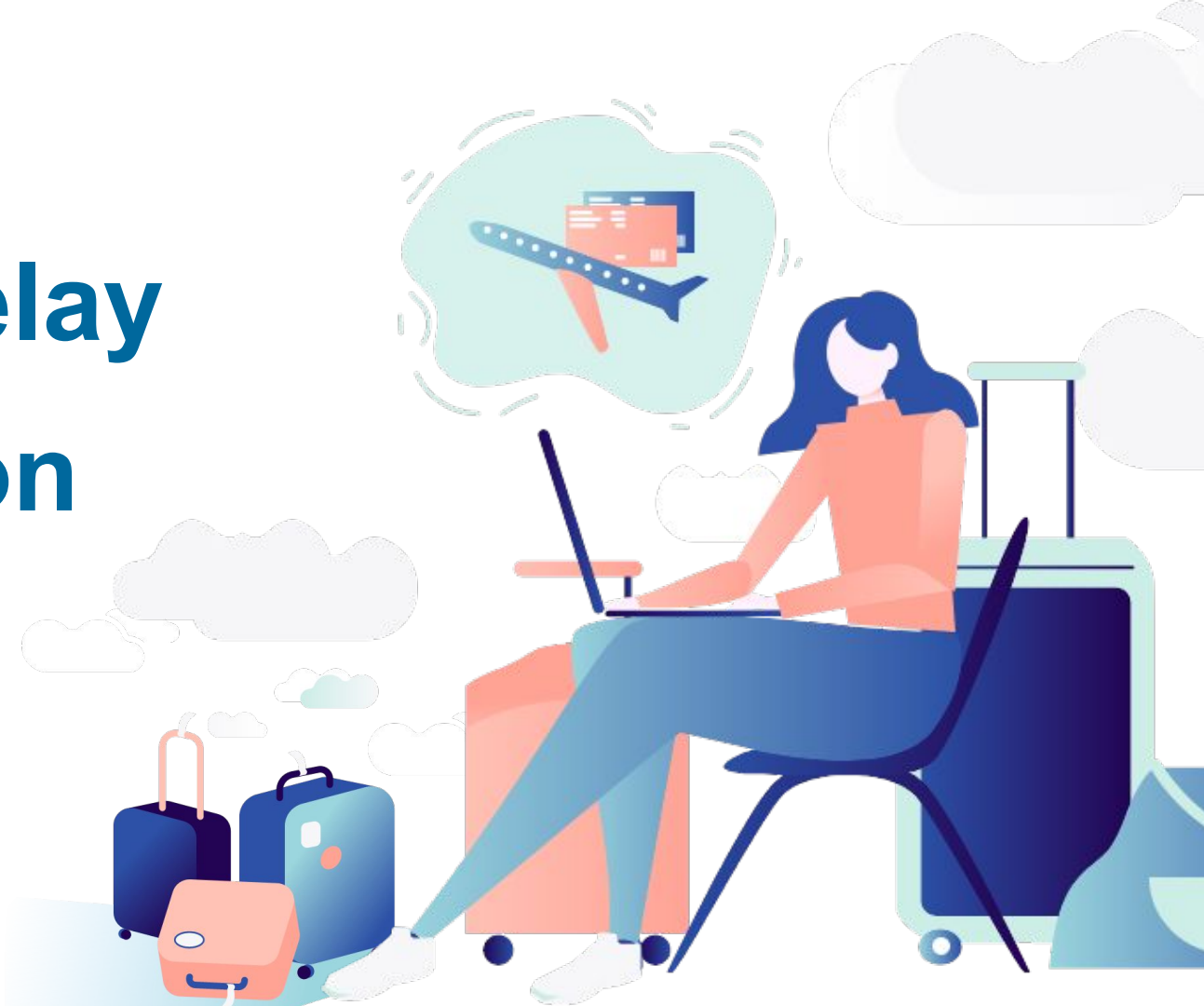
Flight Delay Prediction

Team 3:
ChiaJo Chen (Tina)
TzuYing Yu (Winnie)
Biying Han (Candice)
YuHsin Lee (Kathy)



Project Overview

Part 1: Introduction

Part 2: Data Cleaning

Part 3: Exploratory Data Analysis

Part 4: Model Analysis

Part 5: Key takeaways & Recommendations

Introduction

- **Goal**

  Predict whether a flight will be delayed

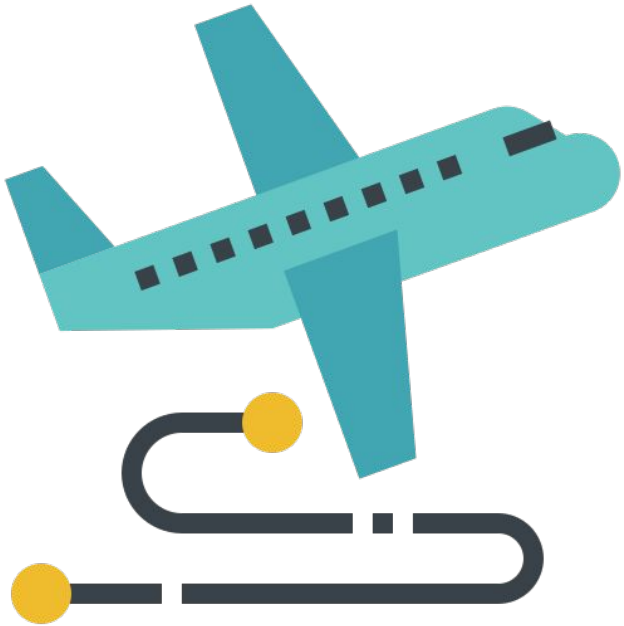
  (Binary variable, 0 = on time, 1 = delayed)

- **Data**

  Flight data and weather information

- **Model**

  Logistic regression, decision tree, and random forest



US Historical Flight Delay & Weather Data

December 2019

The dataset includes **35** variables and **679,996** observations.

