Predicting flight delays based on city, airline, and other factors

By Will Bishop
Springboard Data Science Intensive
February 2017

Introduction

As airborne travel becomes ever more accessible, both in the U.S. and across the world, many companies have stepped in to attempt to help travelers' budget and quality of life by helping them select the best flight options for traveling between given locations. The pioneers are Kayak[1] and Expedia[2], with Google's Flights application following their model.[3] These applications generally involve listing flight itineraries by price, duration, and time of day, helping travelers select the one most amenable. They allow travelers to compare flights across multiple airlines and, if they are flying to or from a larger metropolitan area, multiple airports. These services save the necessity of checking multiple airline websites and show itineraries that the airlines may not choose to display.

While price and duration are the most important factors in selecting a flight itinerary, there are other considerations: for instance, comfort, safety, and the probability of a significant delay. Google Flights has made an attempt to warn customers of the latter by flagging certain flights as "often delayed by 30+ minutes". The purpose of this project is to explore whether we can create a more robust model for predicting delays, which would allow apps such as Google Flights to incorporate those probabilities into their rankings of the best flights. A difference in expected delay time may offset some of the difference in price depending on the traveler's priorities.

Data

The data come from the Bureau of Transportation Statistics' On-Time Performance table.[4] They have data on every flight of major U.S. airlines in monthly tables. The latest available data was from May 2016 at the time I started the project, and the data seem to go back to at least 1990. However, using all the data from 1990 onward would not be very useful because the airline industry and characteristics of airports have changed so dramatically since 1990. Instead I am using a five-year sample from June 2011 to May 2016. From this dataset, I drew a 5% random sample.

Variables included in the dataset

The dependent variables are various measures of arrival delay, the most basic being simply the number of minutes between the scheduled and actual arrival times, as well as a dummy variable simply indicating whether the flight was delayed by 15 minutes or more. There are also five variables that

[1] http://www.kayak.com
[2] http://www.expedia.com
[3] http://www.google.com/flights/
[4] http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

describe types of delay: carrier delay, weather delay, National Air System delay, security delay, and late aircraft delay (meaning that the previous flight using the same aircraft arrived late). However, these variables for the separate types of delay are missing for a large number of observations, so I have focused on the total delay time and the 15-minute dummy variable.

The independent variables in the data set that I identified as likely to hold predictive power are:

- Origin and destination airports
- Airline
- Time of day, week, and year

These variables became the basis for my regression models.

Limitations of the dataset

The data does not include the type of aircraft, number of seats, or number of passengers on the plane. I thought it would be interesting to see whether full and empty flights had different delay rates. The load factor data does not seem essential for the goal of the project, since the purpose of an app like Google Flights is to help passengers pick flights well in advance. A regression model based on the variables mentioned above would presumably account for the probability of a full flight and any effect that would have on delay time. However, the type of plane is often known in advance, so it could theoretically be a factor worth modeling.

The data also does not include weather data, so we cannot answer questions about micro-level effects of specific weather patterns. Again, this variation should be encoded in a model that includes origin and destination cities, with the possible exception of flights booked on short enough notice that it is possible to predict weather beyond seasonal patterns.

Preparation

For the most part, the data came in a clean format (though very large), but a few data cleaning steps were necessary. I merged in a dataset that contained airline names since the original files had only IATA codes for airlines, but BTS provides a separate dataset to join codes and names.[5] I also merged in a dataset that included the names of "markets" rather than cities, which allows analysis by metropolitan area as well as individual airports. I created variables that count the number of departures by city and market, and the number of arrivals by city and market, for descriptive statistics purposes. Finally, I grouped data by month in order to create a time-series plot and see whether an overall time trend in delay times exists.

