Flights Delay Prediction

Yunze Wang, Shuo Li, Xinrui Zhong

1 Introduction

This project aims to analyze patterns in flight delays and cancellations during the holiday season (November to January) using flight and weather data provided by the U.S. Department of Transportation and the National Weather Service. By identifying key factors, we will provide practical advice to help passengers avoid cancellations and arrive on time. Additionally, a predictive model will be developed to estimate flight arrival times, enhancing the travel experience for passengers.

2 Data Preprocessing

2.1 Airport and Weather Station Data Collection and Matching

We began by manually collecting data on over 7,000 airports within the United States and filtered flights to include only those involving approximately 300 airports based on our downloaded flight data. Following this, we used a web scraping tool to gather latitude and longitude information for weather stations across the United States.

To match each airport with its nearest weather station, we utilized the Haversine distance [2] calculation method. Haversine distance is a formula used to determine the spherical distance between two points on the Earth's surface, based on their latitude and longitude. The formula is as follow:

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

where d is the distance between the two points, r is the Earth's radius (approximately 6,371 km), lat_1 and lat_2 are the latitudes of the airport and weather station, and Δlat and Δlon are the differences in latitude and longitude, respectively. After calculating the distance between each airport and all weather stations, we selected the closest station for each airport.

2.2 Time Conversion and Data Merging

Since different regions in the United States use various time standards, while weather station data is in Coordinated Universal Time (UTC) [1], we converted all timestamps to UTC using the `pytz` package. Additionally, most regions in the United States switch to standard time on the first Sunday in November, while Hawaii and Arizona do not observe daylight saving time, which we accounted for during the conversion process.

Next, we merged the flight data with the corresponding weather data. To estimate whether flights are delayed, it is essential to have information about weather conditions before boarding. Therefore, we focused on weather data during the departure period only.

2.3 Handling Missing Values

For numeric variables, missing values were filled with the mean. For categorical (object) variables, one-hot encoding was used. This ensured that the dataset was completed and suitable for model training.

3 Model Selection

To effectively predict flight cancellations and delays, two different models were selected based on the nature of each prediction task.

3.1 Logistic Regression for Cancellation Prediction

For predicting flight cancellations, we employed a logistic regression model, which is suitable for binary classification tasks. In this case, the model predicts whether a flight will be canceled (1) or not (0).

Data Imbalance One challenge in predicting cancellations is the class imbalance, as most flights are not canceled. This imbalance can lead to a bias in the model towards non-cancellation predictions. To address this, oversampling was applied to increase the representation of canceled flights in the training data, thereby improving the model's ability to detect cancellations and enhancing the recall rate. Oversampling was achieved by duplicating canceled flight instances until the dataset had a more balanced ratio of canceled to non-canceled flights.

Feature Selection Relevant features were selected based on domain knowledge and statistical testing. Key factors included weather conditions (e.g., humidity, wind speed), flight schedules (departure and arrival times), and day-specific variables (day of the week, day of the month). By including these features, we aimed to capture the conditions most indicative of cancellations.

Model Performance Evaluation The logistic regression model was evaluated using MSE, recall. To evaluate the effectiveness of the logistic regression model in identifying flight cancellations, we use January 2018 as an example, which has **18976** canceled flights. The model's performance was tested with different oversampling repetitions.

Oversampling Repetitions	**MSE**	**Detected Cancellations**
0	0.031	146
10	0.045	11,970

Table 1: MSE and detected cancellations for different oversampling repetitions in January 2018

As is shown above, the recall is improved and more canceled examples are detected.

3.2 Ordinary Least Squares (OLS) for Delay Prediction

To predict flight delay times, we used an Ordinary Least Squares (OLS) regression model, which is appropriate for continuous outcome variables, such as delay duration.

Data Sampling Strategy Given the large volume of data and seasonal variations, a random sampling strategy was applied. To maintain a representative dataset without overloading computational resources, a subset of flights was randomly sampled each month. This approach allowed the model to capture monthly trends in delays, accommodating seasonal weather patterns and holiday travel surges while ensuring model efficiency.

Feature Selection and Engineering Features for delay prediction included continuous variables like dew point temperature, visibility, wind speed, and humidity, as well as categorical variables indicating the day of the week and the month. To enhance the model's interpretability and capture interactions between these features, polynomial and interaction terms were added selectively based on exploratory data analysis. This helped account for complex relationships between weather conditions and delays.

Model Performance and Validation The OLS model was assessed using standard regression metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), to evaluate how accurately the model could predict delay times. To illustrate the model's performance, we use the data from November 2023 as an example. The OLS model, when applied to this dataset, yielded a Mean Squared Error (MSE) of approximately 1810. This indicates that the model's average prediction error is around 40 minutes. Considering the actual situation that airports do not issue a delay report only once but may do so several times, this means that airports cannot provide an accurate delay time either. Therefore, our error is reasonable.

4 Flight Cancellation Analysis

4.1 Key Dates and Days of the Week

The following dates and days are associated with flight cancellations:

- Day 21 of the Month: Shows a lower probability of cancellation (coefficient = -0.576).
- Day 20 of the Month: Also associated with reduced cancellation risk (coefficient = -0.565).
- Monday: Indicates a higher likelihood of cancellation (coefficient = 0.507).

4.2 Weather Factors

The following weather conditions impact cancellation rates:

- Relative Humidity: Higher humidity levels increase cancellation risk (coefficient = 0.384).
- Dew Point Temperature: Lower dew point temperatures are associated with reduced cancellations (coefficient = -0.337).
- Wind Speed: Higher wind speeds correlate with higher cancellation probability (coefficient = 0.326).

5 Flight Delay Analysis

5.1 Key Dates and Days of the Week

The following days of the week show a significant impact on delays:

- Monday, Friday, and Sunday: These days have a markedly higher probability of delays (coefficient = 3.2546×10^{13}).

5.2 Weather Factors

The following weather factors are associated with delay durations:

- Dew Point Temperature: Lower dew point temperatures are linked to shorter delay times (coefficient = -7.411).
- Relative Humidity: Higher humidity levels contribute to longer delays (coefficient = 4.989).
- Visibility: Lower visibility is associated with increased delay times (coefficient = -3.447).

6 Advice for Flight Management

Based on the analysis, we provide the following recommendations for improving flight operations:

- **Monitor High-Risk Periods:** Given that flights later in the month and on transition days (Friday, Sunday, and Monday) tend to face more delays, consider adjusting schedules or increasing staffing during these times to mitigate impact.
- **Weather-Based Adjustments:** With higher wind speeds and humidity correlating to increased cancellations, real-time weather monitoring systems should be used to proactively manage flights. Planning for alternative flights in extreme conditions can reduce last-minute disruptions.

7 Contribution

- **Shuo Li**: Data cleaning, model building, coding, web scraper, summary writing
- **Yunze Wang**: Model building, coding, summary writing, GitHub maintenance
- **Xinrui Zhong**: Model building, Shiny App development

References

[1] Wikipedia contributors. Coordinated universal time — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Coordinated_Universal_Time&oldid=1255605162, 2024. [Online; accessed 11-November-2024].

[2] Wikipedia contributors. Haversine formula — Wikipedia, the free encyclopedia, 2024. [Online; accessed 11-November-2024].