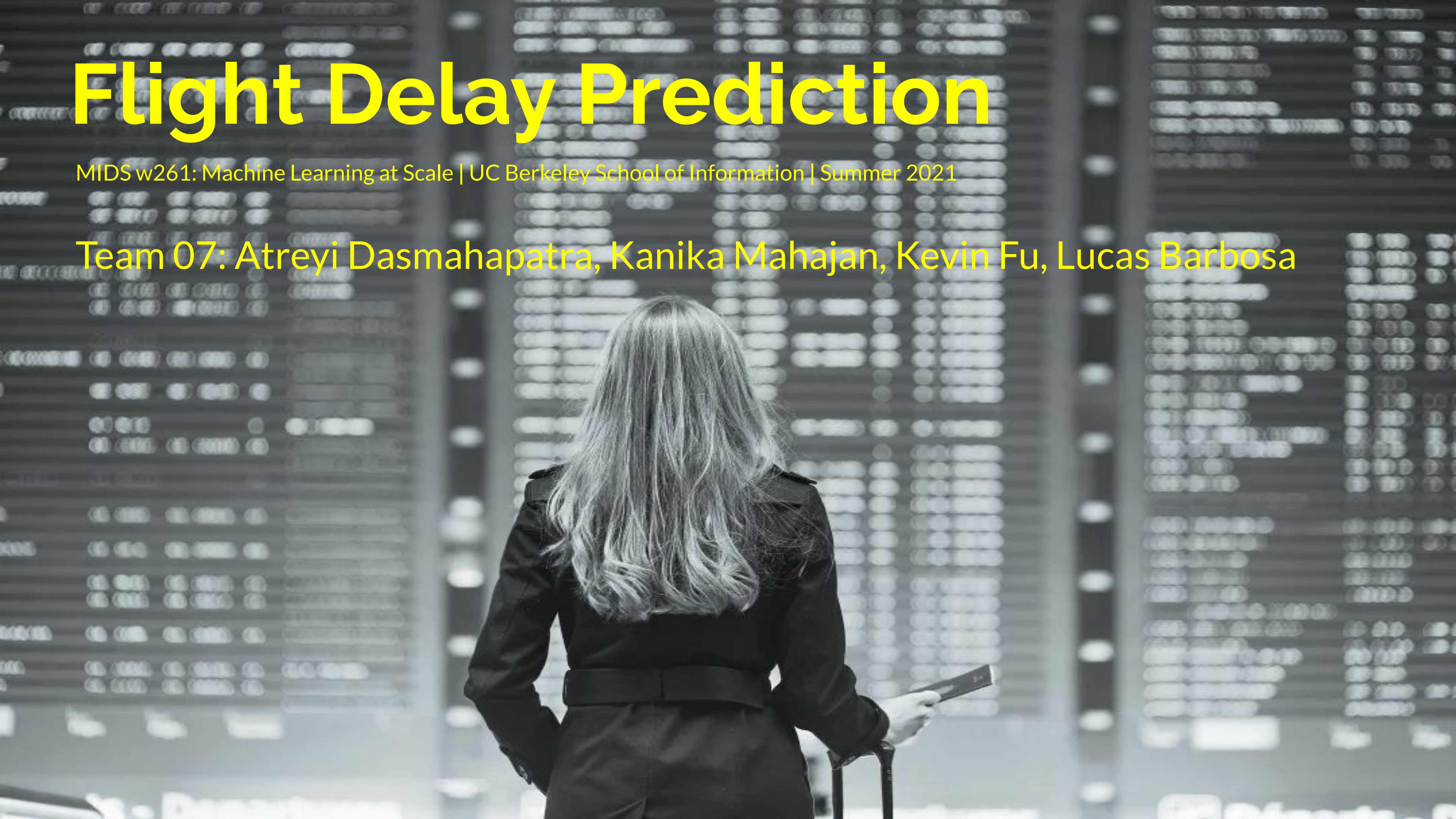
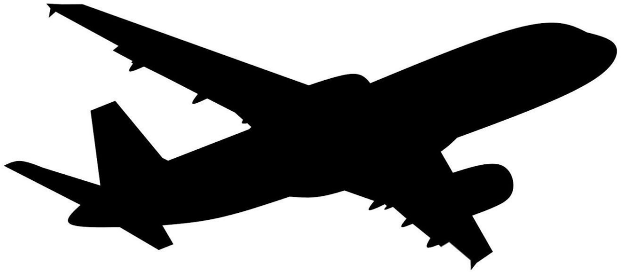
Flight Delay Prediction

