Flight Price Prediction

Rushil Ashish Shah
B20AI036

Abstract – This paper reports our experience with building a Regressor model that predicts flight prices. We are given a dataset which consists of various feature vectors related to flight prices from a number of different cities. We are supposed to apply various regression

I. Introduction

With the increase in the number of people traveling by flights, it has become difficult for airlines to maintain the price of airlines as prices change dynamically. The goal of this project is to predict the price of the flight. However, there is a variety in the shape, design and appearance of glasses which makes it challenging. This flight ticket price predictor can also help customers predict the future flight prices and plan their journey accordingly

Datasets

The file dataset.xlsx is used as the main dataset.
The train dataset contains 10683 rows and 11 columns
The columns contain information about Date of journey,Source,Destination ,Route,Departure and arrival time, Duration,Total Stops, Additional Info, Price etc

II. Methodology

Overview

There are various classification algorithms present out of which we shall implement the following

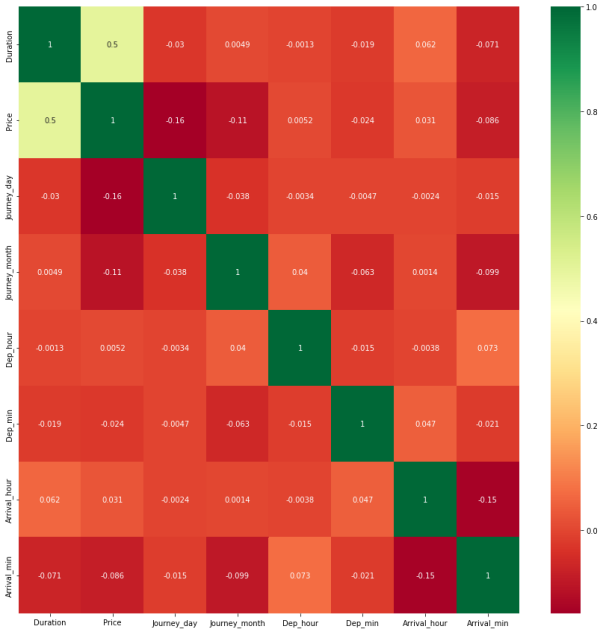
- *Ridge Regression*
- *Lasso Regression*
- *Decision Tree Regressor*
- *Random Forest*
- *XGBoost*
- *LightGBM*
- *Gradient Boosting*

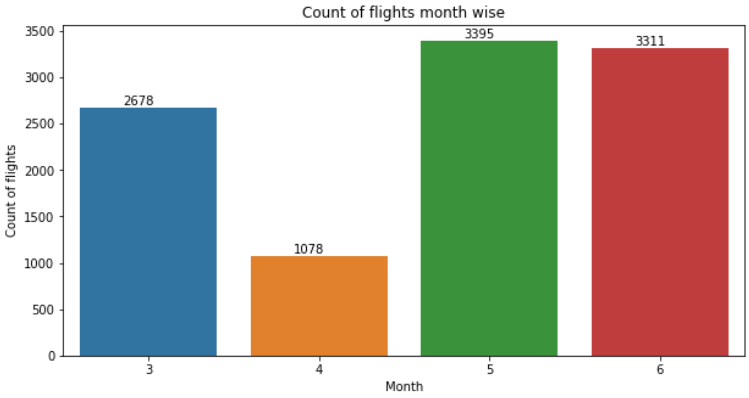
Exploring the dataset and pre-processing

