# Flight's Delay Prediction and Application

Tian Zhang
Department of Computer Science
Boston University
Boston, US
tzhang94@bu.edu

Lu Min
Department of Computer Science
Boston University
Boston, US
mlu821@bu.edu

*Abstract*—**Delayed flights always cause frustration of both passengers and airline companies, and the consequences range from finance (missed connections and cancellations) to environment (wasted fuel) and society (loss of productivity and airport congestion). Thus we use flight arrival and departure data, airport data and aircraft company data to classify whether a flight would be punctual, so that we can conclude which features predict delays and work to mitigate them. Our purpose is to analyze factors leading to delay and predict an optimal flight with given time period and destination.**

***Keywords—flight delay; naïve bayes; random forest; linear regression; data science***

## I. Introduction

Every year there are over 20% airline flights delayed or cancelled around the world. Everyone who is about to take a flight doesn't want flights to be late, especially business people. Our purpose is to predict a flight that would have less probability to delay. Our model would give a record with a specific date, destination and other input variables, and this record would contain a unique flight record as a combination of a flight number and aircraft tail number. Our project would first benefit passengers who would not want flights to delay and second assist aircraft companies to predict factors attributing to delays in order to arrange their flights.

To conduct our analysis and set up our model, we obtain data from US Official Flight Data, Geographical Information of Airport and Aircraft Information with years ranging from 2011 to 2015 and focus on the flights from Boston and New York to Los Angeles. The airport codes are corresponding to 'BOS', 'JFK', 'EWR' and 'LAX'.

About approaches we use in this project: after we pre-process our data, first, we use statistical method to get the tendency of our 2015 flight data, and we get some basic information, such as the quantity of flights monthly, the percentage of flight delay and other general delay information. Second, we build a Naïve Bayes Classifier for calculating metrics for future visualization and analysis. Third, we apply random forest to get the correlation coefficients which can help us to choose features for building linear regression model. Finally, we build the linear regression model, including ordinary least squares (OLS) and ridge regression model, using root-mean-square error (RMSE) and mean squared error (MSE) to evaluate the accuracy of each regression model in train and test dataset.

## II. Dataset of Use

### A. Dataset

First, we collect dataset from the official website of Bureau of Transportation Statistics (BTS)[1] that offers us publicly available data on flight arrivals and departures for major U.S. airports. The downloaded dataset contains 100 variables that describe each flight in terms of departure/arrival date and time, carrier, taxi time, time spent in the air, as well as departure and arrival delays and their causes.

Second, we collect the aircraft information from the FAA website[3], and we import two txt files from this dataset.

Finally, we collect geographical information of airport from the openflights[2].

### B. Data Processing

We collect the data from the above three independent datasets, and merge those data we need and extract useful information for future analysis.

About the aircraft information, except for the basic information we used, we need to add extra aircraft information for better analysis. Sometimes the manufacture year or aircraft manufacturer will influence the punctuality performance of flight, since some older aircrafts are more likely to have mechanical issues than other newly produced aircrafts, and some aircrafts produced by some specific manufacture companies are more likely to have potential problems causing flight delay. So we add the feature "AIRCRAFT_YEAR" for service life of aircrafts.

Then we combine the flight data with airport location information. Meanwhile, with the geo-information of airport, we can analyze the geographical flight delay tendency, and get statistical analysis of monthly, seasonally or yearly flight delay.

Finally, for higher accuracy, we remove the following entries from the raw dataset: First, flights that have been cancelled or diverted. We focus on predicting the delay, so we also remove the columns associated with diverted flights. Second, columns give the answer. This is the case of many columns related to the arrival of the plane. Third, we would drop rows where a value is missing.

# III. BASIC STATISTICAL ANALYSIS

## A. *Purpose*

Before building predicting model for flight delay, we firstly use statistic method to get the general tendency of our 2015 flight dataset, such as the quantity of flights monthly, the percentage of daily flight delays and other general information. Meanwhile, we can use the general information to judge the performance of each airline, airport or flight route.