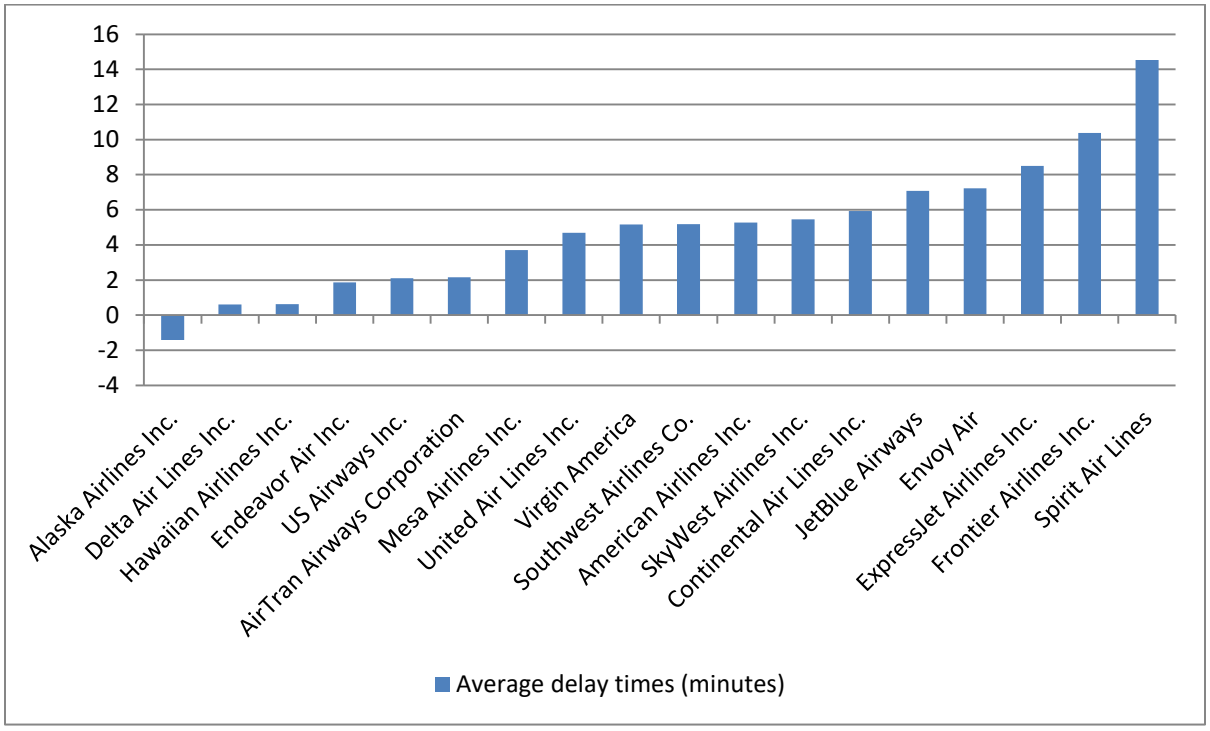
Preliminary Exploration and Findings

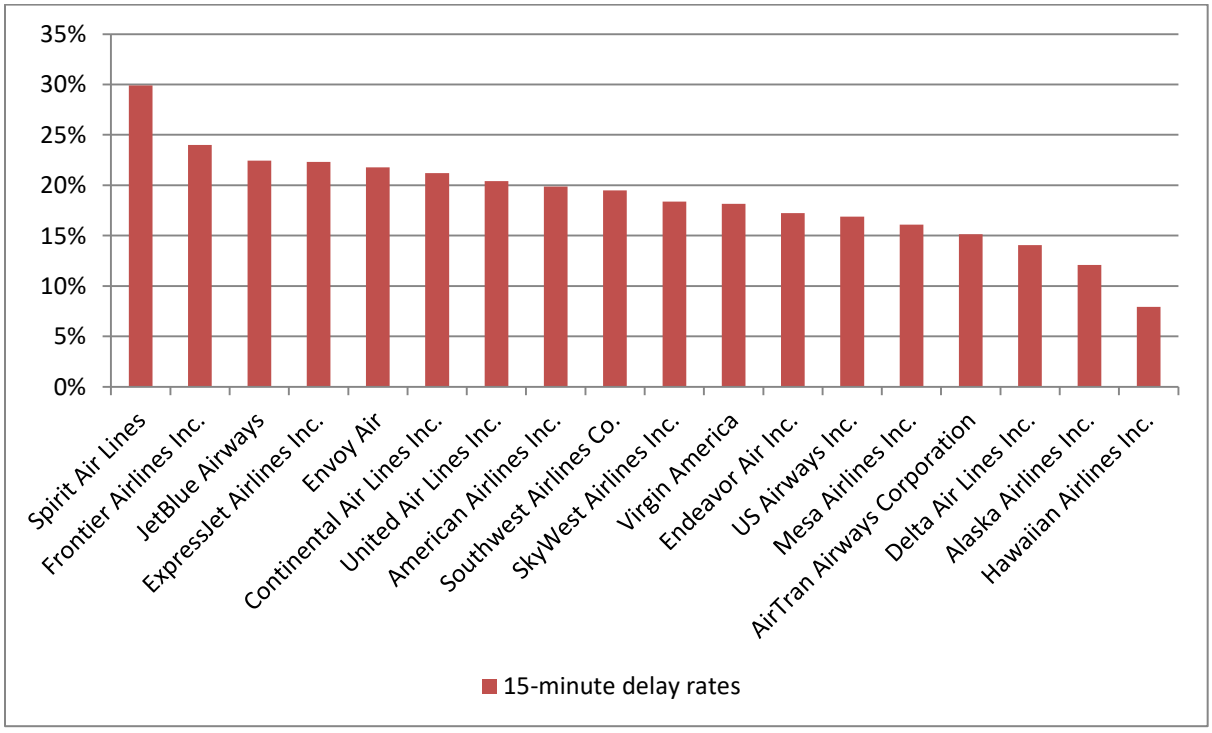
Initial exploration of the data revealed many interesting findings.

Airline name is a very strong predictor of delay. Spirit Air Lines is the least timely airline by a wide margin, both in terms of 15-minute delay rate (30%) and average delay time (14 minutes). Frontier

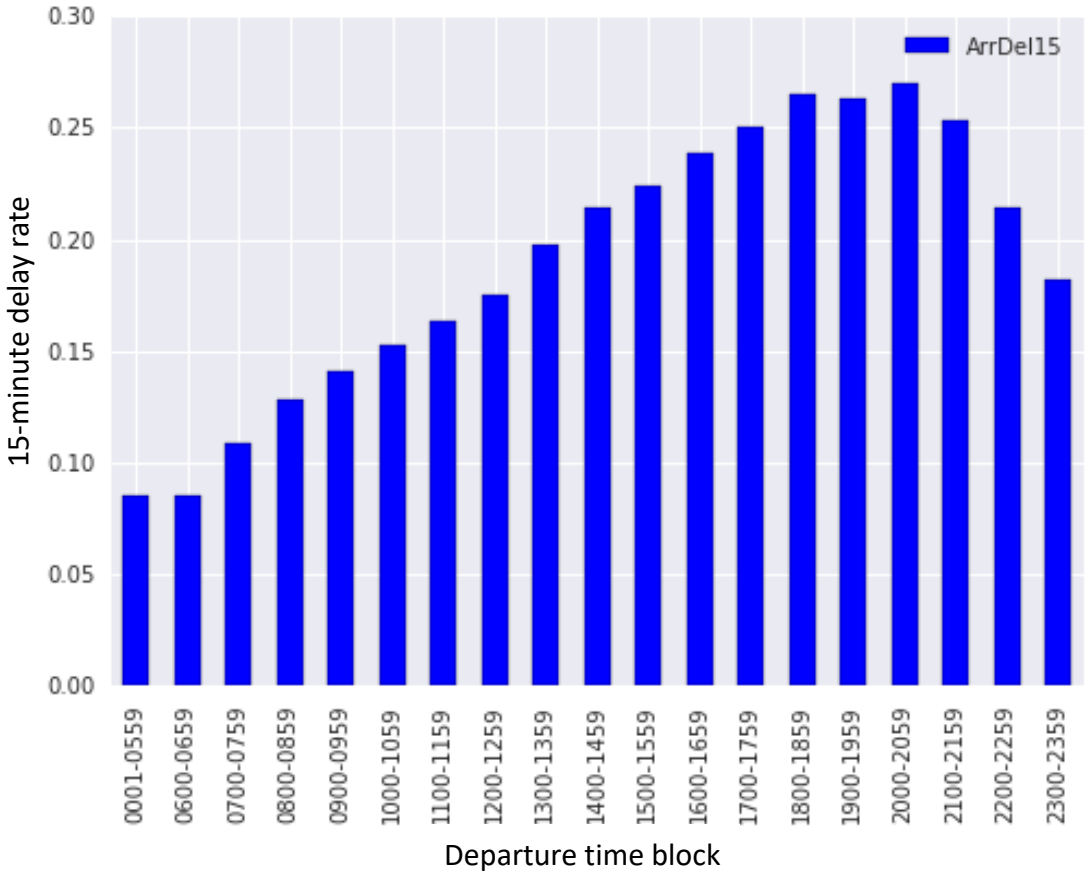
[5] IATA refers to International Air Transport Associations, which assigns two-letter codes to airlines (e.g. AS for Alaska) and three-letter codes to airports (e.g. ANC for Anchorage).

