Flight Delay Model

By Will Bishop
Springboard Data Science Intensive
February 2017

Question and data

- Part 1

Flight delay model goal

- To predict the on–time performance of a flight based on:
  - Origin and destination cities
  - Airline
  - Time of year, time of day, day of week
  - Other variables if possible

Flight delay model data

- Bureau of Transportation Statistics On-Time Performance Table
  - Includes every flight of major U.S. airlines since 1990
  - Used 5% random sample due to computer memory limitations
- Added BTS crosswalks for airline names and market names

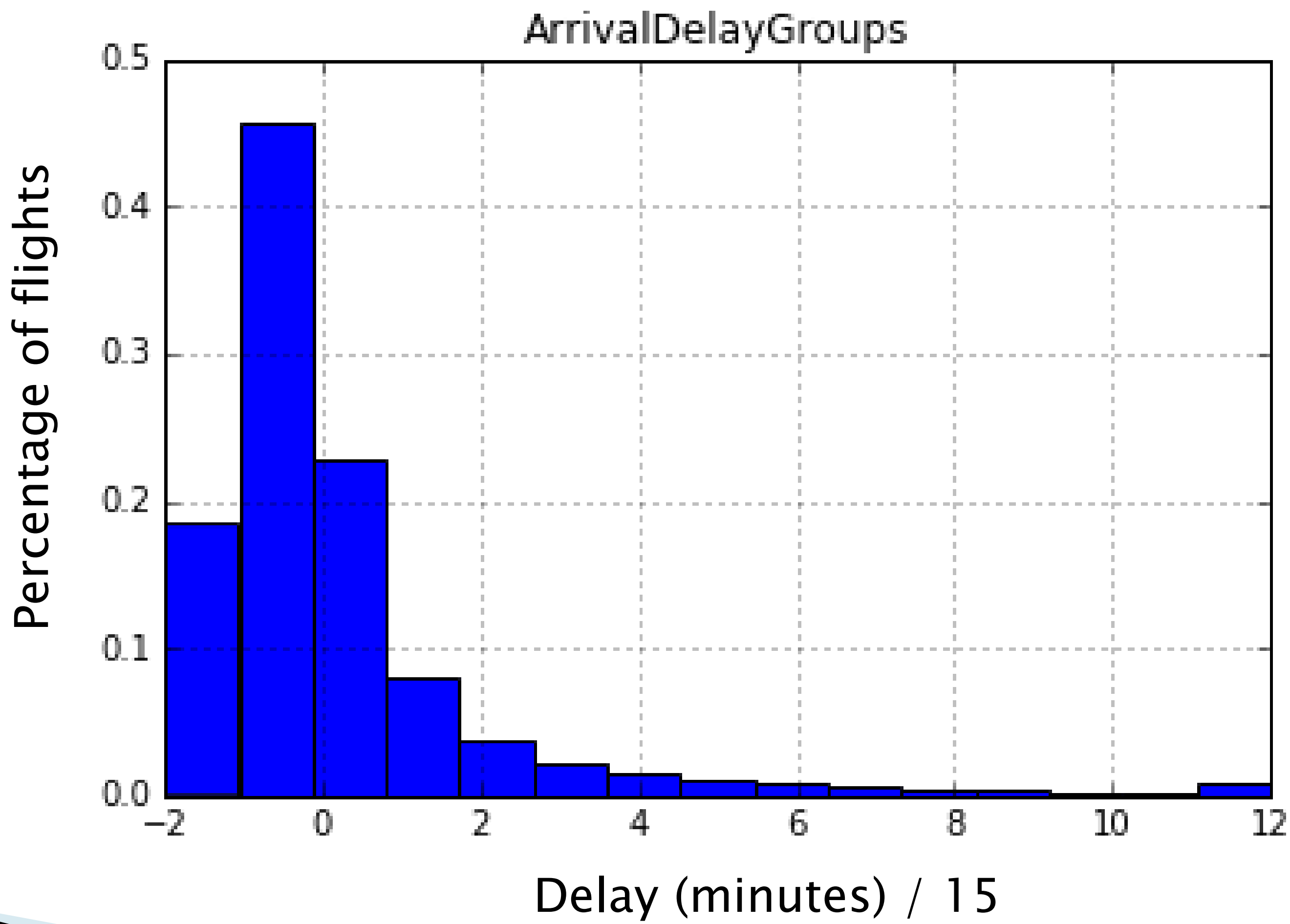
Client and usefulness

- The ideal client is an application developer that helps travelers decide on the best flights
  - Google Flights
  - Kayak
  - Expedia
- With a flight delay model, they can incorporate likely delay times as well as price

