HOUSING PRICE PREDICTION PROJECT

Submitted by:

ANJAY KRISHNA TU

INTRODUCTION

- 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 the business problem and how this problem can be related to the real world. A US bases company has decided to enter into Australian market and collected dataset of sale houses details. The company is looking at prospective properties to buy houses to enter the market. Need to build a machine learning model to predict the price of the house and decide whether to invest in it or not.

- Motivation for the Problem Undertaken

  By understanding the relation between variables, the main variables which plays important role in changing the price can be used for predicting the price. Depending on the price and properties of variables the management can decide whether to invest in or not. This will help them in effective way.

Analytical Problem Framing

- Data Sources and their formats

  The given dataset contains 1460 rows and 81 columns , which includes details like price, street, sale type, sale condition, etc. The datatypes in the dataset are integer, float and object.

- Data Preprocessing Done

  The dataset has null values in 18 variables, some of the null values are filled by using mean median method and few variables are removed from the dataset. Skewness is present in some variables and outliers are also present. Skewness is removed by taking square root.

- Data Inputs- Logic- Output Relationships

  The SalePrice is the target/output variable and all other variables are the features/input variables. The target is highly positive correlated to OverallQual, GrLivArea, GarageCars and GarageArea. Target highly negative correlated to ExterQual, BsmtQual, KitchenQual, GarageFinish and HeatingQC.

- Hardware and Software Requirements and Tools Used
  - Selenium Webdriver – For scraping the data
  - Anaconda jupyter Notebook – For Programming
  - Pandas, Seaborn, Matplotlib – For Visualization.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  From the analysis it is clear that only few input variables are correlated to the output variables. So, if we focusing on the correlated input variables and removing the unwanted data from the dataset will improve the model.

- Testing of Identified Approaches (Algorithms)

  Here 3 algorithms are used.

  1. Random Forest
  2. Ada boosting
  3. Gradient boosting

- Run and Evaluate selected models

  1.Random Forest

```
:  1 rf=RandomForestRegressor(criterion='poisson',
   2                          max_depth=18,
   3                          max_features='log2',
   4                          min_samples_leaf=2,
   5                          min_samples_split=17)
   6 rf.fit(x_train,y_train)
   7 predrf=rf.predict(x_test)
   8 print('R2 score :',r2_score(y_test,predrf))
   9 print('Mean squared error :',mean_squared_error(y_test,predrf))
  10 print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predrf)))
```

```
R2 score : 0.6198926014376446
Mean squared error : 2646550342.4926066
Root Mean squared error  : 51444.633757979136
```

The random forest regressor is an ensemble method machine learning model. The r2 score of this model is 61.98%.

2.Ada Boosting

Ada boosting is an ensemble machine learning algorithm used for both classification and regression problem. The r2 score of ada boosting model is 72.34%.

```
ad=AdaBoostRegressor(learning_rate=1.0,
                     loss='exponential',
                     n_estimators=70,
                     random_state=52)
ad.fit(x_train,y_train)
predad=ad.predict(x_test)
print('R2 score :',r2_score(y_test,predad))
print('Mean squared error :',mean_squared_error(y_test,predad))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predad)))
```

```
R2 score : 0.7234907701037906
Mean squared error : 1925233762.4891856
Root Mean squared error  : 43877.48582689288
```

3.Gradient Boosting

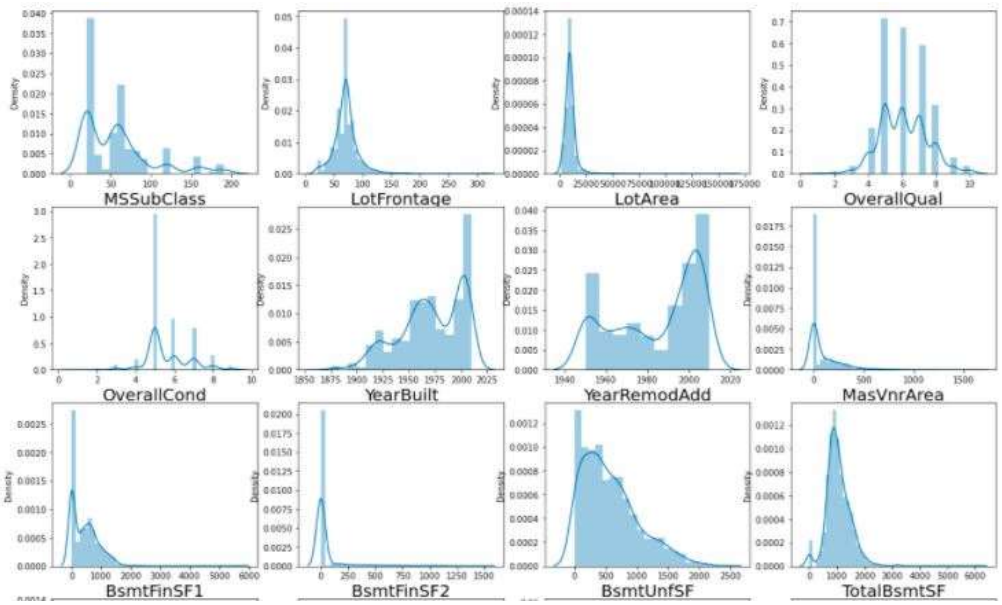
```
gb=GradientBoostingRegressor(learning_rate=0.1,
                             max_features='auto',
                             loss='huber',
                             criterion='mse')
gb.fit(x_train,y_train)
predgb=gb.predict(x_test)
print('R2 score :',r2_score(y_test,predgb))
print('Mean squared error :',mean_squared_error(y_test,predgb))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predgb)))
```