1 Data Cleaning

Top 3 delay reason

- Carrier delay, late aircraft arrival delay, weather delay

Delay = delay minutes > 15 minutes

Transform **binary delay variable (0=ontime, 1=delay)** to all delay reason

Delete columns: station_x, station_y, month, year, date, flight number, etc

2 Filtering

Filter top 30 airports:

MDW, BNA, PDX, TPA, IAD, SAN, MIA, BWI, FLL, SLC, JFK, DCA, BOS, PHL, MSP, MCO, EWR, LAS, LGA, SFO, DTW, IAH, PHX, SEA, LAX, CLT, DEN, DFW, ATL, ORD

3 Training and Testing

Carrier Delay and Late aircraft arrival delay (Weather Delay is different)

- **Training**: Select 10,000 observation →Delay: 5000 & Ontime: 5000
- **Testing**: Select 2,500 observation → Delay: 1,250 & Ontime: 1,250

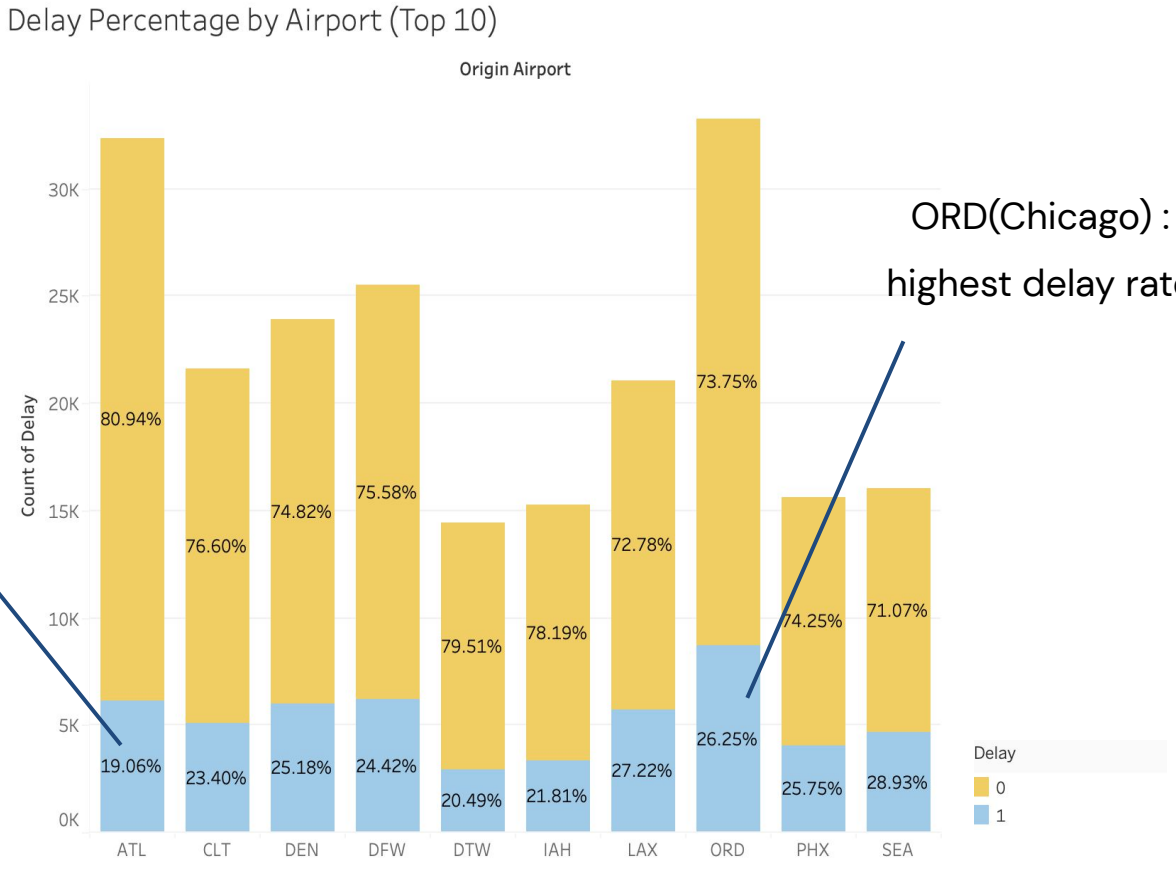
Delay Portion by Airports

ATL(Atlanta):

