Howard Wright
General Assembly
June 23, 2016

Predicting Flight Delays for US Domestic Flights

1. Dataset and Project Goals

Organizing commercial air travel is a logistical issue that impacts a large number of citizens in the US on a daily basis. Every year there are millions of domestic flights in the US and, in 2015, about 39% percent of all domestic flights were delayed. As a Bay Area resident, I wanted to build a tool that would help consumers in the Bay Area avoid delays. My goal in this project was to build a model that could predict the chance of delay for the same flight at each of the three Bay Area airports (San Jose, San Francisco, and Oakland). The Research and Innovative Technology Administration (RITA) has published on time performance data for all US domestic flights from January 1987 to January 2016. I used this data, machine learning techniques, and user input modules to create my model.

2. Describing and Segmenting the Data

RITA's on time flight performance dataset is expansive totaling over 165 million domestic flights since 1987, a massive dataset. The goal of this project necessitated that I segment down the data to flights originating in the Bay Area and landing at one of the 21 shared destinations of the Bay Area airports. Even after using this segmented dataset there were too many rows for the scope of this project. I needed to keep the most relevant information for prediction while still using a smaller segment. I decided that using the most recent data would make the most sense due to changes in airline and air travel logistics trends, so used only flights in 2015 to fit my models. This gave me a reasonably sized dataset of about 185,000 flights. Below are the fields I used from the dataset to help segment the data and predict flight delays:

Table 1: Data Dictionary

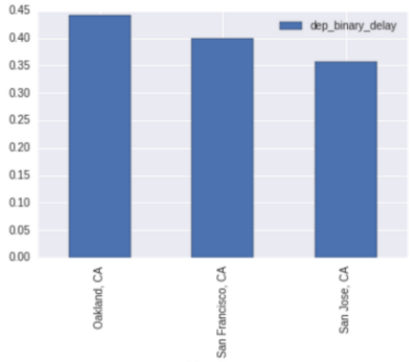
Field Name	Field Description
Month	Departure Month of the Flight
Day of Week	Departure Day of the Week
Day of Month	Departure Day of the Month
Distance	Distance in miles of the flight
Dep Delay	Minutes Delayed
Dep Delay Binary	Binary Flight Delay Status (Created from DEP DELAY)
Dep Time Block	Hour or Time Block of Departure
Origin City Name	Name of the Departure City
Origin City ID	Unique ID number for Departure Airport
Dest City Name	Name of the Destination City
Dest City ID	Unique ID number for Destination Airport
Carrier	Airline Carrier Name (Abbreviation)
Cancelled	Cancelled Status

Note: I ended up excluding cancelled flights from my dataset as they were outliers in the dataset.

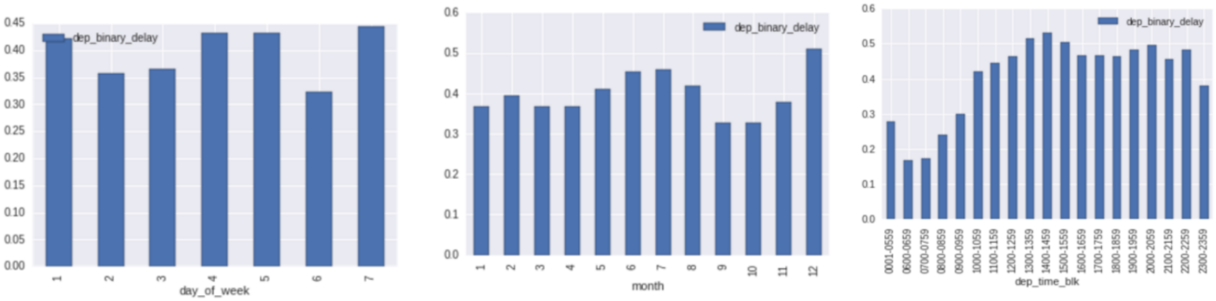
3. Data Observations

Initially I described my data by finding the average delay for each of the variables in my data. Please see the average chance of delay for variables below:

Figure 1: Percentage of Flights Delayed in 2015 at Bay Area Airports



Figures 2-4: Percentage of Flights Delayed by Day of Week, Month, and Time of Day



