Predicting Airplane Departure Delays

Asher Erickson, Linqing Mo, Kyllan Wunder
Michigan State University

April 26, 2024

Abstract

This paper presents an in-depth analysis of factors contributing to the delays of U.S. domestic flights from 2014 to 2018, leveraging a comprehensive dataset that includes both flight details and associated weather conditions. By applying a variety of statistical and machine learning techniques, such as logistic regression, random forests, and neural networks, the study aims to predict flight delays with an emphasis on their duration and influencing factors. The integration of detailed weather data, airport traffic volumes, and temporal features such as proximity to federal holidays enhances the model's predictive power. The findings are intended to inform and improve operational strategies for airlines and airports, helping to reduce the economic impact of flight delays while improving passenger experience. This research not only sheds light on the dynamics of flight delays but also showcases the application of advanced analytics in the aviation industry for proactive delay management.

For access to the complete source code and additional resources used in this project, visit our GitHub repository at: `https://github.com/kyllan16693/flight-delay-prediction`

Contents

1 Introduction

Flight delay is a serious and widespread problem in the United States. We found that nearly 20% of all domestic flights were delayed by more than 15 minutes, according to the Bureau of Transportation Statistics (BTS) data. As demand for air travel continues to grow, the U.S. aviation system is increasingly challenged to maintain timely operations.

Flight delays are costly for all parties involved, significantly impacting airlines, travelers, and the broader economy. In 2019, these delays cost the U.S. economy approximately $32.9 billion, as reported by the Federal Aviation Administration (FAA). Airlines face increased expenses due to higher fuel consumption and staff overtime. At the same time, passengers may incur costs from missed connections and prolonged waits, including additional expenses for hotel stays and meals. Also, the environmental impact of planes idling on the runway for longer times causes more fuel to be burned.

Understanding the reasons behind flight delays is essential for stakeholders in the aviation industry and individual travelers alike. By identifying and analyzing the root causes of these delays, airline operators and airport authorities can develop more effective strategies to mitigate them. Additionally, when passengers understand the causes of delays, they can better manage and minimize disruptions to their travel plans.

This project uses statistical modeling to analyze U.S. domestic flight data and corresponding airport weather conditions from 2014 to 2018. Our goal is to identify patterns and determine the primary factors influencing flight delays. This analysis involves variables such as weather conditions, airport traffic volumes, and the timing of flight departures. To achieve this, we employ various statistical techniques including naïve Bayes, logistic regression, random forest, KNN, and neural network classifiers to address a binary classification problem—whether a flight will be delayed—and a multiple classification problem—when a flight will be delayed how long will that delay be. We aim to provide insights into the field of aviation management, benefiting airline operators, airport authorities, and travelers.

2 Data Description

2.1 Flight Data

The dataset used for this project was sourced from Kaggle. This comprehensive dataset, titled "US Domestic Flights Delay Prediction (2014 - 2018)," was compiled by the Office of Airline Information, under the Bureau of Transportation Statistics (BTS). It includes data on scheduled and actual departure and arrival times, reported by certified US air carriers that account for at least one percent of domestic scheduled passenger revenues. The dataset contains extensive details on flight dates, times, origins, destinations, airlines, distance, and delay status.

2.2 Weather Data

Additional weather data were integrated from NOAA's Climate Data Online service, which offers free access to a global archive of historical weather and climate data. These data were specifically collected from weather stations at each of the main airports involved, enhancing the dataset with hourly measurements of temperature, precipitation, wind, weather conditions, and other climatic variables relevant to flight delay predictions.

3 Data Preprocessing

3.1 Dataset Consolidation and Refinement

The comprehensive dataset we analyzed originated from a merger of 60 individual CSV files covering flight operations from 2014 to 2018. After amalgamating these files into a single dataset, an initial cleansing step was undertaken to remove any columns exhibiting more than 80% missing values, thereby focusing on the most reliable and complete data fields. This process resulted in a reduction from 109 to 56 columns.

Following the column filtration, rows with any missing values, which constituted less than 1% of the total data, were also discarded. This ensures a higher level of data integrity and usability for our predictive modeling.

In an effort to target the most significant sources of data within this refined dataset, our analysis was further narrowed down to concentrate on flights involving the eight largest airports and the top five airlines, based on traffic volume and operational scope. The selected airports and airlines, with their corresponding abbreviations, are listed below:

- Airports:
  - Hartsfield-Jackson Atlanta International Airport (ATL)
  - Chicago O'Hare International Airport (ORD)
  - Dallas/Fort Worth International Airport (DFW)
  - Denver International Airport (DEN)
  - Charlotte Douglas International Airport (CLT)
  - Los Angeles International Airport (LAX)
  - George Bush Intercontinental Airport (IAH)
  - Phoenix Sky Harbor International Airport (PHX)
  - San Francisco International Airport (SFO)
- Airlines:
  - United Airlines (UA)
  - SkyWest Airlines (OO)
  - Southwest Airlines (WN)
  - American Airlines (AA)
  - Delta Air Lines (DL)

This focused approach allows for a more detailed analysis of flight delays related to major hubs and carriers, which are pivotal to understanding broader trends and impacts within the aviation industry.