Part 2

Descriptive statistics

Delay times in 15-minute groups



Airlines by percentage of flights

Southwest	20%
Delta	13%
ExpressJet	11%
SkyWest	10%
American	10%
United	8%
Others	27%

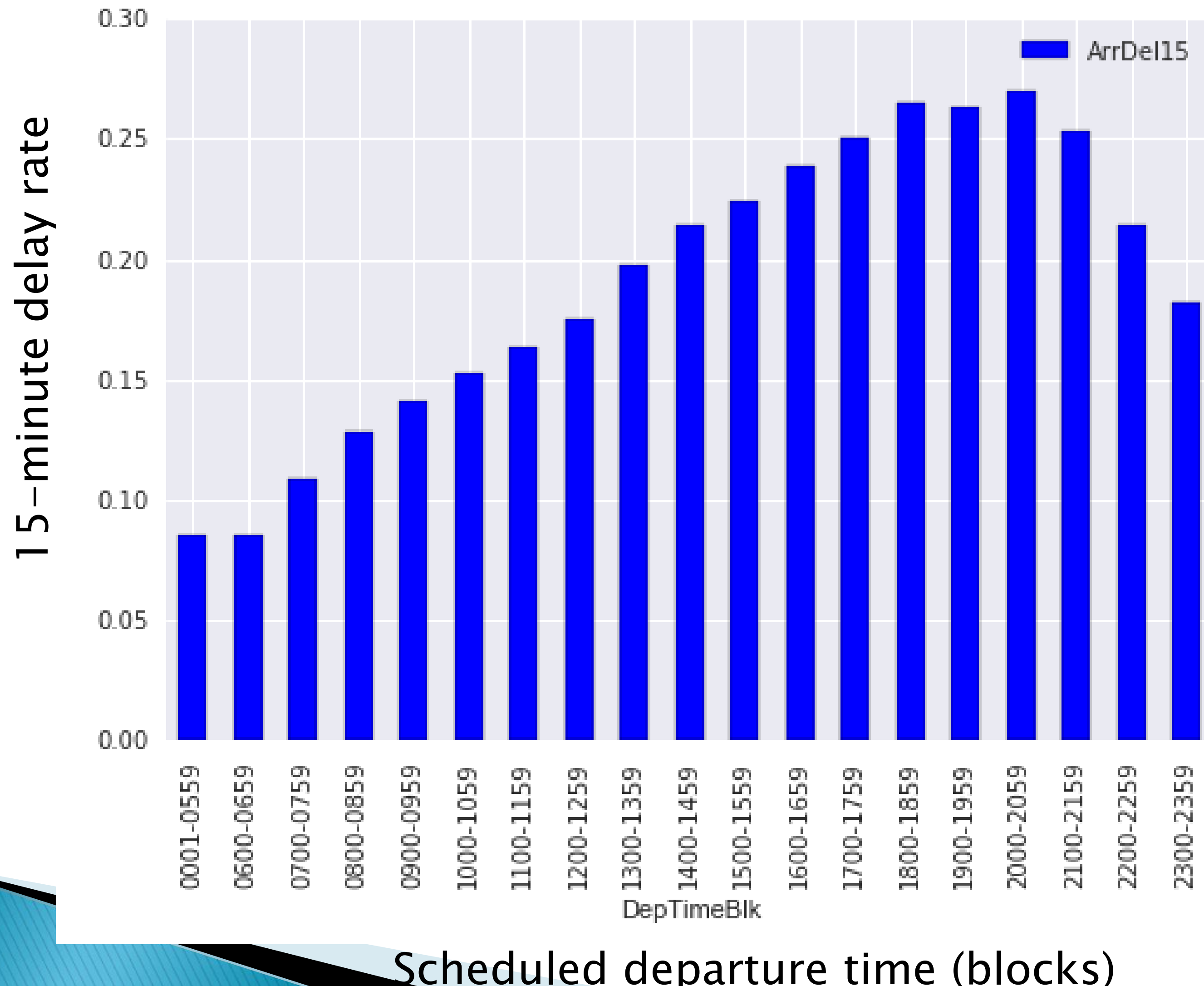
Airlines by 15-minute delay rate

Highest delay rates	
Spirit	30%
Frontier	24%
JetBlue	22%
ExpressJet	22%
Envoy	22%

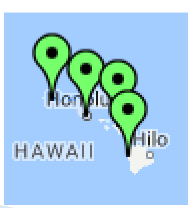
