Prediction of Flight Delays in the United States
Final Report

Guillaume Allège
Fiona Aga
Theodoti Stoikou
Jan Philipp Girgott

Massachusetts Institute of Technology
Sloan School of Management
Advanced Analytics Edge (15.072)

December 7th, 2023

MIT MANAGEMENT SLOAN SCHOOL

Table of Contents

1 Problem Description

The aviation industry deals with the persistent challenge of flight delays, which significantly impact passenger satisfaction, airline operations, and economic efficiency. Recent data from the Bureau of Transportation Statistics highlights an unsettling trend in the fluctuation of delay rates. This project seeks to delve into the various factors contributing to these delays, such as adverse weather conditions, operational complexities, and heightened airport traffic volumes. By utilizing advanced analytical techniques, the project aims to understand the complexities of flight delay prediction and determine additional variables that may enhance the accuracy of delay forecasts. The goal is to not only understand the intricacies behind flight delays but also to provide a predictive model that stakeholders can use to mitigate the adverse effects of these disruptions.

2 Data

2.1 Kaggle Dataset

This project is anchored on a detailed dataset from Kaggle, which provides a year's worth of data on airline delays and cancellations for 2019. It's an extensive repository that includes intricate details such as flight schedules, the extent of delays, reasons for delays ranging from carrier-specific to weather-induced, and the status of cancellations. This comprehensive dataset is crucial for understanding delay patterns and discerning the contributory elements that lead to flight disruptions, serving as a foundation for in-depth classification analysis.

2.2 Selection of Subsets: Three Largest Airports and Airlines

Our analytical goals are streamlined by focusing on subsets pertaining to the three most prominent airports and airlines, selected based on their high passenger throughput and operational intensity. The data preprocessing yielded six distinct subsets, tailored to each airport and airline, enabling targeted analyses with our models. These subsets integrate various dimensions of data, including time-based factors, delay classifications, airline and airport operational specifics, quality of service indicators, aircraft age, and environmental conditions. This segmentation is designed to facilitate a granular analysis of the specific operational dynamics and intricate delay patterns that are characteristic of the airline industry. The uniform structure of these subsets is vital for ensuring the integrity and comparability of our multivariate analyses, which seek to explain the interplay of diverse factors that contribute to flight delays. Through this data architecture, we aim to construct predictive models that are both robust and reflective of the complexities of airline operations.

3 Methodology

Our methodology is structured to leverage statistical and machine learning methods tailored to address specific aspects of flight delays. We use a phased approach, starting with binary classification to determine the likelihood of a delay, followed by multiclass classification to identify the reasons for delays, and concluding with an attempt at regression models to understand their

limitations with our dataset. This comprehensive approach ensures a holistic analysis of the factors influencing flight delays, allowing for actionable insights and robust predictions.

Binary Classification: Will the Flight be Delayed?
Binary classification will serve as the foundation for our predictive analysis, determining if a flight will be delayed or not. This method will process variables such as time of day, weather conditions, and airline operational metrics to output a binary decision. The classification threshold will be fine-tuned to balance sensitivity and specificity, optimizing for the most accurate predictions.

Multiclass Classification: Why Will a Flight be Delayed?
Our multiclass classification model is designed to detect the different reasons behind flight delays. It will classify each potential delay into categories based on the identified variables such as carrier issues, weather challenges, or air system delays. This approach not only highlights the primary factors for delays but also quantifies the relative impact of each factor.

3.4 Regression: By How Much Will a Flight be Delayed?
The regression analysis in our project serves a dual purpose: to attempt to predict delay durations and to show the importance of appropriate method selection. Given that our dataset is tailored primarily for classification tasks, regression models were not as effective. This highlights the crucial nature of choosing suitable modeling techniques in relation to the dataset at hand. To derive accurate delay duration predictions, a different dataset with more granular time-based variables would be necessary.

4 Results

4.1 Results for Three Largest Airports

The evaluation of predictive models for flight delays using Area Under the Receiver Operating Characteristic Curve (AUC) and accuracy as metrics yielded notable results. In binary classification tasks, where the goal was to predict whether a flight would be delayed or not, the Random Forest algorithm achieved the highest AUC scores for Houston Airport (0.88), followed closely by Nashville Airport (0.85) and Minneapolis Airport (0.83). These results are indicative of the Random Forest model's superior performance in differentiating delayed and on-time flights for these airports.

In multiclass classification scenarios, where delays were categorized into multiple classes, Random Forest again proved most effective, with the highest accuracy for Nashville Airport at 76.7%, closely followed by Houston Airport at 76.3%, and Southwest Airlines achieving the top accuracy among carriers with 77.1%. These findings underscore the capability of the Random Forest model in handling more complex, multi-tiered delay classifications, especially for Nashville Airport.

4.2 Results for Three Largest Airlines

Upon analyzing the largest U.S. carriers, Southwest Airlines exhibited robust predictive results across both binary and multiclass classification tasks. In binary classification, it achieved an AUC of 0.87 and maintained high accuracy in multiclass classification with a 77.1% rate. Delta Airlines showed moderate performance with an AUC of 0.81 and a multiclass accuracy of 74.7%, while SkyWest's performance was slightly lower with an AUC of 0.82 and a multiclass accuracy of 74.9%.

