Predicting departure delays of US domestic flights

CSE 258 Assignment 2

Scott Cole & Thomas Donoghue
University of California, San Diego
srcole@ucsd.edu & tdonoghue@ucsd.edu

Abstract

Delays in air travel can be very costly to both passengers and airlines. While many delays due to weather or mechanical failures are unpredictable, it may be possible to predict whether a flight will be delayed based on statistics of past flights. Here we train a logistic regression model to predict if a flight will be delayed by more than 15 minutes. The model was trained using features of the flights known at the time of booking such as the airline, month, week, and hour of departure. The best algorithm trained separate models for each airport and achieved an accuracy of 0.689 (area under the receiver operating characteristic curve). Additional features about the graphical structure of the flight network did not improve performance. While the prediction success is limited due to the lack of knowledge of the weather and the recent status of the flight network, its advantage is that it does not require information immediately before the flight.

Dataset Description and Provenance

In order to train a model to predict flight delays, we acquired data collected by the U.S. Department of Transportation's Bureau of Transportation Statistics of all domestic flights during 2015. This data set, posted by a Kaggle user, contained around 5.8 million flights by 14 airlines that flew among 322 airports. For each flight, the data set included the scheduled dates and times of departure and arrival as well as the actual times of departure or arrival. Additional information about the flight included the airline, origin and destination airports, distance traveled, and duration of taxi from the gate before departure and to the gate after landing.

Some cleaning of the data was required before analysis. For example, two different code systems were used to indicate the origin and destination airports, so the data set was cleaned to use only the 3-letter airport code. Additionally, two codes were initially used for the Austin airport. We restricted the dataset to include only airports from which an average of at least 20 flights departed were included in the final data set (98 airports) in order to restrict the analysis to larger airports. After the data cleaning, about 5.25 million flights remained in the data set.

Exploratory analysis of factors related to flight delay

Prior to creating a predictive algorithm, the data was analyzed to see if and how the metadata of a flight correlates to the delay of its departure. Particularly, we tested how the delay was related to: 1) the flight month, 2) day of the week, 3) time of day, 4) duration, 5) the airline, and 6) the origin airport of the flight.

Because seasonal changes in airport traffic and weather can affect the congestion and flow of airplanes, we expected flight delay to vary systematically throughout the year. For example, more flights flew in July as compared to September (9% and 8% of the total flights, respectively) and so we hypothesize the increased traffic will relate to increased delays in July. **Figure 1A** shows, for multiple airports, that flights are more delayed in the summer months (June - August) compared to the spring and fall. Delay also tends to increase in November and December, likely a combination of increased vacation traffic, and weather, which probably also relates to increased delays in the winter months that disproportionately affects the airports further north (compare, for example, LGA, in New York, in yellow, to SAN, in San Diego, in purple).

The amount of flight traffic also varies over the course of the week, in particular around the weekend, with there being more domestic flights on Fridays (15.0% of weekly flights) and Sundays (14.7%) as compared to Saturday (12.0%). Consistent with our hypothesis regarding airport traffic predicting delays, we observe (Figure 1B) that flights are most delayed right before the weekend (Thursday and Friday) and at the end of the weekend (Sunday and Monday) compared to the middle of the weekend (Saturday). On multiple timescales, we see that traffic predicts delays, where traffic is correlated with both daily and monthly trends, and so we expect these temporal features (day and month) to be useful predictors in our model.

Relatedly, we hypothesized that scheduled departure time would also relate to delay, not only due to traffic, but due to knock-on effects, thus we hypothesize increasing delay with later departure times. This is also observed (Figure 1C), as we see that early morning flights are the least delayed, and that the average delay increases roughly linearly from 5AM to 6PM from 2 minutes to 15 minutes. The average delay then decreases (again, roughly linearly) down to about 5 minutes at 2AM. The later departures around 3AM and 4AM tend to be more delayed, but this estimate is rough, as these flights make up < 0.01% of all flights (reflected by the large error bars). Given these trends, time of day is expected to be a strong predictive feature of delay.

In addition to details about the time of departure, we considered that the duration of a flight may relate to the delay period. Indeed, **Figure 1D** shows that while domestic flights less than 6 hours long average only about 10 minutes of delay, longer flights have longer delays, with 8-9 hour flights having an average delay of 20 minutes. Because the average delay as a function of flight duration is very nonlinear, one-hot encoding will be useful to encode flight duration as a predictor of delay.

