Flight Delays and Cancellations - Trends and Predictions

STAT 628 Module 3 - Group 6
Yifan Zhang, Chenyu Jiang, Amy Merkelz

Motivation

- **Background:** During the holiday season, the number of flights increases significantly, but then there are even more weather extremes that can exist in the winter.
- **Problem:** Airline delays and cancellations cause a great inconvenience to passengers.
- **Goal:** Determine a model that can help the passengers recognize important patterns in flight delays and cancellations to:
    - **Avoid canceled flights**
    - **Arrive early or on time**
    - **Predict the gate arrival time**

Initial Datasets

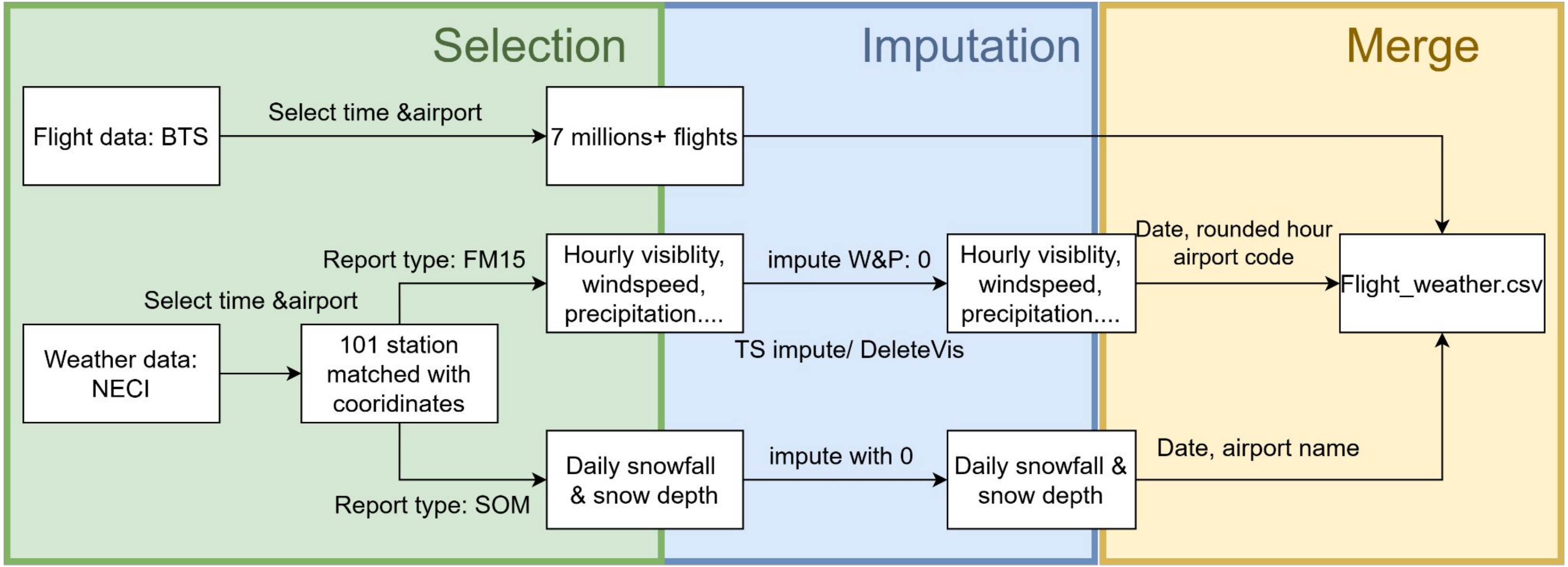
Flight Data:

- Provided by the Bureau of Transportation Statistics
- Flights in Nov., Dec., Jan. from 2018 - 2024, excluding Nov. 2020 - Jan. 2021 due to Covid-19
- Considered only flights from the top 100 airports by 2023 passenger volume, plus Madison

Weather Data:

- Provided by the National Centers for Environmental Information (NCEI)
- Match the nearest weather station for each airport and obtained hourly weather data

