PRML Bonus Project Report

Palaskar Adarsh Mahesh, *B20EE087*

Abstract

This paper reports my experience and all the challenges faced and resolved while building a flight ticket price predictor. The flight ticket dataset of year 2019 was used to predict their prices using various regressor models and they were compared with respect to various aspects to find out the best model for the task.

I. Introduction

THE goal of this project is to predict flight ticket prices by considering a number of factors that can potentially affect the daily variation in their prices. Since there are a lot of factors that play a major role in manipulating them, the prediction becomes challenging. This project can potentially help individuals, travel companies to reduce their commute costs and even the airline companies for setting competitive prices.

II. Methodology

A. End-to-end Machine Learning pipeline :

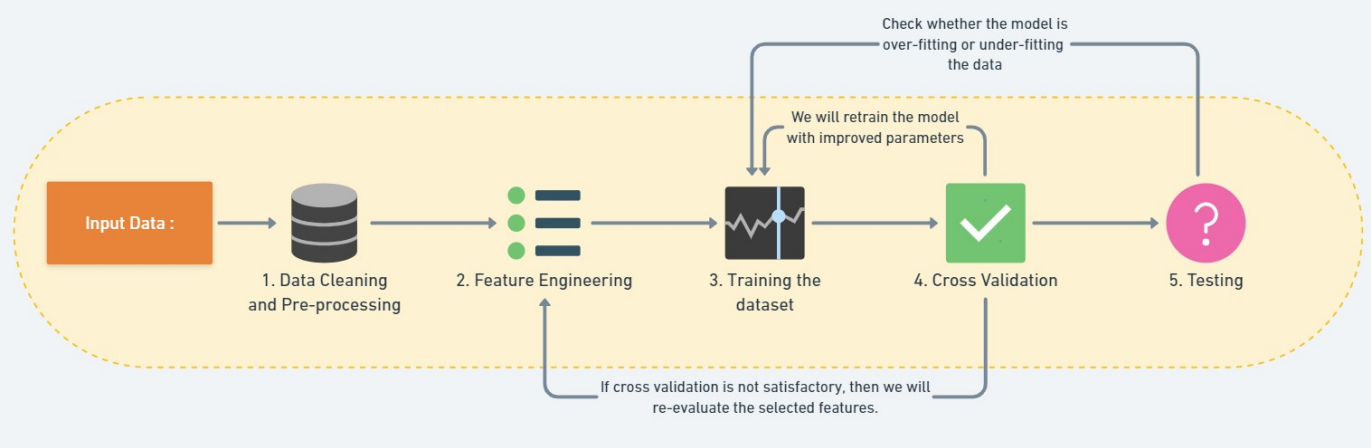

Fig. 1. End to End Pipeline (Made on whimsical.com)

B. Data Cleaning and Pre-processing

Checked for null values and only 2 such rows with incomplete information were found and thus removed from the data. Duplicate values were removed from the data. Spelling and capitalization errors in the data were rectified by converting into uniform values across the dataset.Then extracted day, month year of journey from the date attribute and created 3 different attributes.

C. Feature Engineering

Since flight tickets are generally expensive during weekends and also during holiday periods when large amount of people travel, we have created two attributes which determine the weekday of travel and whether it was a holiday or not. The variation observed is as follows:

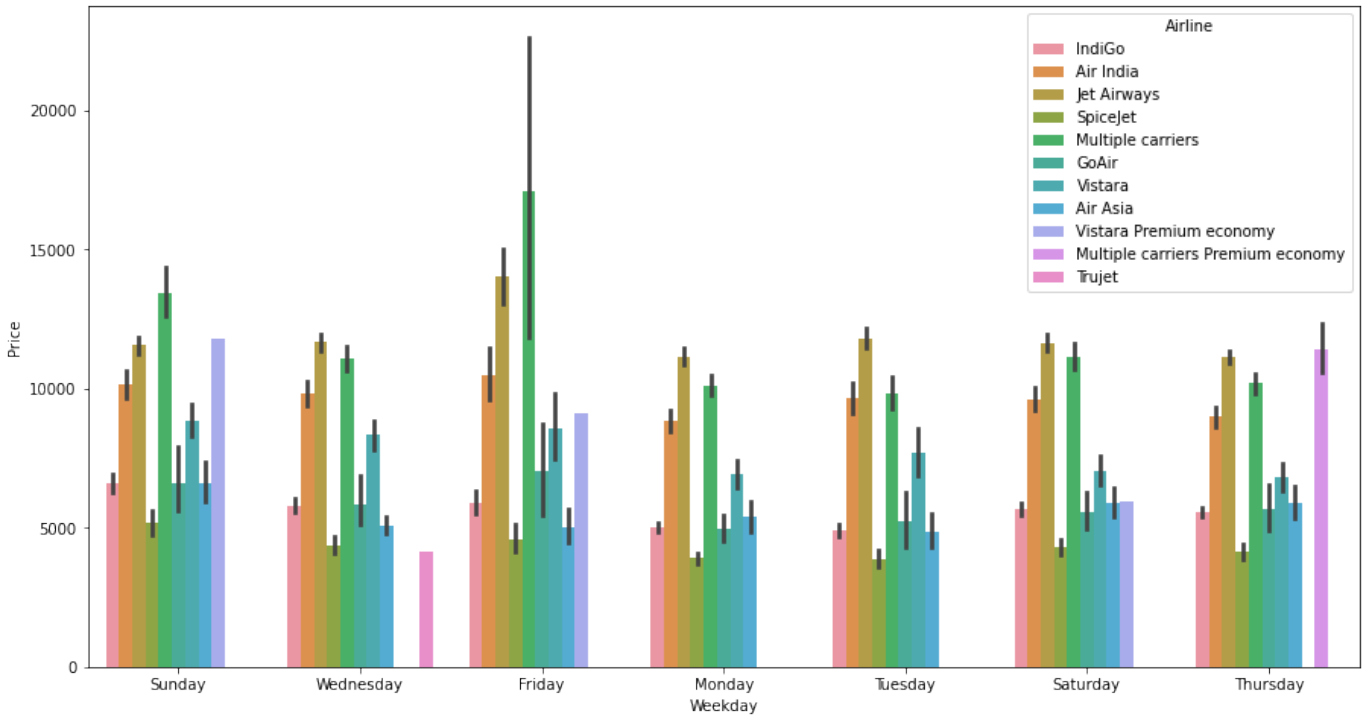

Fig. 2. Variation of Prices with weekdays

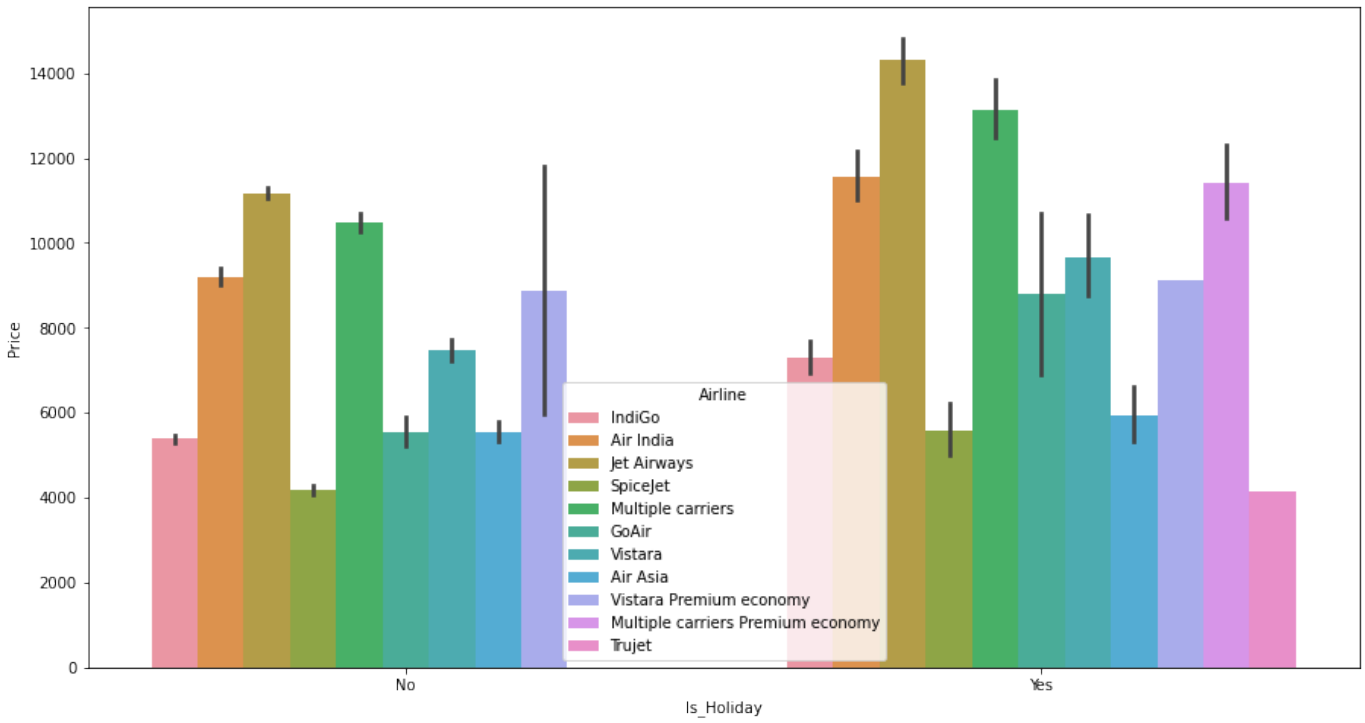

Fig. 3. Variation of prices with holidays

Flight ticket prices also vary on the time of the day the flight departs and lands, and thus we have made two more attributes where the daytime is divided into four parts for both departure and arrival timings. Prices also vary with respect to the number of stops and the route of the flight. All these assumptions are backed by the data and similar visualizations like the above graphs are made in the notebook to prove the same.

All the object attributes of ordinal type were encoded with one-hot- encoding and the redundant features were removed from the dataframe.