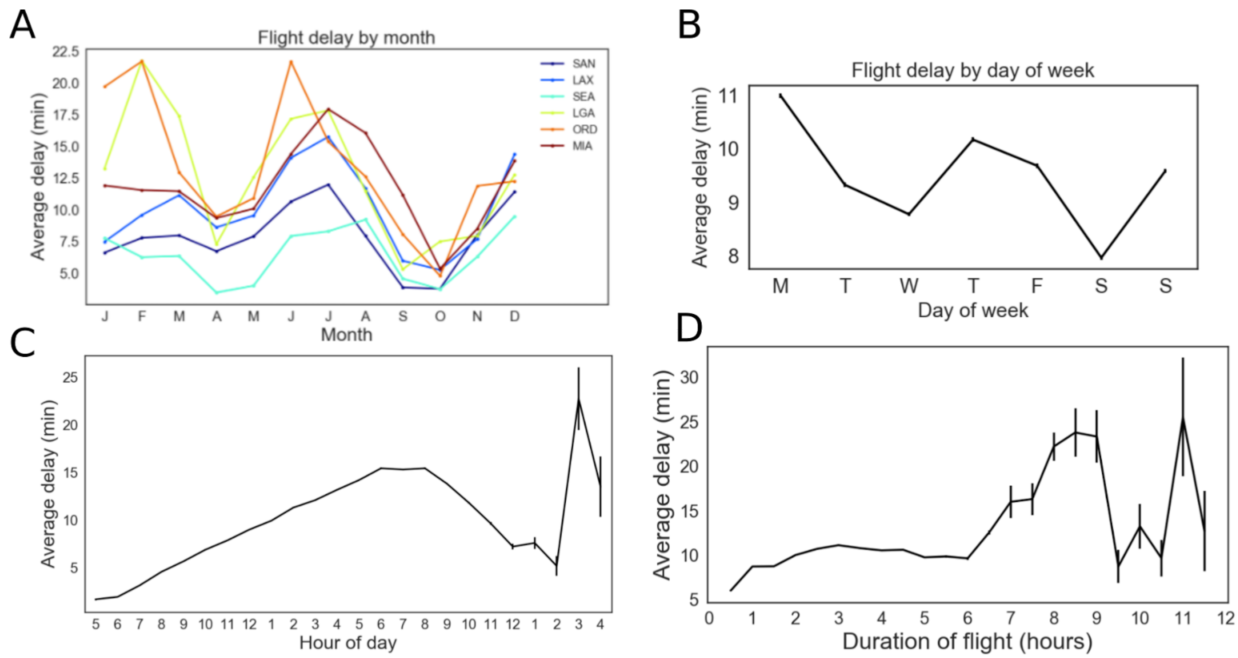

Figure 1. Relation between time of flights and their departure delays. **A.** The average delay of flights from 6 different airports (colors, see legend) over the 12 months of the year. **B.** Average delay of all domestic flights on each day of the week. **C.** Average delay of all domestic flights during each hour of the day. The axis starts at 5AM on the left and ends at 4AM on the right. **D.** Average flight delay as a function of the duration of the flight, binned by each half hour. Error bars in B-D denote standard error of the mean (SEM). Note that most error bars are too small to see.

Network analysis of domestic flights

As well as the flight data, the dataset inherently contains some information about the airlines and airports. To describe the data, and also as a potential source of predictive features we conceptualize the set of airports, and the flights between them as a network, and applied basic graph analysis. Each airport is a node, each flight path an edge, weighted by the number of flights (**Figure 2**). This graph has 98 nodes (airports) and 1614 edges (flight paths).

From this representation, we calculated descriptive features of the network, including node connectivity, centrality, and clustering. Additionally, we computed the average degree of all neighboring nodes and the shortest path length between nodes. Collectively, this analysis describes a highly connected network, as can be seen in **Figure 2**, with high average node connectivity (19.81), low average shortest path length (1.67), and moderate average clustering (0.77). The node specific measures were saved to be used as potentially predictive features.

