Predicting the Likelihood of Flight Cancellations

Aashish Jain, PhD

Data Science Intensive Capstone Project, May 29th 2017 Cohort

Springboard

Thanks to Springboard mentors

Hassan Kingravi, Pindrop Labs

Srdjan Santic, RCC, University of Belgrade

The Problem

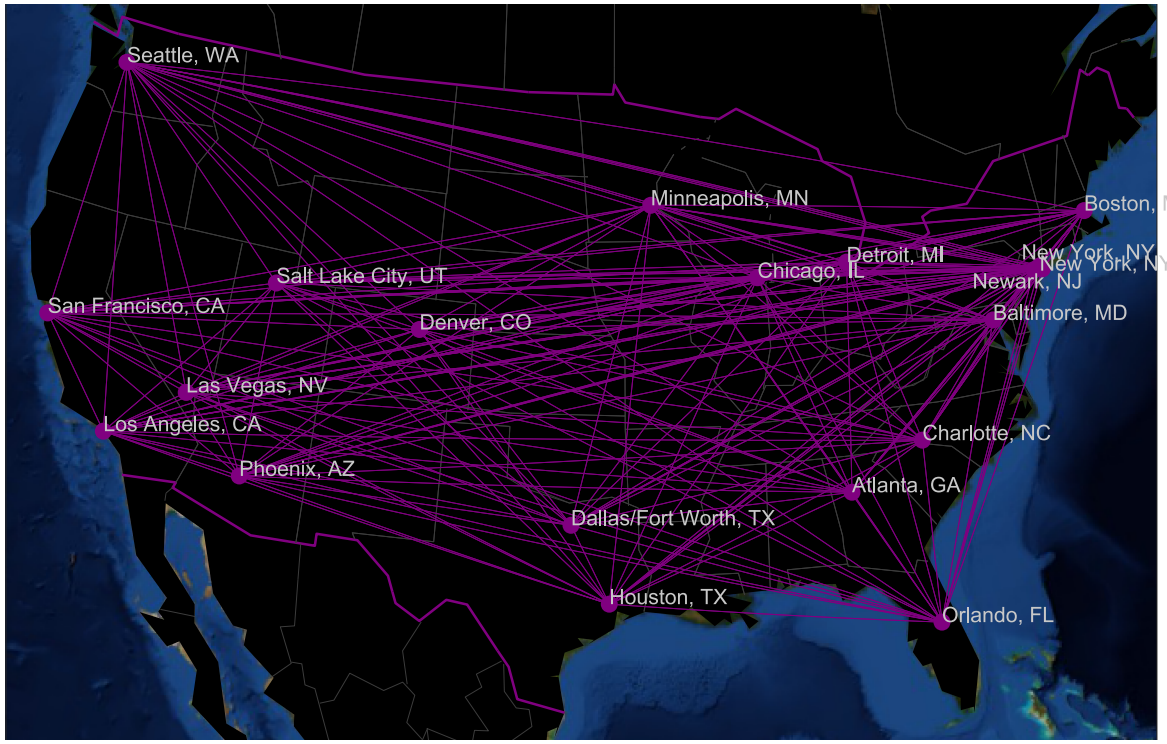
Total flights in two years (2015 and 2016) in top 20 airports:

2.85+ millions

Cancellation rate:

1.14% (≅32,600 flights)

Even though 1.14% is a small number, one rare event causes lot of troubles to passengers



What factors affect flight cancellation rate?

Can we predict the likelihood of a flight getting cancelled?

Who might care?

Travel Agencies

KAYAK

Expedia

Booking.com

priceline.com

travelocity

CheapAir.com

Airlines

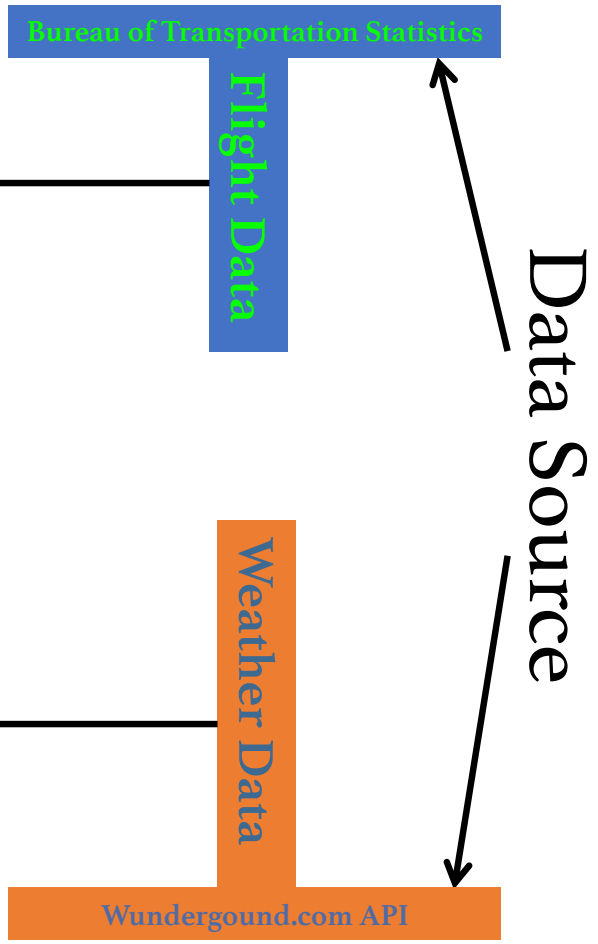
Mobile Apps

...and many more...

What factors might affect flight cancellations?

- Flight schedules: departure and arrival dates and times
- Airline carriers
- Flight length: distance, flight times
- Airport locations: both origin and destination

- Weather conditions: both origin and destination
- Temperature, humidity, wind speed, pressure, etc..: both origin and destination



Data Information

Data acquired for the period: January 2015 – December 2016

Flight and weather data for: Top 20 airports in the US

Number of records: 2,857,139

Number of fields: 90

Data Acquisition and Merging

https://github.com/aajains/springboard-datascience-intensive/tree/master/capstone_project/DataAcquisitionMerging

Flight Data Specifics

- Source: Bureau of Transportation Statistics
- Data can be downloaded for one month at a time
- File format : csv
- Concatenate files from all months into one
- Each record: a unique flight

Weather Data Specifics

- Source: Wunderground.com API
- File format : XML (hourly data available)
- One API call: One day data for a chosen airport
- Concatenate files from days into one, for an airport
- Each record: weather details for an airport at a given hour

Merge them together using following keys:

- Flight date at origin and destination
- Origin and destination airport names
- Scheduled departure and arrival times

Engineering Flight Historical Performance Features From Flight Data

https://github.com/aajains/springboard-datascience-intensive/blob/master/capstone_project/DataAcquisitionMerging/history_calc.ipynb

FlightDate	UniqueCarrier	AirlineID	Carrier	Origin	OriginCityName	Dest	DestCityName	CRSDepTime	DepTime
2015-01-01	AA	19805	AA	JFK	New York, NY	LAX	Los Angeles, CA	900	855.0
2015-01-02	AA	19805	AA	JFK	New York, NY	LAX	Los Angeles, CA	900	850.0
2015-01-03	AA	19805	AA	JFK	New York, NY	LAX	Los Angeles, CA	900	853.0
2015-01-04	AA	19805	AA	JFK	New York, NY	LAX	Los Angeles, CA	900	853.0
2015-01-05	AA	19805	AA	JFK	New York, NY	LAX	Los Angeles, CA	900	853.0

Steps for each flight:

1. Store the date of the flight in question: "thisDate"
2. Get Carrier, Origin, Destination, and Departure time
3. Filter the dataset based on 4 values found in step 2
4. Filter the filtered data further to get a subset of the data with only those dates that are "ndays" prior to "thisDate". Here, "ndays" is the number of days that we want to calculate the history for.
5. Use the ndays filtered dataset and perform aggregations such as count of cancelled flights, median of departure delays etc.. We used ndays = 10, 20 and 30.
6. These aggregated results are then added as new fields/features to original dataset.

Data Exploration

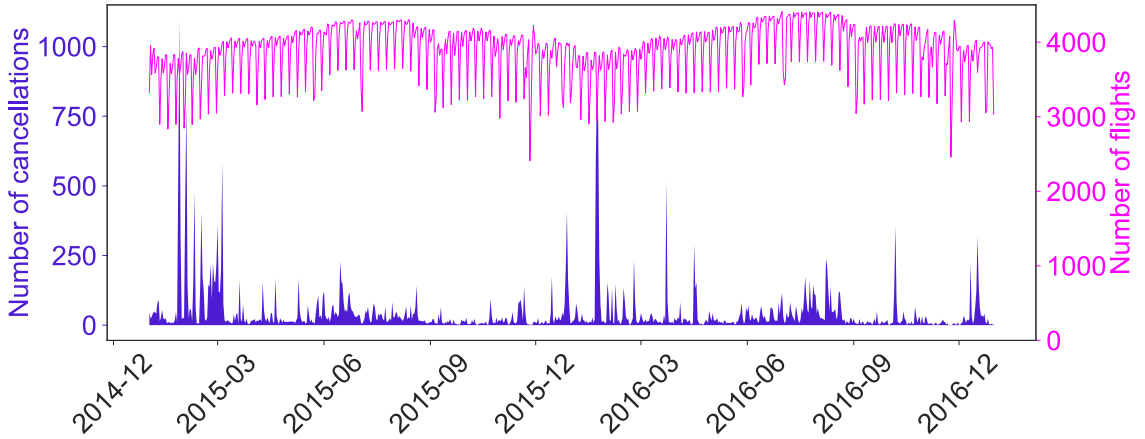
https://github.com/aajains/springboard-datascience-intensive/blob/master/capstone_project/EDA/ExploratoryDataAnalysis_v1.ipynb

