Group 2

Data Cleaning

1. Convert all times to a timezone-aware datetime format

2. Merging with Weather Data

3. Removed potential non-stationary data(2022 Seattle Alaska Issue)

4. Heavy Snowstorm during holiday season

Feature Engineering

- New Features: 1) Flight duration

  2) Time of day(late night, morning, noon, afternoon, night)

  3) Sky condition(CLR, FEW, SCT, BKN, OVC)

- Label Encoding: Airline type, DepAirport, ArrAirport

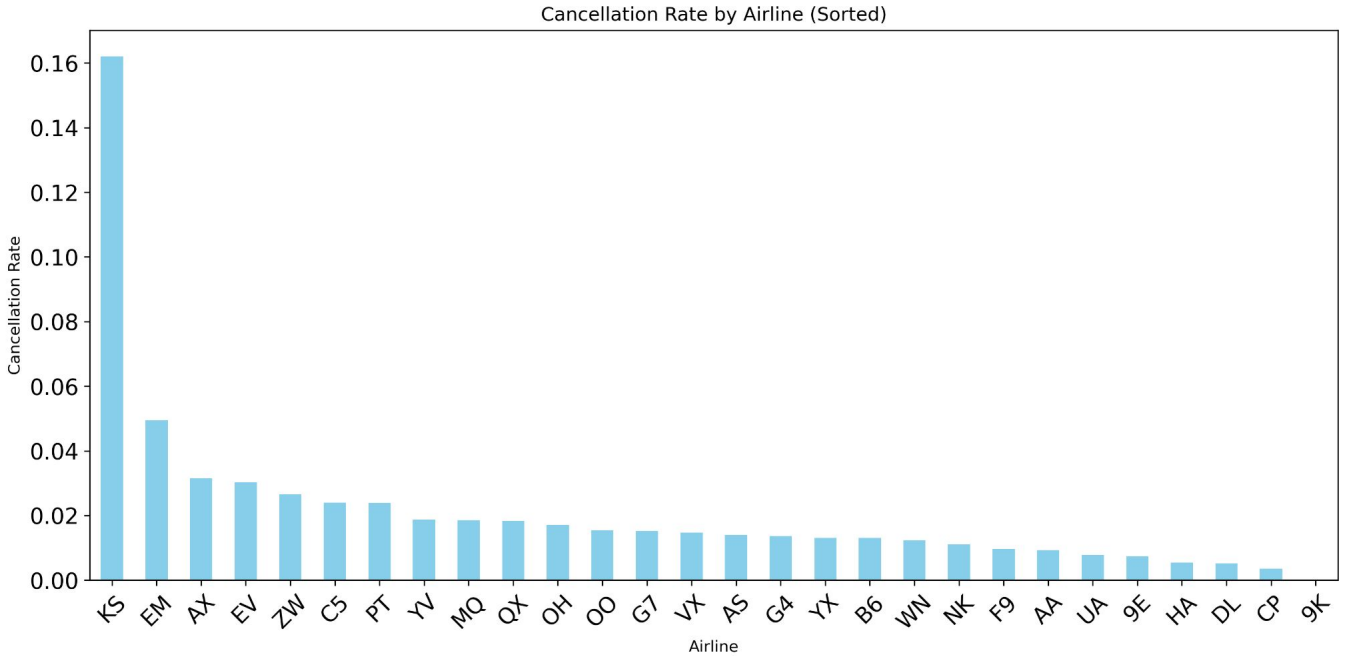
- Final Features: Airline_Type, Flight_Duration, Depart/Arrive Airport, Day of Week, Depart/Arrive Time, Sky Condition, Relative Humiditiy, Temperature, Windspeed, Visibility, Delay Time

  Total 21 Features based on the information will be easy to collect from passengers

