Group 2 Flight Prediction Summary

1. Introduction

Airline cancellations and delays have major impacts, causing financial losses for both passengers and airlines. Various factors like weather, mechanical issues, airport capacity, and holiday traffic affect flight status. This model analyzes seven years of holiday season flight and weather data (Nov. 1 to Jan. 31) to predict flight outcomes, helping passengers anticipate potential cancellations, delays, or on-time arrivals and make informed travel decisions.

2. Data Cleaning & Merging

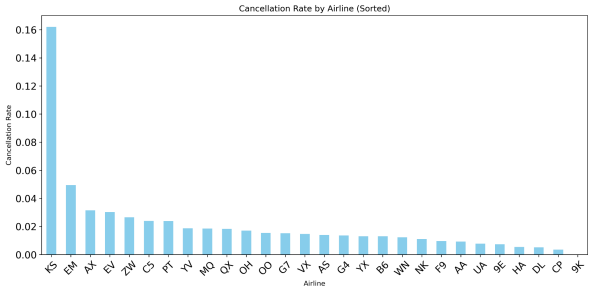
First, we standardize the time format by converting all times, including flight data and weather station data, to a datetime format with timezone information, using the timezone of their respective locations. Then, we match airports with weather stations. Instead of directly using data from the weather station closest to the airport, we chose the station that is both relatively close and has the most comprehensive data. If there is no weather data recorded for a specific time, we use monitoring data from three hours before to one hour after that time to fill in the gap. In the data cleaning, we are planning to clean many outliers including the non-stationary behavior such as the 2022 Seattle-specific Alaska Airlines issue, major snowstorms during the holiday season, etc.

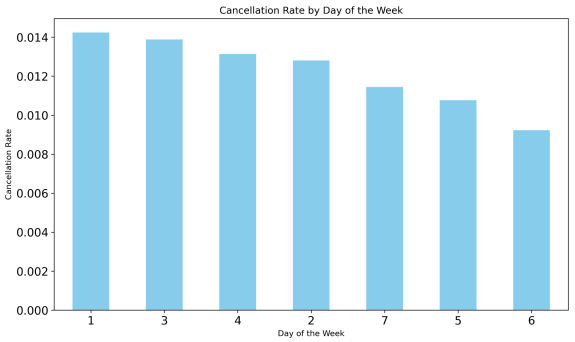
3. Feature Engineering

In this study, feature engineering involved two main aspects. First, categorical features such as airline type, DepAirport, and ArrAirport were transformed using label encoding to convert them into numerical formats suitable for the models. Second, new features were constructed to enhance the predictive power of the model. These include flight duration, calculated as the difference between the scheduled arrival and departure times, time of day, which categorizes flights into five-time segments (late night, morning, noon, afternoon, evening), and sky condition, segmented into five levels based on cloud coverage. These engineered features provided additional context to help improve the model's ability to perform tasks.

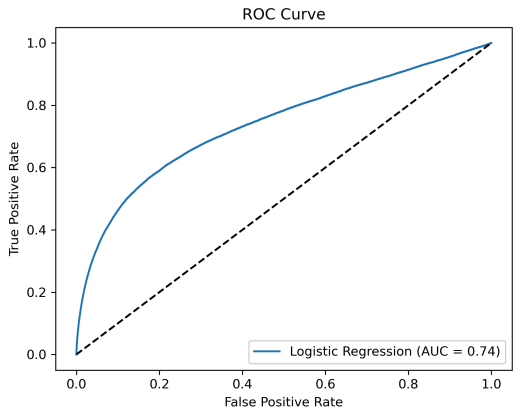
4. Task1:Simple Tips to avoid cancellation flights during the holiday season

