Airlines delays analysis

Overview

Data: 2016, 2017

Total: 9M flights

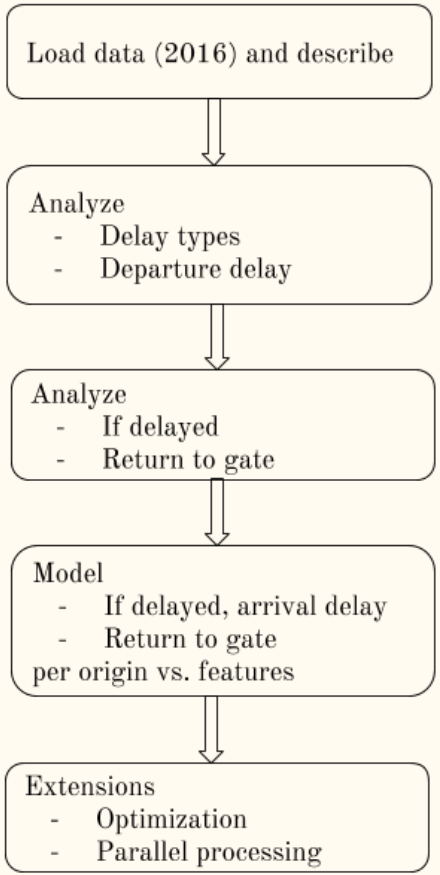
Delayed: 1.6M

Return to gate: 54k

Carriers: 12

Origins: 317

Routes: 8622



Outcomes:

=> when to fly

=> which airline from which origin to avoid

=> predict if delayed and return to gate probabilities

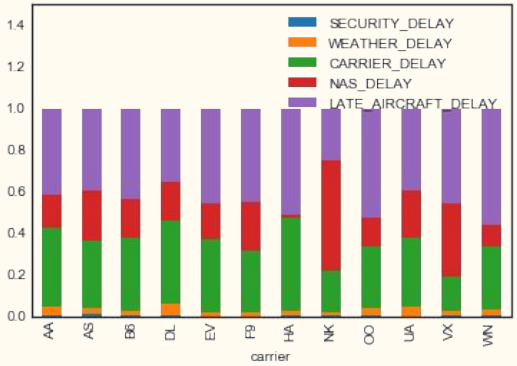
=> predict arrival delay

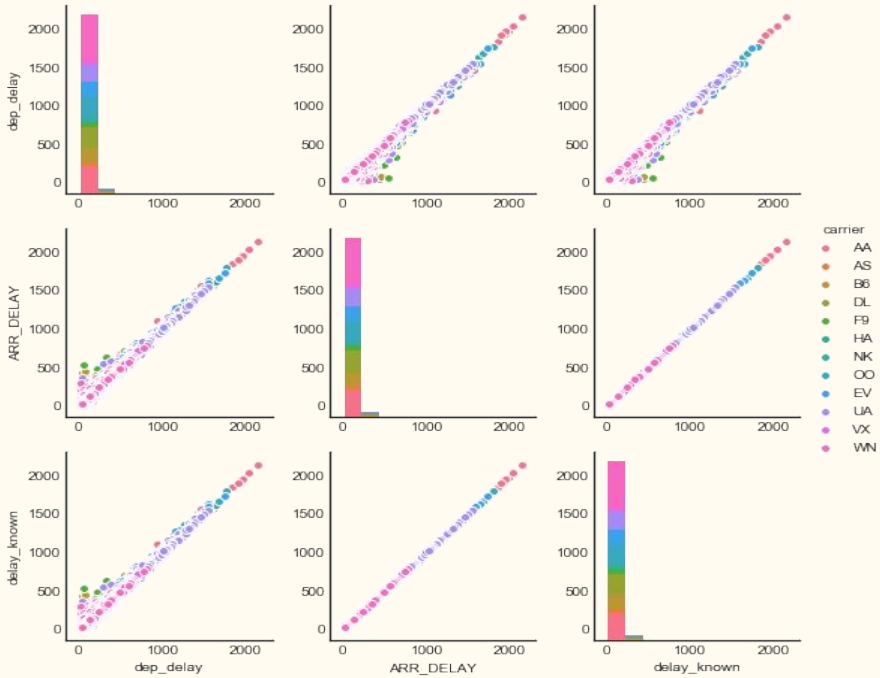
Delay types

For delayed flights:

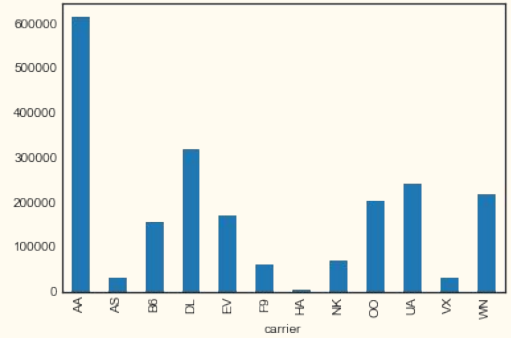
Delay known = sum(delays):

Causes delays:



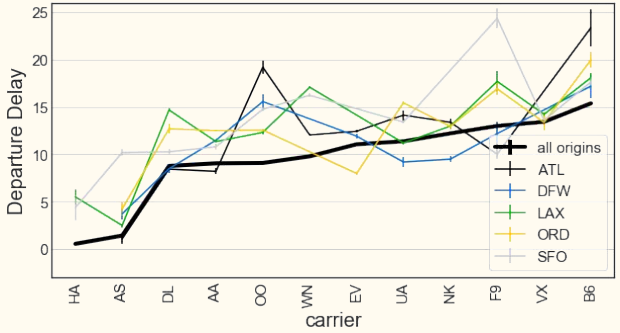


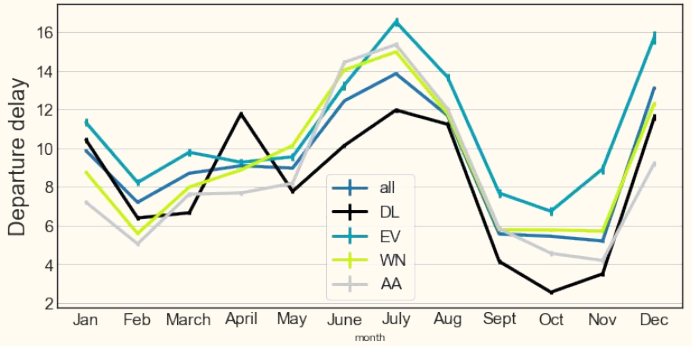
!High # flights with Arrival - Departure delays >1h:

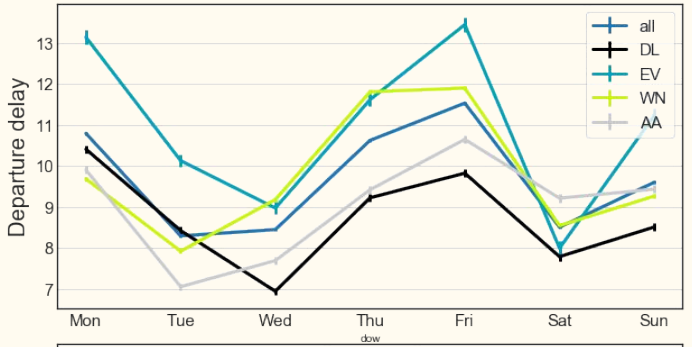


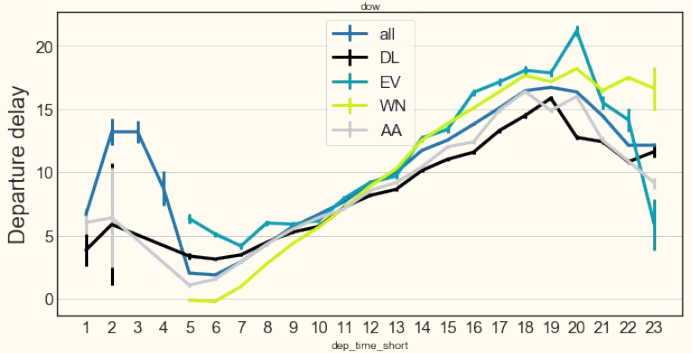
Departure delay

vs. Month, Hour of day, Day of Week, Carrier:







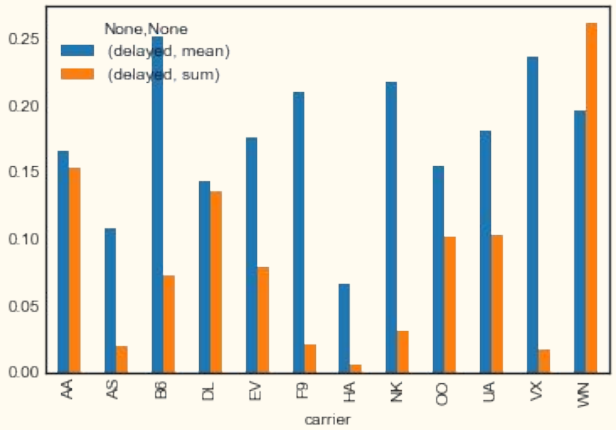


Impact on delay:
similar along carriers and origins.

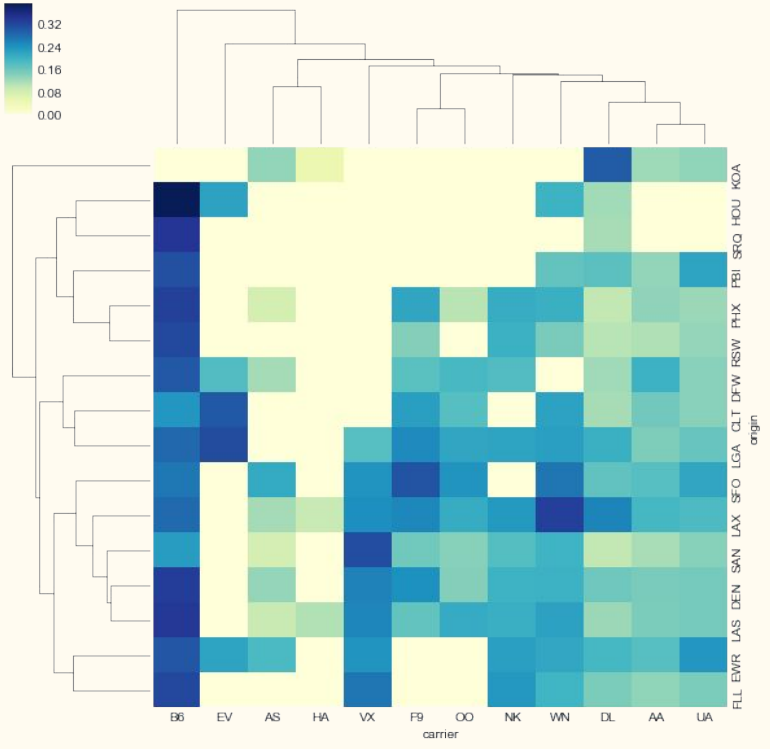
Analyze if delayed

= delay > 15 min

% Delayed, % Total flights per carrier:



Carriers + Origins
>1k flights & delayed >20%:



Models

- Outputs: Delayed, RTG, Arrival delay
- 1 model per Origin
- Features:
  - Categorical (One hot encoded): Hour, Day, Month, Carrier, Destination
  - Continuous: flight_duration, count_flights_carrier, count_flights_origin, count_flights_route, distance