Lowest delay rates	
Hawaiian	8%
Alaska	12%
Delta	14%
AirTran	15%
Mesa	16%

Delay rate across all flights in data set: 19%

Delay rates by time of day

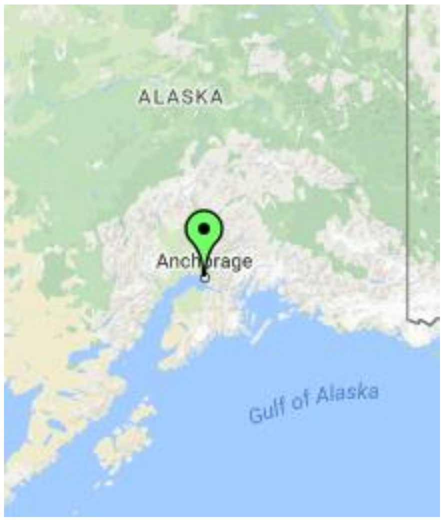


Destination cities by delay rate

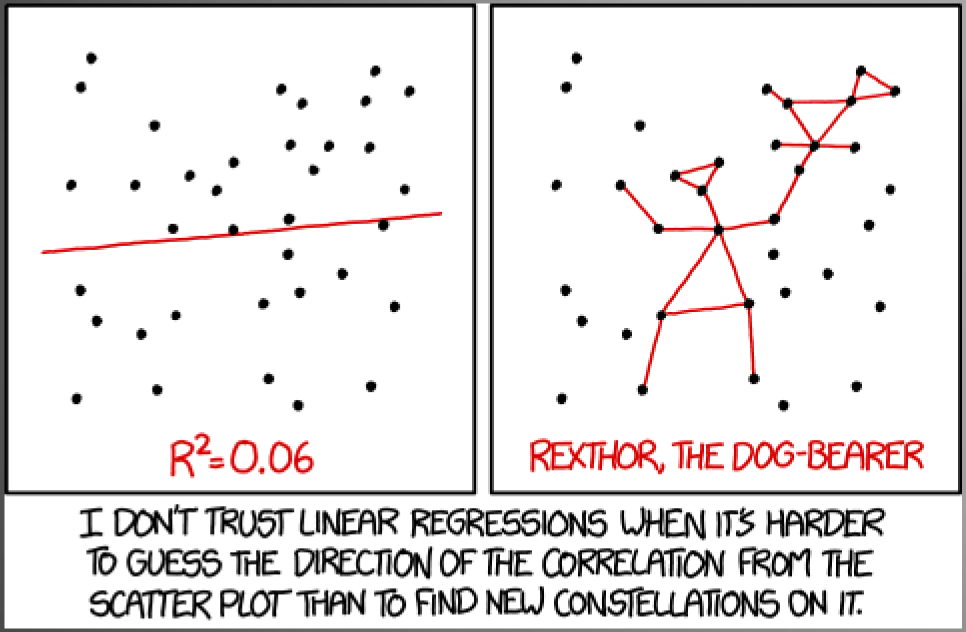


Green: best 10 destination cities
Red: worst 10 destination cities
(At least 2,000 flights in data set)