MIDS w261: Machine Learning at Scale | UC Berkeley School of Information | Summer 2021

Team 07: Atreyi Dasmahapatra, Kanika Mahajan, Kevin Fu, Lucas Barbosa



Business Case



- Predicting flight delays allow airlines to do **better delay management**, optimizing their operations and saving costs
- Ex. should a connecting flight be intentionally delayed if a feeder flight arrives with a delay?
- If not, transferring passengers will lose their connection; if yes, another delay might cause ripple effects in the network

With in advance delay predictions, airlines do have more flexibility to reschedule in an optimized way

Recall is key

- Since most of flights do not delay, a classifier that simply predicts that no flights will be delayed can achieve remarkably high accuracy rates
- Thus, our main challenge is to instead **improve recall**, the percentage of delayed flights that are correctly classified as delayed
- **Key assumption:** costs incurred in anticipation of a potentially delayed flight that doesn't delay are lower than the costs of a delayed flight for which the airline has not prepared itself

The datasets we used

- Airline Dataset: Bureau of Transportation Statistics (BTS).
  - 31 million rows
  - 109 potential features
- Weather Dataset: Integrated Surface Data (ISD)
  - 630 million rows
  - 177 potential features
- Aviation Support Tables: Office of Airline Information, Bureau of Transportation Statistics
  - 18K rows
  - 10 potential features

Key findings from EDA: weather data

- Many unusable/null values
  - 7 predictive features, 42 million rows remain
- Not one single feature indicates bad weather
  - 98.6% normal wind conditions
  - 4.2 m/s avg wind speed (gentle breeze)
  - 0.8 correl (air temp, dew temp)
- Led to creation of Bad Weather Predictor:
  - >50% relative humidity &
  - Wind speed >15.2 m/s |
  - Required visibility <1609 m |
  - Low sea pressure (bottom 10%)