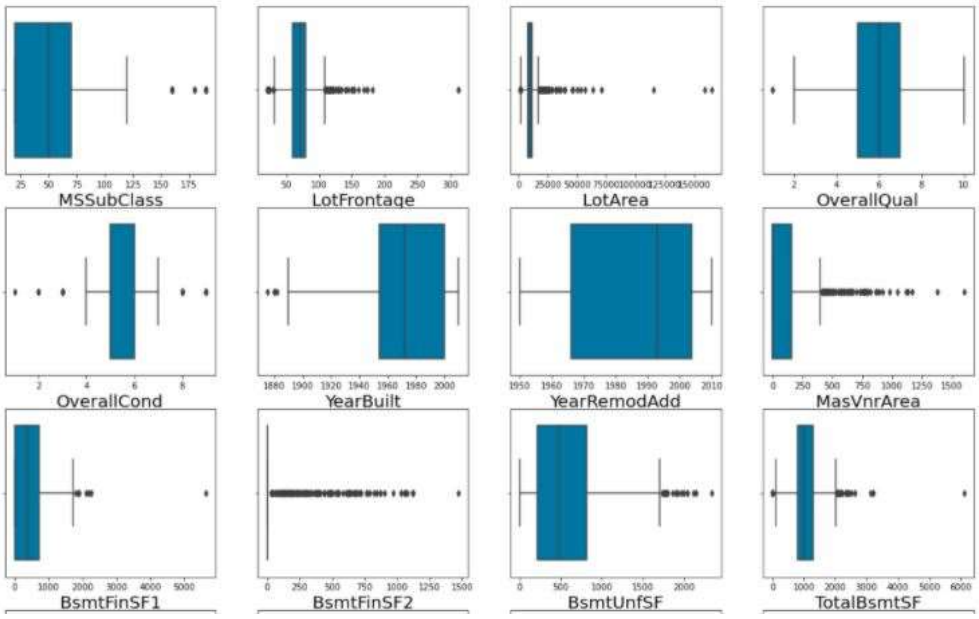
```
R2 score : 0.7883757892355705
Mean squared error : 1473462841.28287
Root Mean squared error  : 38385.711420825195
```

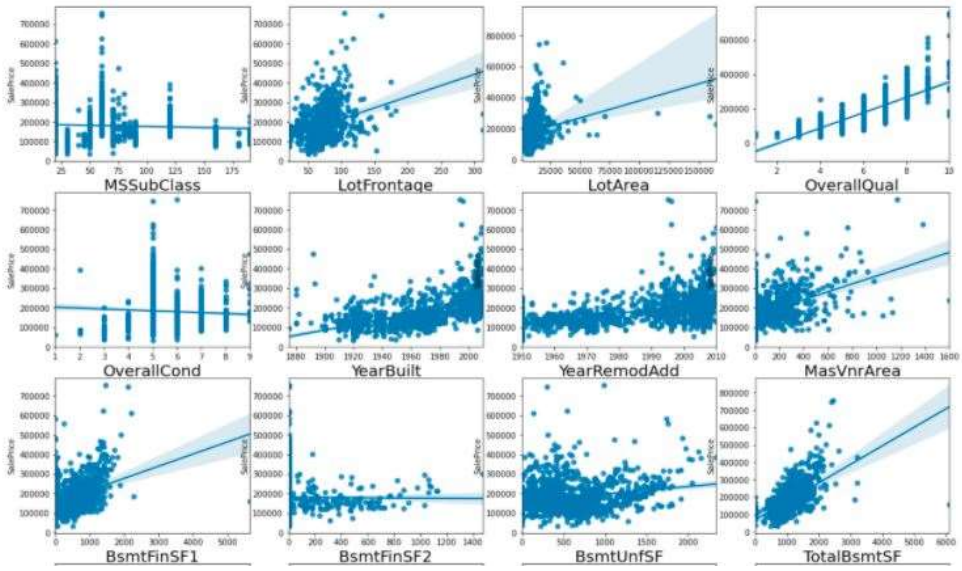
Gradient boosting is an ensemble method used for both classification and regression problem. Weak models are combined to give better result in this algorithm. The r2 score of the model is 78.83%.

