PFA HOUSING REPORT

FEBRUARY 19

FLIP ROBOS TECHNOLOGIES
Authored by: SEEP BANSAL

ACKNOWLEDGEMENT

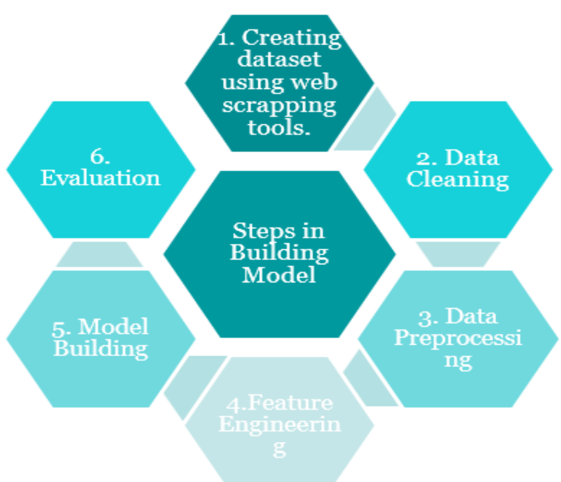
I would like to express my special thanks of gratitude to the team Data Trained and my Mentor MS Deepika Sharma for their exemplary guidance, monitoring and constant encouragement thought the journey of learning Data science and Machine learning techniques. I would also like to express my heartly gratitude to the support team of data trained for their constant support. Last but not the least, I would also like to thank the team of Flip Robo technologies for giving me this opportunity to work on this project and the mentors in Flip Robo Technologies who are constantly guiding me to enhance my knowledge and work. This project helped me not only to learn how to do proper research but also helped me in learning many new things.

INTRODUCTION

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. 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. Dataset has been provided in the form of test and train data where we will be using training dataset to train the data and test dataset has been given where we are supposed to predict the sale price on the basis of various features provided.

PROJECT SUMMARY

Though it is one of the most challenging task to understand what factors actually influence the housing price. The focus of this project is developing machine learning models that can accurately predict the housing price based on its id, msclass, lotarea, street, utilities, neighborhood, basement details and many more. Various machine learning algorithms like Linear Regression, Support Vector Mechanism, Decision Tree Regressor, Random Forest Regressors, K Neighbours Regressor are implemented and evaluated to predict the housing sale price. The best results are given by Random Forest Regressor. Though conventional Linear Regression also gave the good results with the advantage of significantly lower training time as compared to the aforementioned algorithm.



MOTIVATION

Deciding the housing price is actually very difficult. Factors like its id, msclass, lotarea, street, utilities, neighborhood, basement details and many more actually effects the price of the house. It is actually very tough for the sellers to decide upon the price so building one such model will not help the sellers i.e are they quoting the right price for the right product as well the buyers whether it is worthy to buy the product.

ANALYTICAL PROBLEM FRAMING

DATASET

For this project, the company has collected a data set from the sale of houses in Australia. The data for training and testing has been provided in separate csv files. To train the data we will be using the train dataset and for testing purpose will be using the test dataset. Train dataset contains 1168 rows and 81 columns while the test dataset contains 292 rows and 80 columns.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdc
127	120	RL		4928	Pave		IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	Norm	TwnhsE	1Story	6	5	1976	197
889	20	RL	95	15865	Pave		IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	Norm	1Fam	1Story	8	6	1970	197
793	60	RL	92	9920	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	Norm	1Fam	2Story	7	5	1996	199
110	20	RL	105	11751	Pave		IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	6	1977	197
422	20	RL		16635	Pave		IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	7	1977	200
1197	60	RL	58	14054	Pave		IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	2006	200
561	20	RL		11341	Pave		IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1957	199
1041	20	RL	88	13125	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	4	1957	200
503	20	RL	70	9170	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	7	1965	196
576	50	RL	80	8480	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5Fin	5	5	1947	195
449	50	RM	50	8600	Pave		Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	6	6	1937	195

YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	Exterior2nd	MasVnrType	MasVnrArea	ExterQual	ExterCond	Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	BsmtFinType2	BsmtFinSF2	BsmtUnfSF
1976	1976	Gable	CompShg	Plywood	Plywood	None	0	TA	TA	CBlock	Gd	TA	No	ALQ	120	Unf	0	958
1970	1970	Flat	Tar&Grv	Wd Sdng	Wd Sdng	None	0	Gd	Gd	PConc	TA	Gd	Gd	ALQ	351	Rec	823	1043
1996	1997	Gable	CompShg	MetalSd	MetalSd	None	0	Gd	TA	PConc	Gd	TA	Av	GLQ	862	Unf	0	255
1977	1977	Hip	CompShg	Plywood	Plywood	BrkFace	480	TA	TA	CBlock	Gd	TA	No	BLQ	705	Unf	0	1139
1977	2000	Gable	CompShg	CemntBd	CmentBd	Stone	126	Gd	TA	CBlock	Gd	TA	No	ALQ	1246	Unf	0	356
2006	2006	Gable	CompShg	VinylSd	VinylSd	None	0	Gd	TA	PConc	Gd	TA	Av	Unf	0	Unf	0	879
1957	1996	Hip	CompShg	Wd Sdng	Wd Sdng	BrkFace	180	TA	TA	CBlock	Gd	TA	No	ALQ	1302	Unf	0	90
1957	2000	Gable	CompShg	Wd Sdng	Wd Sdng	BrkCmn	67	TA	TA	CBlock	TA	TA	No	Rec	168	ALQ	682	284
1965	1965	Hip	CompShg	MetalSd	MetalSd	None	0	TA	TA	CBlock	TA	TA	No	ALQ	698	GLQ	96	420
1947	1950	Gable	CompShg	MetalSd	MetalSd	None	0	TA	TA	CBlock	TA	TA	No	Rec	442	Unf	0	390
1937	1950	Gable	CompShg	MetalSd	MetalSd	None	0	TA	TA	BrkTil	TA	TA	No	Unf	0	Unf	0	780
2003	2003	Gable	CompShg	VinylSd	VinylSd	BrkFace	223	Gd	TA	PConc	Gd	TA	No	GLQ	483	Unf	0	458
2003	2003	Gable	CompShg	VinylSd	VinylSd	None	0	Gd	TA	PConc	Ex	TA	No	Unf	0	Unf	0	1560
1960	1960	Gable	CompShg	MetalSd	MetalSd	BrkCmn	66	TA	TA	CBlock	TA	TA	No	Unf	0	Unf	0	1065
1955	1955	Gable	CompShg	MetalSd	MetalSd	None	0	TA	Gd	PConc	TA	TA	No	Unf	0	Unf	0	816
1923	1996	Hip	CompShg	Wd Sdng	Wd Sdng	None	0	TA	Gd	PConc	TA	Fa	No	Unf	0	Unf	0	602
1930	2007	Gable	CompShg	Wd Sdng	Wd Sdng	None	0	Gd	Gd	BrkTil	TA	TA	Av	ALQ	538	Unf	0	278
2007	2007	Gable	CompShg	VinylSd	VinylSd	Stone	82	Gd	TA	PConc	Gd	TA	Av	GLQ	724	Unf	0	508
1976	1976	Hip	CompShg	HdBoard	HdBoard	BrkFace	174	TA	Gd	CBlock	TA	Gd	No	BLQ	751	Unf	0	392

Fig:-----snapshots of dataset

Feature Description:

1. MSSubClass: Identifies the type of dwelling involved in the sale.
2. MSZoning: Identifies the general zoning classification of the sale.
3. LotFrontage: Linear feet of street connected to property
4. LotArea: Lot size in square feet
5. Street: Type of road access to property
6. Alley: Type of alley access to property
7. LotShape: General shape of property
8. LandContour: Flatness of the property
9. Utilities: Type of utilities available
10. LotConfig: Lot configuration
11. LandSlope: Slope of property
12. Neighborhood: Physical locations within Ames city limits
13. Condition1: Proximity to various conditions
14. Condition2: Proximity to various conditions (if more than one is present)
15. BldgType: Type of dwelling
16. HouseStyle: Style of dwelling
17. OverallQual: Rates the overall material and finish of the house
18. OverallCond: Rates the overall condition of the house
19. YearBuilt: Original construction date
20. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
21. RoofStyle: Type of roof
22. RoofMatl: Roof material
23. Exterior1st: Exterior covering on house
24. Exterior2nd: Exterior covering on house (if more than one material)
25. MasVnrType: Masonry veneer type
26. MasVnrArea: Masonry veneer area in square feet
27. ExterQual: Evaluates the quality of the material on the exterior
28. ExterCond: Evaluates the present condition of the material on the exterior
29. Foundation: Type of foundation
30. BsmtQual: Evaluates the height of the basement
31. BsmtCond: Evaluates the general condition of the basement
32. BsmtExposure: Refers to walkout or garden level walls
33. BsmtFinType1: Rating of basement finished area
34. BsmtFinSF1: Type 1 finished square feet
35. BsmtFinType2: Rating of basement finished area (if multiple types)
36. BsmtFinSF2: Type 2 finished square feet
37. BsmtUnfSF: Unfinished square feet of basement area
38. TotalBsmtSF: Total square feet of basement area
39. Heating: Type of heating
40. MiscVal: $Value of miscellaneous feature
41. MoSold: Month Sold (MM)
42. YrSold: Year Sold (YYYY)
43. SaleType: Type of sale
44. SaleCondition: Condition of sale

45. GarageCars: Size of garage in car capacity
46. GarageArea: Size of garage in square feet
47. GarageQual: Garage quality
48. Fireplaces: Number of fireplaces
49. GarageType: Garage location
50. FireplaceQu: Fireplace quality
51. 1stFlrSF: First Floor square feet
52. 2ndFlrSF: Second floor square feet
53. LowQualFinSF: Low quality finished square feet (all floors)
54. GrLivArea: Above grade (ground) living area square feet
55. BsmtFullBath: Basement full bathrooms
56. BsmtHalfBath: Basement half bathrooms
57. FullBath: Full bathrooms above grade
58. HalfBath: Half baths above grade
59. Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
60. Kitchen: Kitchens above grade
61. KitchenQual: Kitchen quality
62. HeatingQC: Heating quality and condition
63. CentralAir: Central air conditioning
64. Electrical: Electrical system
65. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
66. Functional: Home functionality (Assume typical unless deductions are warranted)
67. Fireplaces: Number of fireplaces
68. FireplaceQu: Fireplace quality
69. GarageYrBlt: Year garage was built
70. GarageFinish: Interior finish of the garage
71. GarageCars: Size of garage in car capacity
72. GarageArea: Size of garage in square feet
73. GarageQual: Garage quality
74. GarageCond: Garage condition
75. PavedDrive: Paved driveway
76. WoodDeckSF: Wood deck area in square feet
77. OpenPorchSF: Open porch area in square feet
78. EnclosedPorch: Enclosed porch area in square feet
79. 3SsnPorch: Three season porch area in square feet
80. ScreenPorch: Screen porch area in square feet
81. PoolArea: Pool area in square feet

DATA PREPROCESSING/CLEANING

Preprocessing is one of the important steps in building a model. In this phase we usually deal with missing values if any or if there are any unrealistic values. In case of any irrelevant value, we will remove that data. In case if the data loss is huge then removing/dropping of data is not a good practice. We will try to improve the quality of data in this phase so that we can develop a model with high accuracy score. For the dataset, we will first of all fill the missing values if any, then we have most of the data in categorial form we will convert the data in numerical form so that we are able to fed the data into classification algorithms. We can also check if there exist any multicollinearities through VIF FACTOR calculation.

EDA concluding Remarks:

- ✓ There are null values in the dataset.
- ✓ Outliers are there but removing outliers leads to huge data loss so dropping the data may result in important information.
- ✓ Calculation of VIF Factor to check the presence of multicollinearities among the difference variables.
- ✓ As the data is highly skewed using a power transformation method 'yeo-Johnson' to get rid of skewness.
- ✓ Some of the features were dropped while passing the data to the model as they were giving no contribution in predicting the output variable and were affecting the computational time**.**
- ✓ We will be performing standard scaling technique to bring all the columns on the same scale.
- ✓ The data types of the columns were changed from string to numeric type.
- ✓ Feature extraction were performed i.e many new features were extracted from the existing features.

