Airline Project

Zhengyong Chen, Xiangsen Dong, Xupeng Tang, Zhaoqing Wu

Overview

- Data Collection & Cleaning
- Data Modeling
- Shiny App

Data Collection & Cleaning

Data Collection

We collected flight data from *Bureau of Transportation Statistics*, and gathered weather data from *National Centers for Environmental information* based on airports' latitude and longitude.

Data Cleaning & Merging

- Excluding data during the COVID period and flights without matching weather stations or without hourly weather data.
- Imputing missing weather values via forward filling before merging and KNN(N=5) after merging.
- Converting departure and arrival times to CST. In the first two models, departure and arrival times were categorized by periods of the day.
- Adding a *Holiday period* variable to indicate departures during Thanksgiving, Christmas, or New Year's.
- We merged the flight data with the weather data based on the columns for departure airport, arrival airport, and scheduled departure time (CST).

Model 1： Flight Cancellation Prediction

Data Overview

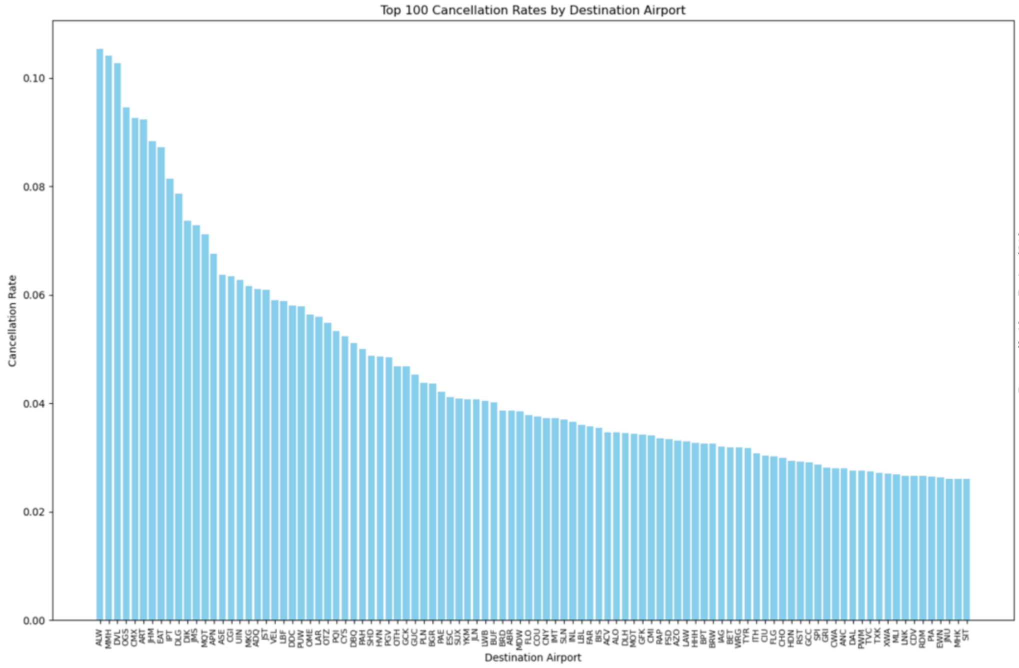
Our data contains 7,248,726 records regarding 378 airports.

