To: Delta Air Lines
From: Iris Brook, Lucy Liu, Krishanu Datta, Devin Wasilefsky
Date: December 8, 2023
Subject: Decoding Delta: A Comprehensive Exploration of Flight Delays and Cancellations

INTRODUCTION

Predicting flight delays and cancellations is an increasingly critical aspect of the aviation industry. These disruptions not only cause inconvenience to passengers but also lead to significant operational and financial challenges for airlines. Understanding and anticipating these disruptions is crucial for enhancing passenger experience and optimizing airline operations. This report presents a detailed investigation into Delta Air Lines' flight delays and cancellations throughout 2021. Our study delves deep into the factors contributing to these operational disruptions.

The primary aim of our research is twofold. Firstly, we seek to augment the precision in predicting flight delays and cancellations. Accurate predictions are essential for airlines to implement proactive measures, thereby mitigating the impact of these disruptions. Secondly, we aim to gain an in-depth understanding of the challenges associated with specific airports, different days of the week, and various time segments. Such an understanding is pivotal for tailoring strategies that are not just generic but also contextually relevant and effective.

In pursuit of these objectives, our study distinguishes itself by significantly emphasizing feature generation. Unlike prior research that predominantly focuses on general factors, we delve into more nuanced elements like flight congestion, the status of airports as central hubs, and the proximity of flights to major holidays. This study holds the potential to benefit not just Delta Air Lines but the broader aviation sector by setting a precedent for data-driven, feature-focused predictive analysis.

METHODLOGY

Our research utilizes a comprehensive dataset obtained from Kaggle. This dataset consists of millions of rows and encompasses 61 features. This data provides crucial information such as the flight's origin, destination, cancellation status, delay indicators, departure and arrival time blocks, among others. Specifically, we extracted data from 2021, allowing for a focused analysis of Delta Airlines' flight operations during that specific timeframe.

We decided to focus on feature generation to differentiate this study from previous research on this subject. We systematically generated new features to capture temporal nuances, introducing variables around holidays. Specifically, we created binary variables signaling whether a flight coincided with specific holidays like Christmas and Memorial Day, recognizing the potential impact of these occasions on flight operations. Furthermore, features denoting whether the flight originated or terminated at a Delta hub were created, along with flight congestion metrics for each airport on a given day. This feature engineering process aims to give our models a better understanding of the factors influencing flight disruptions.

Our analysis employs diverse machine learning models to classify flights into delayed or canceled categories. We utilized Extreme Gradient Boosting (XGBoost), Random Forest, Naive Bayes, Logistic Regression, and Support Vector Machines (SVM). Each model contributes unique strengths to the classification task. XGBoost and Random Forest handle complex relationships and non-linearities, making them ideal for discerning patterns in flight data. Naive Bayes offers simplicity, which is beneficial when dealing with a large dataset with many features. Logistic Regression provides a probabilistic framework for classification, offering interpretability and it is easy to implement. With its ability to handle high-dimensional data, SVM is valuable when the feature space is extensive. Using all of these models allows for a comprehensive exploration of the dataset, leveraging the strengths of each algorithm to achieve accurate classification results.

RESULTS AND ANALYSIS

We primarily assessed our models' performance based on accuracy, as presented in the accuracy table (Figure 1). We observed a significant difference between the best and worst-performing models, while the others exhibited relatively similar accuracies. Specifically, the XGBoost model demonstrated the highest accuracy at approximately 0.724, while the Naive Bayes model yielded the lowest accuracy at 0.591.

The superiority of XGBoost in performance can be attributed to its capability to handle complex relationships within an extensive feature set. Given that our dataset contains intricate patterns and dependencies among features, XGBoost effectively captures these complexities. In contrast, Naive Bayes assumes independence among features, limiting its efficiency in scenarios with complex interdependencies. The stark contrast in performance between XGBoost and Naive Bayes underscores the significance of leveraging models capable of capturing intricate relationships within the data. While Naive Bayes might suffice for more straightforward, less interconnected datasets, the complexities of our data demand a more nuanced approach for predictive accuracy, making XGBoost the optimal choice.

Considering XGBoost is the best performance model, we also intended to explore the feature importance that contributed the most to the model. As shown in Figure 2, Christmas was the most important feature, while Month and Quarter also appeared relatively high.

