Flight Delay Prediction

By Sakina Zabuawala

Problem Statement

Flight delays are frustrating to airline passengers. A 2010 study found that delays cost passengers around $16.7 billion. This number was calculated based on lost passenger time due to flight delays, cancellations and missed connections, plus expenses such as food and accommodations that are incurred from being away from home for additional time.

In this report I propose a flight delay prediction model which would be a great add-on feature for websites like FlightStats.com. Passengers can use this handy tool a few months before their trip all the way to the day of the trip to have a hassle-free air travel.

Data

To build my model I obtained Airline On-Time Performance Data which is collected by the Office of Airline Information, Bureau of Transportation Statistics (BTS). This table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, carrier names, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, air time, and non-stop distance.

For my analysis, I looked at monthly flight data from 2017 and only considered the 30 major USA airports and the 8 major national airlines. After dropping the cancelled and diverted flights, I was left with about 2 million observations.

In addition, I also obtained daily weather data for 2017 from the Climate Data Online (CDO) which provides free access to NCDC's archive of global historical weather and climate data. These data include quality controlled measurements of temperature, precipitation, wind and snow.

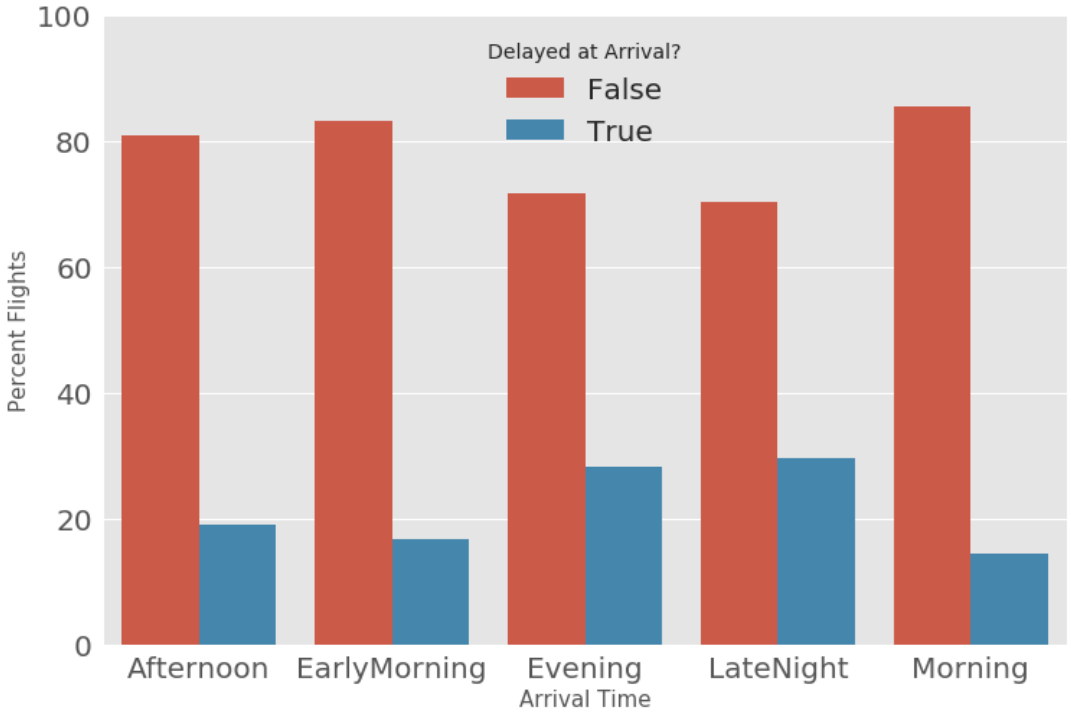
AWS

All processing, model building and analysis was done on AWS using Amazon EC2 Instance Type m5.xlarge

