Predicting Flight Delay

Kaushik Visvanathan: kaushikv@seas.upenn.edu
Tse-Lun Hsu: tsehsu@seas.upenn.edu
Sam Weinberg: wsam@seas.upenn.edu

Abstract

We set out to develop a set of models and visualizations to interpret flight delay at the largest airports in the Continental United States. Flight delays are a critical inconvenience everyone faces in their life and a blemish on the brand of any airline that has them. However, flight delays are not a necessity. Our Machine Learning project sets out to understand the feature space of flight delays as well as their correlation with airport and city specific factors. For our website containing more visual and our findings, please see below:
`https://flightdelay519.herokuapp.com/`

Table 1: Major airports in the United States

LOCATION	TRAFFIC
ATLANTA	882,497
CHICAGO	875,136
DENVER	541,213
HOUSTON	502,844
PHILADELPHIA	411,368
SAN FRANCISCO	429,815
SEATTLE	381,408
MIAMI	412,955
BOSTON	391,222
NEW YORK (LGA)	369,987
NEW YORK (JFK)	438,448
WASHINGTON	264,785

Hypothesis
We expect to see a strong correlation between the overall number of air traffic through an airport and the percentage of total flights that are delayed through a certain airport. Other informative features would be the type of airline and the distance traveled per flight.

1 Final Submission

Motivation and Problems

We began by selecting the ten most highly trafficked airports in major cities in the contiguous United States. In broad strokes, the features we selected would inform our model as well as the specifics of the open-ended problem we approached. Initially, we hoped to answer multiple questions, including: value and brand loss as a result of delay, effect of oil price on delay frequency, seasonal weather and holiday patterns and generic delay by airport/airline. Ultimately, issues with compiling data and time would relegate us to the final question.

Finally, the models and presentations acted as our final defining decision in shaping the problem we sought to answer. We wanted to offer both a granular and definitive estimate across our questions. For this we opted to use a logistic regression and a support vector machine with a Gaussian kernel. The former model offered a binary estimate on each instance, something we needed in casting definite prediction. With the latter, we hoped to offer some more specification without allowing long tails to define our estimate. For this reason, the drop-off nature of a Gaussian Kernel allowed us to characterize our data in a way we found appropriate.

1.1 Approach

The most immediate and obvious feature in assessing flight delay is local weather. Given that we expected to see a strong correlation between poor weather and flight delay, we elected to drop this feature in order to find connections between delay and other factors that might prove to be a bit more subtle.

We were able to obtain an exhaustive dataset consisting of various features for each flight - including departure/arrival airports, time of day, total airport air traffic for the day, flight time and distance to destination, amongst others. In order to incorporate factors not directly related to the airport or flight in question, we also obtained the following data points:

- Fuel Prices: Airplane speed is often held below the aircraft's maximum capability, depending on the necessity to conserve costs on fuel.
- Airline: Budget airline aircraft are often newer and smaller than aircraft operated by larger corporations, possibly impacting the speed and frequency of use.
- Aircraft type: Variance in the size and make of aircraft could impact delay.

Data Preprocessing

For airline information we had datasets from different sources for 2015 and 2016. The source discrepancy meant that one set was more feature rich than the other. At 48 features, the 2016 set was quite cluttered. We ultimately went with this one and the pros of having more features outweighed the negatives as cleaning the set was not overly difficult. We removed state information, quarter, month, flight number, origin id, state abbreviation, destination airport id and a number of other repetitive labels. We then removed all rows that didnt have arrival or departure from an airport we selected. We wanted to have a smaller pool of airports so we could have equally representative data for each. This was as opposed to including as many airports as possible and having different levels of data on each. We ultimately settled on the 10 airports that had the most information in our 2016 dataset.

Flight arrival and departure delay times were expressed as continuous values, which we mapped to binary labels. A positive value for a given sample represents any delay greater than one minute. Though this seems like a rigid definition of delay, the majority of flights that were within a couple of minutes of their expected arrival time were already labeled with a 0, meaning that a delay of 1 represented a legitimate lag behind expected arrival or departure time.

Additionally, delayed flights accounted for approximately 15 percent of all data. Theoretically, our model could achieve a relatively high accuracy at the expense of precision by predicting a 0 for every instance. In order to account for this, we undersampled the dataset, reducing the overall number of instances labeled a 0 to balance the classes a little more evenly. Additionally, setting the class weight parameter of our support vector machine to 'balanced' allowed us to set the penalty parameter for mis classifications ('C') based on the prevalence of the respective classes.

Model Selection

While delay times varied even with instances classified as a positive delay, we elected to use classifiers rather than a regression model. This was motivated by our desire to find causal factors related to the airport, airline and city, rather than information associated with that particular flight. We reasoned that these factors would result in an informative enough model capable of predicting binary labels, given that these factors are more closely tied to the macro factors behind airline delays than specific instances.

Given that our data consisted of a mix of continuous and qualitative features, we believed that a linear model and an SVM would provide us further insight into which category of features were more strongly related with delays.