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements: Hardware Requirements followed while developing this model:

- Intel core i5
- 11th generation
- 16 GB RAM
- Windows 10

Software Requirements: Software required are:

- Anaconda Navigator (64-bit Graphical Installer)
- Juypter Notebook
- Microsoft Edge
- Knowledge of Python Language and Machine learning Algorithms

E. LIBRARIES USED:

```
import pandas as pd   #importing the libraries
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression,Lasso,Ridge
from sklearn.metrics import mean_squared_error, mean_absolute_error,r2_score
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from scipy.stats import zscore
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import roc_curve,roc_auc_score
from sklearn.model_selection import GridSearchCV
import statsmodels.api as sm
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

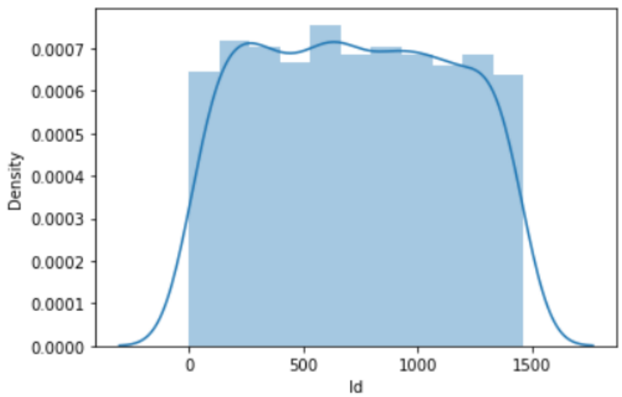
Figure 3: Libraries used

- Import and export of data takes place with the help of pandas
- All the numerical operations are carried out with the help of numpy library.
- Matplotlib.pyplot and seaborn libraries helps in graphical representation of data.
- Warnings library is used to ignore the unwanted warnings
- Sklearn library helps in importing all the machine learning algorithms and evaluation matrix that are required.

