Summary

Introduction

1. To avoid cancellations: Avoid nighttime flights (9 pm - 5 am) and avoid extreme low temperatures and choose days with higher air pressure.

2. To reduce delays: Schedule flights outside nighttime hours and be mindful of weather forecasts, .Avoid extreme low temperatures and choose days with higher air pressure.

3. Model Overview: Using Random Forest models, we achieved 83% accuracy for predicting cancellations, with high precision for non-canceled flights, and identified that nighttime flights and weather conditions are critical factors in flight disruptions.

Background Information about Data

1. Data merge: Our study focuses on all domestic flights in the United States during the Holiday season (from Nov. 1st to Jan. 31st) from 2018 to 2024. We combined historical flight data from the U.S. Department of Transportation with historical weather data from the National Centers for Environmental Information, retrieving climate data via API.

1. Match airports with weather stations: Each airport is matched with its nearest weather station based on location by using geopy; if data from that station is unavailable, we proceed to the next closest station, continuing this process as needed.
2. Match Time: Convert Time Zone. Match the weather conditions of the corresponding weather station based on the departure and arrival time of flights.

2. Data cleaning:

Here, each variable has both "arrival" and "departure" features; for simplicity, only the key identifiers are shown. There are a total of 7112271 records and 26 features.

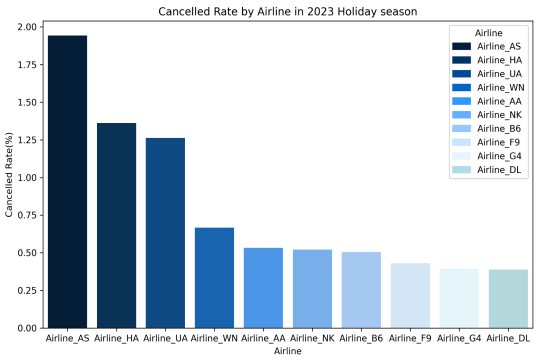
1. Remove irrelevant columns and features with over 70% missing values; ensure all remaining columns have less than 25% missing values.
2. Impute missing values: a) Use mean for *HourlySeaLevelPressure* and *HourlyStationPressure*. b) Use median for *HourlyVisibility*. c) Use 0 for missing values and “T” in *HourlyPrecipitation*. d) Replace "VRB" in *HourlyWindDirection* with 999.0.
3. Convert data: a) Apply one-hot encoding to *HourlySkyConditions* and *airline*. b)The departure and arrival airports into numeric encodings. c) Perform time binning: 5 AM to 12 PM as "morning," 12 PM to PM as "afternoon," 5 PM to 9 PM as "evening," and other times as "night." e) Convert the time column to a Unix timestamp for calculating.
4. Delete samples with missing values and outliers.

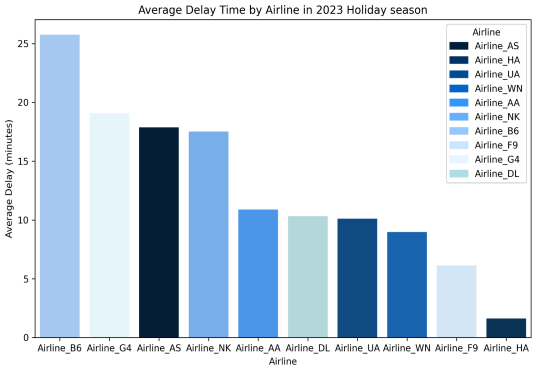
Exploratory Data Analysis (EDA)

We focused our exploratory data analysis (EDA) on the 2023 holiday season.

1. Bar chart for Airline. Two bar charts below show average delay time for different airlines. We used different colors to represent various airlines, with colors ranging from dark to light according to the flight cancellation rate, from high to low. These colors correspond to a bar chart showing the average delay time, where the same color indicates the same airline.

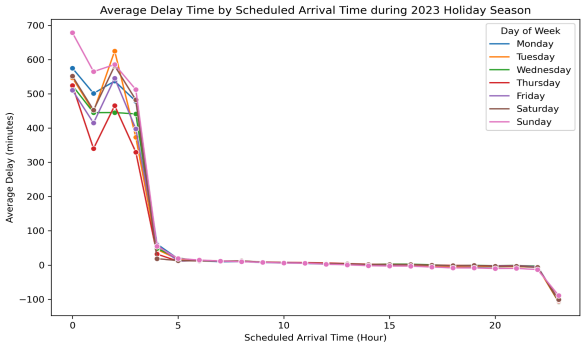
The two charts reveal significant differences in cancellation rates and average delay times across various airlines, underscoring the importance of choosing the right airline. Some flights also

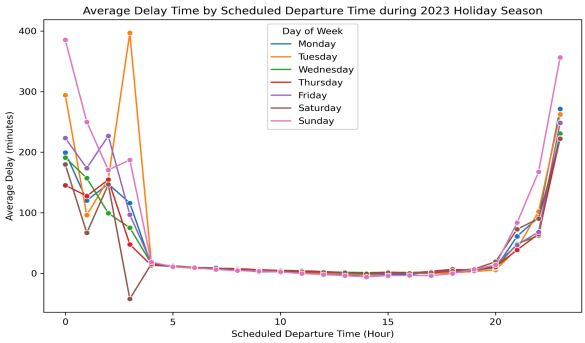




carry higher risks. For example, while Airline HA has the lowest average delay time, it also shows a high likelihood of cancellations. In contrast, Airline A9 stands out as a top choice, with both a low cancellation rate and short average delay time.

2. Line chart showing the average delay time and flight cancellation rate for different scheduled departure times throughout the day across various days of the week.