Data Processing

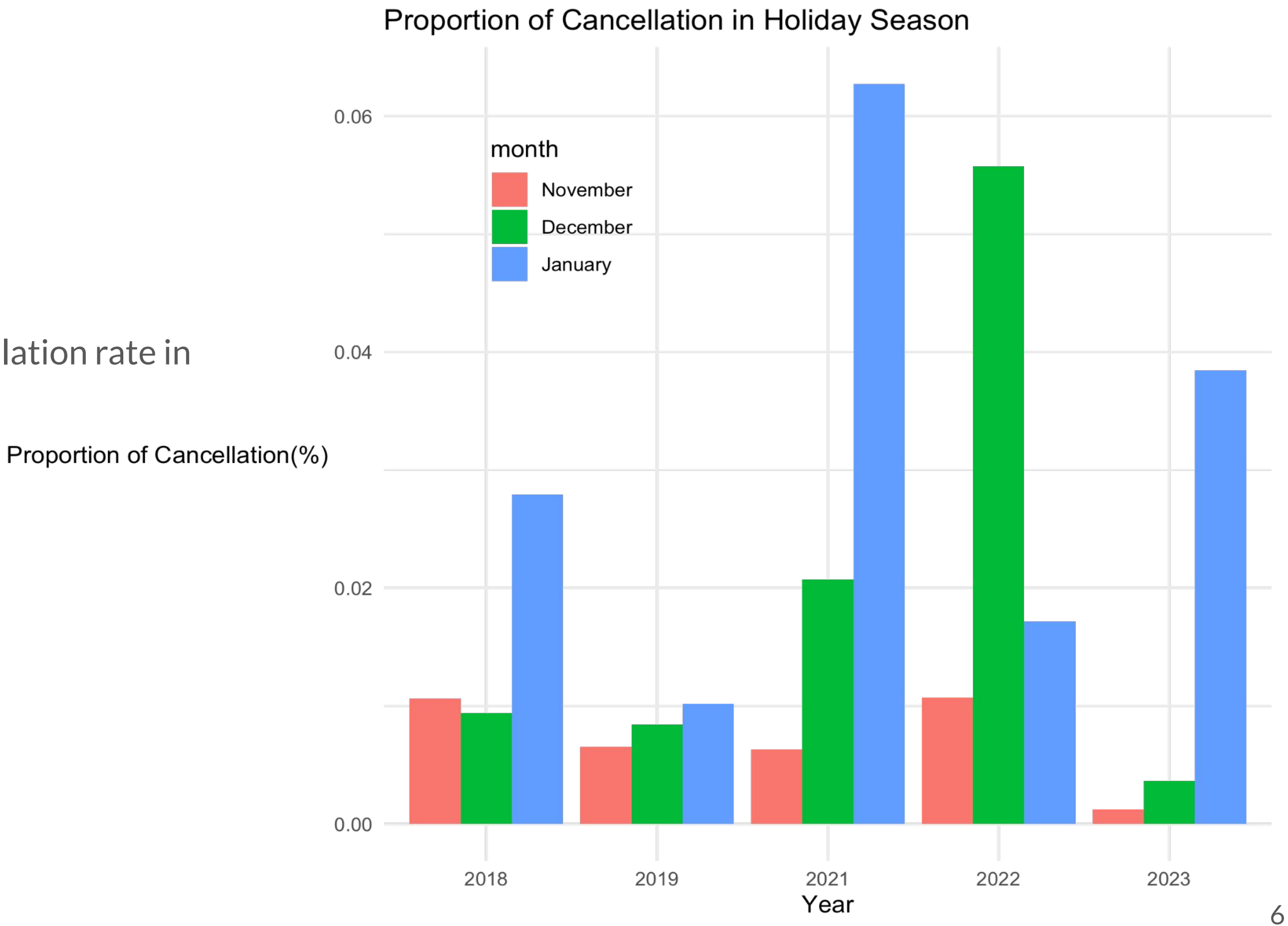


Final Cleaned Dataset

- Includes over 7 million flight records with corresponding weather data
- Have 66 parameters
- Total 10 airlines
- Cancellation rate: 1.88%
- Late arrivals rate: 34%