Airlines is second by both measures. On the other end, Alaska, Hawaiian, and Delta rank as the three best airlines by both measures; Hawaiian has the lowest delay rate (8%) while Alaska has the lowest average delay time (-1.4 minutes, i.e., flights arrive *before* their scheduled arrival on average). But we have to be a bit careful here: is Hawaiian Airlines stellar, or is there something else about flying to Hawaii that makes flights tend to be on time?





Time of day is also a strong predictor, with flights that depart in the afternoon and evening 2-3 times as likely to be delayed as those in the early morning, as shown in the following chart.



Arrival cities have widely varying delay rates. Among cities with more than 2,000 flights in the database, 15-minute delay rates range from 10% to 26% and average delay times range from -0.2 to 12.3 minutes. Departure cities also vary widely: 15-minute delay rates from 9% to 25% and average delay times from -0.9 to 9.2 minutes. Four of the ten cities with the lowest delay rates are in Hawaii.

10 destination cities with highest delay rates*		**10 destination cities with lowest delay rates***	
Knoxville, TN	26.2%	Lihue, HI	10.0%
San Francisco, CA	26.0%	Kona, HI	11.2%
Newark, NJ	25.6%	Kahului, HI	11.3%
Colorado Springs, CO	23.2%	Salt Lake City, UT	12.3%
Tulsa, OK	22.9%	Long Beach, CA	13.2%
Little Rock, AR	22.9%	Honolulu, HI	13.8%
Baton Rouge, LA	22.8%	Santa Ana, CA	14.6%
Fayetteville, AR	22.8%	Phoenix, AZ	14.8%
New York, NY	22.7%	Minneapolis, MN	15.0%
Oklahoma City, OK	22.6%	Burbank, CA	15.2%

*Among cities with at least 2,000 flights in the dataset

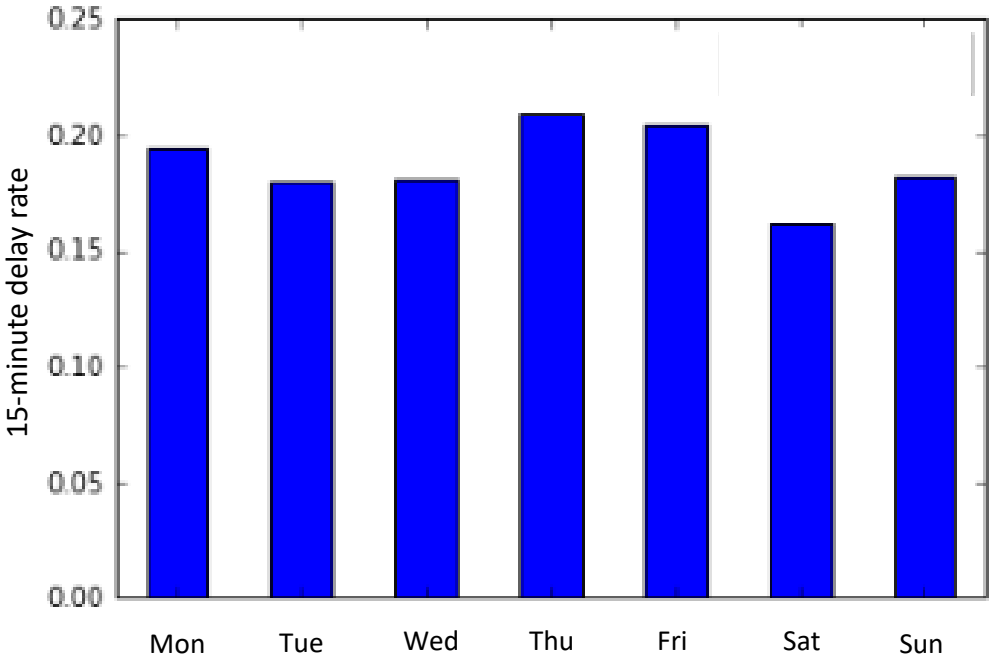
I did the same analysis for origin cities and found similar widely varying results, with some of the best and worst cities being the same (notably, flights both to and from Hawaii tend to be on time) while others are different.