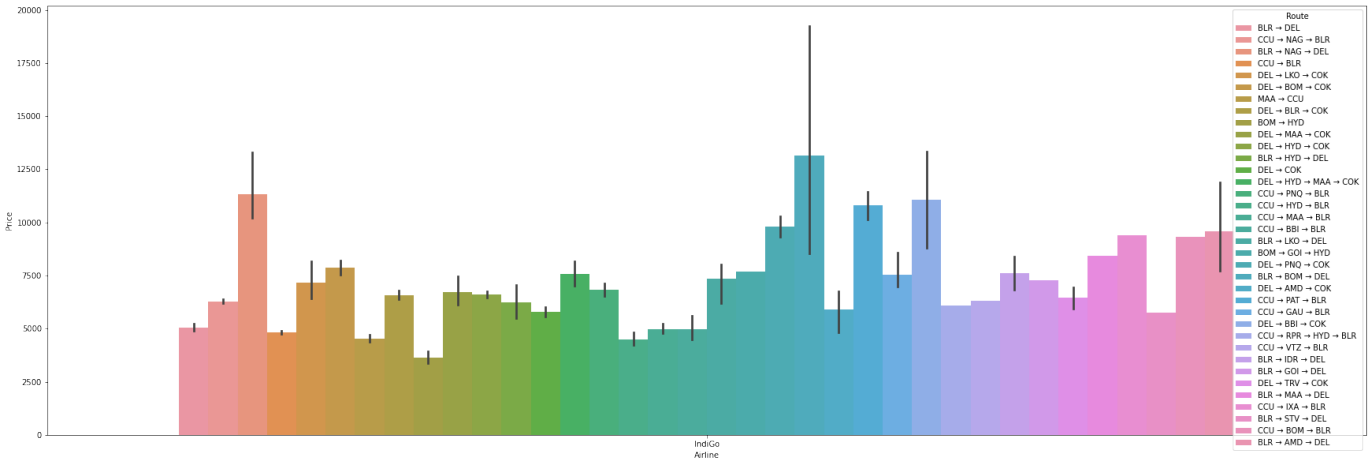

Fig. 4. Variation of prices with different routes

D. Training the dataset, Cross validation and Testing

We trained the dataset using 5 different models and used Grid Search CV as well as Random Search CV and tuned the hyper-parameters accordingly. The results obtained for test are as follows:

- Decision Tree Regressor :

  RMSE : 1958.8101361168476
  R2 Score : 0.8234558980638369

- Gradient Boosting Regressor :

  RMSE : 1428.1363429448627
  R2 Score : 0.906155706309871

- Random Forest Regressor :

  RMSE : 1553.5036870743281
  R2 Score : 0.8889565095976801

- AdaBoost Regressor :

  RMSE : 3093.6299127424595
  R2 Score : 0.5596426552288125

- XGBoost Regressor :