Origin cities by delay rate



Green: best 10 origin cities
Red: worst 10 origin cities
(At least 2,000 flights in data set)



Part 3

Linear regression model

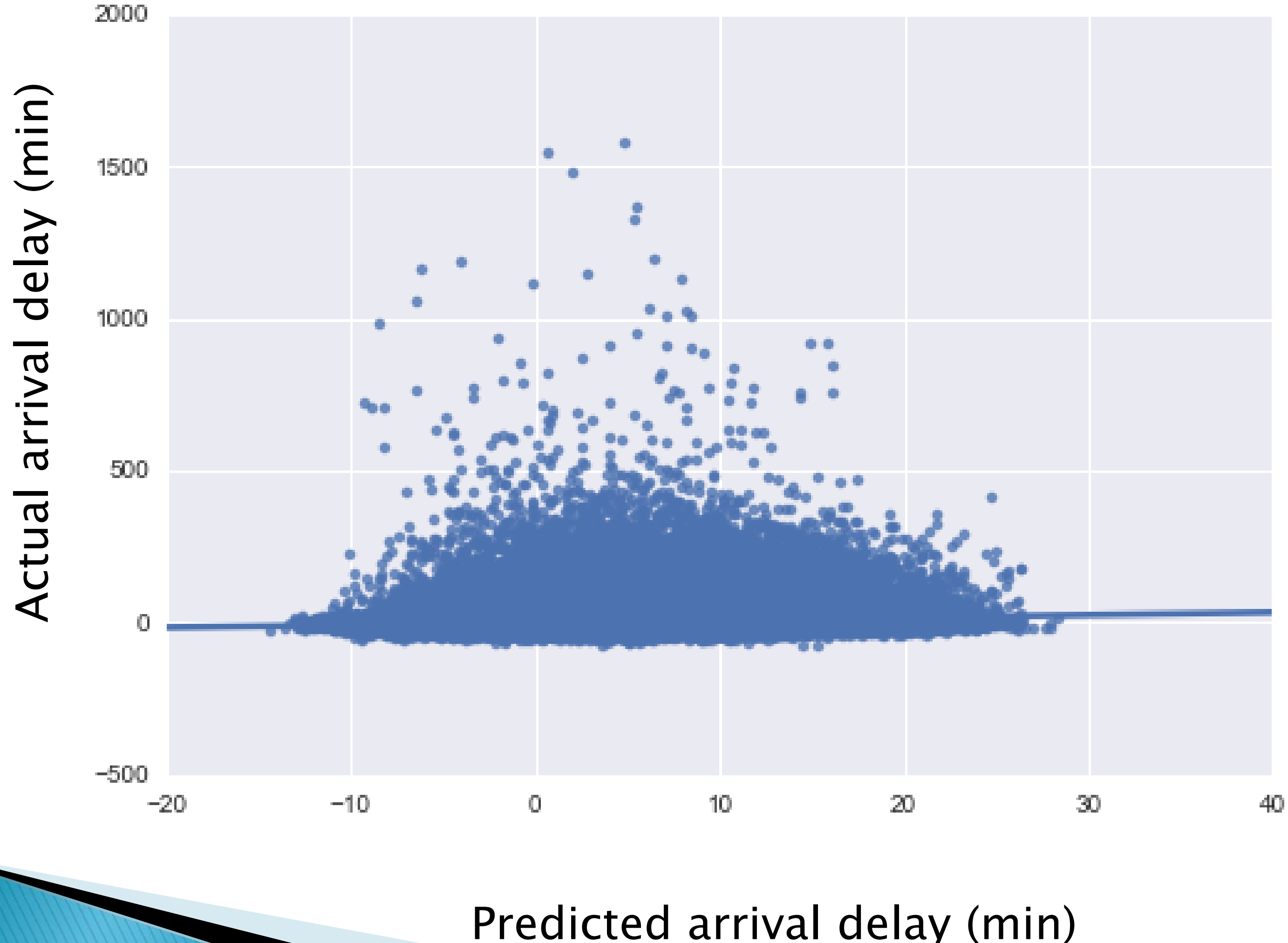
Variables included

- Dependent: Arrival delay time in minutes
- Independent dummy variables:
  - Airline name
  - Origin and destination cities
  - Month
  - Time of day
- Interaction between airline and airport if:
  - Airport has at least 1.5% of flights in data set

