Flight data exercise.

Scott Wolf
Date: May 9, 2016

Objective: Can flight delays be predicted accurately? In this report, I evaluate the time, date, and season in which flights occur, other flight-specific variables, and the variance in weather to predict flight delay. To evaluate the effectiveness of various machine learning methods, I also compare how well a series of algorithms perform in answering this question using the same data and predictors and discuss the trade-off between predictive accuracy and scalability.

Population of observations: Direct / non-stop flights that departed from the New York City metro area and arrived in the Los Angeles metro area during the 2015 calendar year are analysed.* Analyses are restricted to the population of **17,855 flights** with complete data across the following set of variables, split randomly between a training data set (80% of observations, 14,289 flights) and the a training data set (20% of observations, 3,566 flights).

- * Note that direct flights from NYC's LaGuardia airport to the LA metro area were not commercially available in 2015 due to FAA flight distance restrictions at LaGuardia.

Predictors / independent variables:

Time at origin.

Source: US Department of Transportation Airline Traffic On-Time data, 2015
http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

- *Hour of scheduled departure* (nominal: 0-23)
- *Day of week* (nominal: 1-7, Monday to Sunday)
- *Month* (nominal: 1-12, January to December)

Flight information.

Source: US Department of Transportation Airline Traffic On-Time data, 2015
http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

- *Origin airport* (nominal: EWR or JFK; note that LGA had no direct flights to LA)
- *Destination airport* (nominal: BUR, LAX, LGB, ONT, or SNA)
- *Carrier* (nominal: American Airlines, JetBlue Airways , Delta Airlines ,United Air Lines, Virgin America)
- *Total number of flights departing from the origin airport during the hour of scheduled departure* (interval: range: 3 to 34 flights)

Weather at origin (NYC) at scheduled departure time.

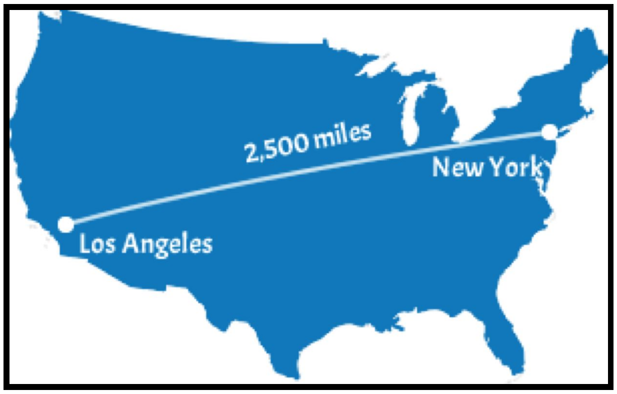
Source: National Oceanic and Atmospheric Administration, Central Park Station, New York City, 2015 hourly data.
https://www.ncdc.noaa.gov/qclcd/QCLCD

- *Visibility* (interval: range 0 to 10 statute miles)
- *Temperature* (interval: range 2 to 96 degrees Fahrenheit)
- *Wind speed* (interval: range 0 to 21 miles per hour)
- *Amount of precipitation during hour of scheduled departure* (interval: range 0" to .32")

Outcome / dependent variable: A delay in arrival of at least 15 minutes (nominal: 1=yes, 2=no)

Github link for code: https://github.com/scottwolf/2015_NYC_to_LA_flight_data

Descriptive statistics (full sample, N = 17,855 flights).

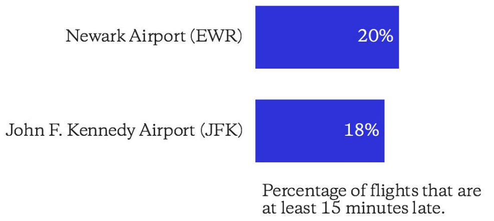


Nearly 1 in 5 flights are late.

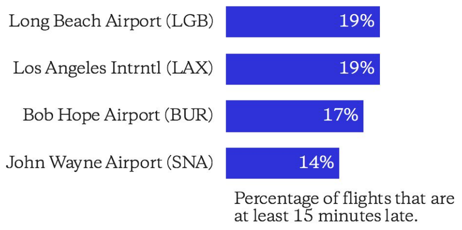
18%

of all LA-bound flights that originate in NYC arrive at least 15 minutes late.