  RMSE : 1488.646374459262
  R2 Score : 0.8980348844306114

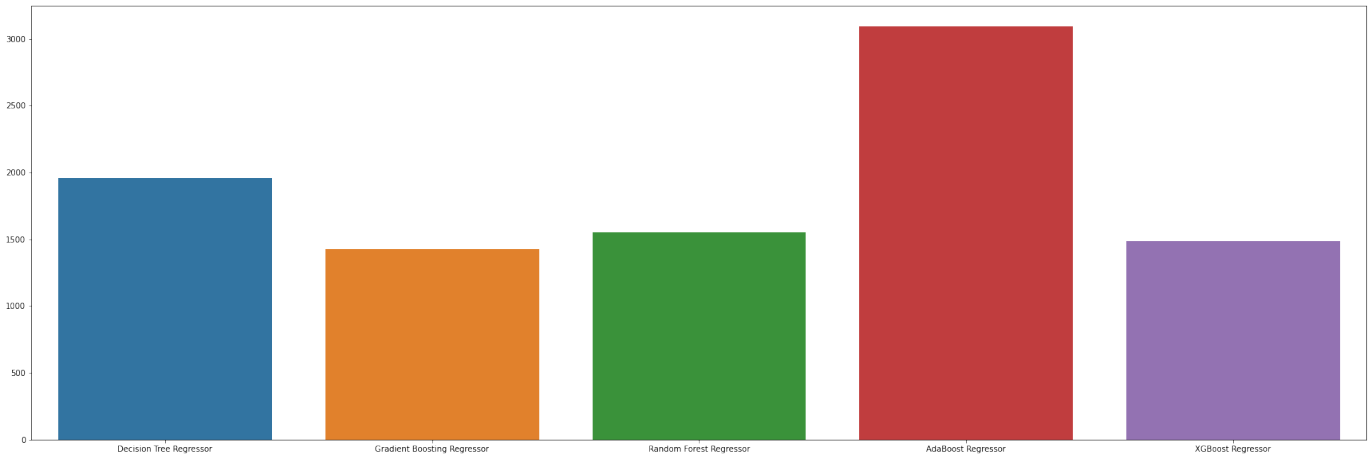

Fig. 5. RMSE scores for different models

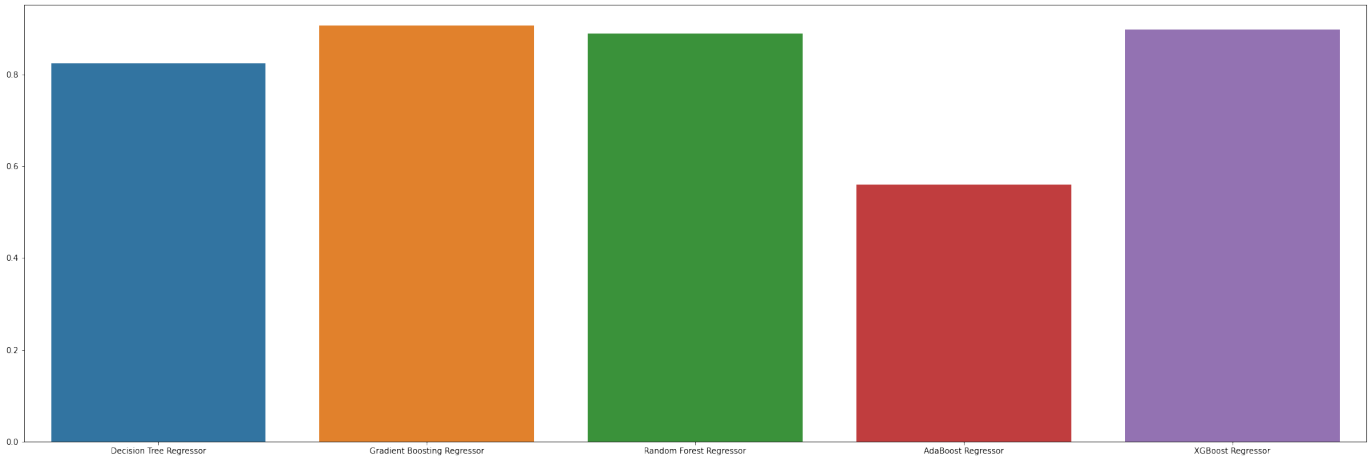

Fig. 6. R2 Scores for different models

III. Conclusion

We observe that AdaBoost Regressor has the worst performance, whereas XGBoost Regressor and Gradient Boost Regressor perform very well, the latter just edging out to be the best performing model by a small margin.

References

[1] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda
[2] Deployed Project - Achyut Joshi
[3] stackoverflow.com
[4] whimsical.com