SVM with Gaussian Kernel

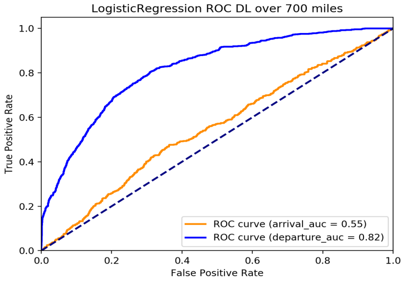
Support vector machine (SVM) works really well in the high-dimensional feature spaces and is able to efficiently perform a non-linear decision boundary. After looking our data set, we decided to use Gaussian kernel since our data set is neither polynomial nor linear. Initially, all the features were input to train and test the model, but the accuracy was slightly above 50 %. In order to improve the accuracy, we decided to assess which features have bigger weights toward the classification and drop those which have less weights. In the process of dimensionality reduction, we reduced our features from 48 features to 7 features, including origin airport, destination airport, departure time, arrival time, actual elapsed time, air time and distance. To create more customized models for different carriers, we decided to group the data by airlines and mainly focused on two biggest airlines which are United Airline (UA) and American Airline (AA). In addition, we delineated arrival delay and departure delay. For United Airline, we were able to obtain 70.07 % and 60.98 % accuracies for arrival delay and departure delay, respectively. The accuracies for AA are slightly lower than UA which are 63.86 % and 65.59 %.

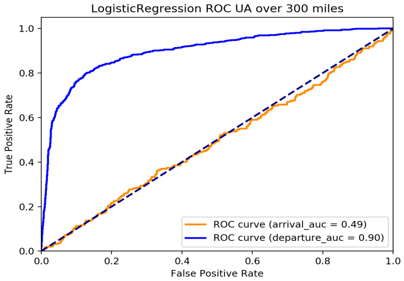
The issues we faced with when implementing the SVM model is that some instances have missing values for the certain features. We solved it by calculating the average for that feature and replacing all the missing values with the average. This could affect our correctness of the models. In terms of the comparison with logistic regression, SVM is able to achieve higher accuracy. This is what we expected since SVM works better in muti-dimentional feature spaces. However, it took much more time to train and predict the instances.

Logistic Regression

A logistic regression was selected as our second model. It provided much more rapid analysis than the SVM. We utilized this to develop granular analysis of our datasets. Initially, we refined our dataset down to mirror that of the SVM. This was done for celerity and uniformity. As with the SVM, we again elected to delineate between arrival delay and departure delay, we built Receiver Operating Characteristic (ROC) curves for each dataset. While these took in probabilistic estimates from our logistic regression, we also found accuracy, precision and recall scores on binary estimates from our logistic model. Further separation was done by screening our datasets down to airline and travel distance combinations.

Results from the logistic regression model were both surprising and informative. First, while accuracy between the models held even across airline and distance sets at around 85%, the ROC curves as well as precision and accuracy scores pointed to a problem. Since the incidence of delay was so low in the arrival category, we received all negatives in our binary output and precision and accuracy was zero for this dataset. For departure delay sets, precision and accuracy was around 80%. The ROC curve, while using a probabilistic estimate instead of a binary one, had similar projections. The Accuracy Under the Curve (auc) for departure delay was around 80% but the auc for arrival delay straddled the dividing line at 50% and even went negative into the region of greater false positive than true positive for United Airlines flights of under 300 miles.

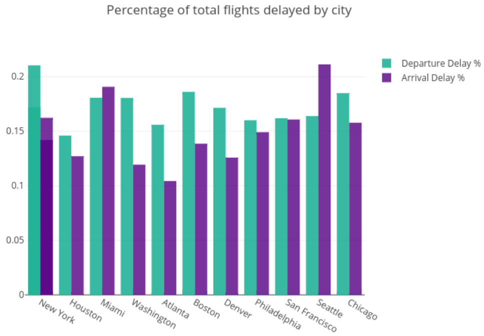




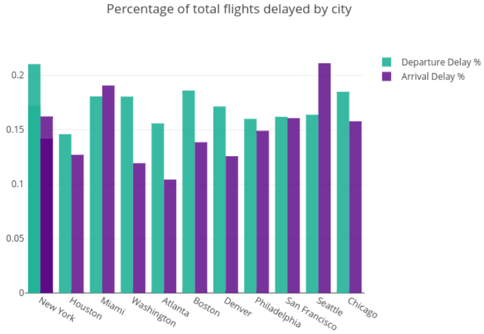
Having developed a set of results for our logistic regression, we developed a number of theories as well as strategies to fix the issues we had encountered. First, there was a strong difference between arrival and departure delays. Further, valid results were only really achieved from the departure delay set. Additionally, Eeven though the logistic regression was faster and achieved higher accuracy it was sensitive to class balances in a way the SVM was not.

1.2 Conclusions

Further analysis of the data indicates that the size of the city the airport is based in may be a more important factor than the overall air traffic of an airport. This is evident when comparing the total traffic handled by a given airport:



with the actual number of flights delayed as a percentage of total:



This is likely related to the number of flights that terminate in a destination versus those that are connective flights through an airport. Airports with a higher percentage of connecting flights might see less delays even if the overall number of flights is higher - a result of connecting flights getting higher priority from air traffic control. Given that our data consisted of a mix of continuous and qualitative features, the improved performance of the linear model over the SVM indicates that the continuous features are more important than qualitative features.

