Predicting Flight Delays

Drew Hibbard

Prepare for the Worst

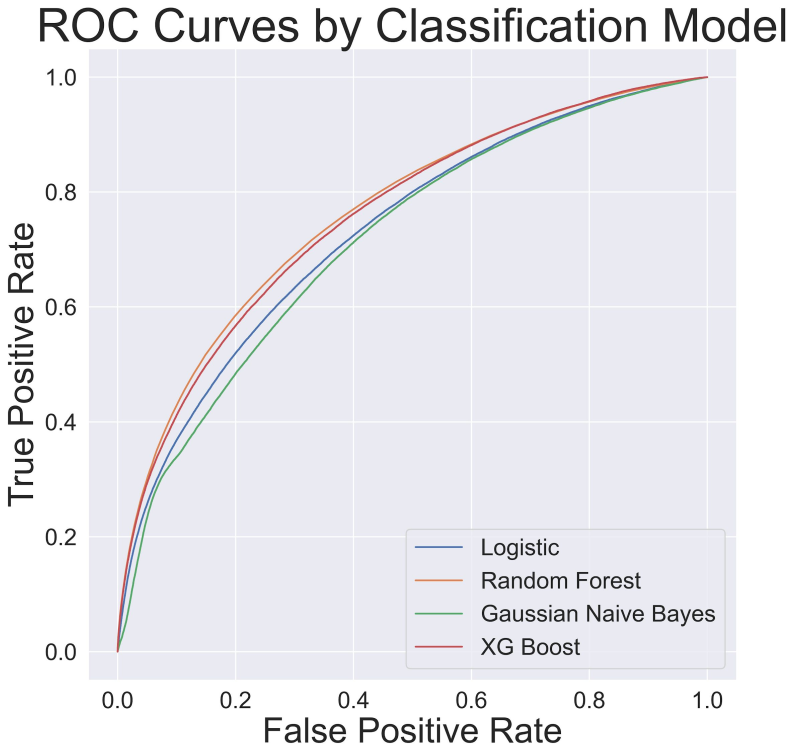
- Flight delays are relatively common
- Understand the factors that affect delays

Obtaining the Data

- Kaggle - all 2015 US flights
- Airfleets.net - aircraft info
- NOAA - weather data
- FlightAware - real-time flight data

Exploratory Data Analysis

Model Testing and Results



Model	AUC	F1 (delays)	F1 (non-delays)
Logistic Regression	0.74	0.43	0.86
Random Forest	0.76	0.49	0.85
Gaussian Naive Bayes	0.71	0.40	0.86
XG Boost	**0.76**	**0.48**	**0.85**

Threshold = 0.25

Recorded Demo

Questions?

Next Steps

- Use more recent data
- Update web app to allow users to change each predictive feature at will

Appendix

Modeling Methods Used

- KNN - too slow
- SVC - too slow
- How to handle imbalanced data?
  - Performed random undersampling
  - Also used balanced class weights
    - Performance between these two methods was roughly equal
-

XG Boost Parameters

- Learning Rate - 0.2
- Max Depth - 8
- Min Child Weight - 1
- Subsample - 1
- Col sample by tree = 0.8

Ensemble Methods

Prediction Correlations	XG Boost	Random Forest	Logistic Regression
XG Boost	1	0.56	0.51
Random Forest	**0.56**	1	0.50
Logistic Regression	**0.51**	**0.50**	1

- Increased AUC to 0.77, but not enough to justify slower speed