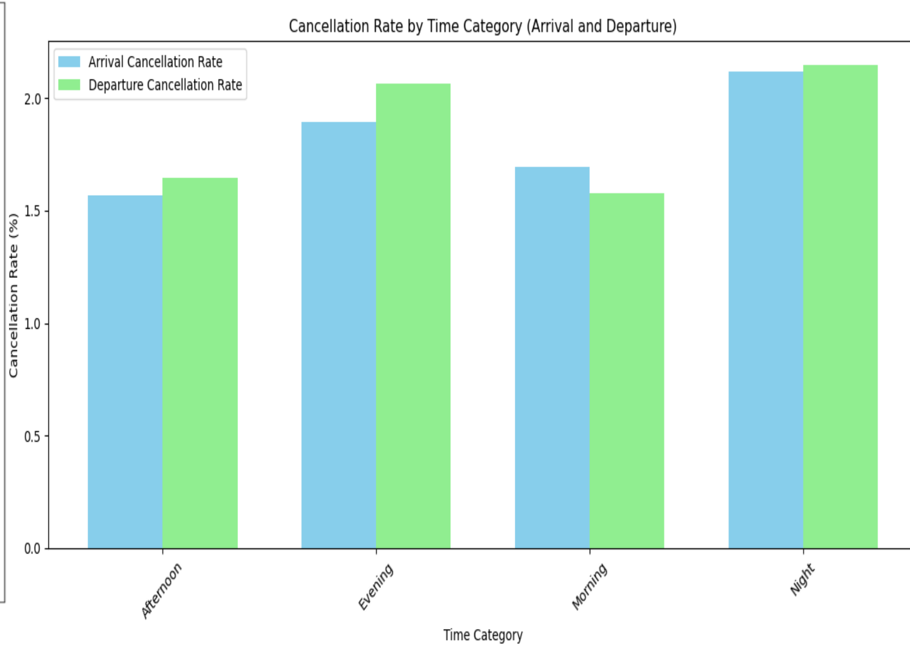
We considered cancelled flight as outcome and 32 predictors:

- ❖ Flight features: Month, Day of Week, Departure Time, Arrival Time, Holiday_Period, Operating Carrier, Origin, Destination, CRS Elapsed Time, Distance
- ❖ Weather features(Origin and Destination): Dew Point Temperature, Dry Bulb Temperature, Precipitation, Pressure Change, Pressure Tendency, Relative Humidity, Sea Level Pressure, Station Pressure, Visibility, Wet Bulb Temperature, Wind Speed

We used One-Hot Encoding for categorical variables.

Exploratory Data Analysis

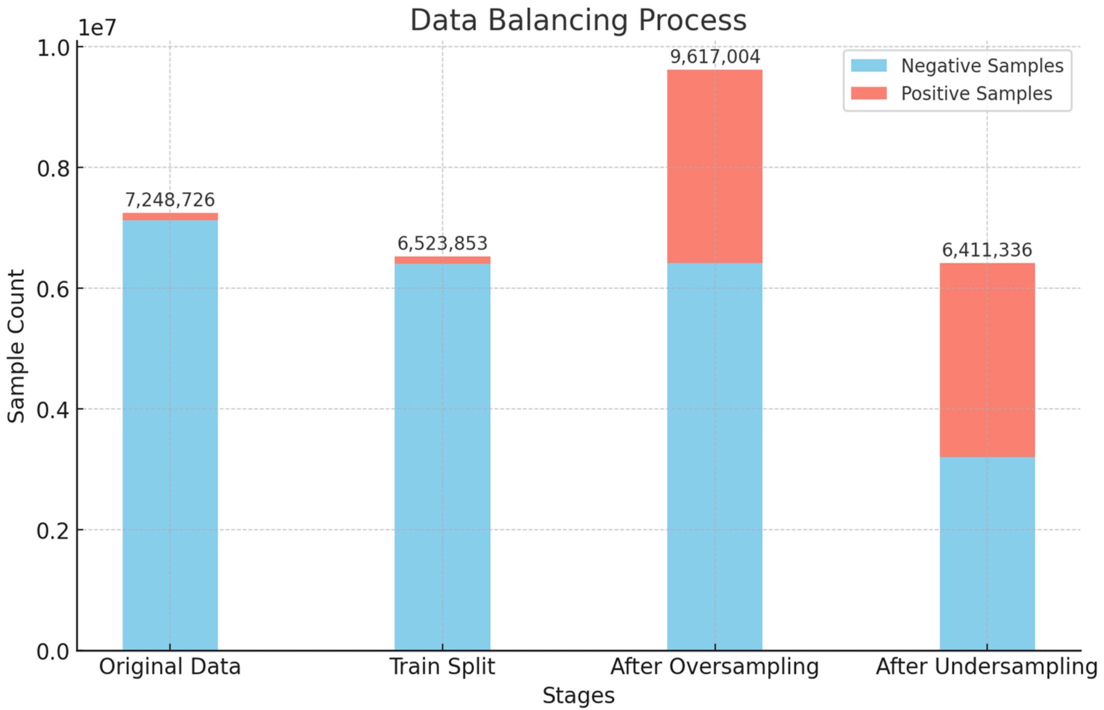




Problem to Solve: Data Imbalance

- Only 1.73% of flights are cancelled, data imbalance was a major issue.
- A logistic regression model fitted on this data yielded 99.98% prediction accuracy for non-canceled flights(negative samples) but only 1.45% for canceled ones(positive samples).

