HOUSING PRICE PREDICTION

Submitted by:

Balu Gumidelli

ACKNOWLEDGEMENT

It's a great pleasure for me to undertake this project. I fell highly doing the project entitled – **"HOUSING PRICE PREDICTION"**.

I would like to express my special gratitude to "Flip Robo Technologies" for giving me this opportunity to deal with a beautiful dataset and it has helped me a lot to improve my data analyzation skills. And I want to express my huge gratitude to Ms.Khushboo Garg (SME), who helped me to get out of all the difficulties I faced while doing the project.

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I would like to Thank my parents & Data Trained Team for their kind cooperation and encouragement which helped me in completion of this project.

INTRODUCTION

- ## Business Problem Framing

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.

- ## Conceptual Background of the Domain Problem

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 data is provided in the CSV file below.

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?

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

Developing an accurate prediction model for housing prices is always needed for socio-economic development and well-being of citizens. In this paper, a diverse set of machine learning algorithms such as Gradient Boost, Random Forest and others, are being employed to predict the housing prices.

The housing price prediction models using machine learning techniques are developed and their regression model performances are compared. Finally, an improved housing price prediction model for assisting the housing market is proposed.

- Data Sources and their formats

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 data is provided in the CSV file.

Data contains 1460 entries each having 81 variables.

- Data contains Null values.
- Data contains numerical as well as categorical variable.
- Two datasets are being provided to you i.e test.csv & train.csv

- Data Preprocessing
  - Initially I've imported all the necessary libraries.
  - Then Test and Train datasets were imported. Then I checked the shape, nunique, value counts, info etc.
  - After that I've checked for null values few records were accommodated with null values.Those values are removed from the dataset in the preparation stage.
  - In data analytics, as part of the data cleaning, outlier detection and removal help in making the training model

more stable.Outliers may make the model unstable and result in increase in variance.

- Percentile method is used for the outlier removal process. Data points having z-score value between -3 and +3 are considered.

- Hardware and Software Requirements

  Hardware Required:
  Processor – i5 or above.
  RAM – Min. 8gb
  Software Required:
  Anaconda – Jupyter Notebook

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  • Imputation Methods were used for null values and Percentile method was used for removing outliers.
  • Categorical columns were encoded using ordinal encoding.
  • Skewness was removed using power transformer..
  • Encoded the object type data into numerical using Ordinal Encoder.
  • Machine learning Algorithms were used to predict the sale price.
  .

  Testing of Identified Approaches (Algorithms)

  Salesprice is our Target variable,since it is a continuous type we've used following algorithms;

  - Random Forest Regressor
  - Decision Tree Regressor
  - Extra Trees Regressor.
  - Gradient Boosting Regressor

- Run and Evaluate selected models

Random Forest Regressor:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

```
RFR=RandomForestRegressor()
RFR.fit(x_train,y_train)
pred=RFR.predict(x_test)
acc=r2_score(y_test,pred)
print('Accuracy_score is ',acc)
print("MAE:",mean_absolute_error(y_test,pred))
print("MSE:",mean_squared_error(y_test,pred))
print("RMSE:",np.sqrt(mean_squared_error(y_test, pred)))
```

```
Accuracy_score is  0.8743722296377414
MAE: 17070.129914529916
MSE: 694461285.6573948
RMSE: 26352.6333723481
```

Random Forest Regressor was performed good with 87% accuracy

Decision Tree Regressor:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

```
DTR=DecisionTreeRegressor()
DTR.fit(x_train,y_train)
pred=DTR.predict(x_test)
acc=r2_score(y_test,pred)
print('Accuracy_score is ',acc)
print("MAE:",mean_absolute_error(y_test,pred))
print("MSE:",mean_squared_error(y_test,pred))
print("RMSE:",np.sqrt(mean_squared_error(y_test, pred)))
```

```
Accuracy_score is  0.6210812323153414
MAE: 27048.252136752137
MSE: 2094635714.7564104
RMSE: 45767.190374288984
```

Decision Tree Regressor Accuracy Score is not so good,so let's see the other algorithms.