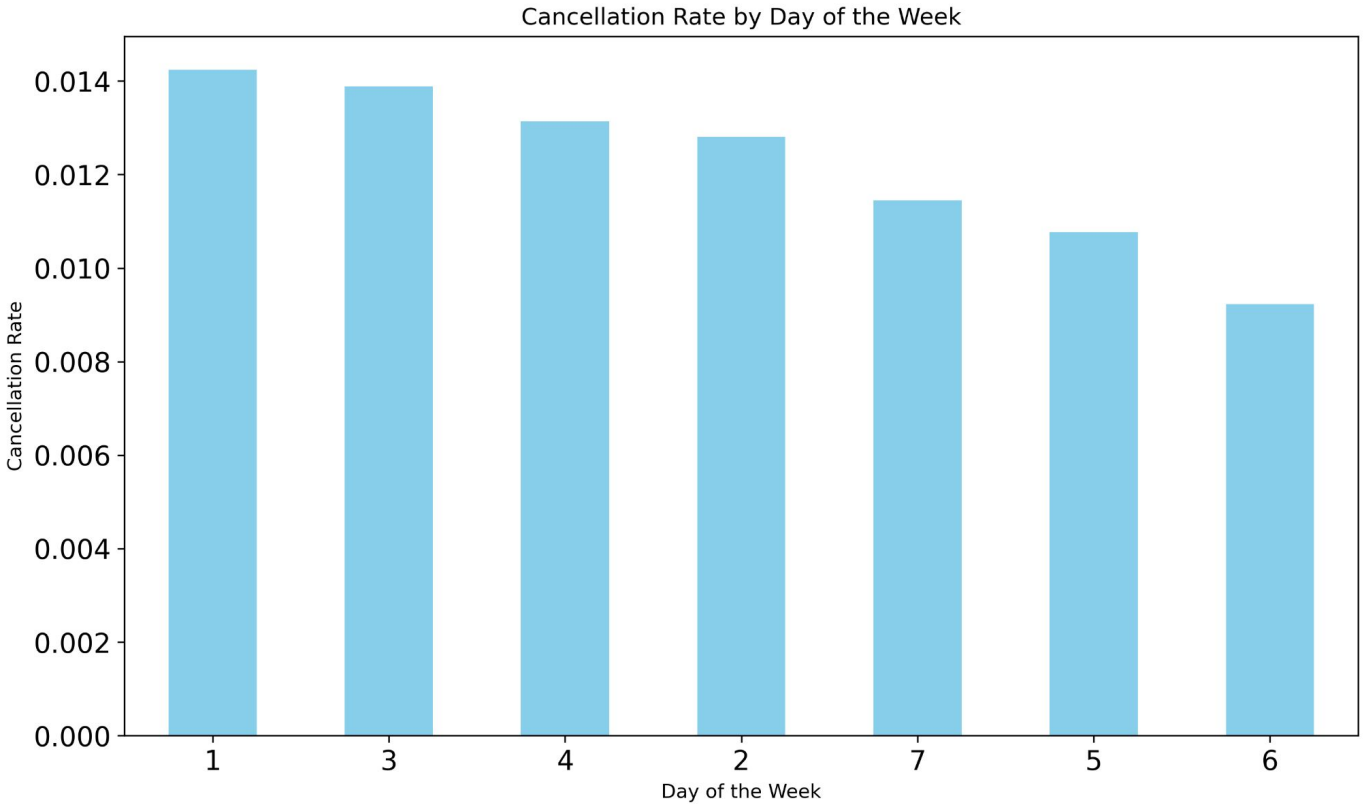
Task1：Tips to avoid cancellation flight during holiday seasons

Visualized Category Features: (Chi-Squared Test) ALL features are significant.

Airline_Type, Depart/Arrive Airport, Day of Week, Depart/Arrive Time,Sky Condition



Task1 :


Cancellation Rate by Day of the Week
Cancellation Rate
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1
3
4
2
7
5
6
Day of the Week

Task1 :

Continuous Features: Flight Duration, Relative Humidity, Temperature, Wind Speed, Visibility

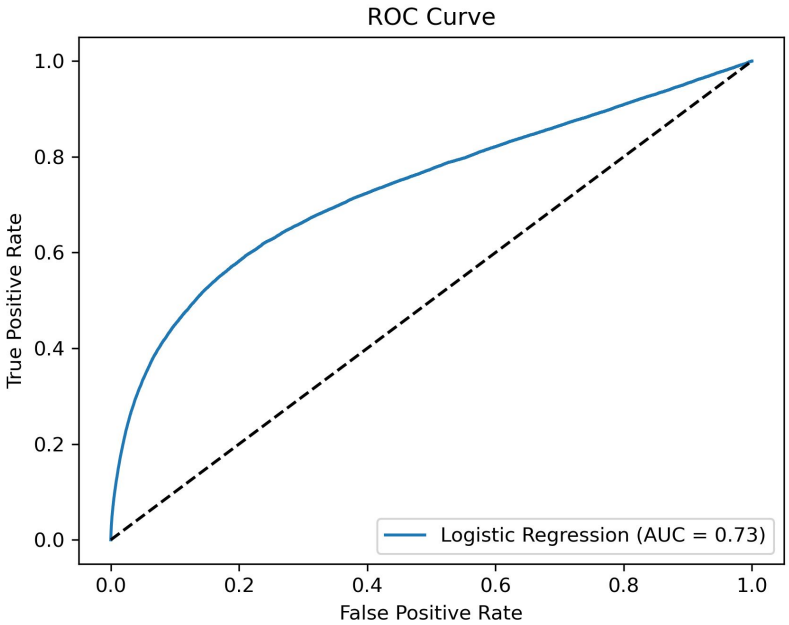
A T-Test was used to determine the significance of these features in relation to flight cancellations.

