FLIGHT PRICE PREDICTION PROJECT

A project report submitted in

partial fulfilment of the requirement for the award of internship provided by

FlipRobo Technologies, Bangalore

Submitted by:

ELTON GRIVITH D SOUZA

engardaelton.work@gmail.com

NOVEMBER 2021

ACKNOWLEDGMENT

I would like to offer my sincere thanks to **FlipRobo Technologies** for providing the opportunity to intern under their company.

I would like to thank **Ms Khusboo Garg** for coordinating and providing valuable suggestions during the process of this project and internship.

Finally, I would like to express my sincere gratitude to my family and friends who have been the embodiment of love and support which helped us carry out this project in a smooth and successful way.

INTRODUCTION

Problem Statement

Flight prices can vary often and is difficult to choose the best flight ticket from the cheapest available option. Therefore this project requires the that this scenario is modelled appropriately to predict the best flight ticket with the cheapest fare and gain additional insights on what really matters.

Data Sources and their formats

The data was sourced from Kaggle. The Dataset consists of more than 10,500 tuples containing 11 features. While there are null values present, it is not to the point wherein it becomes an issue.

Data Pre-processing

The features were first correctly formatted to provide proper input to the model. These features were mainly date and time.

The route feature was converted into number of stops. This reduced unnecessary sparse information to meaningful insight.

VIF was used to check for multi-collinearity. The threshold was taken to be 5. The results acquired, showed that the data showed no signs of the said issue.

The data also consisted of some unrealistic values. But this was ignored to avoid overfitting our model. But we reduced its influence by fixing the skew of our data.

Outliers were removed using Z-Score Analysis. It reduced out data by a bit more than 200 tuples. So data loss did not become an issue in this step.

Skew was treated by using scikits' power transform method. yeo-johnson algorithm used for the treatment. This process was followed by normalisation to make the distribution more Gaussian.

The data was also numerically encoded and scaled using standard scaler to make the data easier to work with and also to accelerate the computations.

Data Relationships

The relationships analysed from the data was done through exploring the correlation, distribution of feature values, distribution of feature sets, distribution of feature values in regard to the target feature etc.

The observations are documented extensively in the analysis notebook provided in the attachments.

Hardware and Software Requirements and Tools Used

The development of this project required the following tools:

- Python 3.8
- CUDA 10
- Scikit-learn 1.0

Model Development and Evaluation

Model Development

The model was trained on a 25% split of the training data which was fed to a pipeline of different classification algorithms. These models are listed below:

- Decision Tree Regressor
- Random Forest Regressor
- AdaBoost Regressor
- SVR
- XGB Regressor
- LGBM Regressor
- ElasticNet
- Gradient Boosting Regressor

The model was evaluated based on the accuracy, precision, recall and CV using AUC/ROC plots.

Hyper Parameter Tuning *(HPT)*

Parameter tuning was done using GridSearchCV on Decision Tree Classifier. This model was chosen based on its CV scores. The grid parameters that were used are given below:

- n_estimators: range(100,1200,12)
- Loss: [‘linear’, ‘square’, ‘exponential’]
- Learning rate: range(0.001,0.1,10)

The best parameters obtained from this process was then used to provide the submission model.

CONCLUSION

The submitted model provides consistent results that can be validated under other circumstances. The result obtained from the provided data can be assumed to be competitive and can provide better insights to the client. The provided solution satisfies all the requirements needed by the client.

Future Work could include cleaner, balanced and localised data with more independent feature sets to provide better estimations and analytical relationships.