PROJECT REPORT

ON

HOUSING:PRICE PREDICTION

Submitted by

ABHISHEK MISHRA

ACKNOWLEDGMENT

As a reference I would like to mention the data source which is provided by the Flip Robo and the dataset which is provided in the form of test and train csv file. I would like to thanks mentors and my SME who support and guided me to completion of this project and not forget to mention some seniors who helped me on GitHub and linkden for this project.

Date-21-01-2023

INTRODUCTION

- ## Housing: Price Prediction

  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.

- ## Conceptual Background of the Domain Problem

  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.

- ## Review of Literature

  As per the comprehensive summary we can clarify that the given problem statement is related to the Regression problem because the target column is "SalePrice". As per evaluation of the Machine learning Theory if the target column is the continuous and float datatype then the model will be predicted as Linear Regression.

  We are going to predict the house prices through Machine Learning and further we are going to compare the actual price and predicted price and get the best accuracy through model building.

- **Motivation for the Problem Undertaken**

To determine the best price prediction of housing and to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. These are the objective to do this project. The motivation behind this project to have the essential knowledge about Machine Learning linear regression model building and to get the job in ML engineer domain.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

These are some important analytical things on which I have done the analysis and treat the things by using the correct methods.

Non-Linearity:- Non linearity is a statistical term which is used to describe the relation between independent variable and dependent variable.

By using the plots we find out the non-linearity between the features.

Skewness:-When the mean value of the column is greater than its median value then the data is skewed. We need to treat the skewness by using some techniques. Here we use the “Log Transform” method to treat the skewness.

Outliers:- when the data goes across -3 to +3 then those datas or values are called the outliers which affect the model so we need to remove these outliers by some techniques. We use the Z-Transform technique to treat the outliers.

PCA technique:- Done the PCA technique to achieve the 100% variance.

- Data Sources and their formats

We had been provided the datasource by Flip robo in the form of CSV file in which the data was separated into two parts that is test data and train data. So, I performed the concatenation function to add the dataset. In the dataset there is 1460 rows and 81 column in the dataset. These data are collected by the companies in Australia from the market. There are some columns which are categorical in nature while some columns are continuous in nature. Target column is the continuous in nature and having Float data type so we are going to build the model on Linear Regression. According to the summary statics there is some skewness present in some columns and outliers are also present so need the treatment of the dataset. The columns in which their mean value is greater than the median value,must be present the skewness. Meanwhile by the plots we can find out the outliers in the columns. We are using the scatterplot to visualize the outliers of the columns .

- Data Pre-processing Done

To handle the missing values in the dataset I have used the mean value of the rest data present in the column. Detect the skewness in the column by using the function 'df.skew()'.Which column is not useful need to drop that column so drop the 'Id' column. Finally, to remove the skewness I have used the log transformation and the boxcox method. Further visualise the data with boxplot to detect the outliers in the dataset and to treat these outliers from the dataset I am going to use the Z-transform method.

By treating the data with skewness and outliers, it is assumed that my data is now cleaned and ready for the further process.

- Data Inputs- Logic- Output Relationships

Describe the relationship between independent and dependent variables we are using the Bivariate analysis by plotting the different plots like histogram and scatterplots.

Finding the correlation between the two varriables.Plot the heatmap to find out that how much the variables are co related with each other. Remove the multicollinearity if any after this analysis.

- Hardware and Software Requirements and Tools Used

To build this project we need some important libraries which all are Listing down as-

Numpy- Treatment of skewness.

Pandas- To create the data in dataframe.

Seaborn- for data visualization

Scipy- Mathematical computation,

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Simply I followed the basic statistics approaches which I have learnt during the classes (mean,median,mode) and offcourse the percentile method to find out the skewness and outliers.