## B. *Results*

We get 10 best departure airports, 10 worst departure airports, 10 best destination airports and 10 worst destination airports. Here "best" means least delay time, and "worst" means longest delay time, and same meaning for the "best" and the "worst" below.

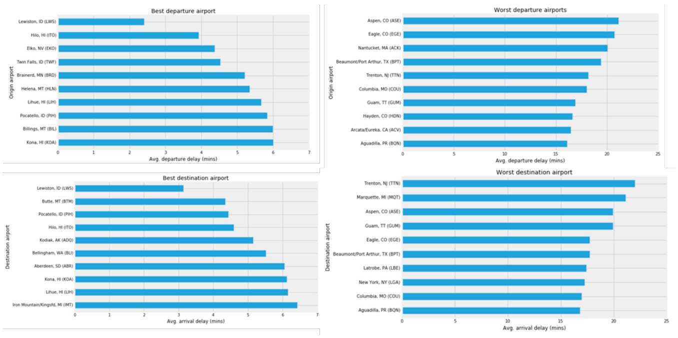


Figure 1

We also get 5 best airlines and 5 worst airlines.

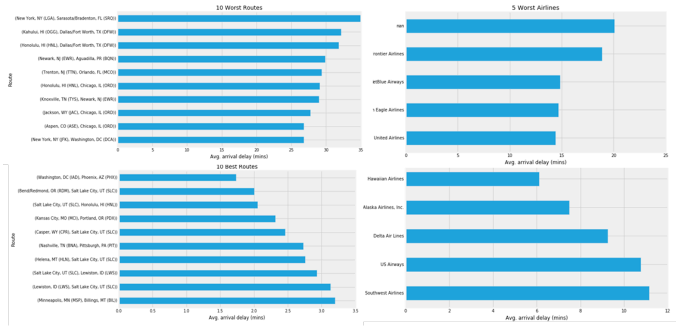


Figure 2

Using the flight data from 2011 to 2015, we perform statistical analysis and figure out the tendency of delay. Meanwhile, we find out the flight delay has monthly, daily fluctuation. So we decide to use regression model to explain those general phenomena and make a prediction for future flight.

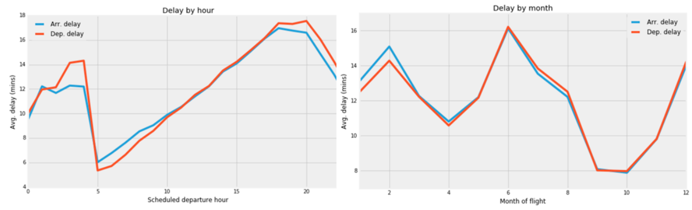


Figure 3

We also get the chart for proportion of all departure delays with carrier delay, aircraft delay and weather delay.

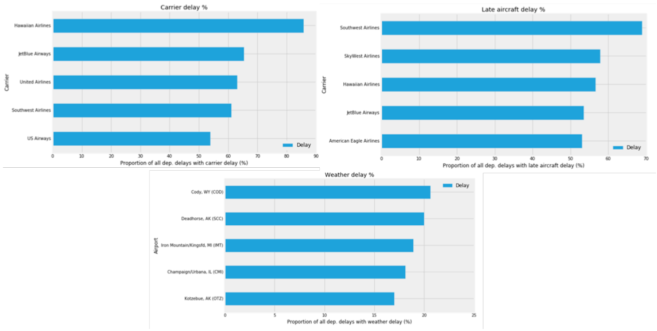


Figure 4

# IV. NAÏVE BAYES

## A. *Model Implementation*

We choose Naïve Bayes Classifier, because Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The Naïve Bayes Classifier was implemented using scikit-learn, and in our project's situation, Bayes theorem states:

$$P(Delay \mid attribute\ a) = [P(attribute\ a|Delay) * P(Delay)]/P(attribute\ a)$$

By applying the model, we would like to figure out the probability of delay given a tuple of attributes. Python's scikit-learn provides a fairly simple function for training, testing, and assessing the results of the model. Because the large size of our dataset, we would use a sample of our data.