$$\text{Relative Humidity } \% = \frac{E}{E_s} \times 100$$

$$E_s = e^{\left(\frac{17.67 \times T}{243.5 + T}\right)}$$

$$E = e^{\left(\frac{17.67 \times T_{dew}}{243.5 + T_{dew}}\right)}$$

T = Ambient Temperature in Celsius

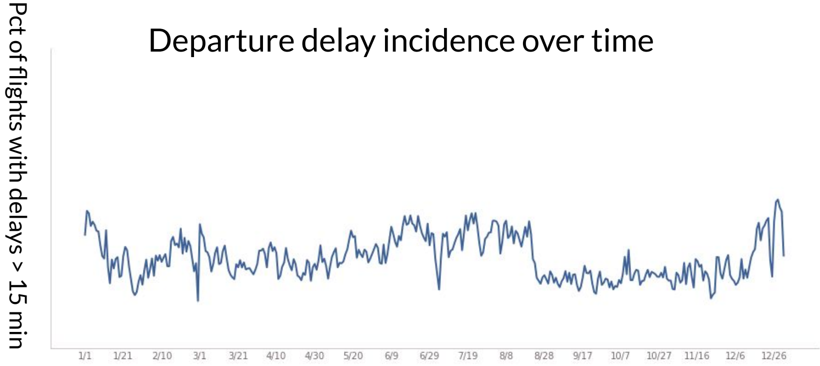
T_{dew} = Dew Point in Celsius

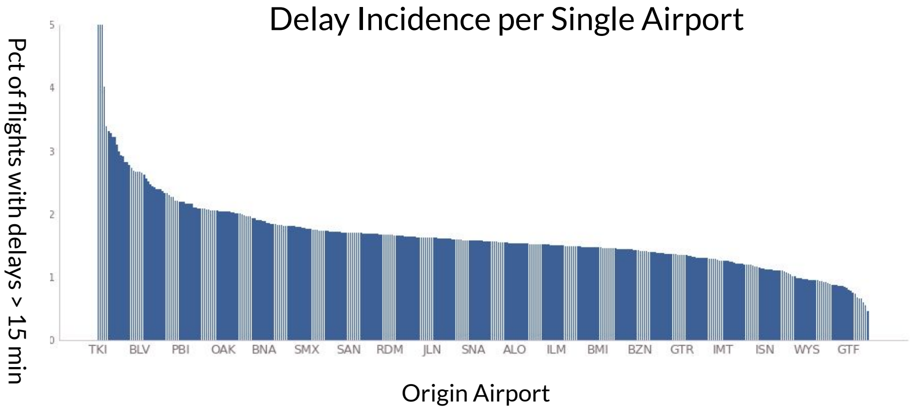
E_s = Saturation Vapor Pressure

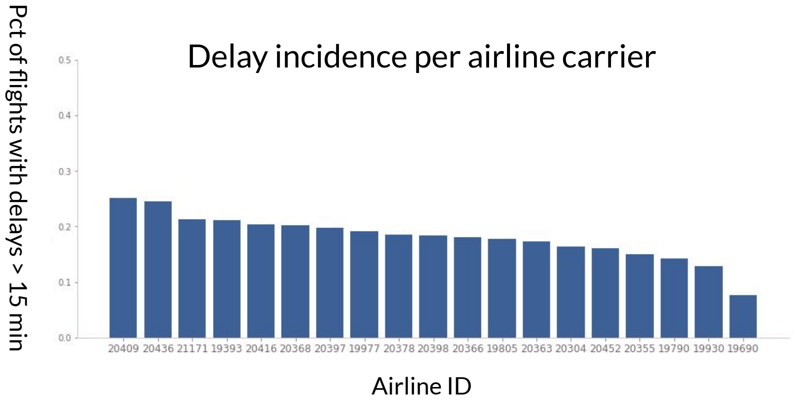
E = Actual Vapor Pressure

Key findings from EDA: airlines data

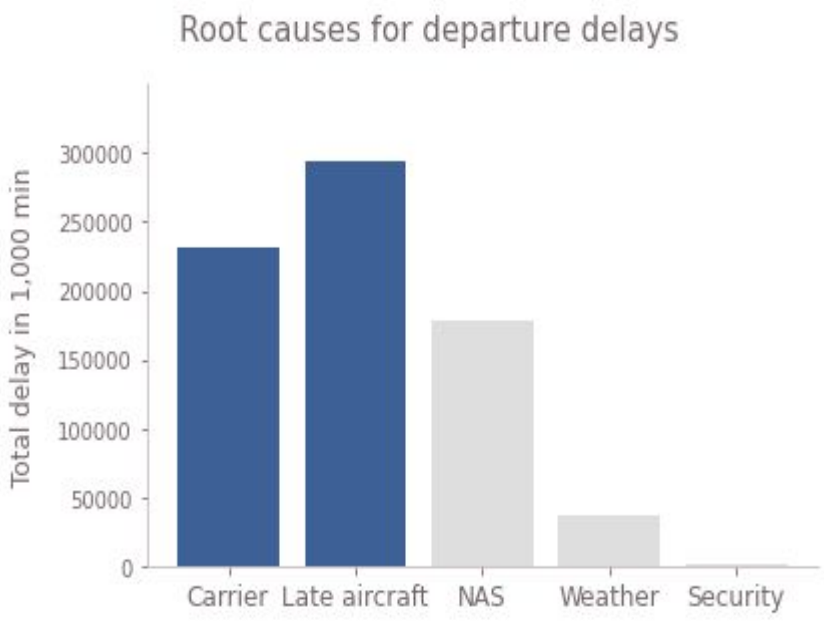
- Duplicate and Missing Data (61 potential features)
- Cancelled and Diverted Flights
- Average delay time ~ 12 minutes
- Average flight duration (~800 miles, 2.5 hours)
- Imbalance in outcome of interest
- Delayed flights not uniform across days, airlines, airports







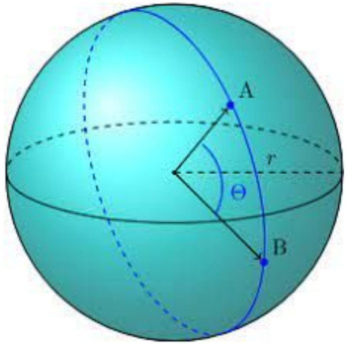
Flight delays and feature creation



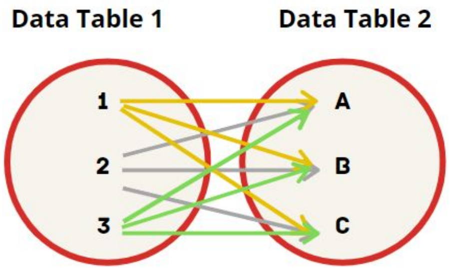
- Air Carrier delays:
    - Average delayed flights (rolling window 30 days) for every aircraft.
- Late Aircraft delays:
    - Frequency of delayed flights at origin, destination airport and by airline carrier in the past 2, 4, 8 and 12 hours
- NAS delays:
    - Frequency of late arrivals in the departure airports
    - Delays in hubs in the past 2,4,8 and 12 hours
    - Part of Day (Morning, Afternoon etc.)