Analysis of the Problem statement and the visualization of the graphs helps a lot. Linear regression model theory analysis was so important while building the model

- Testing of Identified Approaches (Algorithms)

train_test_split for splitting the dataset

```
# Let's separate the input and output variables of train dataset for scaling

x = df_train.drop("SalePrice", axis=1)
y = df_train["SalePrice"]
```

```
# Let's do the scaling

scaler = StandardScaler()
df_x = scaler.fit_transform(x)
df_x
```

from sklearn. preprocessing import StandardScaler

from sklearn.model_selection

import train_test_split,cross_val_score

from sklearn.metrics import r2_score

from time import time

Algorithms are like:-

DecisionTreeRegresor(DTR)

K-NeighborsRegressor(KNR)

RandomForestRegressor(RFR)

GradientBoostingRegressor(GBR)

SupportVectorMachine(SVR)

```
# Let's import the necessary required libraries for model building

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import r2_score
from time import time
```

These all above algorithms are used for the best value prediction in which I found that GradientBoostingRegressor gives me the best values for prediction so I choose the GBR for Hyperparameter Tunning for more accuracy.

- Run and Evaluate selected models

```
from sklearn.ensemble import RandomForestRegressor

RFR = RandomForestRegressor()
beststate(RFR)
```

```
Best Random State      :  74
Best R2_Score          :  0.8507067334231435
Cross Validation Score :  0.7962935546279176

Time taken by model for prediction 86.6546 seconds
```

```
]:  1  # Gradient Boosting Regressor
    2
    3  from sklearn.ensemble import GradientBoostingRegressor
    4
    5  GBR = GradientBoostingRegressor()
    6  beststate(GBR)
```

```
Best Random State      :  74
Best R2_Score          :  0.8769160345428648
Cross Validation Score :  0.8253255263968985

Time taken by model for prediction 49.8110 seconds
```

```
:   1  # K-Neighbors Regressor
    2
    3  from sklearn.neighbors import KNeighborsRegressor
    4
    5  KNR = KNeighborsRegressor()
    6  beststate(KNR)
```

```
Best Random State      :  74
Best R2_Score          :  0.7839475475533682
Cross Validation Score :  0.7197611913309278

Time taken by model for prediction 0.1239 seconds
```

These are the some Regressor algorithms which all I used to get the best values of Price and I found that GradientBoostingRegressor is giving the best value.

- To find the 100% variance I used the PCA technique

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

- Interpretation of the Results

Final Model (Gradient Boosting Regressor)

Saving The Final Model

[35]:
```
# Hyper Parameter Tuning with Gradient Boosting Regressor

x_train, x_test, y_train, y_test = train_test_split(xx, y, test_size=0.25, random_state=74)

```

[36]:
```
from sklearn.model_selection import GridSearchCV

```

[37]:
```
param_grid = {"min_samples_leaf" : [1,2], "min_samples_split" : [2,3], "n_estimators" : [100,200], "learning_rate" : [0.1,0.
grid_search = GridSearchCV(GBR, param_grid=param_grid)
```

- Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

- Key Findings and Conclusions of the Study

From The above project I have learnt a lot that how to perform well as Data Analyst. LinearRegression model building steps well known and as a conclusion got the best value after hyperparameter tunning.

```
# Final Model

Final_Model = GradientBoostingRegressor(learning_rate=0.1,min_samples_split=2,min_samples_leaf=1,n_estimators=200)
x_train, x_test, y_train, y_test = train_test_split(xx, y, test_size=0.25, random_state=74)
Final_Model.fit(x_train, y_train)
y_pred = Final_Model.predict(x_test)
r2_score(y_test, y_pred)
```

```
0.880398996397358
```

- Learning Outcomes of the Study in respect of Data Science

I had faced some issues during the outliers' treatments and also at hyperparameter tunning. I have used the log transform for skewness and outliers.one more important thing I want to discuss that I was stucked a little while dropping the column that which one I should drop who does not affect the match and our target variable.

I preferred the difficult way since last night on this project and get frusted because I was not able to encode the categorical columns then I saw my previous classes lectures which helped me to overcome these difficulties.

which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.