**University of St Andrews**

School of Computer Science

# Elevating Airline Analytics: A Comprehensive Model for Flight Disruption Prediction

Max Riffi-Aslett


Knowledge Discovery and Data Mining

ID5059 - Individual Assignment
Code: **Github**
Friday 24th February 2024

## 1 Introduction

The purpose of this project is to develop a predictive model capable of identifying flight disruptions within the United States. To achieve this, we will utilize a flight status prediction dataset posted on kaggle by Rob Mulla. We aim to create a preprocessing pipeline and modeling strategy to predict flight disruptions, including cancellations, diversions, and significant departure delays, using the approach described by (Géron 2021). By enhancing the accuracy of flight disruption predictions this project seeks to improve airline operations and passenger satisfaction.

## 2 The Big Picture

Flight disruptions induce heavy operational costs for airlines and airports (Fogaça et al. 2022). Our model seeks to help manage this issue by providing airlines and airports with insights to plan flight schedules and allocate resources efficiently. By predicting disruptions accurately, the model seeks to minimize disruptions and additional personnel expenses which in turn seeks to enhance the customer experience and reduce costs.

## 3 Model Overview

The model predicts flight disruptions using only pre-departure information, excluding post-departure data, for reliable predictions. It is a binary classifier with two possible outcomes: disrupted and non-disrupted. Models tested include a Random Forest, XGBoost, and a Neural Network with the neural Network being chosen as the best approach. An in-depth prepossessing pipeline was developed to ensure that the results are reproducible and the model can be retrained with new data. The evaluation of the model's performance was based on accuracy, and other more refined metrics such as precision, recall, and F1-score metrics which shed light on the behavior of the predictions which will be further emphasised later in this document (section 4).

## 4 Interesting Findings

Through thorough data exploration and numerical analysis, we found that departure cities, airlines, and airports were the most influential factors in predicting disruptions, while other factors, such as states, had less influence. (refer to Figure 1 for an overview of the covariates chosen and 2 for a visual representation of the states)

During the data processing stage, grouping locations like cities and airports by their past flight disruption rates proved to be a useful strategy. An overview of the preprocessing methods for each variable is provided in Table 1. Our exploration also revealed that COVID-19 had a significant impact on disruption rates. Figure 2 illustrates the trend in flight disruption rates. To prevent bias, we integrated a variable that considers the impact of COVID-19 during the training process.

In a situation where we are trying to predict flight disruptions, we decided that it is more important to capture all instances of disruption, even if it means occasionally incorrectly labeling a non-disrupted flight as disrupted, rather than missing actual disruptions. This is known as prioritizing recall over precision. (Khan 2021) who utilized Random Forest (RF) for predicting flight disruptions emphasized that recall was crucial for predicting flight disruption. To better understand the specific requirements for a flight disruption prediction system, it would be beneficial to consult with stakeholders who would be affected by the system. This would allow us to tailor the system to meet their unique needs and ensure that it is effective in helping them manage disruptions.

After refining each model's parameters, the Neural Network significantly outperformed the Random Forest and XGBoost models. A comprehensive analysis of errors for each model is presented in Table 2. Ultimately, the Neural Network model was chosen for this task. We utilized a fundamental type of neural network, specifically a Convolutional Neural Network. This model attained an accuracy of 0.6542, a recall of 0.6583, and a precision of 0.6309. The results of the final model are documented in Table 3.

The model's accuracy of 0.6542 shows that it correctly classifies instances better than random chance. The F1 score of 0.6374 indicates a good balance between precision and recall, which is usually preferred. It's worth highlighting that the model's recall of 0.6583 means it can accurately identify around 66% of all disruptions.

The model is still not accurate enough to make reliable predictions. We attribute this mainly to confounding factors that were not included in the analysis. Future efforts in this area could include adding forecasted weather, geopolitical factors, and anticipated number of scheduled flights in the variables.

## 5 Conclusion

With the aid of machine learning, a forecasting model was developed to predict flight disruptions. While room for improvement exists, this model serves as a solid foundation for further advancement and refinement.

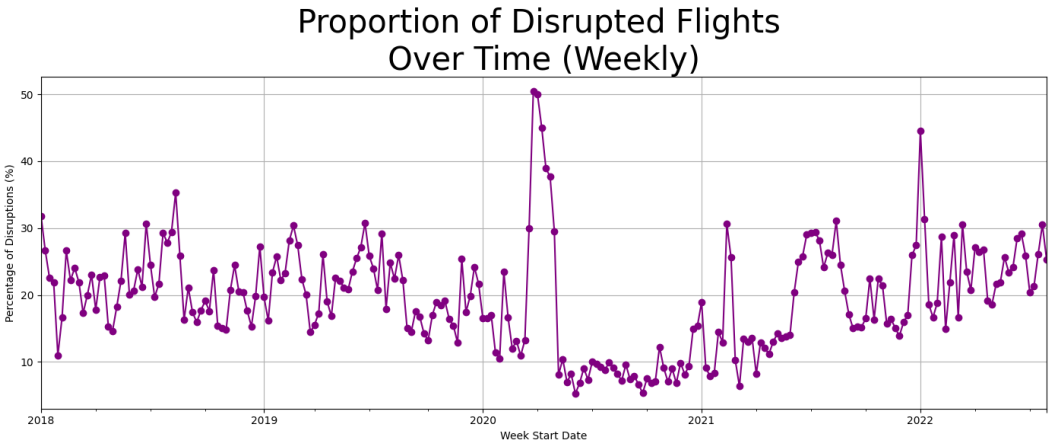


Figure 1: Timeline of the Proportion of Disrupted Flights from 2018 to 2023 from the flight status prediction dataset posted by Rob Mulla on Kaggle. The highlights specific events such as the impact of the COVID-19 pandemic with a pronounced peak in the beginning of 2020.