Result:

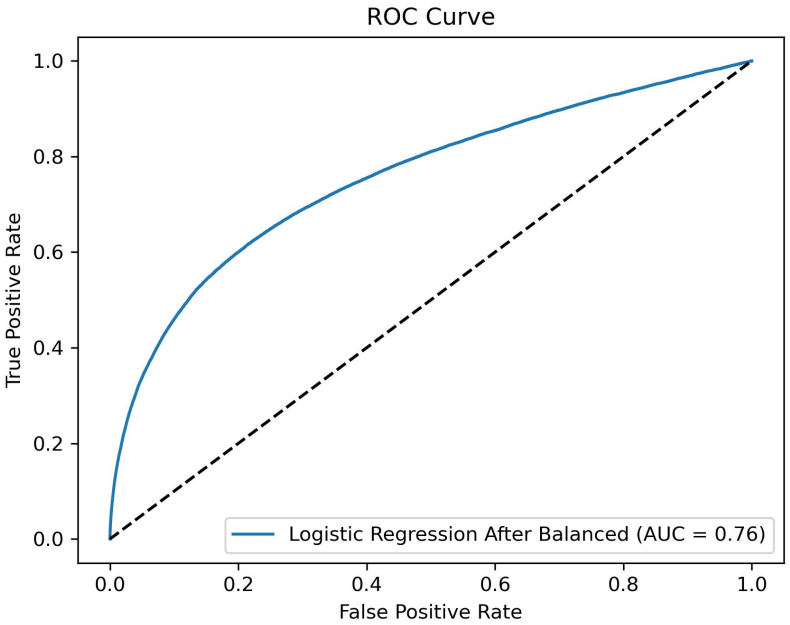
All of the P-Value less than 0.05 (Significant!!!)

Logistics Regression

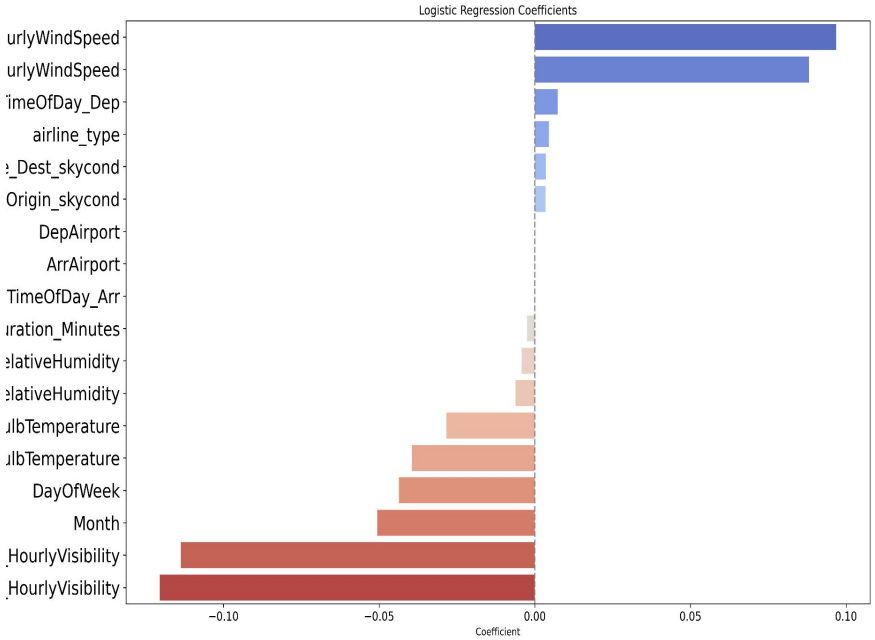
Before Balanced
Accuracy: 0.98
AUC = 0.73