10 origin cities with highest delay rates*		**10 origin cities with lowest delay rates***	
Chicago, IL	24.6%	Kona, HI	9.3%
Newark, NJ	24.3%	Lihue, HI	9.4%
San Francisco, CA	22.6%	Kahului, HI	9.6%
Dallas/Fort Worth, TX	22.6%	Spokane, WA	10.5%
White Plains, NY	22.2%	Honolulu, HI	10.7%
Denver, CO	22.1%	Anchorage, AK	11.2%
Miami, FL	22.1%	Salt Lake City, UT	12.3%
New York, NY	21.8%	Portland, OR	12.8%
Fayetteville, AR	21.6%	Burbank, CA	12.9%
Baltimore, MD	21.6%	Boise, ID	14.2%

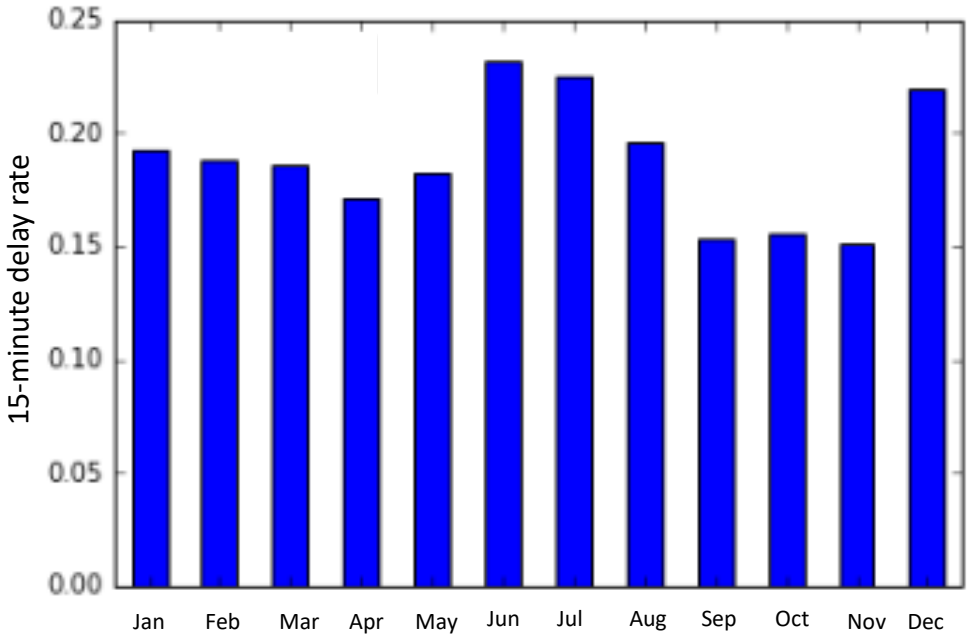
I repeated this analysis for "markets" (roughly corresponding to metropolitan areas) rather than cities. However, since different cities or even different airports within a city can have significantly different delay rates, I decided to use the airport code in the modeling, rather than city or metropolitan area.[6] Since there are so many different cities, a comprehensive picture of their effect on delay rates cannot be achieved by listing individual cities—we need a regression model.

Day of week does not appear to be a strong predictor, as the following chart shows (Note: on the horizontal axis, 1 is Monday, 2 is Tuesday, etc.).

[6] For example, note that San Francisco is second in the "worst destination cities" list while Oakland and San Jose do not appear. Three cities in the Los Angeles metropolitan area are in the "best destination cities" list, but Los Angeles itself is not.

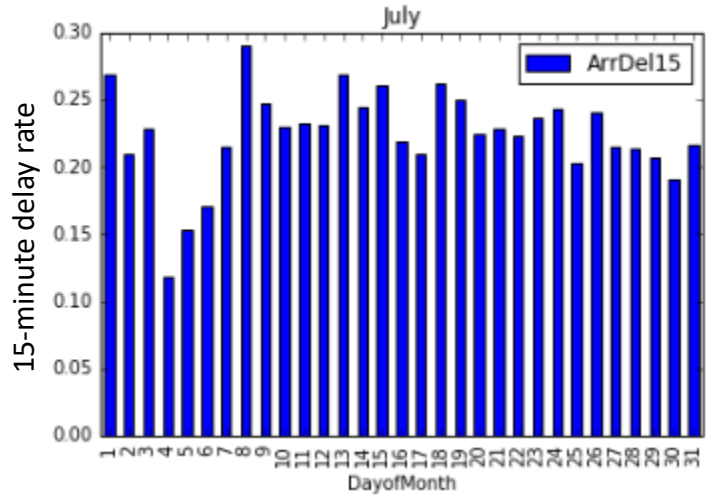


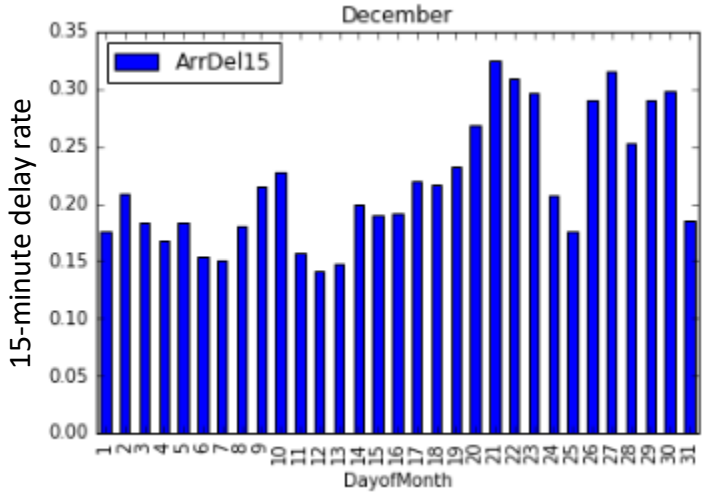
June, July, and December are the months with the highest delay rates, while September, October, and November have the lowest. The differences between months are much smaller than those between airlines, times of day, or cities.



