Module 3: Airline Project

1 Introduction

As air travel becomes more common, it's essential for passengers to choose flights wisely. We analyzed this through logistic regression and neural network models, and here are our findings:

To avoid cancellations, we suggest choosing Delta Air Lines, traveling outside holiday periods, selecting morning flights, avoiding popular tourist destinations, and opting for days with high visibility and low wind speeds. Conclusion is drawn from logistic regression model, with over 70% accuracy on both cancelled and non-cancelled flights.

To avoid arrival delays, travel on warm, clear days with mild wind and no precipitation. Morning arrivals and Delta flights are less prone to delays, while Hawaiian Airlines may have higher delay risks. Conclusion is drawn from logistic regression model, with F1 score 0.51, recall rate 0.57.

Our neural network model predicts delay time reached an RMSE of 39 minutes on the test set. We extracted the first three layers of the five-layer delay prediction neural network model to use as a feature extractor and fine-tuned it to train a flight cancellation prediction model. The recall rate for cancelled case is 0.34.

2 Data Cleaning

To study factors for arrival delay, we first extracted airline IDs from Bureau of Transportation Statistics and used a web driver to scrape airline data. Weather data was gathered from National Centers for Environmental information based on airports' latitude and longitude.

Data during the COVID period was filtered out, and flights without matching weather stations or without hourly weather data were excluded. Two airports (DUT and JHM) with cancellation rates over 15% were also removed. All times were converted to CST. Missing weather values were imputed via forward filling and KNN ($N = 5$). A *Holiday period* variable was added to indicate departures during Thanksgiving, Christmas, or New Year's. In Part 1 and Part 2 below, departure and arrival times were categorized by periods of the day. Lastly, We merged the flight data with the weather data based on the columns for departure airport, arrival airport, and scheduled departure time (CST). The final dataset contains 7,248,726 records regarding 378 airports.

Here are part of our data visualization. Figure 1 shows similar cancellation rates across most airports, with a few exceptions. Figure 2 indicates higher cancellation rates at night and lower rates in the morning.

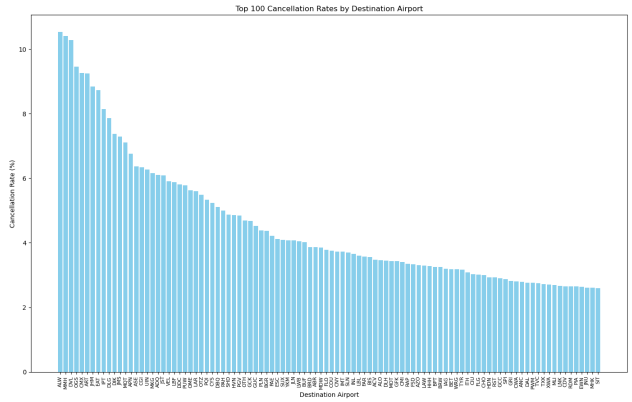

Figure 1: Cancellation Rate by Destination Airport

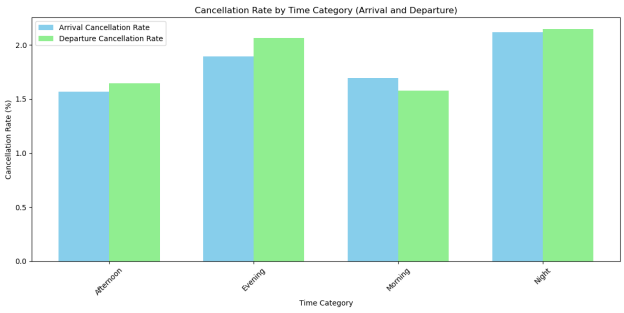

Figure 2: Cancellation Rate by Time

3 Data Modeling and Analysis

3.1 Part 1: Simple Tips to Avoid Cancelled Flights

To predict flight cancellations, we used cancellation as the outcome and included ten flight-specific features (*Month*, *Day of Week*, *Departure Time*, etc.) alongside hourly weather variables (excluding *Weather Type*). One-hot encoding was applied to categorical variables.