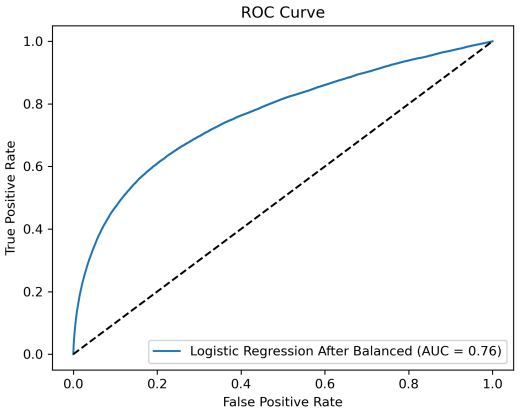
We selected 21 features relevant to flight cancellations, including airline type, flight duration, departure/arrival airport, day of the week, and weather-related factors such as sky condition, humidity, temperature, wind speed, and visibility. To assess significance, we performed a T-test for continuous variables and a Chi-Squared test for categorical variables, both of which showed that all features were statistically significant with p-values below 0.05. Then, we visualized the category data, Airline type, and day of the week to provide some tips for passengers. The results are shown below:



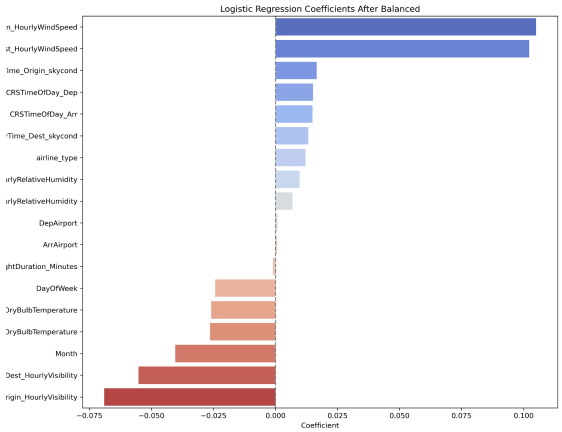


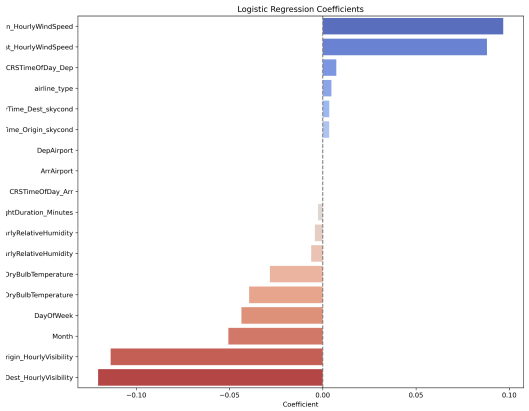
The results show that one of the airline types has a high cancellation rate but the sample is too small in the data set, with only about 600 flights during seven year of holiday season. Thus, the tips in selecting airline type will be selected from the lowest few airline companies such as United Airlines, American Airlines, or Delta Airlines. From the cancellation rate by day of the week, we will recommend selecting flights on Saturdays which have the lowest cancellation rate among all the days. For prediction, we applied logistic regression with flight cancellation as the outcome. Initially, the model had high accuracy (0.98) but a low AUC (0.74), likely due to an imbalance in the dataset, with far more non-canceled flights. After balancing the data, accuracy decreased to 0.73, while AUC improved to 0.76, indicating moderate predictive strength. Key features identified were visibility and wind speed, with increases in visibility and wind speed both significantly reducing the probability of cancellation.





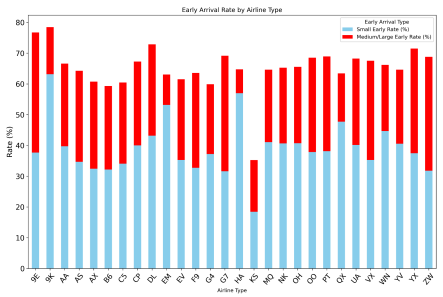
Then, we explored the coefficient of the logistic regression model, we found out that visibility and windspeed are the most important features in the model. Before balancing the data, one unit of visibility increased, the probability of flight cancellation will decrease by about 13%. After balancing the data, one unit of visibility increased, and the probability of flight cancellation decreased by about 7%.