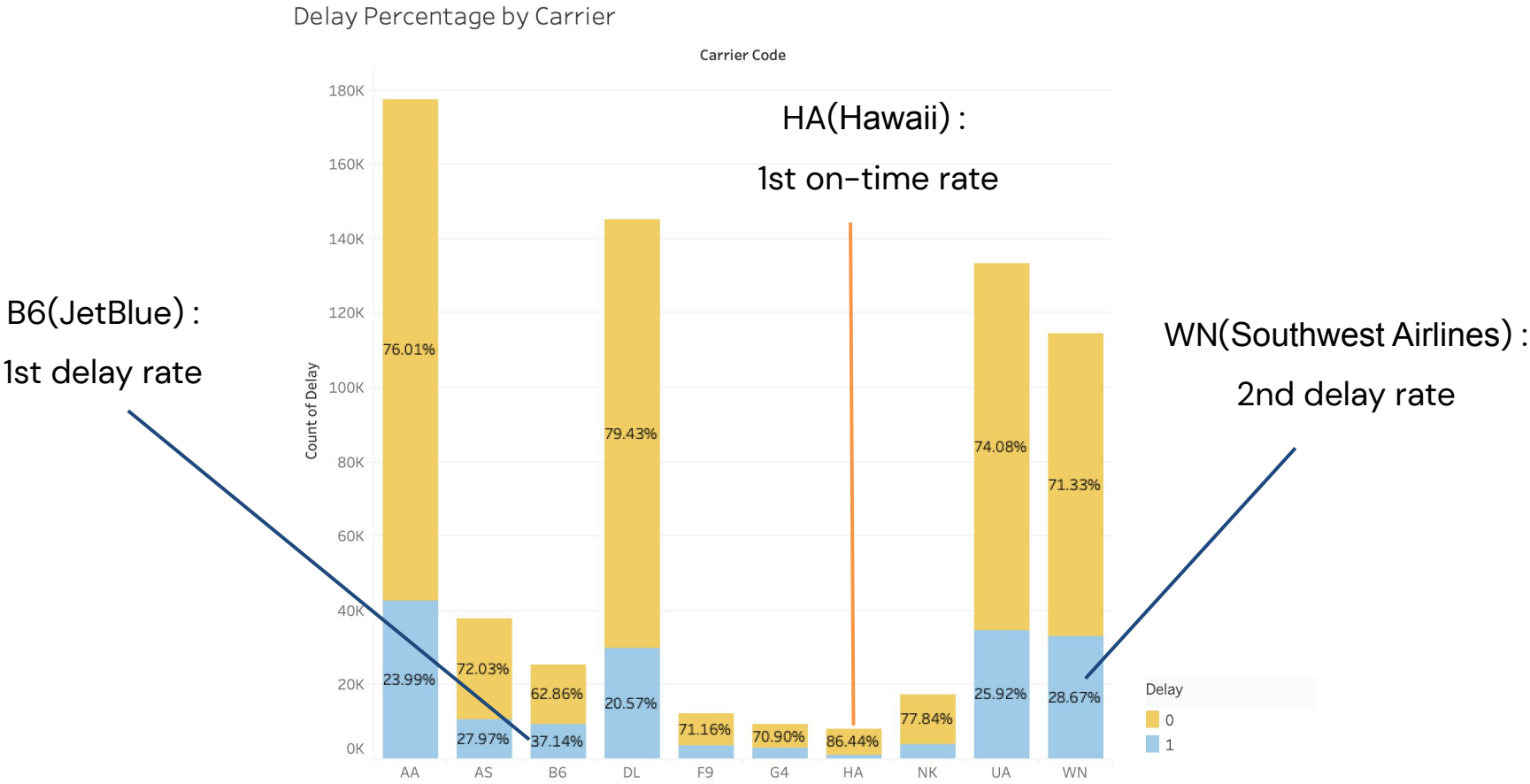
lowest delay rate

ORD(Chicago) :

