price

PoolQC: excellent Pool quality greater will be the sale price

Fence: Good privacy of Fence quality greater will be the sale price

MiscFeature: if has Miscellaneous feature like elevator ,tennis court etc greater will be the sale price

MiscVal: more Value of miscellaneous feature greater will be the sale price

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: if Home just constructed and sold greater will be the sale price

SaleCondition: if sale condition is normal greater will be the sale price

- Libraries Used for this Project include –
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit Learn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

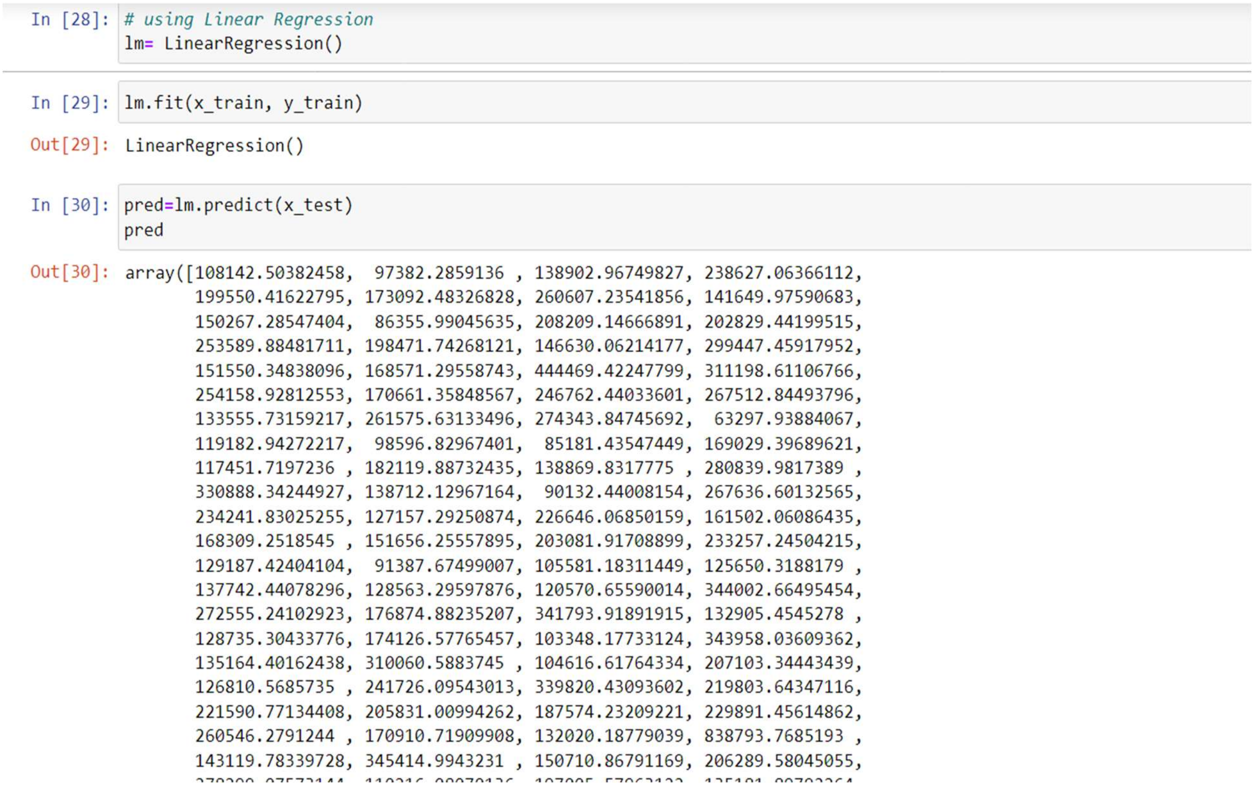
  Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies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

- Testing of Identified Approaches (Algorithms)

  Regression Model

1. Linear Regression is a machine learning algorithm based on supervised learning.
2. It performs a regression task. Regression models a target prediction value based on independent variables.
3. It is mostly used for finding out the relationship between variables and forecasting.

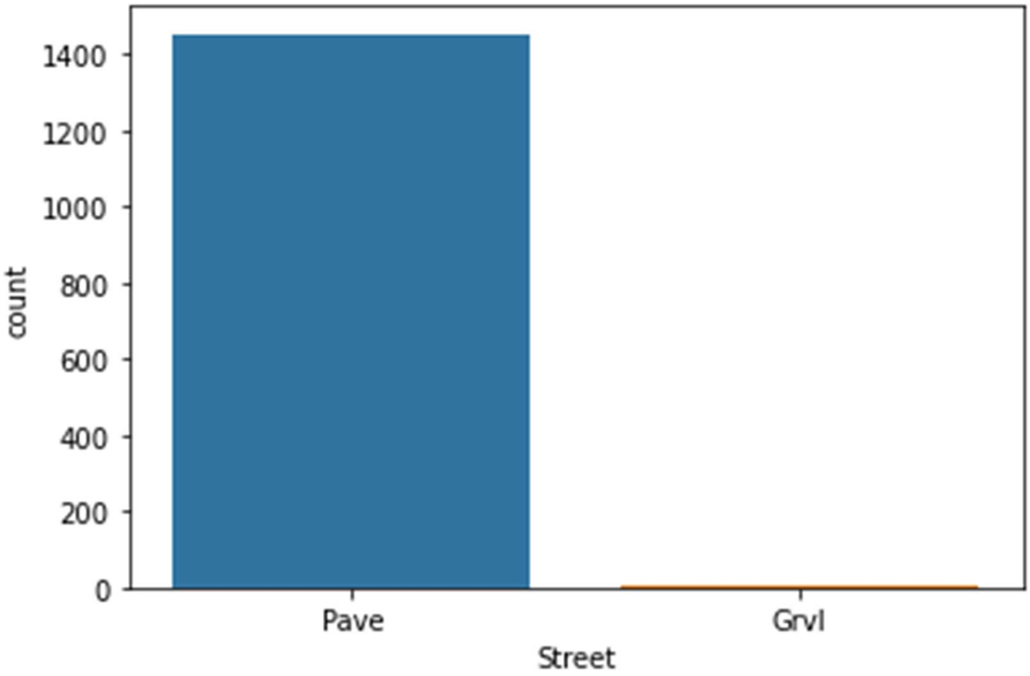
- Evaluating selected model:

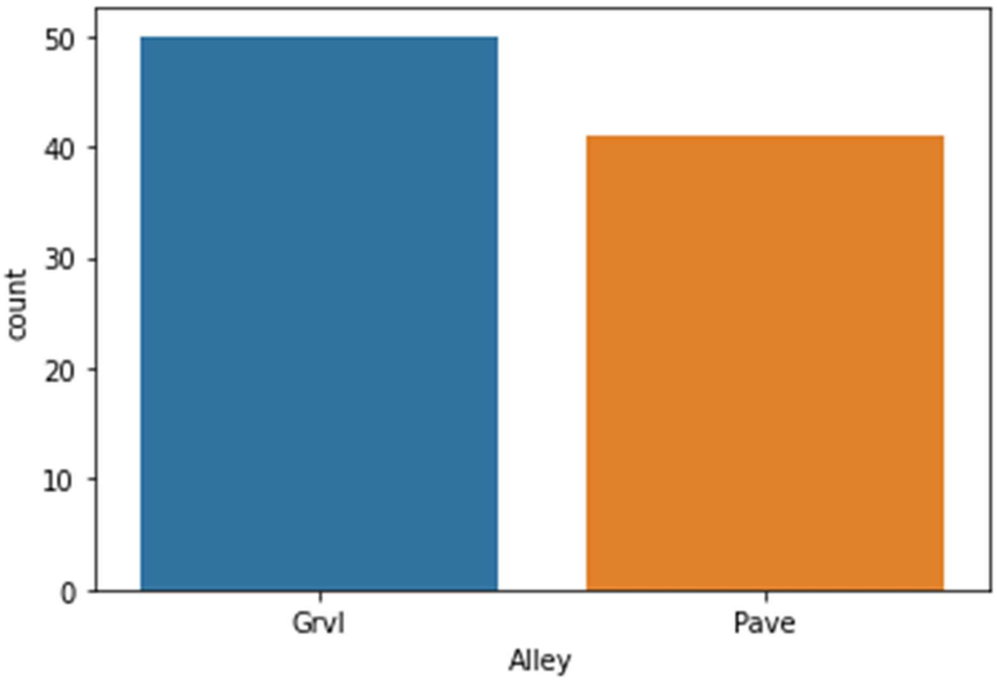


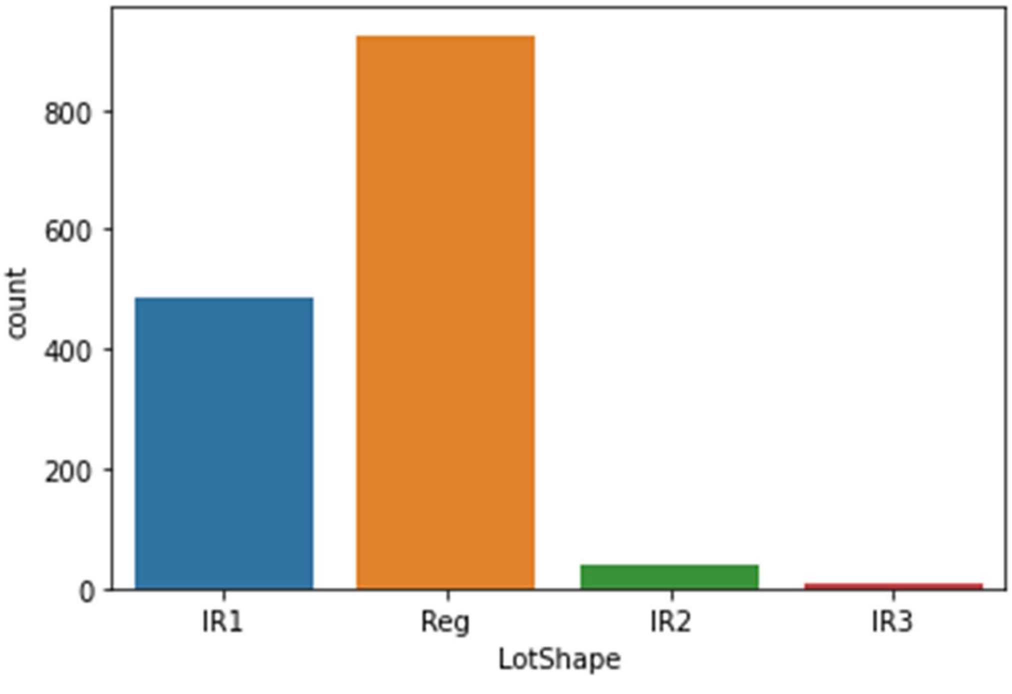
```
In [28]: # using Linear Regression
         lm= LinearRegression()

In [29]: lm.fit(x_train, y_train)

Out[29]: LinearRegression()

In [30]: pred=lm.predict(x_test)
         pred

Out[30]: array([108142.50382458,  97382.2859136 , 138902.96749827, 238627.06366112,
                199550.41622795, 173092.48326828, 260607.23541856, 141649.97590683,
                150267.28547404,  86355.99045635, 208209.14666891, 202829.44199515,
                253589.88481711, 198471.74268121, 146630.06214177, 299447.45917952,
                151550.34838096, 168571.29558743, 444469.42247799, 311198.61106766,
                254158.92812553, 170661.35848567, 246762.44033601, 267512.84493796,
                133555.73159217, 261575.63133496, 274343.84745692,  63297.93884067,
                119182.94272217,  98596.82967401,  85181.43547449, 169029.39689621,
                117451.7197236 , 182119.88732435, 138869.8317775 , 280839.9817389 ,
                330888.34244927, 138712.12967164,  90132.44008154, 267636.60132565,
                234241.83025255, 127157.29250874, 226646.06850159, 161502.06086435,
                168309.2518545 , 151656.25557895, 203081.91708899, 233257.24504215,
                129187.42404104,  91387.67499007, 105581.18311449, 125650.3188179 ,
                137742.44078296, 128563.29597876, 120570.65590014, 344002.66495454,
                272555.24102923, 176874.88235207, 341793.91891915, 132905.4545278 ,
                128735.30433776, 174126.57765457, 103348.17733124, 343958.03609362,
                135164.40162438, 310060.5883745 , 104616.61764334, 207103.34443439,
                126810.5685735 , 241726.09543013, 339820.43093602, 219803.64347116,
                221590.77134408, 205831.00994262, 187574.23209221, 229891.45614862,
                260546.2791244 , 170910.71909908, 132020.18779039, 838793.7685193 ,
                143119.78339728, 345414.9943231 , 150710.86791169, 206289.58045055,
```

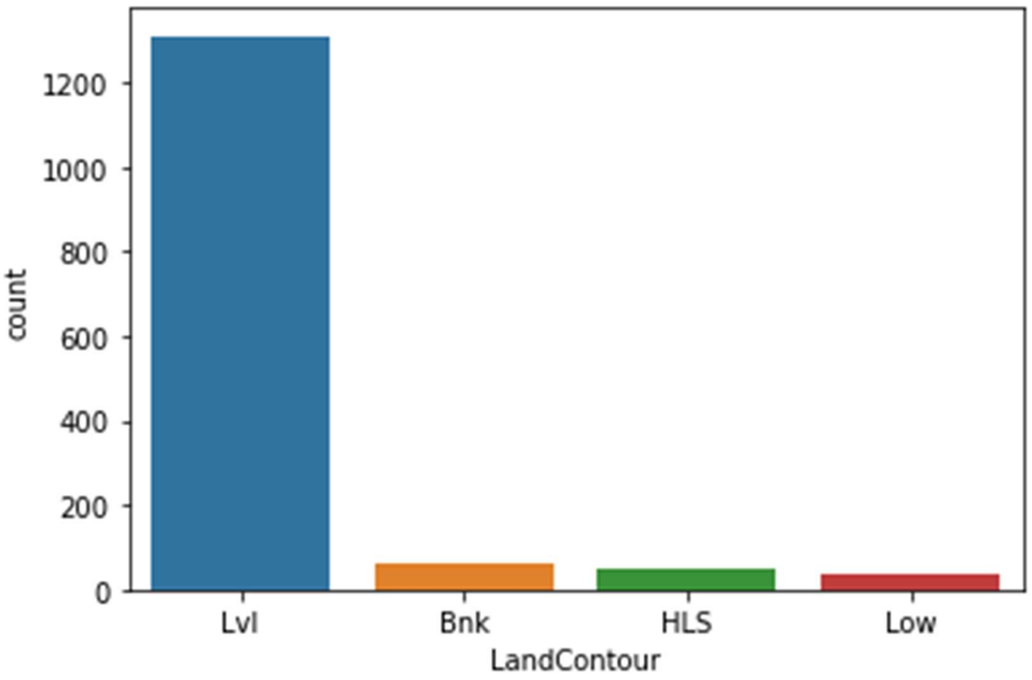
- Key Metrics used: mean_squared_error,mean_absolute_error,r2_score

