Prediction of US Domestic Flight Cancellation

By Albert Cao, Avery Robinson, Lue Li, and Mabel Sekarputri
The Confusion Matrix

TABLE OF CONTENTS

01

Background & EDA

Background

Total number of variables: 44

- Origin and destination airports
- Airlines and flights
- Traffic of flights and airports
- Day and time of flights
- Demographics of passengers

Number of observations in training data: 69,225

Number of observations in testing data: 29,668

EDA - Categorical

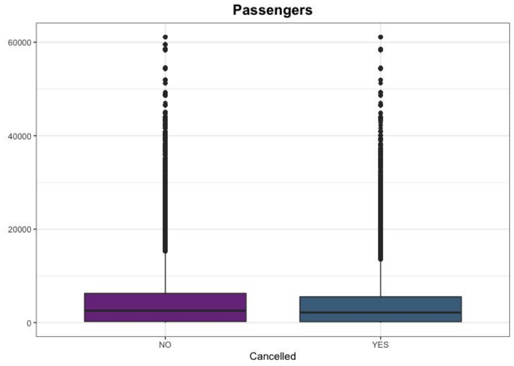
- Variables with exceptionally many levels were eliminated
- Variables with fewer levels were considered for the model

Cat. Var	Levels	Cat. Var	Levels
Destnation_airport	34	DAY	31
O.City	200	DAY_OF_WEEK	7
O.State	50	AIRLINE	14
Origin_airport	214	FLIGHT_NUMBER	4593
Origin_city	209	TAIL_NUMBER	4174
Destiantion_city	30	Rank	35
MONTH	3	Rank.Status	4

EDA - Numerical

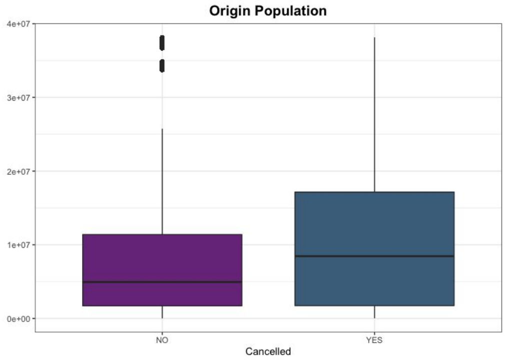
Pattern A

- Many outliers outside the box plot
- Passengers, Seats, Flights, Distance



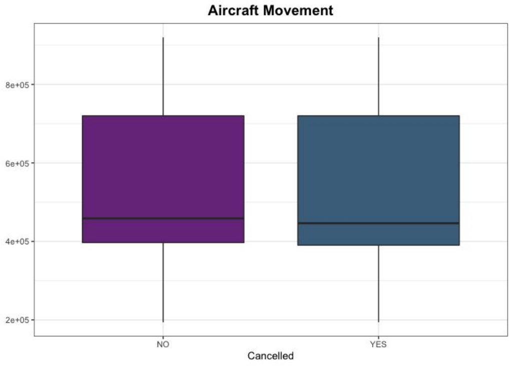
Pattern B

- Significant difference in median and range
- Population, Latitude, Longitude, Average Passengers, Passenger traffic



Pattern C

- Negligible difference
- Aircraft Movement



02

Data Cleaning

Data Cleaning

Variables with a small percentage of missing values were considered for imputation

- Pass.Traffic
- Aircraft.Movement

Concerns for imputing the others:

- inconsistent data
- More work than worth

Approach: Focus on adding new variables instead

Variable	%Missing	Variable	%Missing
Pass.Traffic	0.8	share_white	97.5
Aircraft.Movement	2.1	share_black	97.5
TAIL_NUMBER	9.4	share_native_a merican	97.5
AIR_SYSTEM_DELAY	85.0	share_asian	97.5
SECURITY_DELAY	85.0	hare_hispanic	97.5
AIRLINE_DELAY	85.0	Median.Income	97.5
LATE_AIRCRAFT_DElAY	85.0	poverty_rate	97.5
WEATHER_DELAY	85.0	percent_compl eted_hs	97.5