## B. *Resluts and Analysis*

We apply a Naïve Bayes algorithm with eight-fold cross-validation on the 4 years of data (2006-2009) to train and the 5th year (2010) to test the model. This model is our first attempt to our data, and we will implement other models to get better performance and results catering to our purpose. And this part's training years are different from other models, so the results are our reference for future modeling.

The Naive-Bayes results show us that the classifier performance is far better in predicting non-delayed flights than delayed ones. The F-score on predicting on-time flights is 0.78, while that for delays is only 0.566. Thus, we may use other method to improve the performance of predicting flight delays.

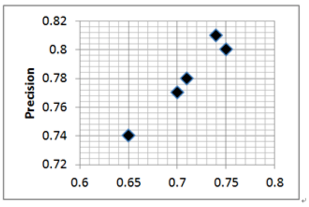


Figure 5 precisions to call

# V. RANDOM FOREST

## A. *Pre-processing*

We focus on flights between Boston, New York and Los Angeles. When cleaning the data set, we have to remove the following entries: flights that have been cancelled or diverted, meaningfulness data columns, rows with null value and flights' destination or origin do not correspond to our need.

We need to normalize the data collecting. The idea of our normalization is that if the variation of one feature A is higher than feature B, we believe the variation of feature A may completely contains the variation of B, which means feature B is less useful for the prediction model.

### B. *Creating Baseline*

We create a baseline predictor that returns the mean of delay time in data set. In our model, the evaluation of predictor will be the Mean Square Error (MSE) against a test set distinct from the training set. And we split the dataset to 70% training set and 30% testing set.

### C. *Model Implementation*

Our final goal is to predict the delay of flight, so we decide to apply random forest first. In general, a random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

Define the departure of flight is late for more than 15 minutes as delay flight. And we use '1' to present 'Delay' and '0' to present not 'Delay'.

### D. *Results and Analysis*

We use a reduced 70% of dataset to train a random forest, and each node contain three features ("auto", "log2", "sqrt"). The following figure 3 shows that the performance of random forest, and we can see that the forest quickly converges after a relatively low number of trees

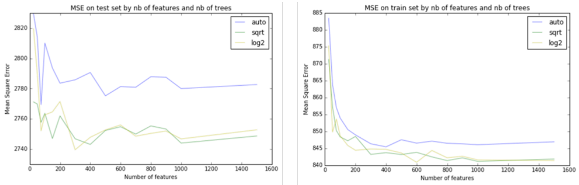


Figure 6 MSE on test and train set

The following figure 8 indicates the features correlation with delays, which we can apply to build another model.

The random forest performs worse than our baseline (2237.042), but we can regard it as feature engineering. As we can see that the main factors in the delay of a flight are the age of the aircraft, the time of departure and the time of arrive. However, the role of service life of the aircraft was not very marked in the explorative analysis. The weekday has a huge influence. The airline and the aircraft model play a very limited role in the delay according to this random forest classifier. With analysis, we can improve the efficiency of feature selection in later model.