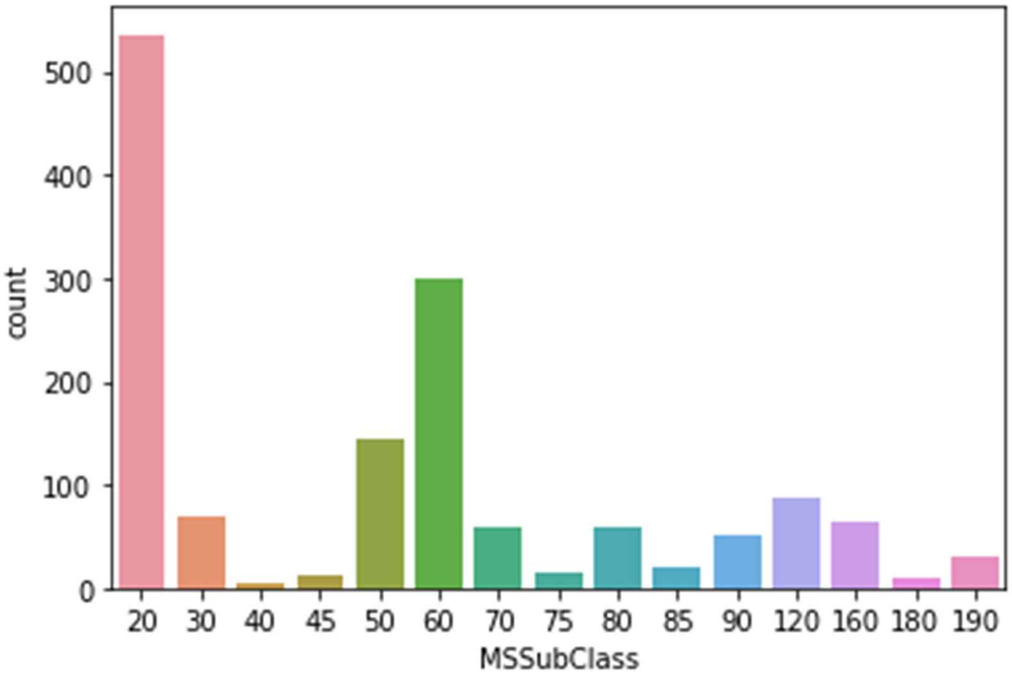
- Visualizations

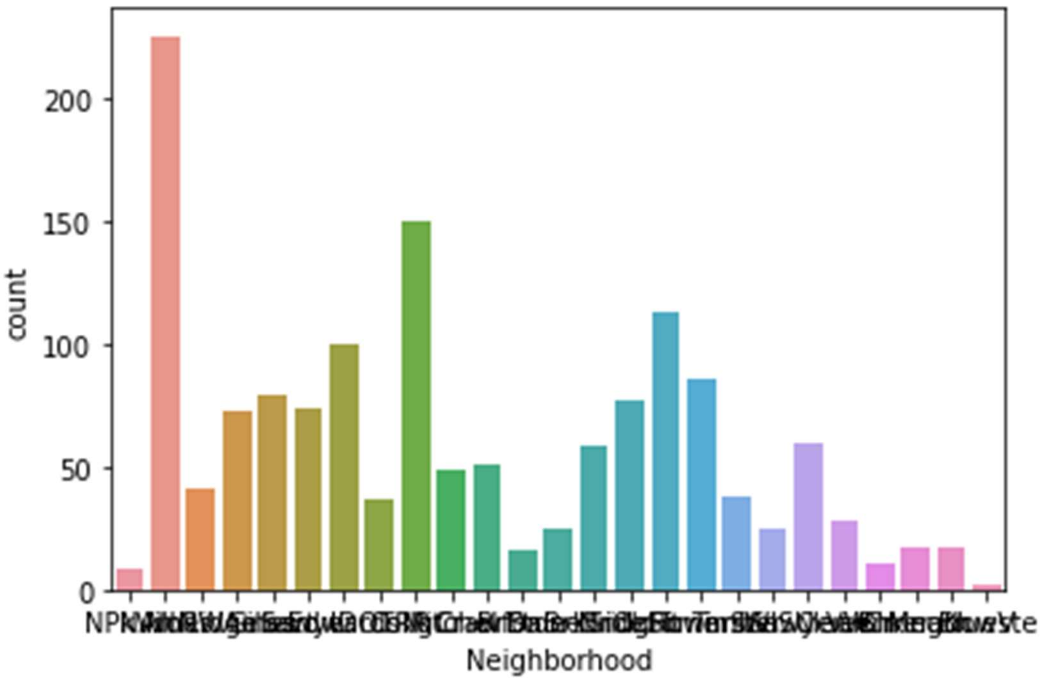


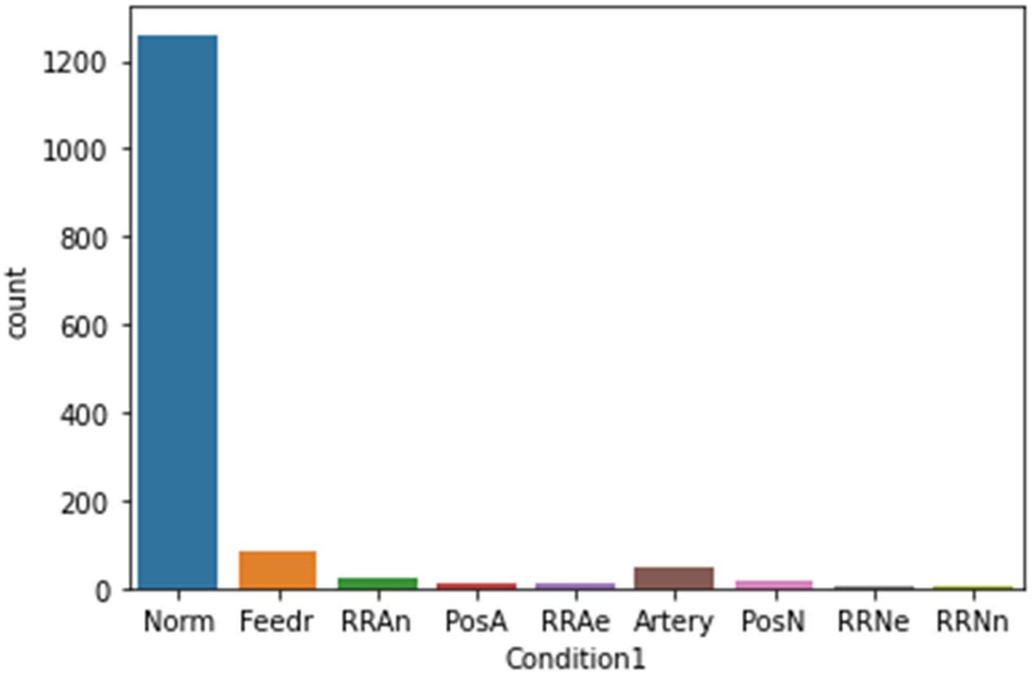

50
40
30
20
10
0
count
Grvl
Pave
Alley

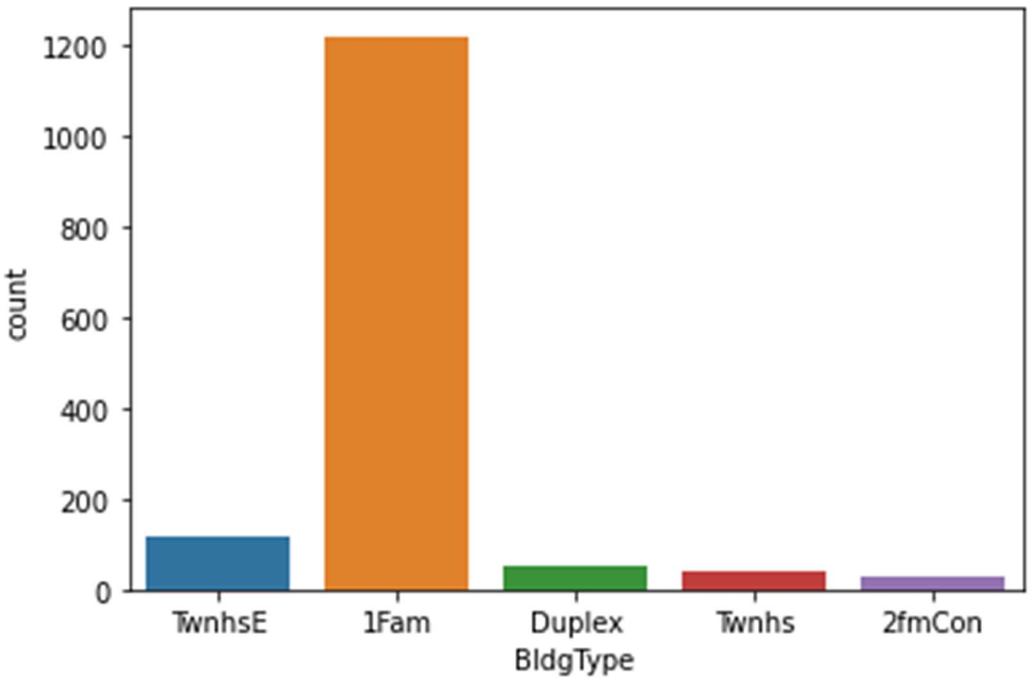

800
600
400
200
0
count
IR1
