Flight Delay Prediction

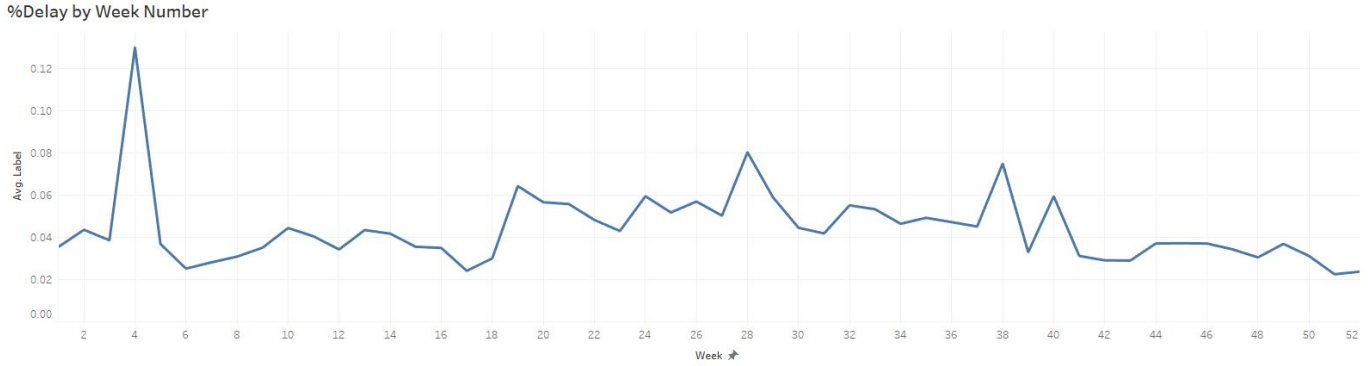
Yuan Liu

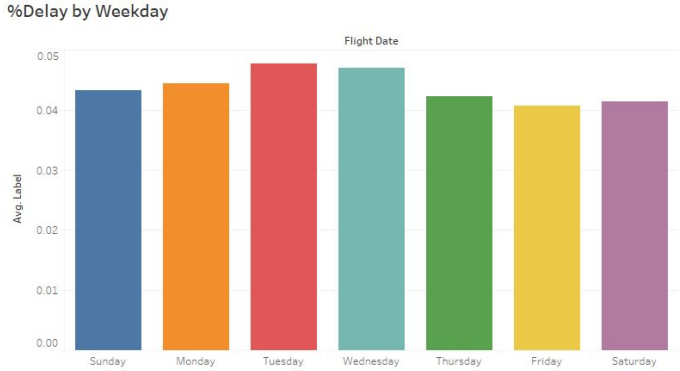
Project Overview

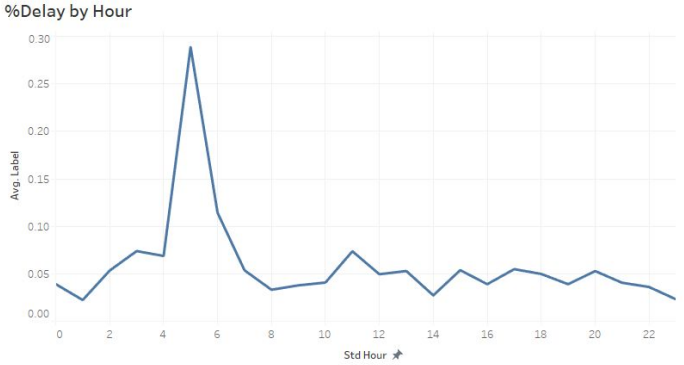
Goal: To predict the claimed amount (0 - $800) assigned to flights

- Explore the parameters that correlate to the flight delays via EDA
- Engineer new features based upon EDA results
- Build up binary classification models
  - Label = 1, if is_claim = 800 (i.e. delay time > 3 hr or cancelled)
  - Label = 0, if is_claim = 0 (i.e. delay time < 3 hr)
  - Due to imbalanced label distribution, SMOTE re-sampling will be used
- Calculate the claimed amount
  - Predicted claimed amount = $800 * probability(label = 1)
- Evaluate the model
  - Check model accuracy (AUC-ROC), mean absolute error, and Brier error on testing dataset