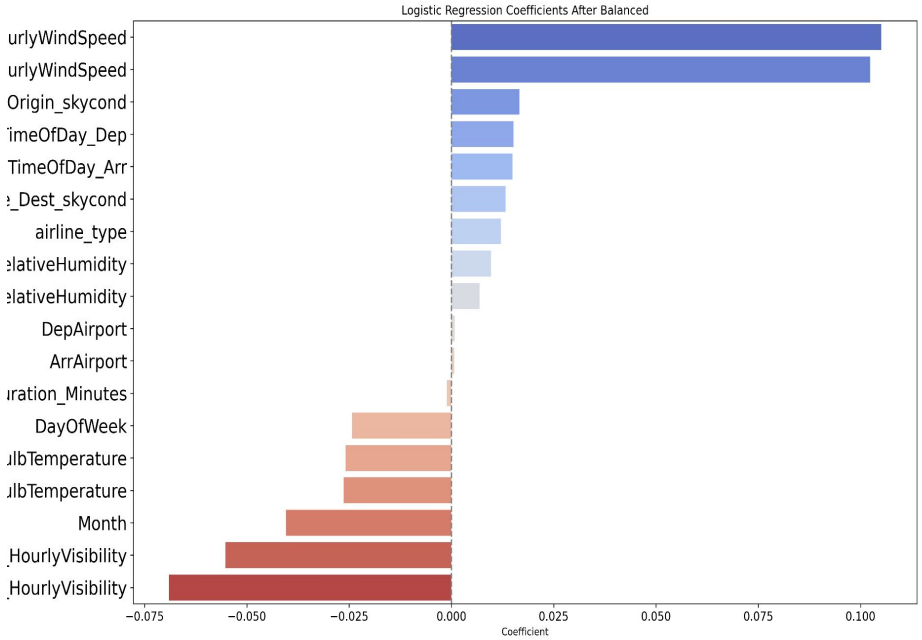


After Balanced
Accuracy: 0.73
AUC = 0.76



Logistic Coefficient





Results:

Before Balanced:

One units of Visibility increase, the probability of cancellation will decrease about 13%

One units of Windspeed increase, the probability of cancellation will increase 9%

After Balanced:

One units of Visibility increase, the probability of cancellation will decrease about 7%

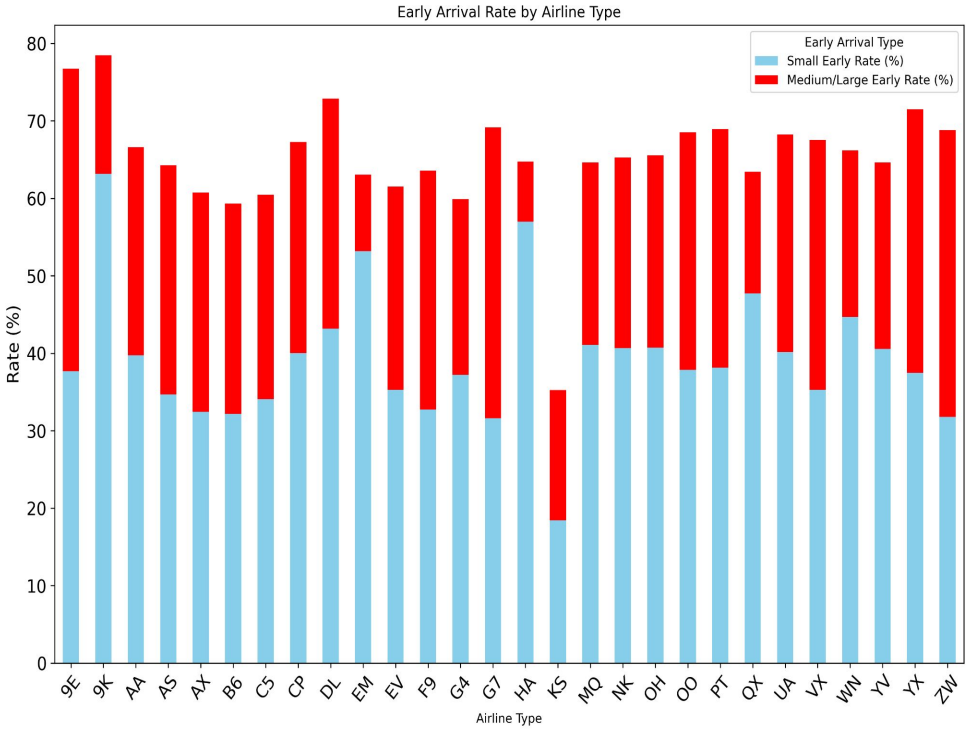
One units of Windspeed increase, the probability of cancellation will increase 10.7%

Task2：Tips to arrive early or on time

- Random Forest

Category in to 6 types:

1. Small Early (< 15 mins)
2. Medium Early (15-45 mins)
3. Large Early (> 45mins)
4. Small Delay (<15 mins)
5. Medium Delay (15-45 mins)
6. Large Delay (> 45 mins)