Our analysis encompassed predictions across crucial variables such as airports, days of the week, days of the month, arrival time block, and whether or not an airport was a hub, revealing intriguing insights. Notably, 'XNA,' 'MSN,' 'DSM,' 'TLH,' and 'JAN' emerged among the top 5 airports, exhibiting the highest accuracy in predictions. Conversely, 'EYW,' 'GNV,' 'MTJ,' 'PWM,' and 'MDT' represented airports with comparatively lower predictive accuracy, warranting closer scrutiny of the unique factors influencing these locations.

Examining predictions over a week, Tuesday displayed the highest accuracy, showcasing an impressive rate of approximately 0.769. Similarly, within the days of the month, the first day remarkably stood out, boasting an accuracy of roughly 0.765. Such distinct patterns indicate the potential influence of specific weekdays or dates on flight disruptions. This is potentially tied to increased passenger volumes and operational factors.

In terms of the arrival time block, block 4 emerged as the segment demonstrating the highest accuracy in our predictions. This finding suggests that flights scheduled within this time frame exhibit more predictable patterns regarding disruptions, potentially influenced by factors such as air traffic conditions or operational efficiency.

Furthermore, intriguing patterns emerged after scrutinizing the accuracy results for Delta hubs. Our model showcased superior performance in predicting flights that did not originate from a Delta hub. Conversely, it displayed a higher accuracy in predicting flights going into a Delta hub. Such different accuracies imply varying complexities and dependencies associated with flight origins and destinations, warranting a deeper exploration of the underlying factors driving these predictions.

DISCUSSION AND RECCOMENDATIONS

Our study identified a subset of airports - XNA (Northwest Arkansas National Airport), MSN (Dane County Regional Airport), DSM (Des Moines International Airport), TLH (Tallahassee International Airport), and JAN (Jackson-Medgar Wiley Evers International Airport) - where there were less flight disruptions. The reasons behind this predictability could be multifaceted, ranging from less complex flight schedules and lower passenger volumes to more efficient airport operations. These findings suggest that operational strategies successful in these airports should be modeled and applied to more challenging locations.

Conversely, airports such as EYW (Key West International Airport), GNV (Gainesville Regional Airport), MTJ (Montrose Regional Airport), PWM (Portland International Jetport), and MDT (Harrisburg International Airport) presented substantial challenges and had many more flight disruptions. The complexity at these airports might be attributable to various factors, including geographical location, airport infrastructure, or specific operational constraints. This insight calls for a more targeted approach to handling operations at these airports, involving enhanced resource allocation and tailored contingency planning.

Our analysis also revealed that predicting disruptions for flights scheduled on Fridays and Sundays was particularly challenging. These days, we see higher passenger traffic and tighter schedules, increasing the likelihood of cascading delays. This pattern underscores the need for Delta to adopt more robust and flexible operational strategies during these peak travel times.

Additionally, we observed that there were more predicted disruptions for flights scheduled later at night than morning flights. This is due to the cumulative effect of daily delays, which tend to exacerbate by night. Hence, focusing on maintaining punctuality earlier in the day will help mitigate the cascading effect of delays into the night.

Furthermore, our study suggests that enhanced investment in predictive analytics could be a game changer for Delta. By refining prediction models and incorporating more nuanced data specific to challenging airports and times, Delta can significantly improve its preemptive strategies against potential disruptions. The study highlighted the importance of reviewing and adapting flight scheduling practices. Delta could reduce the likelihood of disruptions by avoiding scheduling flights during peak congestion times and distributing flights more evenly throughout the day.

CONCLUSION

Our data-driven investigation into Delta Air Lines' flight delays and cancellations in 2021 has produced significant insights that can revolutionize predictive modeling within the aviation industry. Moreover, by leveraging the power of a feature-oriented analysis, we not only made strides toward increasing the accuracy of current iterations of delay analysis, but also uncovered various factors that play a crucial role in current industry disruption patterns.

Unlike previous research that primarily focused on general factors, we delved into nuanced elements such as flight congestion, airport hub status, and the proximity to major holidays. This approach gave us a more detailed understanding of the factors influencing flight disruptions. Notably, the performance of the machine learning models we employed, with XGBoost leading the way, highlighted the effectiveness of this strategy in improving predictive accuracy.