03

Feature Selection & Modelling

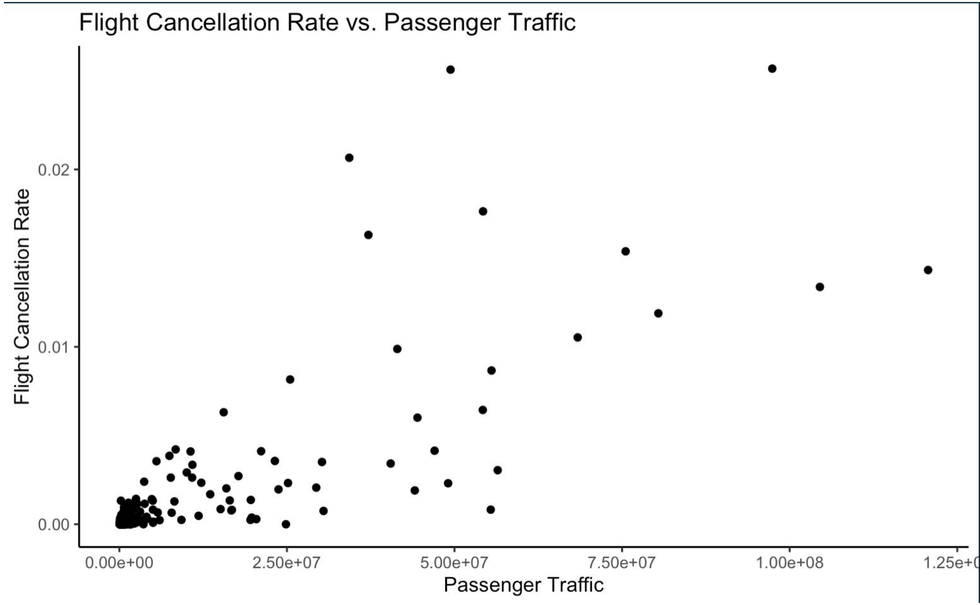
Feature Selection

Mainly used:

- Proportion tables and chi square tests for the categorical variables
- Boxplots and t-tests for the numerical variables
- Variance Importance Plots

Unsuccessful Approaches:

- glm()



Feature Selection

Airport Related	Time Related	Airline Related	Others
Destination_airport	SCHEDULED_DEPARTURE	FLIGHT_NUMBER	DISTANCE
O.City	SCHEDULED_TIME		
Org_airport_long	SCHEDULED_ARRIVAL		
	DAY_OF_WEEK		
	MONTH		

Based on the existing variables, we learned that time and airport can be more important

Feature Engineering

- Added 26 new variables to our data set (FAA 2019)

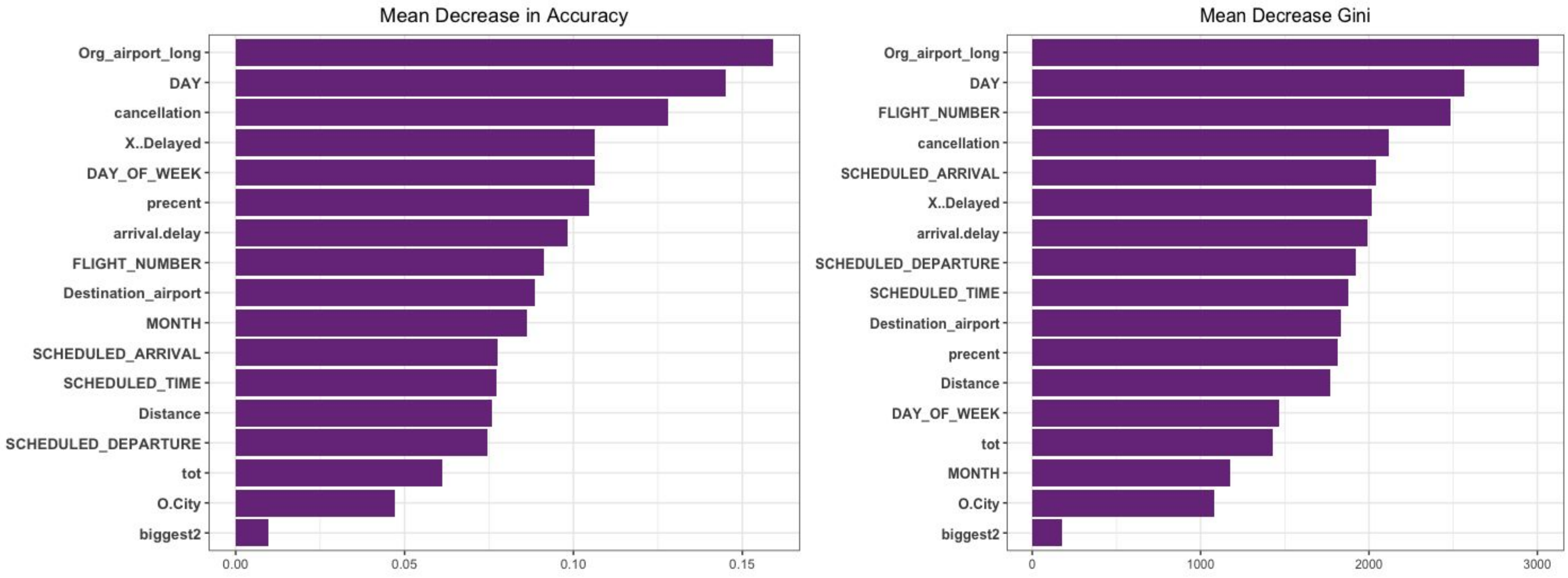
Name	Type	Description
Biggest2	Categorical	“1” if the origin airport is one of the top ten largest airports in the U.S.
tot	Numerical	The sum of # of arrival seats + # of departure seats on the given day by the origin airport
X..Delayed	Numerical	Percentage chance that a flight is delayed on a given day
Percent	Numerical	Percentage chance that a flight is cancelled on a given day
Cancellation	Numerical	Number of cancellation on a given day
arrival.Delay	Numerical	Percentage chance that the flight is delayed upon arrival at the destination airport

