

# 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.



A bar chart titled 'Number of flights from each destination airport to New York City'. The x-axis is labeled 'Destination Airport' and lists 50 airports. The y-axis represents the number of flights, with a scale from 0 to 100. The bars are arranged in descending order of height. The first bar, for LAX, is the tallest at approximately 95 flights. The bars decrease in height as they move to the right, with the final bar, for BNA, being the shortest at approximately 1 flight.

Destination Airport	Number of Flights (approx.)
LAX	95
DAL	90
ORD	85
PHX	80
DFW	75
SEA	70
MDW	65
ATL	60
DTW	55
IAH	50
ANC	45
PHO	40
ABQ	35
ELI	30
DFC	25
MDX	20
LAS	15
POB	10
DBQ	5
PHN	4
OTM	3
CGC	2
ERI	1
STL	1
SLC	1
SFO	1
IAA	1
MDI	1
COU	1
MTT	1
RLI	1
BJS	1
ALG	1
MDT	1
CAK	1
OTF	1
PHF	1
BRW	1
MEG	1
VRG	1
FTT	1
ALC	1
FDN	1
SCF	1
SHR	1
OWB	1
PHL	1
YXK	1
MLB	1
EDW	1
RAH	1
BNA	1
MSK	1
ST	1

A bar chart comparing the Arrival Cancellation Rate (blue bars) and the Departure Cancellation Rate (green bars) across four time categories: Afternoon, Evening, Morning, and Night. The Y-axis represents the Cancellation Rate in percentage, ranging from 0.0 to 2.0. The chart shows that the Departure Cancellation Rate is generally higher than the Arrival Cancellation Rate, with both rates peaking in the Night category.

Time Category	Arrival Cancellation Rate (%)	Departure Cancellation Rate (%)
Afternoon	1.55	1.65
Evening	1.90	2.05
Morning	1.70	1.55
Night	2.15	2.20



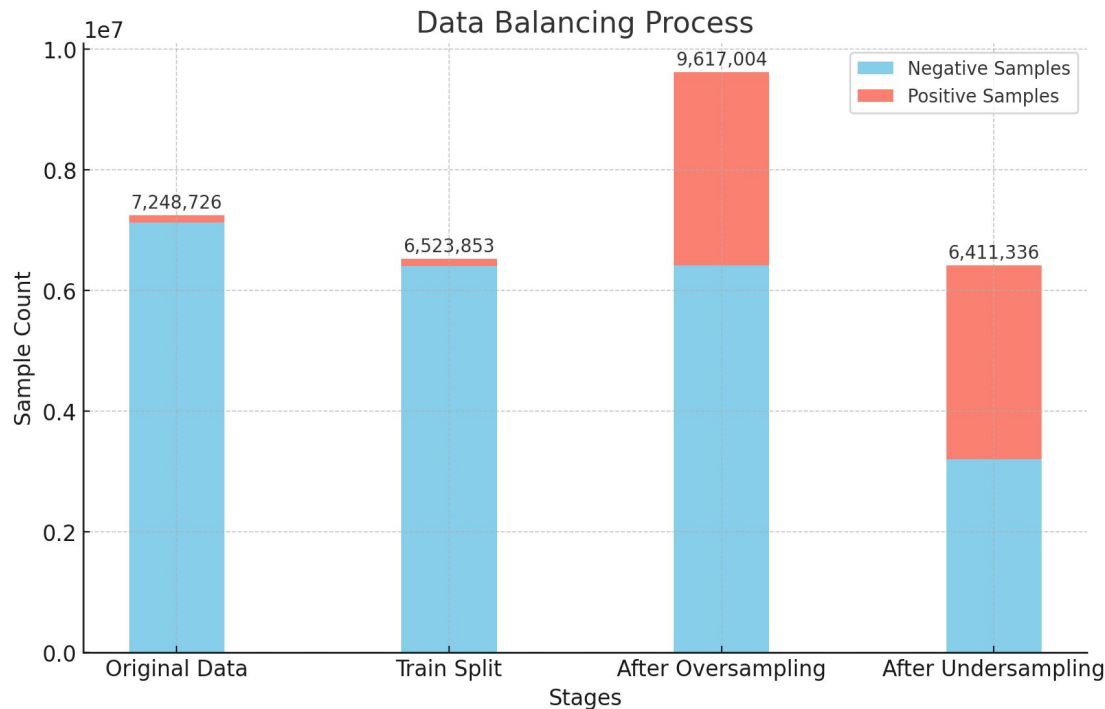


## Problem to Solve: Data Imbalance

- ❖ 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-cancelled flights(negative samples) but only 1.45% for cancelled 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