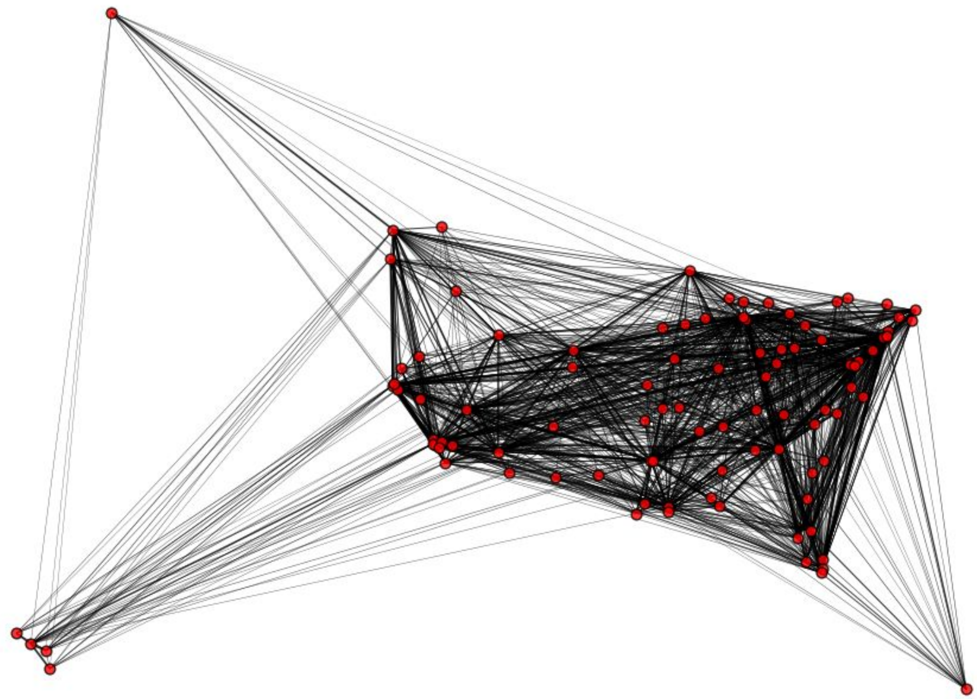

Figure 2. Network representation of the flight data using geographical coordinates, with undirected, weighted edges.

Variance in airlines and airports

In addition to deciding when to fly when booking a flight to minimize the chance of delay, the user also may have the option of several airlines. Therefore, it would be useful to know which airline is best, in terms of minimum delay. This choice may be a function of the airport from which a flight is departing. This could make sense because airlines are nonuniformly distributed across airports, with different airports serving as hubs for different airlines (1). We confirmed from the dataset that airlines do in fact use hubs, operationally defined as the most frequent departing airport for each airline, and quantify there 'hubness' as the percent of total flights an airline flies from its hub airport. For example, Delta airlines use the Atlanta airport as their hub, with 27.7% of their flights departing from there, as compared to Southwest Airlines, who use a distributed flightpath model, only having 6.6% of flights from their most common airport (Chicago Midway). Hypothesizing that airlines may be more (or less) efficient at their hubs, we kept these features for prediction, using the 'hubness' of an airline, and a binary encoding for if the current flight is departing from the hub airport of the current airline.

We also computed the average delay for each airline (across all flights) as well as their average delay at each airport. **Figure 3** shows how the average flight delay across all airlines varies (black line) from less than 1 minute with Hawaiian Airlines (HA) to more than 16 minutes with Spirit Airlines (NK). This ranking of airlines by delay was roughly consistent when analyzing only flights originating from the San Diego airport (SAN, red line). However, this ranking was not conserved for the flights originating from the JFK airport in New York (blue line). Because each airport may differ in the relative ranking of these airlines, this motivates using separate models for each airport when predicting flight delay, if airline is used as a feature.