Joining Data Tables

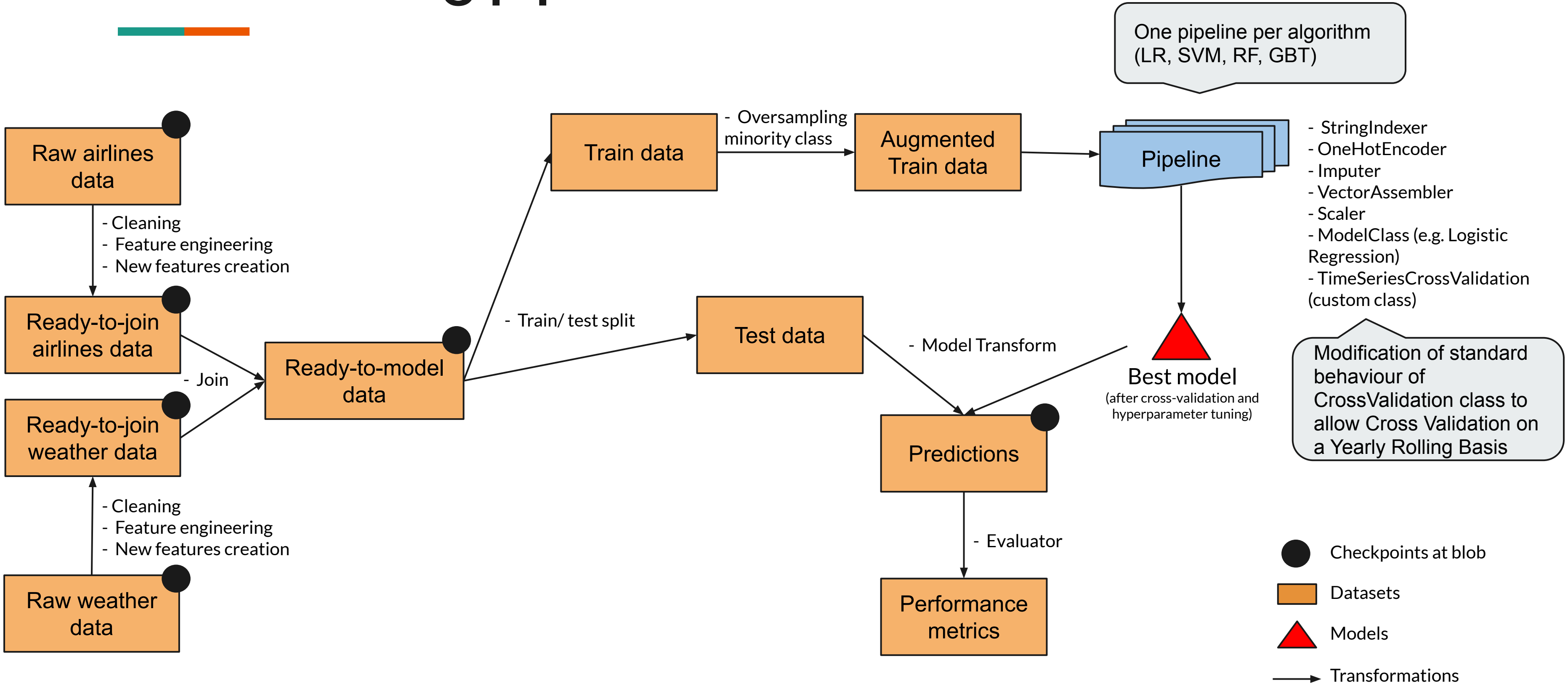
- Nearest weather station to departing airport
  - Use Google geocode api to get latitude and longitude of departing city
  - Map back to matching weather station latitude and longitude using Haversine formula
  - Cross Join first and then inner join
  - Latest weather timestamp
  - 26513705 rows and 128 features
  - Very few empty rows (as expected)



