FLIP ROBO

HOUSING PRICE PREDICTION

Submitted by:

MANISHA

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

CONTENTS

INTRODUCTION

Business Problem Framing:

The market is expanded in various fields but here we talk about houses. Houses are one of the most important requirement of each and every person in their life. There are different variables on which we have to predict the actual value of the prospective properties and decide whether to invest in them or not.

This problem is a real world problem. A seller set a price of their house according to investment and hardwork. The customer also check whether they fulfil their requirements, provide nearby facilities, the property rate etc. After go through each criteria's we take decision that the price rate is appropriate or not and the amount of the house is in a budget or not.

Conceptual Background of the Domain Problem:

We have to predict the SalesPrice so we have to go through all the feature variables and try to understand what information they are giving to us. Then select the feature variable on the behalf of our analysis and they are related or effect on our prices, so go with that meaningful data which gives better values for our target data.

Review of Literature:

In the project we first import the excel file and then check the size of the data. Clean the null values in the data and then visualise the categorical data and continuous data. Convert the object data type into float because analysis do only on numerical data. Describe the data where we get all statistical information, then find correlation. Check the outliers and skewness on the continuous data type. Fit the regression model and select the best model in all of them. Then tuning is applied to get optimize score.

Motivation for the Problem Undertaken:

The objective of the project is to get the efficient Sales price of the houses. In the problem we have large number of features so analyse all the features and find the variables that are important for our prediction. And the suitable price for the house

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

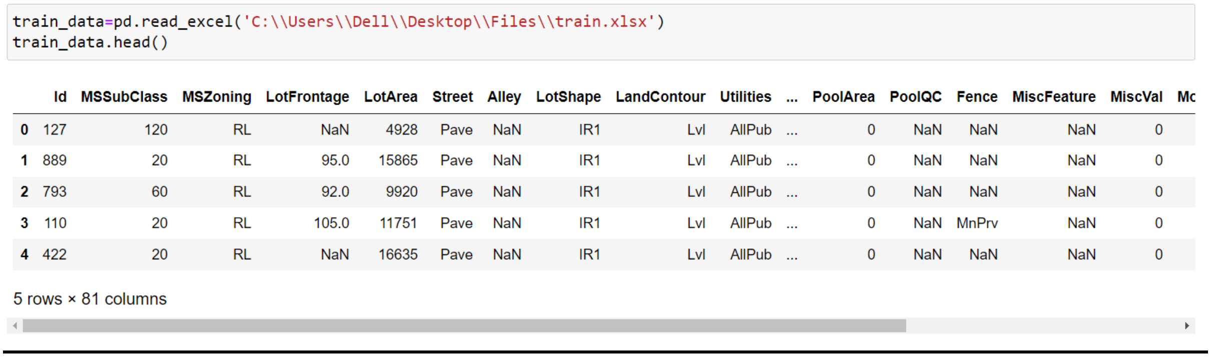
We use Statistical techniques and Supervised Machine Learning models in our projects, such as:

- describe() : use to calculate the statistical values that are mean, standard deviation, quantile deviation, minimum and maximum values.
- corr(): use to calculate the relation between feature variable with the target variable.
- skew(): use to check whether the skewness is present in the continuous data or not.
- Machine Learning Models
  Here we work on the regression problem so the models are:
    - Linear Regression
    - Decision Tree Regression
    - Random Forest Regression
    - Ada Boost Regression
    - Gradient Boosting Regression

Data Source and their formats:

There are two data set one is train and another one is test. In train data we have 81 columns with feature & target variables and 1168 rows. In test data we have 80 columns of feature variables & predict the target values and have 292 rows.