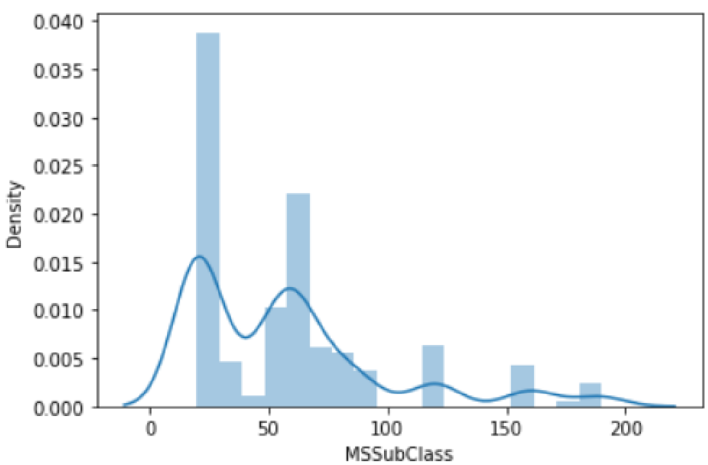
MODEL DEVELOPMENT/EVALUATION

VISUALIZATION

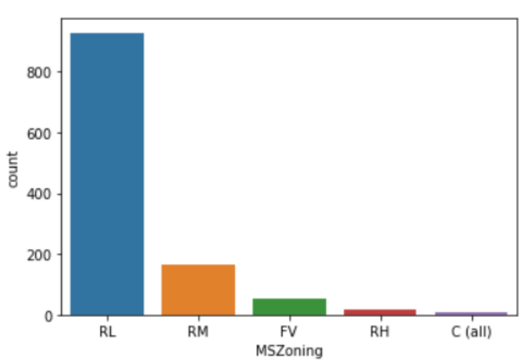
<AxesSubplot:xlabel='Id', ylabel='Density'>



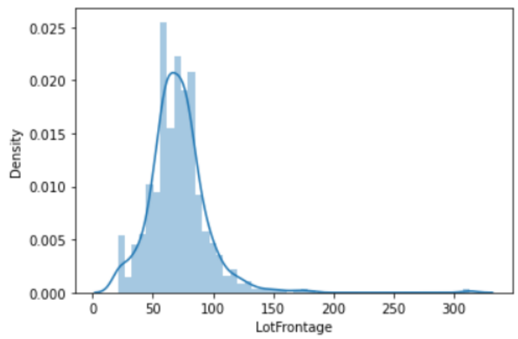
from above plot we can see that ID which is unique identification of each record is normally dirtibuted and values vary between 0-1500.



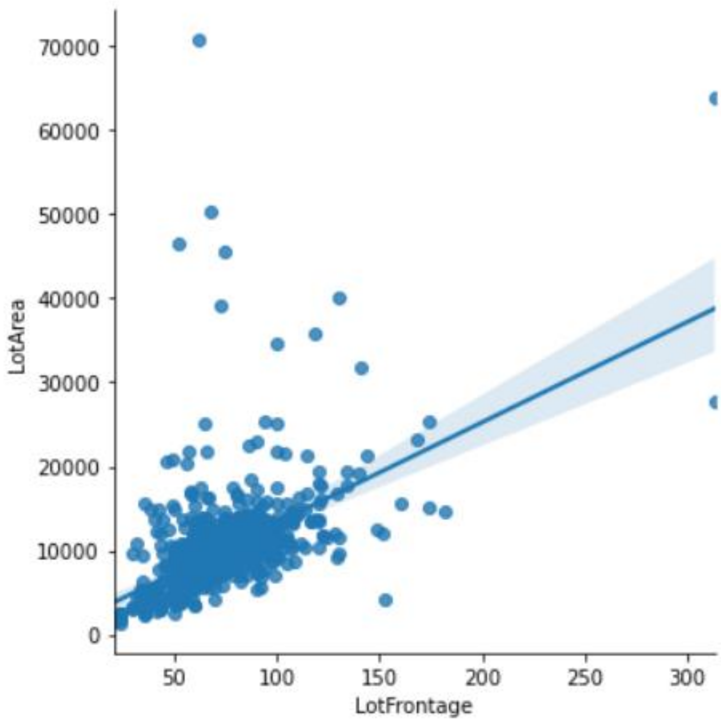
MSSubclass is distributed normally we could not see any skewness and the data is distributed between 0-200



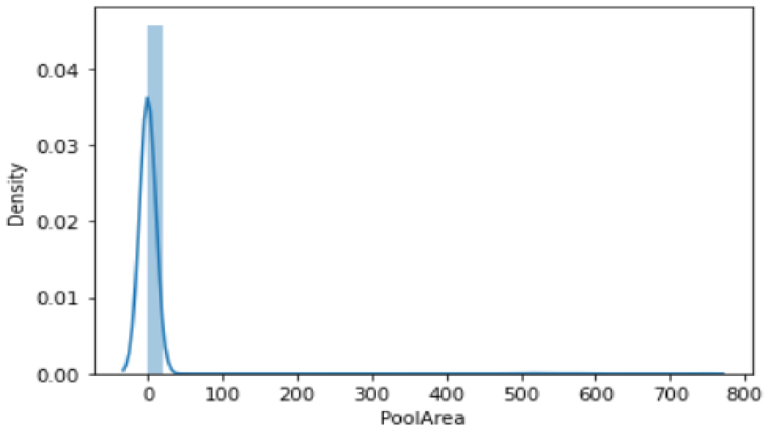
Mszoning identifies the general zoning classification of the sale. RL holds the maximum count. A Agriculture C Commercial FV Floating Village Residential I Industrial RH Residential High Density RL Residential Low Density RP Residential Low Density Park RM Residential Medium Density



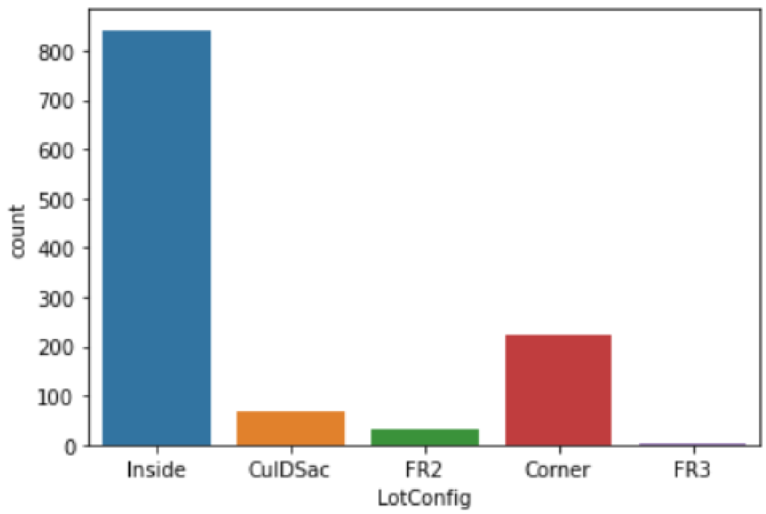
lot frontage means linear feet of street connected to property. from above plot we can see the data is skewed or this could be due to the presence of outliers.



Lot area and lotfrontage are directly propotional to each other.



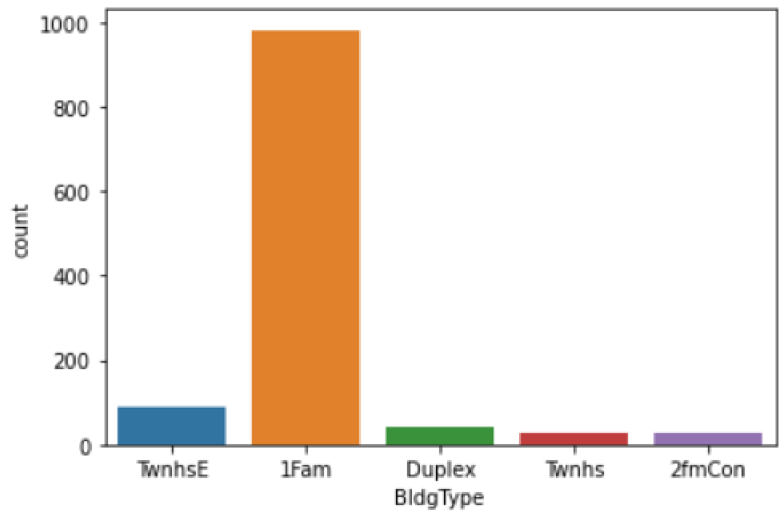
from the above plot we can see that the data is skewed and it could be due to presence of outliers.



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
```

