ID5059 – Individual Assignment

230012908

2024-02-26

University of
St Andrews

Report: Predictive Analysis of Flight Disruptions

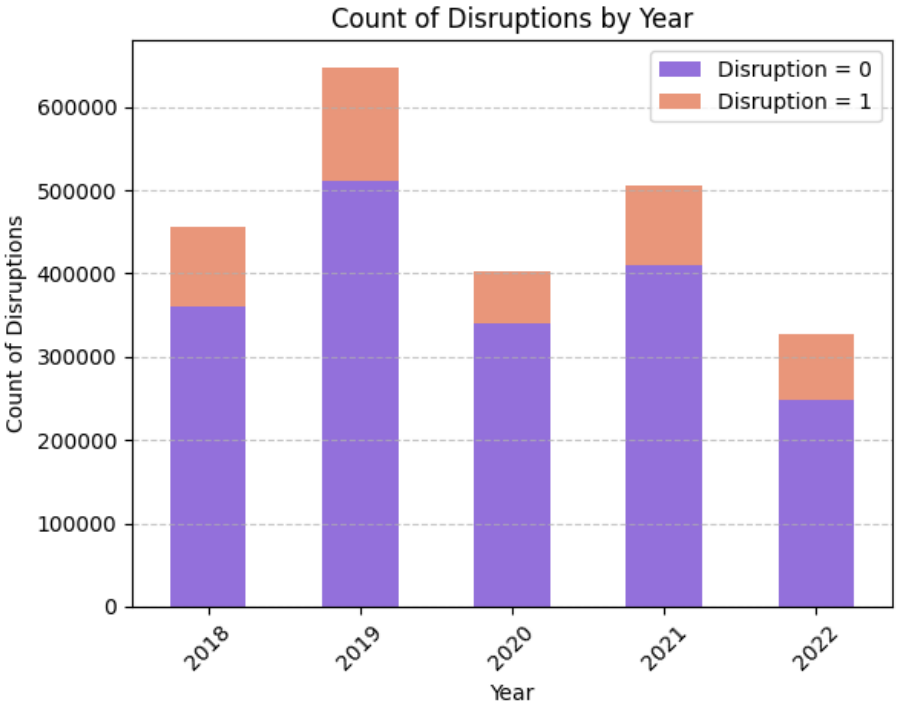
Abstract: The aim of this report is to document the analysis conducted in Python to illustrate and discuss the flight status (Disruption – No Disruption) outcome based on factors such as Airlines, Departure Time, Distance, Month, Day of Week, Tail Number, Origin and Destination City Names and to present the results to anyone who is interested in the topic. The main objective of the analysis is to predict if, before a flight takes off, it will suffer some type of disruption (cancellation, diversion, or a delay). Considering some factors, a small number of variables were selected to predict disruption. Then 3 models were fitted to the training data from which the best model was chosen for making predictions on testing data. The findings were obtained using various statistical methods to determine what parameters impact flight disruptions and what model is suitable for predictions.

Analysis: For the predictive analysis of flight disruptions a large dataset is used and since the data had already been split into training and testing, data exploration and data cleaning were the first steps taken. The data exploration showed that after the data was split, issues arose in the dataframe which was resolved by dropping 2 columns – DestState and OriginCityName and renaming all columns in the dataset. The disruption distribution bar plot grouped by years conveyed that the distribution of disruptions is proportional to the number of flights – more flights more disruptions, regardless of whether a flight was during Covid (2020-2021) or non-Covid time. The data was cleaned – rows with missing values removed and data types were changed to numeric for convenience. Considering the descriptions, data types, missing values in the data, number of unique values of attributes and correlation matrices, 8 features were selected to predict disruption, namely, Airlines, Departure Time, Distance, Month, Day of Week, Tail Number, Origin and Destination City Names. 3 models were fitted to the training data – Logistic Regression, Decision Tree, and Random Forest – and accuracy was estimated for each model. The most accurate model will be taken forward to final analysis against testing data.

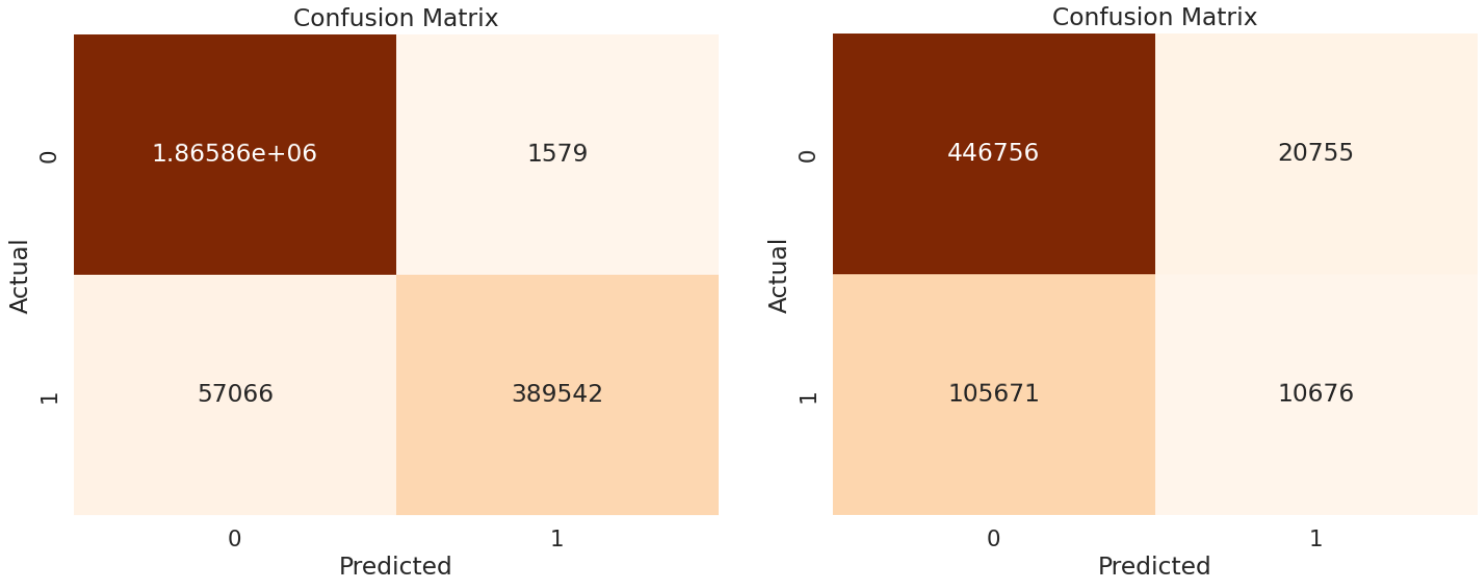
Results: Based on the comparison of the models' accuracy, precision, recall, F1 scores and cross-validation (accuracy and negative mean squared error) results, the Random Forest model was considered to be the best one among others with the accuracy and precision of 97% and 99% respectively. The confusion matrices and the AUC score helped prove the strength of the Random Forest model. Comparing the Random Forest performance on training data and on testing data, as expected the accuracy scores are lower when predicting on testing data (78%) which is still quite high and a good fit of the model.

Conclusion: To sum it all up, the analysis revealed that the best model for predicting flight disruption, among the 3 models fitted, is the Random Forest. Although the accuracy of the model is very high both on training and testing data, there still might be some flaws in the model as it is likely that it is overfitting which could be a strong foundation for further exploration and analysis.

Appendix



1. The impact of COVID-19 on flight disruptions.



2. Comparison between training (left) and testing (right) data confusion matrices results