**Summary Report on the Flight Disruption Prediction Model**

Introduction

To achieve the purpose of improving operational efficiency for airlines, an optimal and accurate model is built based on historical data to predict whether flights will be delayed for airlines before departure.

Data preparation

The predictive model for flight disruptions using data from 2018 to 2022. The aim is to forecast disruptions, categorized as "0" for normal operations and "1" for cancellations, transfers, or delays exceeding 15 minutes. A subset of nine features has been selected, including temporal, geographical, and airline-related features.

Through data exploration, it was found that the overall delay rate is about 20%. The five-year period includes the COVID-19 epidemic period in 2020. In visualization process, it was found that the number of flights during this period was significantly reduced, but the delay rate was also reduced by about 5%. The possible reason is that due to the influence of policies and practical conditions, people's travel is restricted, and flight demand has dropped. So the airlines have more resources to use, which reduces the delay rate.

The data set is divided into 80% training set and 20% test set. And 20% validation set was also divided into the training set. The model was initially trained on the training set and to achieve optimal accuracy before final evaluation on the test set. Data processing included encoding categorical attributes, building feature attributes, and extracting hours from estimated departure and arrival times.

Build and select model

This model aims to address a supervised learning classification problem of predicting flight disruption. Three classifiers, including decision trees, random forests, and logistic regression, were employed. Model evaluation metrics encompassed accuracy, precision, recall, and F1 score. Cross-validation was utilized to assess model stability, while ROC curves and AUC scores assessed different model performances.

The Random Forest Classifier emerged as the optimal choice based on superior performance across evaluation metrics. Then, using a grid search to adjust parameters and try different values to find the best parameter combination and obtain the highest accuracy.

Conclusion

After fine-tuning the parameters, the optimal random forest model is obtained. The F1-Score, which balances precision and recall, is 88% for the non-disrupted class and 12% for the disrupted class. However, the model performance on predicting disrupted flights (class 1.0) is relatively low, especially in terms of recall. This performance gap may be attributed to data imbalance in the target prediction values.

Further improvement of the models may be considered to improve the predictive performance, especially for the positive class. Additionally, addressing data imbalance may be beneficial.