Our research contributes to the aviation field by deepening the understanding of the factors disrupting air travel. By identifying specific airports where prediction is easier and recognizing challenges associated with particular days of the week and times, we provide actionable insights for airlines. Our work encourages a data-driven approach to handling delays and cancellations, emphasizing the importance of airport-specific and temporal factors in operational strategies.

Looking ahead, we could dig deeper into how airports with higher disruptions run their day-to-day operations, to find additional causes for their disruptions. Moreover, referencing external data, such as weather information and global events, may further improve our predictions, as they may serve to impact flight plans significantly. Finally, we may try various models paired with more real-time data, as this could also lead to more accurate and detailed predictions.

The importance of our research extends beyond Delta Air Lines, setting a precedent for data-driven, feature-focused predictive analysis in the broader aviation sector. In managing and forecasting operational uncertainties, our study highlights the application of predictive analytics to the airline industry as one with overwhelming potential. By adjusting scheduling practices, distributing flights more evenly, and investing in targeted strategies for airports burdened by disruptions, airlines can improve passenger experience, address operational challenges, and conquer financial struggles. This research is a foundational step towards a more resilient and adaptive aviation industry equipped to navigate the complexities of modern air travel.

APPENDIX

	Accuracy
Logistic Regression	0.708386
Random Forest	0.713459
Naive Bayes	0.591443
Support Vector Machine (SVM)	0.704056
XGBoost	0.723925

Figure 1: Model Accuracies

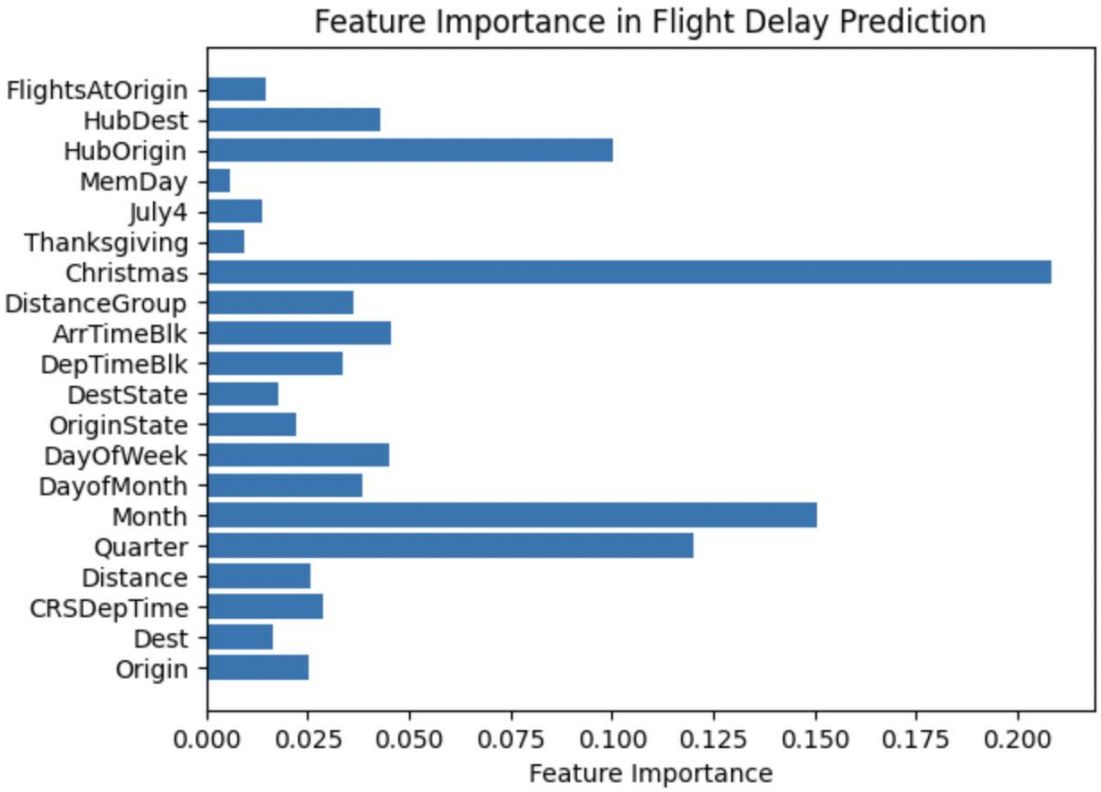

Figure 2: Feature Importance