A statistical model for predicting flight cancellations or delays

A Data-Driven Approach

Shuo Li, Xinrui Zhong, Yunze Wang

Data Science Project
University of Wisconsin-Madison
November 2024

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

1. Data Preprocessing

2. Model Selection

3. Conclusion

Data Preprocessing

- Dataset: `LCD_{station id}_{year}.csv,Airport_Selected.csv`
- Imputed missing values using mean value to retain dataset consistency without losing significant data.
- Matched each airport to the nearest weather station using the Haversine distance method based on latitude and longitude coordinates.

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta \mathsf{lat}}{2}\right) + \cos(\mathsf{lat}_1) \cdot \cos(\mathsf{lat}_2) \cdot \sin^2\left(\frac{\Delta \mathsf{lon}}{2}\right)}\right)$$

- Converted all the time points to UTC.

Data Preprocessing

- Extracted weather data from the nearest time point to the scheduled departure, integrating it into the main dataset.
  - Time-consuming(**Vectorized Operations and Apply**) **ONLY 2h!**
- Created a clean and comprehensive dataset by handling and restructuring columns.

Airport ID	**Nearest Station ID**	**FlightDate**	**HourlyWindSpeed**	...
LAX	ST001	2018-01-01	15 mph	...
JFK	ST002	2018-01-01	10 mph	...
ORD	ST003	2018-01-01	20 mph	...
ATL	ST004	2018-01-01	12 mph	...

Feature Selection

- One-hot encoder applied to each object type.

Year	Object
Month	Object
DayofMonth	Object
DayOfWeek	Object
Marketing_Airline_Network	Object
Origin	Object
Dest	Object
CRSDepTime	Numeric
CRSArrTime	Numeric
HourlyDewPointTemperature	Numeric
HourlyDryBulbTemperature	Numeric
HourlyPrecipitation	Numeric
HourlyPressureChange	Numeric
HourlyRelativeHumidity	Numeric
HourlySeaLevelPressure	Numeric
HourlyVisibility	Numeric
HourlyWindSpeed	Numeric

Model Performance

- Canceling Analysis: Logistic Regression
  - SGD Optimizer
  - l2 penalty
- Delay Analysis: Ordinary Linear Regression
  - subsampling

	Canceling Analysis	Delay Analysis
Model type	Logistic Regression	Ordinary Linear Regression
MSE(in a subset)	0.04477	1810
F-test (p-value)	NA	0(very small)

Table: Comparison of different models

Overview of Findings

- Flights in the later days of the month show a lower likelihood of cancellation.
- Higher wind speeds may indicate extreme weather, increasing cancellation risks.
- Visibility significantly impacts flight delay durations.
- Transition days between workdays and weekends (Friday, Sunday, and Monday) are prone to delays.
- Humidity affects both cancellations and delays, often indicating increased probability of severe weather.

Flight Cancellation Analysis

Key Dates and Days of the Month

- **Later Days of the Month:** Decreased probability of flight cancellations.(coeff=-0.5)

Weather Factors

- **Wind Speed:** Higher wind speeds correlate with a higher likelihood of extreme weather and increased cancellations.(coeff=0.38)
- **Humidity:** Higher humidity levels often lead to more cancellations due to severe weather.(coeff=0.36)

Flight Delay Analysis

Key Days of the Week

- **Friday, Sunday, and Monday:** These days, linking weekends and weekdays, are more prone to delays.

Weather Factors

- **Visibility:** Lower visibility is a significant factor, as it can slow down flight operations.(coeff=-3.447)
- **Humidity:** Higher humidity can indicate rain, contributing to delays.(coeff=4.989)

Summary

- **Flight Cancellations:** Strongly influenced by the time of the month, wind speed, and humidity.
- **Flight Delays:** Commonly impacted by visibility, high humidity, and certain days (Friday, Sunday, and Monday).
- **Weather Impact:** Humidity plays a significant role in both cancellations and delays, indicating a potential increase in severe weather conditions.

Shiny APP

Shiny App link: **click here**

Thank you!