Observations

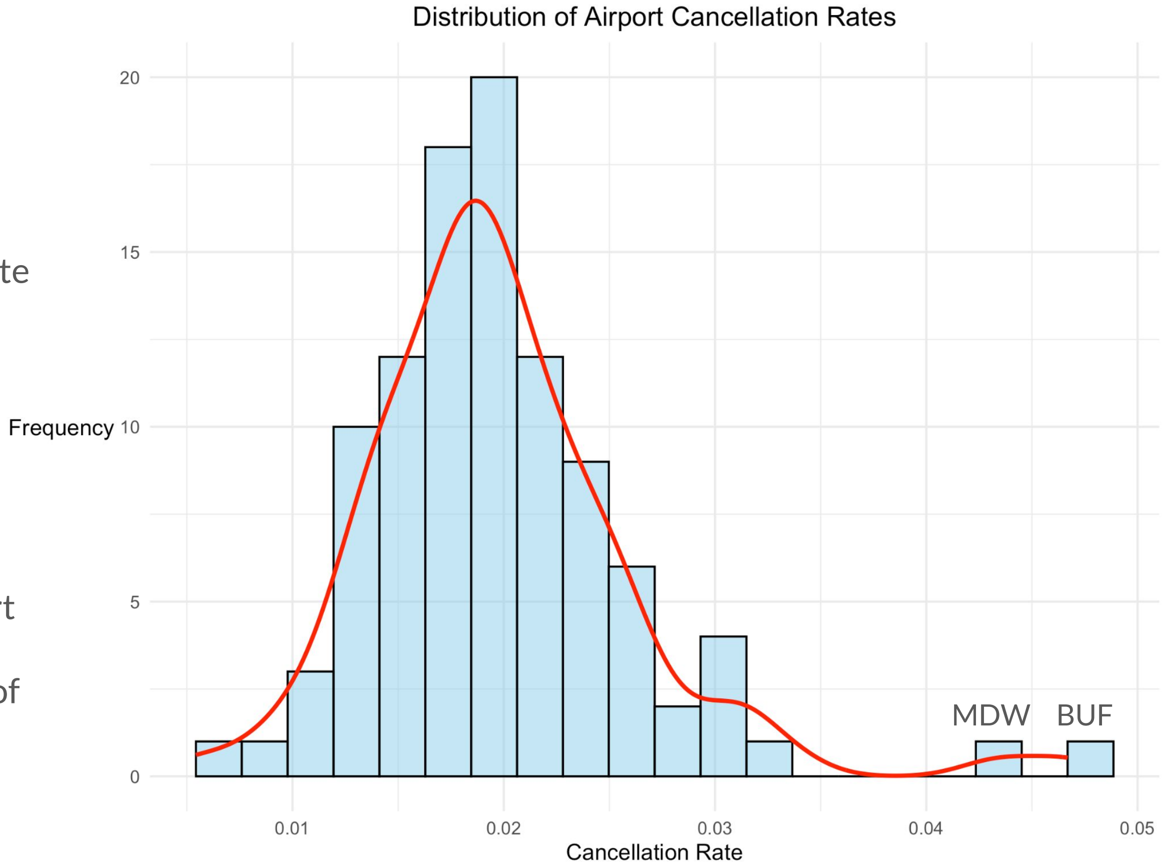
We could see the high cancellation rate in January for most years.