Since only 1.73% of flights are cancelled, data imbalance was a major issue. A logistic regression model fitted on this data yielded 99.98% accuracy for non-cancelled flights (negative samples) but only 1.45% for cancelled ones (positive samples), indicating the model's bias toward predicting non-cancellations. To address this, we applied a resampling method.

We split the data into training and test sets (9:1). On the training set, we first oversampled the positive samples to reach half the number of negative samples, then applied undersampling to achieve a 1:1 ratio, as shown in Figure 3. The logistic regression model trained on this balanced set achieved 74.94% accuracy for negative samples and 71.23% for positive samples on the test set, with the ROC curve shown in Figure 4.

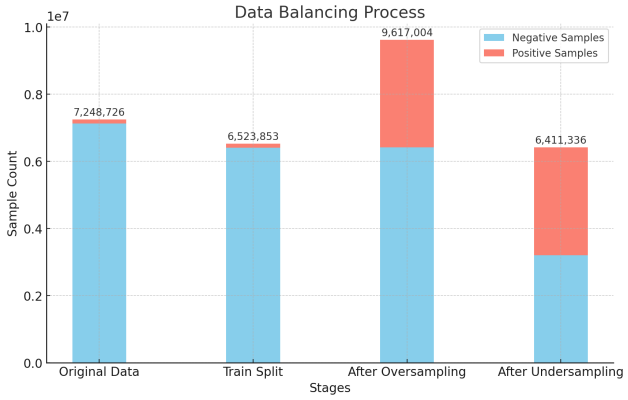

Figure 3: Data Balancing

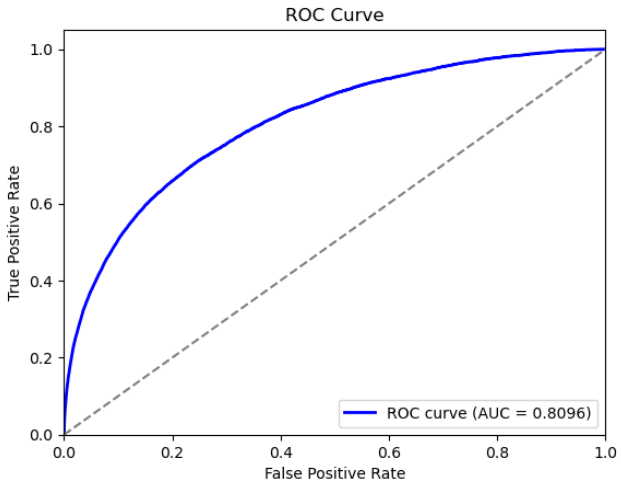

Figure 4: ROC Curve

We calculated the Odds Ratio (OR) for each variable based on its coefficient, using the formula $\text{OR} = e^{\text{coefficient}}$. Based on the OR values, we provided recommendations to avoid cancelled flights. Table 1 shows variables with OR values far from 1.

Variable	**OR Value**
Operating Carrier	WN=2.14, AS=2.03, DL=0.38
Holiday Period	1.73
Origin/Dest	ORIGIN: SAN=1.64, SFO=1.60, DTW=0.49; DEST: MCO=1.63, SFO=1.57, DTW=0.52
Departure/Arrival Time	Arrival: Night=1.20, Morning=0.83; Departure: Evening=1.16, Morning=0.90
Origin Visibility	0.94
Origin Wind Speed	1.14

Table 1: Variables and their Odds Ratios

3.2 Part 2: Simple tips to arrive early or on time to destinations

This part is similar to part 1. We developed a logistic regression model to predict flight delays, complemented by a comparison with a random forest model. The features include hourly weather data (e.g., visibility, pressure, temperature), as well as flight details such as origin and destination airports,

scheduled departure time, distance, elapsed time, and operating carrier. To guarantee both efficiency and accuracy, categorical data are processed. Airport data, which is not significant, are processed with frequency encoding, while other categorical data are encoded into one-hot variables. Also, we checked independence of predictors, results can be seen below.