Data Balancing



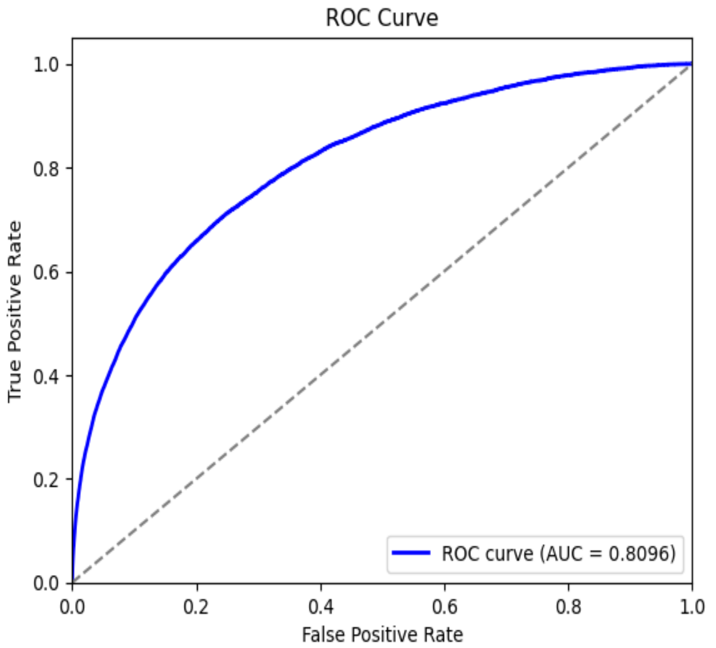
- ❖ Splitting training and test set (9: 1)
- ❖ Oversampling on the training set to increase positive samples to half the number of negative samples
- ❖ Undersampling to achieve a 1:1 ratio

Logistic Regression

The logistic regression model trained on the balanced dataset shows a significant improvement in prediction accuracy for positive samples.

Model	Accuracy for Cancelled Flights	Accuracy for Uncancelled Flights
Model on Unbalanced Data	1.45%	99.98%
Model on Balanced Data	71.23%	74.94%

Table 1: Model Prediction Accuracy Comparison



Odds Ratio

$$OR = e^{coefficient}$$

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

Table 2: Variables and their Odds Ratios

OR>1 : the variable raises the odds of cancellation

OR<1 : the variable reduces the odds of cancellation

Tips to avoid cancelled flights

- ❖ Choose Delta Air Lines
- ❖ Travel outside holiday periods
- ❖ Select morning flights
- ❖ Avoid popular tourist destinations(like Orlando and San Francisco)
- ❖ Opt for days with high visibility and low wind speeds