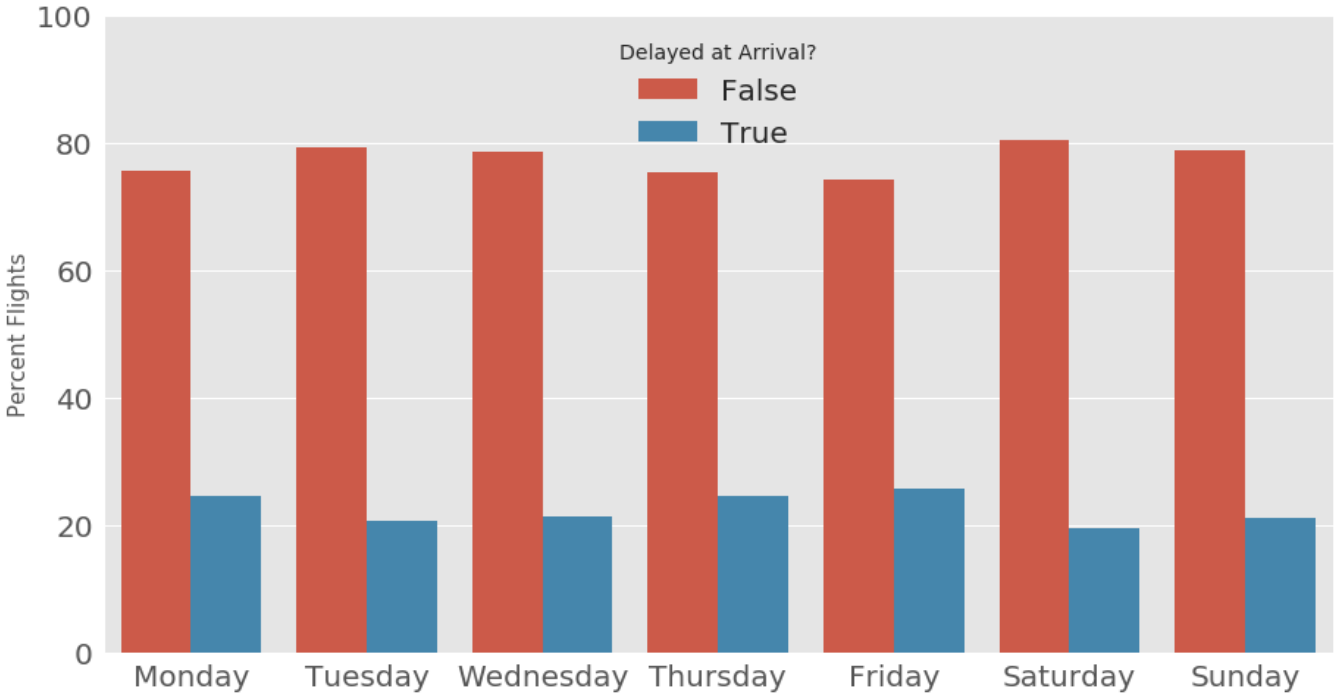
EDA

Next, I looked at the relationship between some of the independent variables and the target variable.