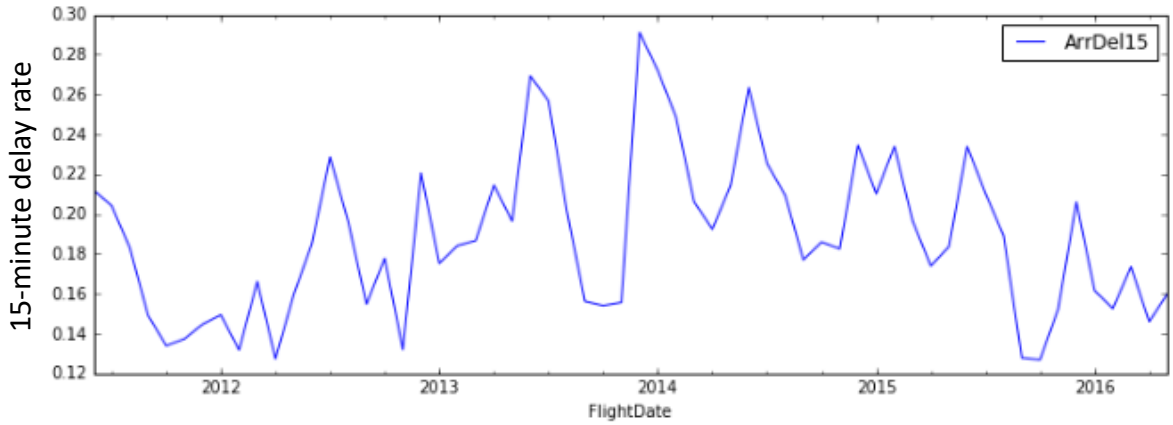
Flights a few days to either side of major travel holidays appear to have high delay rates, while flights on the holidays themselves have low rates; however, there is a good bit of randomness in the data at a

level this granular. To demonstrate this effect for Christmas and July 4, here are the day-by-day delay rates for July and December:





Flights overall don't appear to have improved or worsened over the 5-year period:



Descriptive statistics do not prove anything about causal effects—for instance, the fact that delay rates vary greatly by time of day does not prove that this is the most important causal factor. It is just an initial indicator of where the data might be leading us.

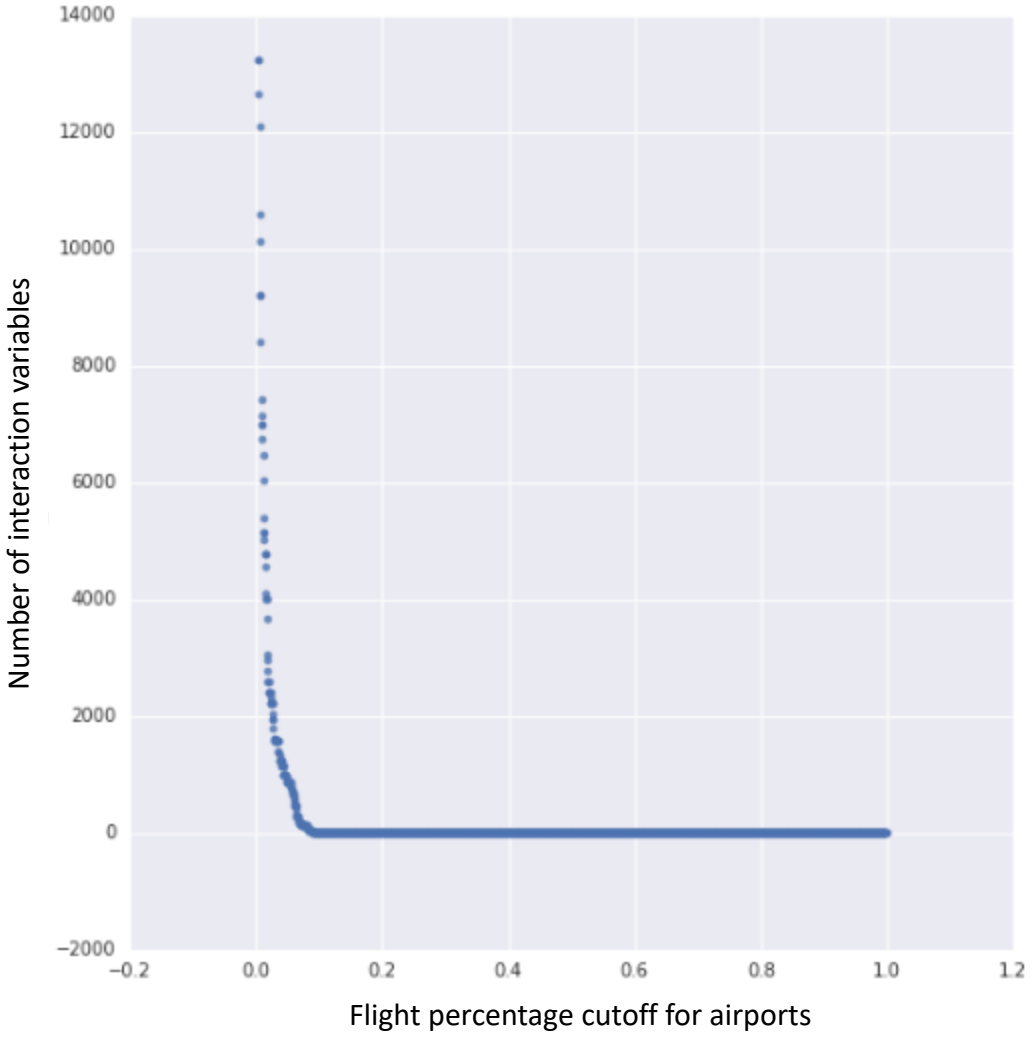
Model Approach

A regression model for this data will require very large numbers of dummy variables and interaction variables to isolate one effect from another. For example, how can we separate the effect of flying on Hawaiian Airlines from that of flying to Hawaii? We need not only a dummy for each Hawaiian airport and for Hawaiian Airlines, but the interaction between the two, so we can see whether flights to Hawaii are earlier on Hawaiian Airlines than other airlines.

My goal was to do a linear regression model for delay time and a logistic regression model for the 15-minute dummy. Due to computing resource constraints, I settled on a strategy that cuts down the number of interactions to only a few hundred:

- Only take interactions for airports with at least 1.5% of the arrivals or departures in the dataset. This is not ideal, as it would be better to find the effect of each airline at each airport, no matter how small. I chose this cutoff based on the scree plot below, which shows the cutoff for percentage of flights versus the number of interaction variables that would be created. The very steep part of the curve ends at between 1.5% and 2%, while the somewhat steep part ends at about 6%—after which there would be essentially zero interaction variables, since very few airports are large enough to have 6% of all flights. Restricting to the larger airports (over 1.5%) is the tradeoff I had to make between model robustness and resources.
- Only use interactions between airlines and airports. These are the ones I am most concerned about, because as stated above, I don't want my results to "punish" airlines for flying out of bad airports.