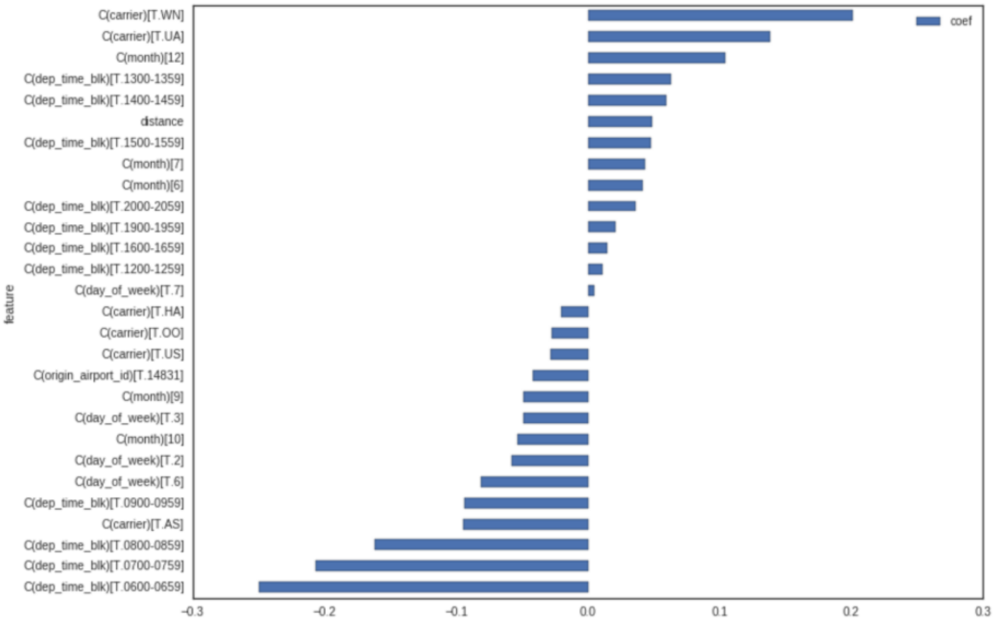
The above graphs show a significant difference in flight delays over the different times of day, days of the week, and months the of the year. Figure 1 also shows a general difference in delays at each Bay Area airport. Based on the difference in averages these categorical variables would be useful to try using them as predictors for flight delays or at least test if they will be useful.

4. Modeling and Results

I used various machine learning techniques to predict flight delays. As my predictions targeted delay classification of on time or delayed status, it was important for me to run a few tests on my predictions to gauge how well my model performed. The simplest way to measure model performance for classification is using accuracy. Accuracy is simply how many of your predictions were correct. The baseline accuracy for departure delays was 60.14%, meaning if I guessed that all flights are on time I would get 60.14% correct predictions. My model's accuracy was 66.06%, an increase of almost 6%. With such a large dataset 6% increase in accuracy is significant.

5. Coefficient Analysis and Implication on Flight Delays

Figure 7: Predictor Coefficients



The graph above shows the extent to which each of my predictors has an effect on delays. Positive values show that the variable increases the chances of delays while a negative value demonstrates a decreases the chances of flight delays. The more positive or negative the value more effect the predictor has in that direction.

i. Airline Carriers

a. Increased Delay: Southwest (0.21), United (0.14)
b. Decreased Delay: Alaska (-0.1), US (-0.04), Hawaiian(-0.03), SkyWest (-0.03)

In this model flying on Southwest and United Airlines are the two strongest influences on delay of any of the predictors. Anecdotally Southwest and United Airlines have both had had very negative public perceptions about their on-time performance. This model gives credence to the public perceptions. As a frequent Alaska Airlines flier, I was not surprised to see that the airline had the strongest negative coefficient of any airline.

ii. Months

a. Increased Chance of Delay: December (.11), June (0.06), July (0.02), August (0.02)
b. Decreased Chance of Delay: October (-0.07), September (-0.05), March (-0.02), April (-0.02), January (-0.01), November (-0.003)

In inspecting the categorical month variables, I was not surprised to see the summer months as contributors of delays. Many consumers enjoy traveling in the summer months to enjoy the good weather around the United States. It also made sense that December received the most delays because many people travel during the winter holiday season, and that time of year is also susceptible to weather delays.

iii. **Day of the Week**

a. Increased Chance of Delay: Sunday (0.02), Friday (0.01)

b. Decreased Chance of Delay: Saturday (-0.08), Tuesday (-0.07), Wednesday (-0.06)

With the standard Monday - Friday 5-day work week in the United States, many people like the travel for the weekends making Friday and Sunday prime time travel days. With more people traveling it make sense that airport would get congested and those two days would correlate with delays. Under the same logic, it makes sense that traveling mid-week on a Tuesday and Wednesday would decrease the chance of delays. Saturday, while a weekend day, is in the middle of the weekend and is a less ideal time to travel because it doesn't allow travelers to maximize on their weekend. With less people traveling on Saturday there is less chance of congestion and delays.

iv. **Time of Day**

a. Increased Chance of Delay: 1400-1459 (0.08), 1500-1559 (0.06), 2000-2059 (0.06),1700-1759 (0.03), 1200-1259 (0.03), 1600-1659 (0.04), 1900-1959 (0.04), 1800-1859 (0.04)

b. Decreased Chance of Delay: 0600-0659 (-0.25), 0700-0759,(-0.20),0800-0859(-0.15), 0900-0959 (-0.08)

This model shows that traveling earlier in the day decreases your chance of being delayed. This manifests because any flight delays can cause backups at the airport making subsequent flights that day delayed.