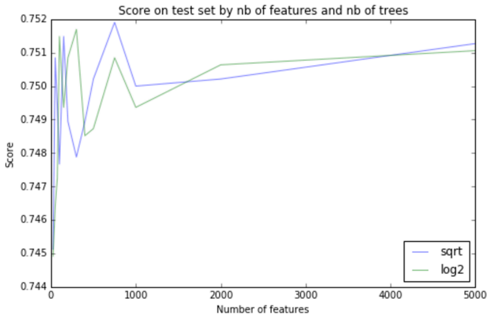
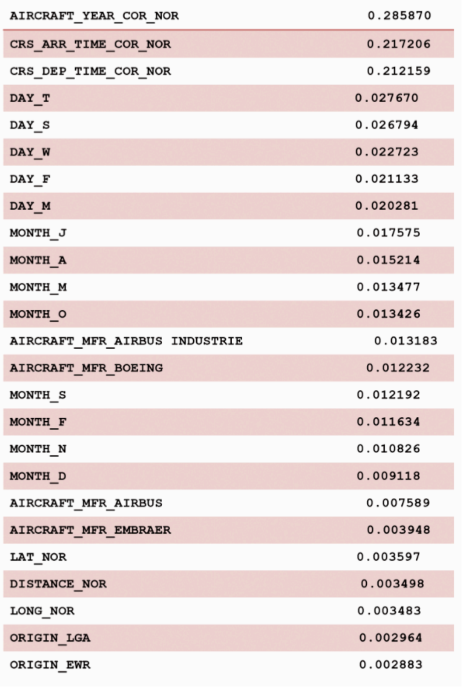


Figure 7 Score on test set



| | |
|---|---|
| AIRCRAFT_YEAR_COR_NOR | 0.285870 |
| CRS_ARR_TIME_COR_NOR | 0.217206 |
| CRS_DEP_TIME_COR_NOR | 0.212159 |
| DAY_T | 0.027670 |
| DAY_S | 0.026794 |
| DAY_W | 0.022723 |
| DAY_F | 0.021133 |
| DAY_M | 0.020281 |
| MONTH_J | 0.017575 |
| MONTH_A | 0.015214 |
| MONTH_M | 0.013477 |
| MONTH_O | 0.013426 |
| AIRCRAFT_MFR_AIRBUS INDUSTRIE | 0.013183 |
| AIRCRAFT_MFR_BOEING | 0.012232 |
| MONTH_S | 0.012192 |
| MONTH_F | 0.011634 |
| MONTH_N | 0.010826 |
| MONTH_D | 0.009118 |
| AIRCRAFT_MFR_AIRBUS | 0.007589 |
| AIRCRAFT_MFR_EMBRAER | 0.003948 |
| LAT_NOR | 0.003597 |
| DISTANCE_NOR | 0.003498 |
| LONG_NOR | 0.003483 |
| ORIGIN_LGA | 0.002964 |
| ORIGIN_EWR | 0.002883 |

Figure 8 Correlation coefficients with delays

## VI. LINEAR REGRESSION

### A. *Model Implementation*

After getting the correlated coefficient value of each independent variable, we found that three of those features influence flight delay observably, including aircraft producing year, departure time and arrival time. Thus, we choose features from our dataset according to these three aspects for our linear regression model.

By considering the size of our enormous dataset, we firstly decide to use a reduced dataset to build the regression model. We limit the origin and destination to Boston and New York , and get a reduced dataset with flight from three airports (BOS, EWR and JFK). And then we tried to fit and normalize all of us selected categorical and numerical features to the linear regression model of the sklearn regression model. Specifically, for numerical features, like depart time and arrive time, we convert them to the Coordinated Universal Time (UTC), and normalize other numerical to unified interval. For categorical feature, like location, day of week or day of month, luckily, sklearn has a routine to do this for us. We first re-index all categorical features and then create lookup tables for the values. (This converting step costs tremendous time to run).

We fit our preprocessed reduced dataset to the linear regression model and ridge regression model by using sklean packages. But we find that the performance of these two models for the reduced dataset are not well, since their RMSEs are higher than our pre-defined baseline (average flight delaying time). Thus, we decide to use the whole dataset of 1 year and 5 years to build the linear regression model as same as above fitting steps. Finally, we tried to use the model to make real-world prediction in pre-defined time schedule.