Model 2:Flight Delay Prediction

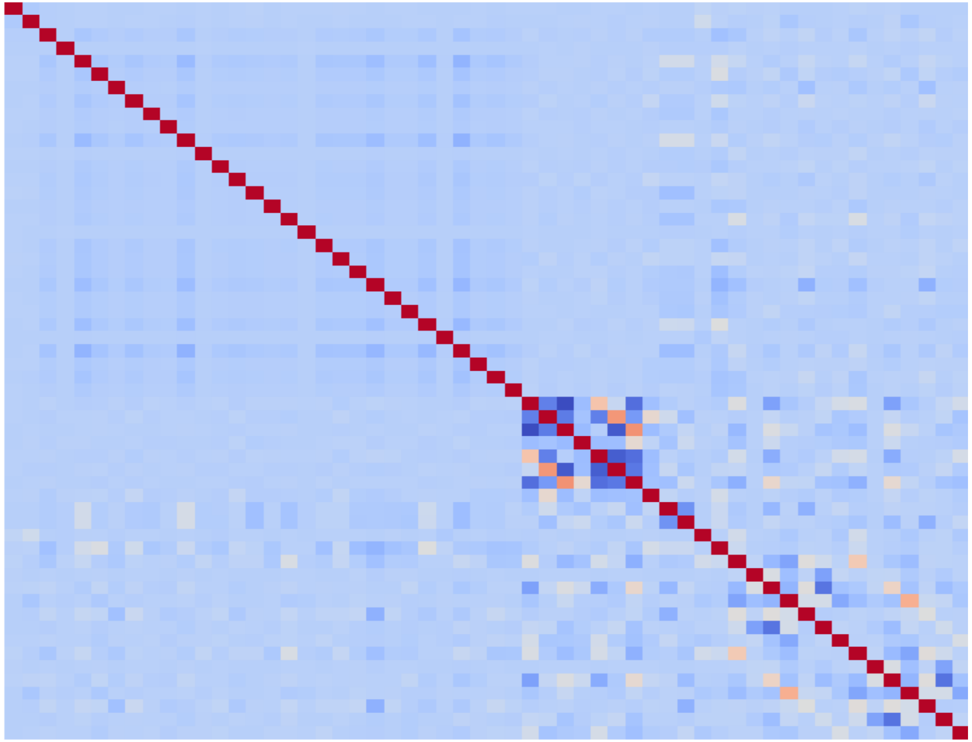
Data Overview

Our dataset contains 7,248,726 records from 378 airports.

We considered cancelled flight as outcome and 32 predictors:

- Flight features: Month, Day of Week, Departure Time, Arrival Time, Holiday Period, Operating Carrier, Origin, Destination, CRS Elapsed Time, Distance
- Weather features(Origin and Destination): Dry Bulb Temperature, Precipitation, Relative Humidity, Sea Level Pressure, Station Pressure, Visibility, Wind Speed

Correlation Checking



Heatmap of variables’ correlation

Categorical data processing

	One hot encoder	Frequency encoder
Strength	Information preserving Easy to handle	Efficient
Weakness	Increase number of features Low efficiency	Information loss

One hot encoded variables	Frequency encoded variables
Operating Carrier, Time period, Day of week	Airport

Influencing variables

Most significant	Moderately significant	Least significant
Origin_Visibility Origin_Dry Bulb Temperature Dest_Visibility Operating Carrier_DL Dest_Wind Speed	Origin_Wind Speed Operating Carrier_HA Arrival time_Morning Arrival time_Evening Dest_Relative Humidity Origin_Precipitation Dest_Precipitation	Airport, Station_Pressure SeaLevel_Pressure

Tips to avoid flight delay

- Choose Delta Airlines
- Travel on a warm, clear day
- Avoid windy or rainy days
- Avoid Hawaii Airlines