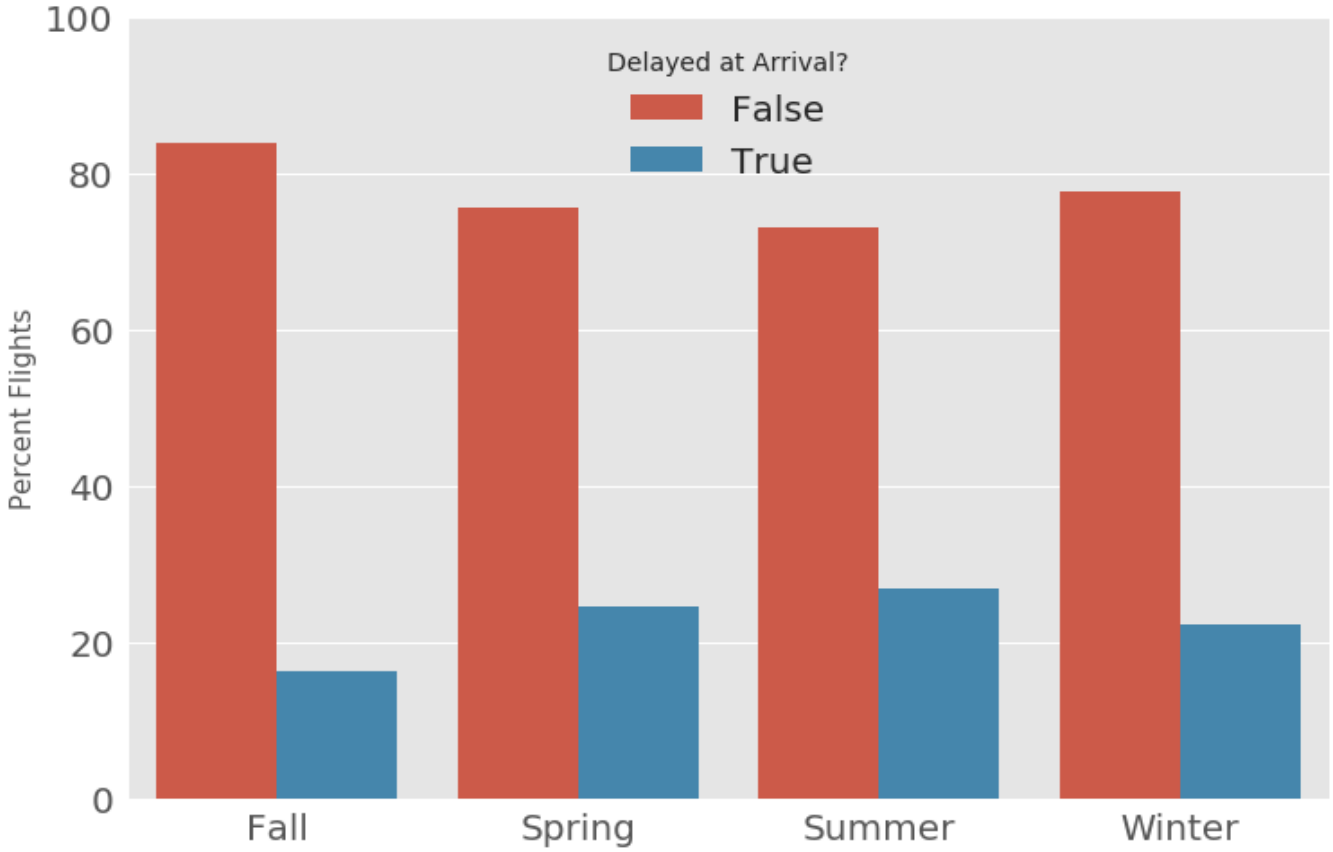
1. Arrival time of day and Arrival delays - Percentage of delays is more during Evenings and Late Nights



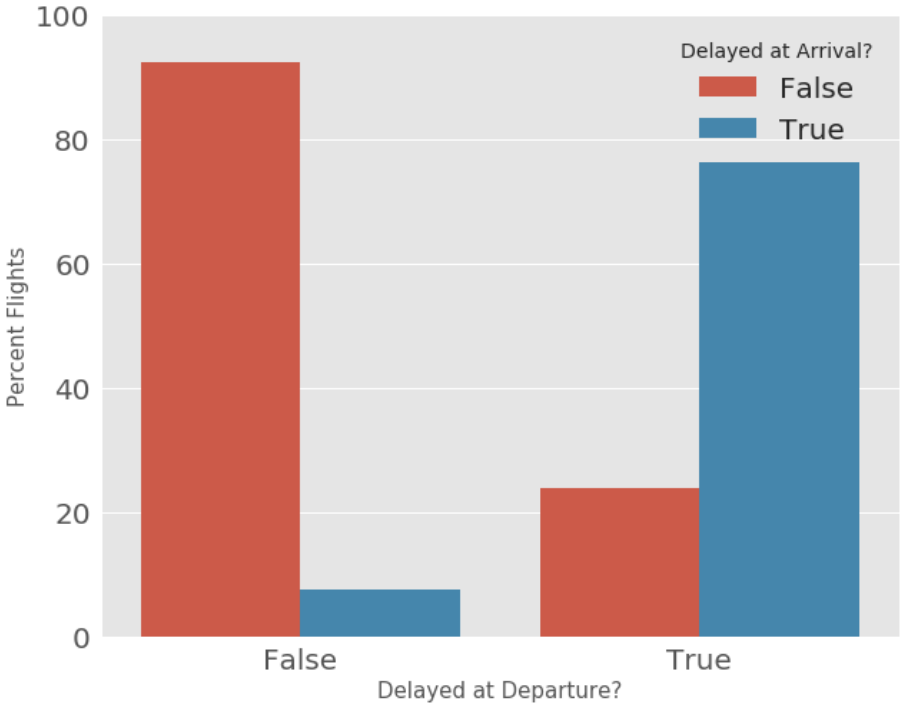
2. Day of Week and Arrival delays - Thursday, Friday and Monday have more delay percentage