The total number of independent variables ended up being 1,430: 718 single-order dummies and 712 interactions. Note that the number of interactions is smaller than the scree plot would indicate because I included only interaction variables between airlines and airports, not between airports and months or times of day.



Model results

Linear regression

Despite including 1,430 dummy variables, including many interactions between airports and airlines, the linear regression model only explains between 1% and 2% of the variation in delay time when tested on batches of the test data set, based on R-squared values. This result is somewhat disappointing but not surprising: flight delays are caused by all sorts of factors external to this data set, such as mechanical problems with airplanes, weather problems that are not characteristic of particular airports, and simple random chance. Unfortunately, other methods of testing the model come to the same conclusion: it has almost no predictive power.

Nevertheless, the model confirms that certain flight characteristics have a significant impact on the expected delay time. According to the model, the 10 variables that most increase the expected delay time are:

1. Departing from Phoenix-Mesa Gateway Airport (AZA): +10.55 minutes delay
2. Departing between 11:00 PM and 11:59 PM: +10.53 minutes
3. Arriving at Dillingham Airport (DLG, in western Alaska): +10.44 minutes
4. Departing from Chicago Rockford International Airport (RFD): +10.13 minutes
5. Being a SkyWest flight arriving in San Francisco: +8.84 minutes
6. Being a Spirit Airlines flight: +8.73 minutes
7. Departing between 7:00 PM and 7:59 PM: +7.34 minutes
8. Departing from Shenandoah Valley Regional Airport (SHD): +7.27 minutes
9. Departing between 10:00 PM and 10:59 PM: +7.12 minutes
10. Departing between 9:00 PM and 9:59 PM: +6.55 minutes

Conversely, the 10 variables that most decrease the expected delay time are:

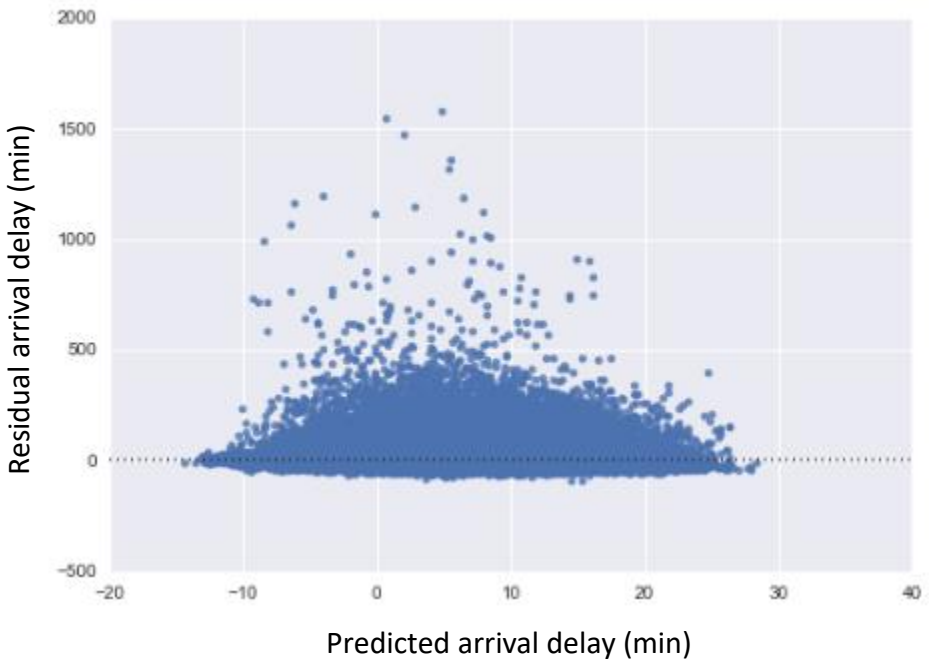
1. Being an Alaska Airlines flight: -4.74 minutes delay
2. Being a SkyWest flight arriving in Salt Lake City: -4.29 minutes
3. Being a flight in October: -3.85 minutes
4. Being a Hawaiian Airlines flight: -3.70 minutes
5. Being a flight in December: -3.45 minutes
6. Arriving at Peoria International Airport (PIA): -3.16 minutes
7. Being a flight in September: -3.15 minutes
8. Being a flight in August: -3.06 minutes
9. Departing between 6:00 AM and 6:59 AM: -2.87 minutes
10. Being a SkyWest flight departing from Salt Lake City: -2.87 minutes

Notice a couple of important points. First, these are not trivial numbers. Seven to 10 minutes of *expected* additional delay per flight adds up to a lot of hours of life for a regular traveler. A fisherman based in Dillingham is probably out of luck—but if you can avoid flying Spirit Airlines or leaving from Phoenix-Mesa Gateway (for a reasonable price), you should. Time of day is very important. Most people are not huge fans of 6 AM flights, but taking one instead of an evening flight means you reach your destination more than 10 minutes sooner. And those of us who grew up in Alaska—where Alaska Airlines is often the only option—are pretty lucky.

Another takeaway from this linear regression model is that despite its limitations, it does improve upon simple descriptive statistics. Notice that interaction variables involving SkyWest appear in both the top and bottom lists. For example, a SkyWest Airlines flight into San Francisco departs 8.7 minutes later than one would expect, even accounting for the effects of SkyWest and San Francisco separately. This suggests—although a deeper probe into the data is needed to confirm—that SkyWest is more susceptible to differences between high- and low-delay airports than other airlines. We would not have seen this by looking at descriptive statistics. We would see that San Francisco is a "bad" airport and Salt Lake City a "good" one in terms of delay time, but since SkyWest's overall delay rate is close to average, we would have expected SkyWest flights to be average for those airports. So if you are a San Francisco tech worker going for a ski weekend in Park City, take SkyWest out but not home.

