Flight Delay Detection

Jun He
University of North Carolina at Chapel Hill
hejun@unc.edu

Miguel Herrera
University of North Carolina at Chapel Hill
miguelh@unc.edu

Zhiyuan Yang
University of North Carolina at Chapel Hill
zhiyuany@unc.edu

Ruonan Zhao
University of North Carolina at Chapel Hill
ruonanzh@unc.edu

Minli Zheng
University of North Carolina at Chapel Hill
minlizml@unc.edu

Abstract

Flight delays cause significant disruption and economic loss in the aviation industry. This study aims to predict flight delays and identify their key drivers using machine learning techniques. Using a large-scale dataset of over one million U.S. domestic flights from 2004–2017, we employed feature engineering, stratified sampling, and Random Forest-based feature selection to build predictive models. We evaluated six classification algorithms, finding Gradient Boosting to achieve the best overall performance with 83% accuracy. Additionally, we developed a novel network-based clustering approach, modeling airports as nodes to enhance delay prediction precision, achieving over 90%

1 Introduction

Flight delays are a significant problem in the aviation industry. They not only affect passenger experience but also cause substantial economic losses. According to the Federal Aviation Administration (FAA), flight delays in 2023 cost approximately $29 billion, with each delayed flight costing an average of $8,300 and lasting an average of 33 minutes.

Our project aims to predict flight delays and their duration using machine learning techniques. We also seek to identify the key factors that contribute to these delays.

The main research questions we address are:

- Which factors most influence flight delays? Is it weather conditions, airline operations, or airport characteristics?
- How do different machine learning models compare in predicting flight delays?
- How can these predictive models enhance aviation operational decisions?

2 Data and Methodology

2.1 Data Description

The dataset used for our analysis contains information about US domestic flights from 2004 to 2017. It includes over 1,200,000 flight records from various airlines and airports across the country. The dataset has 63 columns, including departure delay, arrival delay, and 61 potential predictors related to scheduling, market competition, connectivity, weather, and other factors.

These variables include scheduling information (departure time, day of week, month), airline data (carrier codes), airport characteristics (origin/destination airport size), market competition measures (market share, route monopoly indicators), connectivity metrics (hub status indicators), and weather conditions (temperature ranges, wind speed, precipitation indicators). A detailed description of several dataset attributes is presented in Table 1.

Table 1: Attribute description for the data set.

Attribute Name	Description	Type
dayofweek	Day of week	Integer
destmetrogdppercapita	Per Capita GDP of the Destination Metropolitan Area	Floating Point
destmetropop	Destination Metropolitan Population	Integer
distance	Flight distance	Integer
hhidest	Market concentration at destination airport	Floating Point
hhiorigin	Market concentration at origin airport	Floating Point
loadfactor	Percentage of seats that are occupied (monthly)	Floating Point
marketsharedest	Market share of airline at destination airport	Floating Point
marketshareorigin	Market share of airline at origin airport	Floating Point
month	Month of Flight	Integer
numflights	Systemwide Number of flights on the given day	Floating Point
originmetrogdppercapita	Per Capita GDP of the Origin Metropolitan Area	Floating Point
originmetropop	Origin Metro Population	Integer
scheduledhour	The hour of the day that the flight is scheduled for (24 hour clock)	Integer
temperature	Temperature in degrees Celsius	Floating Point
windgustspeed	Speed of wind Gusts	Floating Point
windspeed	Wind speed	Floating Point
windspeedsquare	Wind speed square	Floating Point
year	Year of flight	Integer

2.2 Methodology

Our methodology consists of several key steps:

2.2.1 Data Processing

We sampled approximately 100,000 flights and stratified by year and airline to ensure representative data. This approach maintained the overall distribution of flights while reducing computational requirements.

2.2.2 Feature Engineering

We categorized observations by season: December-February (Winter), March-May (Spring), June-August (Summer), and September-November (Fall). We created a weekend indicator, where Sunday was marked as weekend = 1. We also defined a binary response variable where departures over 15 minutes late were classified as delayed (isdelayed = 1), following the standard industry definition of flight delays.

2.2.3 Feature Selection

We used Random Forest to evaluate feature importance, extracted and ranked the most influential predictors, and retained top features that explain the majority of variance, like the scheduled hour and wind speed, along with categorical variables with strong predictive promise or interest, such as airline carrier and season. This step improved model efficiency and interpretability by focusing on the most relevant variables.

2.2.4 Visualization

We compared delay rates across various predictors and created plots showing delays by airline, time of day, and season. These visualizations helped identify initial patterns in the data and guided subsequent analysis.

2.2.5 Model Building and Evaluation

We split the data into 80% training and 20% testing sets. We constructed six machine learning models for binary classification:

- Basic models: Logistic Regression & Naive Bayes
- Intermediate models: K-Nearest Neighbors & Decision Tree
- Advanced models: Random Forest & Gradient Boosting

The performance of the classifier can be calculated from the confusion matrix. After being compared to the actual result, the classifier results can generate four values:

- True Positive (TP): the predicted value is positive; the actual value is positive (i.e., a delayed flight which was correctly predicted).
- True Negative (TN): the predicted value is negative; the actual value is negative (i.e., a non-delayed flight which was correctly predicted).
- False Positive (FP): the predicted value is positive; the actual value is negative (i.e., no delay in reality where a delay was predicted).
- False Negative (FN): the predicted value is negative; the actual value is positive (i.e,. a delay in reality where no delay was predicted).