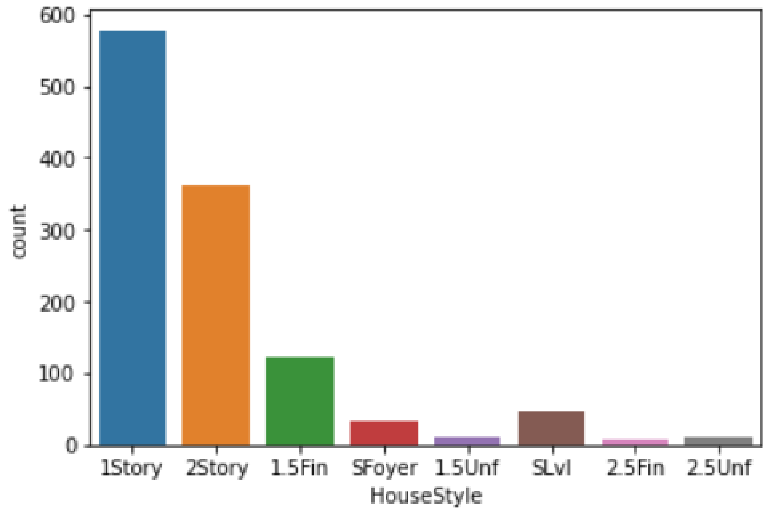
maximum count stands for inside lot.