Extra Trees Regressor:

The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset.

```
ETR=ExtraTreesRegressor()
ETR.fit(x_train,y_train)
pred=ETR.predict(x_test)
acc=r2_score(y_test,pred)
print('Accuracy_score is ',acc)
print("MAE:",mean_absolute_error(y_test,pred))
print("MSE:",mean_squared_error(y_test,pred))
print("RMSE:",np.sqrt(mean_squared_error(y_test, pred)))
```

```
Accuracy_score is  0.870980649292878
MAE: 17630.241709401707
MSE: 713209698.0499127
RMSE: 26705.98618381116
```

Extra Trees Regressor is also performing good with 87% accuracy score, so let's check with other algorithms.

Gradient Boosting Regressor:

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value.

```
GBR=GradientBoostingRegressor()
GBR.fit(x_train,y_train)
pred=GBR.predict(x_test)
acc=r2_score(y_test,pred)
print('Accuracy_score is ',acc)
print("MAE:",mean_absolute_error(y_test,pred))
print("MSE:",mean_squared_error(y_test,pred))
print("RMSE:",np.sqrt(mean_squared_error(y_test, pred)))
```

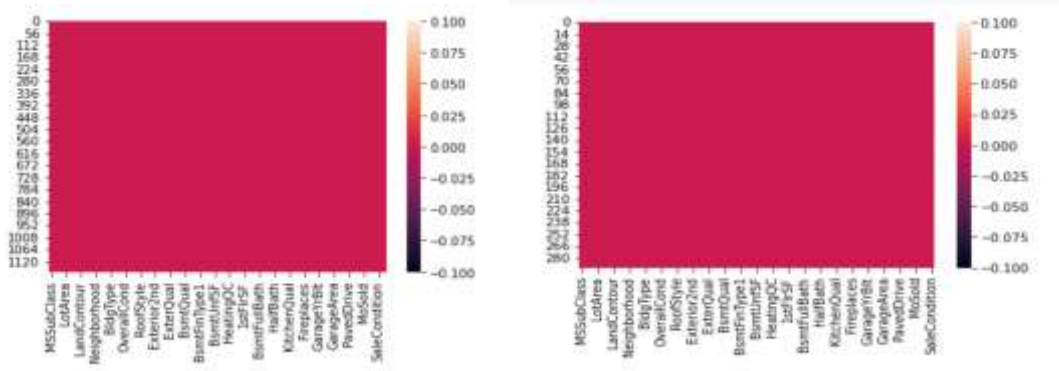
```
Accuracy_score is  0.9040545601857497
MAE: 15891.781137790587
MSE: 530379495.66591793
RMSE: 23029.9695107466
```

Gradient Boosting Regressor is performing great with 90% accuracy score. So, let's proceed with this algorithm.

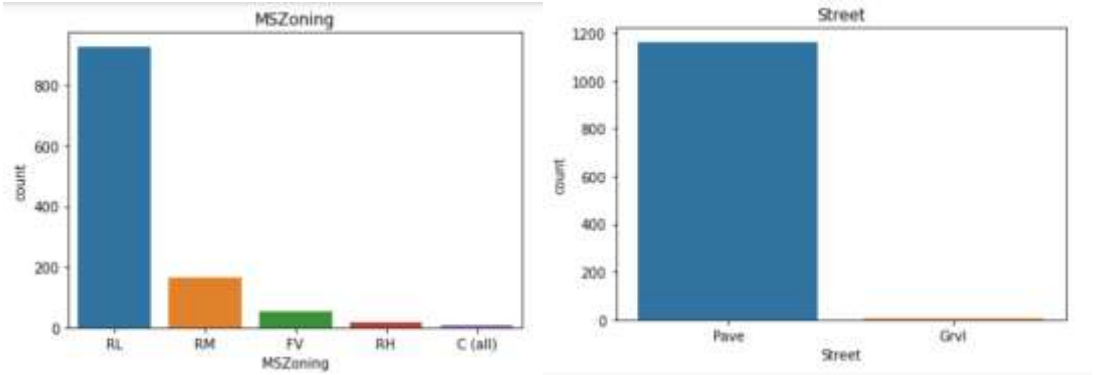
- Key Metrics for success in solving problem under consideration