Observations

Two outliers for cancellation rate

- **Midway**: a Southwest Airlines hub that was disproportionately impacted by the 2022 meltdown
- **Buffalo**: the major airport for upstate NY, with an average yearly snowfall of 68.8 inches



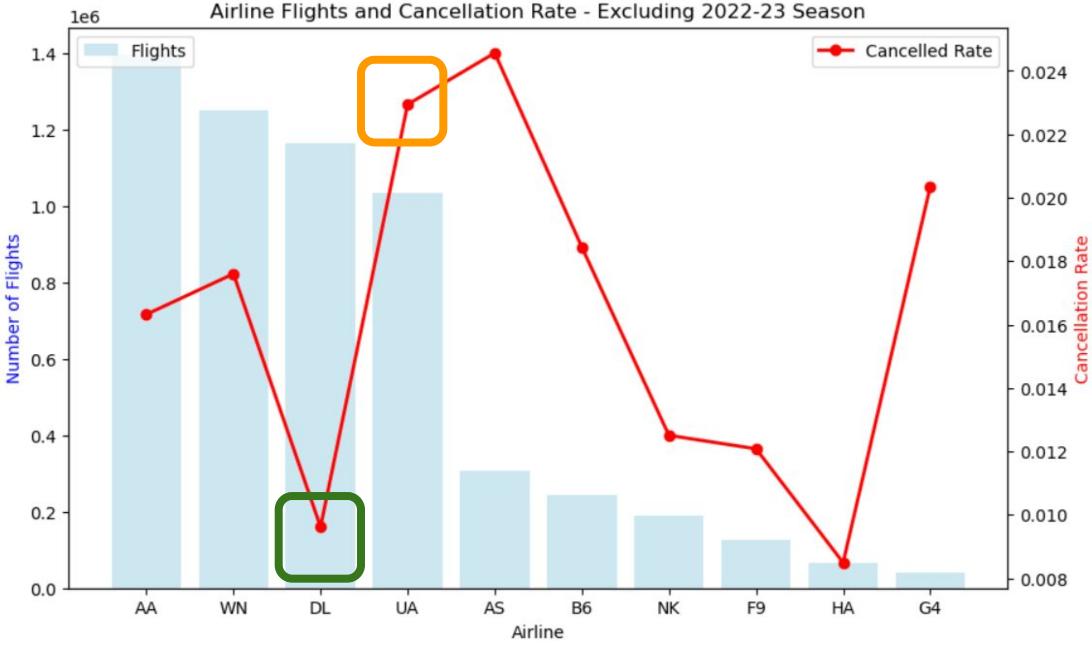
Observations

Cancellation rate over all flights (excluding 2022-23): **1.66%**

United Airlines has the highest cancellation rate (**2.2%**) of the 4 major airlines.

Delta has the least cancellations of the 4 major airlines (**0.9%**)