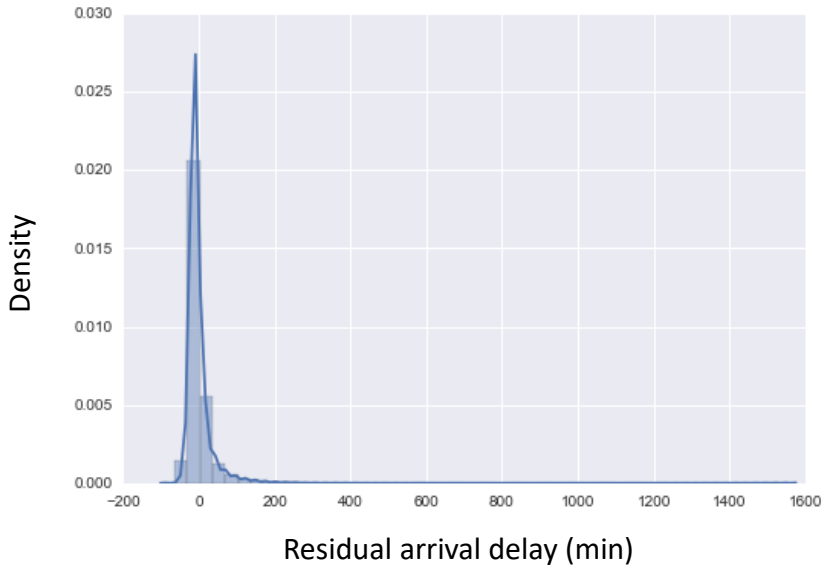
Linear regression diagnostics

The residuals-versus-fitted plot for the linear regression, using the test data rather than the training data, is as follows:



Most importantly, the residuals from the test set do not appear correlated with the predicted values. However, a few flights with extreme delays have very large residuals. There are no corresponding outliers on the negative side. It is hard to have a residual of -1500 minutes, as no flights are predicted to be delayed that much. This could make it appear as if the average residual is not zero. However, when I tested the average residual, it was in fact -0.53 minutes. It is not exactly zero because it was calculated using the test data; however, the negative value shows the positive outliers are not biasing the model.

The histogram of the residuals looks as we would predict based on the residuals-versus-fitted plot:



The peak is slightly below zero, but there is an extremely long tail above zero. Because of the lower bound on the delay time, the residuals are of course not normally distributed. This should not bias the coefficient estimates because of the very large sample size.[7] However, we should be careful about interpreting the confidence intervals around the coefficients and the predictions they generate, because the assumption of normally distributed residuals does affect the standard errors on the estimates.

Logistic regression

At first glance, the logistic regression model appears much more powerful than the linear one. However, its accuracy score of roughly 81% is merely due to the fact that about 81% of flights are not delayed by more than 15 minutes—it is no more accurate than just naively guessing "0" for every flight. The F1 score is 0.0002, or essentially worthless. This result is consistent with the linear model's findings that flight delay is primarily determined by factors external to the variables used in the model.

Like the linear model, the logistic model does have some power to show us which variables most affect probability of flight delay. The interpretation of the coefficients in a logistic regression is the impact of each variable on the log odds—in this case, the natural logarithm of the odds of being delayed by 15 minutes, where the odds equals the ratio of the probability of being delayed to not being delayed. According to the logistic model, the 10 variables that most increase the odds of a 15-minute delay are:

1. Departing from Shenandoah Valley Regional Airport (SHD): odds 81% higher
2. Departing between 11:00 PM and 11:59 PM: odds 81% higher
3. Departing from Phoenix-Mesa Gateway Airport (AZA): odds 74% higher
4. Departing from Chicago Rockford International Airport (RFD): odds 67% higher
5. Arriving at Dillingham Airport (DLG): odds 62% higher
6. Departing between 4:00 PM and 4:59 PM: odds 62% higher
7. Departing between 6:00 PM and 6:59 PM: odds 60% higher
8. Being a Spirit Air Lines flight: odds 59% higher
9. Departing between 7:00 PM and 7:59 PM: odds 58% higher
10. Departing between 5:00 PM and 5:59 PM: odds 54% higher

The 10 variables that most decrease the odds of a 15-minute delay are:

1. Being a Hawaiian Airlines flight: odds 82% lower
2. Being a flight in December: odds 73% lower
3. Departing between 6:00 AM and 6:59 AM: odds 41% lower
4. Departing between 7:00 AM and 7:59 AM: odds 36% lower
5. Arriving at San Angelo Regional Airport (SJT): odds 33% lower
6. Being an Alaska Airlines flight: odds 32% lower
7. Arriving at Newport News/Williamsburg International Airport (PHF): odds 32% lower
8. Being a flight in August: odds 28% lower
9. Being a Delta Air Lines flight: odds 24% lower