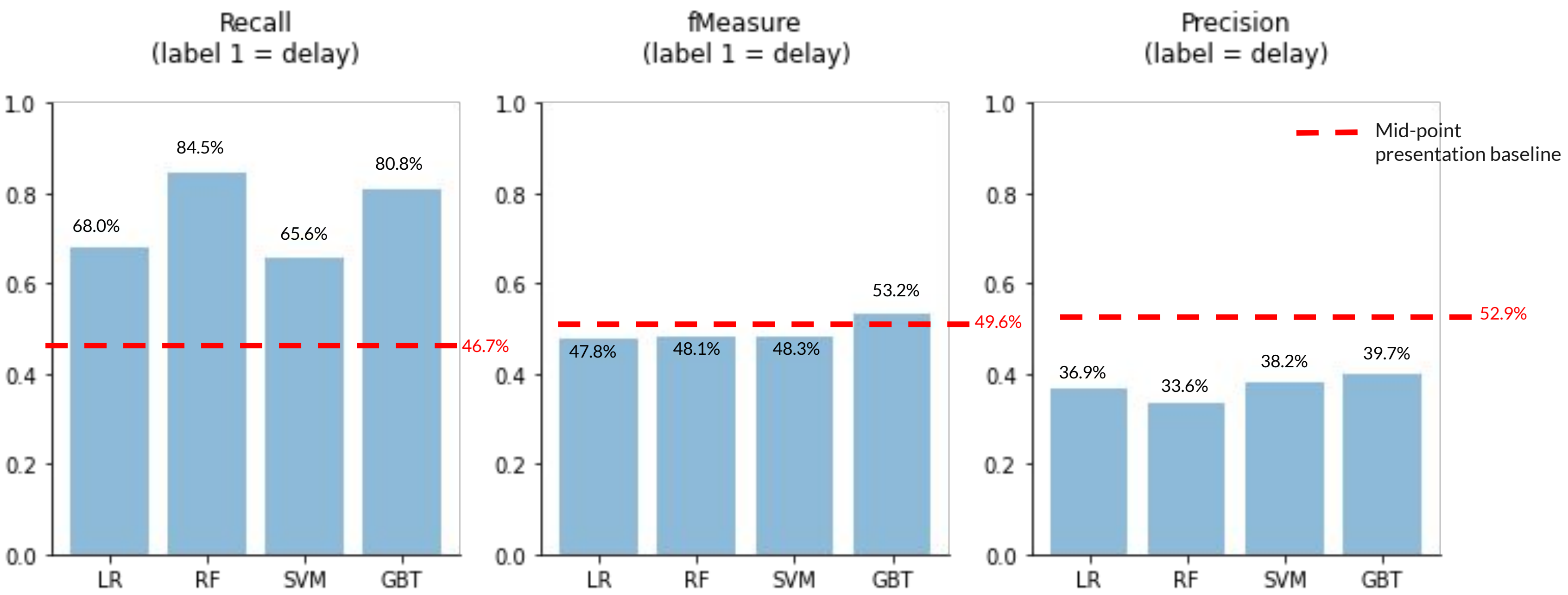
Great-circle distance between two points on a sphere



Our modelling pipeline



Results



Recall
(label 1 = delay)
fMeasure
(label 1 = delay)
Precision
(label = delay)
Mid-point
presentation baseline
1.0
0.8
0.6
0.4
0.2
0.0
68.0%
84.5%
65.6%
80.8%
46.7%
47.8%
48.1%
48.3%
53.2%
49.6%
36.9%
33.6%
38.2%
39.7%
52.9%
LR
RF
SVM
GBT

Conclusions and Challenges

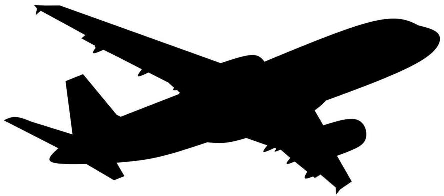
- First ever ML project with datasets that couldn’t be tweaked in Pandas
- Application of LR, RF, SVM and GBT algorithms
- LR from scratch: recall of 0.43 (vs 0.68) and a precision of 0.29 (vs. 0.37)
- Flight data lacked timestamps and had local time - calculate local time zones convert both flight & weather data to unix timestamps
- A common column between the two data tables to join did not exist
- The joined dataframe gets saved as a parquet file without any complaints but count of the data frame is drastically lower upon reloading

