FLIP ROBO

NAME OF THE PROJECT :-

PFA Housing Project

Submitted by: -

Pranavkumar Ravindrakumar Pandey

ACKNOWLEDGMENT

In this accomplishment of project completion, I would like to express my special gratitude to all my teachers, and most importantly to Udemy courses, without their clear and practical explanations it is not possible to complete this regression PFA Housing Project.

Finally, I would like to thank our main guru Google and friends who helped me in finishing this project.

INTRODUCTION

Business Problem Framing

In this “Housing Price Prediction” project we need to build a machine learning regression model which can predict the house prices accurately.

We have many independent/input attributes present in given dataset from which we need to find the best attributes which effect the property prices and nature of those attributes on property price.

This project will help the companies in real estate investments in the Australian market, since it will predict the price of the houses located in any particular area with the help of which investor can buy the houses in good prices.

Conceptual Background of the Domain Problem

As this project is about the price prediction of houses located in a city of Australia, so knowledge of housing or real estate plays important role in this project while choosing the main attributes for the price prediction. Like we all know that housing prices are mainly dependent on the locality in which house is located, various facilities like road, electricity supply, transportation connectivity also plays major role in the housing prices.

Review of Literature

For this PFA Housing project I have took guidance many blogs written by various data scientists/Machine learning engineers on websites like Analytics Vidhya, medium.com etc.

Motivation for the Problem Undertaken

The main objective of this project is to build a good machine learning model which can predict the housing prices accurately, by which I want to contribute my ml model to our real estate industry.

The main reason to build this project is to understand the main aspects of the real estate industry and I wanted to apply my data science knowledge to this real estate domain.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

In this project we have many independent variables or features (predictors) with the help of which we want to predict our target feature Sales Price using different machine learning algorithms.

Data Sources and their formats

Dataset is provided by Surprise Housing Company.Dataset is in structured/tabular format.

Data Pre-processing Done

- We have removed those columns which have more than 60 % of null values in them.
- Column which has few null values, we replace those null values with suitable mean/median if column is numerical and with mode if column is categorical.
- If the numerical column has some kind of skewness, we have replaced the null values with the median of that column since median is less prone to the outliers.
- If the numerical column has almost no skewness or very less skewness in that case, we may replace null values of those columns with the mean value of that column.

- We have created the dummy columns for the nominal categorical features.
- Deleted one of the columns from each dummy columns to cope up with multicollinearity issue.

- We have checked for the outliers present in the numerical columns using interquartile range.
- Splitted the data into independent feature and dependent feature for model training.
- After splitting the data, we formed the standard scaling on the dataset so that our machine learning model will treat each features/independent variable equally.

Data Inputs- Logic- Output Relationships

In this project we have many input/independent features some of them have numerical datatype, some of them have categorical datatype. By means of dummy encoding we converted all categorical datatype into binary 0/1 datatypes. Later we fed this independent feature to our machine learning model to predict our target variable “Sales Price”.

Hardware and Software Requirements and Tools Used

We have used jupyter notebook as an environment of model building.

Programming language python 3 is used in this project,along with that we have used many python packages also which are listed below

a. Numpy→Package for scientific computing

b. Pandas→Package for data manipulation and analysis.

c. Matplotlib→Visualisation package

d. Seaborn→Visualisation package

e. Sklearn →software machine learning package

Model Development and Evaluation

Testing of Identified Approaches (Algorithms)

- *Linear Regression Model*
- *AdaBoost Regressor Model*
- *Gradient Boosting Regressor Model*
- *Support Vector Regressor Model*
- *Random Forest Regressor Model*
- *Decision Tree Regressor Model*
- *KNearest Neighbour Regression Model*
- *Ridge Regressor Model*
- *Lasso Regressor Model*