Flight schedules
Airports
Airlines
Flight distance

Weather factors

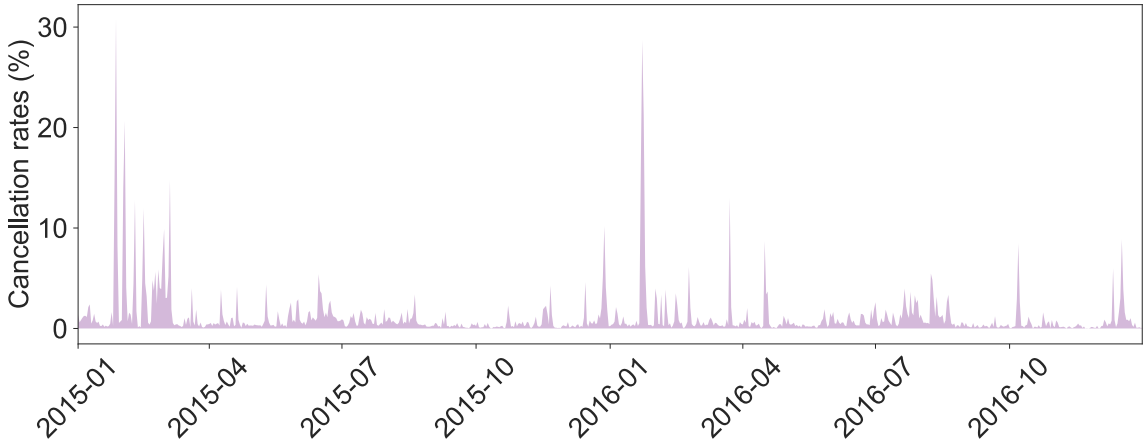
Flight historical performances

Flight Cancellation Rate: Definition



Number of flights and cancellations on a daily basis

$$\text{Flight cancellation rate} = \frac{\text{Number of flights cancelled for a given scenario}}{\text{Total number of flights for a given scenario}}$$

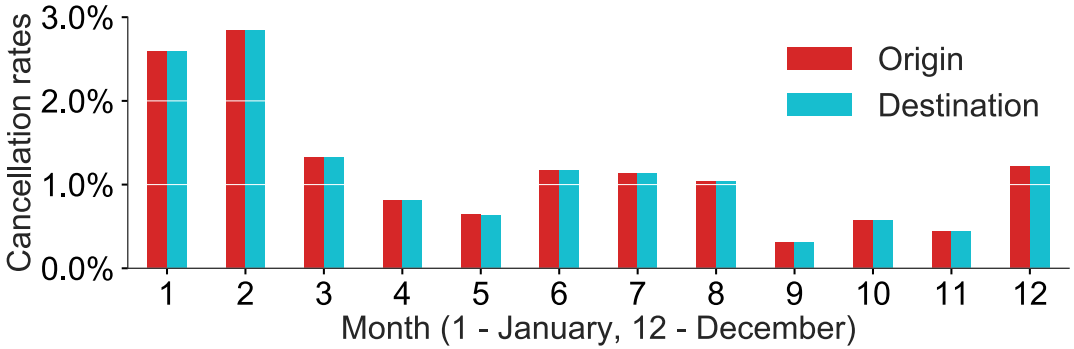


Cancellation rates on a daily basis

Here, a given scenario corresponds to a date. A scenario could be a weather condition, or a specific airline, or an airport, or a day of week, etc..

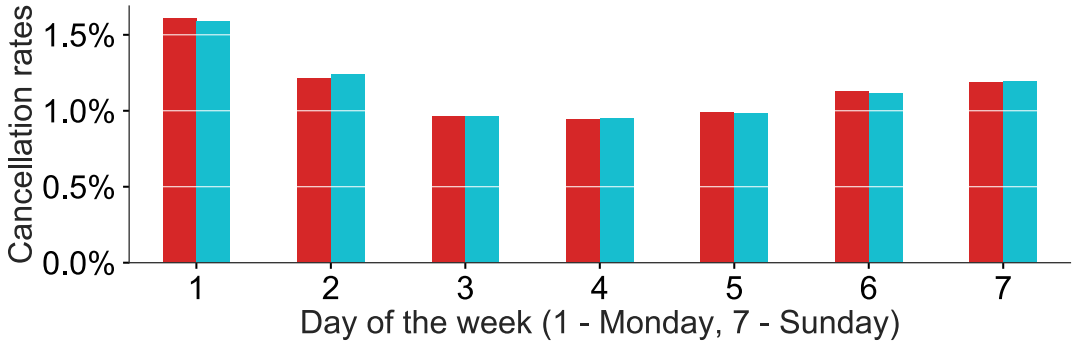
Flight schedules
Airports
Airlines
Flight distance