In the resulting dataset of 56 columns, many columns contained redundant information. For example, the columns OriginAirportSeqID (1039705), Origin (ATL), OriginCityName (Atlanta, GA), and OriginState (GA) all indicated that a flight departed from Atlanta International Airport in the state of Georgia. To simplify the dataset, we retained only the Origin(ATL) variable. Furthermore, since the project's objective is to predict flight delays before departure, we removed all arrival statistics, such as the actual arrival time. After these adjustments, the final comprehensive dataset contains 1,639,429 rows and 17 columns.

3.2 Feature Engineering

According to the FAA's 2022 report, spanning from 2017 to 2022, approximately 75% of flight delays were caused by weather conditions, 15% by airport volume, and 6% by closed runways. To enhance the predictive accuracy of our model, we engineered several new features based on this data. This process involved creating variables that provide deeper insights into the factors influencing flight delays.

3.2.1 Federal Holiday

Using 2014/01/01, as the reference date, we calculated the number of days from this date for each flight. We also identified all federal holidays between 2014 and 2018. For each flight, we created a new variable called is_holiday_week to indicate whether the flight occurred within a week of a federal holiday, defined as a period extending three days before and three days after the holiday. This variable helps in assessing the impact of holidays on flight schedules and potential delays.

3.2.2 Airport Volume

Instead of using the refined comprehensive dataset, we decided to better represent the real airport volume by calculating all flights' departures and arrivals at each main airport for any given period. To do this, we needed to revert to the original 60 data files.

First, for each airport, we identified all departing flights and their respective departure times. We then created a new data frame that included every minute from 2014 to 2018. For each minute, we used a 120-minute window—60 minutes before and 60 minutes after—to calculate the number of flight departures during this period for each major airport. We applied the same strategy to the arrival data.

Finally, we integrated this data back into our main flight dataset. This integration shows the departure and arrival density at the exact time of each flight, providing a more dynamic and precise representation of airport traffic volume.

3.2.3 Weather

For the weather data, we concentrated on three key metrics at each airport station: hourly wind speed, hourly visibility, and hourly weather conditions. The hourly wind speed and hourly visibility are numerical variables, which can be used directly without additional processing. However, the hourly weather conditions are represented as strings describing the weather for each hour, separated by vertical lines. For example, ’TS:7 +RA:02 FG:2’ indicates a thunderstorm, heavy rain, and fog during that hour. To utilize this information effectively, we partitioned these strings and converted them into categorical variables.

We categorized weather conditions into three levels to better assess their impact on flights, as recorded by the FAA:

- **Normal Weather:** Conditions that do not fall into the below categories.
- **Bad Weather:** Includes less severe but impactful conditions, such as:
  - Fog (FG)
  - Snow Pellets (PL)
  - Snow (SN)
  - Haze (HZ)
  - Blowing Snow (BLSN)
  - Blowing Spray (BLPY)
  - Freezing Drizzle (FZDZ)
  - Freezing Fog (FZFG)
  - Ground Fog (MIFG)
- **Severe Weather:** Encompasses extreme conditions that significantly affect flights:
  - Thunderstorms (TS)
  - Hail (GR)
  - Glaze (GL)
  - Blowing Dust (DU)

- Tornadoes (FC)
- Damaging Winds (WIND)
- Freezing Rain (FZRA)

To integrate this weather data into the flight dataset, we matched each flight with the closest weather record time.

4 Exploratory Data Analysis

4.1 Initial Visualizations

The initial phase of our exploratory data analysis involved generating visualizations to better understand the distribution and relationships within our dataset. We first created a simple histogram of the departure delay that we are trying to predict(Figure 1).

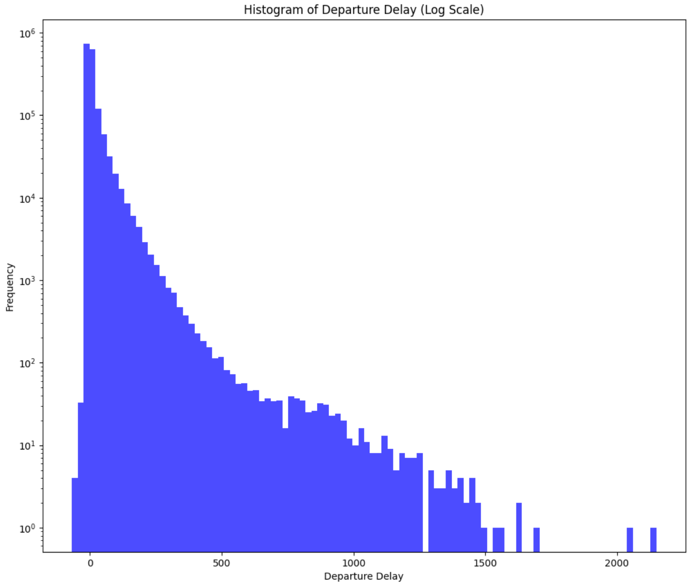

Figure 1: Departure Delay

Binary Classification of Delays: In our exploratory data analysis for binary classification, we selectively used only those features deemed most critical for predicting whether a flight was delayed or not. For this analysis, we randomly sampled 100,000 rows, ensuring a 50/50 split between delayed and non-delayed flights to provide a balanced view. This strategic sampling helped us effectively illustrate the potential predictors of delays, as shown in Figure 2. These pair plots explored the relationships between different variables and the binary status of flight delays, aiding in the identification of significant predictors.