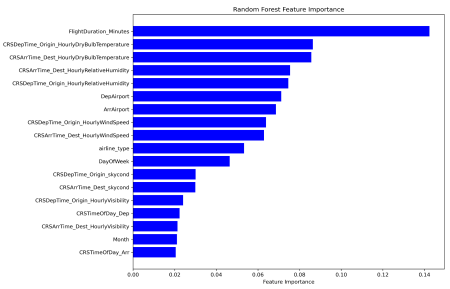




5. Task2: Simple Tips to arrive early or on time to their destination during the holiday season

We implemented a random forest model with the aim of analyzing flight arrival types, specifically focusing on whether flights arrive early or late. The model selection included a variety of predictors, such as *Flight Duration*, *Hourly Dry Bulb Temperature* at both the origin and destination, *Relative Humidity*, *Wind Speed*, *Airline Type*, and other weather and scheduling factors. The feature importance analysis revealed that *Flight Duration* is the most significant factor affecting flight arrival times, with other top contributors being *Dry Bulb Temperature* and *Relative Humidity* at both the origin and destination airports. This aligns with expectations since flight duration directly impacts arrival timings, and environmental conditions such as temperature and humidity can influence flight performance and scheduling. The random forest model's ability to identify and rank important features helps us understand which variables most influence early or delayed arrivals. While the model is robust, it can be challenging to interpret compared to simpler models like logistic regression.





6. Task3: Prediction Model for Arrival Times

Methods: In this task, three machine learning algorithms were utilized for classifying the type of flight delay based on various features, including RF, XGBoost, and LightGBM. The independent variables continued from those used in Task 1, while the dependent variable followed from Task 2. Due to class imbalance, we applied a full resampling technique to balance the dataset, adjusting the sample size of each class to match the size of the smallest class. Evaluation metrics like Accuracy and micro F1-score were utilized to assess model performance. All samples were divided into disjoint training and test samples in a ratio of 4:1.

Results: Performance of the leveraging models is shown in the following table. RF achieved an Accuracy of 0.395 and micro F1 score of 0.376, outperforming the other two models. XGBoost and LightGBM exhibited similar performance, with both models achieving comparable accuracy and micro F1 scores.

Model	Accuracy	Micro-F1
RF	0.395	0.376
XGBoost	0.385	0.367
LightGBM	0.383	0.362

Discussion:The results demonstrate that RF outperformed XGBoost and LightGBM in both accuracy and micro F1 score, indicating its superior capability in classifying flight delay types given the selected features. The use of a full resampling technique effectively balanced the dataset, though it might be the one responsible for low overall performance in this task.. Future work could explore alternative ensemble techniques and data balancing methods to further enhance classification performance.

Conclusion:

This report analyzed key factors influencing flight cancellations and delays during the holiday season to help passengers plan smarter travel. Our logistic regression model found that visibility and wind speed significantly reduce cancellation risks, while our random forest model showed that flight duration, temperature, and humidity affect arrival timing. And our final predicted model shows that compared with Random forest, XGBoost, and LightGBM, Random forest have the highest accuracy and F-1 scores. But the predicted results has a long way to go. Therefore, we recommend choosing flights with airlines that have low cancellation rates, such as United, American, or Delta, and consider flying on Saturdays for better on-time performance based on the data visualization. While our models provide valuable insights, future work could further improve prediction accuracy by refining data balancing techniques.

Contribution

Task	Zhenke Peng	Zekai Xu	Jiren Lu
Data Cleaning and merging	Contributed to downloading weather data years 20-21, Cleaning outliers	Contributed to downloading weather data years 22-24, reviewed data cleaning and merging	Contributed to downloading weather data for years 18-19, Merging flight data and weather data
Task1	Contributed to data exploration and Logistic regression model	Review the logistic regression model	Review the data exploration and logistic regression model
Task2	Contributed to develop the Random forest model	Contributed to develop the random forest model	Contributed to the review of the Random forest model
Task3	Contributed to review and comment on the final flight prediction model	Contributed to developing the final flight prediction model	Contributed to review and comment on the final flight prediction model
Shiny App	Review and comment on the shiny app	Review and comment on the shiny app	Contributed to developing the Shiny app