Reg
IR2
IR3
LotShape

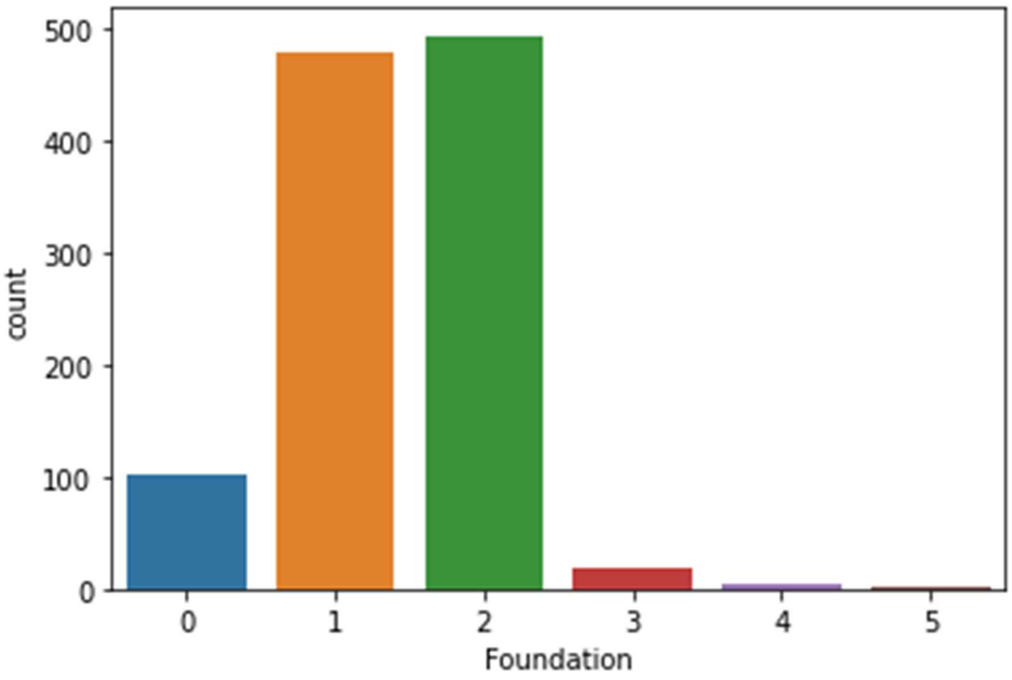

1200
1000
800
600
400
200
0
count
Lvl
Bnk
HLS
Low
LandContour

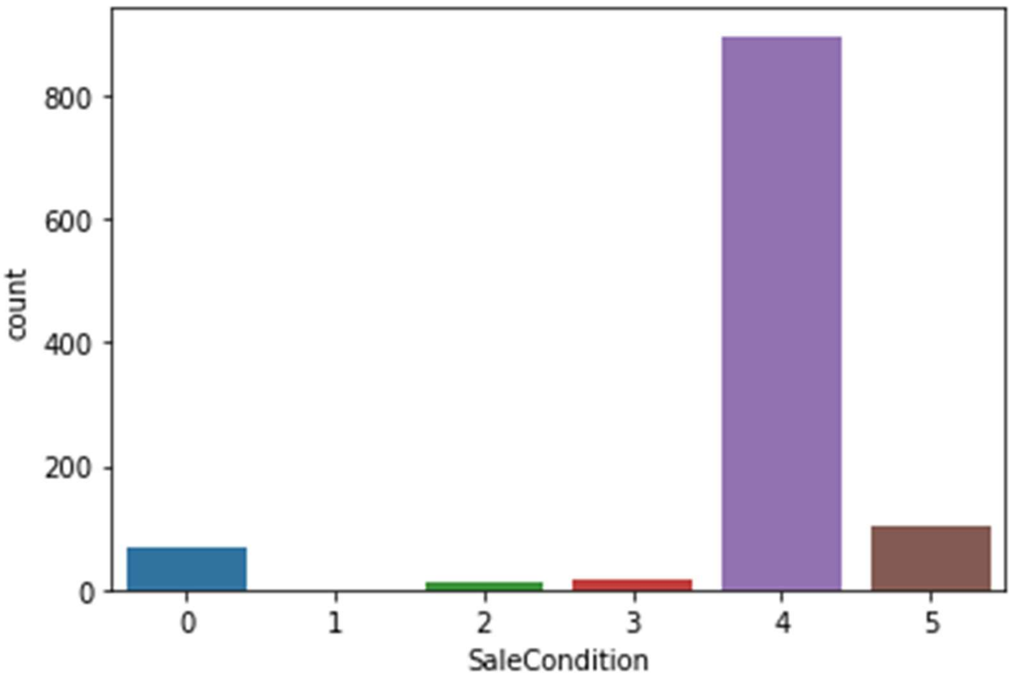

500
400
300
200
100
0
count
20
30
40
45
50
60
70
75
80
85
90
120
160
180
190
MSSubClass