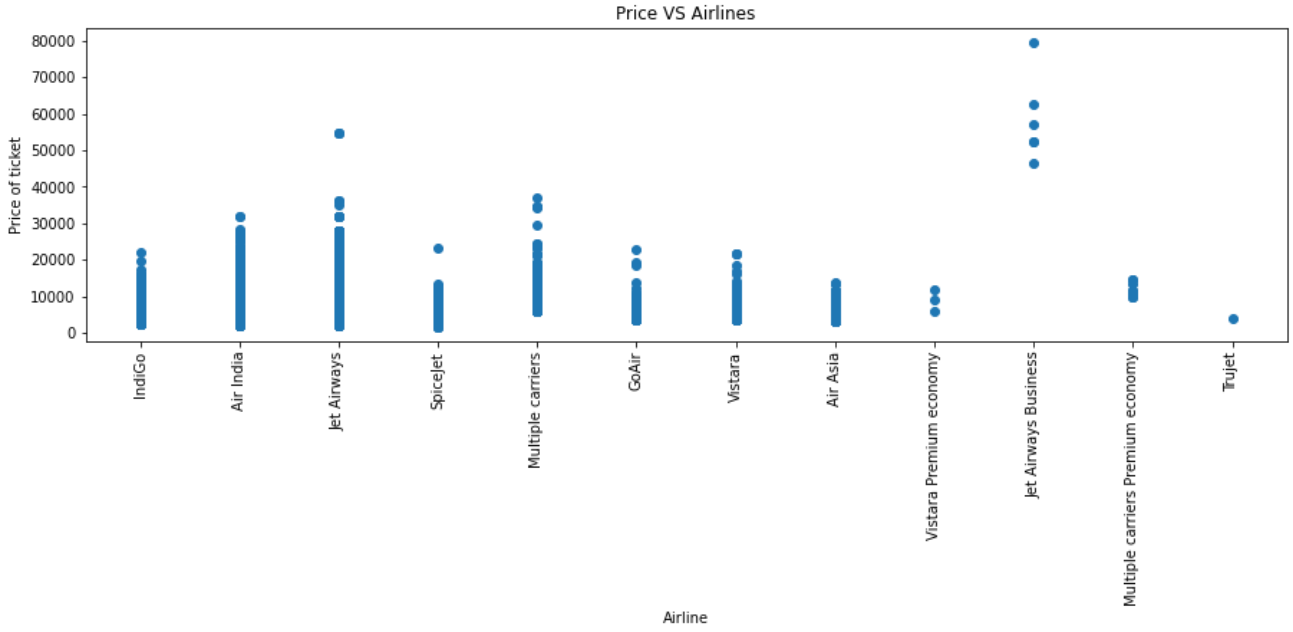
We counted the number of Null values present in the dataset and deleted the rows containing them. I plotted the category plots , converted the duration to int. I also dropped the arrival and departure time of the flights. I did label encoding for the columns containing str datatype. I also counted the number of flights monthwise and plotted the heatmap to correlation of the features.

I also split the dataset into X,y and then train test dataset separately

Visualization







Implementation of classification algorithms

I implemented all the algorithms mentioned above using the inbuilt sklearn library.

- *Ridge regressor :* Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.
- *Lasso Regression : lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable.*

- *Decision tree regression :*Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- *Random Forest Regression :*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.
- *XGBoost :*The most common loss functions in XGBoost for regression problems is reg:linear , and that for binary classification is reg:logistics . Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoost is one of the ensemble learning methods.
- *Lightgbm regressor :*LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage.

III. Improving the models by HyperParameter tuning

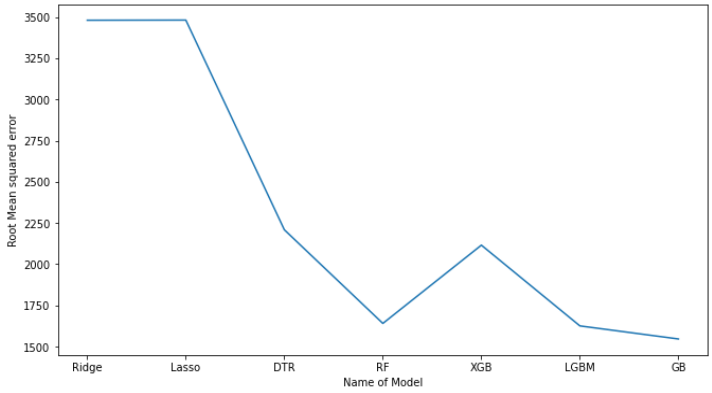
I have used the RandomizedsearchCV To find out the best Hyperparameters for all the models used. I defined the parameter range for all the models and then randomizedsearchcv trained the model through all possible combinations of the parameters to find out the best parameters. I have printed the r2_score , rmse and mean absolute error for all the best models of each regressor.