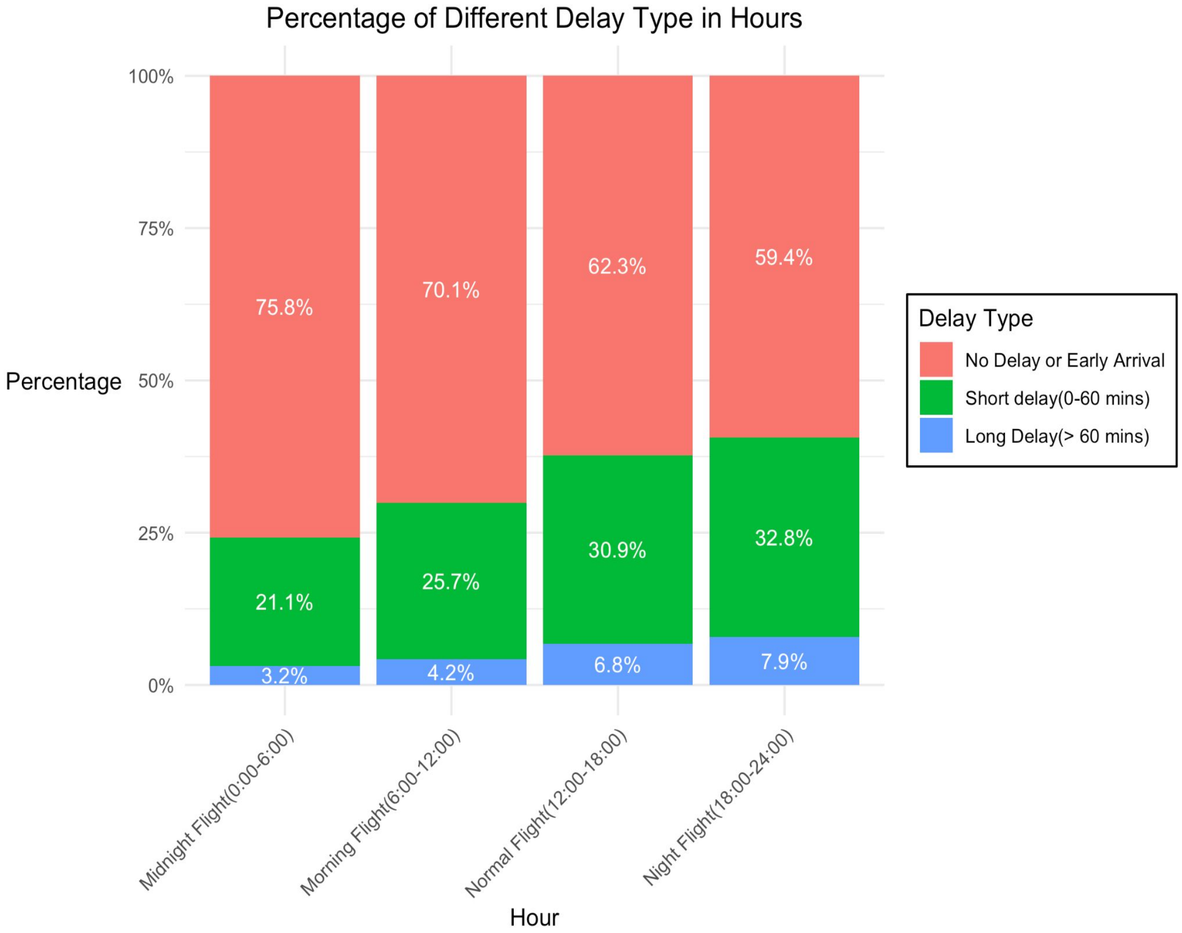
Delta also has the shortest average delay as well.



Observations

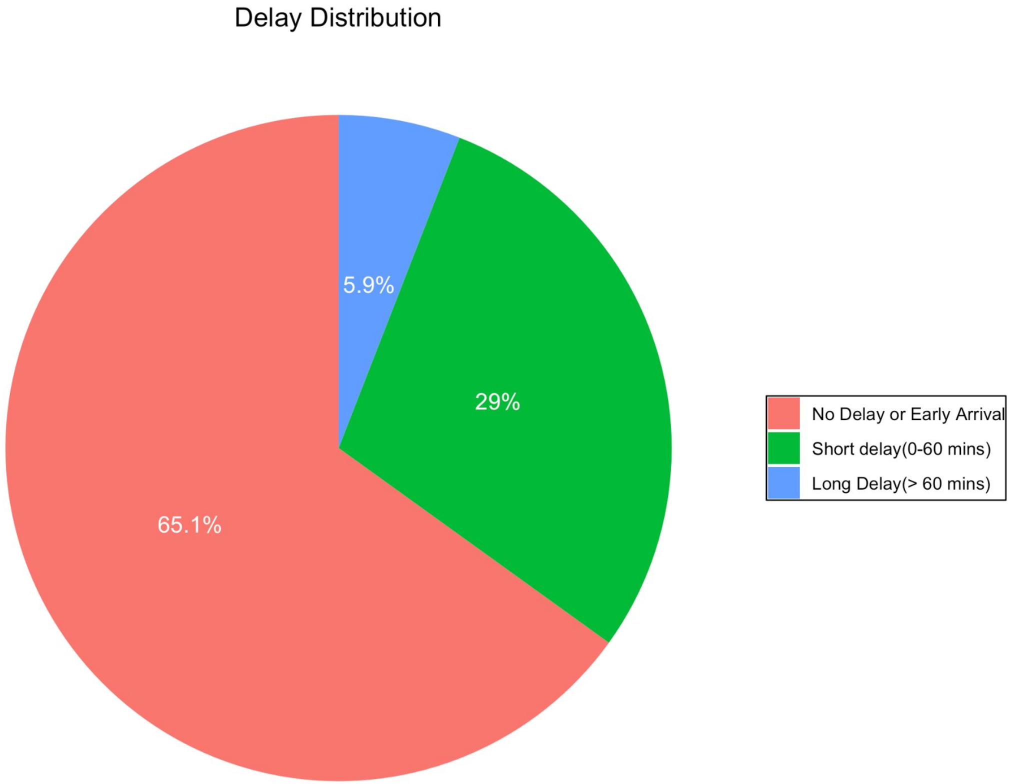
Flights departing after 6 pm have the lowest percentage of on time arrivals.

