Predicting Flight Disruptions in the Post COVID-19 Era
by 220029955

Introduction

Analysis was performed on a dataset containing 23 million North American flights and disruptions (i.e., flights that were delayed, diverted or cancelled). The dataset is inclusive of the years 2018 to 2022, and so it covers flights before, during and after the COVID-19 pandemic.

Forecasting disruptions would be useful for airline companies and passengers looking to avoid costly flight replacements or rebookings. In addition, accurate forecasts can also help governments in planning for future pandemics. The analysis here will determine whether COVID-19 had any effect on flight disruptions and predict disruptions for future flights.

Methods

The analysis was done in Python using the machine learning library 'Sci-Kit Learn'. Some data were missing in the target column 'Disruption', so this was addressed by inferring from the columns 'Delayed', 'Diverted' or 'Cancelled'.

Graphs were created with proportions of disruptions, and a statistical tool called 'chi-square' was used to determine if there were significant differences in COVID versus non-COVID years. Correlations were also computed to determine which columns provided the best information for flight disruptions. Out of the 46 columns, only eight were selected that had data on departure time, calendar schedule, airline and airport IDs. These were then transformed through numerical scaling (adjusting the values so they are easier to compare), and categorical one-hot encoding (giving each value its own numeric column).

To predict future flight cancellations, three machine learning classification models were created: Stochastic Gradient Descent (SGD), Logistic Regression, and Decision Tree. These were trained using cross-validation (i.e., diving the training set to have its own 'test' sets), and assessed using 'F1' (i.e., balanced scoring for imbalanced data). The SGD model was further tuned by applying different 'learning rates', and the unseen test set was then used to assess the models. As a final test, the models were fed records of actual disruptions to see whether they could correctly predict them.

Results and Recommendations

The overall percentage of non-disruptions (for all years) was 79.95%, however this varied per year (*Table 1*) from a low of 75.78% in non-COVID 2022 to a high of 84.51% in COVID 2020. The chi-square test on all years resulted in a p-value of 0.999, which means that there is no significant difference in COVID and non-COVID years in terms of flight disruptions.

Cross validation was performed, and the SGD model performed marginally better than the rest with a score of 0.3572 (*Table 2*).

Fine tuning the SGD model's *eta* values showed very small improvements, with *eta* value=0.0005 having the best performance with 0.3534 (*Table 3*). This model was then used to evaluate with the unseen test set, however, the model predicted more True Positives than True Negatives (*Table 4*).

Further evaluation showed the SGD model very performed poorly with an F1 score of 0.1621. The model had low 'precision' (only correctly predicted 'Disruptions' 21% of the time), and even lower 'recall' (only 13% of all 'Disruptions' were captured). (*Table 5*).

When tested against past flights, the model did predict 2 out of 3 scenarios, however we cannot predict future flights at this stage. The poor F1, precision, and recall scores show that further data wrangling and training are needed for an effective predictive model. On a lighter note, we can advise airline companies that COVID-19 did not have a significant effect on airline disruptions, so future pandemics should not severely affect these future flights.

Appendix

Table 1 Proportion of Non-Disruptions per Year

Year	Non-Disruptions
2018	79.07 %
2019	78.96 %
2020	84.51 %
2021	81.08 %
2022	75.78 %

Table 2 Cross Validation F1 Scores of Different Classifiers

Classifier	F1 Score
Logistic Regression	0.3547
SGD	0.3572
Decision Tree	0.3446

Table 3 F1 Scores from Fine Tuning the SGD Classifier

Eta Value	F1 Score
0.0001	0.3531
0.0005	0.3534
0.001	0.3508
0.005	0.3461

Table 4 Final SGD Classifier Confusion Matrix

	0	1
0	4,096,600	571,542
1	1,016,848	153,644

Table 5 Final SGD Classifier Scores

Metric	Score
F1 Score	0.1621
Precision Score	0.2119
Recall Score	0.1313