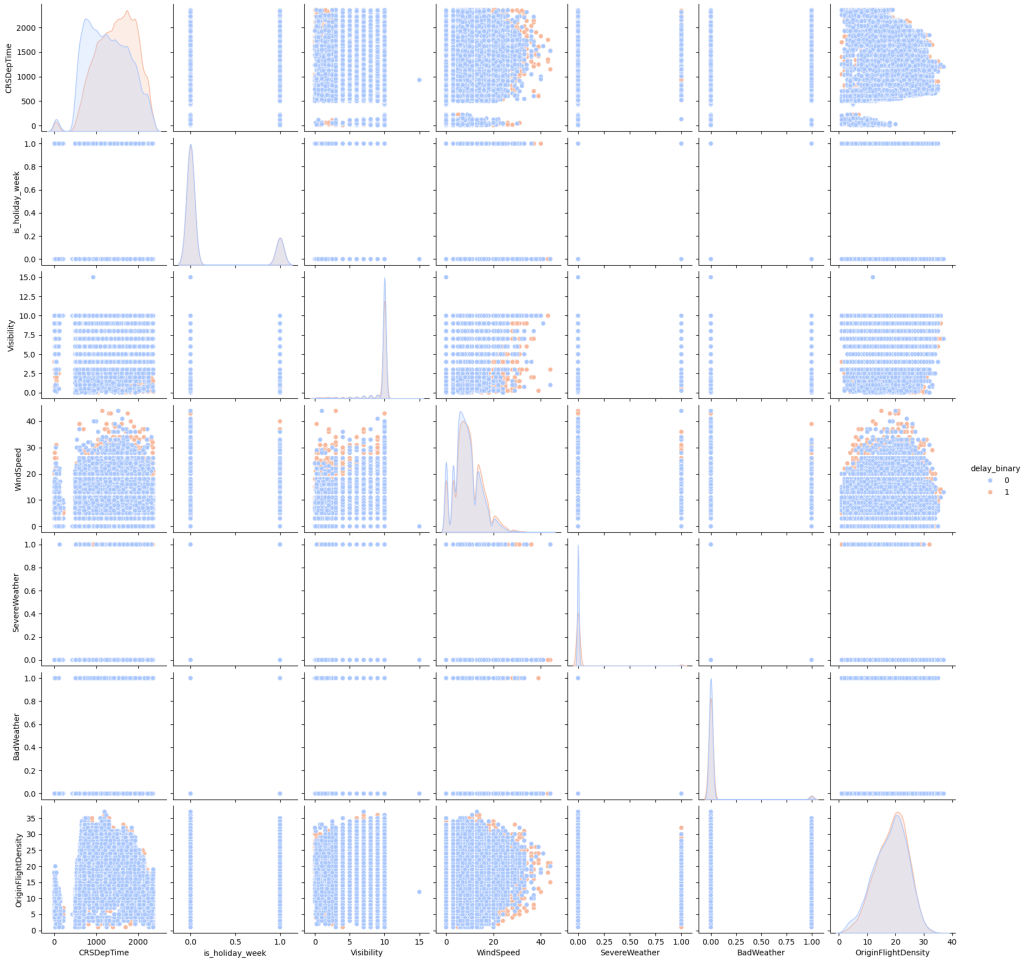

Figure 2: Pairplot of Binary Delay

Delay Duration Intervals: For the interval classification of delays, we focused on pair plots that examined how various features interact with the categorized delay intervals. Specifically, we randomly selected 100,000 rows that were not delayed to specifically analyze the characteristics influencing the duration of delays, depicted in Figure 3. This analysis helped us understand the factors that most significantly affect the duration of delays, thereby aiding in the refinement of our multi-class classification model.