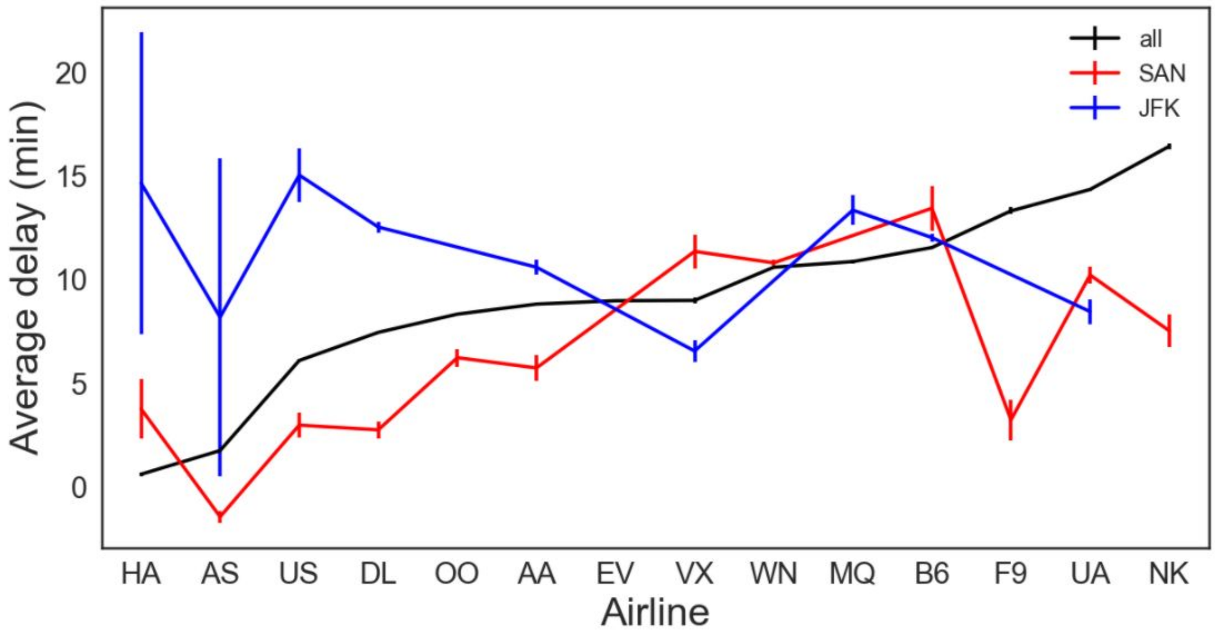

Figure 3. Average flight delay for each airline, aggregated across the entire data set (black), only the San Diego (red), and only the JFK airport (blue). Error bars denote standard error of the mean (SEM).

Additionally, **Figure 3** hints that the average delay may not be the same across all airports. Therefore, we calculated the average delay for each individual airport and plotted a subset of the results in **Figure 4.** This plot demonstrates that delays are indeed a function of airport, for example flights are considerably less delayed in Knoxville (TYS, average delay = 2.6 minutes) as compared to Salt Lake City (SLC, average delay = 11.9 minutes).

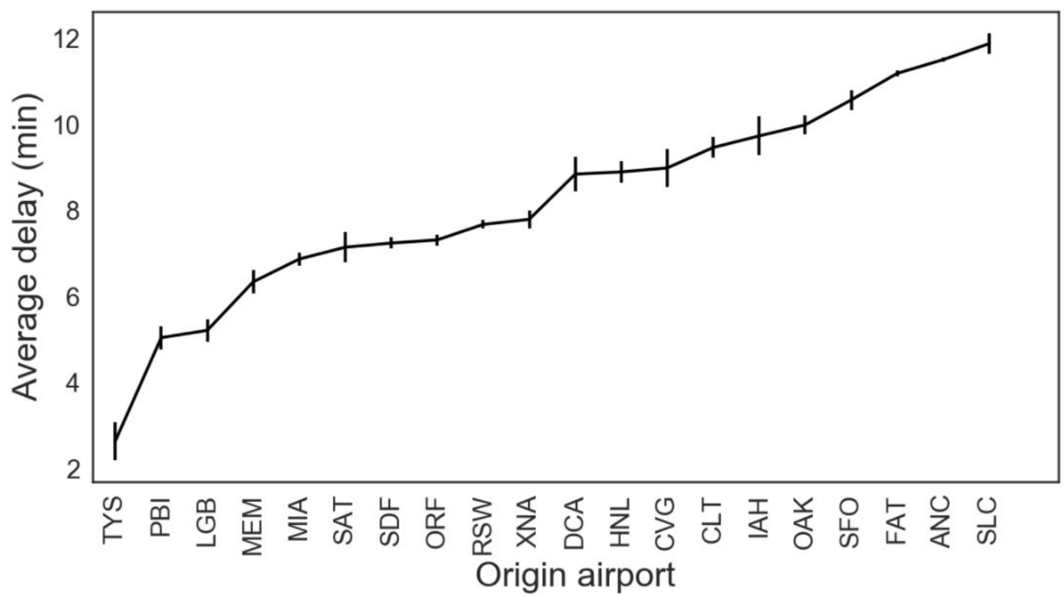

Figure 4. Average delay of flights departing from 20 randomly chosen airports. Error bars denote standard error of the mean (SEM).