OLD SLIDES REPOSITORY

Flight delays cause extra costs for passengers and airlines

- Longer trips
- Missed connections
- Lengthy stays at airports

Personal and professional costs



- Crew/aircraft availability
- Passenger reaccommodation
- Penalties/fines

Operational costs and brand reputation

Airlines Dataset: EDA

Data Overview

- Started with airlines 3 months data: Total 109 features and 161057 rows

Missing Data

- Removed features that had > 96% of data missing (mostly diverted flights)
- Usable features went down to 61

Data Cleaning

- Cancelled flights removed from dataset (unable to classify if delayed or not)
- Diverted flights left as such

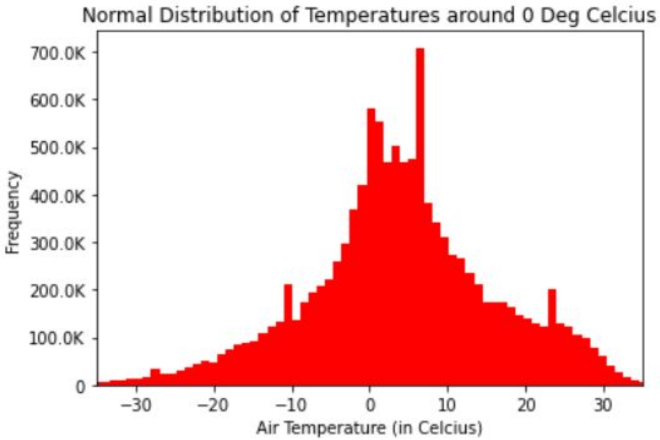
Feature engineering - Airlines data

- Part of the day (snowball effect of delayed flights early in the day)
- Average delayed flights by aircraft (tail_num) - rolling window 30 days
- Frequency of delays in the departure airport (for all airlines) in the past 2, 4, 8 and 12 hours
- Frequency of delays in the destination airport (for all airlines) in the past 2, 4, 8 and 12 hours
- Frequency of delays of the same airline (in the departure airport) in the past 2, 4, 8 and 12 hours
- Frequency of late arrivals in the departure airport (for all airlines) in the past 2, 4, 8 and 12 hours

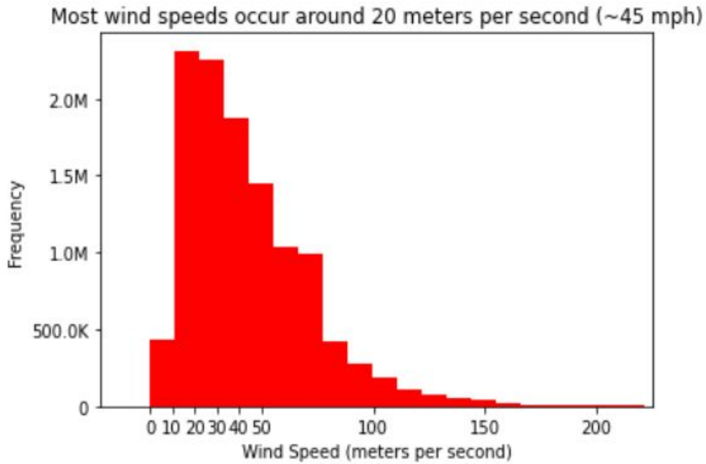
Weather Dataset: EDA

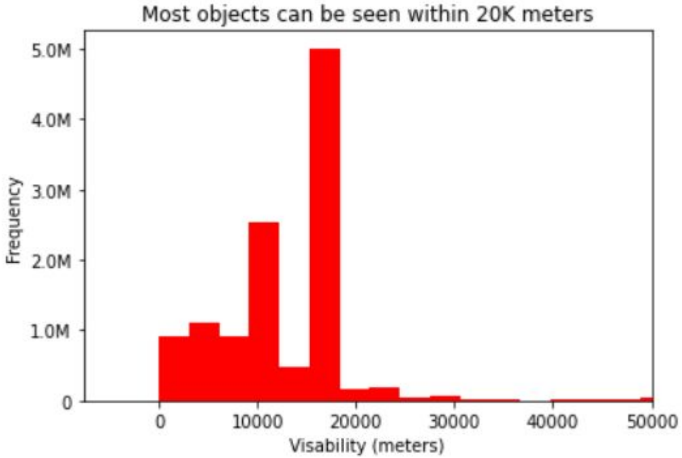
- Remove Missing/Erroneous Features (Columns)
  - 161 out of 177 features had 50%+ missing data (91% reduction)
- Remove Missing/Erroneous Rows
  - 11,588,530 rows remain out of 29,823,926 (61% reduction)
- Data Cleaning
  - Total of 19 useable features, un-nested from 6 columns