3. Season and Arrival delays - Summer has the most percentage delays

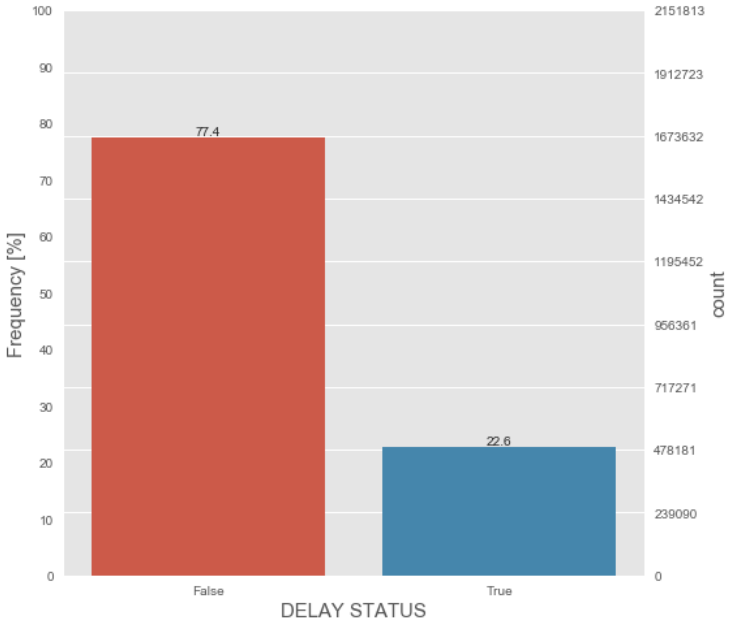


4. Departure delays and Arrival delays - If a flight is delayed at departure, it is highly likely to be delayed at arrival.



Target Variable

The target for my classifier is a binary class, a flight is either delayed at arrival or not. A flight is considered delayed when it arrives 10 minutes after it's scheduled arrival time. The positive class for my classifier is the delayed flight. The data contained 77% on-time flights and 23% delayed flights.



Feature Set

I built 3 models with increasing predictive power, which can be used at different stages of the passenger's trip planning. I also divided the baseline features into 3 sets - airport based, flight based and time based.

Model A (Baseline):

This model can predict delays about 2-3 months before the flight date. At this time reliable weather information is unavailable and we only have the passenger's flight information. The feature set for this model includes:

- Flight based - Airline name, Flight distance and Flight duration
- Airport based - Origin and Destination airport, Inter flight arrival time
- Time based - Arrival and Departure time of day, Day of week, Season of travel, Proximity to a holiday

The Arrival and Departure time of day are categorical variables with values Early Morning, Morning, Afternoon, Evening and Late Night.

Proximity to a holiday is a binary feature which indicates whether a flight date is a holiday or close to it or not.

Day of week is a categorical feature with values Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday.

Season is a categorical feature with values Spring, Summer, Fall and Winter.

Inter flight arrival time looks at all the flights arriving at an airport in chronological order and finds the difference between two consecutive arrivals. This gives an idea of how the busy the airport is.

Model B (with Weather):
This model predicts delays about 1 week before the flight date. Since a good weather forecast is available at this time we can include information like expected amount of snow/precipitation and wind speed.

Model C (with Departure delay):
This model predicts delays on the day of travel. At this time, in addition to the features in Model A and weather information, we also have the know whether the flight has departed or will depart the origin airport on time or not (which is a binary variable).

Memory Optimization

I converted all my categorical features to type 'category', the binary features to 'bool' and downcasted the 'float64' variables to' float32'. This brought down the size of my pandas dataframe from ~3GB to 600MB.