Best-Performing Model

Out of all the techniques used, random forest with 500 trees and 5 variables randomly selected at each split gave us the highest accuracy.

Technique	Accuracy %
Logistic	75.66448
KNN	95.50751
LDA	80.52292
QDA	79.0735
RandomForest	99.841
GBM boosting	99.29218
Voting	~78

Variance Importance Plot



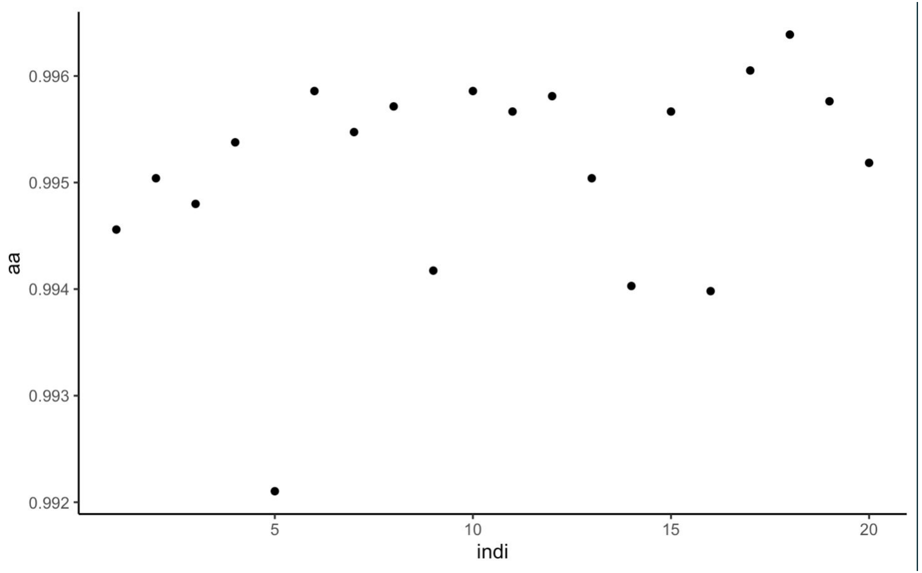
Most Efficient Model

Parameters tuned: mtry and ntree

Used loops to find the optimal values

Processing Time: 152.34s vs 1.59s

Accuracy: 0.9975 vs. 0.9985



flight.efficient <- randomForest(data=tr2, as.factor(Cancelled) ~ Destination_airport + SCHEDULED_DEPARTURE + DAY + O.City + Distance + DAY_OF_WEEK + MONTH + Org_airport_long + FLIGHT_NUMBER + SCHEDULED_TIME + SCHEDULED_ARRIVAL , mtry = 4, ntree = 10, importance = TRUE)

04

Results

Results - Best Result

Public Leaderboard: 99.850% (6th place)

Private Leaderboard: 99.831% (7th place)

- The low accuracy rates from LDA and QDA and high accuracy rates from KNN (three neighbors), random forests, and boosting suggest a more complex decision boundary

Future Research Directions

- Impute more of the missing data, such as AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY, LATE_AIRCRAFT_DELAY, WEATHER_DELAY
- Incorporate data on passengers' perceptions of airport reliability on flight bookings, service rating, and other operational factors
- Collect traffic volume data on all airport so that we could join the dataframe by origin airport instead of destination airport
- Acquire information on aircrafts by using the tail number provided in the dataset
- Implement different machine learning techniques

Resources

Federal Aviation Administration

FAA Operations & Performance Data

FAA Operations and Performance Data provides access to historical traffic counts, forecasts of aviation activity, and delay statistics.

Database Access Systems

- Aviation System Performance Metrics (ASPM)
- Operational Network (OPSNET)
- Traffic Flow Management System Counts (TFMSC)
- Airline Service Quality Performance (ASQP)
- Terminal Area Forecast (TAF)

- System Descriptions

Reporting Systems

- Business Jet Reports

aspm.faa.gov

Thank you!
THE CONFUSION MATRIX
12/17/2020