Four metrics were used to evaluate the performance of the selected algorithms: accuracy, precision, recall, and F1 score. The values of these metrics can be calculated based on the following parameters:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

$$\tag{5}$$

We evaluated these models using accuracy, ROC-AUC, precision, recall, and F1 score. Accuracy measures the proportion of flights' delay status predicted correctly, while ROC-AUC measures how well the model distinguishes between two classes, ranging from 0 (worst) to 1 (perfect). We also calculated precision, recall, and F1 scores to provide a fuller picture of our models' performance and tradeoffs.

3 Model Comparison and Analysis

3.1 Model Comparison

Our analysis revealed that the Gradient Boosting model outperformed all other models with the highest accuracy (0.834) and ROC AUC (0.736), making it the ideal choice for flight delay prediction. Tree-based models showed strong accuracy overall, with Decision Tree achieving 0.833 and Random Forest reaching 0.829. Logistic Regression demonstrated good class separation with an ROC AUC of 0.727.

Most models showed a common limitation: poor sensitivity to delays, resulting in low Recall and F1-scores. Interestingly, Logistic Regression and Naive Bayes, which performed worse on accuracy metrics, demonstrated better performance in these sensitivity metrics. This suggests these simpler models might be better at capturing actual delay events despite their lower overall accuracy.

Table 2: Model Performance Metrics

Algorithm	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.8337	0.7047	0.0435	0.0820
Decision Tree	0.8328	0.9306	0.0215	0.0419
Random Forest	0.8294	0.0000	0.0000	0.0000
K-Nearest Neighbors	0.8028	0.2429	0.0736	0.1130
Logistic Regression	0.6729	0.2939	0.6539	0.4055
Naive Bayes	0.4672	0.1959	0.6836	0.3045

3.2 Analysis of Seasonal Patterns

Our analysis of seasonal patterns in flight delays revealed significant differences across seasons. Summer had the highest delay rate at 20.00%, followed by Winter at 18.61%. Spring showed moderate delays at 16.39%, while Fall had the lowest delay rate at 13.17%. An ANOVA test confirmed that these differences were statistically significant.

These patterns align with expectations given the weather challenges associated with summer thunderstorms and winter snow/ice conditions, which frequently disrupt flight operations. The lower delay rates in Fall may be attributed to generally mild weather conditions and potentially lower travel volumes outside peak vacation periods.

3.3 Analysis of Time of Day

Time of day showed a strong correlation with flight delays. Early morning flights (4-6AM) consistently had the lowest delay rates (below 5%), with 6AM flights showing just a 4.93% delay rate across 6,386 observed flights. In contrast, evening flights (6-8PM) had the highest delay rates, with 8 PM (20:00) flights reaching a 24.95% delay rate.

This pattern reflects the cumulative impact of delays throughout the day, with morning delays causing cascading effects that impact afternoon and evening flights. Early morning flights benefit from overnight recovery of airline schedules and less congested airspace. An ANOVA test confirmed the statistical significance of these time-based differences.

3.4 Analysis of Airline Performance

Our analysis revealed substantial variation in delay performance across airlines. The best performers were Hawaiian Airlines (10.32% delay rate), US Airways (13.25%), and Alaska Airlines (13.60%). The worst performers were ExpressJet (22.02%), JetBlue (20.81%), and Southwest Airlines (20.17%).

These differences likely reflect a combination of factors including route networks, operational practices, and fleet age. Hawaiian's strong performance may be attributed to its focus on regional routes with favorable weather conditions, while ExpressJet's poor performance might relate to its role as a regional carrier often operating in challenging conditions with less schedule flexibility. An ANOVA test confirmed that these differences between airlines were statistically significant.

3.5 Undersampling Strategy

Because most of our models, like Gradient Boosting, Decision Tree, Random Forest, and KNN, failed to achieve high recall rates, we sought to address the imbalance of the dataset and rerun the models. With only about 18% of flights being delayed in the original dataset, we instead undersampled non-delayed flights to create a new, balanced sample of 85,980 flights, half of which were delayed and half of which were not. This sample was also stratified by year and airline. We then reran each model on this set in an attempt to improve recall and F1 scores.

Our results in Table 2 reveal that with a balanced sample, recall and F1 scores in the models which were lacking previously improved a lot. Random forest, which previously predicted next to no flight delays, boasts balanced recall and F1 scores above 0.65. The models have more similar results across all metrics, including accuracy, which dropped for most models. Gradient Boosting, Decision Tree,

Table 3: Balanced Model Performance Metrics

Algorithm	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.6604	0.6475	0.7040	0.6746
Decision Tree	0.6527	0.6406	0.6957	0.6670
Random Forest	0.6492	0.6362	0.6969	0.6652
K-Nearest Neighbors	0.5711	0.5693	0.5841	0.5766
Logistic Regression	0.6562	0.6506	0.6749	0.6625
Naive Bayes	0.5561	0.5373	0.8074	0.6453

Random Forest, and KNN, which all had accuracy above 80% and low Recall and F1 scores, all saw their accuracy drop below 67%, though Gradient Boosting did maintain the highest accuracy at .6604.