Categorical Variable Representation

I used three techniques to convert the strings to numeric values in the categorical features.
1. Dummy variables: Each category in the variable has a new column in the feature set, with 1 indicating its presence and 0 its absence. So the new feature set has number of categorical feature times the number of categories in that feature.
2. Numerical labels: Each category is given a numerical label. So it's just one column for each categorical feature.
3. Frequency count: Each category is represented by the number of times it occurs. Here again the number of columns is the same as the original feature set.

Classifier Models

Before trying out my models, I first divided the dataset into two sets - a Cross Validation set (70% flights) and a Holdout set (30% flights).

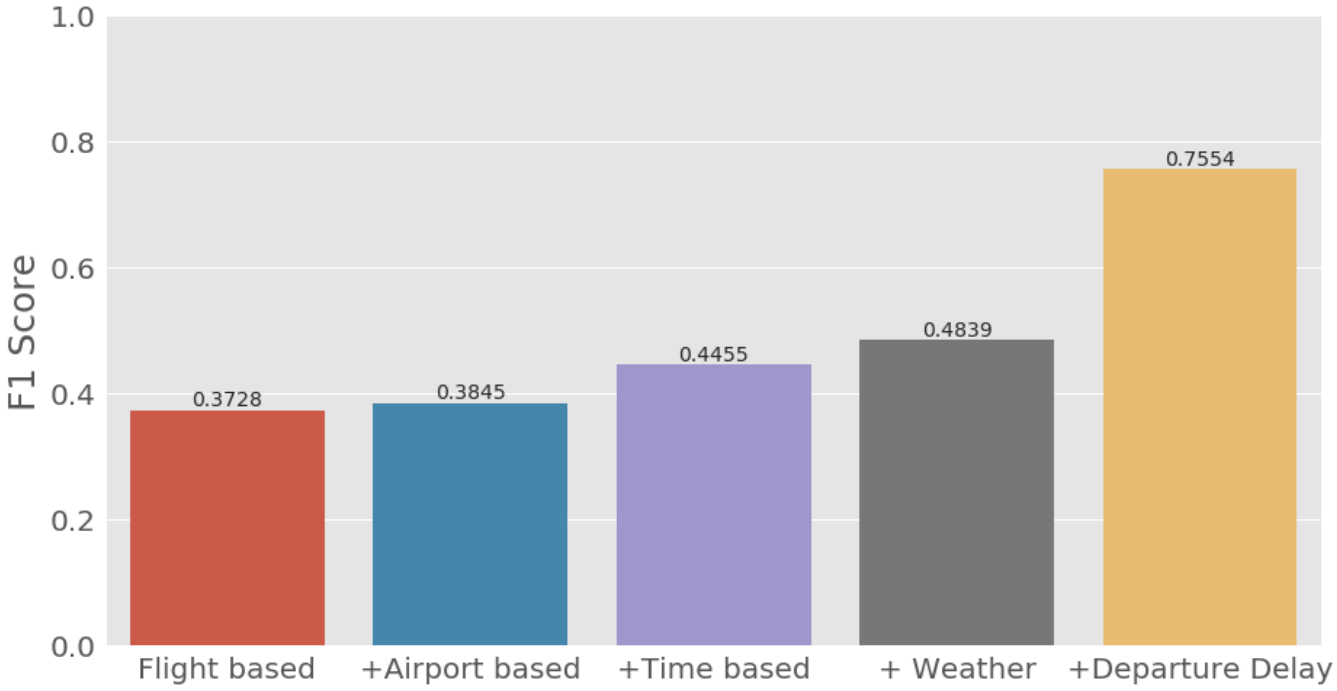
The models that gave the best results for all three prediction stages of my use case are

1. Logistic Regression with no regularization and class weights 4:1
2. Random Forest Classifier with 100 estimators and balanced class weights