Model metrics

Accuracy: 0.5887

Precision: 0.4487

Recall: 0.5679

F1 Score: 0.5013

ROC AUC: 0.6192

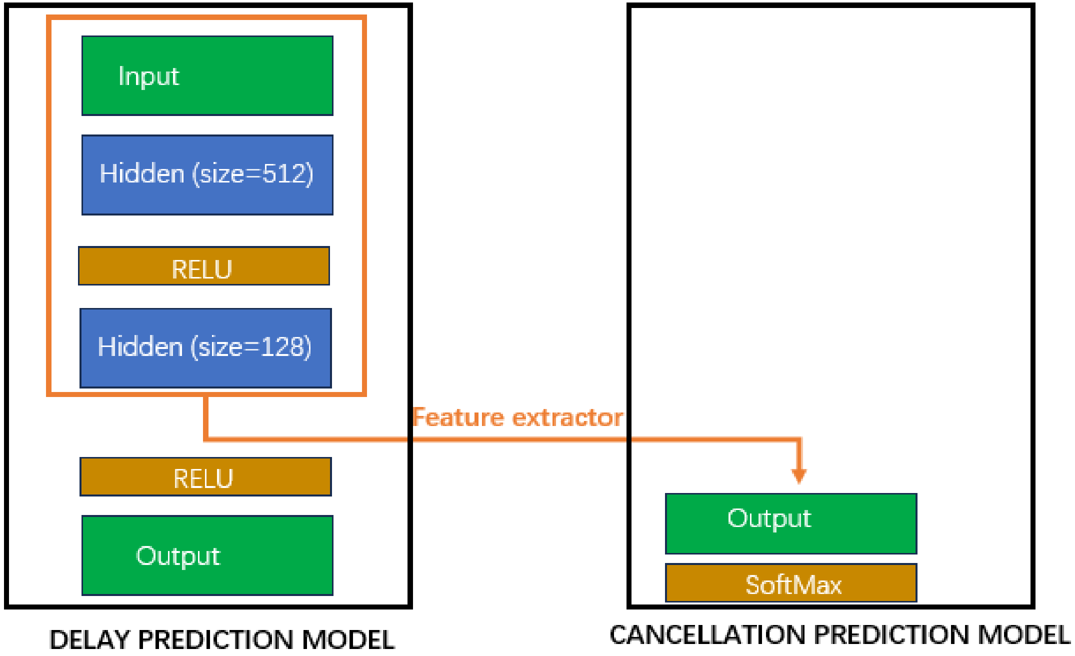
Average Precision: 0.4833

Model 3 : Delay Time prediction

Feature

- Weather feature: Almost all weather features except weather type
- Airline feature:
  - Operating Carrier(encoded)
  - Distance (numeric)
- Location feature:
  - Longitude/Latitude (numeric)
  - Origin/Destination Airport (encoded)
- Time feature:
  - Month/Day of the week(encoded)
  - Date/Date_CST (Here, I should have encoded the hours, but I processed them as continuous data)

