PREDICT FLIGHT DEPARTURE DELAY

Po-Yan Tsang

Delayed aircraft are estimated to have cost the airlines several billion dollars in additional expense.

-Airlines for America

(http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/)

PROJECT GOALS

- Predict flight delay time using linear regression based on 2017 United States flights data
- Understand factors that impacts flight delay time

METHODOLOGY

DATA SOURCES

- Bureau of Transportation Statistics
  - 2017 Flight delay and cancellation data
  - 2015 airport volume data scraped from website
- Iowa State University website
  - Historic airport weather data from

TOOLS

- Data analysis and visualization: *pandas, numpy, statsmodel, matplotlib, seaborn*
- Web scraping: *beautifulsoup*
- Web data gathering: *requests*

WORKFLOW

2017 Flight Data
(5.5 mil rows)

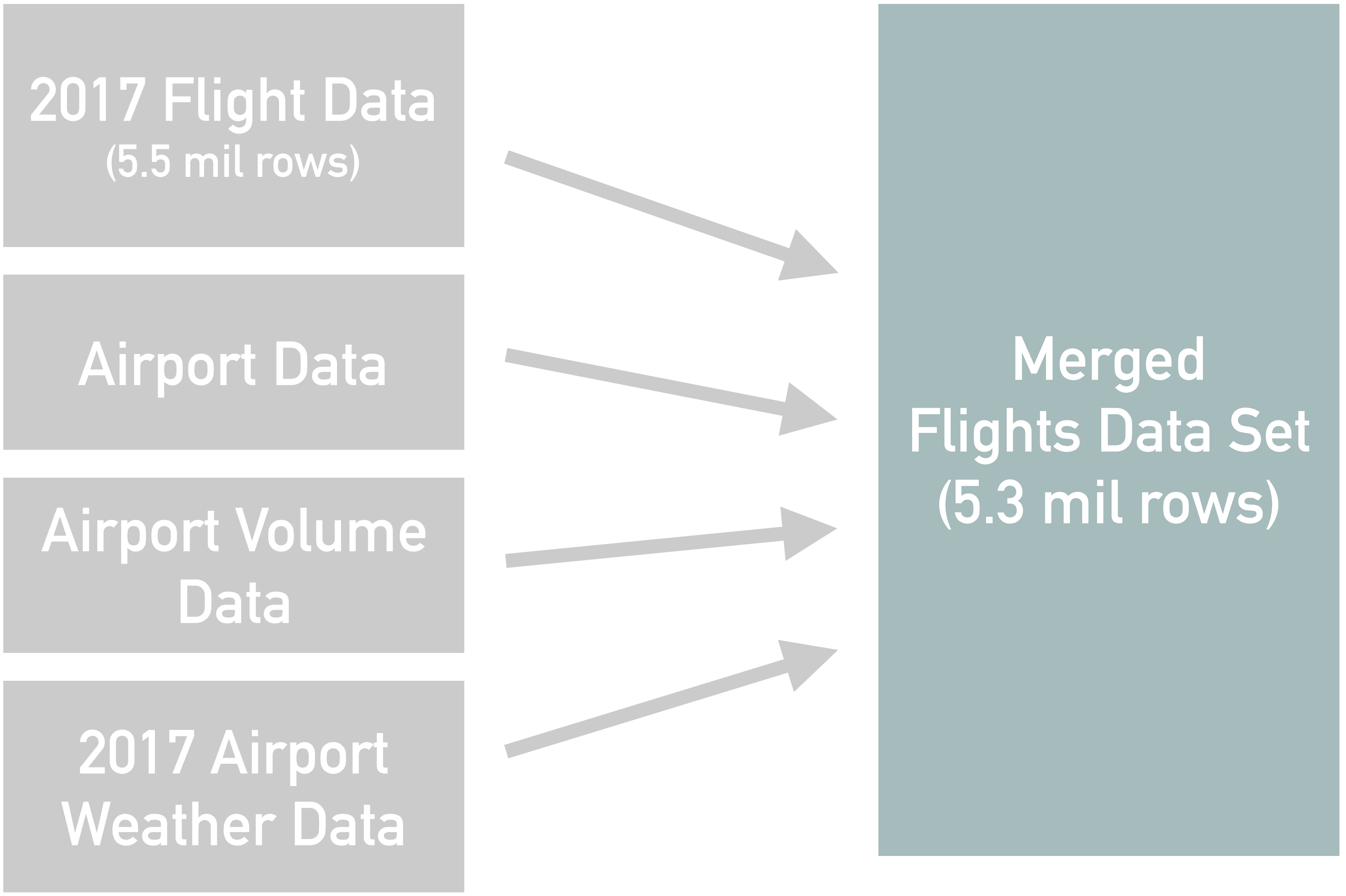
Airport Data

Airport Volume Data

2017 Airport Weather Data

- Sampled from hourly to 3-hour interval weather data
- Computed new features:
  - Inbound delay
  - Turnaround time

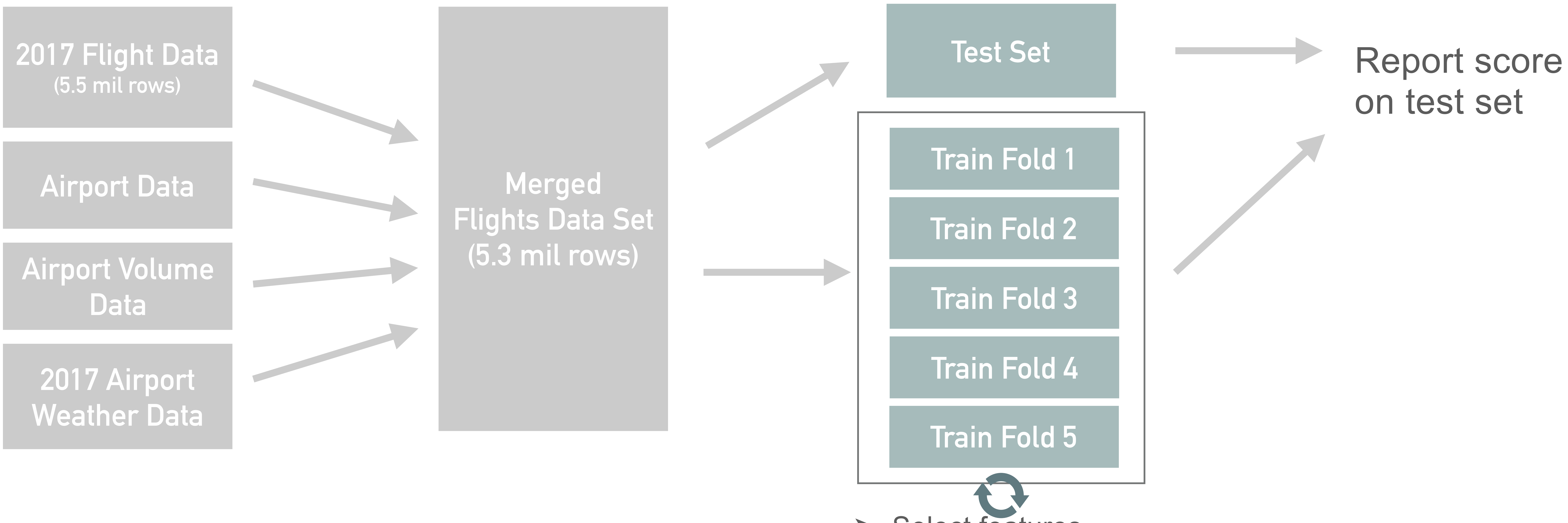
WORKFLOW



- Sampled from hourly to 3-hour interval weather data
- Computed new features:
  - Inbound delay
  - Turnaround time

- Dropped rows with n/a weather data
- Randomly sampled to smaller data set (50K rows)

WORKFLOW



- Sampled from hourly to 3-hour interval weather data
- Computed new features:
  - Inbound delay
  - Turnaround time