%Delay Varies by Time

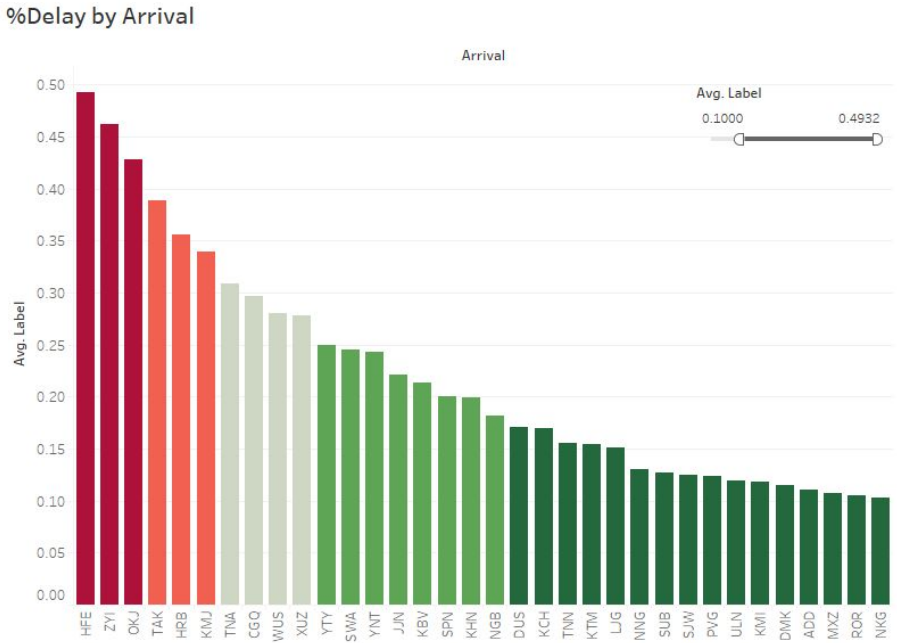


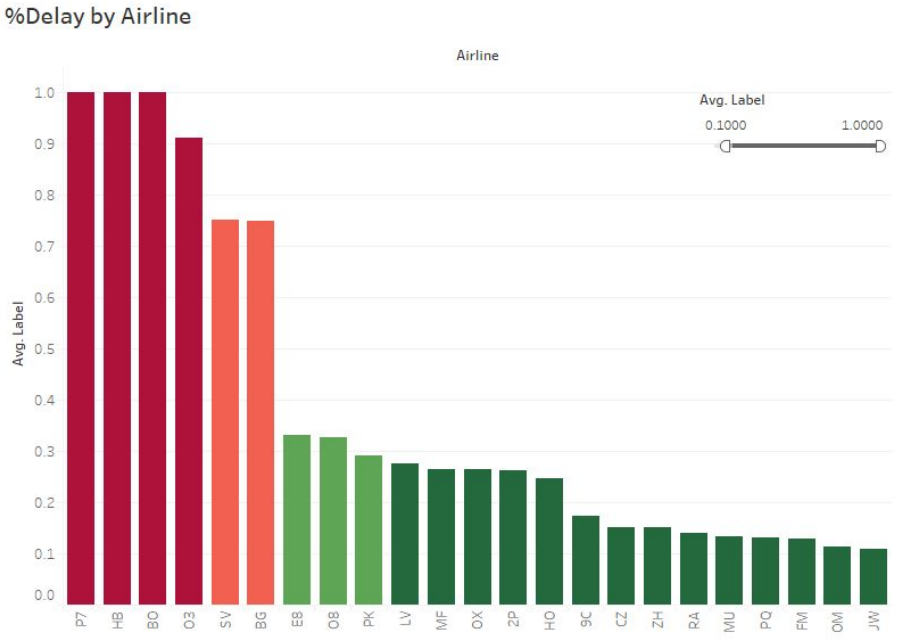




Takeaways: Week 4 of the year, Tuesday and Wednesday, and 4am-6am of the day have higher %delay than other time

%Delay Varies by Arrival and Airline





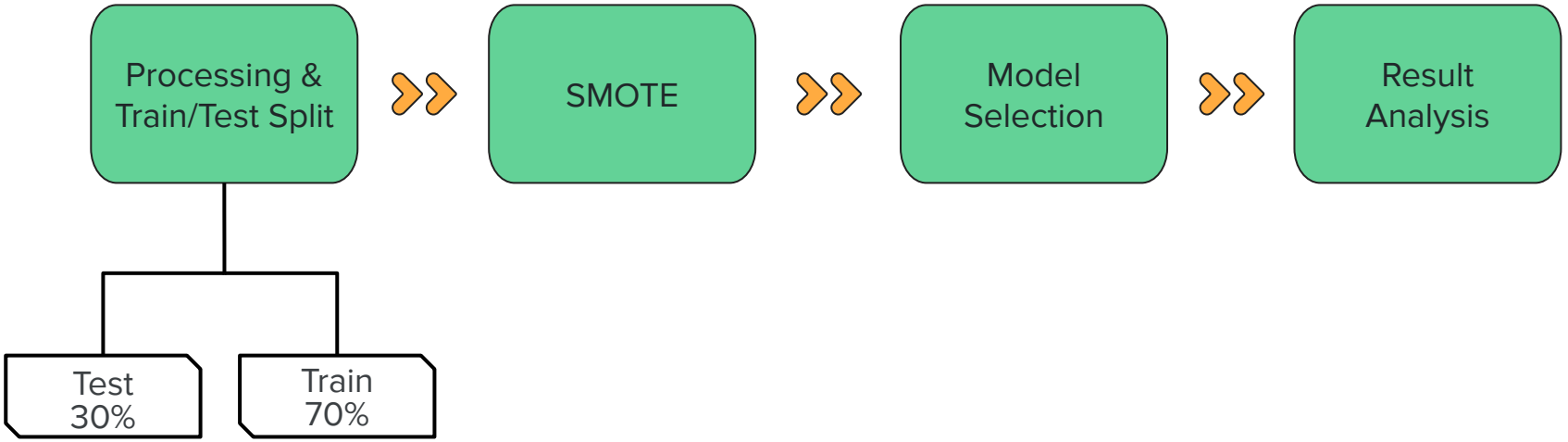
Takeaways: Certain arrivals and airlines have significantly higher %delay than others

