Final

Ben Kean

2/17/2019

problem

Most of us have been there before. We are on the way to the airport for a vacation, business trip or a long weekend out of town and *it* happens. A text appears on your phone. It is from a strangely formatted number and obviously written by a bot. Upon realizing this you know all too well what has happened: **your flight has just been delayed.** This is NOT the way you wanted to start your trip.

Is it possible that this could have been avoided? I believe that there is and I want to explore if there is a way to predict which flights, in terms of departure time and/or city should be avoided to reduce the likelihood of a flight delay.

client

These days, flight delays seem to be unavoidable. Flight delays cost business travelers millions of dollars each year in added travel costs, lost revenue and missed opportunities. **If there was a way for business travelers could predict which flights to avoid, they could save themselves time, money and stress.** They could select higher performing flights and better connection cities.

data

The data used for this project will be from the Kaggle website. It is sourced from the Department of Transportation and contains flight data for the year 2015. The data will need to be formatted and joined to airport data and airline data.

link to data (https://www.kaggle.com/usdot/flight-delays#flights.csv)

approach

For this study, I focused on which time of day the delay occurred, and at which airport/region the delay occurred. I thought it would be necessary to consider these two variables together due to environmental issues that affect specific airports (e.g. afternoon thunderstorms in Miami). Time of year was also considered to account for weather that only occurs during certain times of the year. In addition to the challenges that weather can pose, I wanted to see if avoiding certain flight times can improve one's chance of avoiding a flight delay, despite the time of year.