Classification task

Our goal is to correctly predict whether a flight's departure will be delayed, where delayed is operationally defined as the actual departure being at least 15 minutes later than scheduled. Therefore, classification models were applied in order to discriminate negative examples (flight not delayed) from positive examples (flight delayed by at least 15 minutes).It is important to note that this is an unbalanced dataset, with a median departure delay of -1 minute (i.e. the plane leaves the gate 1 minute before the scheduled departure time), and only 18.04% of all flights being delayed by at least 15 minutes. Therefore, a raw accuracy measure may be misleading, because a naive classifier that solely predicts that every flight is not delayed achieves an accuracy of about 82%. Though this accuracy is higher than a coin flip, this model is uninformative.

Therefore, we chose to evaluate models by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC). This is possible for models that output a probability of a training example corresponding to each class, such as logistic regression. From this property, we can choose a cutoff probability other than 0.5 in order to manipulate our true positive and false positive rates. For this evaluation, an uninformative model will have AUC = 0.5, while a perfect model will have AUC = 1.0.

For evaluation, the entire data set was randomly split into a training set with 70% of the flights (3.7 million flights) and the remaining 30% of the flights were in the test set (1.6 million flights). Note that no validation set was used. This was because a wide range of regularization parameters for our logistic regression models did not seem to have any effect on the model training. Additionally, the models performed roughly equally well on the training set and the validation set, showing that they did not overfit to the training data. This lack of a need for precise regularization may be because our data set is very large compared to the size of our feature space.

A simple baseline model for comparison uses only the flight number, predicting the delay of a flight by checking how often that flight has been delayed in the past. For each flight number in our training set, we calculated the fraction of flights that were delayed by at least 15 minutes (e.g. flight AA1024 from ABQ is delayed with a probability of 0.27). This model provided predictions slightly better than chance (AUC = 0.505 on the test set).

We designed the features for the predictive model based on the patterns seen in the exploratory analysis above. One-hot encoding was used to transform the following flight details into useful features: 1) month, 2) day of week, 3) origin airport, 4) airline, 5) duration, and 6) hour of departure. For encoding, the hour of departure was rounded down (e.g. flights departing at 6:00AM through 6:59AM were assigned to the 6AM bin). For duration, the flights were binned into hours (i.e. 0-1 hours, 1-2 hours, 2-3 hours, etc.).

The importance of each feature set was evaluated by adapting the logistic regression model in two ways. First, logistic regression models were trained using only one feature set (e.g. only the one-hot encoding for day of week). The AUC measure was computed on the test set for each model and compared to a 0.5 value to judge if the model using this feature set performed better than chance. Second, logistic regression models were trained using all feature sets except one of them, and the performance of this model was compared to the model that included all feature sets. If removing a feature set made little or no difference in the AUC measure on the test set, it was concluded that this feature set was redundant to the other feature sets and did not add any additional information for predicting flight delay.

Model description

Logistic regression was chosen to model flight delay for multiple reasons. First, the weights of each feature trained by logistic regression are easily interpretable, as the sign of the weight indicates if a flight is more or less likely to be delayed if it has a high value for that feature. Second, logistic regression outputs a measure of confidence in its output through the probability of belonging to each class. This allows us to calculate the AUC measure mentioned above, as opposed to if the only output was a label.

It was important that the model does not assume independence between features (like the Naive Bayes method) because there exist correlations between the features described above. Most notably, the distribution of airlines is not uniform across airports (e.g. Hawaiian Airlines flights are common in Honolulu but not in Atlanta). Additionally, the duration of flights may be biased across airlines that fly different distances. For example, the average duration of a flight on Virgin America is 210 +/- 0.5 minutes (mean +/- standard error of the mean) compared to 105 +/- 0.5 minutes on Hawaiian Airlines.

Using the logistic regression model above, we did not have any issues of overfitting, as judged by the roughly equivalent accuracy on the training and testing sets. Therefore, there was no need to optimize the regularization of the model. Additionally, there were no issues of model scalability, as training the model with all features using the entire training set took at most a few minutes on a laptop.