The consistent performance of the Random Forest algorithm across different airports and airlines underscores its reliability as a predictive tool for flight delays. However, the variance in predictive accuracy between different airports and airlines indicates that external factors specific to each may significantly influence the outcomes, necessitating tailored strategies for delay mitigation.

5 Discussion of Results

5.1 Managerial Implications and Recommendations

The results from the predictive models provide a probabilistic assessment of flight delays, which can be leveraged to enhance decision-making processes for both airports and airlines. For airports experiencing high volumes of traffic, predictive insights can inform gate assignment strategies. By analyzing patterns of predicted delays, airport management can better allocate gates and plan for potential delays, potentially reducing their frequency and impact.

Airlines can utilize these predictions to improve crew scheduling by assigning more experienced crews to flights that are predicted to have a higher probability of delay. This can be particularly impactful in mitigating the effects of delays when they do occur. Moreover, the inclusion of weather data within the models allows for more informed decisions regarding equipment readiness and maintenance, helping to avoid technical delays.

Optimizing time padding involves a strategic balance. Time padding is the extra time airlines add to flight schedules to account for unforeseen delays. While too little padding can lead to a higher incidence of late arrivals, too much padding can result in inefficient operations and increased costs. By utilizing predictive models, airlines can tailor the padding time to the historical performance of specific routes, times of day, and even specific aircraft types. This approach would allow for a more dynamic and data-driven allocation of padding time, potentially reducing unnecessary wait times for passengers while still safeguarding against the operational impact of delays.

5.2 Limitations

The predictive models, while incorporating historical data and weather variables, do not in themselves suggest actions but rather provide a statistical likelihood of delay occurrences. This probabilistic output is invaluable for planning and operational adjustments; however, it is imperative to recognize that these models are not prescriptive.

Furthermore, the models may not fully capture real-time operational nuances or sudden, drastic weather changes. The severity and duration of delays are not differentiated within the scope of this model, which could be critical in assessing the operational and economic impact of delays on both airports and carriers.

While the inclusion of weather data provides a more comprehensive view, there remains a degree of unpredictability in weather patterns that can affect the accuracy of predictions. Hence, it is important to continuously update models with the most current data available and to maintain flexibility in operational plans to accommodate for this variability.

6 Appendix

6.1 In-Depth Methodology and Results

Multiclass Classification

The target variable for classification, ARR_DEL_CAT, is derived from the maximum delay type, and is categorized into five classes representing different delay reasons. Given the class imbalance, we apply the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes in the training set. This approach generates synthetic samples for minority classes, thereby providing a more balanced dataset for model training. We then use StandardScaler within a pipeline to standardize features by removing the mean and scaling to unit variance, ensuring that all features contribute equally to the model's performance.

We employ a variety of classification models, each chosen for their suitability in multi-class classification scenarios:

1. **Decision Tree Classifier (CART):** A tree-based model providing interpretable classification rules.
2. **Random Forest Classifier:** An ensemble of decision trees, typically more robust and accurate than individual trees.
3. **Logistic Regression:** A linear model for classification, extended to multi-class scenarios.

For each model, a range of hyperparameters is explored using GridSearchCV, which performs exhaustive search over specified parameter values. This search is coupled with cross-validation to evaluate the performance of each parameter combination, ensuring robustness and generalizability of the models. The primary evaluation metric is accuracy, calculated through cross-validation during hyperparameter tuning. Additionally, confusion matrices are generated for each model to understand the classification performance across different classes. Each model's performance is assessed on the test set, which is kept separate and not involved in the training or tuning process.

Regression: By how Much Will a Flight be Delayed?

We utilized six distinct regression techniques to assess their efficacy on the dataset:

1: Linear Regression: A basic method that models the linear relationship between a dependent variable and one or more independent variables. It's useful for predicting outcomes and understanding the influence of predictors.
2. Elastic Net: A linear regression model that combines L1 and L2 regularization to prevent overfitting and to handle collinearity.

3. Partial Regression: This method was employed to isolate the effect of a single independent variable while controlling for the impact of others.
4. Polynomial Regression: Used to model a non-linear relationship between the independent and dependent variables by transforming the predictors using polynomial features.
5. Regression Trees: A non-linear model that divides the dataset into branches to make predictions.
6. XGBoost: An advanced ensemble technique known for efficiency and performance, particularly good at uncovering complex patterns in large and diverse datasets.

The dataset was preprocessed to fit a regression framework, including normalization of features and handling of missing values. The evaluation metrics selected were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared.

MODEL NAME	MAE	MSE	R^2	RMSE	MAPE
LINEAR REGRESSION	54.736	11474.804	0.007946	-	-
ELASTIC NET	52.641	-	0.0096	96.007	122.994%
PARTIAL REGRESSION	120.286	40172.582	0.0365	-	-
POLYNOMIAL REGRESSION	52.665	-	0.0159	95.920	123.012%
REGRESSION TREES	53.760	-	-0.0000569	100.686	124.824%
XGBOOST	50.402	-	0.055988	93.733	115.995%

As visible from the results, each model encountered specific challenges with the dataset, with XGBoost showing a relative, though limited, improvement in performance. The main difficulty was the nature of the dataset, which is inherently designed for classification tasks. The challenges in adapting regression methods to this dataset were evident in the performance metrics, particularly the high MAPE values and low R-squared scores. Additionally, the dataset's variables, including their distribution and the presence of null values, posed significant challenges to achieving a good model fit. The results suggest that the dataset's characteristics might be better suited for classification models rather than regression models.