train_data

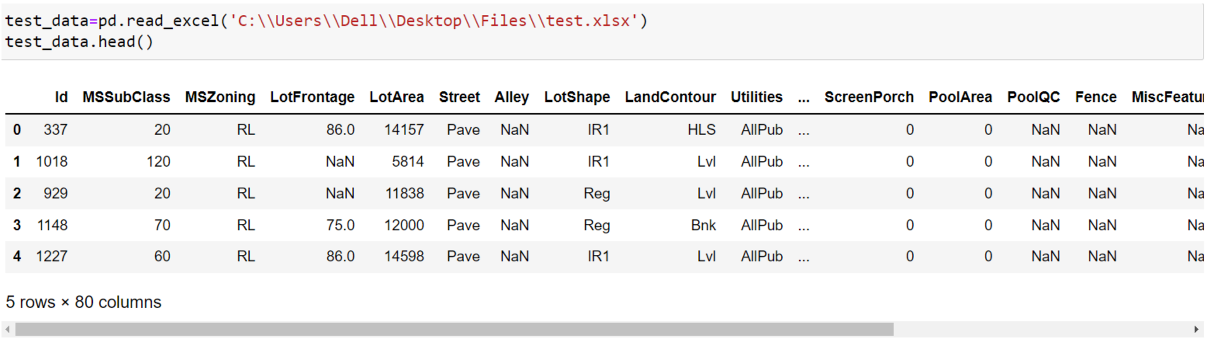


```
train_data=pd.read_excel('C:\\Users\\Dell\\Desktop\\Files\\train.xlsx')
train_data.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	Mo
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	

5 rows × 81 columns

test_data



```
test_data=pd.read_excel('C:\\Users\\Dell\\Desktop\\Files\\test.xlsx')
test_data.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeatu
0	337	20	RL	86.0	14157	Pave	NaN	IR1	HLS	AllPub	...	0	0	NaN	NaN	Na
1	1018	120	RL	NaN	5814	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	Na
2	929	20	RL	NaN	11838	Pave	NaN	Reg	Lvl	AllPub	...	0	0	NaN	NaN	Na
3	1148	70	RL	75.0	12000	Pave	NaN	Reg	Bnk	AllPub	...	0	0	NaN	NaN	Na
4	1227	60	RL	86.0	14598	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	Na

5 rows × 80 columns

Data Pre-processing

We check the number of null values on both the train and test data and we get number of columns having null values but in four columns have more than 80% missing data so we drop them. And fill the continuous values by mean() and nominal by mode(). We check outliers and skewness of the feature variable and trying to remove them. After removing we again gets some columns having skewness so those columns also dropped out.

Data inputs-Logic-Output Relationships

In the data there are 80 columns of input variable and SalesPrice is the Output variable. There are various input data, some of them have its own importance and some of them have less relation so we check and calculate the effect on our price.

Hardware and Software Requirements and Tools used

Hardware:

- Memory 16GB minimum
- Hard Drive SSD is preferred 500GB
- Processor intel i5 minimum
- Operating system Windows 10

Software:

- Jupyter notebook (Python)

Libraries:

pandas (used to create the data and read the data)

numpy (used with the mathematical function)

seaborn (used to create a different types of graphs)

matplotlib (used to plot the graph)

variance inflation factor (used to check the existence of multicollinearity)

zscore (used to remove outliers)

GridSearchCV (give the best parameters for tuning the model)

r2_score (used to calculate the train and test score)

These are the regression models:

LinearRegresion

DecisionTreeRegression

RandomForestRegression

AdaBoostRegression

GradientBoostingRegression

Model/s Development and Evaluation

Identification of possible problem:

We approach to both statistical and analytical problem

- Plot a bar graph for nominal data and distribution graph for continuous data
- describe () use to calculate mean, standard deviation, minimum, maximum and quantile deviation
- corr() used to calculate the correlation of input variable with the output variable.
- zscore used to calculate the outliers
- skew() used to calculate the skewness of the data

Testing of Identified Approaches:

All the algorithm used for training and testing data

Linear Regression

Decision Tree Regression

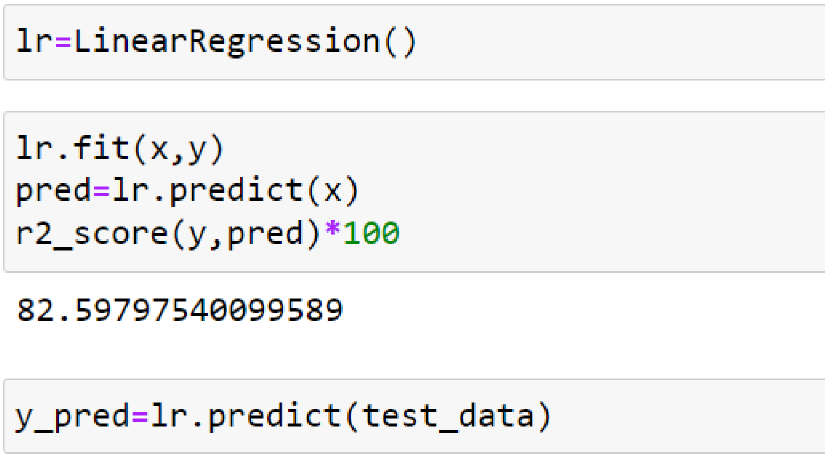
Random Forest Regression

Ada Boost Regression

Gradient Boosting Regression

Run and Evaluate selected models:

Linear Regression



```
lr=LinearRegression()
```

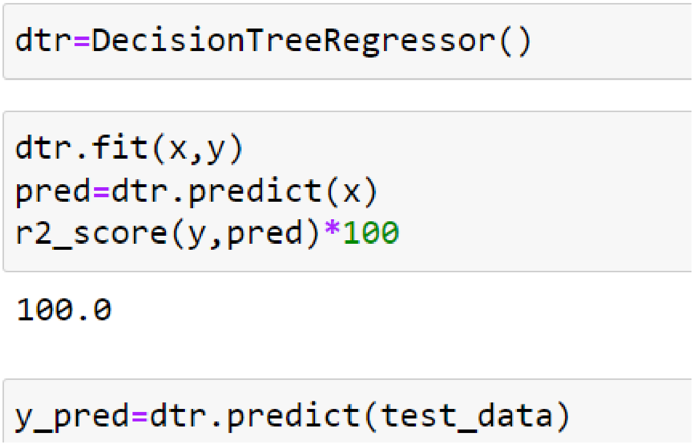
```
lr.fit(x,y)
pred=lr.predict(x)
r2_score(y,pred)*100
```

```
82.59797540099589
```

```
y_pred=lr.predict(test_data)
```

Linear regression’s r2_score for the train data is 82.59 and the test value is predicted according to the train data.

Decision Tree Regression



```
dtr=DecisionTreeRegressor()
```

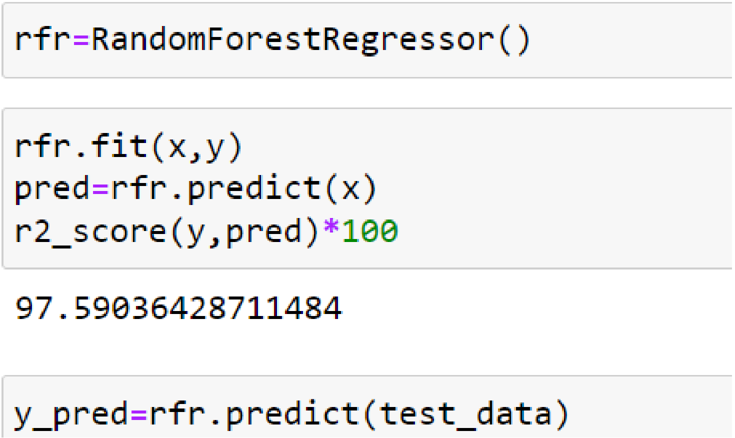
```
dtr.fit(x,y)
pred=dtr.predict(x)
r2_score(y,pred)*100
```

```
100.0
```

```
y_pred=dtr.predict(test_data)
```

The r2_score for the decision tree regression is 100 which shows that there is no variation in input and output variable.

Random Forest Regression



```
rfr=RandomForestRegressor()
```

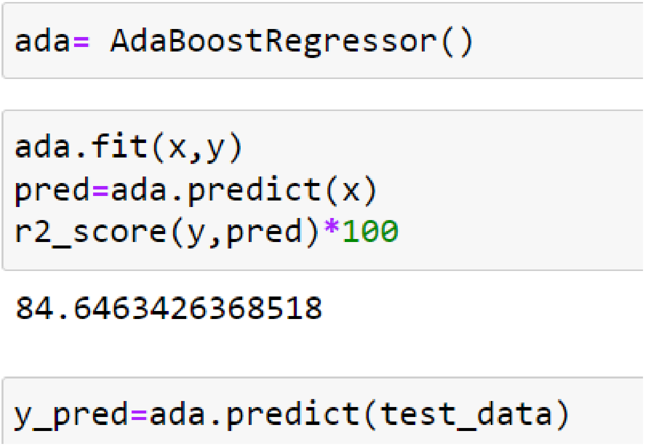
```
rfr.fit(x,y)
pred=rfr.predict(x)
r2_score(y,pred)*100
```

```
97.59036428711484
```

```
y_pred=rfr.predict(test_data)
```

The r2_score of the random forest regression is 97.59 which is less than the decision tree regression.