Task2：

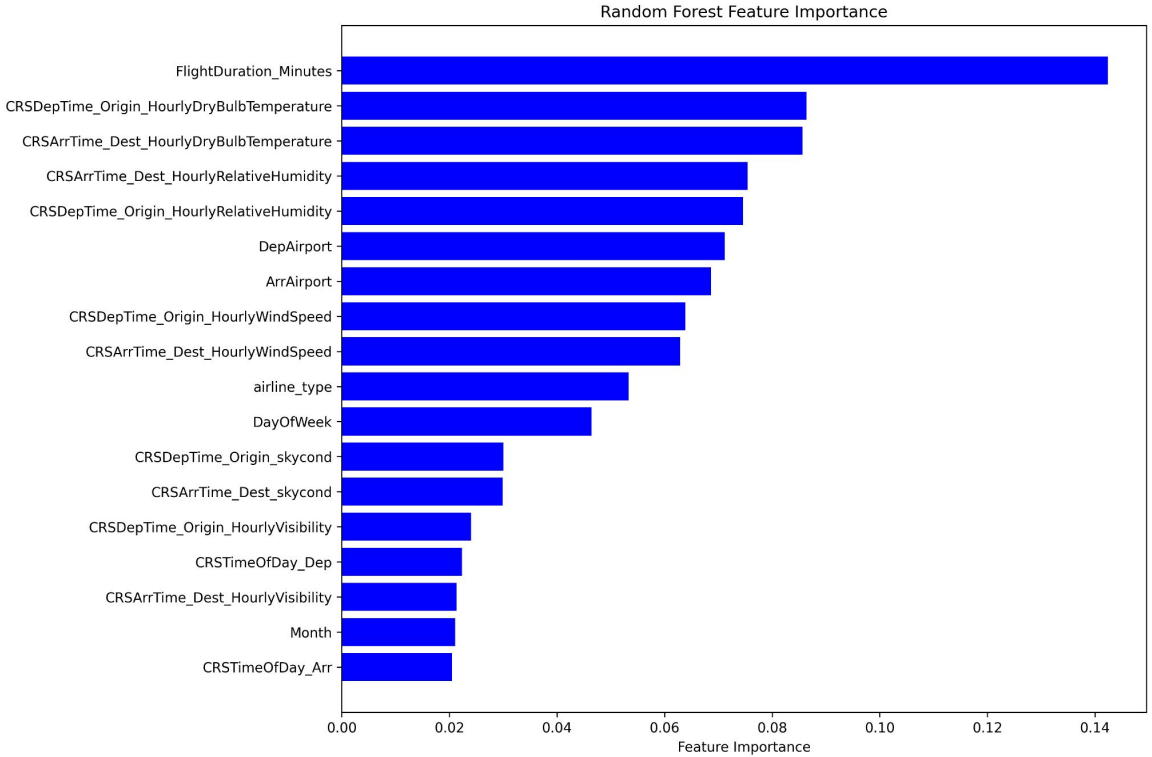
- Feature Importance

What we find out the feature importance?

Flight Duration?!?!?

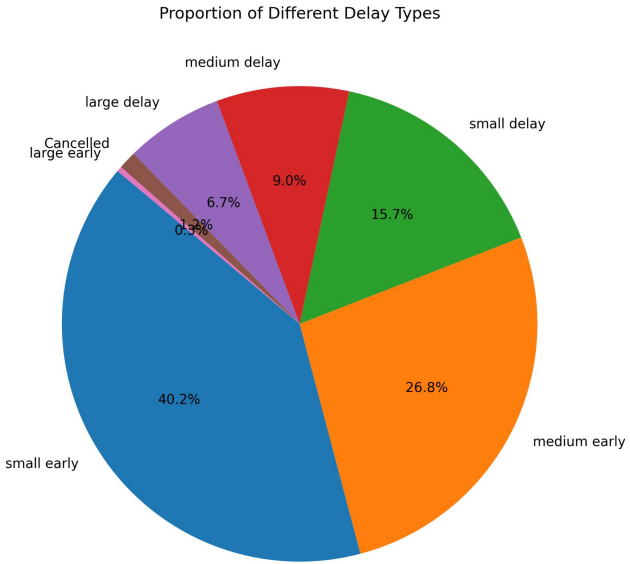
Wind Speed?!?!?

Visibility?!?!?



Task3：Prediction Model for Arrival Times

- Balancing data



Large early(0.3%): 28461 samples

resample：28461 * 6

Task3：Prediction Model for Arrival Times

- Model Performance

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

- Why low performance?

Shiny App

- https://jirenlu.shinyapps.io/flight_prediction/