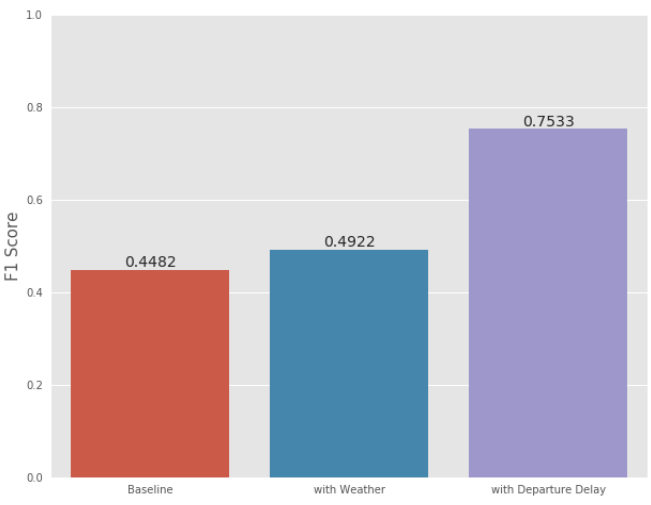
Performance Metrics

Since my target variable is imbalanced and I care more about the Recall, I used the F1 score and F2 score to choose the best model.

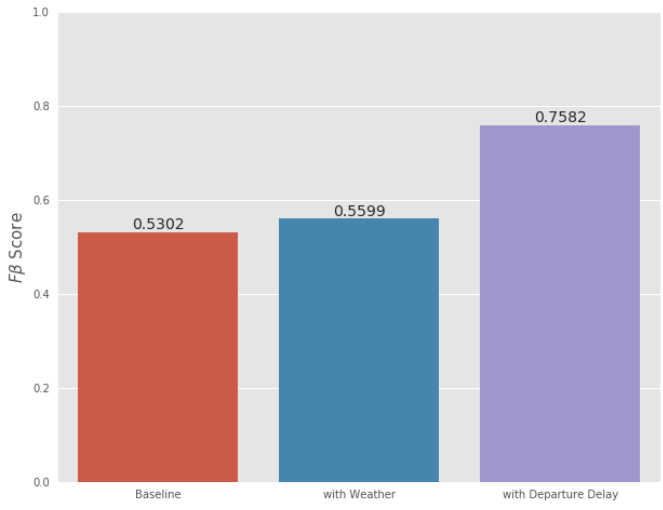
Results on CV Set using Random Forest Classifier



Results on Hold out Set using Random Forest Classifier



F1 Score



F2 Score

Interpretation of the Results

The variables with the 4 highest and 4 lowest values of the exponential of the coefficients from Logistic Regression are:

	Baseline Model	**+Weather**	**+Departure Delay**
Increases the odds of flight being delayed	Flights to SFO Flights to EWR Flights to LAX Flights out of EWR	Rain Forecast at Origin Rain Forecast at Destination Flights to SFO Flights to EWR	Delay at Departure Rain Forecast at Destination Flights to LAX Flights to SFO
Decreases the odds of flight being delayed	Flights in Fall Early Morning Departures Early Morning Arrivals United Airlines	Flights in Fall Early Morning Departures Early Morning Arrivals United Airlines	Flights in Fall Early Morning Arrivals Delta Airlines United Airlines

Features with the highest values of the feature importance scores from Random Forest Classifier are:

Baseline Model	**+Weather**	**+Departure Delay**
Flight Distance Evening Departures Inter-flight Arrival Time Flights in Fall Evening Arrivals Early Morning Departures Flights to SFO Flights in Summer	Rain Forecast at Origin Rain Forecast at Destination Wind Forecast at Origin Wind Forecast at Destination *Flight Distance* *Evening Departures* *Evening Arrivals* *Inter-flight Arrival Time*	*Delay at Departure* Rain Forecast at Origin Rain Forecast at Destination Wind Forecast at Origin Wind Forecast at Destination *Flight Distance* *Evening Departures* *Inter-flight Arrival Time*

The features with high logistic regression coefficient shows that they have a strong linear relationship with the target variable. Since the F1 score with Random Forest Classifier is significantly higher than the Logistic Regression shows that there are significant interactions between features and non-linear relationships with the target that are better captured by a complex model. Hence the features with high feature importance scores from the random forest classifier better predict the target.

Future Work

- Add more features
    - Airline staff at each airport
    - Number of terminals at various airports
    - Age of aircraft and airport
- Add more years of data
- Predict the arrival delay as a multi-class problem