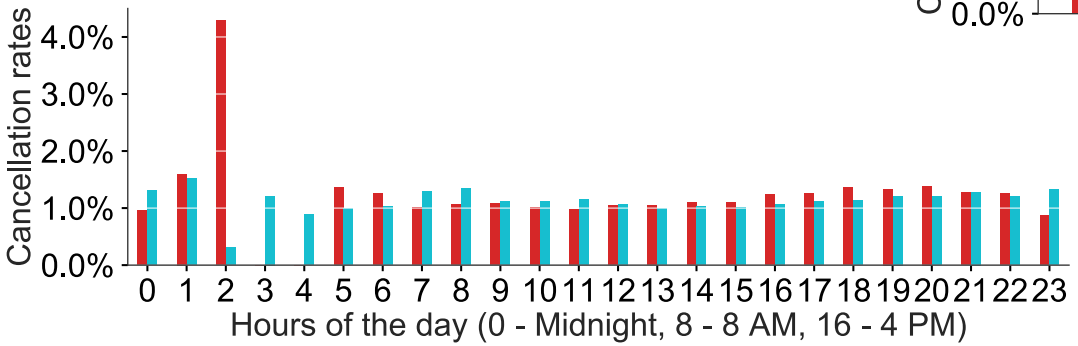
Flight Schedules



Monthly cancellation rates

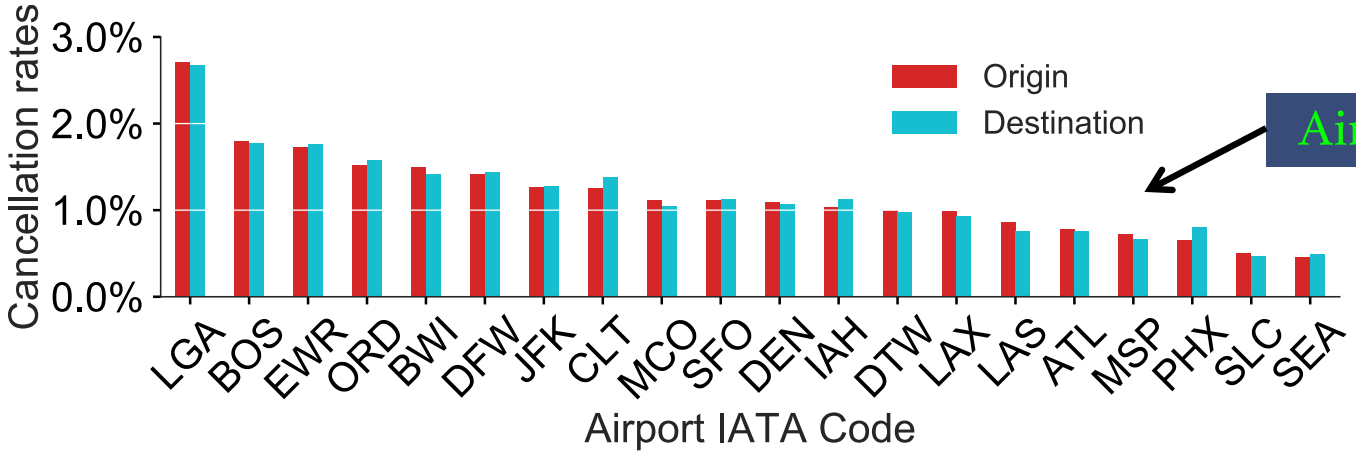
Weekly cancellation rates





Hourly cancellation rates

Airports and Airlines

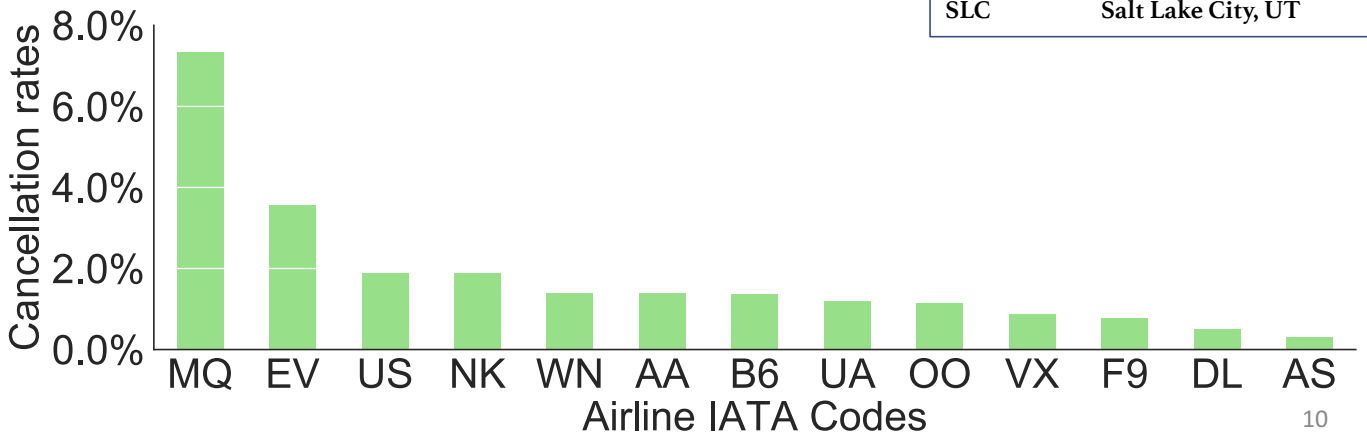


Airports

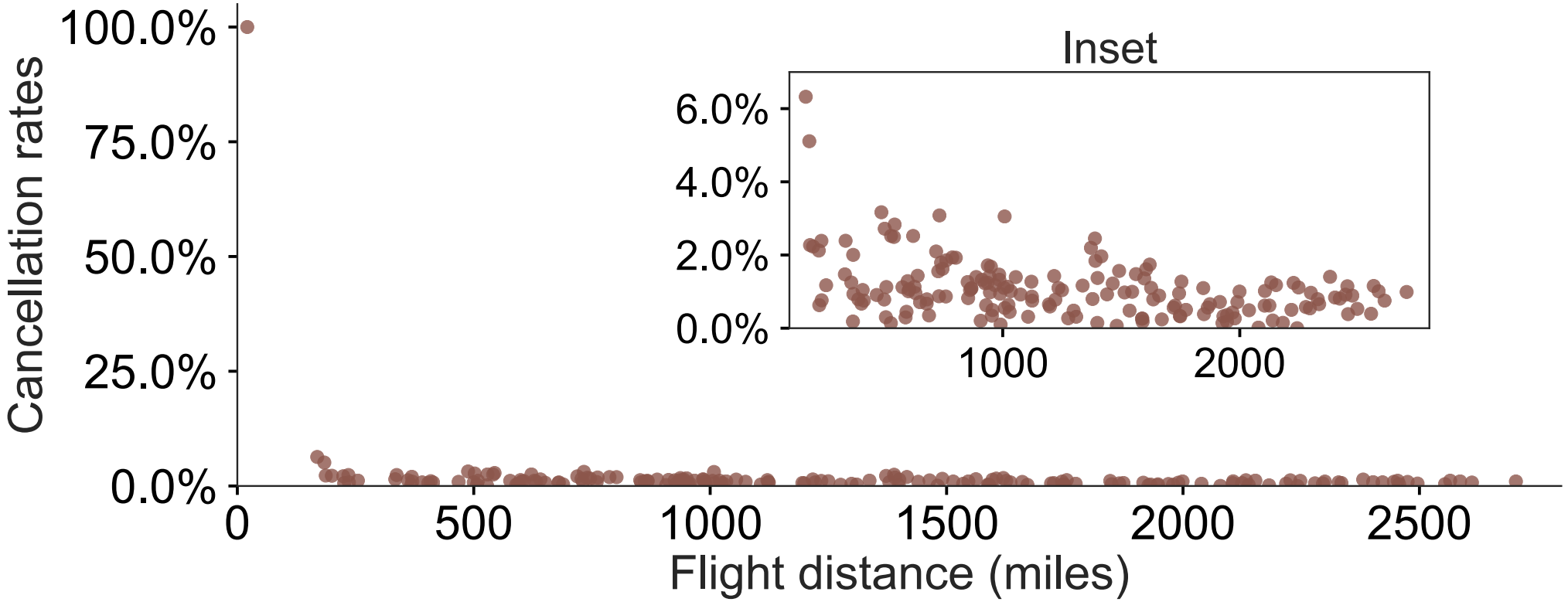
ATL	Atlanta, GA
BOS	Boston, MA
BWI	Baltimore, MD
CLT	Charlotte, NC
DEN	Denver, CO
DFW	Dallas/Fort Worth, TX
DTW	Detroit, MI
EWR	Newark, NJ
IAH	Houston, TX
JFK	New York, NY
LAS	Las Vegas, NV
LAX	Los Angeles, CA
LGA	New York, NY
MCO	Orlando, FL
MSP	Minneapolis, MN
ORD	Chicago, IL
PHX	Phoenix, AZ
SEA	Seattle, WA
SFO	San Francisco, CA
SLC	Salt Lake City, UT

Airlines

MQ - Envoy Air
EV - ExpressJet Airlines
US - US Airways
NK - Spirit Airlines
AA - American Airlines
B6 - Jetblue Airways
UA - United Airlines
OO - SkyWest Airlines
WN - Southwest Airlines
VX - Virgin America
DL - Delta Air Lines
F9 - Frontier Airlines
AS - Alaska Airlines