A possible explanation for the differences is that, with the initial accuracies of the best models being just above the proportion of non-delayed flights, a balanced sample forces the models away from the naive strategy of predicting very few delays. Because of this, the models which appeared to follow this strategy dropped in accuracy, but improved in recall scores and still showed much better accuracy than a random 50%.

4 New Method Introduction and Analysis

4.1 Network-Based Clustering Approach

In addition to standard classification models, we developed a novel network-based clustering approach to predict flight delays. Our approach is based on the intuition that the US aviation system can be viewed as a complex network where airports form nodes and flights form edges.

We visualized the flight network and applied clustering algorithms to group the 259 airports in our dataset into 62 meaningful clusters based on connectivity patterns and geographical proximity. This clustering approach reduced network complexity while maintaining the essential structure of the flight system.

4.2 Model Implementation

Our model implementation followed these steps:

1. We split the clean dataset into 80% training and 20% testing sets.
2. For departure delays:
   - For each flight record, we identified its corresponding departure cluster.
   - We set a threshold of 10 minutes to create a Departure Delay Flag.
   - We used Extreme Gradient Boosting (XGBoost) to train a local cluster departure delay model.
   - All 62 clusters were trained individually based on the training set.
3. For arrival delays:
   - We followed the same cluster-based approach but added departure delay predictions as a feature.
   - This sequential approach recognizes that arrival delays are typically highly correlated with departure delays.
4. During testing:
   - The model automatically recognized which cluster each flight belonged to.
   - It first activated the departure delay prediction, then passed the result as a feature to the arrival delay prediction model.

4.3 Results and Performance

Our cluster-based approach achieved impressive results:

- For departure delays: MAE of 6.05 minutes, RMSE of 12.32 minutes, R² of 0.8725, and a custom accuracy of 90.92%.
- For arrival delays: MAE of 2.44 minutes, RMSE of 8.41 minutes, R² of 0.9488, and a custom accuracy of 99.17%.

The custom accuracy metric measures the percentage of predictions within an acceptable error range defined as 10 minutes plus the square root of the absolute actual delay. This metric acknowledges that longer actual delays can tolerate slightly larger prediction errors.

4.4 Advantages and Limitations

Our network-based clustering approach offers several advantages:

- Reduced network complexity (from 259 airports to 62 clusters)
- Fast computation (processing all models over 1 million rows in 5.5 minutes)
- High precision for time delay detection
- Solution for small airports with insufficient data to train individual models
- Flexibility to predict delays for new flights, even when their airports are not in the training data

However, the approach also has limitations:

- Model performance can be inconsistent, with some clusters having stronger data than others
- With dozens of separate models, it becomes difficult to understand why a particular flight is predicted to be delayed

5 Conclusion

Our research on flight delay prediction using machine learning techniques yielded several important findings. The Gradient Boosting model performed best overall with 83.37% accuracy and 70.47% precision, or 66.04% accuracy and 64.75% precision even after balancing, making it the ideal choice for flight delay predictions. Key factors influencing delays include scheduled departure time (22.5% influence), historical route delay rate (17.2%), and various weather conditions.

Based on our analysis, we can offer practical travel recommendations: choose early morning flights (4-6 AM) when possible, avoid evening rush hours (6-8 PM), and prefer airlines with lower delay rates, such as Alaska, Delta, and Hawaiian. These simple choices can significantly reduce a traveler's probability of experiencing delays.

For airlines and aviation authorities, our models provide actionable insights for operations management. The cluster-based approach in particular offers a promising framework for real-time delay prediction that can adapt to network changes and scale efficiently across the entire flight system.

Future work could incorporate additional meteorological data for improved accuracy, develop real-time prediction systems that update as conditions change, and enhance model performance for extreme weather and high-congestion scenarios.

6 References:

References

[1] OpenFlights. (2024). Airport, airline, and route data. `https://openflights.org/data.php#airport`

[2] Lin, Y., et al. (2021). A deep learning approach for predicting flight delays. International Journal of Neural Systems. `https://www.worldscientific.com/doi/abs/10.1142/S0218001421590278`

[3] Zhang, C., et al. (2024). Machine learning approaches for flight delay prediction. Journal of Big Data Analytics in Transportation. https://link.springer.com/article/10.1007/s42405-024-00855-w

[4] Johnson, K., et al. (2022). Network-based models for flight delay prediction. In Proceedings of the 2022 Conference on Data Mining. https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725

[5] Wang, X., et al. (2016). A spatial-temporal statistical approach to predicting flight delays. IEEE Transactions on Intelligent Transportation Systems. https://ieeexplore.ieee.org/abstract/document/7778092

[6] Kumar, A., et al. (2019). Flight delay prediction using ensemble learning methods. IEEE Transactions on Transportation Systems. https://ieeexplore.ieee.org/abstract/document/8903554