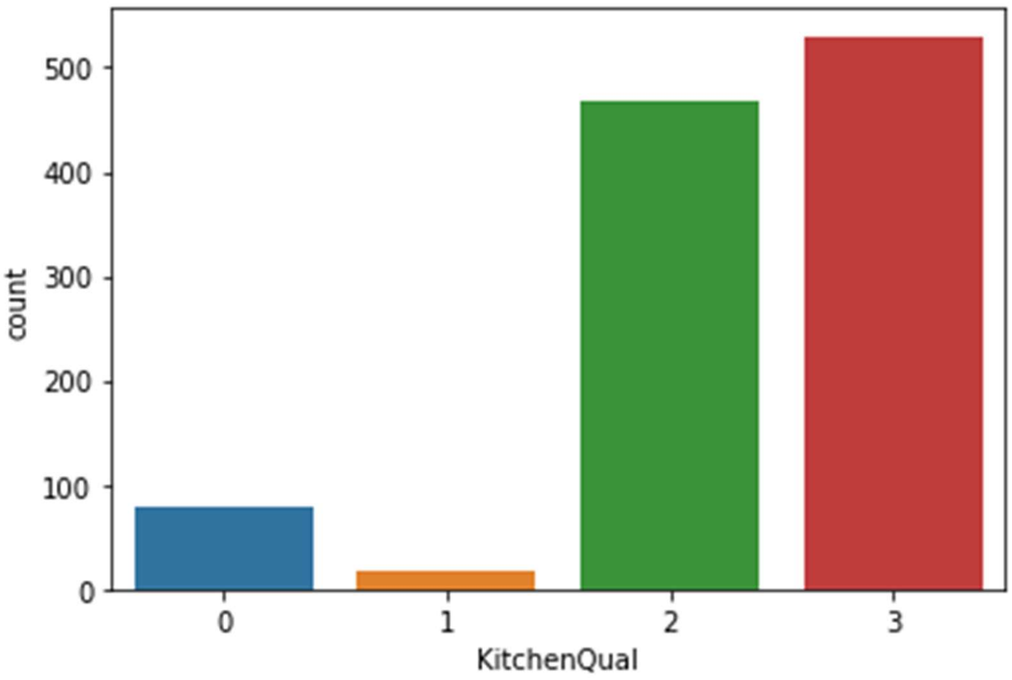

count
0
50
100
150
200
Neighborhood

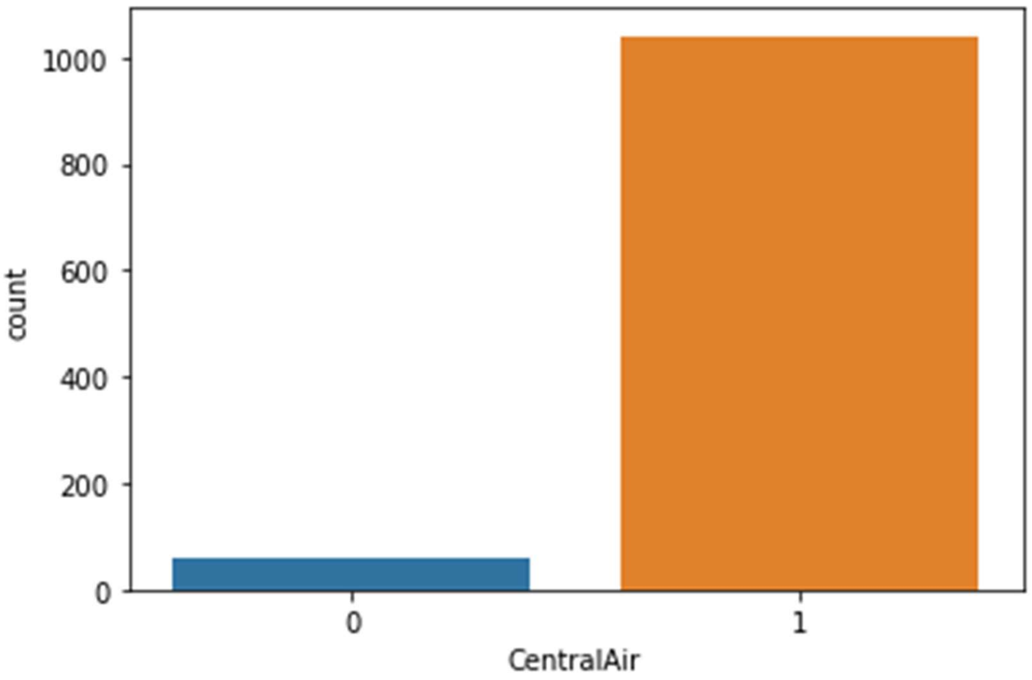

count
0
200
400
600
800
1000
1200
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition1

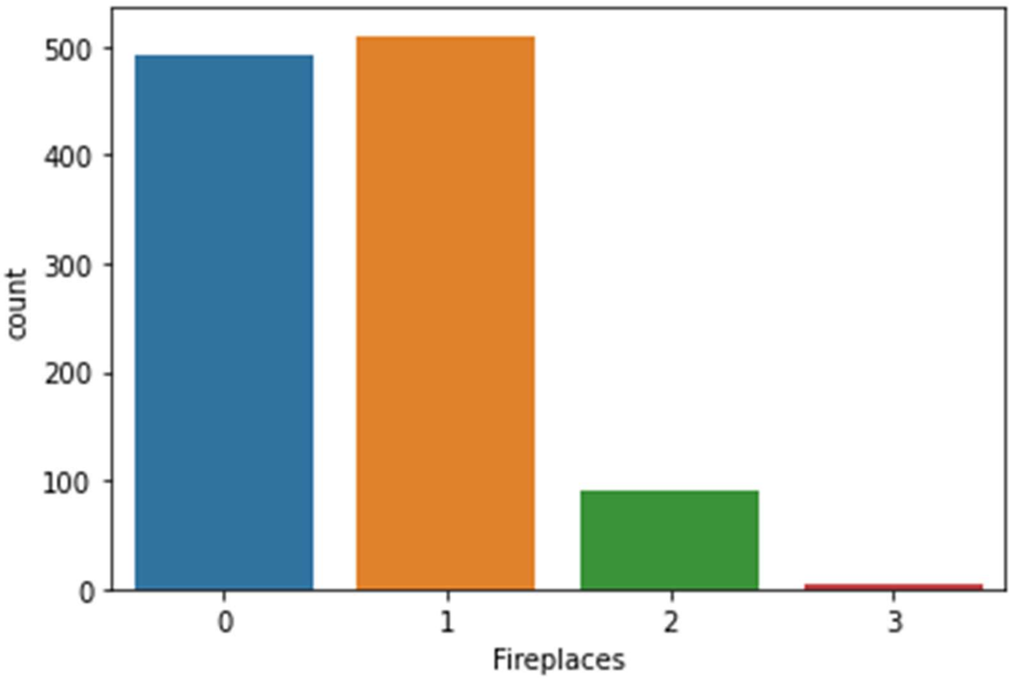

0
200
400
600
800
1000
1200
count
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType

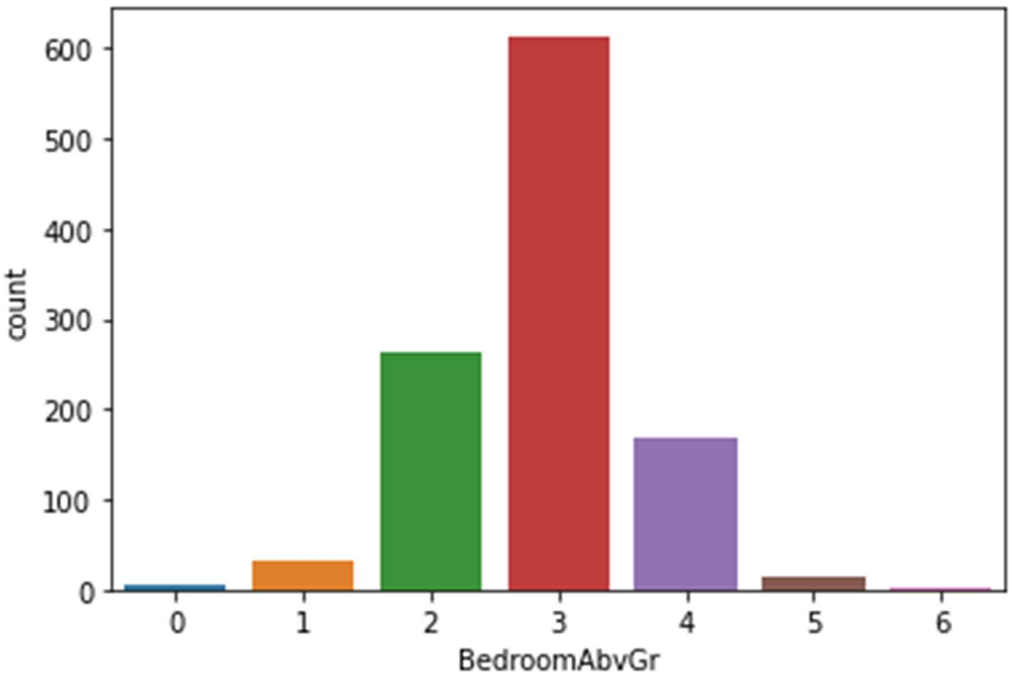

0
100
200
300
400
500
count
0
1
2
3
4
5
Foundation

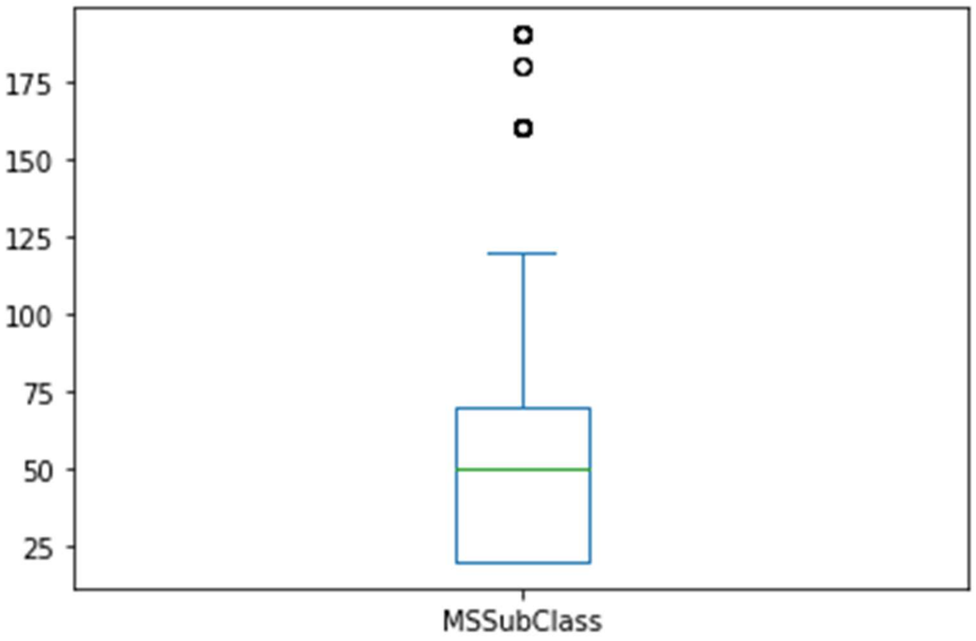

count
800
600
400
200
0
0
1
2
3
4
5
SaleCondition

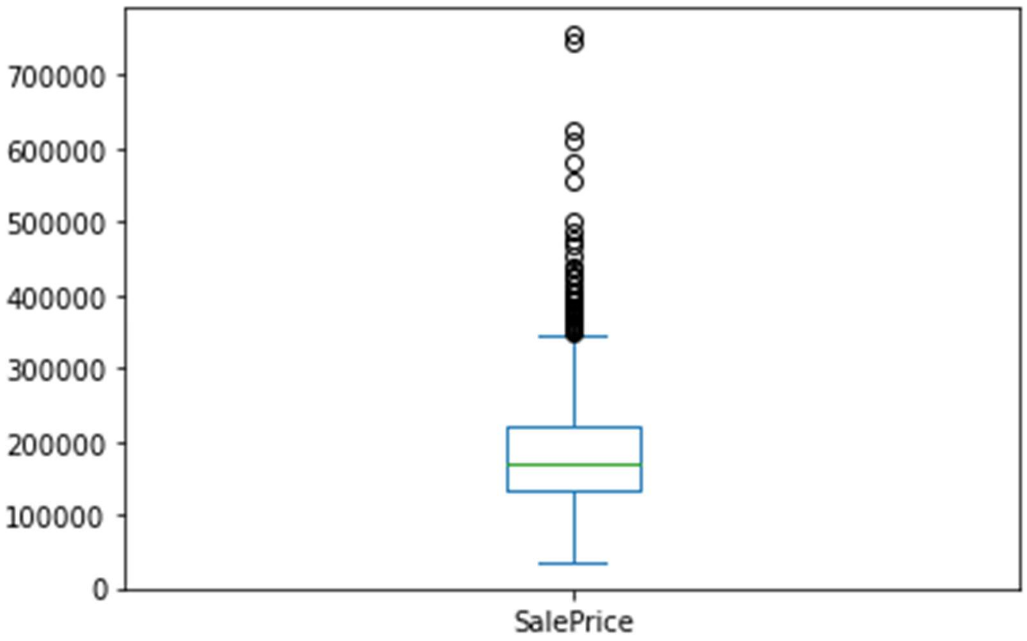

count
500
400
300
200
100
0
0
1
2
3
KitchenQual

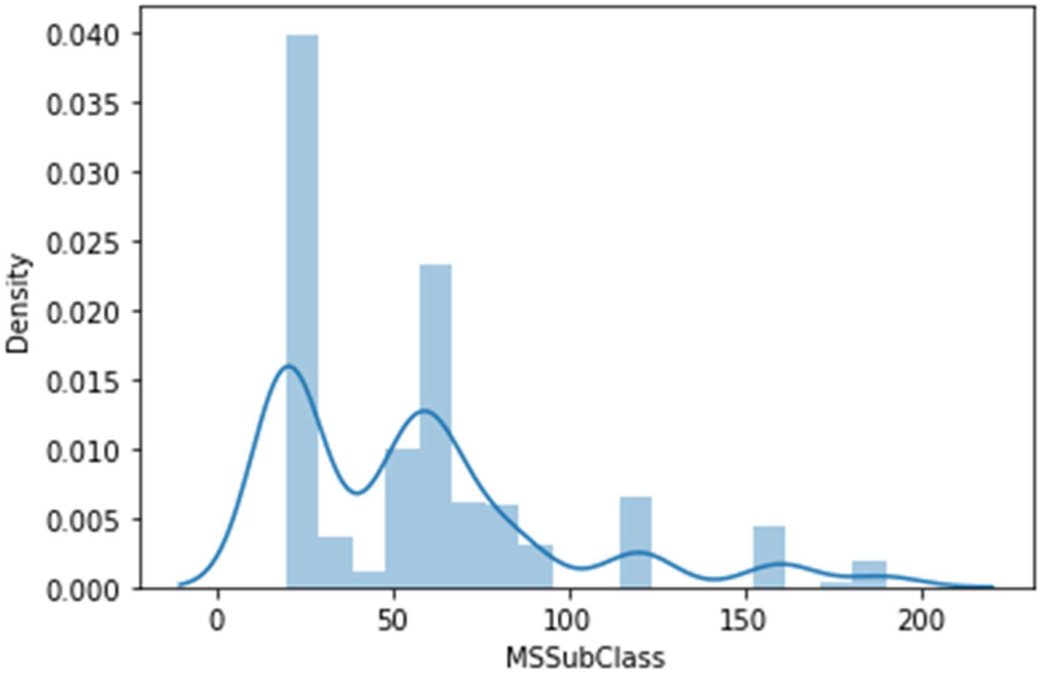

count
0
200
400
600
800
1000
0
1
CentralAir

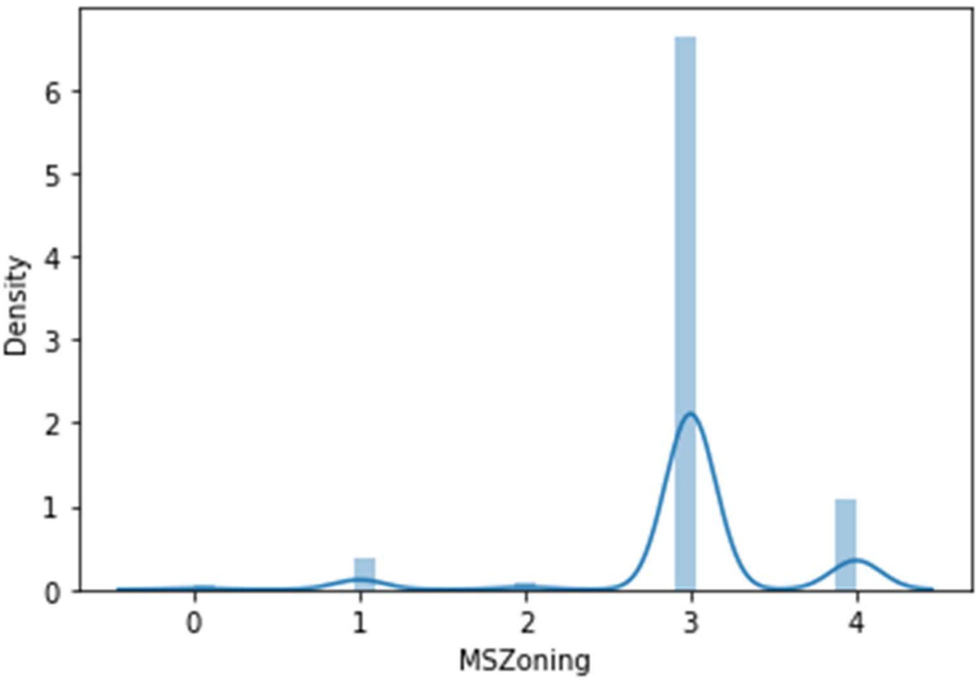

count
0
100
200
300
400
500
0
1
2
3
Fireplaces

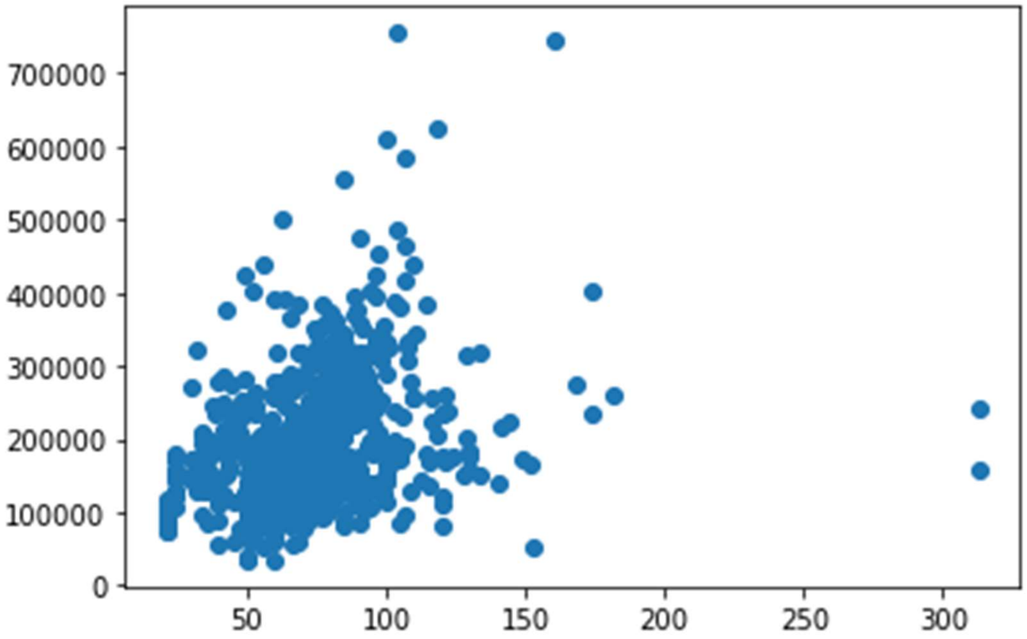

count
600
500
400
300
200
100
0
0
1
2
3
4
5
6
BedroomAbvGr

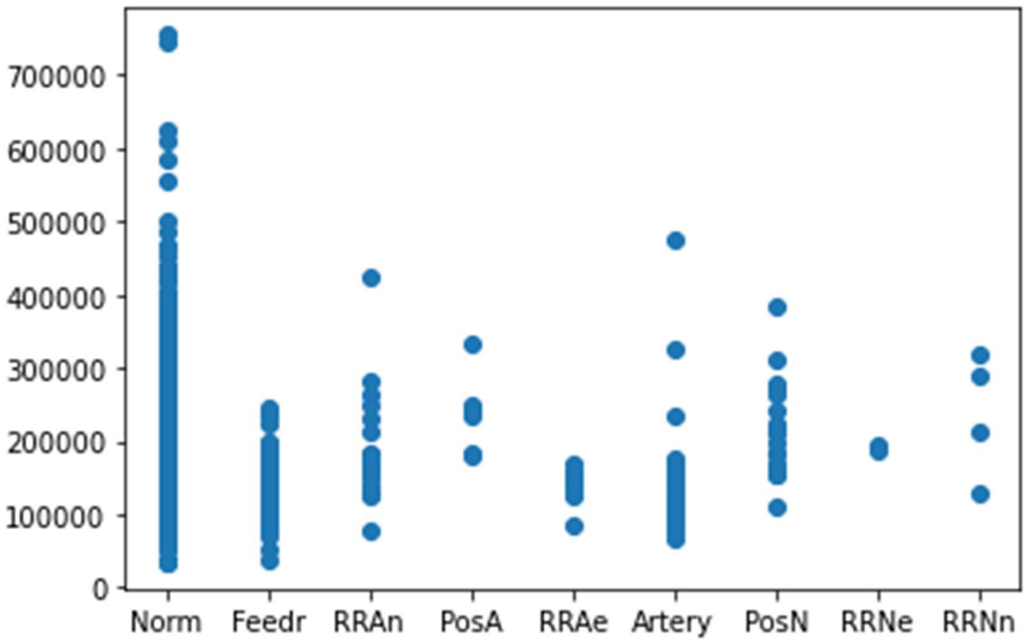

175
150
125
100
75
50
25
MSSubClass

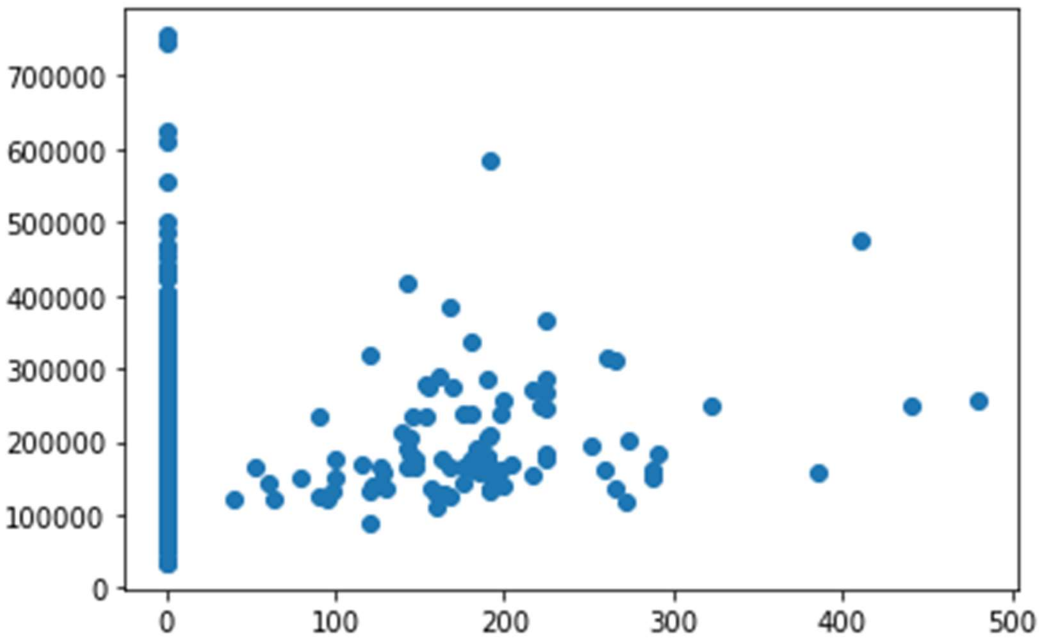

700000
600000
500000
400000
300000
200000
100000
0
SalePrice

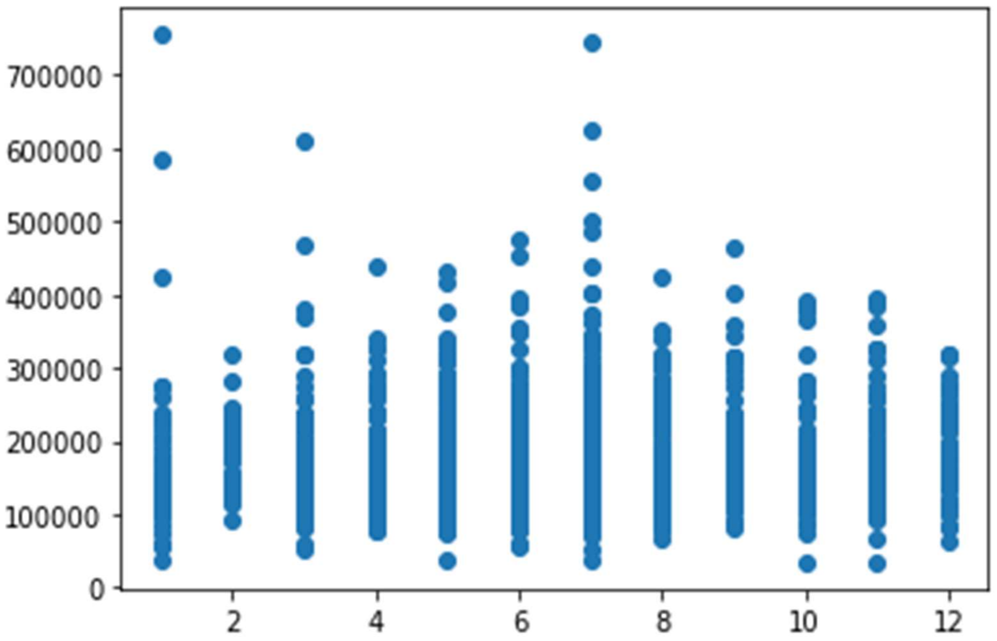

0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
50
100
150
200
MSSubClass


Density
0
1
2
3
4
5
6
0
1
2
3
4
MSZoning


0
100000
200000
300000
400000
500000
600000
700000
50
100
150
200
250
300


700000
600000
500000
400000
300000
200000
100000
0
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn


700000
600000
500000
400000
300000
200000
100000
0
0
100
200
300
400
500



- Interpretation of the Results

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

The performance score is 86%

CONCLUSION

So, our Aim is achieved as we have successfully ticked all our parameters as mentioned. It is seen that the Linear Regression is the most effective model for our Dataset with 86% score.