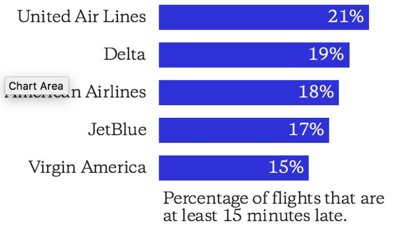
Flights from Newark are somewhat more likely to arrive late.



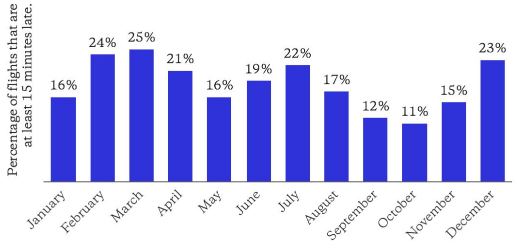
Flights to John Wayne Airport less likely to arrive late.



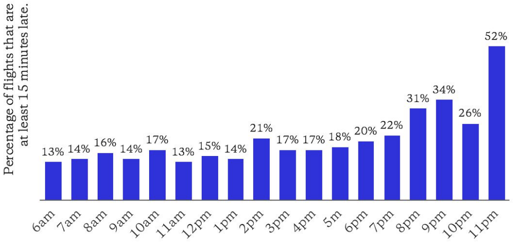
Flights on Virgin America are less likely to arrive late.



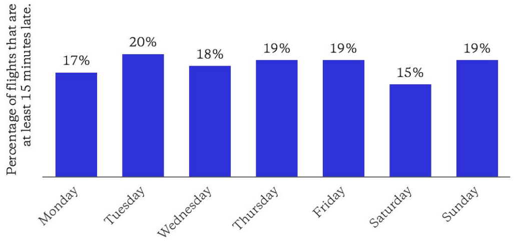
Flights in October & September are less likely to arrive late.



Flights that depart later in the day are more likely to arrive late.



Flights on Saturday are less likely to arrive late.



Comparison of machine learning techniques using the same data and predictors.

	In-sample (training set, n = 14,289)				**Out-of-sample (testing set, n = 3,566)**			
	Accuracy	Specificity	Sensitivity	Kappa	**Accuracy**	**Specificity**	**Sensitivity**	**Kappa**
Baseline model.	81.3%	100%	0%	.00	**80.7%**	**100%**	**0%**	**.00**
Logistic multiple regression.	81.8%	98.7%	8.3%	.11	81.0%	98.7%	7.4%	.09
Boosting: ADA.	82.1%	82.5%	65.6%	.12	81.2%	81.7%	60.5%	.10
Boosting: GBM.	82.7%	83.0%	73.2%	.16	81.5%	82.1%	64.8%	.13
Naive Bayes.	80.8%	95.8%	15.5%	.15	80.1%	95.4%	16.1%	.15
Random forest.	96.0%	95.4%	99.7%	.86	**83.1%**	**84.3%**	**67.6%**	**.28**

The **Random Forest** technique provides the highest predictive lift when all independent variables (see Page 1) are used to predict late flights (i.e., those that arrived at least 15 minute late).

- *Kappa:* best lift on predictive power over a baseline model where all flights are predicted to be “on time.”
- *Accuracy:* 83.1% overall correct classification, the highest of all techniques.
- *Specificity:* 84.3% of “late” flights are correctly identified, an average percentage.
- *Sensitivity:* 67.6 of “on time” flights that are correctly identified, the highest of all techniques.

Advantages of Random Forest: Its superior predictive power is accomplished by creating multiple random decision trees from multiple samples and variable combinations. Conceptually, the “average” parameters for the group of decision trees are computed, also avoiding overfitting.

Disadvantages of Random Forest: The high degree of out-of-sample accuracy requires far more computing power and time. As such, it is often difficult to scale this process with larger datasets or those which need to be processed quickly.

Conclusion: Although the predictive models in this investigation did not provide an exceptionally impressive lift over the baseline model, we are missing a more refined set of highly correlated independent variables. Ideally, I would also make significant modifications to the set of predictors informed by the diagnostic testing of the models (in order to make comparisons across models, they are unmodified in this analysis). More time allotted to this task would enable these explorations. Nonetheless, I compared a series of common machine learning algorithms using the same predictors and outcome variable to determine which was able to achieve the highest degree of predictive lift over the baseline model. While the Random Forest model was the optimal solution for this set of variables and dataset, this will not always be the case. Multiple techniques should be explored for any investigation, with special attention paid to the need for accuracy vs. the need for scalability.