In addition to logistic regression, two other methods for training models were applied: random forests and support vector machines (SVMs). All models were implemented using scikit-learn. Despite varying multiple hyperparameters for the random forest algorithm, the model continuously overfit the training set and did not outperform the logistic regression model on the test set. Training SVMs led to scalability issues, as the model took very long to train on a laptop when using more than a million samples in the training set.

In addition to a single logistic regression model, we also split up the data set by origin airport (98 total airports) and trained a model for each airport. This was motivated by the result shown in **Figure 3**, in which a feature set (in this case, one-hot encoding of airline) may warrant different weights in order to well predict flight delay at different airports. For each airport, 70% of flights were chosen randomly for the training set, while the remaining flights were used in the test set. A single AUC measure was then computed across all the tests sets from all airports and compared to the accuracy obtained using a single logistic regression model. An improvement in flight delay prediction by training separate models for each airport would be indicated by an increase in AUC. In addition to the potential improvement in predictive accuracy, training many small models for each airport is actually about 30% faster than training one model on all training examples.

Related work in predicting flight delay

The data used here, collected by the US Bureau of Transportation Statistics, has also been analyzed by other data scientists. Several projects have been published online with similar goals of predicting flight delays, but their methods and data set sizes were diverse. Although these models are informative for our problem, they are solving a slightly different problem, so not all aspects of them are relevant. Our problem setup assumes trying to predict delay some time significantly prior to the departure date (for example, at time of booking), at which point detailed information about the weather and previous flight delays are unavailable.

Several groups have tried different models and other data features to predict flight delay. For example, one group used an artificial neural network (ANN) to predict flight delays at JFK airport (2). The main development of this paper is towards using ANNs with interpretable nominal features, but the small training set (~1000 flights) from a single airport, and the evaluation of this model making continuous predictions for only 5 flights make it quite difficult to compare to our approach.

A related project explored using Naive Bayes, SVMs and Random Forests to predict flight delay as a categorical variable, using a larger dataset (multiple years of data) of about 135 million flights (3). They also used weather data to predict flight delays. Overall, the results were not very promising - all three methods performed about equivalently, but with error rates that were not clearly much better than a trivial predictor, given the unbalanced dataset (also evidenced by poor recall performance). There may well be significant improvement possible with feature design (as well as model choice), but there is scant description of their feature design to be able to evaluate what to change from their approaches.

Another project used a similar dataset from the Bureau of Transportation Statistics, comprised of around 7 million flights during 2007 and 2008 (4). As in our analysis, they operationalized a categorical delay label as being 15 minutes late. Using a subset of the data (flights from Chicago O'Hare airport), they used regression on small subsets of the data in order to guide feature selection for classification models using random forests and SVM. This group also used day-of features to predict flight delays, including weather data and text data scraped from Twitter. They report that the Random Forest performed better, compared to the computationally expensive SVMs. The final results were again unbalanced - while the model was relatively successful of predicting which flights were not delayed (91% of non-delayed flights correctly predicted), it was much less successful at predicting delayed flights (only 41% of delayed flights correctly predicted) corresponding to a precision and recall of 0.66 and 0.37 respectively.

A final related project, by Microsoft Research (5) as a demonstration of the Cortana analysis platform, used approximately half a year of flight data (April to October 2013) to predict a binary delayed label, operationalized as being 15 minutes late. They also scraped and added weather data for their predictions. They tried two classification models, a logistic regression model, which performed with AUC of 0.675 and boosted decision trees which performed slightly better with AUC of 0.697.

Collectively, the prior literature gives us a basis to start from, but suggests that this is very much an unsolved prediction problem, in particular in our use case of not using day-of features, with a large dataset. One of the main difficulties stems from the unbalanced data, it being fairly difficult to increase overall accuracy above trivially predicting all flights to be on time. Random Forests and SVMs have been tried several times, but often with limited success (as we replicate). Neural networks show promise, but are much more complicated to work with and have only been shown on much smaller datasets than we are using here. Our approach is to instead use the simpler model of logistic regression, putting most of the focus into feature design. Since prior work used different datasets and a variety of performance measures, comparison between them is difficult, but it would appear that the 'gold standard' to beat is from Microsoft.

