Predicting Flight Disruptions Report - 200018630

Introduction

The goal of this project was to predict whether a flight will experience disruption (cancellation, diversion, or a delay) before it takes off. The complete data between 2018-2022 contain more than 3 million rows, so we train and test our model on the medium subset with 30 thousands representation.

Design

The initial exploration revealed a brief increase in flight disruptions during the COVID-19 pandemic, before returning to average levels. Consequently, we excluded the pandemic as a variable in our model due to its transient impact. We preprocess the data, excluding 10 rows where the 'disruption' column contained null values. We then visualized the proportion of disrupted vs. non-disrupted flights within each potential feature category (e.g., Airline, Month, DayOfWeek) using stacked bar charts. Additionally, we examined the correlation between the target variable and factors like scheduled departure time, flight distance, origin, and destination. Finally, we selected 10 features that capture essential flight details, temporal aspects, and distance-related information, including 2 engineered new features that categorize scheduled departure times and flight distances: ['Month','DayOfWeek','Airline', 'Origin', 'Dest', 'FlightType', 'OriginState', 'DestState', 'PartOfDay', 'Quarter']. Since all variables are categorical, we used one-hot encoding to convert them into numerical form.

Implementation

Ensemble methods were chosen as they typically outperform single models. We first fit a non-tuned bagging logistic regression model and a random forest model as benchmarks. To place equal weight on disruption classes, we accessed the models via the macro average F1-score (Arithmetic mean of F1-scores for the two classes). The F1-scores of the models are as follows:

Model	Macro Average F1-Score %	Accuracy(Overall proportion of correct predictions %)
Model 1	0.46	0.51
Model 2	0.51	0.76
Model 3	0.57	0.67

The best performing model was a random forest model with the following parameters:

Parameter	Best Model (Model 3)	Raw Model (Model 2)
n_estimators	120	100
min_samples_split	3	2
min_samples_leaf	4	1
max_depth	15	None
class_weight	{0: 1, 1: 4}	{0: 1, 1: 4}

Testing

The model's performance was evaluated using a hold-out test set. The learning curve plot indicated convergence of training and testing errors as the sample size increases, suggesting good generalization. The precision-recall curve provides insights into the trade-off between precision and recall for the disruption 1 class, revealing a significant decrease in precision for minor improvements in recall. Lastly, the ROC-AUC curve demonstrates that the model outperforms a random classifier.

Difficulties

One difficulty was that the data was imbalanced, with around 80% non-disrupted flights, making the prediction for the disrupted class difficult. As shown from the Explanatory Data Analysis, the correlation between the target variable and the features was weak, making it challenging to predict the disruptions accurately.

Conclusion

Overall, the project developed a basic model that can predict flight disruptions with reasonable accuracy. However, future work could focus on improving the model's accuracy by incoperating more relavant factor such as number of disruption before departures, weather, and the live passenger traffic.