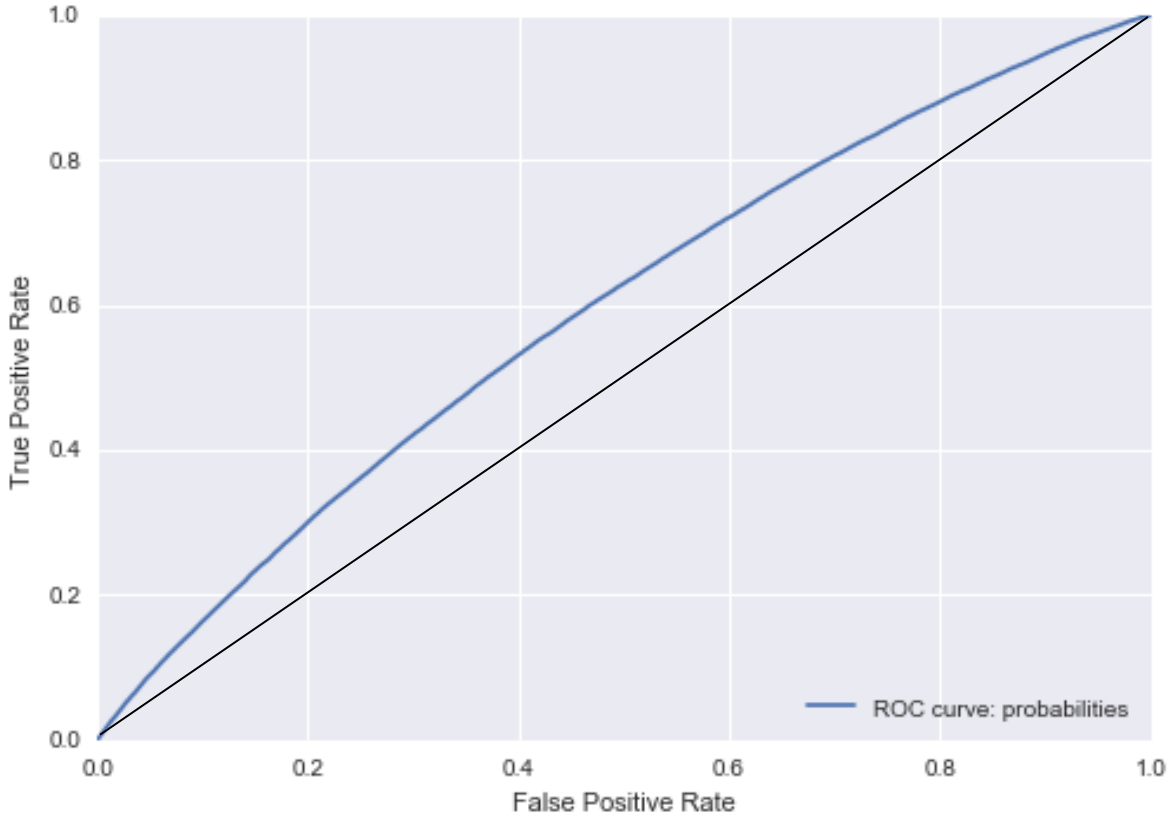
[7] http://blog.minitab.com/blog/adventures-in-statistics-2/how-important-are-normal-residuals-in-regression-analysis

10. Being a flight in October: odds 21% lower

Many of these variables are the same as those identified in the linear model, such as the best and worst airlines and airports. But there are some interesting differences—for instance, while flights in the late evening have the highest average delay times, it appears that flights in the early to mid-evening (between 4 and 8 PM) have a higher probability of short delays.

Another note about the predictive power of the logistic model is that it is important not to look simply at the predicted values, but the probabilities. Very few flights in the data set have enough red flags that the model gives them more than a 50% chance of being delayed by 15 minutes. Therefore, when we look at the confusion matrix and F1 score using only the predicted values (classifications), the model appears almost useless: true positives are no more likely than false positives. However, the model performs better—albeit leaving much to be desired, just like the linear model—when we look at confusion matrices with different cutoffs for what is considered a "positive" result, with the best performance coming when the cutoff is 20%, very close to the percentage of delayed flights in the whole data set. And when we run a ROC curve using probabilistic predictions rather than 0/1 classifications, there is a modest but noticeable bend to the upper left corner, meaning true positives are more likely than false positives.

The ROC curve for the logistic model:



The area under this ROC curve is 0.588, which does not indicate a strong model (0.5 would be worthless), but does indicate the model explains a modest amount of variation.

The confusion matrix using a 50% cutoff rate for a positive result makes the model look useless:

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

The confusion matrix using a 20% cutoff rate looks better—true positives now outnumber false negatives, though are regrettably still outnumbered by false positives:

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

Finally, regarding our question about Hawaiian Airlines, it looks as if they really do deserve credit. The coefficient for Hawaiian Airlines is the single best predictor of low delay rates and the fourth-best for low delay times, while the coefficients for various Hawaiian airports do not appear in the top-10 lists. A different airline is less likely to get you to tropical paradise on time. This may still be an indirect result of the favorable weather in Hawaii: Since other airlines fly to more different locations, flight delays due to late-arriving aircraft are probably more common. Still, this speaks to the benefits of Hawaiian Airlines' business model.

Conclusions and areas for future study

This capstone project has taught me a great deal about the technical work needed to produce a model—not just machine learning techniques, but also about practical computer memory issues and the time needed to research coding questions (although I do have experience debugging code in other contexts). The overall lesson I have learned is that if I seriously want to pitch an airline delay model to a client like Google Flights, I need to use a much more powerful computer and train a model with many more variables to account for the subtle interactions that govern flight delay. I would also need to include a much higher percentage of the flight data set, preferably all of the data rather than a 5% sample.

Google Flights already does provide warnings on certain flights, such as "Often delayed by 30 minutes," but it does not appear to incorporate a predicted delay time or probability for every flight into its algorithm for sorting the best flights. Based on what I have been able to find, I would recommend considering overall delay probabilities for airlines and airports, rather than just identifying particular flights that are often delayed. However, my model needs to improve its predictive power greatly before it would be ready for the market.

In addition to the improving the predictive power of the variables I already used, another avenue for improvement of this model would be several factors I considered but ultimately did not include. Load factor data would be valuable for predicting whether full flights are more likely to be delayed than

empty ones. The BTS does have free load factor data, but only found it aggregated by airline and time period rather than for every individual flight. Perhaps more useful would be data on the type of airplane used, since unlike load factor this is often known in advance (although subject to change on short notice). Finally, although I had latitude and longitude data for each airport, I did not use it in this model. It is not essential to predicting delay time by airport since the information about geographic variation is already encoded in the airport dummies; however, latitude and longitude (as well as elevation) could be useful for figuring out whether airports in cold and snowy locations more generally are likely to have longer delay times.

Finally, an area for improvement of the model would be to compare airlines against a standard flight time for each route instead of the scheduled arrival times published by the airlines. This is what FiveThirtyEight does in its more sophisticated version of the same analysis I did; they refer to their metric as a "target time".[8] FiveThirtyEight also has a special method of dealing with canceled and diverted flights, whereas I only considered the arrival delay time. It is interesting to note that FiveThirtyEight's model was created before Spirit Airlines began reporting, but they expected that company to finish "near the bottom of the table"—as it does in my analysis, ranking worst in average delay time and delay rate.

[8] https://fivethirtyeight.com/features/fastest-airlines-fastest-airports/