Flight Delays and Cancellations Project Executive Summary

Introduction

The holiday season, spanning November to January each year, is one of the busiest periods for the airline industry. Increased passenger traffic during this time places heightened demand on airline operations, resulting in greater strain and higher rates of flight delays and cancellations. These disruptions can have a significant impact on travelers, causing inconvenience and financial costs. By identifying critical patterns in delays and cancellations, passengers can make informed decisions, anticipate potential disruptions, and plan their journeys with greater confidence.

This report aims to explore key trends and factors influencing flight delays and cancellations during the holiday season, using data from the U.S. Department of Transportation and the U.S. National Weather Service. Through data analysis and pattern recognition, we examine the primary causes of delays, identify high-risk periods and destinations, and assess the impact of weather-related conditions. Additionally, this report includes the development of a predictive model designed to estimate the likelihood of flight delays and cancellations. By building a reliable model, we aim to provide insights that can help both airlines and passengers proactively manage the challenges of holiday travel, ultimately enhancing the travel experience during one of the most demanding times of the year.

Data Cleaning Process

To prepare the flight data for analysis, we undertook a thorough data cleaning and preprocessing process aimed at building a structured and reliable dataset for examining flight delays and cancellations during the busy holiday season. The process began with loading the data, followed by essential transformations and cleaning tasks. We removed unnecessary and redundant columns to streamline the dataset and focus on relevant information, and carefully filled in missing details, such as time zone information, to ensure consistency across all records.

One major challenge involved time zone discrepancies between origin and destination airports. To address this, we standardized all arrival and departure times to Coordinated Universal Time (UTC), which was crucial for maintaining chronological accuracy, especially when analyzing delays across different time zones. Additionally, we integrated weather data with the flight data, initially matching weather records to flights using IATA codes (three-letter airport identifiers) and, where necessary, using ICAO codes (four-letter identifiers) to ensure complete weather coverage.

A notable limitation arose with Kapalua–West Maui Airport (IATA: JHM), a regional airport in Hawaii, for which no METAR weather data was available. After exploring alternatives, we decided to exclude observations involving JHM to preserve data quality. Through this systematic data cleaning and integration process, we established a high-quality dataset with consistent temporal alignment and complete weather information, providing a strong foundation for analyzing patterns in flight delays and cancellations, and offering valuable insights to improve holiday travel planning

Exploratory Data Analysis

Our Exploratory Data Analysis (EDA) revealed several notable trends in flight delays and cancellations. Among the primary reasons for cancellations, weather-related cancellations represented a significant portion, with the cancellation rate due to bad weather being approximately four times higher than cancellations for other reasons. This finding highlights the substantial impact of adverse weather on flight disruptions, particularly during peak travel times. By comparing cancellation patterns on days with poor weather to regular days, we confirmed a strong association between adverse weather conditions and higher cancellation rates.

Our classification of airports by size further revealed that larger airports, with higher flight volumes, generally experienced more cancellations and delays than smaller airports. This suggests that operational strain at high-traffic airports may contribute to increased disruptions. Additionally, an analysis of the time of day showed that flights scheduled at night had a notably higher cancellation rate than those scheduled at other times, likely due to operational constraints and limited recovery options for late flights.

We also tested logistic regression models to identify which variables were most predictive of cancellations. Factors such as airport size, weather conditions, and time of day emerged as strong predictors, supporting the predictive model's ability to estimate cancellation probabilities based on these key variables. These insights provide valuable guidance for both travelers and airline operators, allowing them to anticipate and prepare for potential disruptions.

The EDA phase not only uncovered key trends in flight delays and cancellations but also validated essential features for our predictive model. This groundwork enables us to accurately forecast disruptions in air travel during peak holiday seasons, offering a data-driven approach to enhancing the holiday travel experience for passengers and improving operational efficiency for airlines.

Final Model Selection and Rationale

The final logistic regression model predicts flight cancellations using a combination of operational and weather-related variables. Key predictors include:

- Operational Factors: Airline carrier, National Air System (NAS) delays, and security delays capture airline-specific operational patterns and broader system disruptions. These factors help identify unique trends in cancellations tied to specific carriers and national air traffic conditions.
- Weather Conditions: Variables like temperature, relative humidity, wind speed, and visibility are included due to their impact on flight safety and scheduling. Extreme temperatures, high winds, or low visibility significantly increase cancellation likelihood. Additionally, flights at night showed higher cancellation rates, likely due to limited rescheduling options.