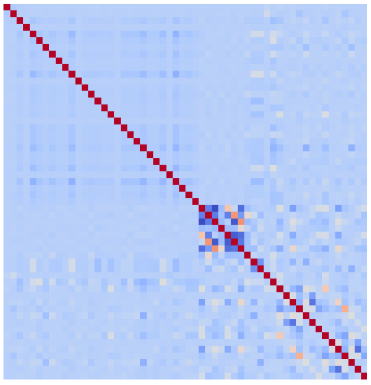

Figure 5: Heatmap of predictors' correlation

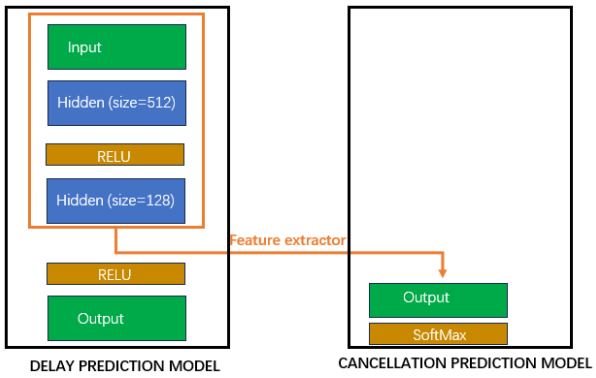

Figure 6: Prediction Model Structure

Based on the coefficient, we can pick the most influencing factors:

Variable	**Coefficient**
Origin visibility	-0.20
Origin bulb temperature	-0.16
Operating Carrier	DL=-0.12, HA=0.09
Origin wind speed	0.11

Table 2: Influencing Variables and their coefficients

3.3 Part 3: Prediction model for arrival times

We used a neural network structured as shown in Figure 6, treating cancellations as infinite delays to capture the relationship between cancellation and delay. We split the data 70/30 for training and testing, using features such as *Dest/Origin*, *Longitude*, *Latitude*, *Date* (Local and CST), *Operating Carrier*, and almost all the weather features (excluding *Weather Type*). The model was optimized with *ADAM* and a learning rate scheduler that exponentially reduced the rate. The model also uses *Xavier initialization* as a good starting point for optimization.

Our delay model achieved an RMSE of 39 minutes on the test set. For cancellations, the model showed strong performance for uncancelled flights with high recall (0.6796), precision (0.9830), specificity (0.9647), and F1-score (0.8037), but struggled on cancelled cases with low recall (0.3408), precision (0.0186), specificity (0.0353), and F1-score (0.0352).

Overall, our model is strong in delay time prediction with effective feature utilization and optimization. Additionally, Our joint training approach effectively reduces resource consumption while maintaining high performance. The weakness is that the cancellation model struggles with cancelled cases. So we chose the model in Part 1 for cancellation prediction in the Shiny app.

4 Conclusion

To conclude, we developed reasonable models to provide tips for passengers. However, the results are far from satisfactory. More advanced models and refined data processing methods are needed to improve the accuracy of the predictions.

Contributions

Contribution	**Zhengyong Chen**	**Xiangsen Dong**	**Xupeng Tang**	**Zhaoqing Wu**
Presentation	Responsible for Shiny App	Responsible for Data Cleaning and Model 2	Responsible for Model 1	Responsible for Model 3
Summary	Reviewed and provided feedback	Responsible for Introduction and Part 2	Responsible for Data Cleaning and Part 1	Responsible for Part 3
Code	Responsible for Model 3	Responsible for Data Cleaning and Model 2	Responsible for Data Cleaning and Model 1	Responsible for Data Cleaning and Model 3
Shiny App	Responsible for Shiny app implementation	Reviewed/edited and provided feedback	Reviewed/edited and provided feedback	Responsible for Shiny app implementation

Table 3: Team Contributions