Results: accuracy of flight delay prediction models

Each classifier trained to predict the delay of flights was evaluated using the AUC on the rest set. The performance of all classifiers trained without using graph-based features are shown in **Table 1**. Training a logistic regression model using any one of the 6 feature sets previously described (month, day of week, airline, departure hour, origin airport, or flight duration) outperformed the baseline predictor trained only with flight numbers. Combining all features into a single model resulted in a test AUC of 0.672.

Using flight duration and day of week only slightly resulted in model predictions slightly above chance (test AUC ≅ 0.52). Including features for month, origin airport, or airline was better (test AUC ≅ 0.56), and the most informative feature was the hour of day that the flight departed (test AUC ≅ 0.63). Consistent with these results, the test AUC was only mildly decreased when removing any of the first five feature sets (test AUC > 0.658). However, removing the feature set encoding the hour of departure significantly lowered the test AUC to 0.611. Because the test

AUC only decreased by 0.001 when removing either the duration or day of week feature sets, we can conclude that these features are redundant with the others in the model.

As shown in the final line of **Table 1**, training separate models for each airport further increased the test AUC to 0.689. Therefore, our ability to predict whether or not a flight would be delayed by at least 15 minutes was improved by considering the features separately for each airport. Notably, this accuracy is higher than the previously mentioned logistic regression model by the Microsoft Research team (AUC of 0.675), and is only slightly lower than their best performing boosted decision tree (AUC of 0.697), although notably both of these models also included additional weather information.

	Features	Train AUC	Test AUC
Baseline	Flight number		0.505
One feature set	Duration	0.518	0.518
	Day of week	0.518	0.518
	Month	0.565	0.565
	Origin airport	0.560	0.560
	Airline	0.568	0.568
	Hour of day	0.628	0.629
All but one feature set	Duration	0.671	0.671
	Day of week	0.671	0.671
	Month	0.658	0.658
	Origin airport	0.667	0.668
	Airline	0.662	0.662
	Hour of day	0.611	0.611
Single model	All features	**0.672**	**0.672**
Separate model for each airport	All features	**0.691**	**0.689**

Table 1. Performance of various logistic regression models to predict flight delay.

Feature weights interpretation

After training the single linear regression model on all feature sets, we analyzed the values of the trained coefficients. **Table 2** shows the most extreme coefficients for each set of features and the particular feature associated with those coefficients. Because the feature values are boolean, positive weights correspond to an increase in the probability of a flight delay when a flight has this property, relative to whichever category was left out of the one-hot encoding. For example, a flight being operated by Spirit Airlines will be predicted to have a greater probability of being delayed compared to any other airline because this airline obtained the highest weight after one-hot encoding. In contrast, if the flight were operated by Alaska Airlines, its probability of delay would be decreased because the weight corresponding to this feature is highly negative.

The most extreme feature weights shown in **Table 2** are all consistent with the exploratory analyses shown in **Figures 1-4**. As expected from the exploratory analysis, the weights for flight duration, hour, day of week, and month are significantly nonlinear when mapped to a linear axis (i.e. January is 1, February is 2, etc.). Therefore, the one-hot encoding was able to extract significantly more information from these features than would have been possible using simple linear encoding or any type of monotonic encoding (e.g. logarithmic transform).

Feature set	Lowest weight	Feature	Highest weight	Feature
Duration	0.08	< 1 hour	0.93	7-8 hours
Day of week	-0.10	Saturday	0.11	Monday
Month	-0.64	October	0.20	June
Origin airport	-0.02	Westchester County (HPN)	0.17	Lihue, Hawaii (LIH)
Airline	-0.84	Alaska Airlines (AS)	0.32	Spirit Airlines (NK)
Hour of day	-1.46	5AM-6AM	0.60	3AM-4AM

Table 2. Most extreme feature weights within each feature set.

Accuracy interpretation

While a raw accuracy measure isn't very informative due to the nonuniform distribution of all flights among the two classes, the AUC measure isn't perfectly informative either. Though the AUC gives a measure of the tradeoff of the true positive and false positive rates (the two axes of an ROC curve), the user may still be interested in other performance measures of a classifier, specifically precision, recall, and specificity (a.k.a. True negative ratio, TNR). **Table 3** shows these metrics of performance in addition to the AUC and accuracy measures for the two baselines and logistic regression models (together or separate for each airport). All models have roughly the same accuracy scores when applying a cutoff of 0.5 (i.e. a flight is predicted as delayed if the probability of delay exceeds 50%).