The model results show that both operational factors and weather conditions significantly affect cancellation probabilities, providing a data-driven tool to anticipate and manage disruptions. This model offers valuable insights for travelers and airlines, helping them better understand and respond to cancellation risks.

Statistical Analysis and Key Metrics

The logistic regression model was evaluated using a confusion matrix and key performance metrics to assess its effectiveness in predicting flight cancellations. The model achieved an **overall accuracy** of 95.41%, with a 95% confidence interval of (94.88%, 95.9%), indicating strong predictive performance.

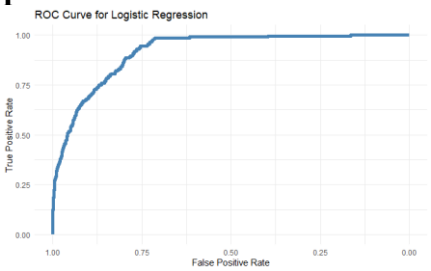
In terms of **sensitivity**, the model correctly identified 97.17% of non-cancelled flights, demonstrating high reliability in predicting flights that would proceed as scheduled. However, the **specificity** was 42.86%, indicating a moderate ability to correctly identify cancelled flights. The **positive predictive value** (PPV) was 98.07%, showing that most of the flights predicted to be non-cancelled were indeed on schedule, while the **negative predictive value** (NPV) was 33.7%, reflecting the model's more limited capacity to predict actual cancellations.

The model's **balanced accuracy** of 70.01% provides a more comprehensive measure of its performance by balancing sensitivity and specificity, making it a reliable tool for anticipating flight disruptions. Additionally, McNemar's test yielded a **p-value** of 0.0009322, suggesting a statistically significant difference between the model's predictions and a random classifier,

underscoring the model's effectiveness in predicting flight status.

These metrics demonstrate that the model is highly accurate in predicting non-cancellations, making it a valuable resource for travelers and airlines to manage expectations and plan accordingly. However, the lower specificity suggests that further refinement could improve its accuracy in predicting actual cancellations.

Model Diagnostics and Assumptions



To evaluate the logistic regression model's performance, we examined key metrics, the ROC curve, and underlying assumptions. The ROC curve shows strong model discrimination, with a high area under the curve (AUC) indicating effective separation between cancelled and non-cancelled flights. The confusion matrix metrics reveal high overall accuracy (95.41%) and sensitivity (97.17%) for predicting flights likely to proceed, though specificity (42.86%) for detecting actual cancellations is moderate.

The model assumes a linear relationship between predictors and the log odds of cancellation. We included carefully selected operational and weather-related variables (e.g., carrier type, NAS delay, temperature, visibility, wind speed) that are known to impact flight disruptions. This selection aligns with domain knowledge, ensuring that our model captures significant predictors of cancellations. Overall, these diagnostics confirm that the model is reliable and effective for predicting flight cancellations, particularly in identifying flights likely to proceed.

Strengths and Weaknesses of the Model

The logistic regression model for predicting flight cancellations shows strong overall accuracy, particularly in identifying flights likely to proceed as scheduled. By incorporating both operational and weather-related variables, it provides a well-rounded view of the factors that impact cancellations, making it useful and interpretable for decision-making.

However, the model's ability to correctly identify actual cancellations is moderate, which may limit its effectiveness in fully anticipating disruptions. Additionally, the linear assumption of logistic regression may not capture complex interactions between factors, suggesting that further refinement with more advanced models could enhance its predictive accuracy.

Conclusion

This project analyzed factors influencing flight delays and cancellations during the holiday season through data cleaning, exploratory analysis, and predictive modeling. The logistic regression model effectively predicts non-cancellations, providing valuable insights into key operational and weather-related factors. While there is room to improve its ability to identify actual cancellations, the model serves as a useful tool for airlines and passengers to better manage holiday travel challenges and enhance planning during peak travel times.

Contributions

Parts	Hengyu Yang	Yi Ma	Leyan Sun	Tianle Qiu
Presentation	App demonstration	Responsible	Model part	Intro part
Summary	Review and some advice	Responsible	Check and add some parts	Format
Code	Responsible Data download, clean and merge	Some advice	Fitting model, exploratory analysis	Some details
Shiny App	Responsible	Some advice	Format adjustment	Some advice