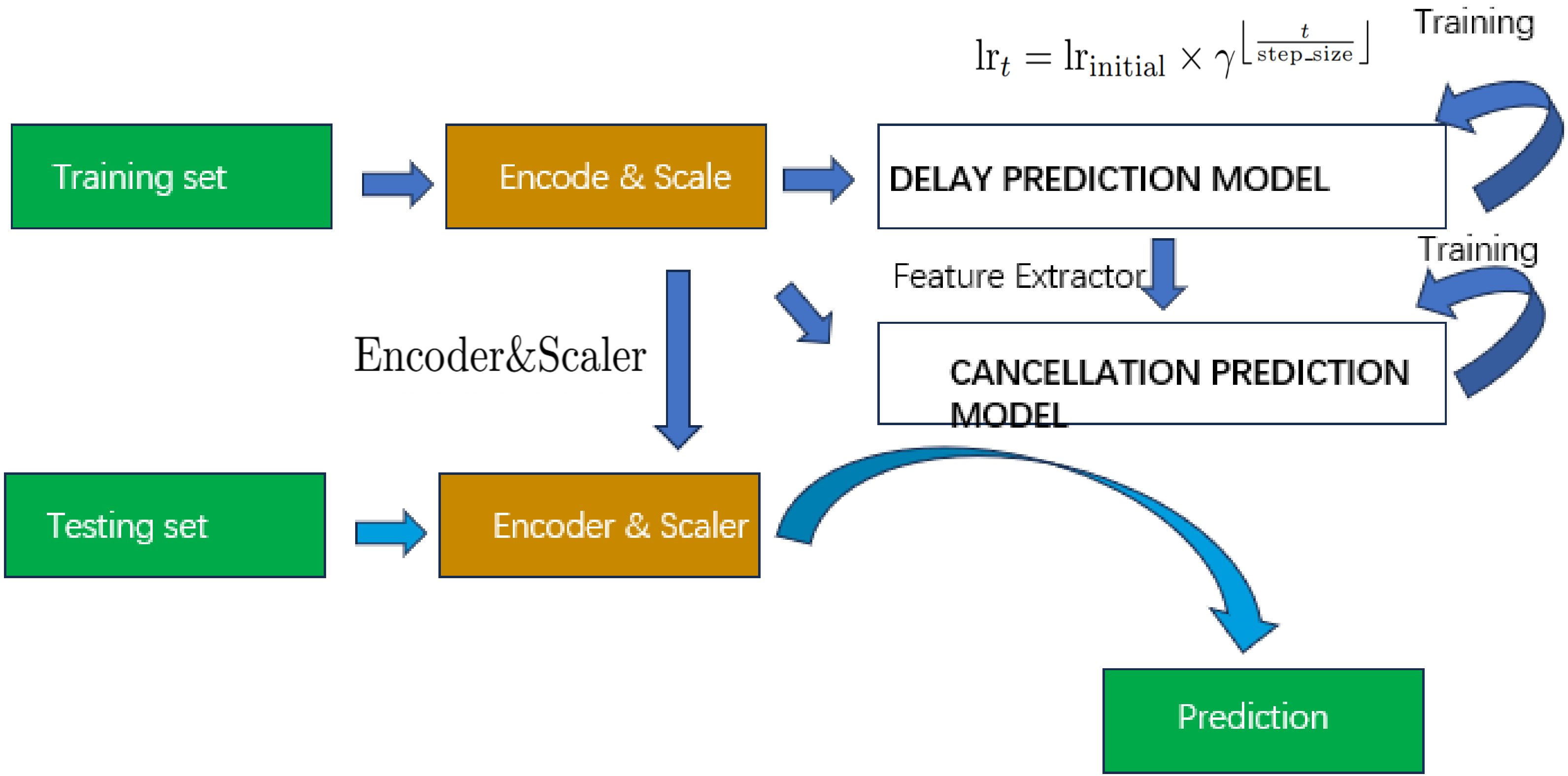
Model structure



Process

Optimizer = ADAM
Scheduler = stepLR

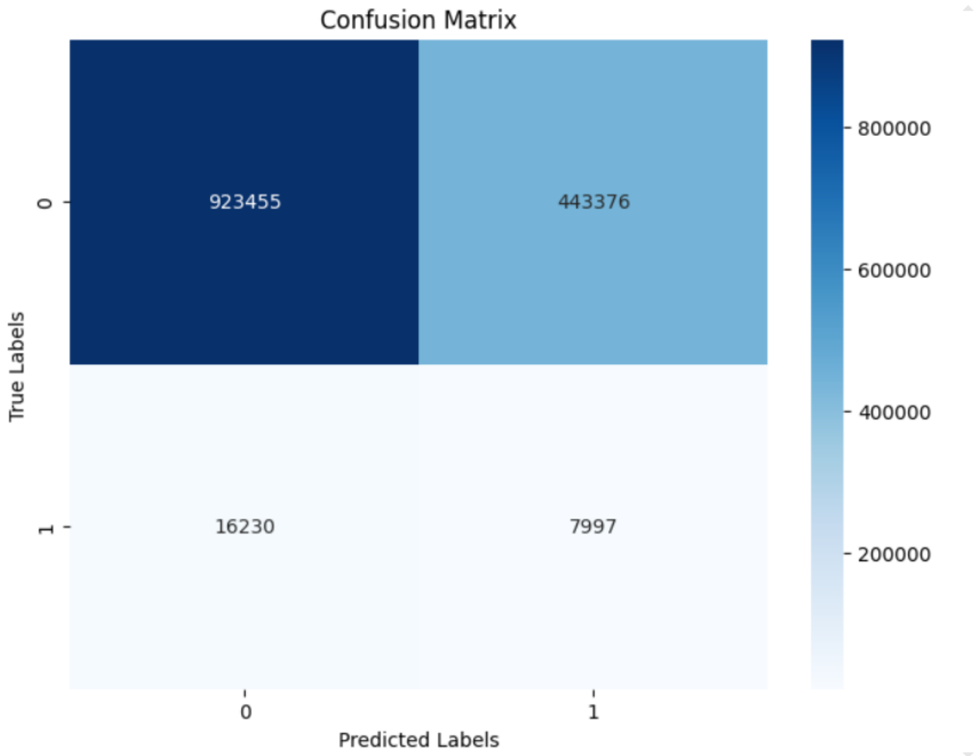
$$\mathrm{lr}_t = \mathrm{lr}_{\mathrm{initial}} \times \gamma^{\left\lfloor \frac{t}{\mathrm{step\_size}} \right\rfloor}$$



Results

- Our delay prediction neural network reached an RMSE of 39 minutes on the test set.

- Our combination flight cancellation prediction network reached recall (0.6796) for cancelled case and recall (0.3408) for uncancelled case.



Strength and weakness

- Advantages:
  - Our method can reduce the cost required to train classification models.
  - Our model architecture can be easily expanded and has great potential.

- Disadvantages:
  - The performance of our model is not sufficient. We should add more layers.
  - Poor interpretability.