highest delay rate

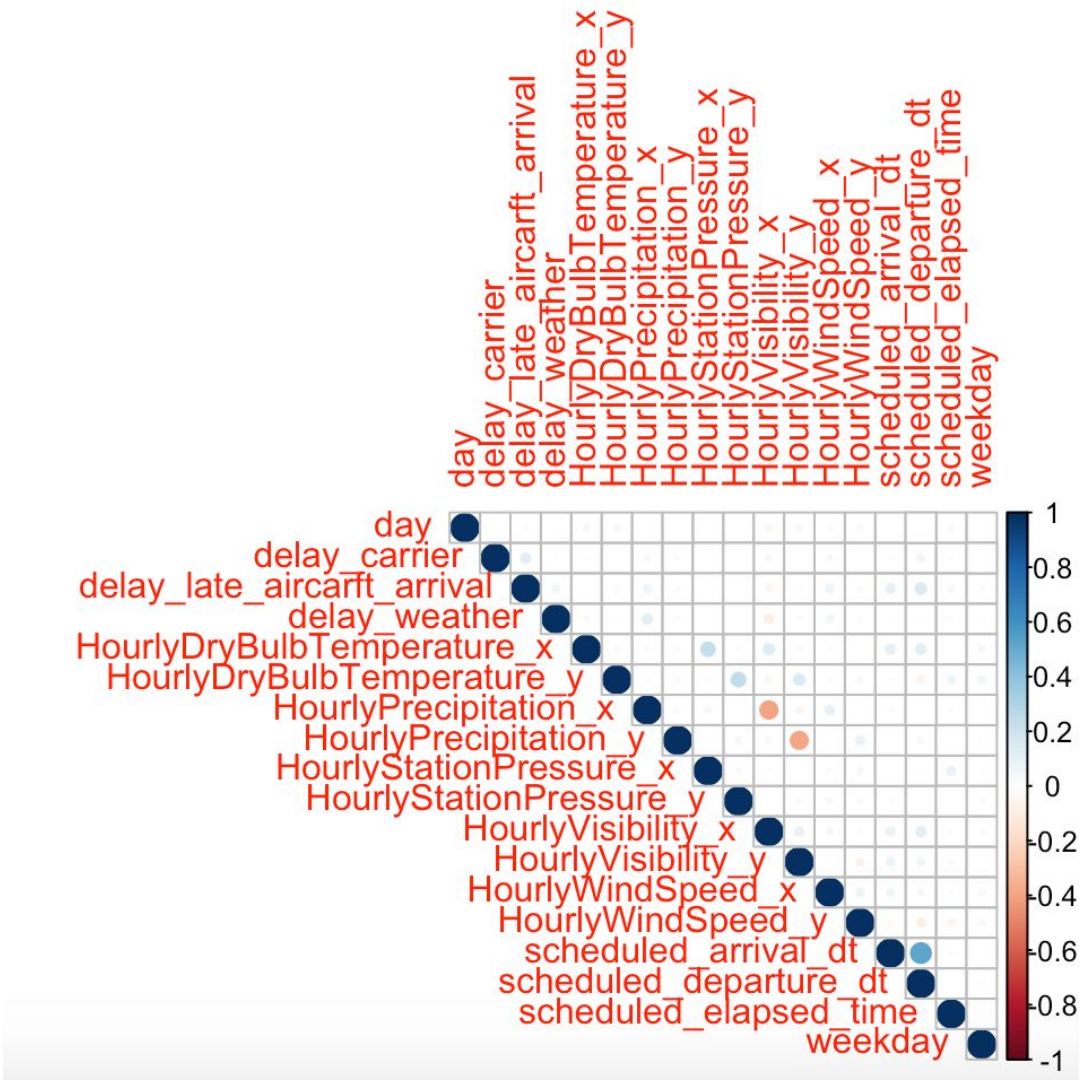


Delay Portion by Carriers



Correlation

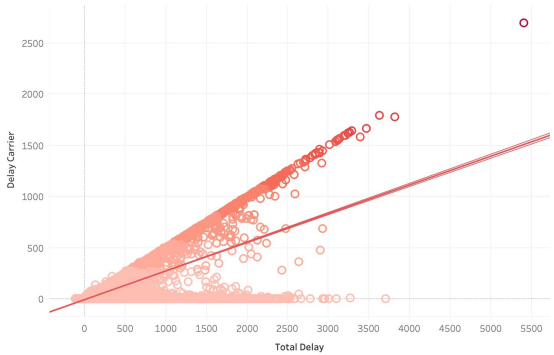
- No strong correlation between delay reason and other variables
- Delay reasons are weakly auto-correlated with each other
- Weather variables has correlation with each other



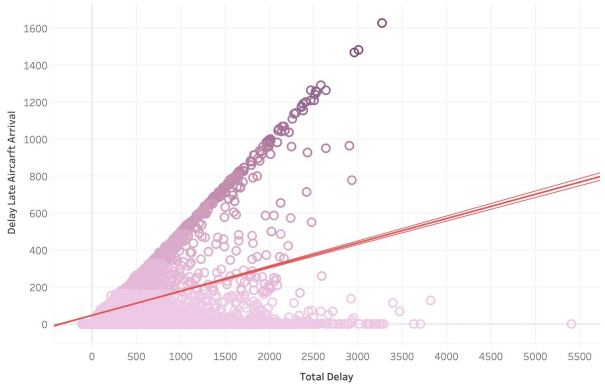
Correlation – Delay Reason

- Total delay = departure delay + arrival delay
- Most of the delay are caused by carriers and late aircraft arrival
    - Larger coefficients and higher R-squared value for carrier delay

