STAT 628 Module 3 - Flight Cancellation and Delay: Executive Summary

Introduction

Airline delays cost airlines multi-billion dollars per year and cause a great inconvenience to passengers. During the holiday season, the number of flights increases significantly, but then there are even more weather extremes that can exist in the winter. According to the Bureau of Transportation Statistics (BTS), approximately twenty percent of the entire scheduled commercial flights are delayed. Weather is not only one of the main reasons for delays but it is also closely related to other delay causes[1]. Rule-of-thumb flight models allow people to easily predict the departure time of the flight and if the flight will be canceled or not. This summary describes the development of a flight delay and cancellation model during the holiday season.

Data Extraction and Processing

For raw flight data, we selected all flight data from top 100 airports by 2023 passenger volume[2], as well as Madison, during November, December, and January for 2018-2024. Considering the Covid-19, the data from Winter 2020-2021 are excluded. Next, we looked for weather data based on the geographic location of the 101 airports.

Based on the coordinates of the airports and the weather stations, we match the nearest weather station for each airport. We selected report type FM-15 to obtain hourly weather data and specific weather data columns. Among these columns, visibility,wind speed, and daily snow depth and snowfall are further used in our model.

For imputation, we replaced missing values in precipitation with 0, treating both missing values and 'T' (indicating trace precipitation) as zero precipitation. Similarly, we imputed missing values in wind speed and snowfall with 0, assuming that the lack of recorded values indicates no precipitation, snowfall, or wind. For visibility data, we initially applied time series interpolation. For sections with extended missing values and instances of variable visibility, we removed them from the dataset, as they constitute only 0.4% of the data. Additionally, their cancellation rate of 1.4% is comparable to the overall cancellation rate of 1.9%.

We categorized visibility and wind speed into three categories (0/1/2) to balance simplicity and accuracy. The classification thresholds were determined by selecting values that maximized the distribution difference between canceled and non-cancelled datasets. Snowfall and snow depth are categorized into 2 categories, using 0 as threshold.

Flight data was joined with weather data at both origin and destination on 3 keys: flight date, scheduled departure hour, and airport code. Before joining the data sets, the temporal columns of the flight and weather data were converted to CST and times were rounded down to the nearest hour, because the previous hour's weather report would be available at the scheduled departure time. We also created 2 columns to represent days until and days after a US holiday.

Our final dataset includes 7 million flight records with corresponding weather data, from 101 airports and 10 airlines, featuring a cancellation rate of 1.88% and a 34% rate of late arrivals.

Exploratory Data Analysis (EDA) & Recommendations

Based on our analysis, the following recommendations are suggested to avoid cancellation: 1. Avoid airports with high cancellation rates and go to other airports nearby- for example, book tickets in ORD instead of MDW (Fig.1). 2. Avoid flying in January, which has a higher cancellation rate than other months (Fig. 2). This may be due to more snowfall occurring in January. 3. Fly Delta - they have the lowest cancellation rate (about 1%) of the 4 major airlines.

Here are some tips to arrive on time: 1. If you need to make a connection, it's best to allow an hour or more. About 29% of flights may be delayed up to sixty minutes (Fig.4). 2.

Avoid flights at night - less than 60% have no delay. Instead take early morning flights (Fig.3). 3. Fly Delta - they have the lowest average delay time.

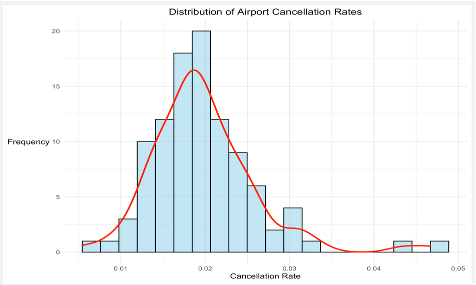

Figure1. Distribution of cancellation rates for different airports

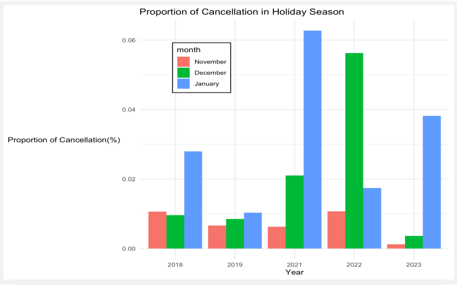

Figure2. Cancellation rates in different months

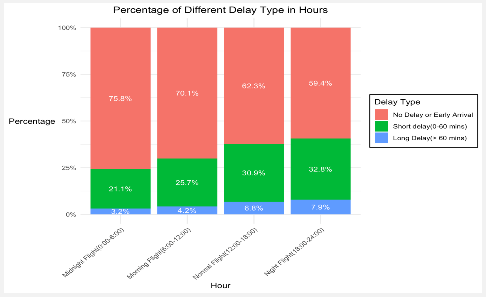

Figure3. Percentage of delay classifications for different hours

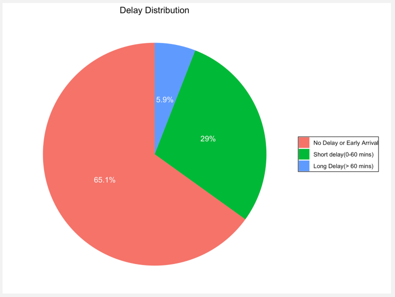

Figure4. Percentage of different delay classifications

Model Selection

For both the cancellation model and delay model, we chose **logistic regression** and compared it with other models.