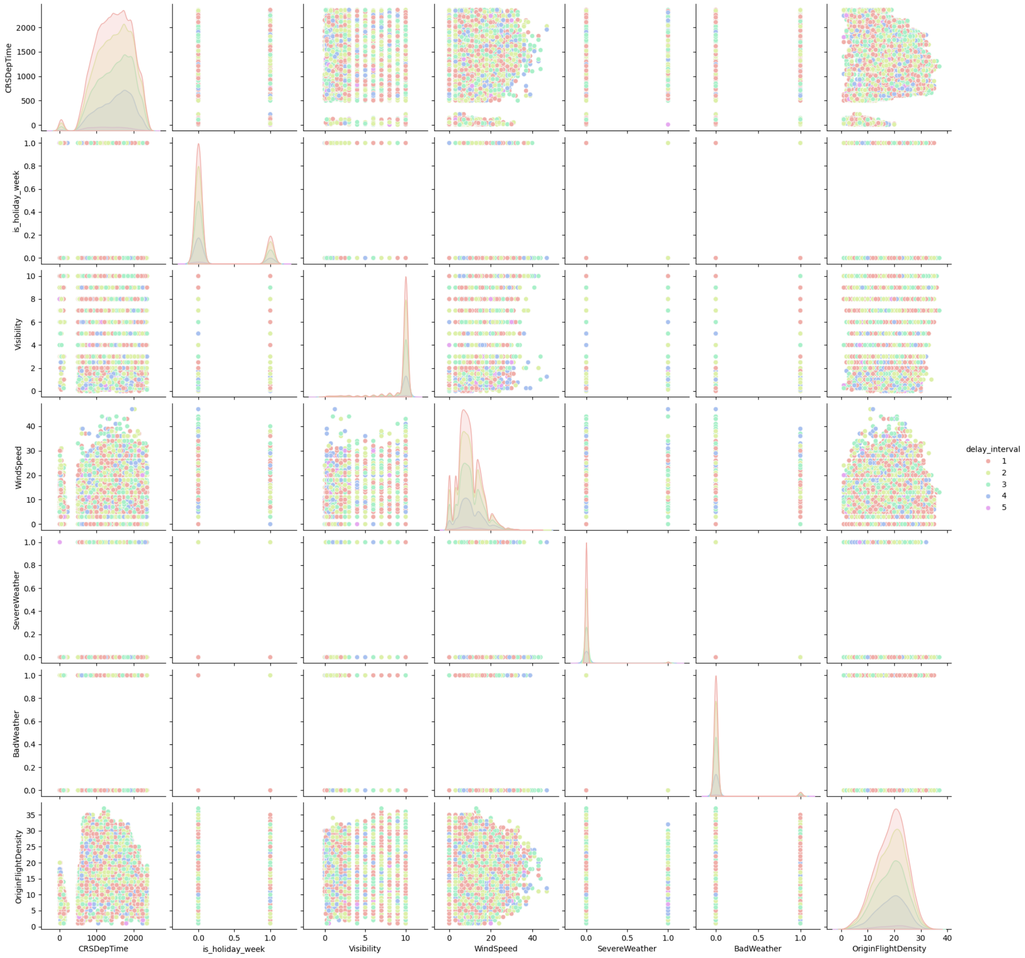

Figure 3: Pairplot of Interval Delays

The initial visualizations have highlighted key patterns and relationships, setting the stage for a deeper exploration of specific factors influencing flight delays.

4.2 Operational Impact on Delays

Analysis indicated that the prevalence of delays varies significantly across different airports and airlines. We found distinct percentages of delays at each airport and for each airline, suggesting that operational factors at specific airports and within certain airlines have a substantial impact on delay duration. The investigation into every combination of airline and airport delays revealed that these combinations are statistically independent of each other. Figure 4 and 5 illustrate these findings.