Early morning flights departing before 6 am have lowest percentage of delayed flights.



Observations

Most flights are on time or have less than 1 hour delay.



Recommendations

To avoid having a cancelled flight:

1. **Avoid traveling through airports with high cancellation rates** (e.g., Buffalo) when possible.
2. **Avoid traveling in January**, which has higher cancellation rates than November or December.
3. **Fly Delta!** Lowest cancellation rate of the 4 major airlines.

To arrive on time:

1. **Try to have 60+ minutes between flights**, to avoid delays causing a missed connection.
2. **Avoid evening flights**- only 59.4% are early or on time. Instead, take flights that depart before 6 am.
3. **Fly Delta!** Lowest average arrival delay (< 1 minute).

Model Selection

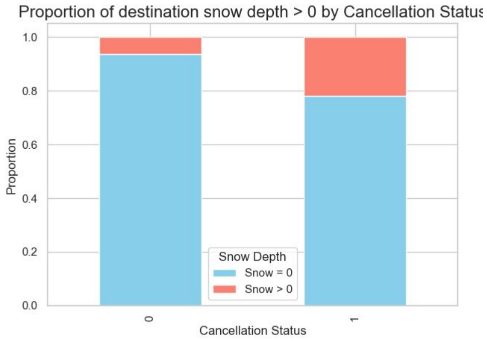
Logistic Regression for both model: **Simplicity**, **interpretability** and **computational complexity**.

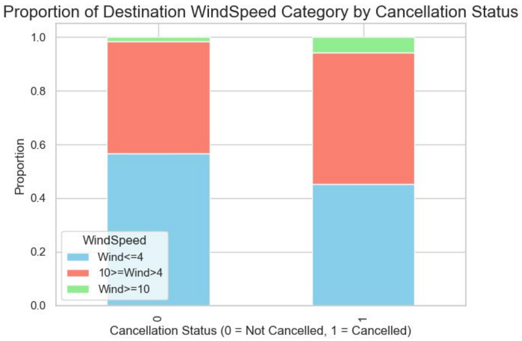
- **Undersampling** on both datasets to balance the dataset to a 1:1 or 1:1:1 ratio.
- Training :testing=8:2.
- Compared with other models(Decision Tree, Random forest...)
- Evaluated by **accuracy** and **F1-score**.

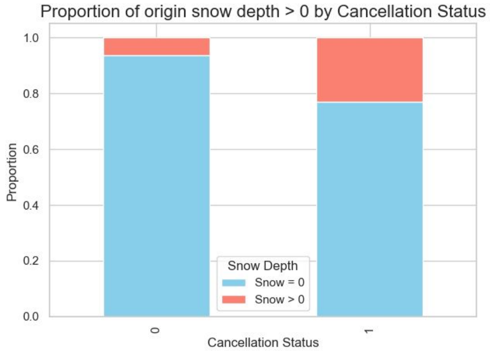
Model Selection

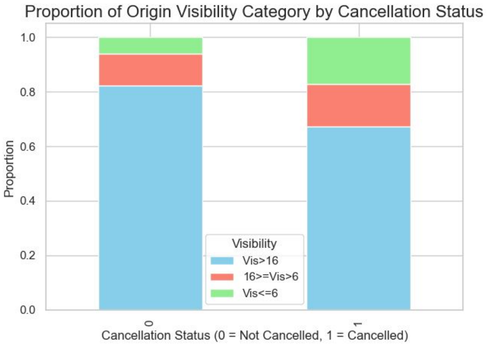
- Focused on **weather-related cancellations**.
- **Binary classification** model:canceled (1) or not canceled (0).
- **5 features**: month, visibility at origin airport, wind speed at destination, daily snow depth at origin and destination.