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
```

maximum count stands for 1Fam.



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
```

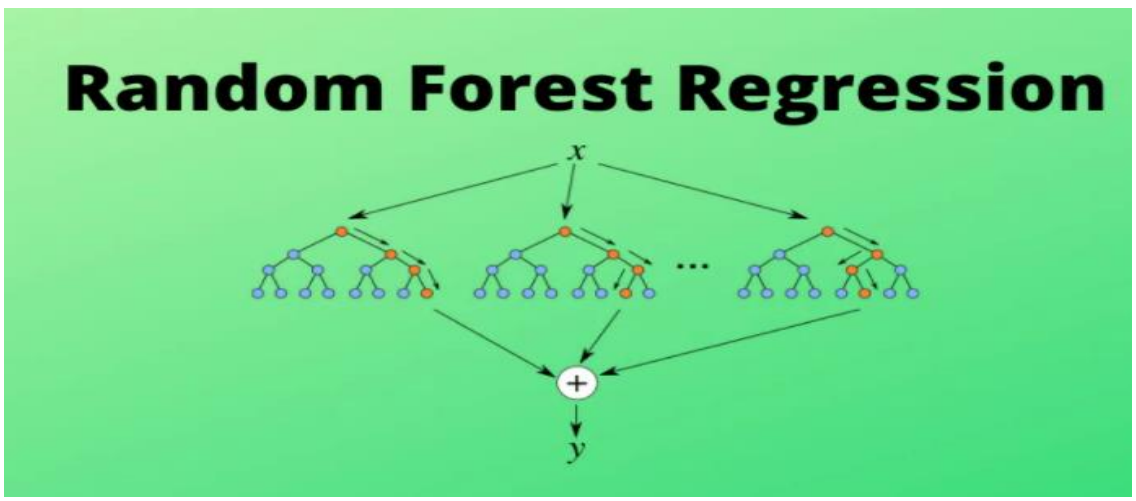
One story holds the highest count.

ALGORITHMS USED For Training and Testing the Model:

- Linear Regression Model
- Decision Tree Regressor
- Random Tree Forest Regressor
- K Neighbors Regressor
- Support Vector Regressor

REGRESSION EXPERIMENTS

We have performed various Regression algorithms like Linear Algorithm, Decision Tree Algorithm, Support vector Regressor, K Neighbors Regressor, Random Tree Regressor to check the MSE, MAE, RMSE. The main algorithm which has predicted the best results is the Random Forest Tree Regressor. Random forest regression is an ensemble learning technique which takes multiple algorithms or same algorithm multiple times and put together a model that's more powerful than the original.



Random Forest Regressor.

EVALUATION

Evaluation of model plays a very important role in evaluating the performance of any Regression. The metrices that are evaluated here are the R2 score, MSE, MAE, RMSE.Time taken to test the model on dataset plays a very crucial role.Here, for Random Tree Forest Classifier the accuracy score we are getting is 97% and the RMSE is 35927 which is least in any of the model build and the computational time is very less.

```
final=RandomForestRegressor(max_features='log2', bootstrap=True, oob_score=False, max_depth=19)
final.fit(x_train1,y_train1)
sw1=final.score(x_train1,y_train1)
#print("Coffecient is: ",dtr.coef_)
#print("Intercept is: ",dtr.intercept_)
print("Score is: ",sw1)
pred=rfr.predict(x_test1)
print("Mean Squared Error is:",round(mean_squared_error(y_test1,pred),2))
print("Mean Absolute Error is:",round(mean_absolute_error(y_test1,pred),2))
print("R2 Score is:",round(r2_score(y_test1,pred),2))
print("RSME",np.sqrt(mean_squared_error(y_test1,pred)))
```

```
Score is:  0.9788391489796187
Mean Squared Error is: 1290796785.95
Mean Absolute Error is: 18840.35
R2 Score is: 0.8
RSME 35927.66045753526
```

Figure: Summary of the evaluation metrices

CONCLUSION

Though it is very difficult to decide the price of the house by the various variables provided by the client by developing this model and performing different algorithms was aimed to get the different perspectives. The main aim of building this model was to quote the price of car in such a way that makes easy for sellers to sell and buyers to buy. The various data visualization techniques were used. Data was analyzed from different point of views many preprocessing techniques like scaling, label encoding etc were followed. The relation between different features were examined and the best model Random Forest Regressor was used to predict the best price of the house.

LIMITATIONS AND FUTUTRE SCOPE

The above model is used to predict the price of the house on the basis of the dataset provided by the client. However, this was relatively a small dataset with around 1168 rows was used make a strong inference.Room of improvement is more data preprocessing and more data cleaning techniques to be followed which could help is reducing RMSE. We could actually build this model with different scaling technique like min-max scaler which may give us better results or we could also try different encoding technique which may enhance the results.