- Limitations of this work and Scope for Future Work

1. LIMITATIONS

There is a notable amount of research done in the house price prediction department but very research has come up to any real-life solutions. There is very little evidence of a working house price predictor set up by a company. For now, very few digital solutions exist for such a huge market and most of the methods used by people and companies are as follows:

Buyers/Customers:

1. When people first think of buying a house/Real estate they tend to go online and try to study trends and other related stuff. People do this so they can look for a house which contains everything they need. While doing these people make a note of the price which goes with these houses. However, the average person doesn't have detailed knowledge and accurate information about what the actual price should be. This can lead to misinformation as they believe the prices mentioned on the internet to be authentic.

2. The second thing that comes to mind while searching for a property is to contact various Estate agents. The problem with this is these agents need to be paid a fraction of the amount just for searching a house and setting a price tag for you. In most cases, this price tag is blindly believed by people because they have no other options. There might be cases that the agents and sellers may have a secret dealing and the customer might be sold an overpriced house without his/her knowledge.

Seller/Agencies:

3. When an individual thinks of selling his/her property they compare their property with hundreds and thousands of other properties which are posted all around the world. Determining the price by comparing it with multiple estates is highly time-consuming and has a potential risk of incorrect pricing.

4. Large Real estate companies have various products they need to sell and they have to assign people to handle each of these products. This again bases the prediction of a price tag on a human hence there is room for human error. Additionally, these assigned individuals need to be paid. However, having a computer do this work for you by crunching the heavy numbers can save a lot of time money and provide accuracy which a human cannot achieve.

SCOPE:

Machine Learning (ML) is a vital aspect of present-day business and research. It progressively improves the performance of computer systems by using algorithms and neural network models. Machine Learning algorithms automatically build a mathematical model using sample data also referred to as training data which form decisions without being specifically programmed to make those decisions.

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.

The primary aim of this paper is to use these Machine Learning Techniques and curate them into ML 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.

Additionally, there exists a possibility of under-pricing the product. If the price is predicted for these users this might help them get estates for their deserving prices not more not less.