** For clarity, only the arrivals or airlines with %delay >= 0.1 were shown in the figures*

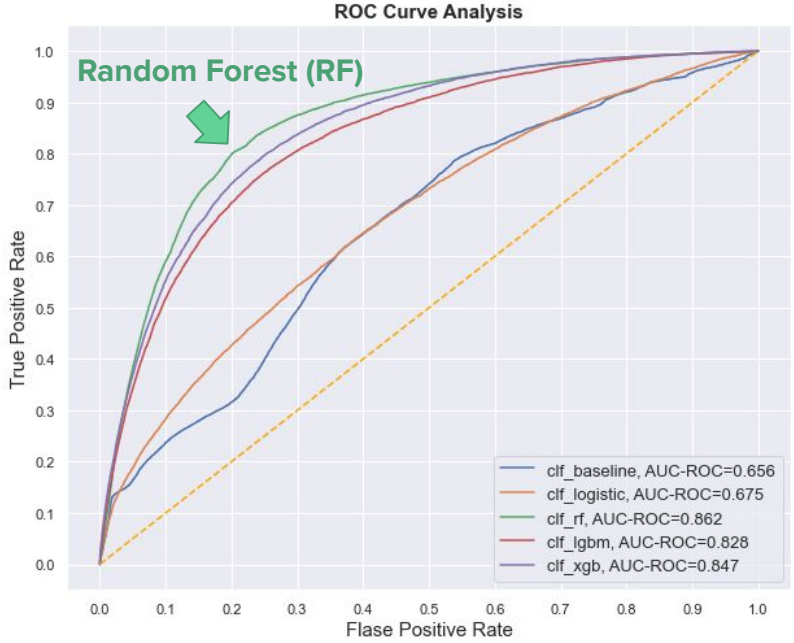
Summary of Modeling Features

- **Time Domain**
  - Week of the year
  - Day of the week
  - Hour of the Day
- **Location Domain**
  - Arrival airports (high, medium, low delay groups)
  - Geographical distance between departure and arrival
- **Airline Domain**
  - Airlines (highest, high, medium, low, lowest delay groups)

Modeling Process Overview



Modeling Results



Random Forest as the best-performing model

- AUC-ROC = 0.86
- Mean absolute error= 35.21
- Brier error = 28011.43

** The results were calculated using the testing data set*

Feature Importance of RF Model

Weight	Feature
0.3177 ± 0.0819	distance
0.0234 ± 0.0055	weekday_Tuesday
0.0223 ± 0.0053	weekday_Wednesday
0.0221 ± 0.0072	weekday_Sunday
0.0216 ± 0.0072	weekday_Monday
0.0189 ± 0.0177	arrival_low_delay
0.0182 ± 0.0104	weekday_Saturday
0.0168 ± 0.0109	weekday_Thursday
0.0148 ± 0.0080	week_4
0.0140 ± 0.0221	airline_low_delay
0.0120 ± 0.0217	airline_medium_delay
0.0114 ± 0.0029	hour_20
0.0112 ± 0.0030	hour_18
0.0105 ± 0.0059	hour_15
0.0103 ± 0.0033	hour_16
0.0102 ± 0.0038	hour_12
0.0101 ± 0.0026	hour_21
0.0099 ± 0.0046	hour_19
0.0096 ± 0.0028	hour_13
0.0096 ± 0.0024	week_40
… 67 more …	

Strong delay indicators:

- **Long distance** between arrival and departure
- Flights on **Tuesday, Wednesday, and Sunday**

Conclusion

- **What has been done**
  - Identified the underlying factors related to flight delay via EDA
  - Constructed the predictive modeling process to forecast the claimed amount due to flight delays, achieving the delay classification accuracy of 0.86, mean absolute error of 35.21, and Brier error of 28011.43
  - Analyzed the modeling results to identify delay indicators
- **What has been found**
  - Among the factors from time domain, location domain, and airline domain, **long distance** between the departure and arrival airports is the strongest indicator for delay more than 3h
  - In addition, flights during **Tuesday, Wednesday,** and **Sunday** are also more likely to be delayed for more than 3 hours
- **Where to improve**
  - Explore different grouping methods of airlines and arrival airports
  - Explore flight number as the additional feature
  - Further tune the model hyperparameters via detailed grind search