Run and evaluate selected models

1. ***Linear Regression Model:-***

   R2_score of LinearRegressor model is : -5.479463540787324e+17

   Adjusted_r2_score of LinearRegressor model is : -6.491912641724677e+17

   Mean_squared_error of LinearRegressor model is : 3.567348025097778e+27

2. ***AdaBoost Regressor Model :-***

   R2_score of AdaBoost Regressor model is : 0.7359

Adjusted_r2_score of AdaBoost Regressor model is : 0.6872

Mean_squared_error of AdaBoost Regressor model is : 1718852746.5531664

3. ***Gradient Boosting Regressor Model :-***

R2_score of GradientBoosting Regressor model is : 0.8743

Adjusted_r2_score of GradientBoosting Regressor model is : 0.8516

Mean_squared_error of GradientBoosting Regressor model is : 815040721.80

4. ***Support Vector Regressor Model :-***

R2_score of Support Vector Regressor model is : -0.0556

Adjusted_r2_score of Support Vector Regressor model is : -0.2506

Mean_squared_error of Support Vector Regressor model is :- 6872513508.865689

5. ***Random Forest Regressor Model :-***

R2_score of RandomForestRegressor model is : 0.8189

Adjusted_r2_score of RandomForestRegressor model is : 0.7854

Mean_squared_error of RandomForestRegressor model is: 1178761345.8827558

6. ***Decision Tree Regressor Model :-***

R2_score of DecisionTreeRegressor model is : 0.7197

Adjusted_r2_score of DecisionTreeRegressor model is : 0.6679

Mean_squared_error of DecisionTreeRegressor model is : 1824560965

7. ***KNearest Neighbour Regression Model:-***

R2_score OF KNN regressor model is : 0.6922

Adjusted_r2_score of KNN regressor model is : 0.6353

Mean_squared_error of KNN regressor model is : 2003840516.85641

8. ***Ridge Regressor Model :-***

   R2_score of RidgeRegressor model is : 0.6533

   Adjusted_r2_score of RidgeRegressor model is : 0.5892

   Mean_squared_error of RidgeRegressor model is : 2257085682.643661

9. ***Lasso Regressor Model :-***

   R2_score of LassoRegressor model is : 0.6533

   Adjusted_r2_score of LassoRegressor model is : 0.5892

   Mean_squared_error of LassoRegressor model is: 2257122179.3836536

From all of models' performance mentioned above we can clearly see that gradient boosting regressor model is performing very good.

After gradient boosting regressor model, Random Forest regressor model is also a great choice.

Key Metrics for success in solving problem under consideration

We are using adjusted r2 score as key metric for the model building because it is more reliable than r2 score.

Since it has penalising factor for irrelevant independent features while r2 score increases when we add an independent features in model whether it is relevant or irrelevant.

Visualizations

We have used histogram plots for the checking the skewness of the continuous variable columns in which we have found that few columns are heavily skewed, few are lightly skewed and rest are approximately normally distributed.

We have used scatterplots for checking the correlation between the sale price and each independent variable.

We have used box-plots for visualising the outliers present in any continuous variable column.

Interpretation of the Results

After hyperparameter tuning of gradient boosting regressor model we found that the following results.

R2 score of tuned gradient boosting regressor model is 0.884.

Adjusted R2 score of tuned gradient boosting regressor model i s 0.863

CONCLUSION

Key Findings and Conclusions of the Study

OverallQual, GrLivArea, GarageArea, TotRmsAbvGrd, TotalBsmtSF,1stFlrSF, YearBuilt, these are the most important features for predicting the sales price of the houses.

We should not remove the outliers present in above columns since they have some real and important information about the prices.

Learning Outcomes of the Study in respect of Data Science

Gradient boosting algorithm is working best on our dataset and it is giving an adjusted r2 score of 0.8636.

Limitations of this work and Scope for Future Work

While working on the project I have deleted/neglected few columns for the sake of simple model building, if we consider those columns too then we may need little more computational power but it will surely increase the adjusted r2 score of the model.

One more thing is that we have not considered all the parameters available for the hyperparameter tuning, if we consider those parameters also then we may get better result but for that we need more computational power and the computational time.