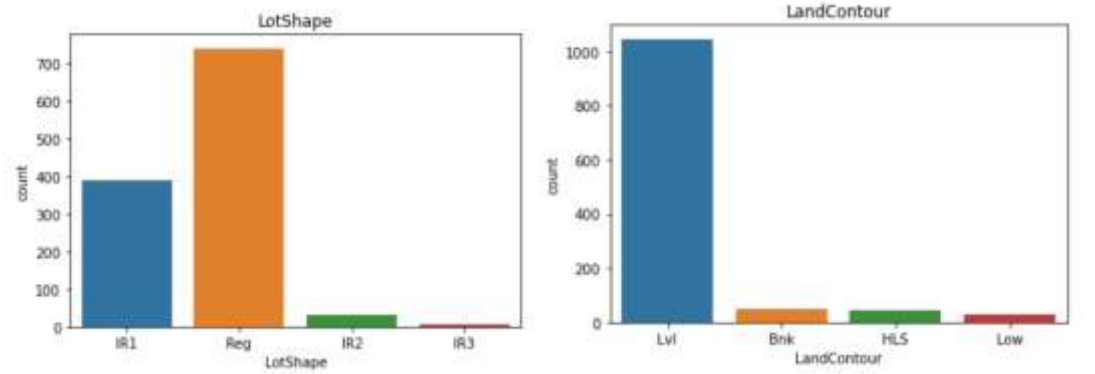
  In this ML project we've used Mean Absolute Error(MAE),Mean Squared Error(MSE) and Root Mean Squared Error(RMSE)

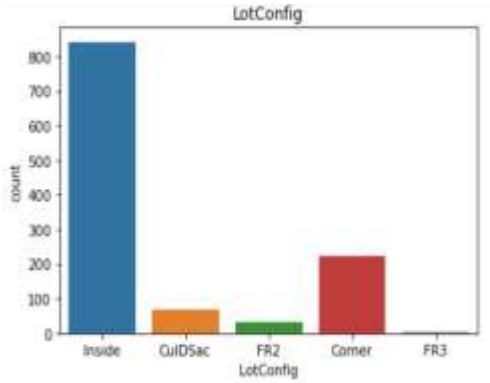
- Visualizations

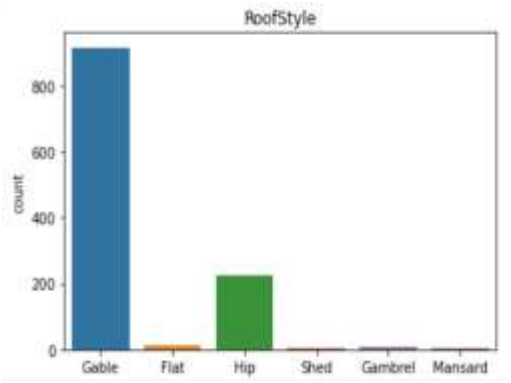


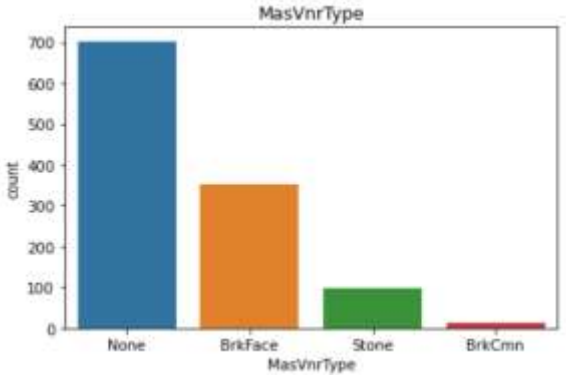
After the data cleaning process,the dataset is checked for null values using heat map.As we can observe that our data is clear from null values.