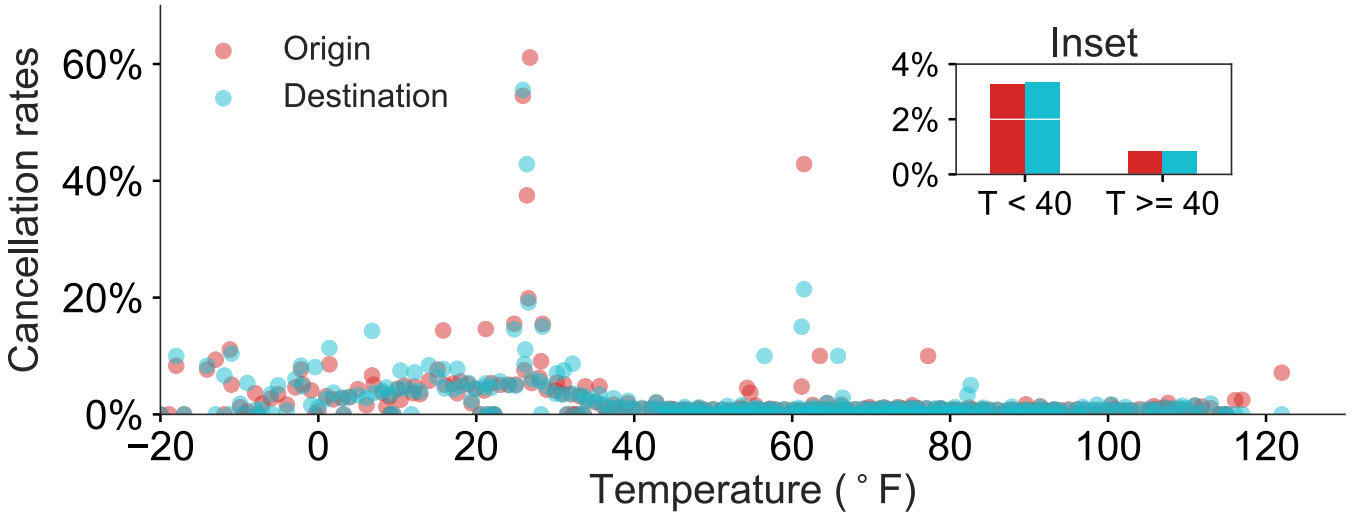


Flight Distance



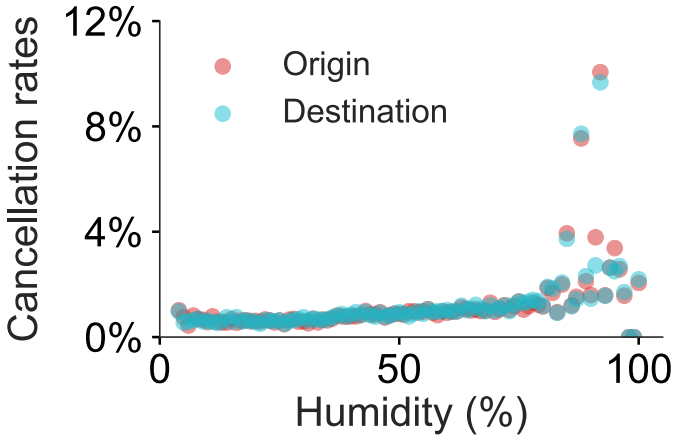
Weather factors

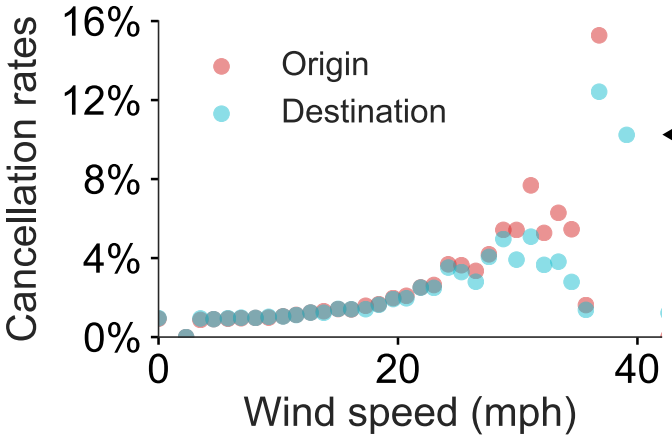
Temperature, Humidity and Wind Speed



Temperature

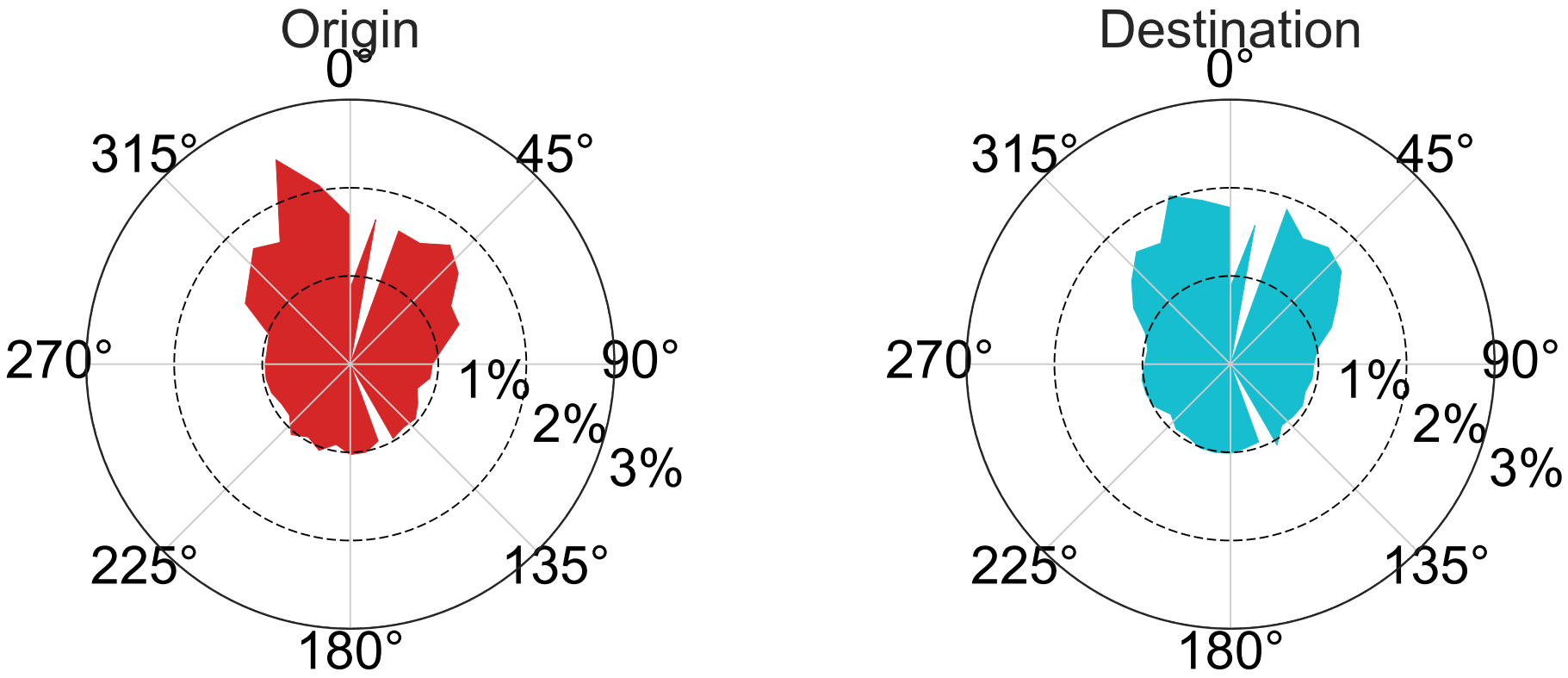
Humidity





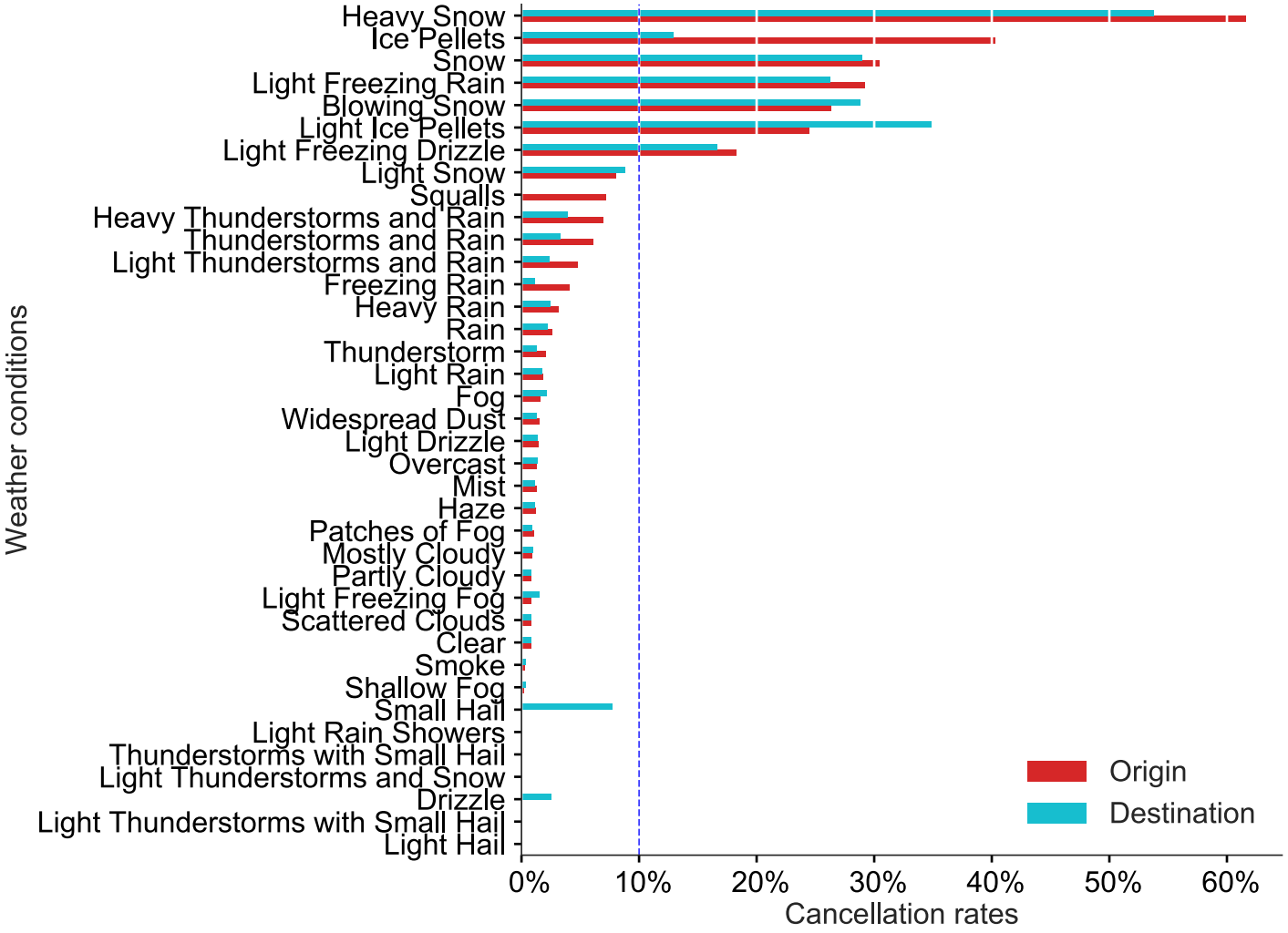
Wind speed

Wind Direction



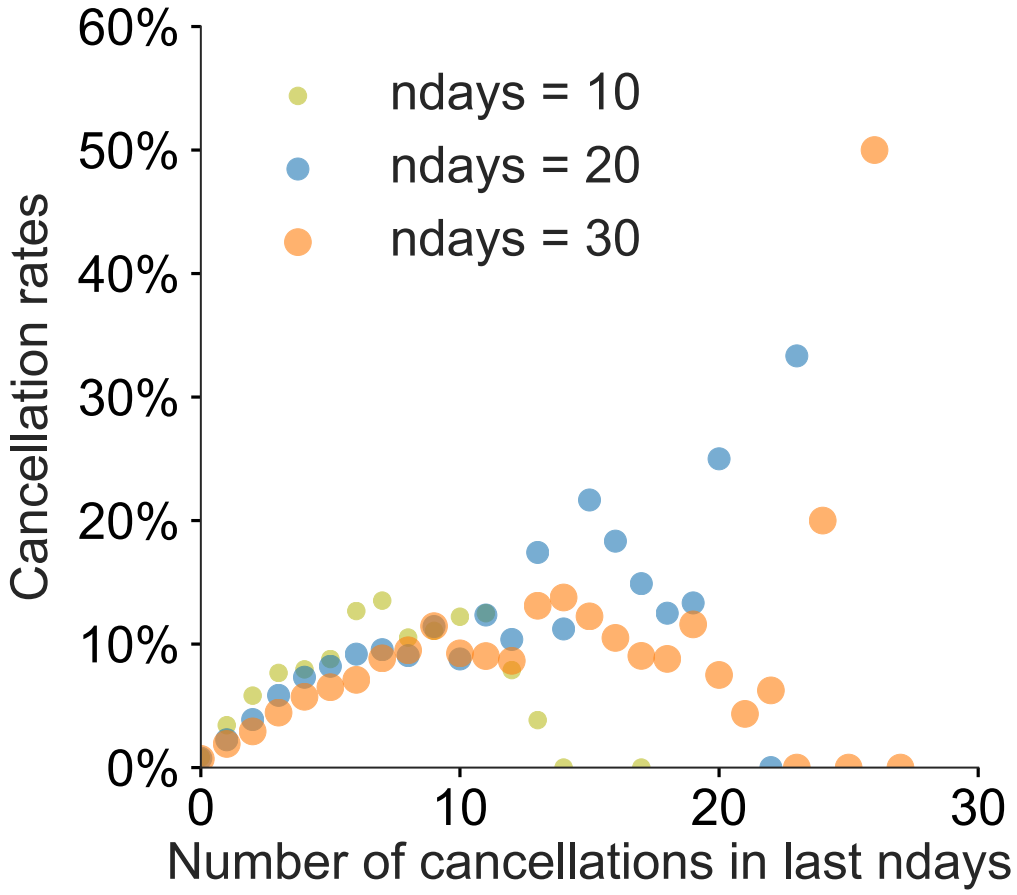
0° : North, 90°: East, 180°: South, 270°: West