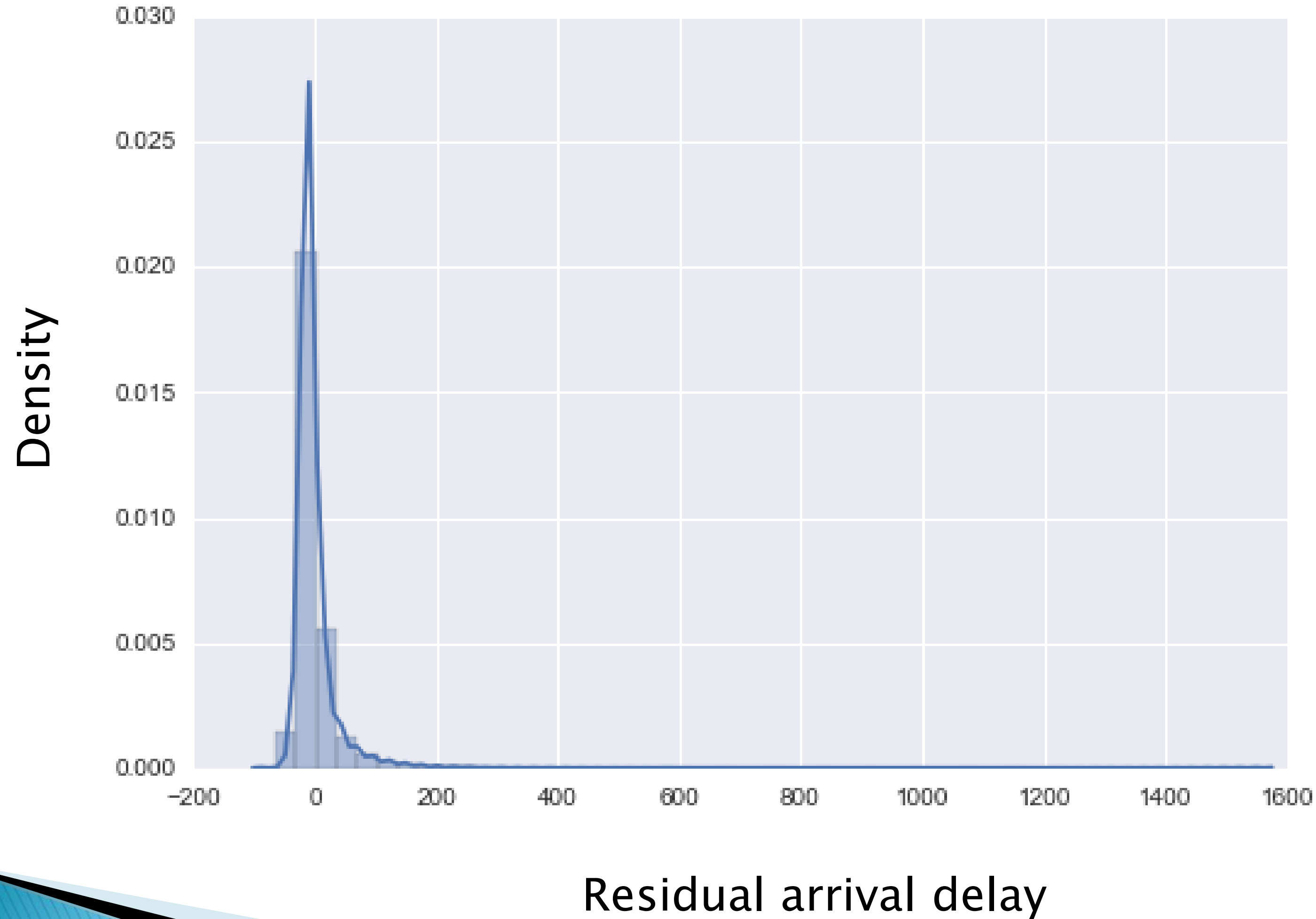
Linear regression technique

- SGDRegressor from scikit-learn
- Used partial_fit to allow large data set
- Broke data into 15 batches of about 100,000 observations each
  - Each batch was 70% training data and 30% test data