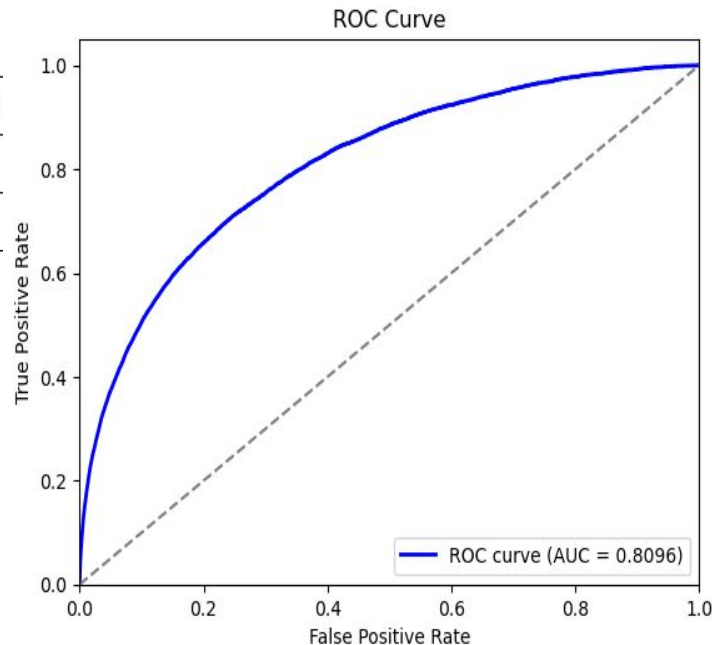


# Logistic Regression

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

$$OR = e^{\text{coefficient}}$$

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

OR>1 : the variable raises the odds of cancellation

OR<1 : the variable reduces the odds of cancellation

Table 2: Variables and their Odds Ratios