1.3 Citations and References

Acknowledgments

@TechReport, author = A. Sternberg, J. Soares, D. Carvalho, E. Ogasawara, title = A Review on Flight Delay Prediction, institution = CEFET/RJ, year = 2017, address = Rio de Janeiro, Brazil,

@Article, title= Which airline should you fly on to avoid significant delays?, author= Department of Transportation, year= 2017,

@Article, title= 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset, author= Jason Brownlee, year= 2015,

@Article, title= Support Vector Machines vs Logistic Regression, author= Kevin SVersky, year= 2015,

@misc, title= Quandl.com, author= NA, year= 2016,

@misc, title= SEC.Gov, author= NA, year= 2016,

@misc, title= stackexchange.com, author= NA, year= 2017,

CIS 519 Intro to Machine Learning: Flight Delay

Problem

Flight delays are affected by many factors and hurt airlines and passengers.
We aim to understand the feature space of flight delays as well as develop application to predict flight delays.

Approach

Datasets Collection and Clean Up- Flight_data2016.csv, Airlines.csv and Airports.csv from Kaggle.

Front End Setup- Flask

Feature Assessment and Dimensionality Reduction from 42 to 7 features

Model Building – Logistic Regression and Support Vector Machine with Gaussian kernel

Application

Flask with data visualization
SVM with Gaussian kernel and Logistic regression mode for predicting American Airline and United Airline

Web Application:

Flight Delay

Welcome to Flight Delay, an application showcasing how likely you are to see your flight delayed at various airports.

Findings: local population most correlates to flight delay

Penn Engineering

Professor: Eric Eaton
Team Member: Sam Weinberg, Kaushik Visvanathan, Tse-Lun(Paul) Hsu

Results Across Models

Support Vector Machine (with Gaussian Kernel)

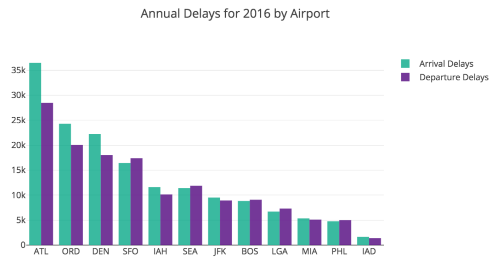
Accuracy	Airline	ARR/DEP
0.638692353593	AA	ARR_DELAY
0.700707964602	UA	ARR_DELAY
0.60982300885	UA	DEP_DELAY
0.655911686983	AA	DEP_DELAY

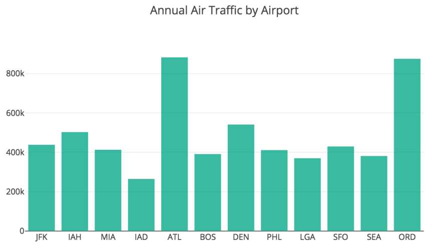
Logistic Regression

Accuracy: 82% Average

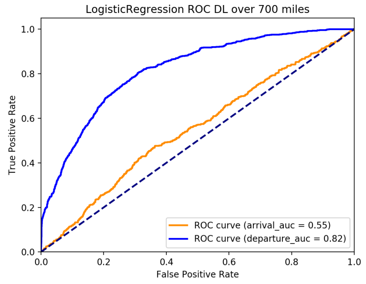
Flight Delay

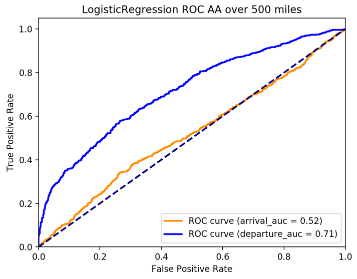
Welcome to Flight Delay, an application showcasing how likely you are to see your flight delayed at various airports.

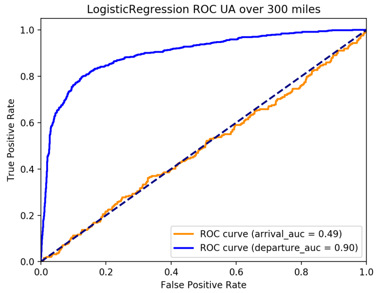

Annual Delays for 2016 by Airport


Annual Air Traffic by Airport

Logistic Regression ROC Curves


LogisticRegression ROC DL over 700 miles
True Positive Rate
False Positive Rate
ROC curve (arrival_auc = 0.55)
ROC curve (departure_auc = 0.82)


LogisticRegression ROC AA over 500 miles
True Positive Rate
False Positive Rate
ROC curve (arrival_auc = 0.52)
ROC curve (departure_auc = 0.71)


LogisticRegression ROC UA over 300 miles
True Positive Rate
False Positive Rate
ROC curve (arrival_auc = 0.49)
ROC curve (departure_auc = 0.90)