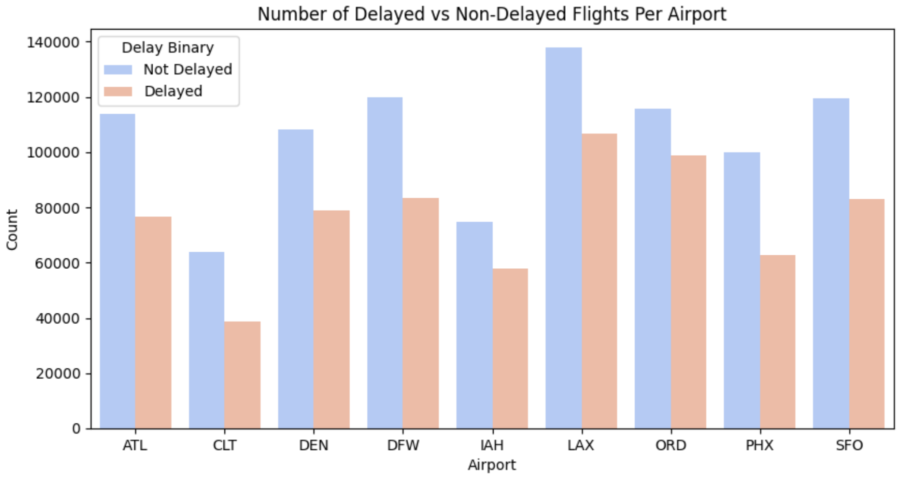

Figure 4: Airport Delays

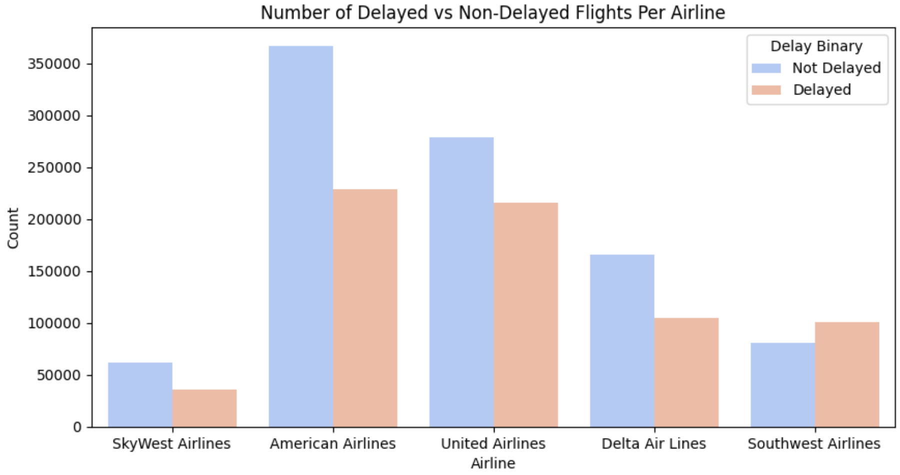

Figure 5: Airline Delays

4.3 Temporal Patterns of Delays

The analysis of delays over time revealed notable fluctuations in the incidence of delays. A time series analysis, as illustrated in Figure 6, demonstrates how delays have evolved over the years, showing a general decline in their frequency. This trend suggests improvements in operational efficiencies or adaptations in scheduling practices. Seasonally, we observe significant spikes in delay occurrences during the summer months, which corroborates the FAA's findings that the majority of delays occur in this season, likely due to increased air traffic and weather disruptions.

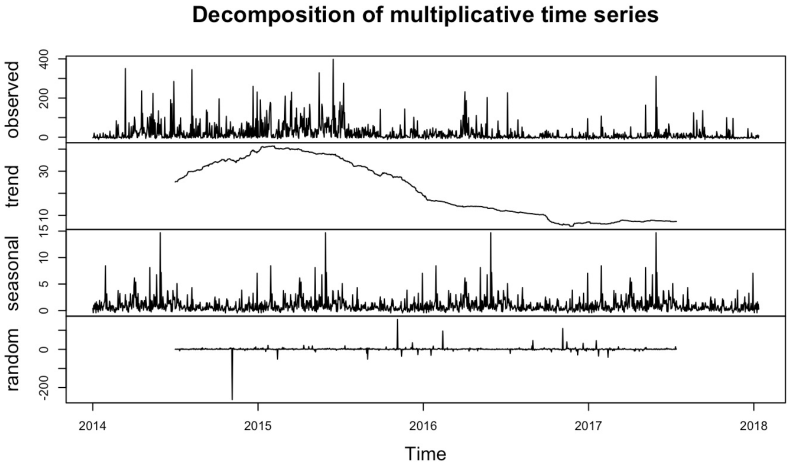

Figure 6: Time Series Analysis of Flight Delays

Moreover, an analysis of hourly delay patterns presented in a histogram (Figure 7) reveals a clear daily cycle. Delays tend to increase throughout the day, peaking between 6 PM and 9 PM when the number of delayed flights exceeds those on schedule. This pattern is attributed to the cumulative effect of earlier delays within the day, commonly known as the 'snowball' effect. The frequency of delays decreases overnight as fewer flights operate, allowing the system to reset by the next morning.

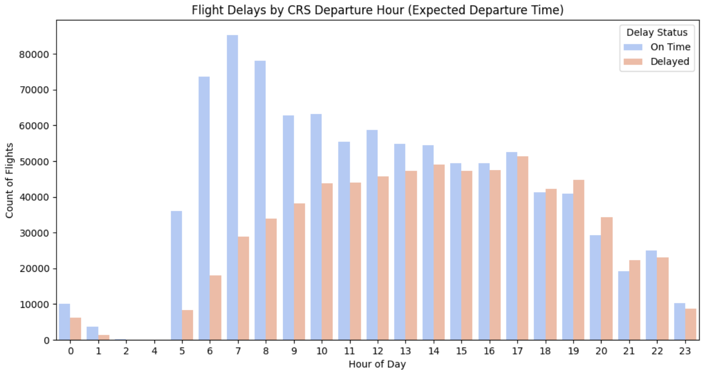

Figure 7: Hourly Distribution of Flight Delays

Weekly Patterns: While our findings also indicated that the day of the week influences the frequency of delays, this factor was less significant compared to the hourly patterns. Figure 8 provides a visual representation of this trend.

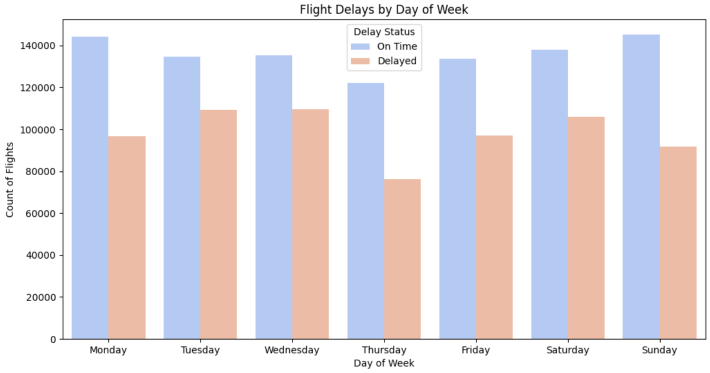

Figure 8: Weekly Patterns of Flight Delays

4.4 Weather Impact on Delays

As previously noted, the FAA reported that around 75% of flight delays from 2017 to 2022 were due to weather conditions. Our analysis supports this finding. As shown in Figure 9, in both binary and multi-class classifications, there is a distinct difference in the proportion of flight delays under different weather conditions.