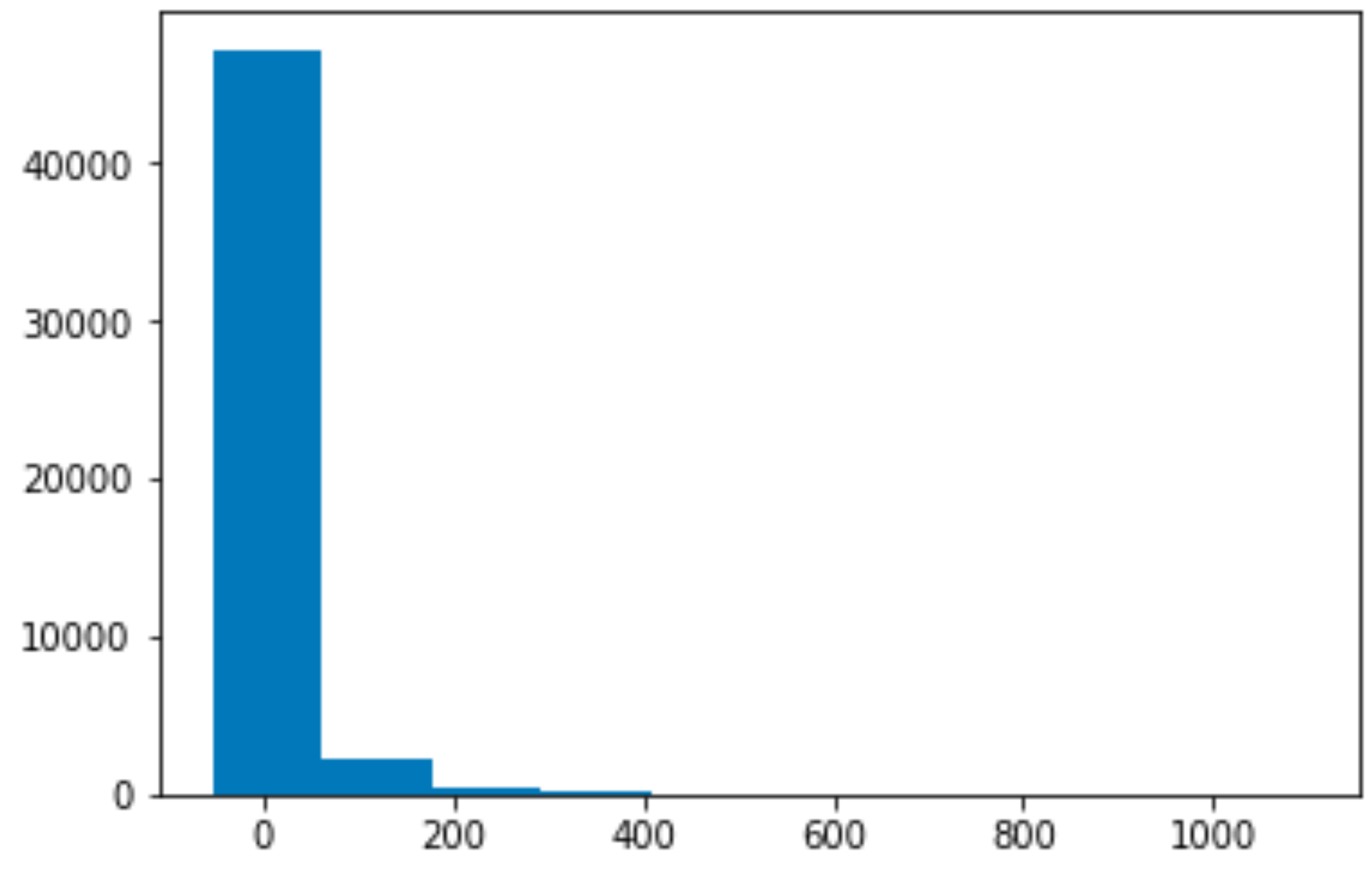
- Dropped rows with n/a weather data
- Randomly sampled to smaller data set (50K rows)

- Select features
- Perform transforms
- Train on models:
  - LassoCV
  - RidgeCV
  - ElasticNetCV

MODELING

- Departure Delay data is highly skewed and needed to be log transformed
- Adding features gave glimpse into what could be predictive (e.g. airlines), but often causes overfitting elsewhere

Histogram of Departure Delay (min)



RESULTS

	Lasso	Ridge	ElasticNet
R2	0.232	0.234	0.233
Alpha	0.01	81	0.01
Coefficients			
Inbound Delay	0.136	0.145	0.140
Departure time	0.059	0.068	0.063
Precipitation	0.006	0.016	0.011

Low R2 value indicates better predictive features needed!