Torch model to skearn model

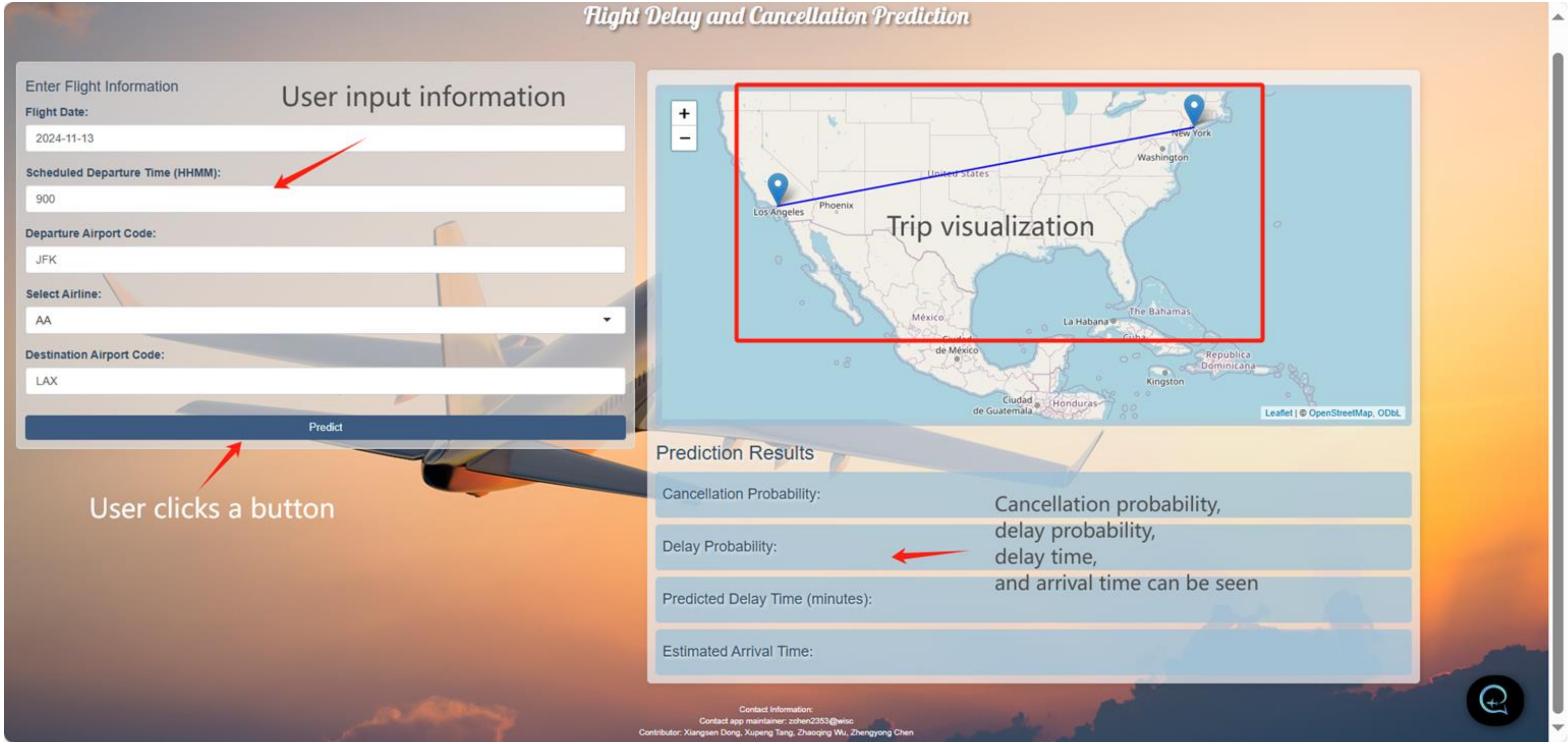
- ❖ Question Statement:
  - ➢ Our online Shiny application needs to call an online Python environment, but Shiny for Python does not have the PyTorch package pre-installed.
  - ➢ Every time we download the PyTorch package online, the environment gets cleared when the webpage is closed.