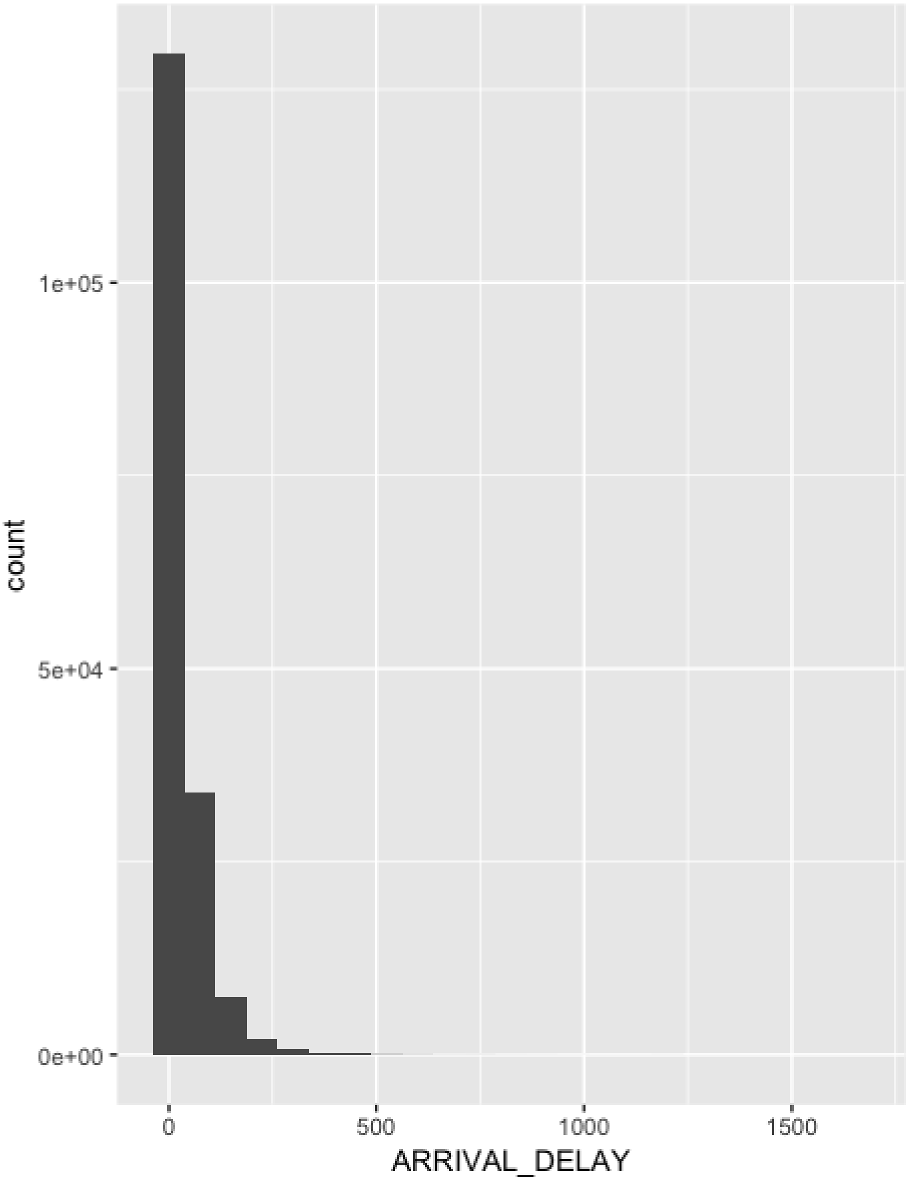
Data Wrangling

Application of Statistics

Summary Statistics

Below are summary statistics of the data when a delay occured. If a delay did not occur on a flight, then it was not included in the summary below. Below the summary statistics is a histogram showing the distribution of the delay times. Clearly, I am dealing with a left-skewed data distribution.