FUTURE WORK

- Look for features to account for “busy-ness” of an airport, or air traffic conditions, for example:
  - # of hourly traffic at airport around scheduled departure time
  - Public holidays
- Further breakdown of how specific airline carrier can impact delay

Q & A

APPENDIX

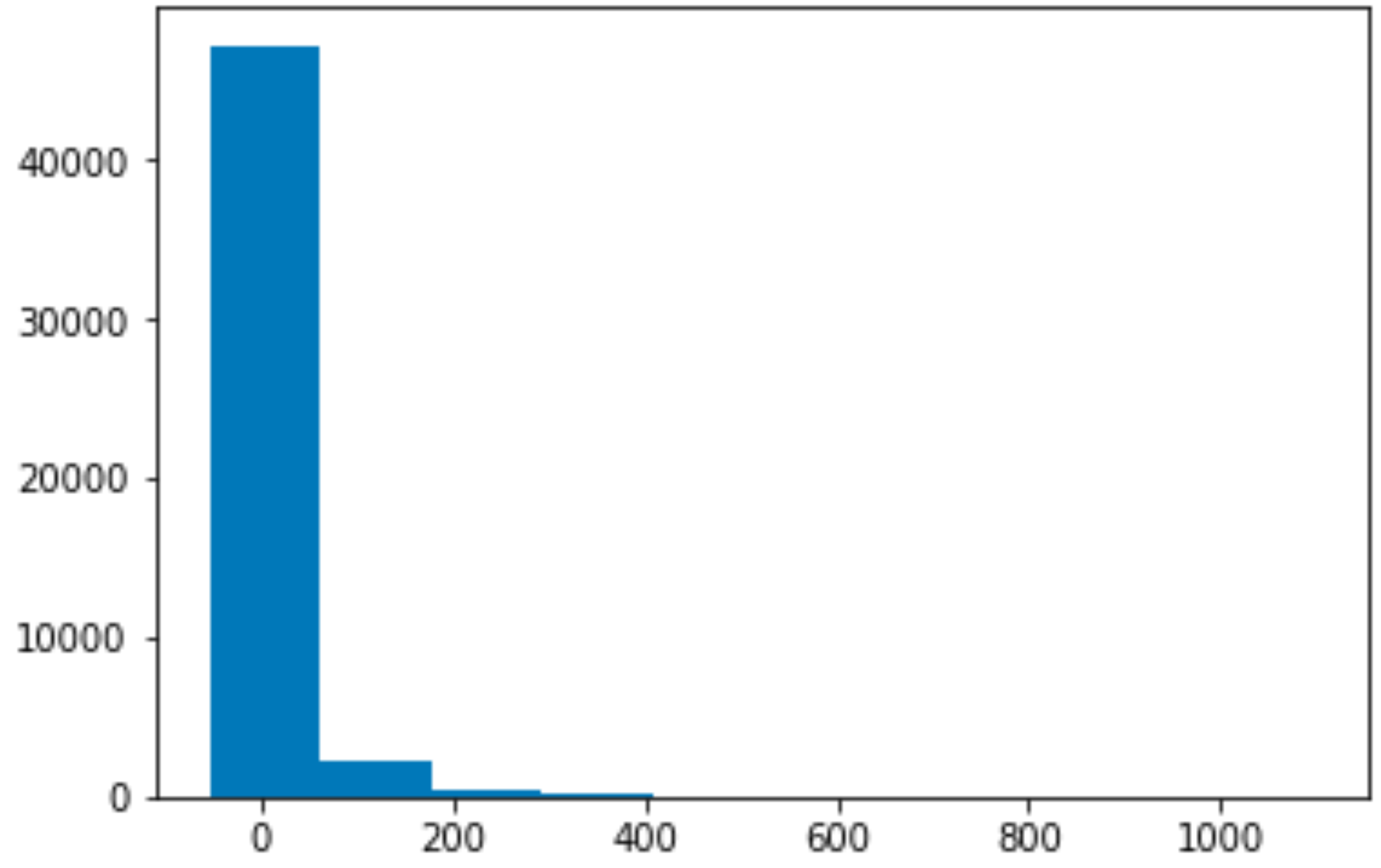
FEATURES CONSIDERED

1. Inbound plane delay*
2. Departure time*
3. Precipitation*
4. Month
5. Day of month
6. Airline carrier (as dummy variables)
7. Origin airport departure volume
8. Plane turnaround time from last flight
9. Temperature (F)
10. Wind speed