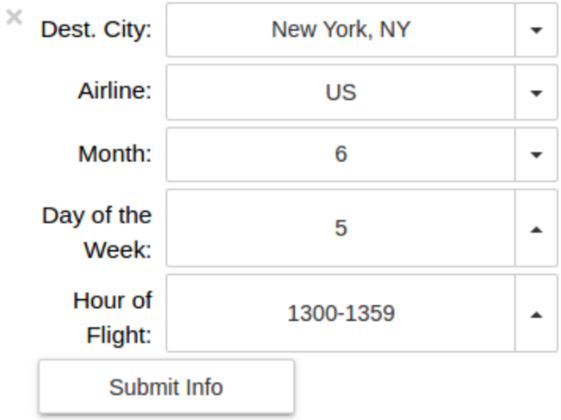
v. **Distance**

a. Increased by .07

Distance had a positive coefficient meaning the farther you fly the more likely for there to be a delay. The farther an aircraft has to fly the more preparation an airplane requires and the less aircrafts are eligible to make the flight. It can be harder to find replacement aircrafts and to do any maintenance that may be required.

6. Incorporating User Input

After cleaning my data and finding the optimal model, the final step in this project was to incorporate user input to fully complete the consumer tool that was my ultimate goal. Using the a programming package called ipywidgets package I was able to implement my model with user input dropdowns options of all of the major attributes of the flight. A snapshot of these dropdowns can be seen below:



Dest. City: New York, NY

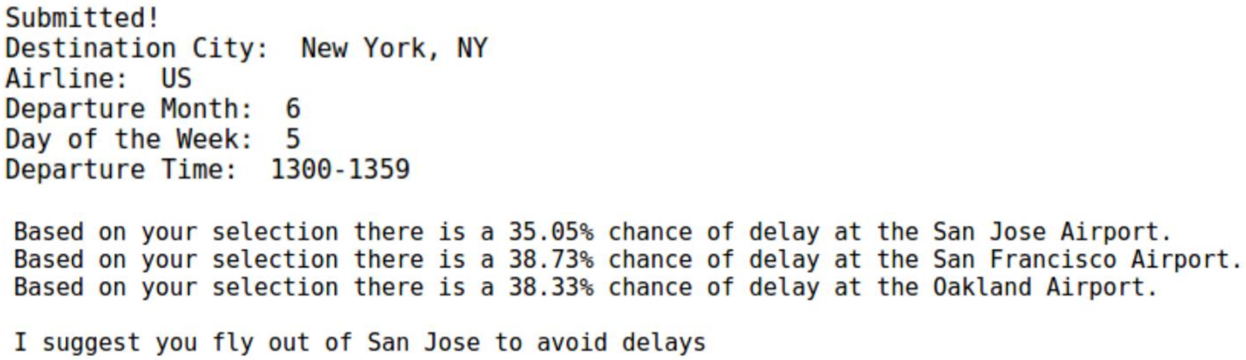
Airline: US

Month: 6

Day of the Week: 5

Hour of Flight: 1300-1359

Once the user selects the input they want and presses the submit button, my tool prints the selected outputs, gives the percent chance of delay for those specific flight details at each of the Bay Area Airports, and gives a recommendation of which airport to fly from. A sample output is printed below:



```
Submitted!
Destination City:  New York, NY
Airline:  US
Departure Month:  6
Day of the Week:  5
Departure Time:  1300-1359

 Based on your selection there is a 35.05% chance of delay at the San Jose Airport.
 Based on your selection there is a 38.73% chance of delay at the San Francisco Airport.
 Based on your selection there is a 38.33% chance of delay at the Oakland Airport.

 I suggest you fly out of San Jose to avoid delays
```

7. Potential Next Steps

Adding more types of data to my dataset could really help improve my model. If I could find hour by hour weather conditions at each of the Bay Area Airports I would be able to predict weather delays much easier and would add to my predictive power. I also think that being able to have more computing power and being able to use all the full dataset would also increase my model's accuracy in prediction. I would also like to incorporate a visual representation of flight delays in any future projects.