Predicted vs. actual values (test data)

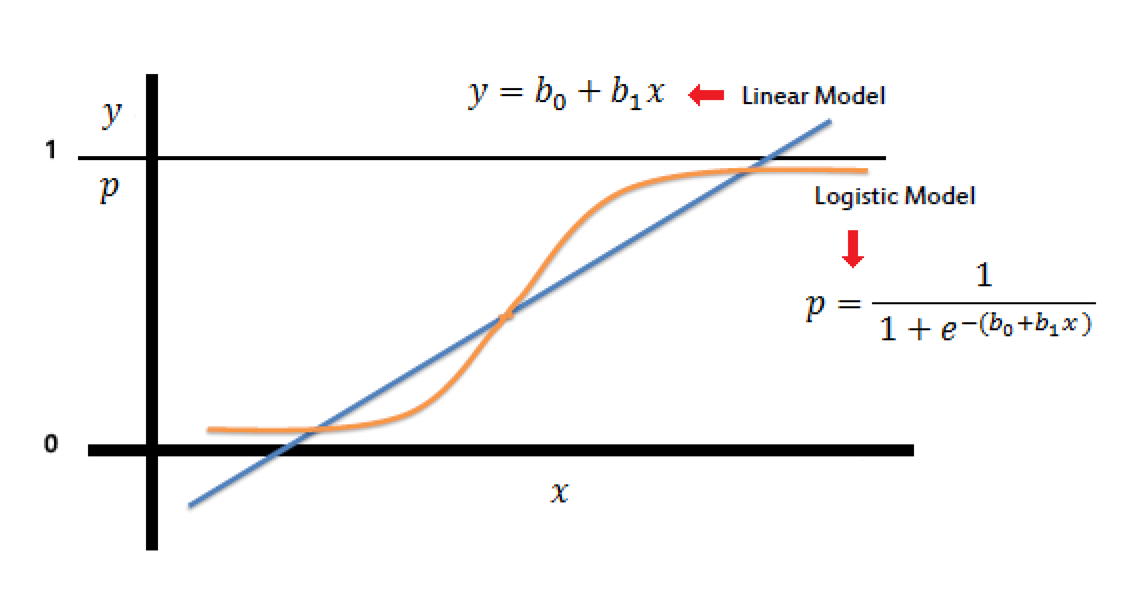


Histogram of residuals (test data)



Linear regression conclusions

- Most variation in flight delays driven by factors outside this model
- Extreme outliers (500+ minutes delay) are hard to predict and have no analogue on the negative side
- However, the model does provide insights beyond descriptive statistics



Part 4

Logistic regression model

Logistic regression technique

- Dependent variable: 15-minute delay binary
- Independent variables: same as linear model
- SGDClassifier from scikit-learn
  - Used loss = log to estimate probability of delay
  - Used L1 penalty to drive more coefficients to 0
- Used partial_fit with 15 batches as in linear model

Logistic regression results

- Simple measures of classifying power look very poor
  - Prediction score: 81% – no better than predicting "0" for every flight
  - F1 score: 0.0002
  - Confusion matrix:

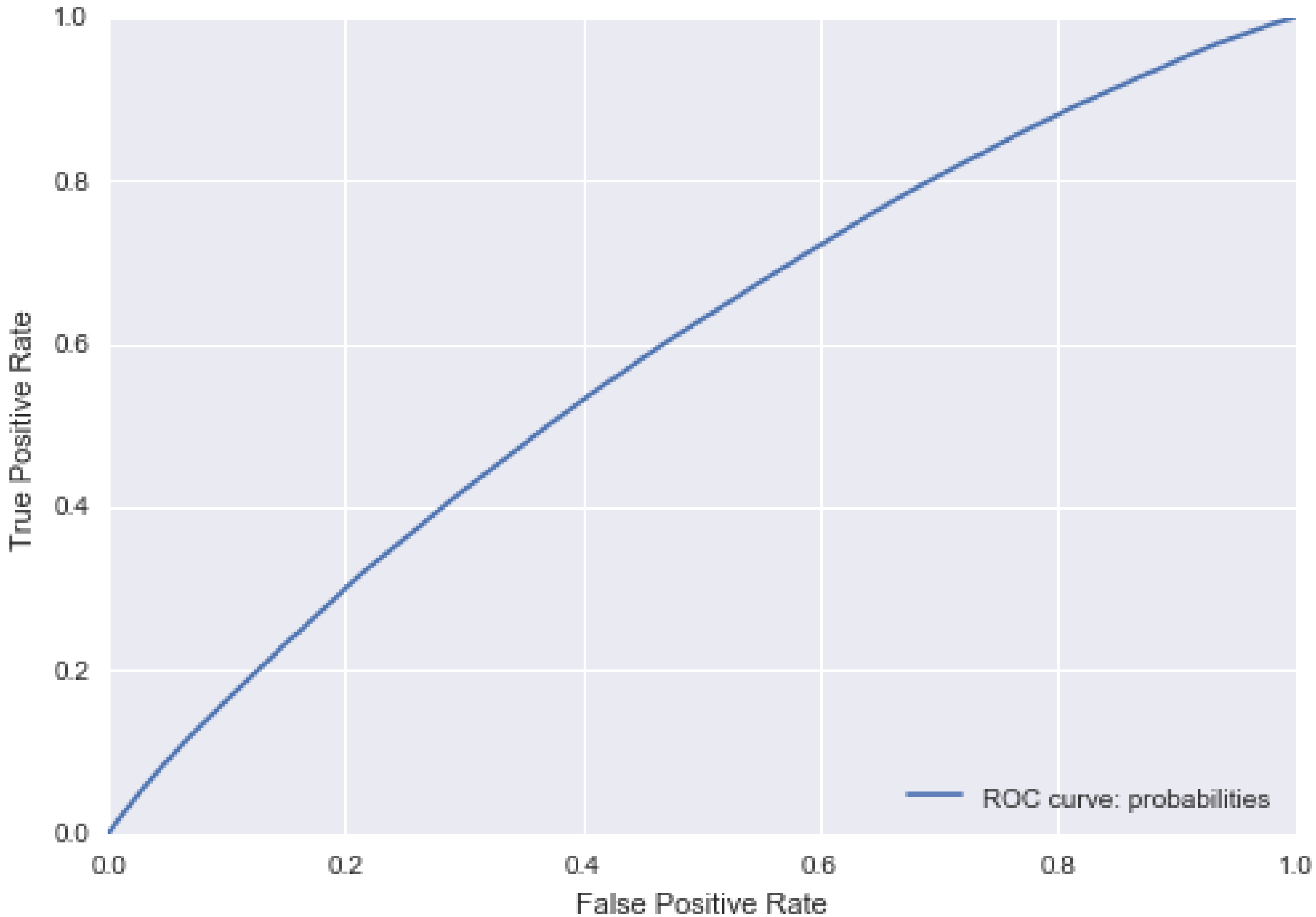
15-minute delay?	Predicted no	Predicted yes
Actual no	297,822	19
Actual yes	68,475	7

Logistic regression results cont.

- Consider different probabilities of delay as cutoff for a “positive” result
- Example: confusion matrix with 20% probability cutoff is

15-minute delay?	Predicted no	Predicted yes
Actual no	174,377	123,064
Actual yes	31,105	37,377

Logistic regression ROC curve



Overall conclusions

- Flight delays are often driven by random events
- Model needs more data and more variables in order to have predictive power
- Our exploration and modeling still provides useful information on airlines, airports, and times of day to avoid flight delays