Team ID: PTID-CDS-JAN22-1265

Project ID: PRCP-1025-FlightPricePrediction

Members:

1. Atharvdesh Pande

2. Nazeef Shaikh

3. Lukmanul Hakeem M

4. Sahil

5. Sumit Jangale

Project Analysis:

In this project we are trying to build a machine learning regression model for predicting Indian Domestic Flight Price. Data set used has 10683 records and 11 features including dependent ‘Price’ feature. We have found sklearn’s ensemble algorithm ‘ExtraTreesRegressor’ is most suited for our problem. On this algorithm we build a model that able to achieve adjusted R2-score of 0.971 on train data and 0.920 on test data.

Workflow:

Since this data set has several features with unstructured format which includes date, time and routes, we pass data through several pre processing and feature engineering steps before feeding into cross checking. There are some missing values and outliers present for some features.

Half of the features are categorical, with several trail-error steps we found use of label-encoding and one-hot encoding on perticular feature enhance performance of model in cross checking step. We use XGBoost, LGBM and several sklearn algorithms to cross check. We found ensemble algorithm ‘ExtraTreesRegressor’ perform better compared to other.

On cross validation in cross checking step ‘ExtraTreesRegressor’ model shows R2-score of 0.906 with standard deviation of 0.022, which is better and tells model is more consistent in prediction compared to others. After hyper parameter tuning of model, model achieved adjusted R2-score of 0.971 on train data and 0.920 on test data.

Files:

* ‘main.ipynb’: main notebook file where modeling done.
* ‘Flight\_Fare.xlsx’: data set file.
* ‘Route\_labelencoder\_ref.csv’: can be used as reference for converting 'Route' categorical values while deploying or external use.
* ‘etr\_model.gzip’: this file used to saved final ExtraTreesRegressor model, can be used while deploying.