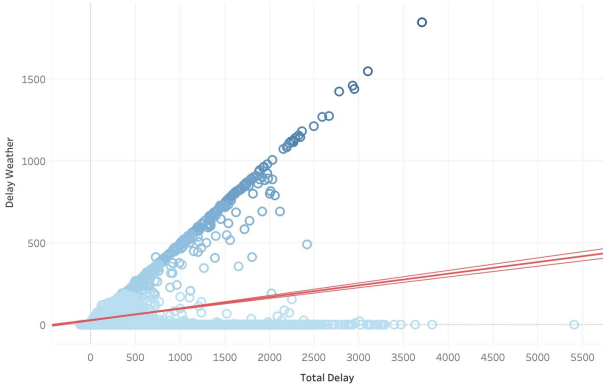
Carrier Delay



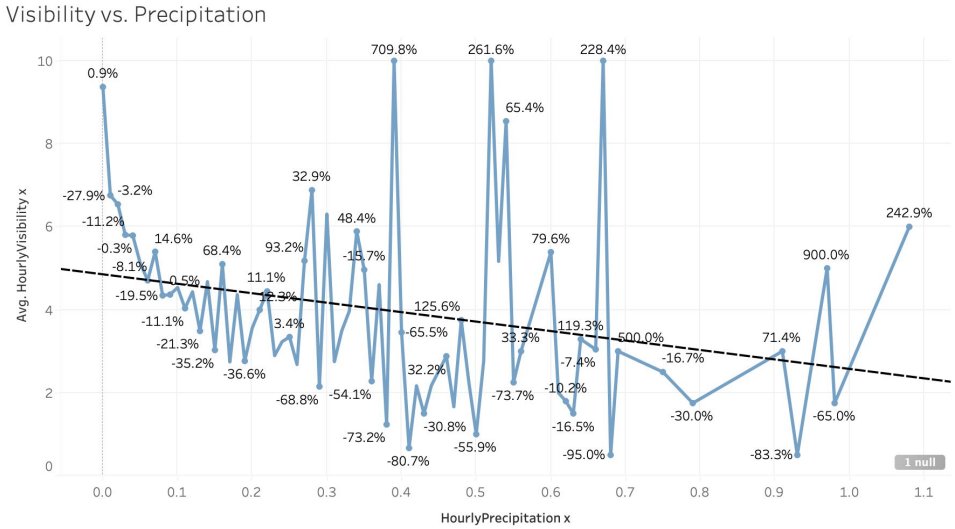
Late Aircraft Arrival Delay



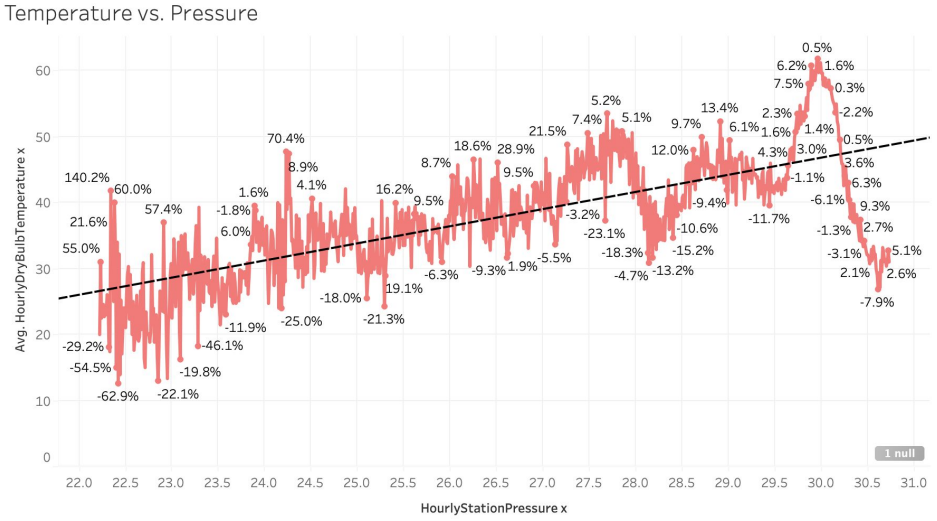
Weather Delay



Correlation – Weather Variable



- Visibility vs. Precipitation - Negative
- Temperature vs. Pressure - Positive



Carrier Delay

Logistic Regression

carrier_code B6	carrier_code F9	carrier_code WN	carrier_code AS	carrier_code NK	carrier_code DL	carrier_code UA
120%	73%	24%	23%	-39%	-24%	-21%
scheduled_ departure_dt	day6	day15	day25	WindSpeed_x	Temperature_y	Visibility_x
5%	-54%	-53%	-54%	2%	1%	-4%