From the two charts above, we can preliminarily conclude that time of day may be a key factor influencing both cancellation rates and average delay durations. Different days of the week tend to show similar patterns across time periods (morning, afternoon, evening, night). The charts suggest that flights scheduled for evening departures are more likely to be canceled or delayed. Therefore, in our subsequent modeling, we transformed the scheduled arrival and departure times to better account for these patterns.

Part 1: Simple tips to avoid canceled flights during the holiday season

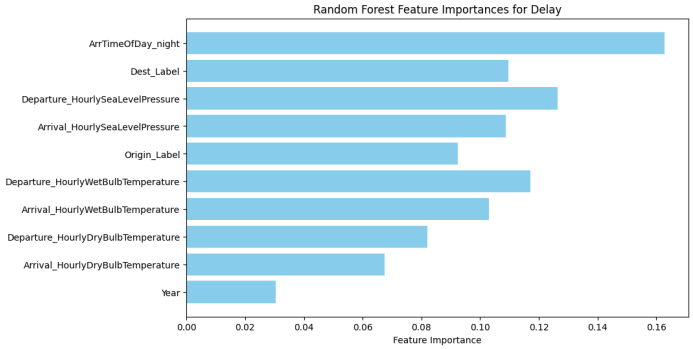
We tried a **logistic regression model, decision tree** and **random forest classifier**. The logistic regression model was tested but did not perform well. We used a Random Forest classifier to predict flight cancellations, with predictors including *nighttime arrival, destination, origin, sea-level pressure*, and *wet/dry bulb temperatures* at departure and arrival airports. The model achieved an 83% accuracy, with high precision for non-canceled flights (99%) but lower precision for canceled flights (7%), and a recall of 62% for cancellations. This indicates the model is more effective at identifying non-canceled flights, likely due to class imbalance. Feature importance analysis showed weather factors, especially temperature and pressure, as key predictors, while airline-specific details were less impactful.

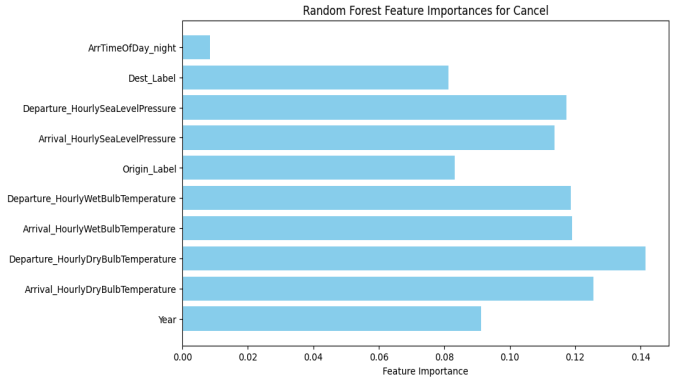
Part 2: Simple tips to arrive early or on time to their destination during the holiday season

We used a **Random Forest regressor** to predict flight delay time, yielding an R^2 score of 0.086 on the test set. Nighttime arrival (9 pm - 5 am) was the most important factor, with weather conditions like pressure and temperature also influential. While Random Forest provides feature importance, it does not indicate the direction of each feature's influence, making specific impact interpretation challenging.

Part 3: Prediction Model for Arrival Times/Canceled Flights

Overall, we built 2 models to predict delay time and cancelation probability by Random Forest regressor and Random Forest classifier, respectively. We use predictors as 'wheather arrive at night','Dest','HourlySeaLevelPressure' for both departure or arrival airport, 'Origin', 'WetBulb/DryBulb Temperature' for both departure and arrival airport, where response are delay time and cancelation, respectively.





We trained models on an 80/20 train-test split, initially using a decision tree but later adopting Random Forest to reduce overfitting and improve accuracy. Predictor selection was based on importance scores from a Random Forest model, with a new predictor, "time of day," added to capture arrival/departure times. Our final delay prediction model, a Random Forest regressor, achieved an R^2 of 0.067 and MSE of 25,918 on the test set, while the cancellation prediction model, a Random Forest classifier, reached an accuracy of 83.3%. Strengths of these models include reduced overfitting, stability across diverse data, and capturing non-linear relationships. However, weaknesses include limited interpretability and lack of straightforward hyperparameter tuning options.

Conclusion

In summary, we developed two Random Forest models to predict flight delay times and cancellation probabilities, achieving reasonable accuracy and identifying key predictors like nighttime arrivals and weather conditions (temperature and pressure). The classifier performed well for non-canceled flights but struggled with cancellations due to class imbalance. While Random Forest's ability to capture complex, non-linear relationships and reduce overfitting proved valuable, interpretability and hyperparameter tuning remain challenging. Our findings highlight the impact of weather and timing on flight disruptions, offering insights for minimizing delays and cancellations.

Contributions and References

Bureau of Transportation Statistics. “OST_R | BTS | Transtats.” www.transtats.bts.gov, 2024, www.transtats.bts.gov/.
“NCEI Object Store Explorer.” Noaa.gov, 2024, www.ncei.noaa.gov/oa/local-climatological-data/index.html#v2/.

Contributions	Meiyi Yan	MinYuan Zhao	Siyu Wang
Presentation	Responsible for slides part 3 Final Model.	Reviewed and provided feedback on slides.	Responsible for slides part 1, 2 and 4.
Summary	Responsible for Part 1, Part2, Part 3 and Conclusion.	Reviewed and provided feedback on summary.	Responsible for introduction, background, EDA and references.
Code	1)Responsible for random forest code. 2)Responsible for data cleaning code.	1)Responsible for random forest code. 2)Responsible for data merging and data cleaning code.	1)Responsible for data download API code. 2)Responsible for EDA code.
Shiny App	Reviewed and provided feedback on Shiny app.	Responsible for Shiny app.	Reviewed and provided feedback on Shiny app.