A greater proportion of flights are cancelled when there is snow on the ground at origin & destination


Proportion of destination snow depth > 0 by Cancellation Status


Proportion of Destination WindSpeed Category by Cancellation Status


Proportion of origin snow depth > 0 by Cancellation Status


Proportion of Origin Visibility Category by Cancellation Status

Model Selection

- Cancellation Model

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

Model	Accuracy	F1 Score	Number of Parameters
Logistic Regression	**0.753**	**0.757**	**5**
Decision Tree	**0.762**	**0.762**	**5**

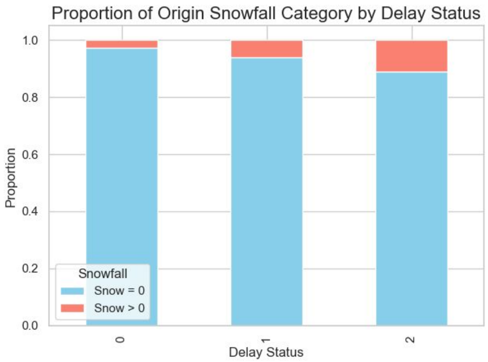
Model Selection

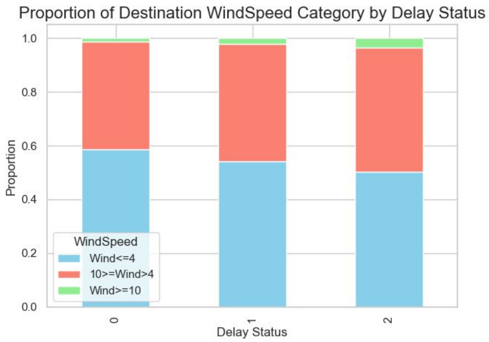
Three-class classification model:
more **realistic** and better captures the different levels of delay:

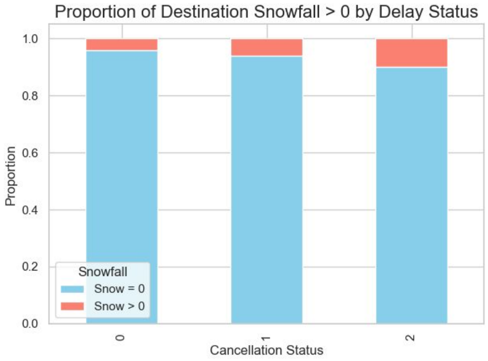
- **0**:on time or early , **no need to worry**
- **1**:< 1 hour delay , **need to hurry**
- **2**:> 1 hour delay , **require rescheduling**

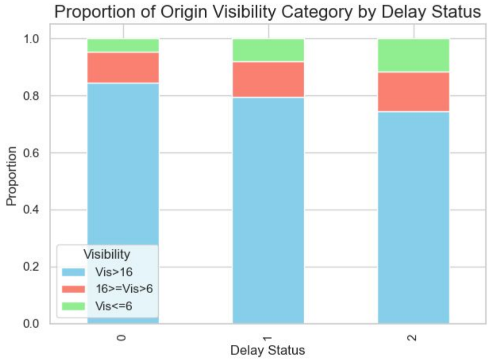
might be more helpful for passengers.

6 features: month, schedule departure hour, visibility at origin airport, wind speed at destination, daily snowfall at origin and destination.