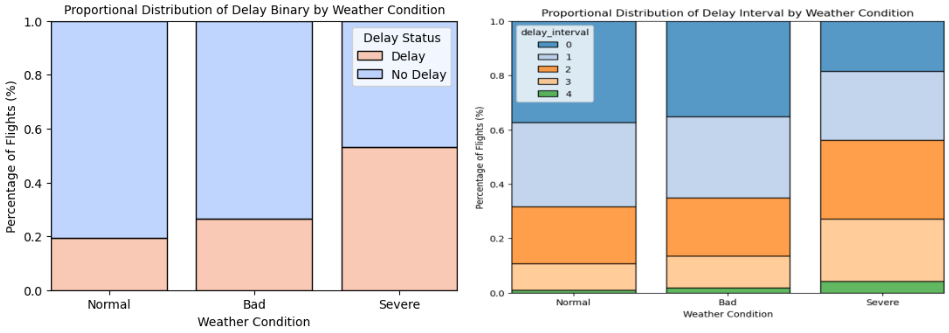

Figure 9: Flight Delay under Different Weather Conditions

The data reveal that 2.8% of flights occur under bad weather conditions and 0.8% under severe weather conditions. When weather conditions are normal, only 20% of flights experience delays. However, this percentage increases to 27% for flights under bad weather. Most notably, under severe weather conditions, the delay rate jumps to 54%.

5 Results

5.1 Model Development

To address the objectives of predicting flight delays, two distinct models were developed. The first model is a binary classifier designed to predict whether a flight will experience a delay. The second model, a multi-class classifier, which predicts the duration of the delay categorized into specific time bins.

5.2 Binary Classification Model

The binary classification model was designed to identify flights that would be delayed by 15 minutes or more, in accordance with FAA definitions. The model development process included initial data preparation, training, and validation. We introduced a binary variable, ‘Delay‘, where ‘1‘ indicates a delayed flight and ‘0‘ signifies no delay, derived from the ‘DepDelay‘ column. To facilitate robust evaluation, the dataset was divided into training and testing subsets, allocating 20% for testing to assess model performance.

5.2.1 Evaluation and Results

The effectiveness of the binary classification models was assessed primarily through their accuracy in predicting whether a flight would be delayed. This evaluation compared the performance of several models using a naive approach as a baseline, followed by more sophisticated machine learning models.

- **Naive Approach:** Initially, a simple prediction model was employed, which achieved an accuracy of 52.61%. This model was straightforward in its implementation, focusing primarily on predicting the occurrence of a delay using the origin and destination only.

- **Advanced Models:**
  - **Logistic Regression:** This model improved upon the naive approach, achieving an accuracy of 79.4%.
  - **Random Forest:** Utilizing an tree-based method provided a further boost, with an accuracy of 79.6%.
  - **XGBoost Classifier:** A gradient boosting framework that offered better performance, attaining an accuracy of 79.8%.
  - **Neural Network:** We utilized a multi-layer perceptron with five hidden layers in our model, each followed by a batch normalization layer and a dropout layer with a dropout rate of 0.5. This configuration proved to be the most effective, achieving an accuracy of 80.6%.

5.3 Multi-class Classification Model for Delay Duration

The initial approach to the multi-class classification model aimed to predict a range of delay durations, starting from no delay (0-15 minutes) to significant delays. However, due to the heavily skewed nature of the initial bin (0-15 minutes), which predominantly captured flights without substantial delays, we revised our approach to focus on more impactful delay durations. This adjustment led to the definition of the following bins:

- Bin 1: 15 - 30 minutes
- Bin 2: 30 - 60 minutes
- Bin 3: 1 - 2 hours
- Bin 4: 2 - 5 hours
- Bin 5: Over 5 hours

5.3.1 Naive Approach

As a preliminary step, we employed a Naive Bayes classifier to establish a baseline for our model's performance. This naive approach was crucial in setting a benchmark for further model development. After fine-tuning the smoothing factor, the classifier achieved an accuracy of 36.98%. Analysis of the feature importance from this model revealed a significant reliance on just two variables: the Destination Airport and the Origin Airport, which accounted for 50.03% and 49.72% of the feature importance respectively. This limited scope in feature utilization partly explains the lower accuracy level. Nevertheless, this model served as a baseline for comparison, providing a clear benchmark against which the effectiveness of more complex models could be measured.