61.3%

Testing Accuracy

Confusion Matrix	On Time	Delay
On Time	796	513
Delay	454	737

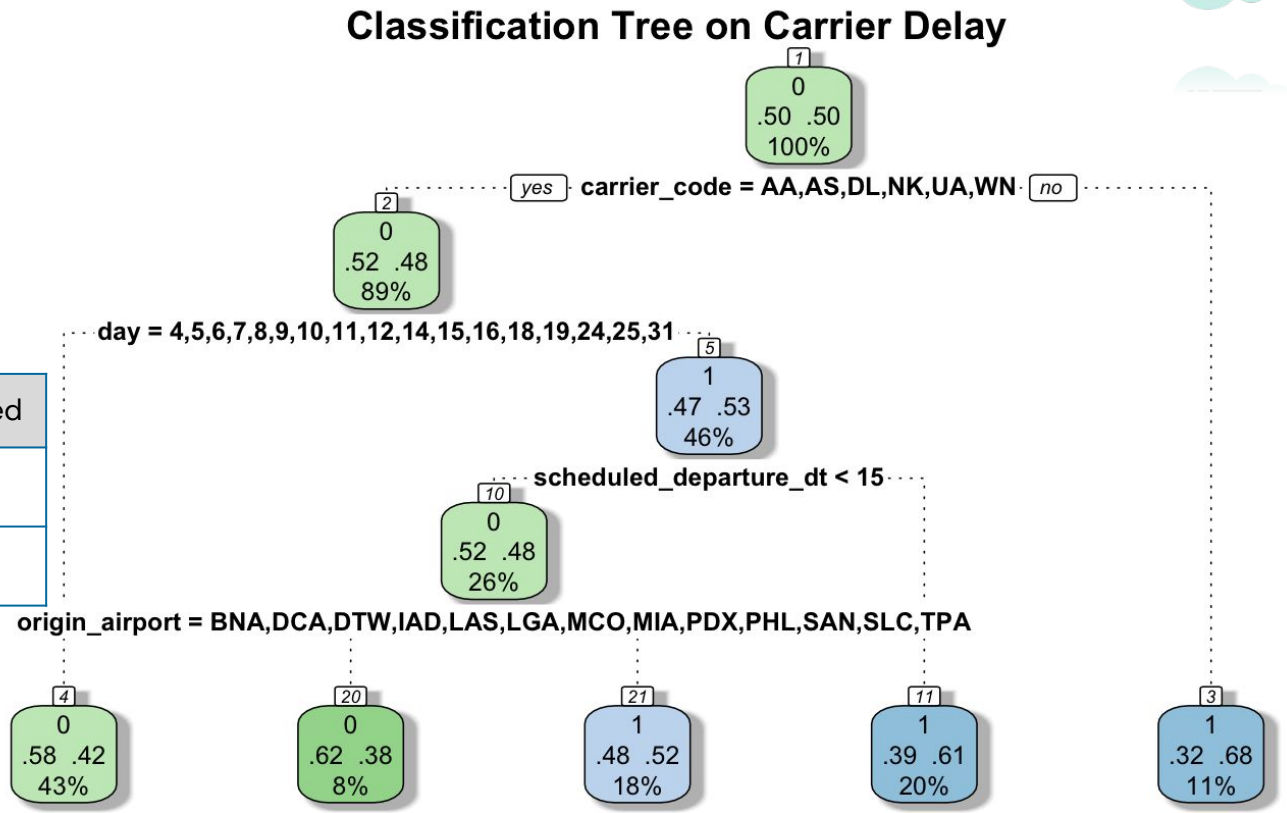
Decision Tree

Wiith Boosting

64.3%

Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	763	487
Delayed	406	844



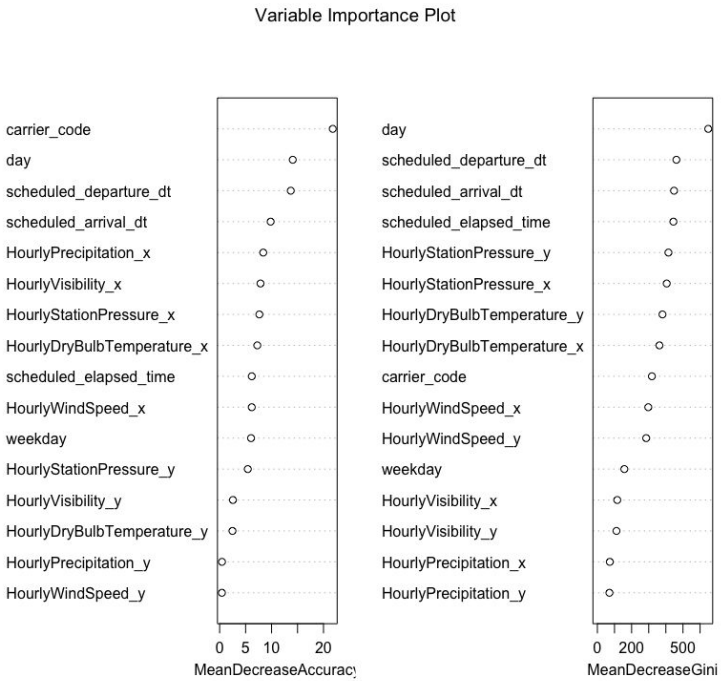
Random Forest

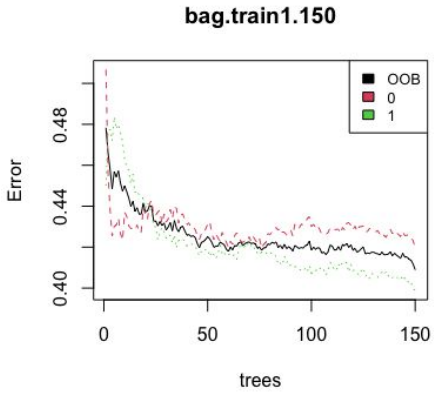
- ntree = 150

59.1%
Training Accuracy

66%
Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	708	542
Delayed	306	944





Late Aircraft Arrival Delay

Logistic Regression

origin_airport SFO	origin_airport EWR	origin_airport DEN	origin_airportS LC	day11	day15	day25
799%	706%	-96%	-92%	63%	-68%	-70%
carrier_codeF9	carrier_code NK	scheduled_ departure_dt	Hourly Precipitation_x	HourlyStatio nPresure_x	Hourly Visibility_x	Hourly WindSpeed_x
78%	-44%	12%	651%	-55%	-5%	3%