- Visualizations

The distribution plot shows skewness is present in some of the variables. The boxplot shows outliers are present in LotFrontage, LotArea,MasVnrArea,BsmtFinSF2,BsmtUnfSF,LowQualFinSF,GrLivArea,GarageArea,WoodDeckSF,OpenPorchSF,EnclosedPorch,ScreenPorch and SalePrice. Regplot shows that some variables are positive linear relation and some other negative linear relation.


MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF

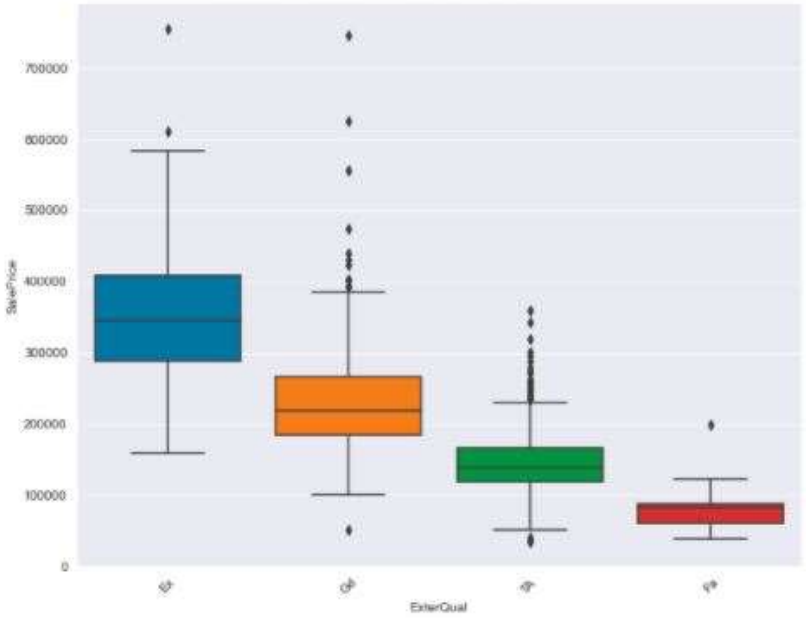

MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF


SalePrice
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF

The price range is different for the houses with different exterior quality.

```
: 1 plt.figure(figsize=(10,8))
  2 plt.xticks(rotation='45')
  3 sns.boxplot(x='ExterQual',y='SalePrice',data=df.sort_values('SalePrice',ascending=False))
```

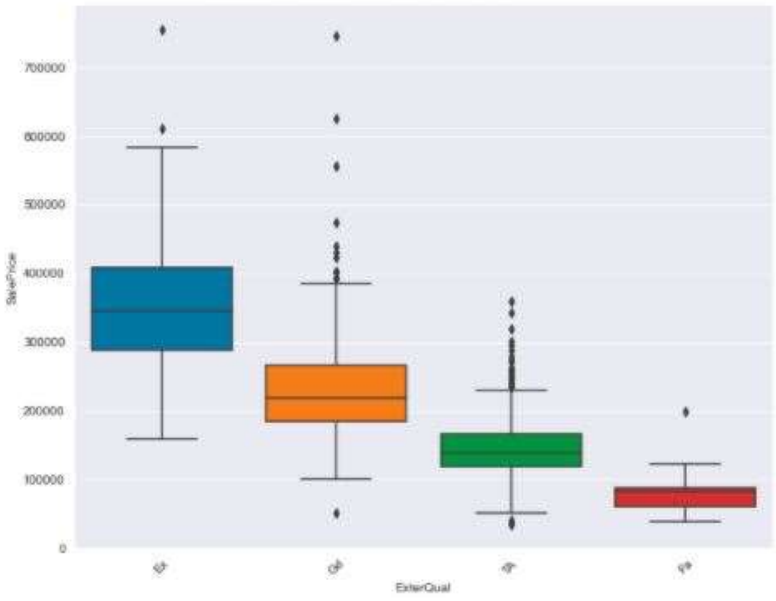
```
: <AxesSubplot:xlabel='ExterQual', ylabel='SalePrice'>
```



The price range also different for the house with different kitchen quality.

```
: 1 plt.figure(figsize=(10,8))
  2 plt.xticks(rotation='45')
  3 sns.boxplot(x='ExterQual',y='SalePrice',data=df.sort_values('SalePrice',ascending=False))
```

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



- Interpretation of the Results

  The variables are positively correlated to the price are plays an important role in predicting the price. Small change in these variables can make an effect in the target variable price.

CONCLUSION

- Key Findings and Conclusions of the Study

  The variables OverallQual, GrLivArea, GarageCars and GargeArea are highly correlated to the price, these variables will decide price of the house. The overall condition and the location will make a change in the price.

- Learning Outcomes of the Study in respect of Data Science

  From the all used algorithm for predicting the price of the house Gradient boosting is the best model with an r2 score of 78.83% and cross validation score of 83%.