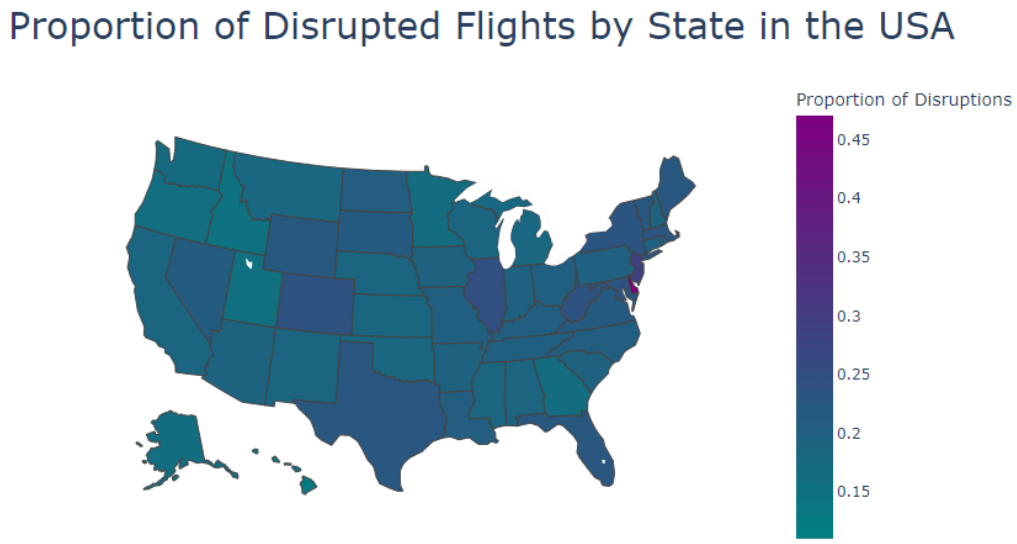


Figure 2: Proportion of Disrupted Flights from 2018 to 2023 in the United States from the flight status prediction dataset posted by Rob Mulla on Kaggle.

| Covariate | Description |
|---|---|
| Disruption | Dropped null values to avoid bias. |
| Distance | Imputed negative values and log-transformed due to skewness. Scaled values between 0 and 1 for the neural network using min-max scaler. |
| OriginCityName | Binned cities according to the proportion of disrupted flights and one-hot encoded resulting in 3 extra columns. |
| Airline | Created a binary variable to highlight airlines with a disruption rate over 25 percent. |
| Month, DayOfWeek | One-hot encoded to avoid adding too many columns. |
| DepTimeBlk | Split into morning, afternoon, evening, and night, then one-hot encoded. |
| CRSArrTime | Created 5 bins based on the proportion of disruption and then one-hot encoded. |
| FlightDate | Created a binary variable for the impact of COVID on disruption rates to avoid bias. |

Table 1: Overview of Covariates Used in Flight Disruption Prediction Model: This table details the preprocessing and encoding strategies for each covariate in our model.

| Metric | RandomForest | XGBoost | Neural Network |
|---|---|---|---|
| Training Accuracy | 0.6124 | 0.6049 | 0.6802 |
| Val Accuracy | 0.6042 | 0.6492 | 0.6492 |
| Val Precision | 0.6049 | 0.6429 | 0.6429 |
| Val Recall | 0.6052 | 0.6452 | 0.6452 |
| Val F1 Score | 0.6050 | 0.6411 | 0.6411 |
| Val True Positives (TP) | 51098 | 54812 | 54812 |
| Val False Positives (FP) | 33382 | 29575 | 29575 |
| Val True Negatives (TN) | 50771 | 54701 | 54701 |
| Val False Negatives (FN) | 33340 | 29503 | 29503 |
| Average Cross-Val Accuracy | 0.6007 | 0.6022 | - |
| Std Dev in Cross-Val Acc | 0.0014 | 0.0013 | - |

Table 2: This table presents the performance of three models: Random Forest, XGBoost, and Neural Network. The key performance metrics used for comparison are Accuracy, Precision, Recall, and F1 Score. The table also includes counts of True Positives, False Positives, True Negatives, and False Negatives. It's important to note that cross-validation was not performed for the Neural Network model due to its high computational cost

| Metric | Neural Network (Test Set) |
|---|---|
| Accuracy | 0.6542 |
| F1 Score | 0.6374 |
| Recall | 0.6583 |
| Precision | 0.6309 |
| True Positive Rate (TPR) | 0.65 |
| False Positive Rate (FPR) | 0.35 |
| True Negative Rate (TNR) | 0.65 |
| False Negative Rate (FNR) | 0.34 |

Table 3: This table presents the performance evaluation results of our final Neural Network model for predicting flight disruptions. Key metrics such as Accuracy, F1 Score, Recall, Precision, as well as the rates of True and False Positives and Negatives, are included.

# References

Fogaça, Lucas Bertelli et al. (Mar. 2022). "Airline Disruption Management: A Naturalistic Decision-Making Perspective in an Operational Control Centre". In: *[Insert Journal Name Here]*. URL: https://www.researchgate.net/publication/349776875_Airline_Disruption_Management_A_Naturalistic_Decision-Making_Perspective_in_an_Operational_Control_Centre.

Géron, Aurélien (2021). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media.

Khan, Waqar Ahmed (2021). "Hierarchical Approach for Flight Delay Prediction". In: *Journal of Air Transport Management* 90, p. 101933.