Ada Boost Regression



```
ada= AdaBoostRegressor()
```

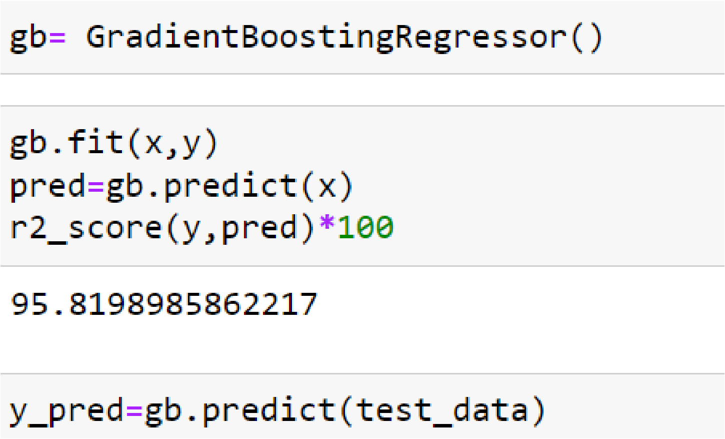
```
ada.fit(x,y)
pred=ada.predict(x)
r2_score(y,pred)*100
```

```
84.6463426368518
```

```
y_pred=ada.predict(test_data)
```

The r2_score of the ada boost regression is 84.64 The Linear regression and Ada boost regression is not much good regression.

Gradient Boosting Regression



```
gb= GradientBoostingRegressor()
```

```
gb.fit(x,y)
pred=gb.predict(x)
r2_score(y,pred)*100
```

```
95.8198985862217
```

```
y_pred=gb.predict(test_data)
```

The 2_score of the gradient boosting regression is 95.81

After trained all the models we get the better model for predicting the SalesPrice for test_data is Decision Tree Regression.

Visualisation:

We visualise nominal data by bar graphs and continuous data by distribution graph. The large data is of categorical type so we plot number of graphs and try to analyse the columns whom are beneficial for modelling in predicting the better score and Prices. On visualising we select important feature variables.

Interpretation of the Results:

On visualising the dataset we saw that there are lots of feature such as Area, available rooms, bathrooms, kitchen, Garage year, built year etc., they all are helpful in predicting the sales price.

On the pre-processing technique we dropped number of features on the basis of missing values, have no correlation with the target variable and mostly the continuous data is skewed which shows that it will not effect on predicting the SalesPrice for the test data.

We fit the regression model on the train data and try to find the better price for the test data. Here we get Decision Tree Regression is the better in all of the model so we tune the parameters and predicts the SalesPrice for the test data.

Conclusion

Key Findings and Conclusions of the Study:

When checking null values there are four columns having 80% to 100% data missing. In graphs some columns having mostly same data. In continuous data there are some columns having zeroes in large amount thats why the skewness is present. After transforming if they are skewed so we have to drop those features. Model the data by taking important feature variables.

Learning Outcomes of the Study in respect of Data Science:

In the data there are missing values first we impute by fillna() method and convert object data into float data type. We describe the data, find the correlation, skewness and outliers in continuous data. Apply five regression models and select Decision Tree Regression because it gives better r2_score and predicts the SalesPrice for the test_data.

The feature variables are large so it takes time to go through all of the variables and understand them.

Limitations of this work and Scope for Future Work:

Here we have not get the SalesPrice for the test_data so we can't calculate the r2_score for the testing. If we don't do this so metrics are also not calculated.

For the business we have to take some target values before investing, and after analysing the data then we check those values with given values. And check the variation between the actual and predicted. Also get that whether it is in our budget or not.

References

- Hastie, Friedman, and Tibshirani, The Elements of Statistical Learning, 2001
- Bishop, Pattern Recognition and Machine Learning, 2006
- UCI Machine Learning Repository
- By Kaggle