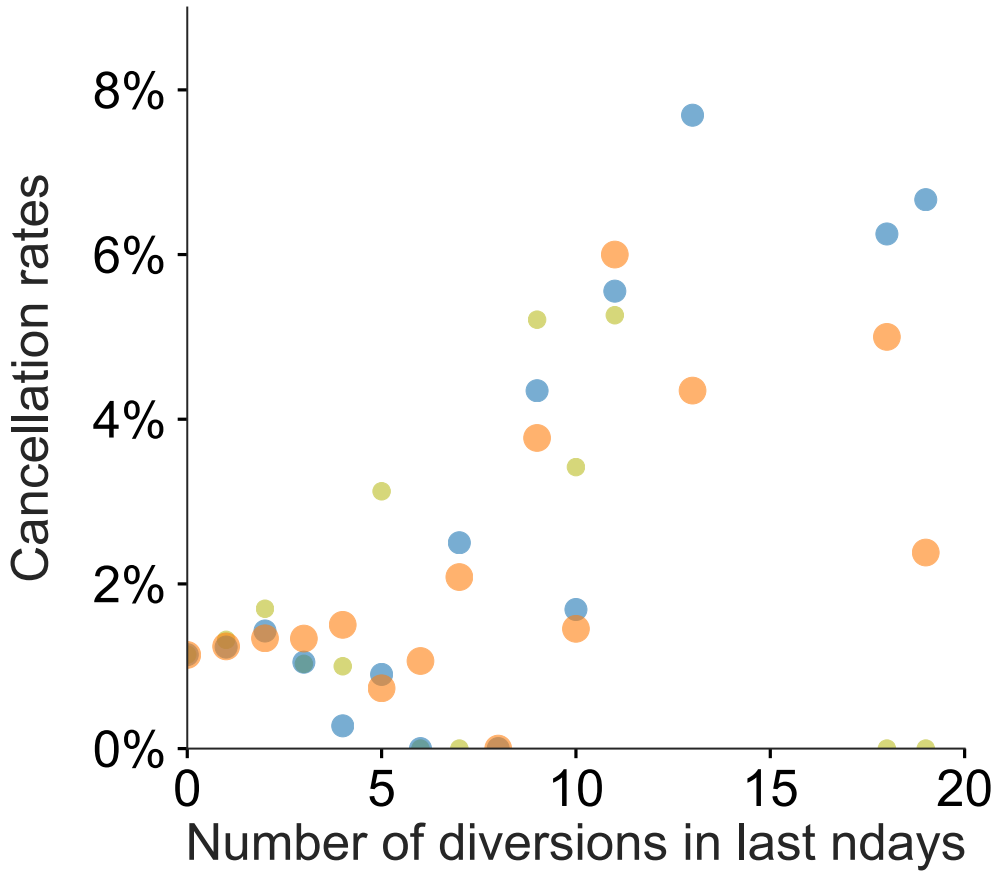
Weather Conditions



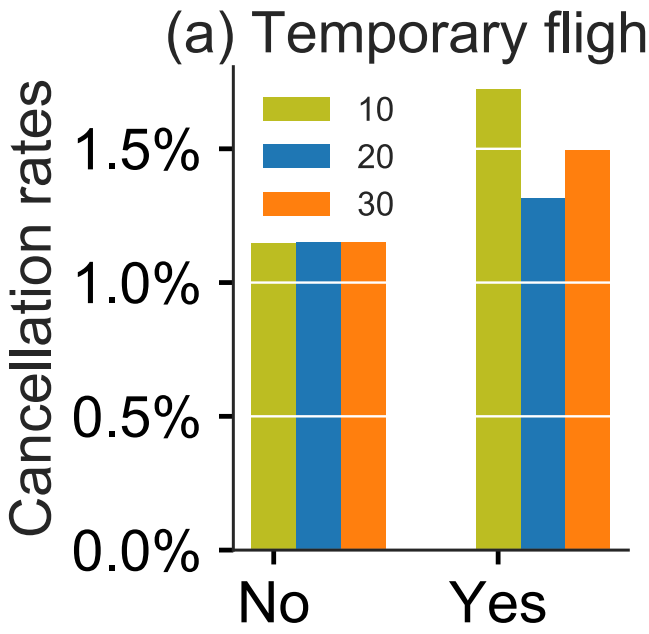
Flight historical performances

Cancellation and Diversion History

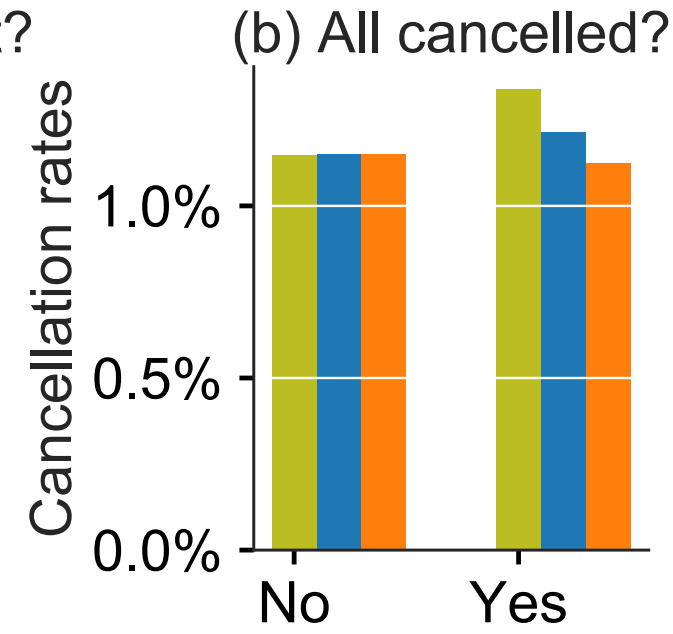




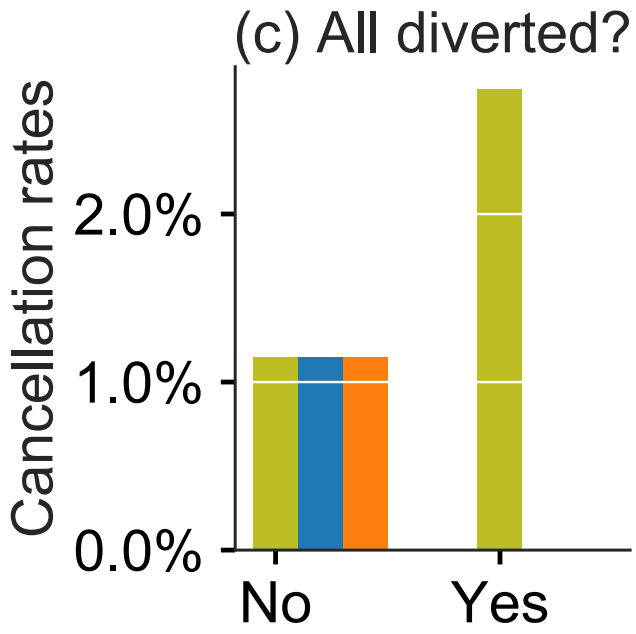
Some Historical Indicators



Temporary flight: A flight which has history of no flights in lasy ndays

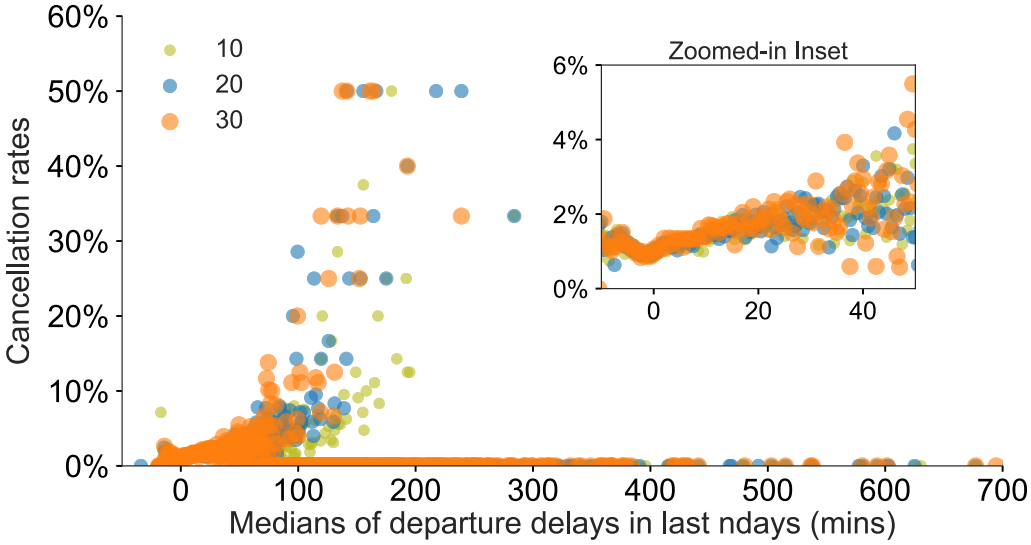


All cancelled: A flight which has history of 100% cancellations



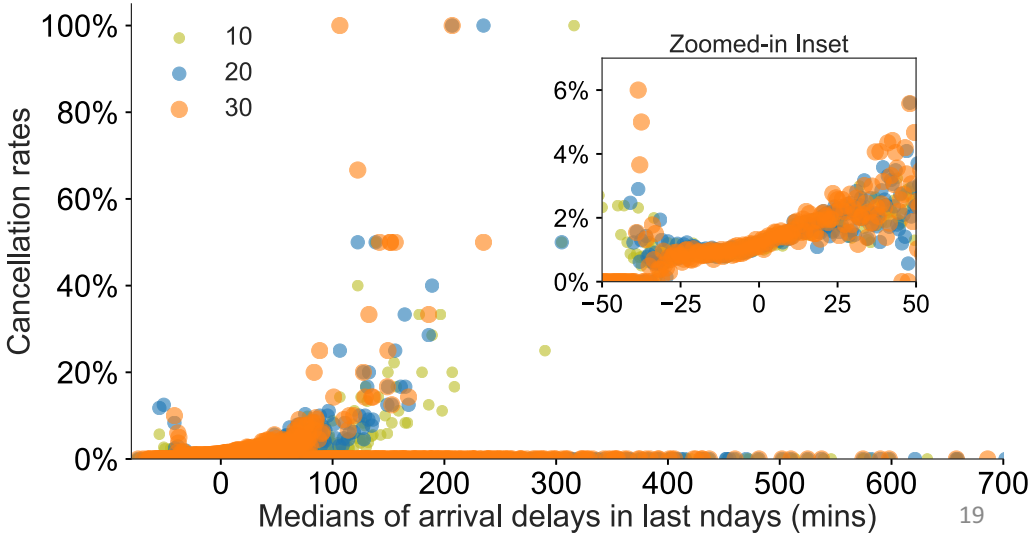
All diverted: A flight which has history of 100% diversions