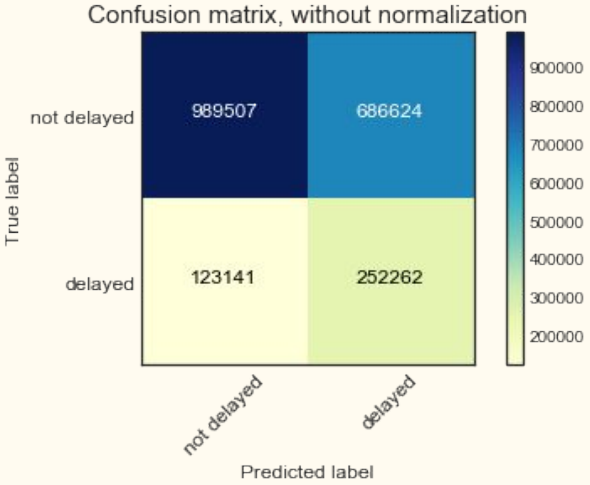
Model if delayed

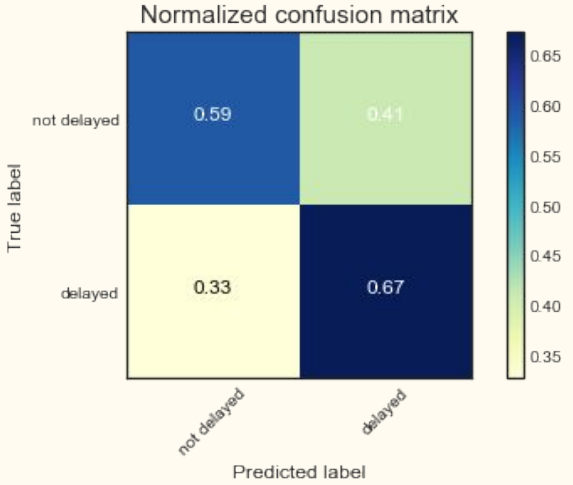
Model per origin

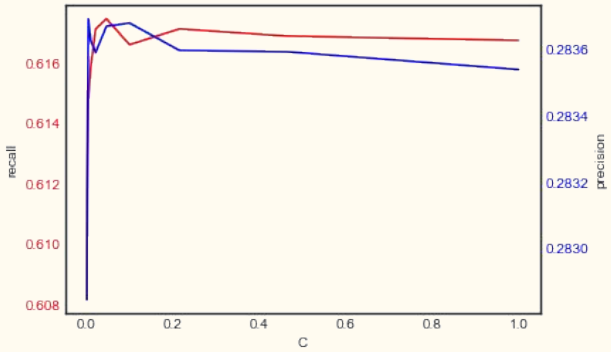
Logistic regression

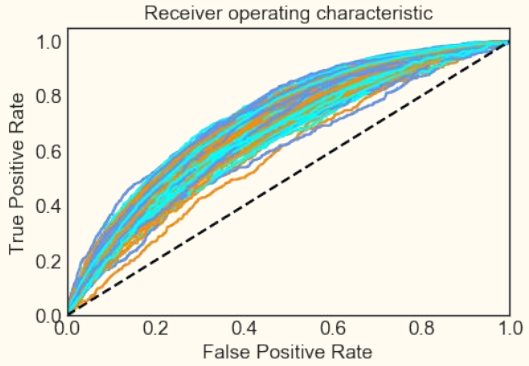
Weight = Balanced

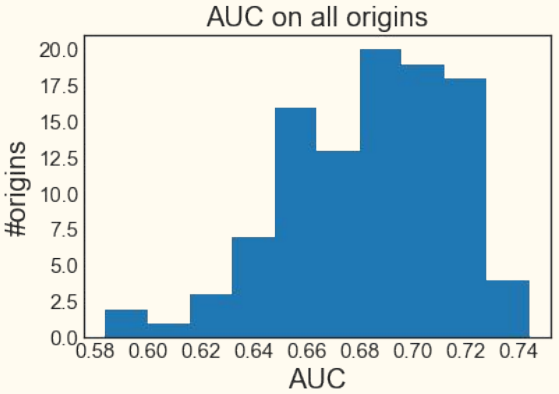
Test set: 25%









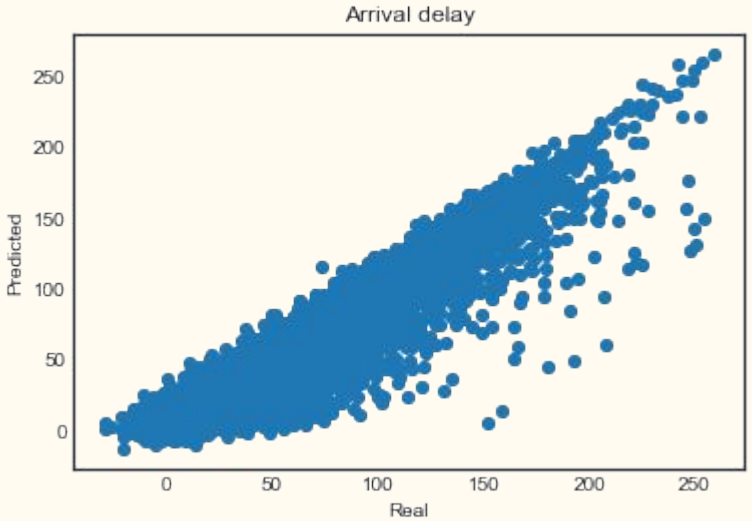


Arrival delay

- Linear regression (statsmodels) with categorical features

  ARR_DELAY ~ C(dow) + C(dep_time_short) + C(carrier) + C(dest) + dep_delay + count_flights_carrier + count_flights_origin + count_flights_route + distance + TAXI_OUT + flight_hours
- Dep_delay >15 min, <3 h
- E.g. origin = ATL:
- Test score: R^2: 94%,

  Mean abs. err: 6.3 minutes
- Good for over-approximation limit
- Can help to announce follow-up delays