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.00 6.00 15.00 33.12 39.00 1636.00
```



Delay by flight hour

Probability of having a delay

The first data point I wanted to look at was percentage of flights that are late (delayed) upon arrival. Initially, I thought this would be 10-15%. However, according the data that I looked at, around 37.4% of flights are delayed - over 150% higher than I expected! A slightly higher number of flights were delayed for Departure. Departing flights were delayed around 39% of the time for the sample I looked at.

At a rate of 37.4%, a person would have the following expected chances of having a delay:
P(delay given n number of flights) = 1-(1-0.374)^5 1 flights - 37.4%
2 flights - 60.8%
3 flights - 75.5%
4 flights - 84.6%
5 flights - 90.4%

Probability of delay given an event

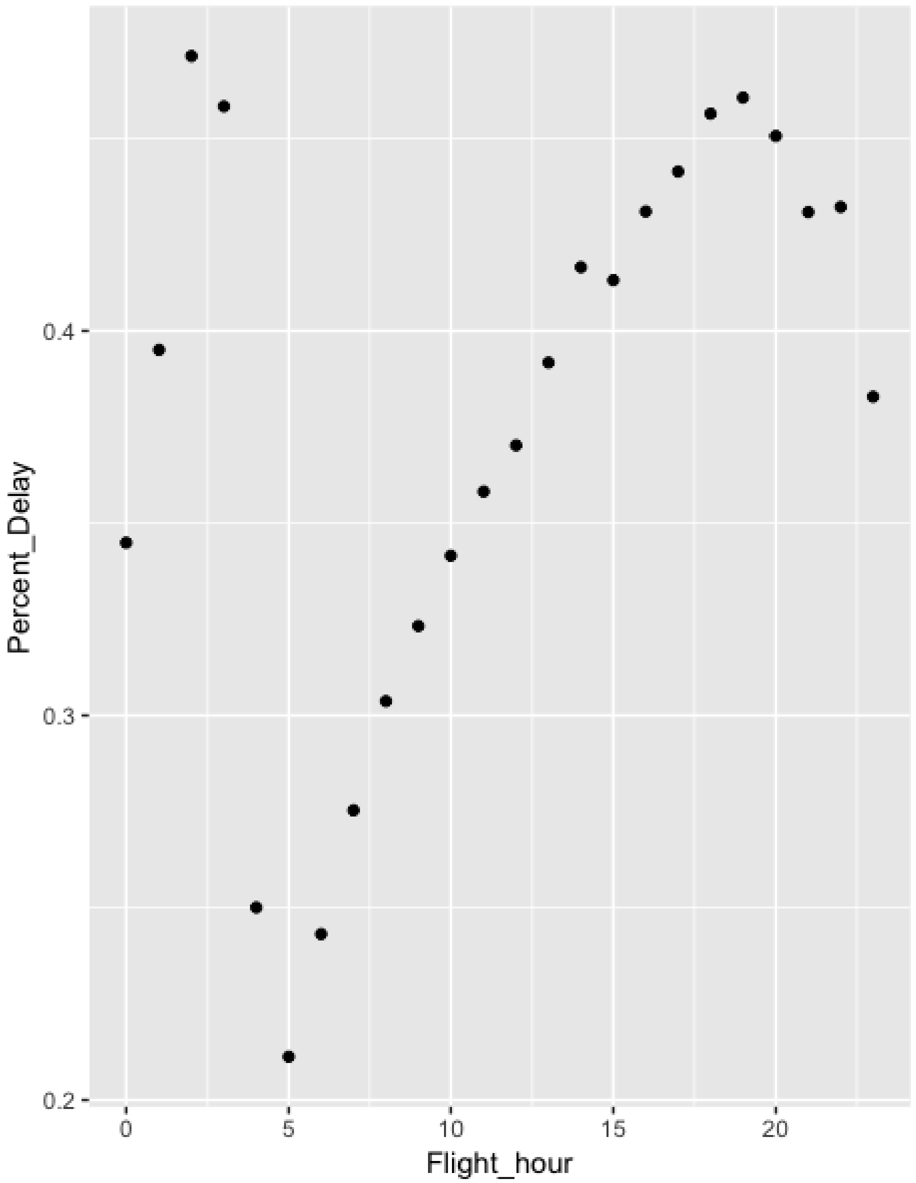
The next data point I looked at was determining the probability of an arrival delay given that a departure delay occured.

P(arrival delay|departure delay)

The results were not suprising. When there was a delay on the departure end of flight, there was likely a delay on the arrival side. The probability of a delay on arrival give there was a delay on departure was observed was 71.1%. Where there was not a delay on departure, you were only expected to be delayed 16.3% of the time. This expected outcome seemed obvious, but I wanted to review it because it might be a good indicator to use in a regression model.

Probability of delay given hour of departure

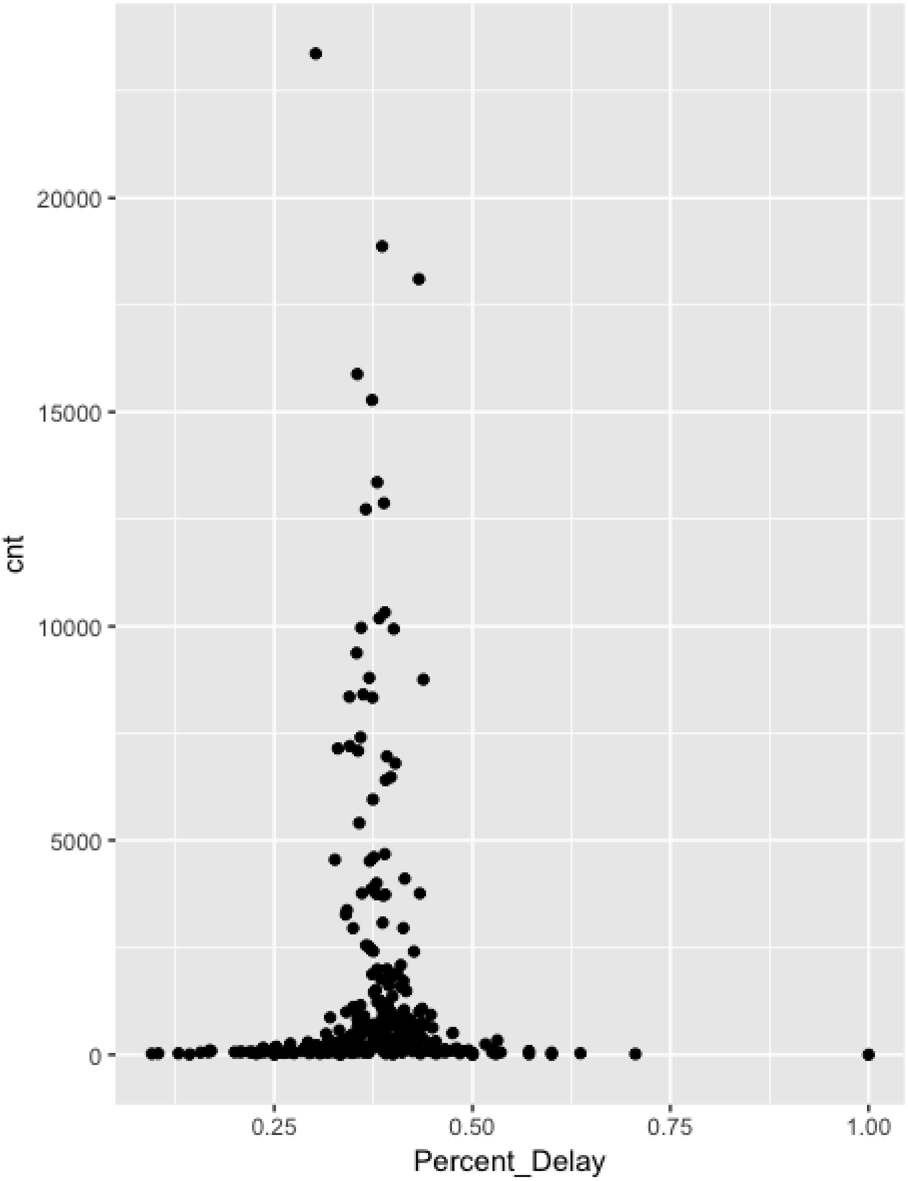
This statistic is the one of the main themes of my analysis. I expected that the results for the morning would show a lower probability of delay. My expectations generally follow what was shown in the data. The results show that early morning hours, that is 4 or 5 AM, show a lower expected probablity of delay (~25%) than later in the day and those hours after midnight (+40%). Due to the difference in time vs. how humans, and subsequently airlines, structure their day, I plan to create some extra elements for future analysis (morning, midday, evening, overnight). I believe this makes more sense and this is how major travel sellers present their products. Additionally, flights after midnight made up less than 1% of the data so it may make sense to eliminate them.



Delay by flight hour

Size of Airport

I wanted to see if the size of an airport was correlated to the probablility of a delay. I expected busier airports would have more delays. However, initial results do not support this idea. It seems that regardless of size, most of the larger airports seems to hover in the 37% range according to the graph.



Airport size