Departure and Arrival Delay History



Departure Delay

Arrival Delay



Machine Learning Modeling

https://github.com/aajains/springboard-datascience-intensive/tree/master/capstone_project/Modeling

Modeling Overview

Type: Supervised learning

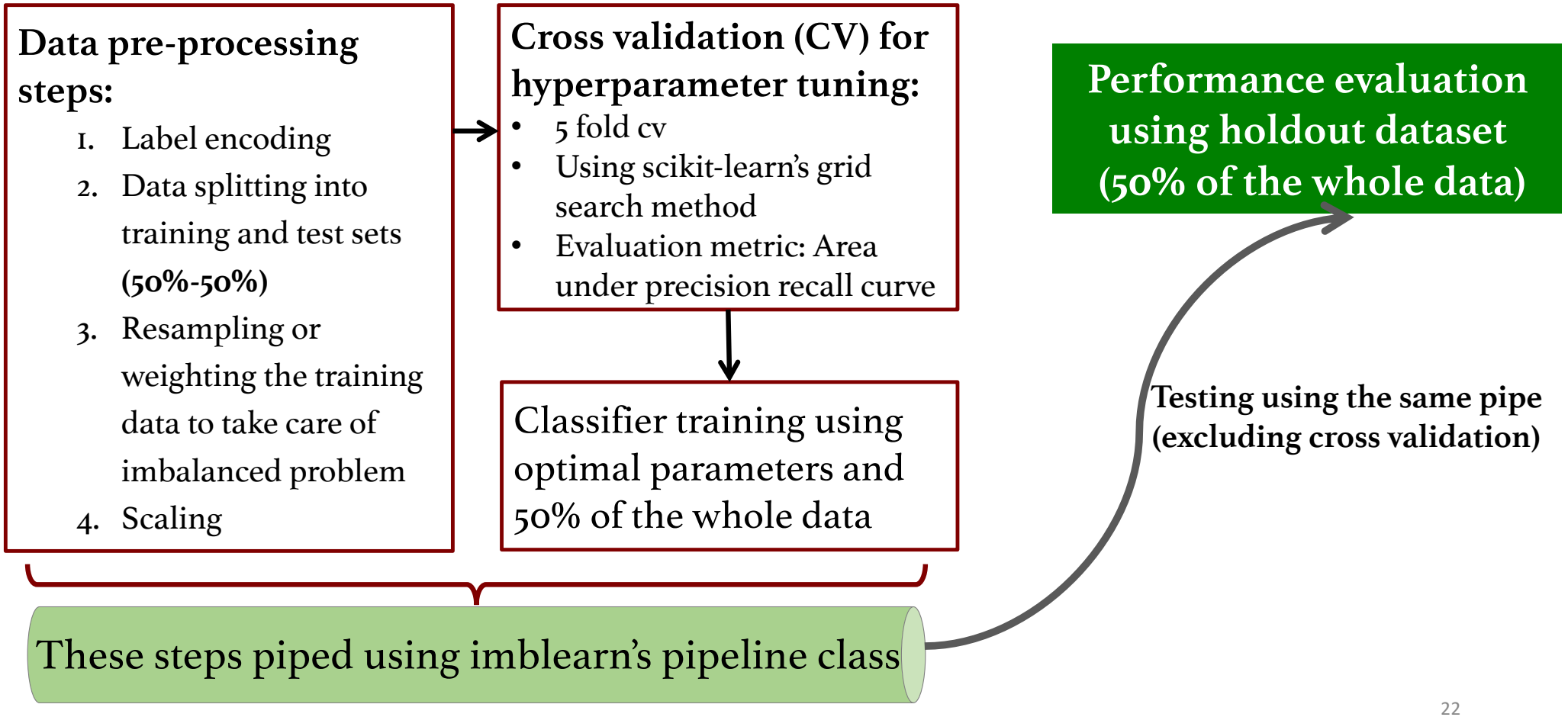
Binary classification: 1 for cancelled and 0 for non-cancelled flights

Highly imbalanced data: 1.14% data tagged with class 1

Needs special attention

Tools: Python's scikit learn and imblearn

Modeling Steps



Resampling/Weighting Techniques Used:

Resampling techniques (imblearn):

- Random under-sampling (RUS)
- Synthetic Minority Oversampling Technique (SMOTE)

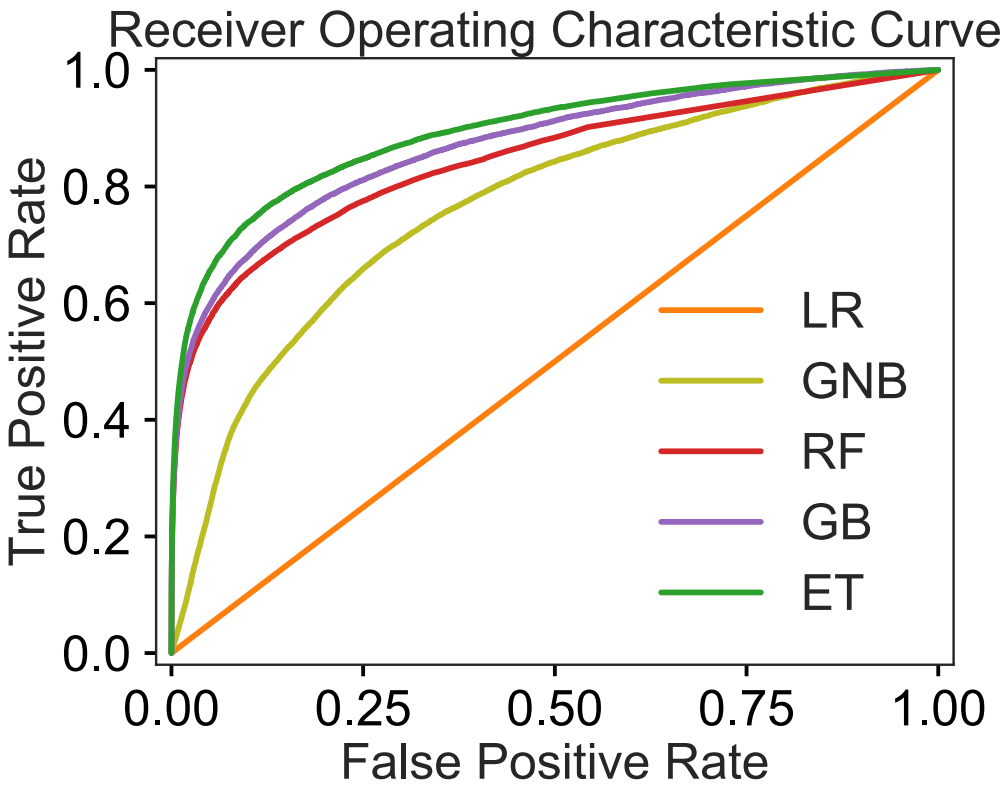
Weighting technique (sklearn):

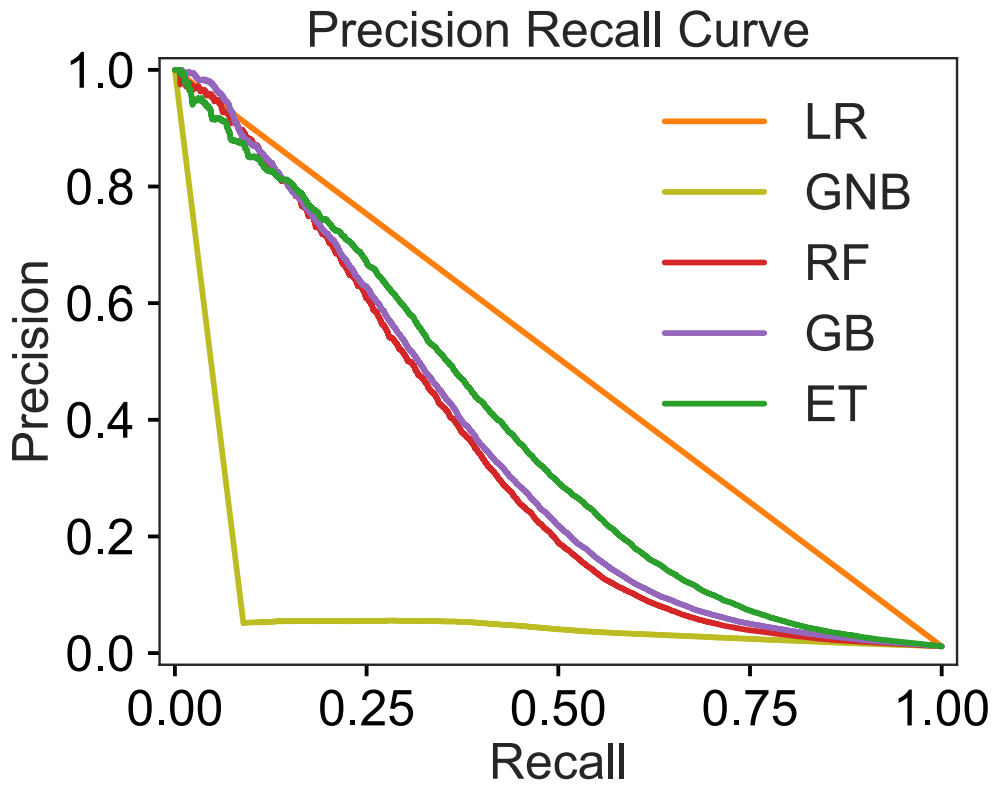
- Using class_weight (='balanced') parameter in several scikit-learn's classifier implementations