## 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**



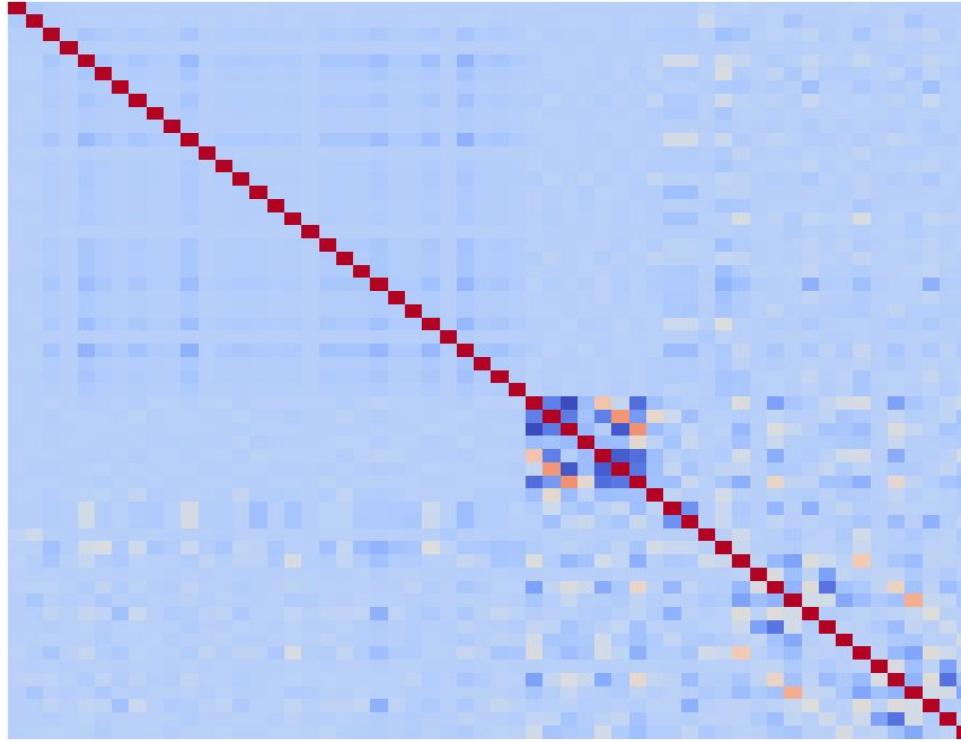
# Data Overview

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

Operating Carrier, Time period, Day of week

Frequency encoded variables

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**

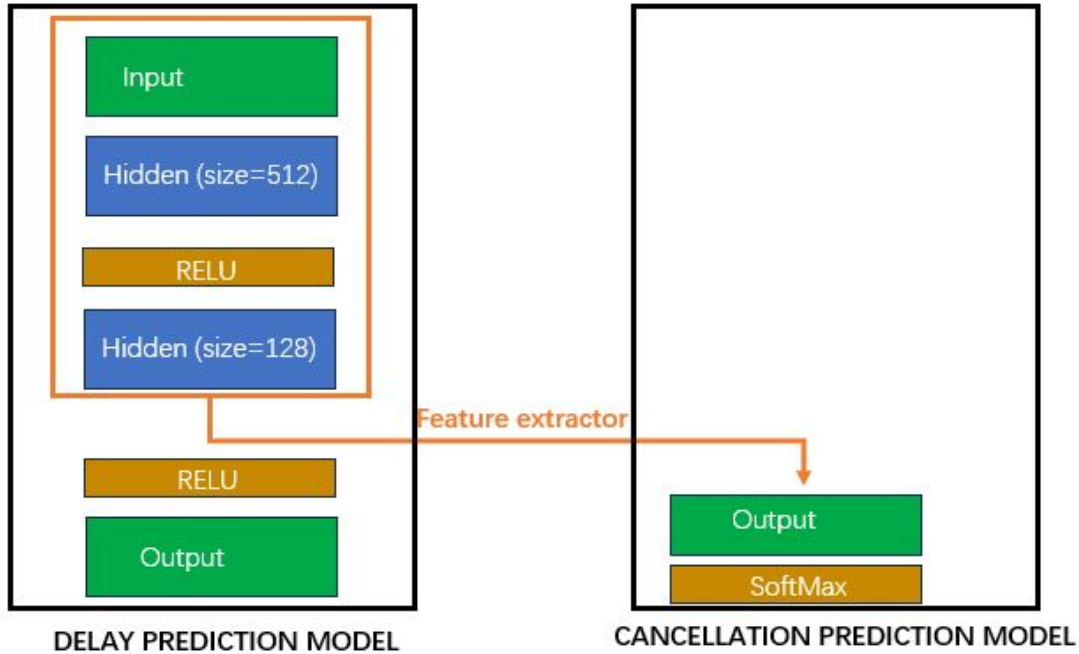


# Feature

- ❖ 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)