- ❖ Proposal Step:
  - ➢ Step1 Redefine the model using sklearn's neural network implementation.
  - ➢ Step2 Extract all the parameters from the PyTorch model, convert them to arrays, and import the parameter arrays into the sklearn model.

Shiny App

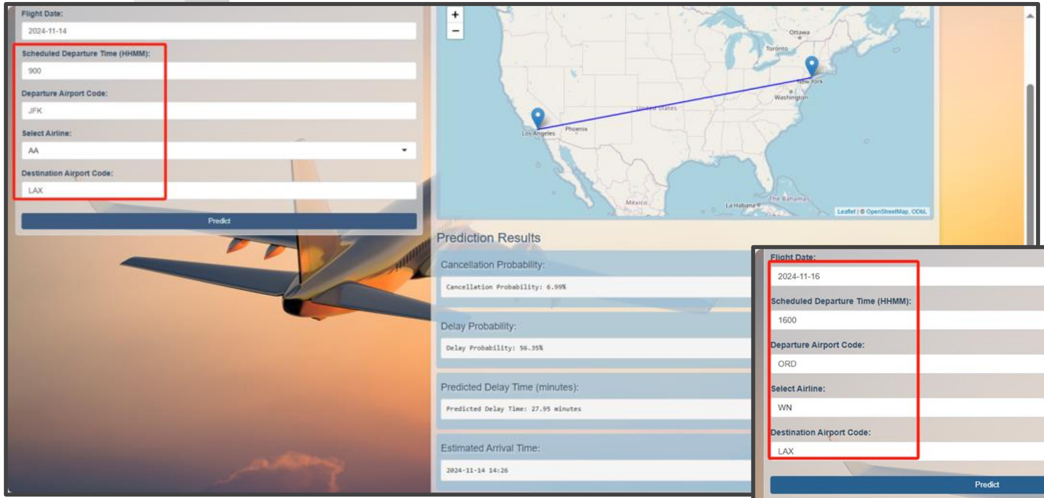
Shiny App - Flight Delay and Cancellation Prediction

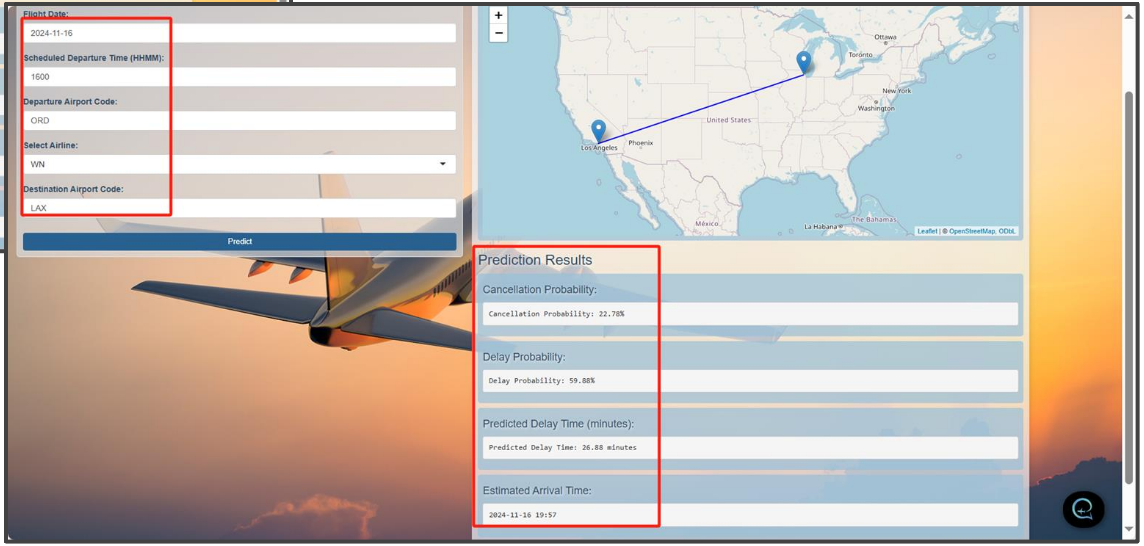
https://andrewchanshiny.shinyapps.io/Group10_P3/



Shiny App - Flight Delay and Cancellation Prediction

https://andrewchanshiny.shinyapps.io/Group10_P3/





Think you!