Classification Algorithms Used:

1. Logistic Regression – RUS works best
2. Gaussian Naïve Bayes – RUS works best
3. Random Forest – class_weight = 'balanced' works best
4. Gradient Boosting – SMOTE works best
5. Extremely Randomized Trees – class_weight = 'balanced' works best

Model Comparisons





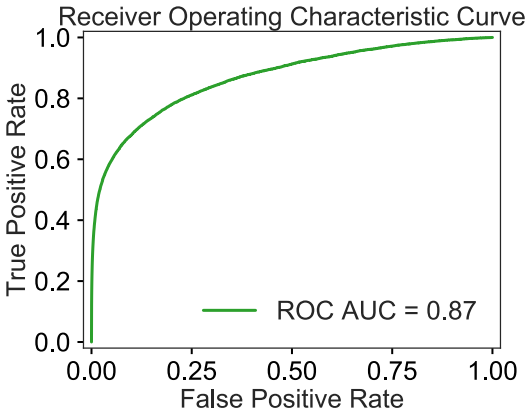
LR: Logistic Regression, GNB: Gaussian Naïve Bayes, RF: Random Forest, GB: Gradient Boosting, ET: Extremely Randomized Trees

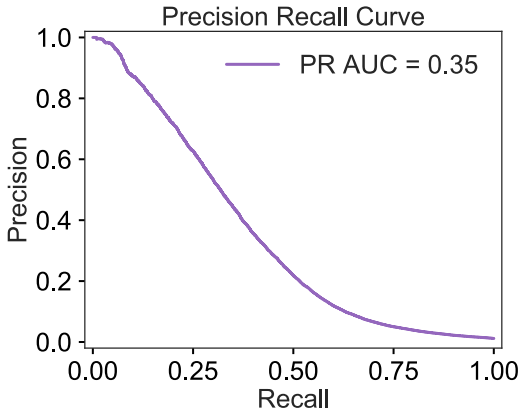
Model Comparisons

Model	PR AUC	ROC AUC	Brier Score	Log Loss
Logistic Regression	0.51	0.50	0.01	0.69
Gaussian Naive Bayes	0.08	0.76	0.14	1.55
Random Forest	0.33	0.84	0.01	0.08
Gradient Boosting	0.35	0.87	0.01	0.14
Extremely Randomized Trees	0.38	0.89	0.01	0.08

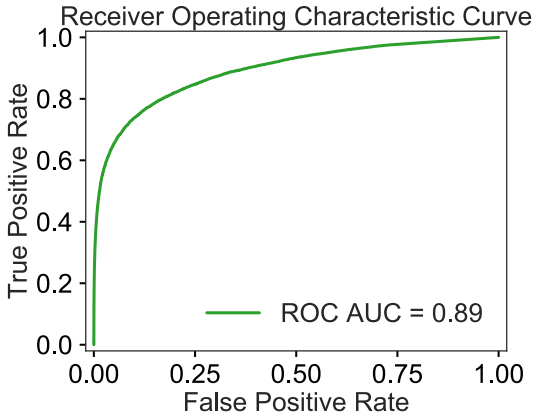
Logistic Regression is the worst and Extremely Randomized Trees is the best

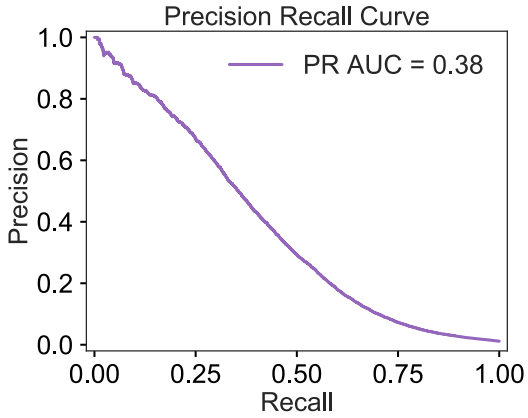
Some Details on the Best Model (ET)





- Before optimizing number of trees and set of features.
- We used 50 trees and all 67 features (note that we removed some features after merging various datasets)



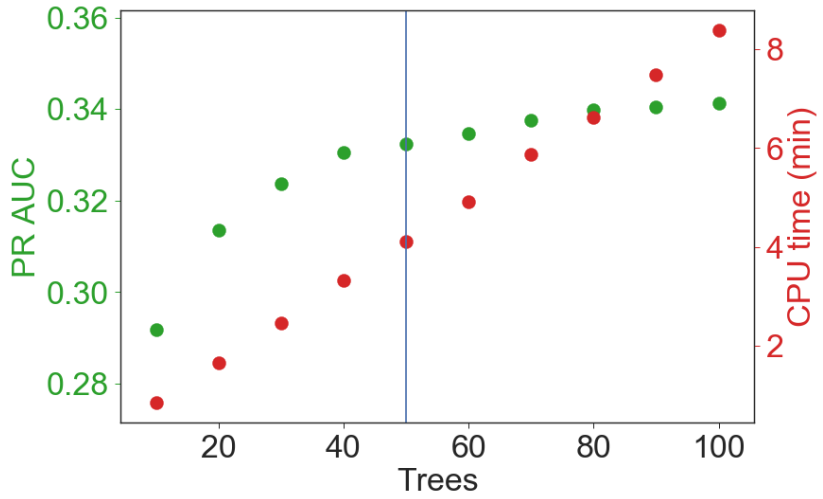


- After optimizing number of trees and set of features.
- Optimum number of trees: 50
- Optimum number of features: 31
- Used one-hot encoding (OHE), leading to 295 features

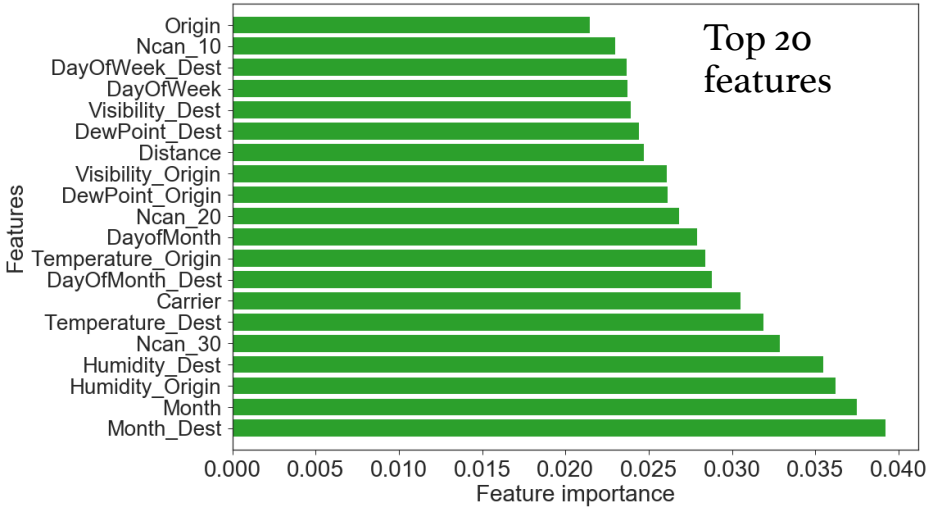
How did we optimize number of trees and set of features?

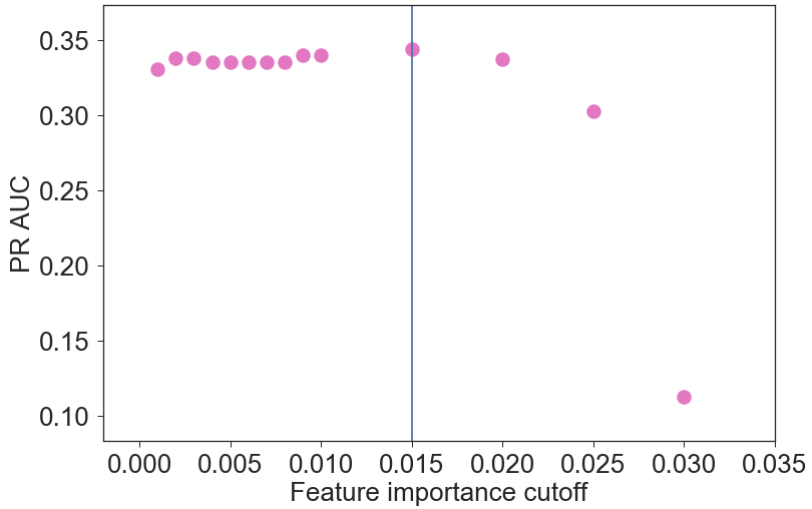
Some Details on the Best Model (ET)