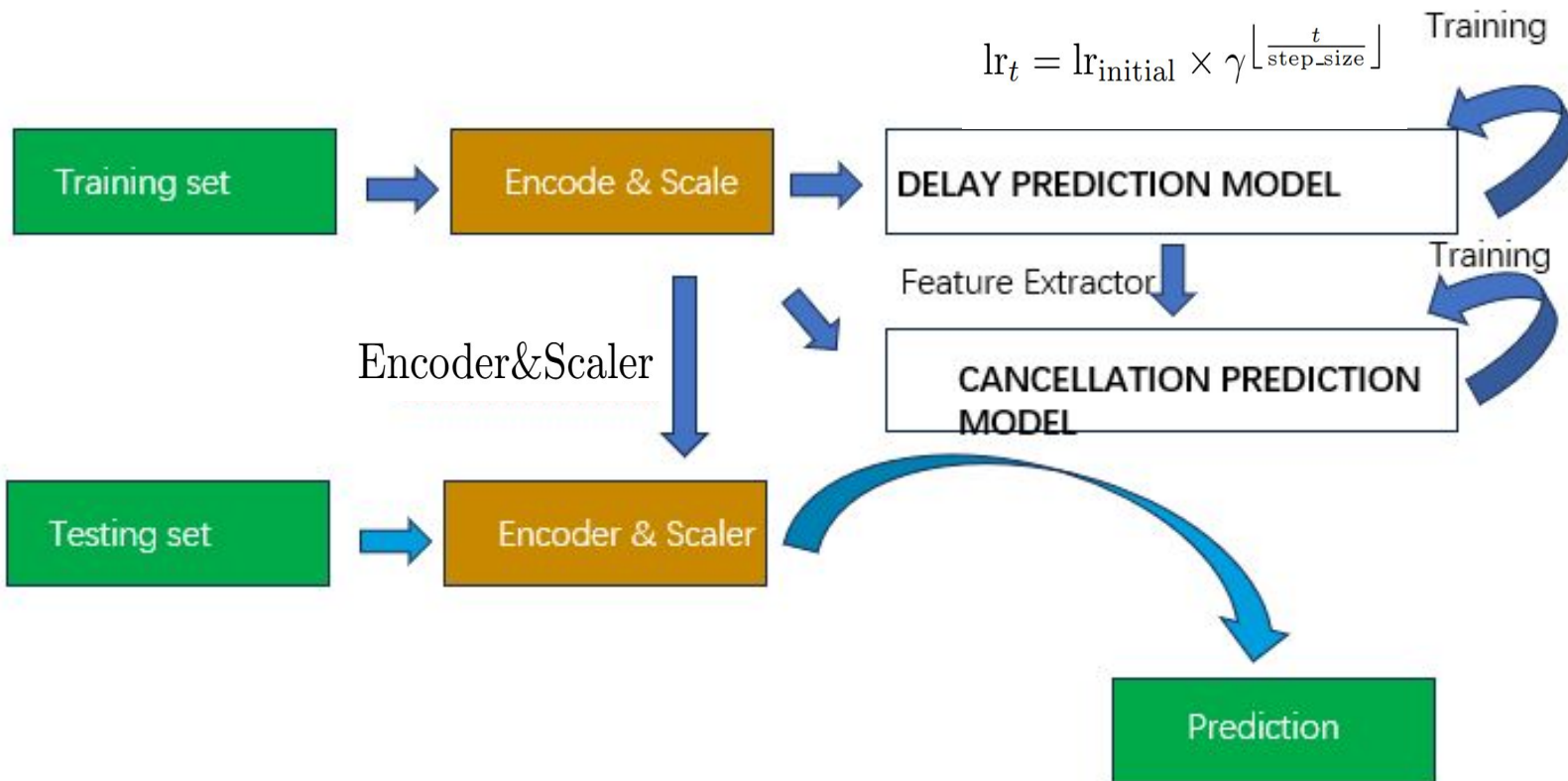
# Model structure





Scheduler = stepLR

$$\text{lr}_t = \text{lr}_{\text{initial}} \times \gamma^{\lfloor \frac{t}{\text{step-size}} \rfloor}$$





# 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 uncanceled 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.



# Shiny App

# Shiny App - Flight Delay and Cancellation Prediction

[https://andrewchanshiny.shinyapps.io/Group10\\_P3/](https://andrewchanshiny.shinyapps.io/Group10_P3/)

*Flight Delay and Cancellation Prediction*

Enter Flight Information

Flight Date:

2024-11-13

Scheduled Departure Time (HHMM):

900

Departure Airport Code:

JFK

Select Airline:

AA

Destination Airport Code:

LAX

Predict

User input information

Trip visualization

Prediction Results

Cancellation Probability:

Delay Probability:

Predicted Delay Time (minutes):

Estimated Arrival Time:

Cancellation probability, delay probability, delay time, and arrival time can be seen

User clicks a button

Contact Information:  
Contact app maintainer: zhen2953@eisc  
Contributor: Xiangsen Dong, Xupeng Tang, Zhaoping Wu, Zhengyong Chen

# Shiny App - Flight Delay and Cancellation Prediction

[https://andrewchanshiny.shinyapps.io/Group10\\_P3/](https://andrewchanshiny.shinyapps.io/Group10_P3/)

Flight Date: 2024-11-14

Scheduled Departure Time (HHMM): 900

Departure Airport Code: JFK

Select Airline: AA

Destination Airport Code: LAX

Predict

Prediction Results

Cancellation Probability: Cancellation Probability: 6.99%

Delay Probability: Delay Probability: 56.39%

Predicted Delay Time (minutes): Predicted Delay Time: 37.95 minutes

Estimated Arrival Time: 2024-11-14 14:26

Flight Date: 2024-11-16

Scheduled Departure Time (HHMM): 1600

Departure Airport Code: ORD

Select Airline: WN

Destination Airport Code: LAX

Predict

Prediction Results

Cancellation Probability: Cancellation Probability: 22.78%

Delay Probability: Delay Probability: 59.88%

Predicted Delay Time (minutes): Predicted Delay Time: 26.88 minutes

Estimated Arrival Time: 2024-11-16 19:57



**Think you!**