67.8%

Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	846	400
Delayed	404	850

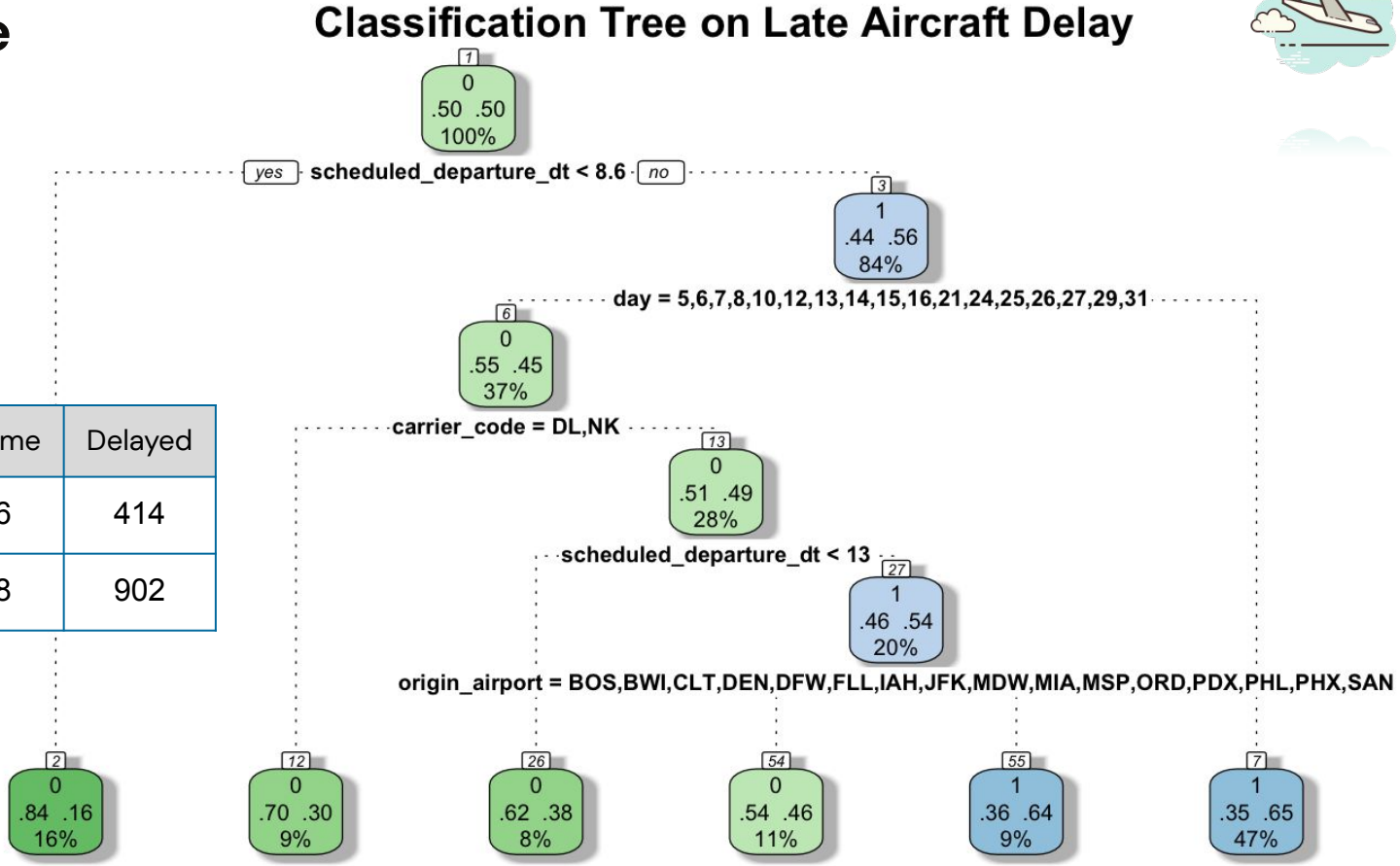
Decision Tree

Wiith Boosting

69.5%

Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	836	414
Delayed	348	902



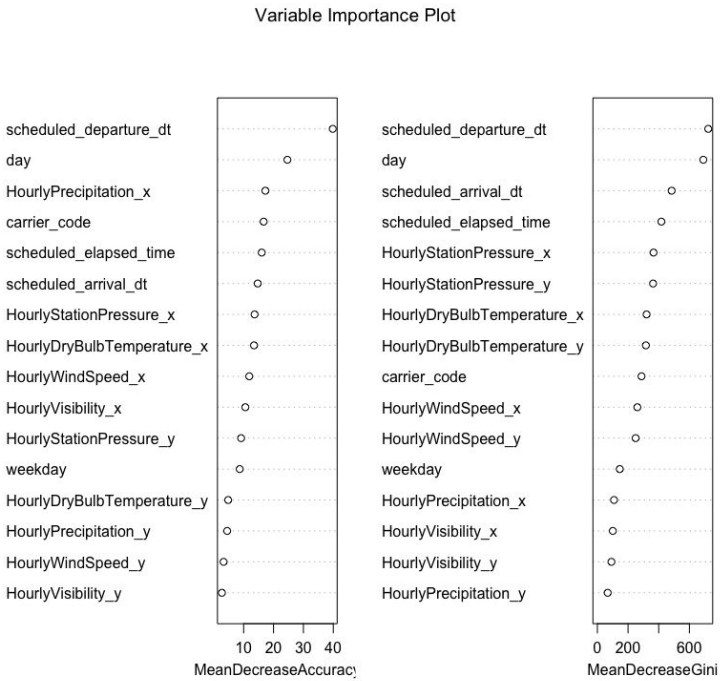
Random Forest

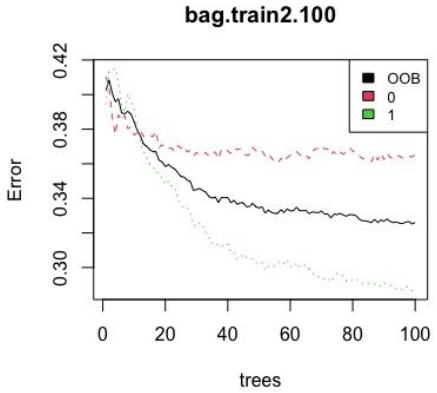
- ntree = 100

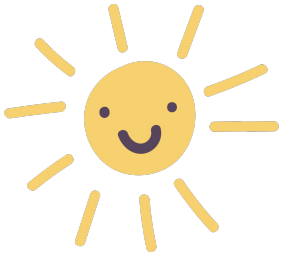
67.4%
Training Accuracy

71.7%
Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	814	436
Delayed	271	979







Weather Delay

Logistic Regression

origin_airport FLL	origin_airport IAH	origin_airport SAN	origin_airport TPA	origin_airport MCO	origin_airport MIA	origin_airport SFO
6993%	3584%	3048%	1930%	1604%	1556%	1478%
day4	day9	day11	day14	day24	day25	day27
-64%	-53%	137%	-66%	-88%	-86%	-84%
WindSpeed_x	Temperature_x	Visibility_x	Station Pressure_x	Visibility_y	Hourly Precipitation_x	scheduled_ arrival_dt
3.56%	-2.28%	4.56%	8.77%	6.64%	1689507%	3%