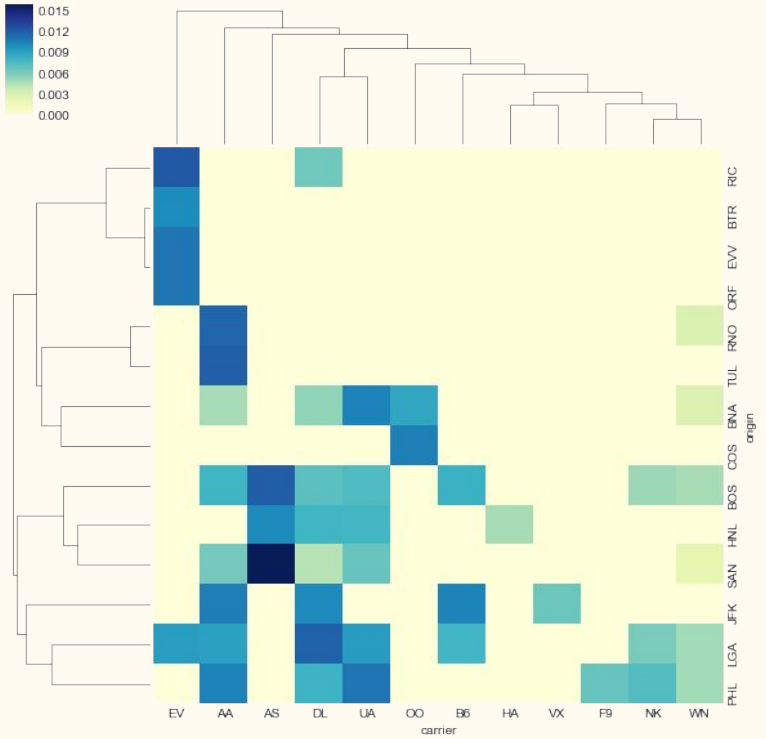


Analyze Return to gate

%RTG, for groups airline-origin with

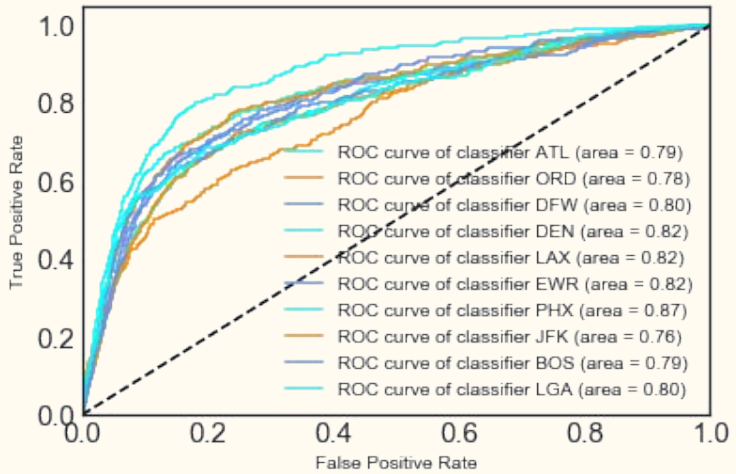
>1k flights, >1%RTG, >30 RTG

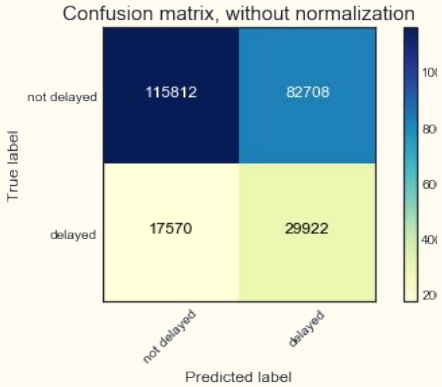
=> groups of carriers + origins

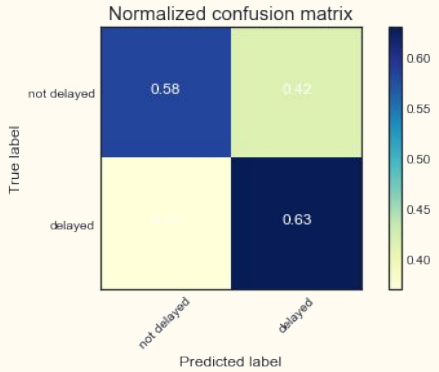


Model Return To Gate

+ Extra features:
  late aircraft delay, carrier delay







Parallel processing & Performance

- Sandbox Hortonworks + Pyspark
- Load data in Hdfs (csv)
- Distributed ML alternatives
  - 1: build feature matrix and labels (X, y) using Spark, train in Python
  - 2: train sequentially by mini-batches using partial fit on SGD classifier
  - 3: use Pypark MLlib
  -

- Difficulties & potential solutions
  - Many categ features. Embedding could be better than one hot encoding.
  - Many obs if using historical data. Reduce to stats or batch training could help.

Optimization

- Rank best times & airlines by probability (delayed + RTG)
- Cluster carriers/routes by performance
- Add features
  - Weather (wind, etc.)
  - Location
- Use historical information (time series)
- Try other techniques
  - Xgboost, random forest, deep neural networks
  - RTG outlier analysis
  - Graph (clustering) => suggest best routes

Summary

- To minimize probable delay
  - Certain Times to fly and Origin - Carriers should be avoided
- To notify arrivals more accurately and estimate follow-up delays
  - Arrival delay can be estimated on departure delay

- More interesting outcomes could be investigated
  - How late aircraft delays propagate through the network
  - How to group carriers/routes/times best
  - Why some specific routes are more often delayed/affected by RTG

Delay causes, Airline codes

- Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.
- Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

AA American Airlines Inc.
AS Alaska Airlines Inc.
B6 JetBlue Airways
DL Delta Air Lines Inc.
EV Atlantic Southeast Airlines
F9 Frontier Airlines Inc.
HA Hawaiian Airlines Inc.
MQ American Eagle Airlines Inc.
NK Spirit Air Lines
OO Skywest Airlines Inc.
UA United Air Lines Inc.
US US Airways Inc.
VX Virgin America
WN Southwest Airlines Co.