Cancellation:
$$P(\text{Cancelled}=1)=\frac{1}{1+e^{-(\beta_0+\beta_1\cdot\text{Month}+\beta_2\cdot\text{Vis\_ORIGIN}+\beta_3\cdot\text{Wind\_DEST}+\beta_4\cdot\text{Snow\_ORIGIN}+\beta_5\cdot\text{Snow\_DEST})}}$$

We focused on **weather-related cancellations**. The model is a **binary classification model**, where the response variable is whether the flight is canceled (1) or not canceled (0). The model used 5 features: month, visibility at origin airport, wind speed at destination, daily snow depth at origin and destination.

Delay:
$$P(\text{delay\_category}=k|X)=\frac{e^{\beta_{0,k}+\beta_{1,k}\cdot\text{Month}+\beta_{2,k}\cdot\text{Hour}+\beta_{3,k}\cdot\text{Vis\_ORIGIN}+\beta_{4,k}\cdot\text{Wind\_DEST}+\beta_{5,k}\cdot\text{Snow\_ORIGIN}+\beta_{6,k}\cdot\text{Snow\_DEST}}}{1+\sum_{i=0}^{2}e^{\beta_{0,i}+\beta_{1,i}\cdot\text{Month}+\cdots+\beta_{6,i}\cdot\text{Snow\_DEST}}}$$

For predicting flight delays, we use a **three-class classification model**: on time or early (0), under 1 hour delay (1), over 1 hour delay (2). The model used 6 features: month, schedule departure hour, visibility at origin airport, wind speed at destination, daily snowfall at origin and destination.

Unlike previous studies that use one 15-minute/60-minute threshold for binary classification [3,4], we believe this approach is more realistic and better captures the different levels of delay. For class (0) , there's no need to worry (early or on-time); for class(1), passengers need to hurry to catch connecting flights; for class (2), there's a required rescheduling or significant planning adjustments. We believe this categorization might be more helpful for passengers.

For model training, we first perform undersampling on both models to balance the dataset to a 1:1 or 1:1:1 ratio. Since the dataset is highly imbalanced, even with weighting, the model tends to predict all cancellations/delays or no cancellations/no delays. The dataset for both models was split into 80% training and 20% testing.

The motivation of choosing logistic regression is based on the consideration of simplicity, interpretability and computational complexity of large datasets. We evaluated the models using accuracy and F1-score. As F1-Score is a harmonic mean of recall and precision, it can comprehensively describe the model's ability to identify positive examples.

The model performance is listed in the table below. The accuracy and F1 value of the logistic regression are comparable to those of other methods such as decision trees and random forests. Since logistic regression is simple and has good interpretability, it was selected as the final model.

Cancel Model	accuracy	F1_score	predictors
Logistic Regression	0.753	0.757	5
Decision Tree	0.762	0.762	5

Table1. Cancellation Model Comparison

Delay Model	accuracy	F1_score	predictors
Logistic Regression	0.412	0.395	6
Decision Tree	0.420	0.411	6
Random Forest	0.417	0.403	6
Gradient Boosting Machine	0.413	0.386	6

Table2. Delay Model Comparison

Strengths and Weaknesses

One of the major strengths of both the cancellation model and the delay model is the limited number of inputs required to generate a prediction, meaning a user does not need to know a lot of details about the flight to gain value from the Shiny app. Additionally, both models use logistic regression, which means that they are highly interpretable. It is clear to see how changing each individual feature would influence the prediction result.

A major weakness is the limited dataset used - our analysis only used data from the top 100 airports by passenger volume, but there are roughly 1000 passenger airports in the United States. Many mid-sized regional airports similar to Madison were not considered by our analysis, meaning our recommendations and model have limited scope. Another weakness is that the models do not consider airlines as a feature, despite trends in cancellation rates per airline found during EDA. We attempted to use airline as a feature, but found it wasn't a significant variable in our models.This may be caused by either confounding between an airline's hub airports and typical snowfall for those airports, or the limited number of airports considered.

Conclusion

In conclusion, our chosen models use logistic regression to predict cancellation and delay time of a flight between 101 airports in the US during the holiday season. Both models use month, visibility at origin, wind speed at destination, and snowfall at both origin and destination as features, with the delay model additionally using hour as a feature. Our cancellation model results are similar to others, but the delay model needs improvements to become more effective at prediction. Further improvements could be made by expanding the dataset used to develop the models and improve their usability for people flying between mid-sized airports. Our tips for avoiding cancellations and delays focus on non-weather trends, such as month and day trends. We did not explore, but believe there may be insights in studying trends between an airlines' cancellation rates, and their hub airport locations and weather.

References

1. Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2016). Prediction of weather-induced airline delays based on machine learning algorithms. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. https://doi.org/10.1109/dasc.2016.7777956
2. Travelmag. (2024, June 26). *The biggest 100 US airports by passenger traffic - Travelmag*. Travelmag. https://www.travelmag.com/articles/biggest-us-airports/#google_vignette
3. Kim, S., & Park, E. (2024). Prediction of flight departure delays caused by weather conditions adopting data-driven approaches. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-023-00867-5
4. Kiliç, K., & Sallan, J. M. (2023). Study of delay prediction in the US airport network. *Aerospace*, *10*(4), 342. https://doi.org/10.3390/aerospace10040342

Contributions

Contributions	Amy Merkelz	Chenyu Jiang	Yifan Zhang
Presentation	Slides 3, 8, 11, & 17, and text for slides 6, 7, & 9.	Slides 2 & 5, and plots for slides 6, 7, & 9.	Slides 4 & 12-16.
Summary	Wrote data merging information, strengths and weaknesses.	Wrote introduction and EDA.	Wrote data processing and model selection.
Code	Responsible for merging the data, and script to load models in the Shiny app.	Responsible for selecting the flight data and visualization code.	Responsible for collecting weather data, imputation and model selection.
Shiny App	Responsible for the Shiny app.	Provided feedback on the Shiny app.	Provided feedback on the Shiny app.