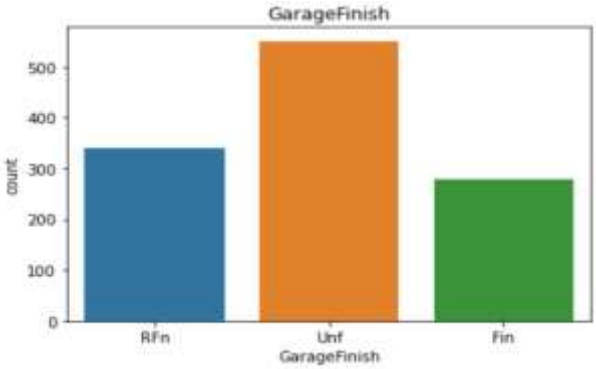


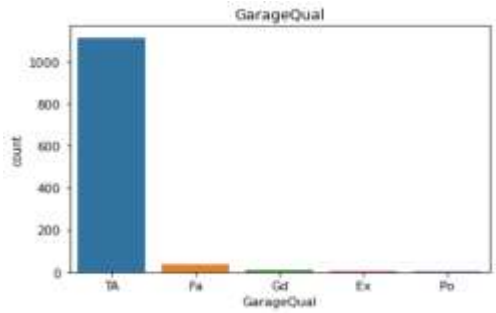


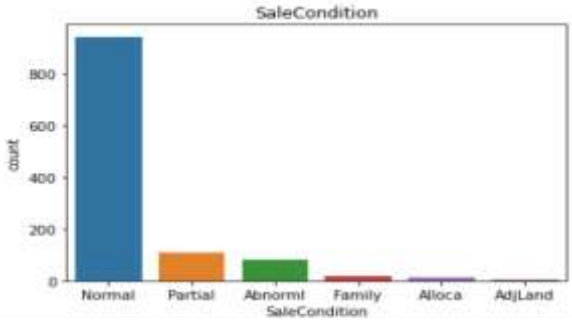




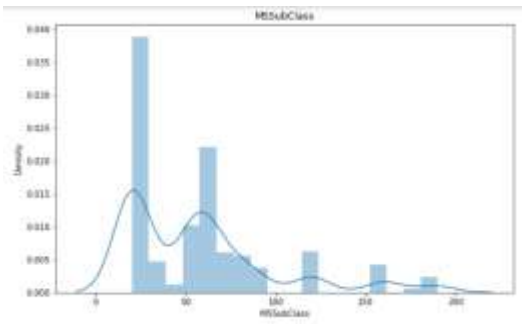


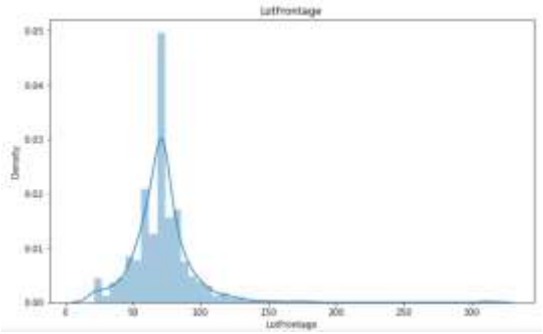


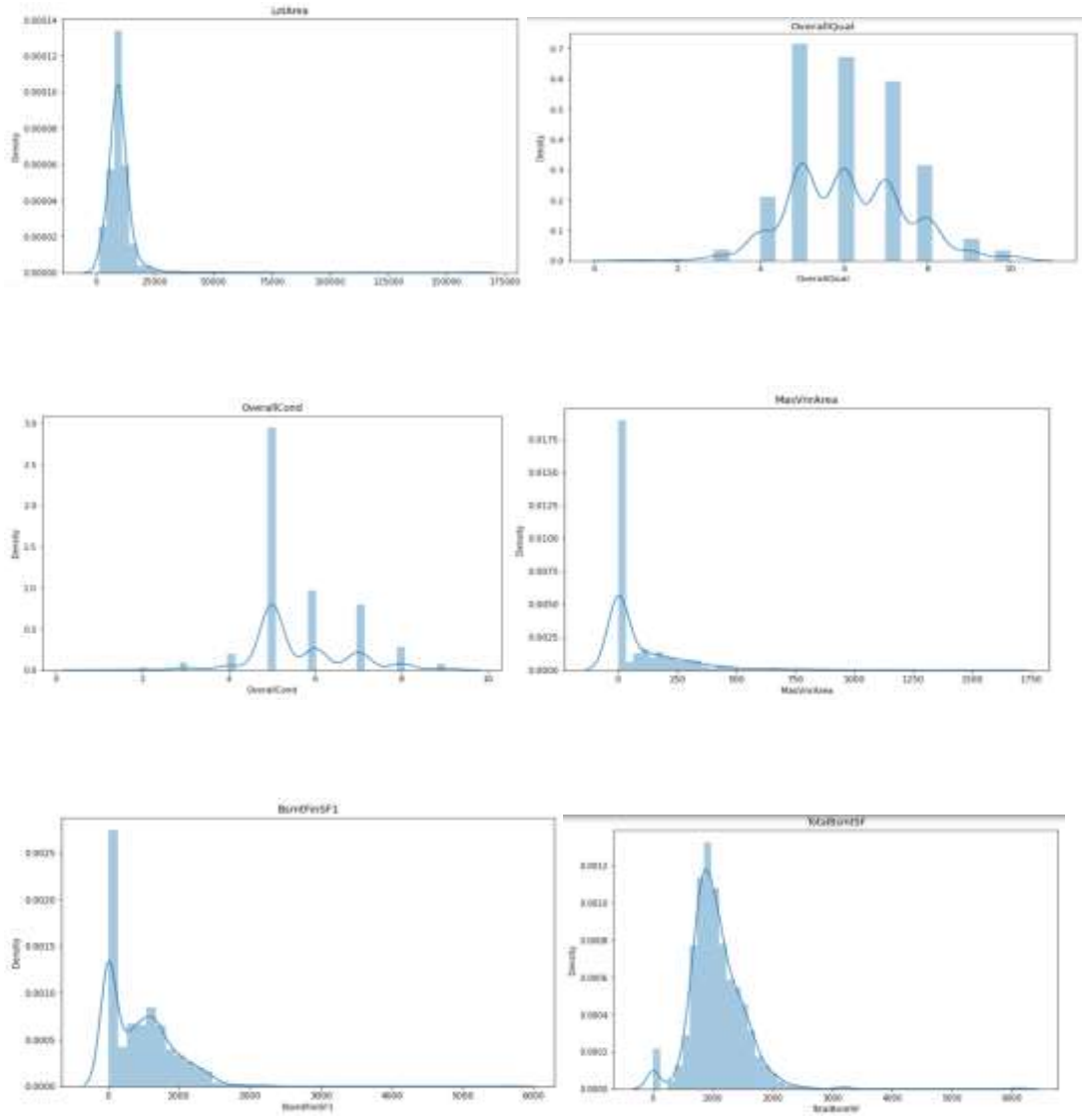




The ordinal columns were visualised using countplot







- Interpretation of the Results
  - From data visualisations,we came to know that the data is distributed Normally in some columns and abnormal in some columns.
  - The Null values in the dataset are identified using Heatmap.
  - The Nominal columns were in linear relationship with target column.
  - There were outliers and skewness in the data,the outliers were represented using box plot.
  - After removing outliers,the data has scaled using standard scaler.

- Then based on accuracy, Gradient Boosting Regressor is considered as best model with accuracy score around 90% among other algorithms.

CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

  Dataset has many null values which was solved using imputation Techniques. For better understanding, several types of plotting was used for visualisation. Variance Inflation Factor was checked and unwanted columns were dropped.Finally Several Machine Learning Algorithms were used for getting best Accuracy,finally Gradient Boosting has great accuracy i.e 90%.

Limitations of this work

- The dataset has much outliers which reduces model's accuracy.
- Due to data leakage concatenation of train and test datasets are not recommended.
- The dataset doesn't contain many details which can used for detail investigation of model.