Weather Dataset: EDA



- 0.70 correl(air temp, dew temp)
- -0.23 correl(sea level, visibility)





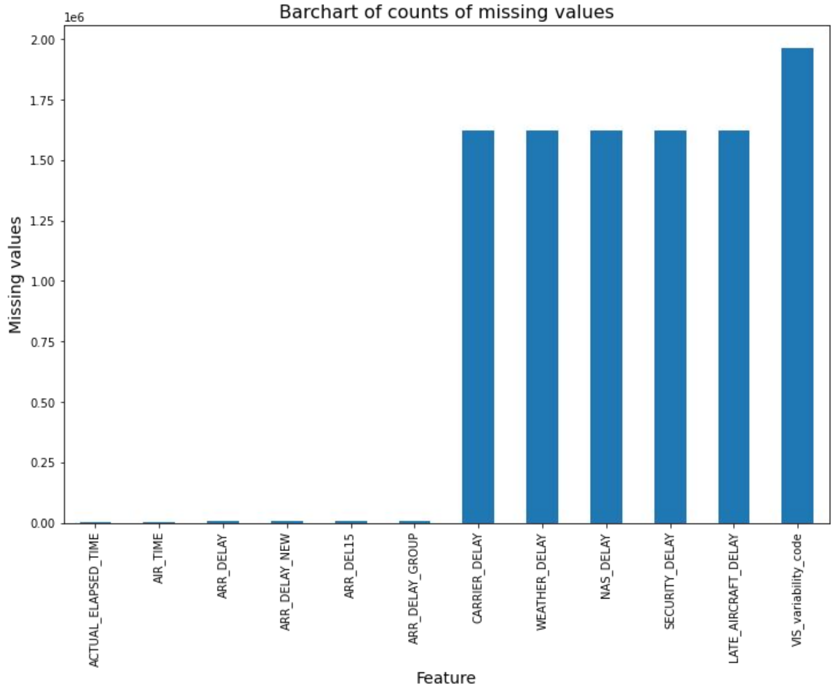
Feature engineering - Weather data

- 6 out of the 19 features could be predictive
    - Others were identifying features
    - To use Lasso as final judge to determine if var is sig predictor for flight delays
    - Filling in vs. Removing x million rows may make features more significant
- Good vs. Bad Weather
    - Within explicitly labeled variables, 90%+ “Good”, <10% “Bad”
    - Need more “Bad” weather data points: must look into 6m/Full dataset for a complete Season/more Winter data points

Joined Dataset: EDA

- 26513705 rows and 128 features
- Very few empty rows (as expected)

Name of Origin City	Bad Weather Prediction
Chicago, IL	53932
DFW, TX	35192
Atlanta, GA	32135
Phoenix, AZ	29293



Logistic Regression: baseline results

Only 4 features survived after L1 reg.; none from the weather dataset

Features	Weights
Intercept	-1.680632
Total number of **departure delays** in the **previous 2hrs** in the **origin airport**	0.002778
Total number of **departure delays** in the **previous 4hrs** in the **destination airport**	0.228222
Total number of **arrival delays** in the **previous 2hrs** in the **origin airport**	0.008814
Total number of **arrival delays** in the **previous 4hrs** in the **origin airport**	0.000651

Recall of ~60%: too good to be true?

```
Performance metrics
-------------------------------------------------
Accuracy: 0.7984
Weighted Precision: 0.7902
Weighted Recall: 0.7984
F1-Score: 0.7938
Precision By Label: [0.8607, 0.5286]
Recall By Label: [0.8878, 0.4670]
F1-Score by Label: [0.8740, 0.4959]
```

Next steps

EDA and join on complete dataset

More feature engineering - Airport hubs and frequency of delays, Late arrivals and airports, Weather features

Lasso on additional combinations of features

Test on ensemble models based on classification trees