In particular, a user may be particularly interested in the recall measure, which is the fraction of delayed flights which are correctly predicted to be delayed. The naive predictor that always predicts that a flight is delayed has a perfect specificity, but achieves 0 recall because it never predicts a positive result. The logistic regression model that is trained across all airports only achieves slightly improved recall (0.0004) at a small cost to the specificity (0.9999). Adding graph-based features to the model slightly increased the recall but did not significantly improve classification. Separating the models by airport significantly improved the recall, but it remained low (0.01). Further improvements to recall could be gained by lowering the threshold on the probability output of logistic regression for which we predict that a flight will be delayed. As seen in the bottom rows of **Table 3**, decreasing the cutoff to 0.3 increased the recall to 0.27, so now we can now predict 27% of the flights which are delayed. However, the precision now drops to 37%, indicating that we are now predicting more false positives than true positives (i.e. a flight that is predicted to be delayed is actually most likely not going to be delayed). If the cutoff is further lowered to 0.2, the model achieves a recall of 0.60. The precise cutoff to be used will depend on the priorities of the user, and what their relative penalty is for false positives compared to false negatives.

	AUC	Probability cutoff	Accuracy	Precision	Recall	TNR
Baseline - always predict not delayed			0.82		0	1
Baseline - flight numbers	0.505	0.5	0.81	0.20	0.02	0.98
Single logistic regression model	0.672	0.5	0.82	0.54	0.0004	0.9999
Single logistic regression model with network features	0.672	0.5	0.82	0.54	0.0006	0.9999
Multiple logistic regression models	0.689	0.5	0.82	0.54	0.01	0.99
		0.3	0.79	0.37	0.27	0.90
		0.2	0.66	0.29	0.60	0.67

Table 3. Additional performance metrics for flight delay classifiers.

From **Table 3**, we conclude that our precision and recall performance is worse than some previous work (4), although this is to expected given that this work only used data from a single airport. Additionally, the scope of our model differs from that of this report (and many others), in that we did not use weather data. Our goal is to create a model to help users predict whether a flight will be delayed at the time of purchasing a ticket, which may be several months before the flight, and so accurate weather forecasts will not be available. Because weather is one of the major causes of flight delays, and weather reports strongly correlate with flight delay (6), it is reasonable that our models were not capable of predicting as many delayed flights as other approaches that used this additional information..

Conclusion

Overall, our models are only of limited utility since none were capable of correctly predicting flight delays with both precision and recall greater than 50%. This seemingly low performance is likely due to the many causes of flight delays being outside the scope of our data. It is unclear if it is even possible to predict whether or not a flight will be delayed so far in advance, as we have set up the problem, because so many of the causes of delays (e.g. mechanical issues and weather) cannot be known in advance. Despite this, we were successful in creating models that outperform baseline models, and perform at least about as well as prior work, even when we often use less information, and generalize to more airports. Although imperfect, this model still makes potentially useful predictions about which flights are more or less likely to be delayed. Future work may well be able to further improve this kind of flight delay prediction at time of booking, perhaps by further work on feature design and collecting other informative features about flights, and/or work on more sophisticated modeling techniques.

References

1. Sarkis J (2000) An analysis of the operational efficiency of major airports in the United States. *Journal of Operations Management* 18(3):335–351.
2. Khanmohammadi S, Tutun S, Kucuk Y (2016) A new multilevel input layer artificial neural network for predicting flight delays at JFK airport. *Procedia Computer Science* 95:237–244.
3. Lawson D, Castillo W (2012) Predicting flight delays. *Stanford CS229*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download.pdf.
4. Huang S, Dowey D, Tashman M (2014) Airline Flight Delay Prediction. *Airline Flight Delay Prediction*. Available at: http://ddowey.github.io/cs109-Final-Project/#!index.md [Accessed March 13, 2017].
5. AzureML Team (2014) Cortana Intelligence Gallery. *Binary Classification: Flight delay prediction | Cortana Intelligence Gallery*. Available at: https://gallery.cortanaintelligence.com/Experiment/Binary-Classification-Flight-delay-predicti

on-3 [Accessed March 13, 2017].

6. Allan S, Gaddy S, Evans J (2001) Delay Causality and Reduction at the New York City Airports Using Terminal Weather Information Systems. *Lincoln Laboratory, Massachusetts Institute of Technology*.