5.3.2 Advanced Modeling Techniques

Following the naive model, we explored a range of more sophisticated machine learning models, each subjected to extensive hyperparameter tuning to optimize their performance. To address the imbalance in the delay duration bins and ensure more robust model training, we implemented an oversampling strategy to achieve more even bin splits. This method was crucial for enhancing the accuracy and generalizability of the models. The best accuracies achieved and the parameters tuned for each model are detailed below (settings that yielded the highest accuracy are highlighted in bold):

- **XGBoost:** Achieved the best accuracy overall of 73.69%. The hyperparameters tuned included:
  - Number of estimators: [50, 100, 200, 300, 1000, **2000**]
  - Maximum depth: [2, 3, 5, 10, **15**]
  - Learning rate: [**0.01**, 0.1, 0.2]
  - Subsample: [0.5, 0.7, **0.9**]
- **Random Forest:** Best accuracy recorded was 56.14%. The hyperparameters tuned included:
  - Number of estimators: [50, 100, 200, 300, 1000, **2000**]
  - Maximum depth: [2, 3, 5, 10, **15**]
  - Minimum samples split: [**2**, 5, 10]
  - Minimum samples leaf: [**1**, 2, 4]
  - Criterion: [gini, **entropy**]

- **Decision Tree:** Achieved an accuracy of 52.84%. The hyperparameters tuned included:
  - Maximum depth: [2, 3, 5, 10, 15, **20**]
  - Splitter: [**best**, random]
  - Minimum samples split: [**2**, 3, 4, 5]
  - Random state: [**None**, 4]
  - Minimum samples leaf: [**1**, 2, 4]
- **K-Nearest Neighbors (KNN):** We tested the model using both normalized and unnormalized data but found that normalization did not significantly affect the performance. The highest accuracy achieved with normalized data was 63.28%. The hyperparameters tuned included:
  - Number of neighbors: [**2**, 5, 10, 20, 50, 100, 250]
  - Algorithm: [**ball_tree**, kd_tree]
  - Leaf size: [2, 4, 5, 10, **30**]
- **Neural Network:** This model reported an approximate best accuracy of 37.64%. Due to its deep-learning nature, specific layers and nodes were configured for performance optimization, beyond typical hyperparameter tuning.

This systematic approach to model enhancement through hyperparameter tuning has allowed us to substantially refine the predictive accuracy and reliability of our multiclass classification models.

The confusion matrix presented in Figure 10 from the tuned XGBoost model provides valuable insights into the model's performance across different delay duration bins. Notably, the model demonstrates particularly high accuracy in predicting delays for bins 3 (2-5 hour delay) and bin 4 (over 5 hours delay). This superior performance may be attributed partially to the oversampling technique used during the model training process, which aimed to balance the dataset and improve model training for these typically underrepresented classes. However, it is also plausible that this high accuracy in predicting longer delays can be explained by the confluence of significant predictive factors such as extreme weather conditions, the occurrence of a holiday, and the specific time window of 6-9 PM. Such conditions, individually or combined, are likely to result in more substantial delays, making them easier to predict using the model. This suggests that while oversampling has helped improve model sensitivity to longer delays, the nature of the events leading to these delays also inherently makes them more predictable.

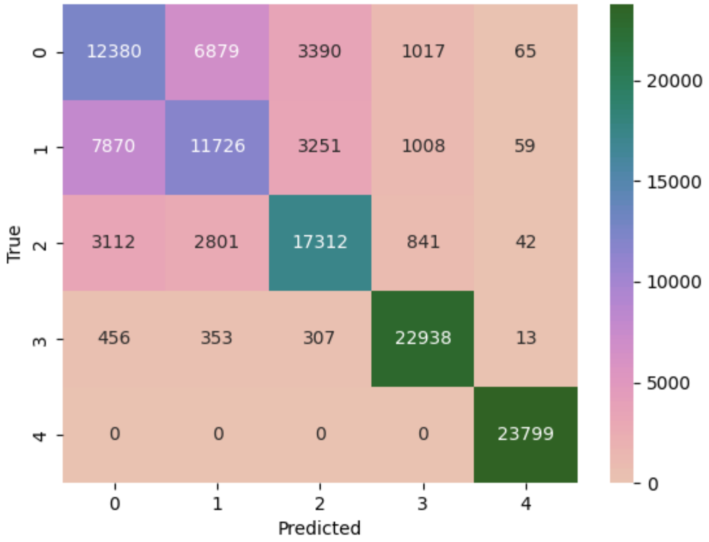

Figure 10: Tuned XGBoost Confusion Matrix

5.3.3 Feature Importance in Advanced Models

Since the XGBoost model demonstrated the highest effectiveness among our advanced models, we will focus on its feature importance analysis. This analysis indicated a diversified set of predictors influencing delay durations. To facilitate a clearer understanding of their collective impact, we grouped these features into relevant categories and calculated the cumulative importance for each group. It's important to note that while we highlight the XGBoost model, the feature importance profiles of the other models were similar in composition.

Weather-Related Features: Features that directly relate to weather conditions, which have a substantial impact on flight delays as highlighted by the FAA.

- SevereWeather: 19.39%
- BadWeather: 6.23%
- Visibility: 5.03%
- WindSpeed: 4.85%

Total Importance: 35.50%

Temporal Features: Time-related aspects that affect flight scheduling and operations.

- CRSDepTime: 5.46%
- is_holiday_week: 5.10%
- DayofMonth: 5.08%
- Year: 5.06%
- Month: 5.02%
- DayOfWeek: 4.83%

Total Importance: 30.55%

Operational Features: Variables related to the specific operational conditions of flights.