Proportion of Origin Snowfall Category by Delay Status


Proportion of Destination WindSpeed Category by Delay Status


Proportion of Destination Snowfall > 0 by Delay Status


Proportion of Origin Visibility Category by Delay Status

Model Selection

- Delay Model

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

Model	Accuracy	F1 Score	Number of Parameters
Logistic Regression	**0.412**	**0.395**	**6**
Decision Tree	**0.420**	**0.411**	**6**
Random Forest	**0.417**	**0.403**	**6**
Gradient Boosting	**0.413**	**0.386**	**6**

Model Strengths and Weaknesses

Strengths:

- Both models use few features - easy to build a user-friendly tool to predict flight delay/cancellation.
- Logistic regression models are interpretable - easy to understand how each feature influences the prediction

Weaknesses:

- Limited dataset: Only considered 101 airports out of ~1000 airports
- Unimportance of airline in final models, despite seeing trends in exploratory data analysis
  - Possible confounders: an airline's 'hub' locations and typical snowfall there

Shiny App Demo

https://amerkelz.shinyapps.io/628_module3_group6/

(if the app does not load quickly, try pausing and refreshing the page)

References

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

Travelmag. (2024, June 26). The biggest 100 US airports by passenger traffic - Travelmag. Travelmag. https://www.travelmag.com/articles/biggest-us-airports/#google_vignette

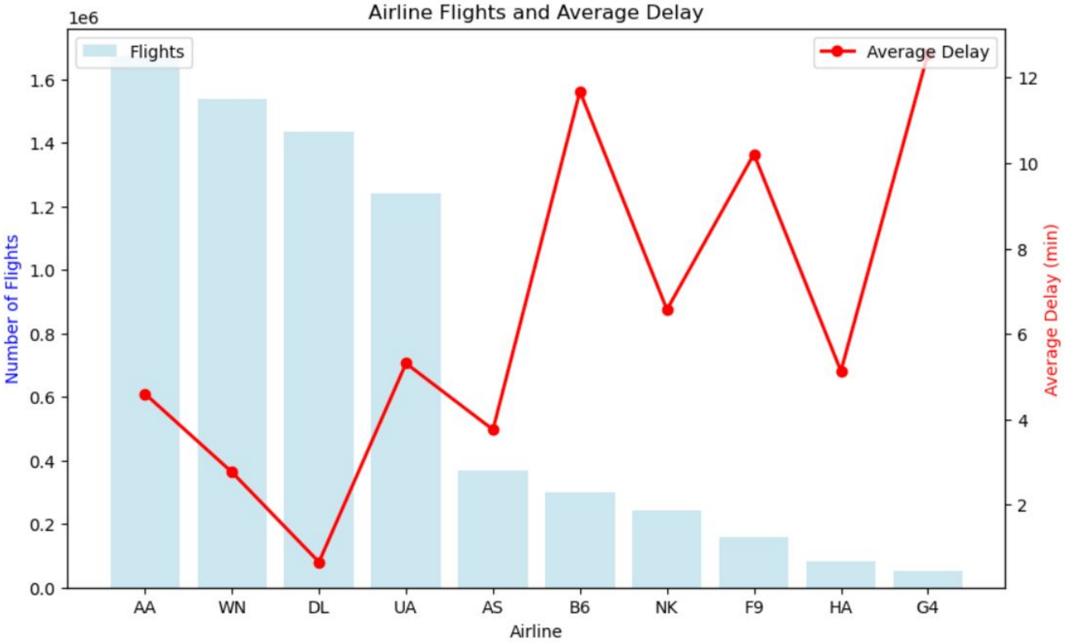
Kim S, Park E. Prediction of flight departure delays caused by weather conditions adopting data-driven approaches[J]. Journal of Big Data, 2024, 11(1): 11.

Kiliç K, Sallan J M. Study of delay prediction in the US airport network[J]. Aerospace, 2023, 10(4): 342.

Giblin, P. (2023, January 19). WIVT - News 34. WIVT - News 34.. www.binghamtonhomepage.com/news/the-snowiest-cities-in-the-u-s/

Thank you!

Appendix: Airline Delay Rates


Airline Flights and Average Delay
1e6
Flights
Average Delay
Number of Flights
Average Delay (min)
AA
WN
DL
UA
AS
B6
NK
F9
HA
G4
Airline