82.1%

Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	190	43
Delayed	39	186

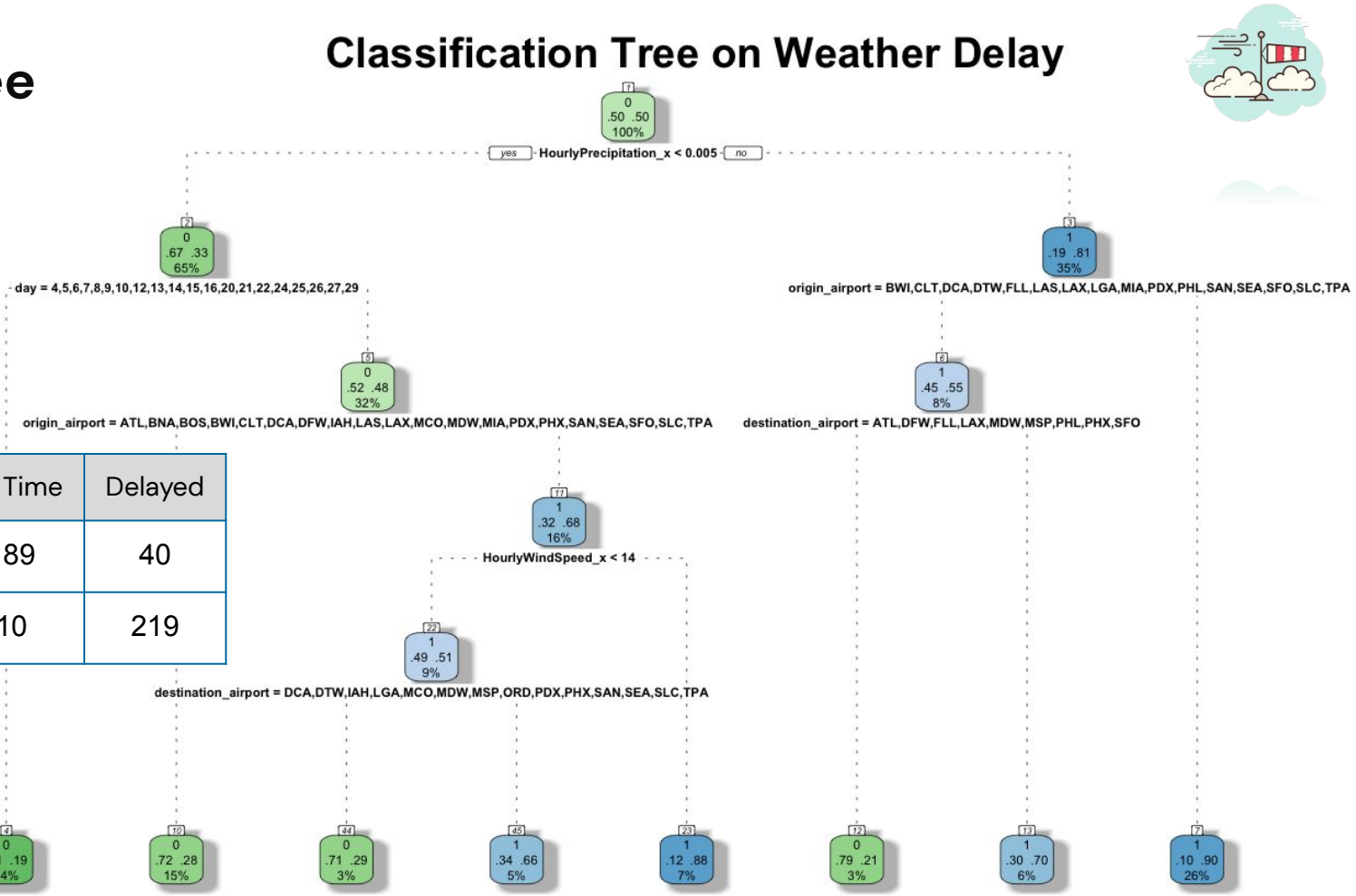
Decision Tree

Wiith Boosting

89.2%

Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	189	40
Delayed	10	219



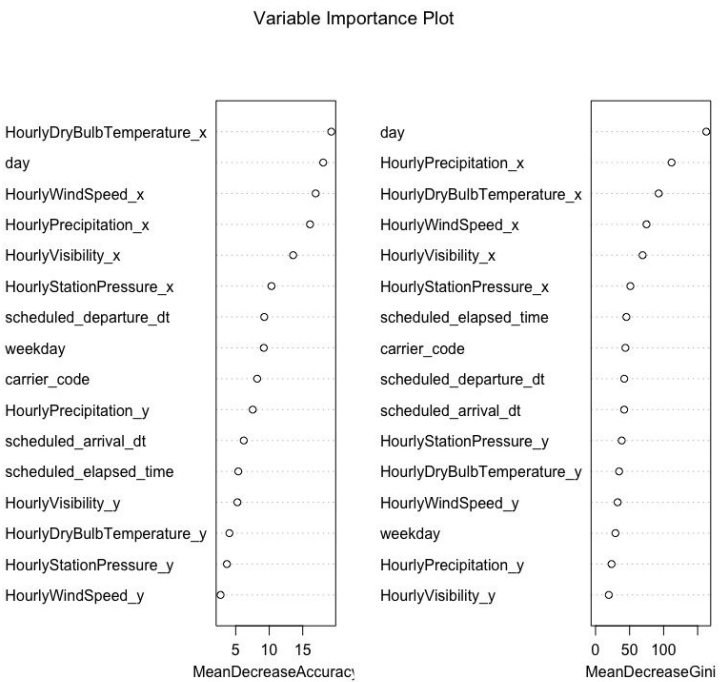
Random Forest

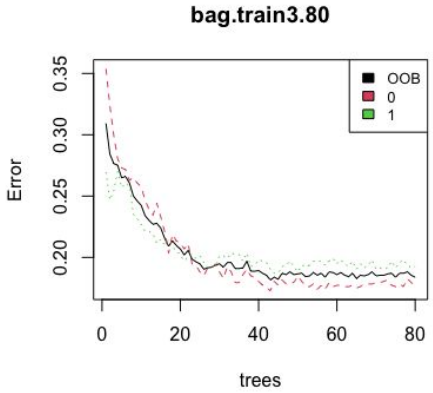
- ntree = 80

81.6%
Training Accuracy

90.6%
Testing Accuracy

Confusion Matrix	On Time	Delayed
On Time	198	31
Delayed	12	217



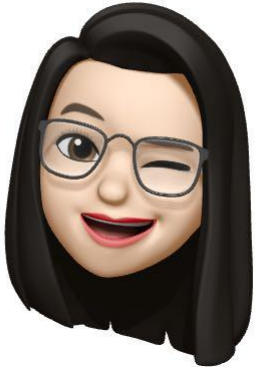


Model Results

	Carrier Delay	Late Aircraft Arrival Delay	Weather Delay
Logistic Regression	61.3%	67.8%	82.1%
Decision Tree with Boosting	64.3%	69.5%	89.2%
Random Forest	66%	71.7%	90.6%

Key Takeaways & Recommendation

- Avoid carriers: Jetblue, Southwest Airlines, Alaska Airlines
- Most popular destinations: Chicago, San Francisco, New York, Miami, Boston, Los Angeles have higher delay percentage
- Holiday like Christmas and New Year have lower chance to delay
- **Carrier delay:** carrier code, day, scheduled departure time
- **Arrival delay:** scheduled departure time, day
- **Weather delay**: wind speed, precipitation, visibility
- Need more data on carriers to improve model performance



Team 3

Thank You