HOUSING PROJECT

Submitted by:

HARI KRISHNAN J

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.

INTRODUCTION

Business Problem Framing

- 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.

Conceptual Background of the Domain Problem

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.

Motivation for the Problem Undertaken

To Build a model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

- The data set consist of Train data and Test data.
- Train dataset has 1168 records with 81 features.
- Test dataset has 292 records.

Data Sources and their formats

- Dataset has all type data like object, int and float.

Data Preprocessing Done

- Categorical variables are encoded using ORDINAL ENCODER
- Null values haves have been treated with mode method.

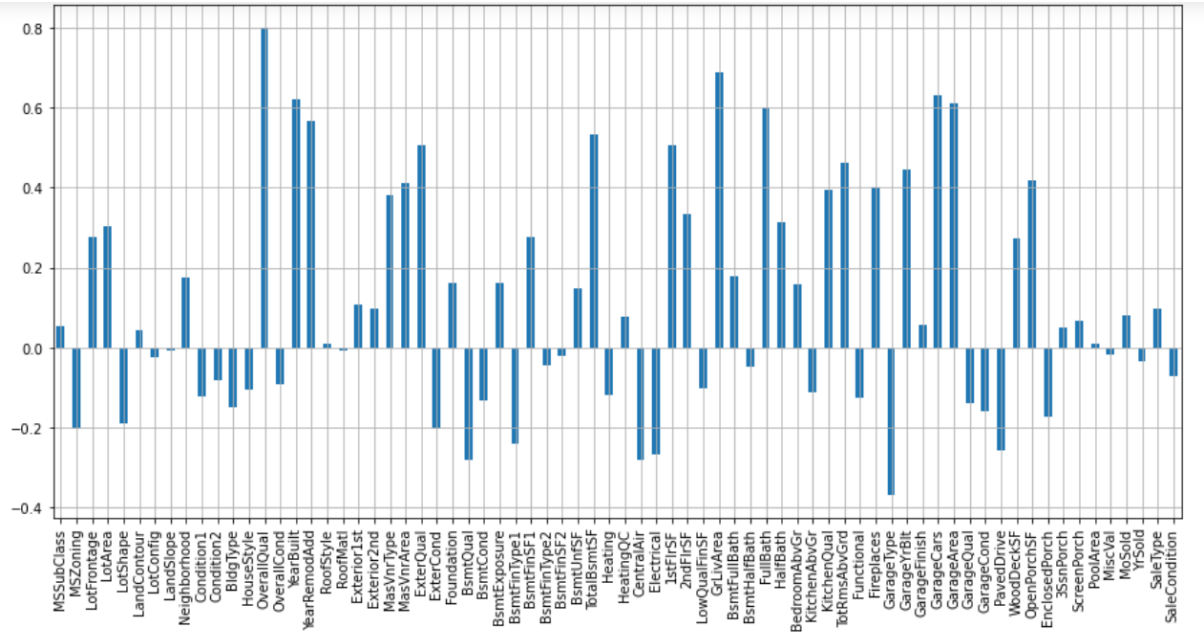
Data Inputs- Logic- Output Relationships

- Some features does not have signification contribution to target label. These variables have all unique values or above 50% records have null values. So, dropping these labels as below:
  - 'Alley'
  - 'PoolQC'
  - 'Fence'
  - 'FireplaceQu'
  - 'MiscFeature'
  - 'Id'
  - 'Utilities'
  - 'Street'
- Null values haves have been treated with mode method.
- The Features have high correlation with Target Label (SalePrice)

- Below mentioned features are some variables that have high correlation.
    - OverallQual
    - GirlivArea
    - GarageCars
    - GarageArea
    - YearBuild and so on.
- These features are treated as important for model building.

State the set of assumptions (if any) related to the problem under consideration

- Some features have multicollinearity problem that is both varible have some relation with each other and its sufficient to have only one feature from both which has high relation with target label.
- Some features has significantly very low relation with target label. So we will drop those features.
- These decisions are taken based on heat chart and correlation chart.



Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

- For building the model we have splited Train-Test Split\
  - x = data.drop(columns = ['SalePrice'],axis =1)
  - y = data.SalePrice
- Scaler tranformation carried out to break the variance in numerical ranges.
  - from sklearn.preprocessing import StandardScaler
  - scale = StandardScaler()
  - x_scaled = scale.fit_transform(x)
- For building the model, we need to define the best random state. Here, we have used LINEAR REGRESSION for random state selection.

Testing of Identified Approaches (Algorithms)

- We have used many regression models to find the best model.
  - *from sklearn.ensemble import RandomForestRegressor*
  - *from sklearn.tree import DecisionTreeRegressor*
  - *from sklearn.neighbors import KNeighborsRegressor*
  - *from sklearn.ensemble import GradientBoostingRegressor*
- After running various model, **GRADIENT BOOST DECISION TREE** algorithm gives the comparatively best performance as shown below.

```
==================Train Results============
R2_score: 96.5111022306828 %

Mean Squared Error: 104079625.40632255

Mean Absolute Error: 7823.889735713238


==================Test Results============
R2_score: 89.41275628175894 %

Mean Squared Error: 318207888.6709046

Mean Absolute Error: 13106.578130256388
```

Run and Evaluate selected models

- All the algorithms have been cross validated using K-Fold method having r2 as score.
    - *print(cross_val_score(gbr, x_train, y_train, scoring='r2', cv=5))*
    - *print(cross_val_score(gbr, x_train, y_train, scoring='r2', cv=5).mean())*
- We have Cross Validation score as 86.8% for Gradient Boosting DT algorithm.

Key Metrics for success in solving problem under consideration

- R2 Score, Mean Square Error and Mean Absolute Error had been treated as success key metric.

Interpretation of the Results

- Hyperparameter tuning has been done for selecting the best parameters using GRIDSEARCHCV method.
- We have attained optimal performance after hyperparameter tuning as shown below

```
==================Train Results============
R2_score: 95.08231341551196 %

Mean Squared Error: 146702773.03062946

Mean Absolute Error: 8899.32897187954


==================Test Results============
R2_score: 86.80286251666666 %

Mean Squared Error: 396650286.5930851

Mean Absolute Error: 14406.42000113272
```

CONCLUSION

Key Findings and Conclusions of the Study

- The final tuned model has been saved and used for predicting test data.
- Same processing has been done as same done with train dataset to set the same result
- Some of the features have no impact on target variable and some feature has high relation with target variable.