### B. *Results and Analysis*

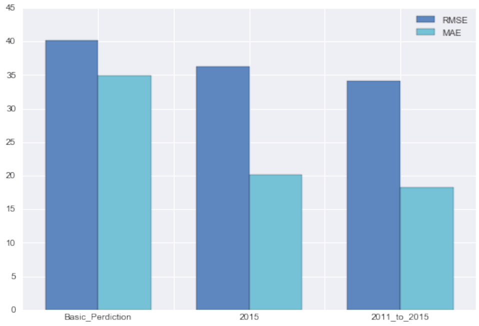
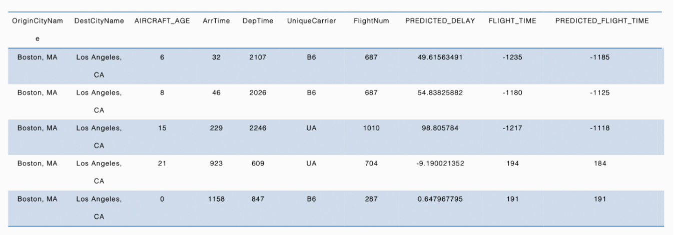


Figure 7

We try to apply our model in both one year and five years dataset. We can conclude that this model is better than our baseline and has a tendency of improving accuracy among more years dataset. We get a 1-year model and 5-year model, and try to use these two models to predict real-world flight delay.

We get several records of flights information by predicting the punctuality performance of flights from BOS to LAX in December 17th, 2016. The following table shows some recommended flights.



| OriginCityNam e | DestCityName | AIRCRAFT_AGE | ArrTime | DepTime | UniqueCarrier | FlightNum | PREDICTED_DELAY | FLIGHT_TIME | PREDICTED_FLIGHT_TIME |
|---|---|---|---|---|---|---|---|---|---|
| Boston, MA | Los Angeles, CA | 6 | 32 | 2107 | B6 | 687 | 49.61563491 | -1235 | -1185 |
| Boston, MA | Los Angeles, CA | 8 | 46 | 2026 | B6 | 687 | 54.83825882 | -1180 | -1125 |
| Boston, MA | Los Angeles, CA | 15 | 229 | 2246 | UA | 1010 | 98.805784 | -1217 | -1118 |
| Boston, MA | Los Angeles, CA | 21 | 923 | 609 | UA | 704 | -9.190021352 | 194 | 184 |
| Boston, MA | Los Angeles, CA | 0 | 1158 | 847 | B6 | 287 | 0.647967795 | 191 | 191 |

Figure 8 Records of prediction by linear regression

## VII. Discussion

We explored three models, including Naïve Bayes, Random Forest and Linear Regression, in attempting to predict the delay of the flights. The first attempted Naïve Bayes model performed well, but the model did not converge to a global solution for flight prediction. And although the random forest has high RSME, it helped us to get an insight about the parameters that come to influence the flight delay, especially the time of flight and weather. Meanwhile, we encountered multiple problems when developing several models, mostly around complex data transformations necessary to ensure meaningful inputs and computational complexity encountered during training. After data format transformation and normalization, we apply it to an applicable predicting model, and make a prediction for the flights from BOS to LAX in December 17th, 2016. In addition, our classification (>15 min) shows that flights longer than 15 minutes late do not have strong distinguishing characteristics in the dataset, so better features may improve the model.

Consequently, after trying three models for prediction, we decide to use linear regression model to predict the punctuality performance of flights by our given time and destination.

## VIII. Future Work

Our linear regression model could be further improved by augmenting the flights feature of dataset with additional data sources such as weather data, since weather patterns are likely important factors influencing departures delays, and additional geographical analysis of weather is necessary for explaining the monthly and daily flight delay in clustering areas. Also we want to explore the holiday influence for flight delay, since huge population is more likely to take flights during holiday, which definitely will increase the possibility of delay. Overall, we believe that a mixed prediction model with more additional features will increase the accuracy of flight prediction model.

## References

[1] The official flight database for every domestic flight in the US 2011-2015: < http://www.transtats.bts.gov/>

[2] Airport information with geographical data:
< http://openflights.org/data.html>.

[3] Aircraft and airline information :
<http://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/>

[4] Random forests:
<http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>