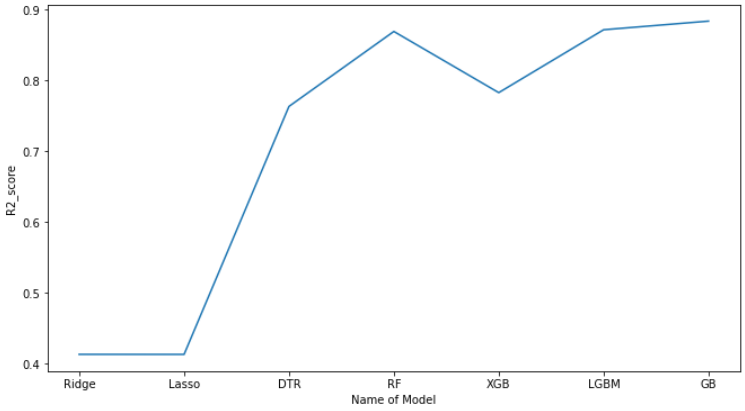
IV.Evaluation of models

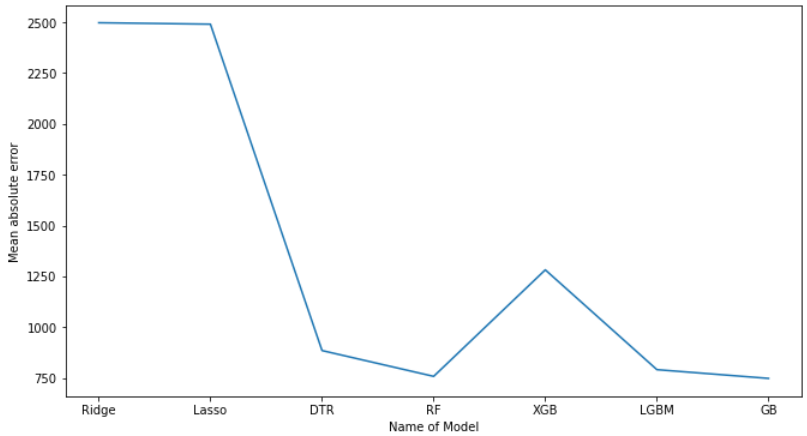
I have used three metrics to evaluate the performance of the models that I have used. These three metrics involve Root Mean squared error, Mean absolute error and R- Squared Error

Model Name	Root Mean squared error	Mean Absolute Error	R-Squared
Ridge Regression	3480.035	2498	0.413
Lasso Regression	3480.907	2491	0.41293
Decision Tree Regressor	2208.94	885	0.7635
Random Forest	1641.023	758	0.8695
XGBoost	2115.711	1282	0.783
LightGBM	1625.96	791	0.8719
Gradient Boosting	1546.63	748	0.88410

Plots







VI. Results and Analysis

The Table shows how all regressors have performed on different metrics. It is clearly visible from the table that Gradient Boosting and LightGBM have performed exceptionally well. It can also be seen from the plot that mean absolute error and the root mean squared error is higher for models like ridge lasso and is lower for models like decision tree, random forest , xgboost and lightgbm. The same can also be seen from their r2_scores. Techniques for dimensionality reduction were not used since the number of dimensions were already so few.Lasso gave the least r2_score.This might show that its model was incomplete. The highest r2_score obtained for Gradient Boosting is equal to 0.8841.

References

[1]Gradient Boosting (sklearn website)
[2] Lasso Regression(sklearn website)
[3] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda
[4] Ticket price prediction (Kaggle)