** Features chosen in model*

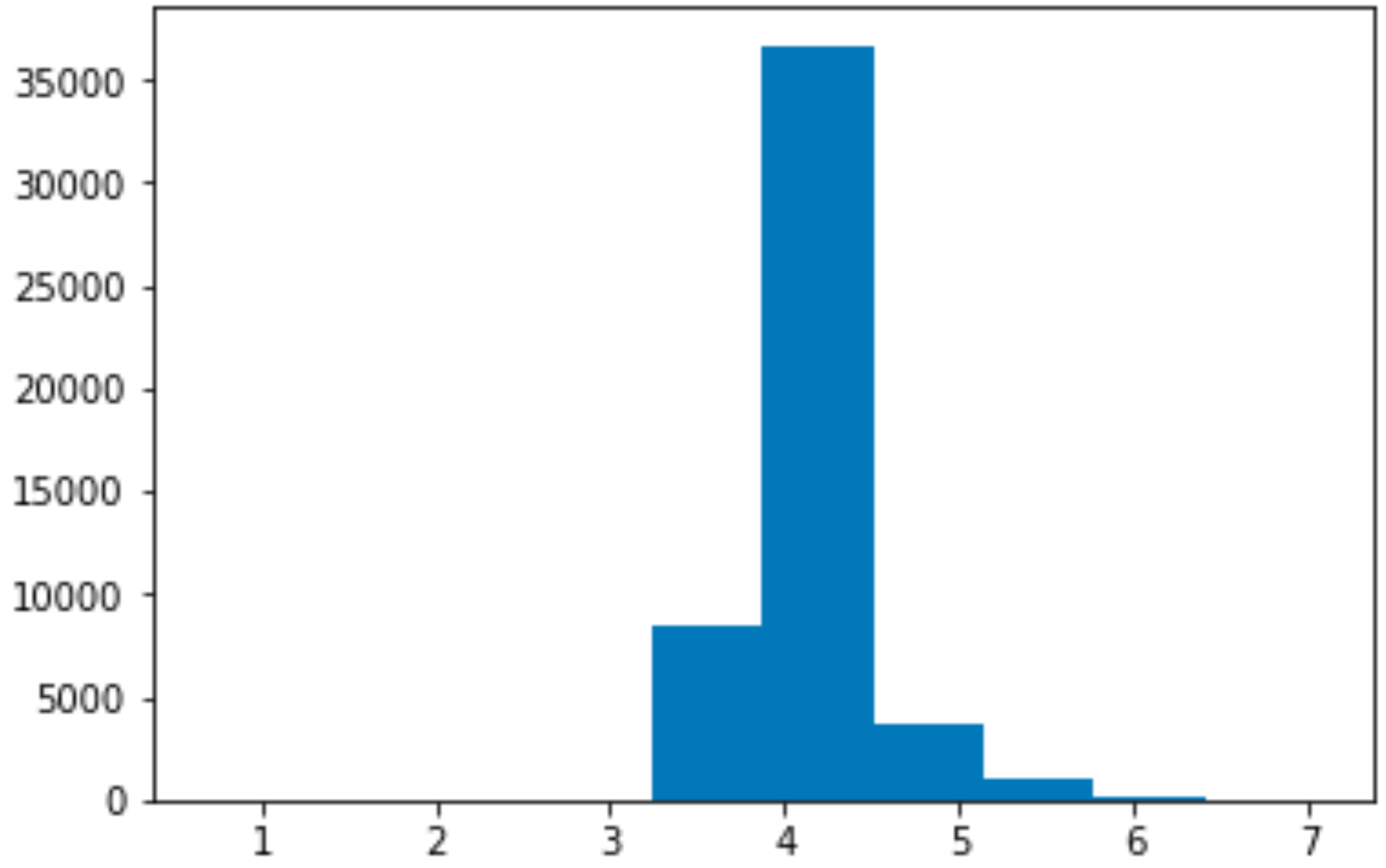
LOG TRANSFORM ON DEPARTURE DELAY (Y)

Histogram of Departure Delay (min)



Log Transform

Histogram of Log of Departure Delay (min)



CHOSEN MODEL (RIDGE) RESIDUAL PLOTS