Optimizing number of trees

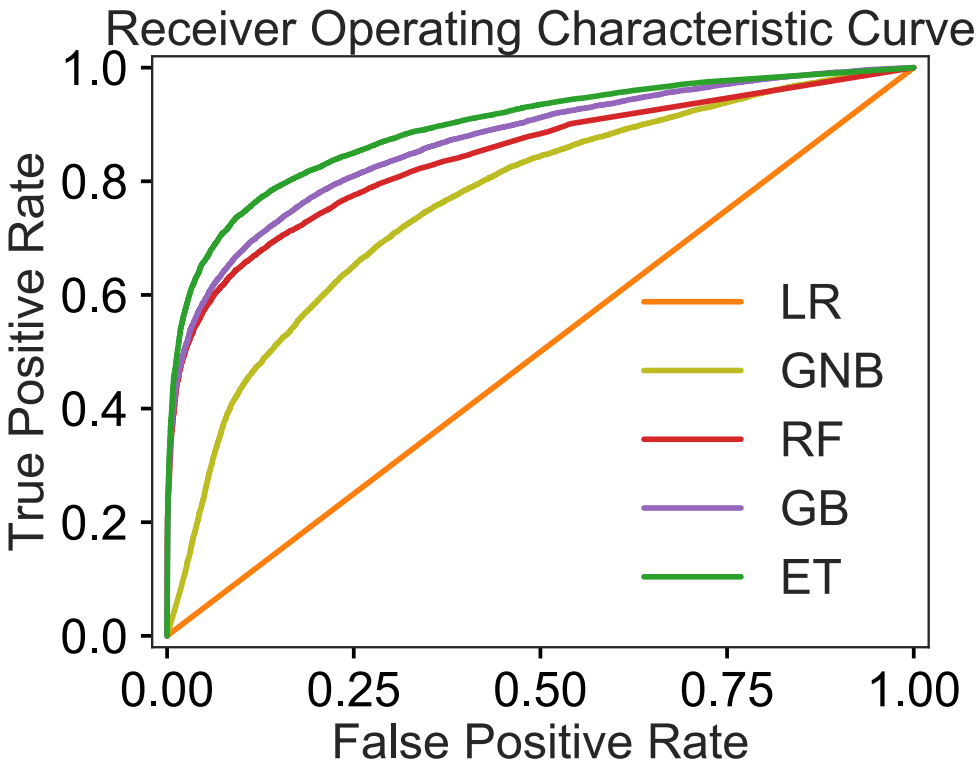


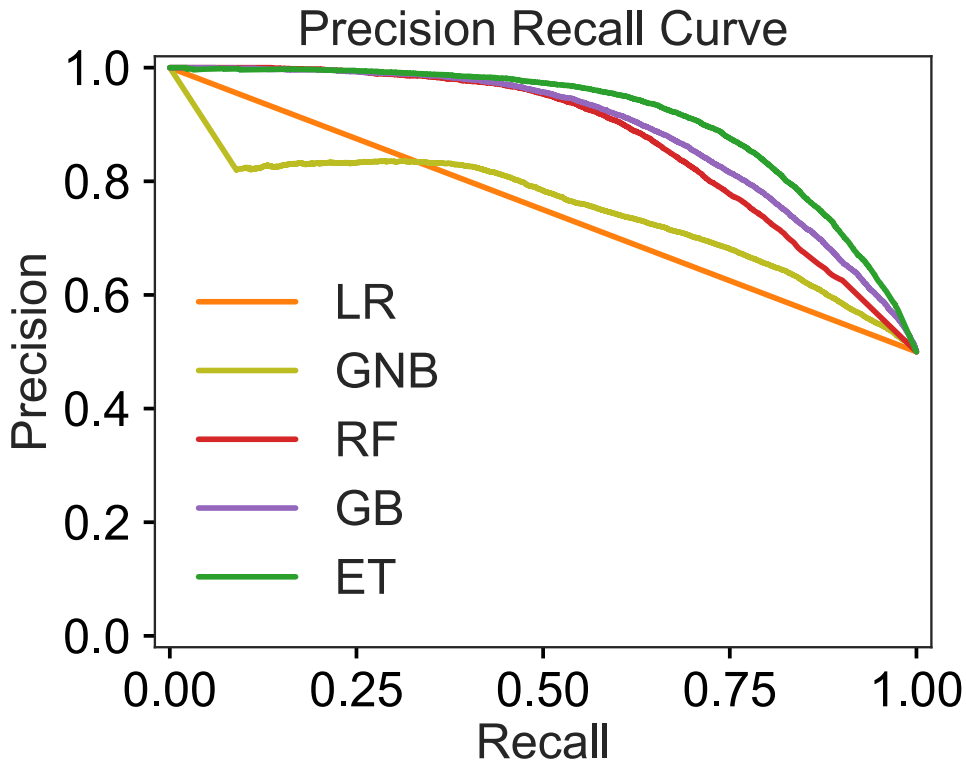
Optimizing set of features





Testing on Under-sampled Test Data





LR: Logistic Regression, GNB: Gaussian Naïve Bayes, RF: Random Forest, GB: Gradient Boosting, ET: Extremely Randomized Trees

Testing on Under-sampled Test Data

Model	PR AUC	ROC AUC	Brier Score	Log Loss
Logistic Regression	0.75	0.50	0.5	0.69
Gaussian Naive Bayes	0.75	0.76	0.32	2.42
Random Forest	0.88	0.84	0.39	2.56
Gradient Boosting	0.89	0.87	0.39	5.69
Extremely Randomized Trees	0.91	0.89	0.30	1.09

Logistic Regression is the worst and Extremely Randomized Trees is the best

Using the Model

- Select the top 31 features from the dataset.
- Note that Ncan_10, Ncan_20 and Ncan_30 are not originally present in any dataset. One needs to calculate (or engineer) them beforehand
- Perform one-hot encoding on categorical variables
- Use model pipeline on the new data and predict probabilities for flight cancellation

Features	Source type	Data type	OHE required?	Description
Month_Dest	Flight data	Categorical	Yes	Month at destination
Month	Flight data	Categorical	Yes	Month at origin
Humidity_Origin	Weather data	Numerical	No	Humidity at origin (%)
Humidity_Dest	Weather data	Numerical	No	Humidity at destination (%)
Ncan_30	Flight history data	Numerical	No	Number of cancellations in last 30 days
Temperature_Dest	Weather data	Numerical	No	Temperature at destination (°F)
Carrier	Flight data	Categorical	Yes	Airline carrier
DayOfMonth_Dest	Flight data	Categorical	Yes	Day of month at destination
Temperature_Origin	Weather data	Numerical	No	Temperature at origin (°F)
DayOfMonth	Flight data	Categorical	Yes	Day of month at origin
Ncan_20	Flight history data	Numerical	No	Number of cancellations in last 20 days
DewPoint_Origin	Weather data	Numerical	No	Dew point at origin (°F)
Visibility_Origin	Weather data	Numerical	No	Visibility at origin (miles)
Distance	Flight data	Numerical	No	Flight distance (miles)
DewPoint_Dest	Weather data	Numerical	No	Dew point at destination (°F)
Visibility_Dest	Weather data	Numerical	No	Visibility at destination (miles)
DayOfWeek	Flight data	Categorical	Yes	Day of week at origin
DayOfWeek_Dest	Flight data	Categorical	Yes	Day of week at destination
Ncan_10	Flight history data	Numerical	No	Number of cancellations in last 10 days
Origin	Flight data	Categorical	Yes	Origin airport
Dest	Flight data	Categorical	Yes	Destination airport
WindDirection_Origin	Weather data	Numerical	No	Wind direction at origin (degrees)
WindSpeed_Origin	Weather data	Numerical	No	Wind speed at origin (mph)
WindDirection_Dest	Weather data	Numerical	No	Wind direction at destination (degrees)
Condition_Origin	Weather data	Categorical	Yes	Weather condition at origin
WindSpeed_Dest	Weather data	Numerical	No	Wind speed at destination (mph)
Condition_Dest	Weather data	Categorical	Yes	Weather condition at destination
Pressure_Origin	Weather data	Numerical	No	Pressure at origin (inHg)
CRSDepHr	Flight data	Categorical	Yes	Scheduled departure hour
Pressure_Dest	Weather data	Numerical	No	Pressure at destination (inHg)
CRSArrHr	Flight data	Categorical	Yes	Scheduled arrival hour

An Example of Model Usage: Possible Recommendations

Probability ranges	**Alert messages**
$p(\text{cancelled}) > 0.9$	Extremely high chance
$0.75 < p(\text{cancelled}) \leq 0.9$	Very high chance
$0.5 < p(\text{cancelled}) \leq 0.75$	High chance
$0.25 < p(\text{cancelled}) \leq 0.5$	Moderate chance
$0.1 < p(\text{cancelled}) \leq 0.25$	Low chance
$p(\text{cancelled}) \leq 0.1$	Very low chance

Assumptions, Limitations and Disclaimers

- We assume that all flights are independent, though there might be some time-correlations
- Used only 20 airports and two years of data
- The model will behave poorly if we try to predict cancellations of flights that are scheduled too far in future
- The flight data does not contain all airlines data (e.g. Sun Country Airlines data is missing in the current flight dataset)

More Ideas to Improve the Model in Future

- Engineer more features related with airports such as number of runways, airport capacity, airport infrastructure, etc.
- Extract more information about airlines such as their ratings, stock market performance etc., to get more features
- Similar to flight historical performances, generate features with airport historical performances
- Use social network and news media to extract sentiments about airlines and airport to get more features

Conclusions

- All sources of datasets contributed to the predictive power of the model.
- Out of 5 supervised classification models, the Extremely Randomized Trees provided the best results.
- Out of 67 features, we used only 31 features for the best model with 12 from the flight data, 16 from the weather data and 3 from the flight historical performances data (which we engineered).
- With 50%-50% splitting, the test data set gave ROC AUC = 0.89.
- With more ideas, the model can be improved in the future.

Thank you!

Aashish Jain, PhD
Email: aajains@gmail.com
https://www.linkedin.com/in/aashishjain/
https://github.com/aajains/
Project report: https://github.com/aajains/springboard-datascience-intensive/blob/master/capstone_project/Report/CapstoneReport.pdf