Machine Learning

Logistic Regression

The first machine learning technique I looked at was logistic regression. The main goal I was hoping to achieve was to determine whether or not a flight was delayed. Therefor, my dependent variable was a field I created Departure_Delay_BOOL that is a boolean where 0 represents no delay and 1 represents a delay. The independent variables I used to test for a delay were:

Flight_hour: the 2-byte hour for the scheduled departure. This was converted to a factor. **State**: this is the US State where the departure airport is located. **season**: the meteorological season. For example Jun, Jul & Aug are summer Dec, Jan, Feb are winter, etc. **Day_of_Week**: the day of the week the flight departed. **Taxi_Out**: the amount of time it takes for a airplane to taxi into position for departure

The baseline model, where it was predicted that there were no delays gave an accuracy of around 61%. Based on this, I was hoping to get at least 70% accuracy. However, depsite all the combinations I tried, I was not able to get better than 62%. Or, a 1% increase in performance over the baseline.

Initially, I classified *all* departure delays greater than 0 as a delay. However, this creates two issues in the analysis. The first issue is that not all delays are posing issues and need to be tracked. A delay of 1 minute was being treated the same as a delay of 1 hour. These are two different severity levels and should be treated different. The next issue is that the factors I was test were trying to predict different kinds of delays. Season and to some extent Flight_hour and State are mainly factors to consider for weather delays while Day_of_Week and Taxi_Out are more related to the airlines operational delays or air_system_delays. After splitting up the delays and identifying attributes seperately, I was able to get overall delay predictions of 62%, however, overall model accuracy was around 50-55%.

Linear Regression

The other ML techinique I attempted was linear. I wanted to take another angle and assess arrival delay to see if I would have better results. The following elements are the independent variables that I compared to arrival delay:

ELAPSED_TIME: the amount of time the flight took enroute + Taxi time **DISTANCE**: the length, in miles, of the flight

Both variables I picked turned out to be significant with 3 starts each from the R summary. However, results from this comparison were quite low and only about 5% of the variation is explained by the model.

Recommendations

1 Travelers should aim to leave earlier in th emorning.

body

2 Travelers should avoid the following airports, if possible: ORD, ATL

body

Avoid flying with these airlines

body