- Origin: 6.43%
- Reporting_Airline: 6.02%
- Distance: 5.86%
- TotalDensity: 5.37%
- CRSElapsedTime: 5.14%
- Dest: 5.10%

Total Importance: 33.92%

The feature importance analysis from the XGBoost model highlights the significant predictors of flight delays, such as weather, temporal, and operational factors. Grouping these features demonstrates their collective impact and guides more effective predictive modeling. This insight is crucial for improving airline delay management strategies, emphasizing the integration of these critical factors into operational planning.

5.3.4 Impact of Newly Added Features

Looking at the features we added to the original data, the addition of specific weather-related and operational features has significantly enhanced the model's predictive capabilities. The breakdown of their importance is as follows:

- **SevereWeather:** 19.39%
- **BadWeather:** 6.23%
- **TotalDensity:** 5.37%
- **is_holiday_week:** 5.10%
- **Visibility:** 5.03%
- **WindSpeed:** 4.85%

Total Importance of Newly Added Features: 45.97% — Highlighting their substantial role in enhancing delay predictions.

5.4 Key Findings

The comprehensive analysis of flight delay predictions through both binary and multi-class classification models has yielded several pivotal findings:

- The binary classification model effectively predicted whether a flight would be delayed with a high degree of accuracy, especially when advanced models like XGBoost and neural networks were applied.

- The multi-class classification model demonstrated its strength in predicting longer delays with considerable accuracy, particularly for delays over two hours, which was significantly improved through techniques like oversampling.
- Feature importance analysis from the XGBoost model confirmed that weather conditions are the most influential predictors of flight delays.
- The performance enhancements from hyperparameter tuning underscored the critical nature of model optimization in achieving higher accuracy in delay prediction.

5.5 Implications

The insights derived from the predictive models have profound implications for the aviation industry:

- Enhanced predictive accuracy allows for better resource allocation, including gate assignments and crew scheduling, potentially reducing operational costs and improving efficiency.
- By accurately forecasting delay durations, airlines can more effectively communicate with passengers, potentially improving customer satisfaction and reducing the stress associated with travel disruptions.
- The identification of significant predictors such as severe weather and peak operational times provides airlines with actionable intelligence to preemptively manage and mitigate delays.
- Airlines and airports can use these insights to revise their strategies during high-risk periods, such as during severe weather forecasts or high traffic volumes, to minimize the cascading effects of delays.

These implications suggest that the integration of advanced predictive models into daily operations could revolutionize how delays are managed, offering a clearer pathway toward more resilient and responsive airline operations.

6 Conclusions and Discussions

Our comprehensive analysis and application of machine learning techniques have yielded significant insights into the prediction of flight delays. The binary and multi-class classification models developed in this project have effectively quantified the impact of various predictors, such as weather conditions, airline operations, and temporal factors. These models achieved high accuracy levels, particularly in scenarios involving complex delay patterns, validating the efficacy of advanced analytical techniques in addressing real-world problems in aviation logistics.

The implications of this research are far-reaching, offering airlines and airport authorities actionable strategies to enhance scheduling, improve customer communication, and optimize resource management. By implementing these predictive models, the aviation industry can achieve higher operational efficiency, reduce economic losses, and provide a more reliable travel experience for passengers.

7 Future Work

The successful development and implementation of predictive models for flight delays pave the way for several exciting avenues of future research and enhancement. The following objectives are proposed to build on the current findings and extend the applicability of our models:

1. **Enhanced Weather Data Integration:** Improve the precision of delay predictions by incorporating detailed weather data that reflects the exact conditions expected at the airport at the time of the scheduled departure. Further, modeling the duration of weather events could refine predictions of delay times based on these factors.

2. **Inclusion of International Flights and Airports:** Expand the dataset to encompass international flights at international airports to provide insights into global delay dynamics and their unique causes.

3. **Increased Computing Resources:** Obtain more powerful computing resources to process the full dataset rather than a reduced version. This enhancement will allow for more robust data analyses, potentially uncovering subtler patterns that could improve predictive accuracy.

4. **Development of a User Interface:** Design and implement a user-friendly interface that allows airlines, airport staff, and passengers to access real-time predictions of flight delays. This tool would be invaluable in managing operational decisions and passenger expectations effectively.

5. **Exploration of Advanced Modeling Techniques:**

   - Investigate the application of deep learning techniques to detect complex patterns within the data that traditional models may overlook.
   - Explore a graph-based approach to model the cumulative effects of delays, such as the dynamics of aircraft traffic into and out of airports, to better understand and mitigate the cascading impacts of initial delays.

These initiatives aim to not only refine the accuracy and reliability of our predictive models but also to enhance their practical utility in real-world scenarios, thereby contributing to improved operational efficiency and passenger satisfaction in the aviation industry.

8 References

Data Sources:

1. AWS Academy. (2024). *US Domestic Flights Delay Prediction (2013 - 2018).* Retrieved from Kaggle: `https://doi.org/10.34740/KAGGLE/DSV/7423794`

2. National Oceanic and Atmospheric Administration, National Centers for Environmental Information. (n.d.). Climate Data Online (CDO). Retrieved from `https://www.ncdc.noaa.gov/cdo-web/`

License:

1. The dataset from Kaggle is licensed under CC0: Public Domain, permitting unrestricted use in research and analysis.