Final Project - 36-462: Data Mining

Sivan Mehta, Mary St. John, and Graceanne Wong

5/12/2017

Introduction

The dataset at hand concerns flights to and from the Pittsburgh International Airport (PIT). Each observation is a single flight either entering or leaving Pittsburgh. While the predictive focus will be on departing flights, we will also look at flights overall to get a general structural overview of the dataset.

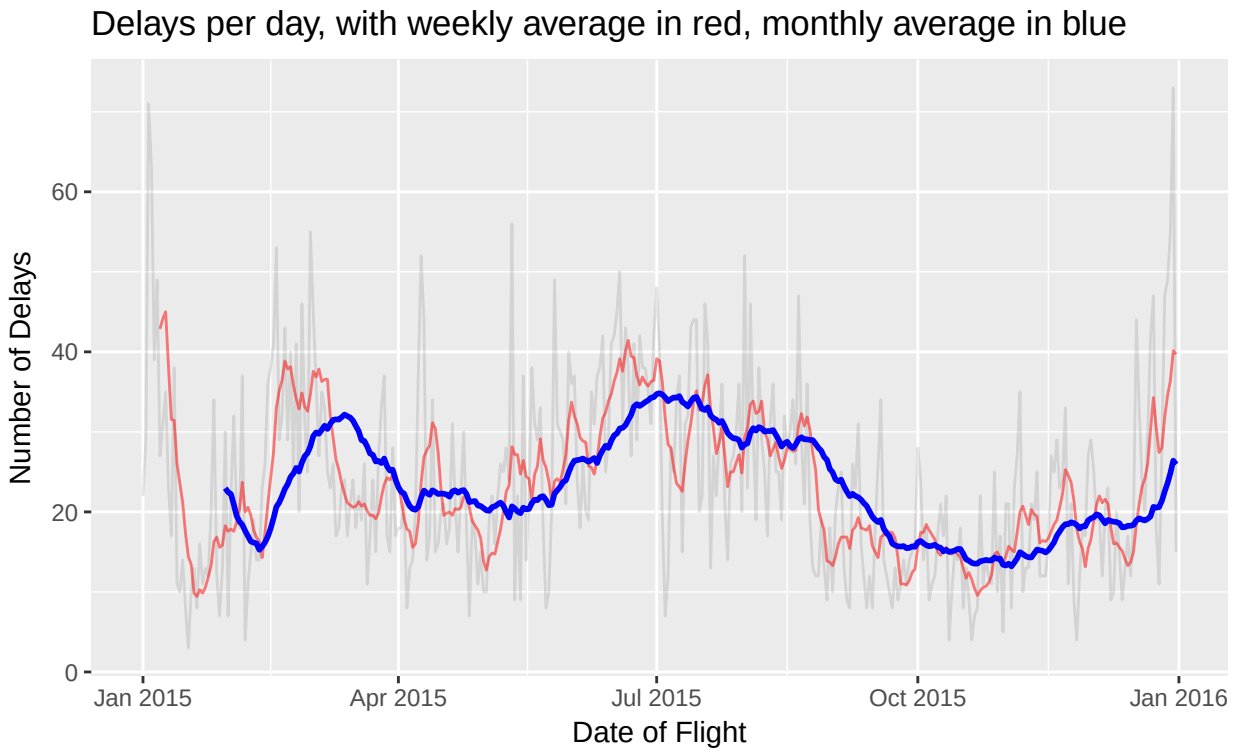
The 57 variables provided can be split into a few primary groups. Time Period information concerns the date of the flight. Airline information pertains to the specific airline and carrier of the flight. Origin information concerns origin airport details, and Destination information contains similar details. Departure and Arrival performance information contains metrics on the delays, taxiing, etc. Finally, there are general delay metrics, detailing the breakdown of what contributed to the overall delay, if any.

To predict flights, we will be working with three data sets with flight data from the Pittsburgh airport. Our training dataset is all the arrivals and departures from 2015. Our test data is arrivals and departures up to a random timepoint in the day. After that time point, we have unlabeled data on which we will form our guess dataset.

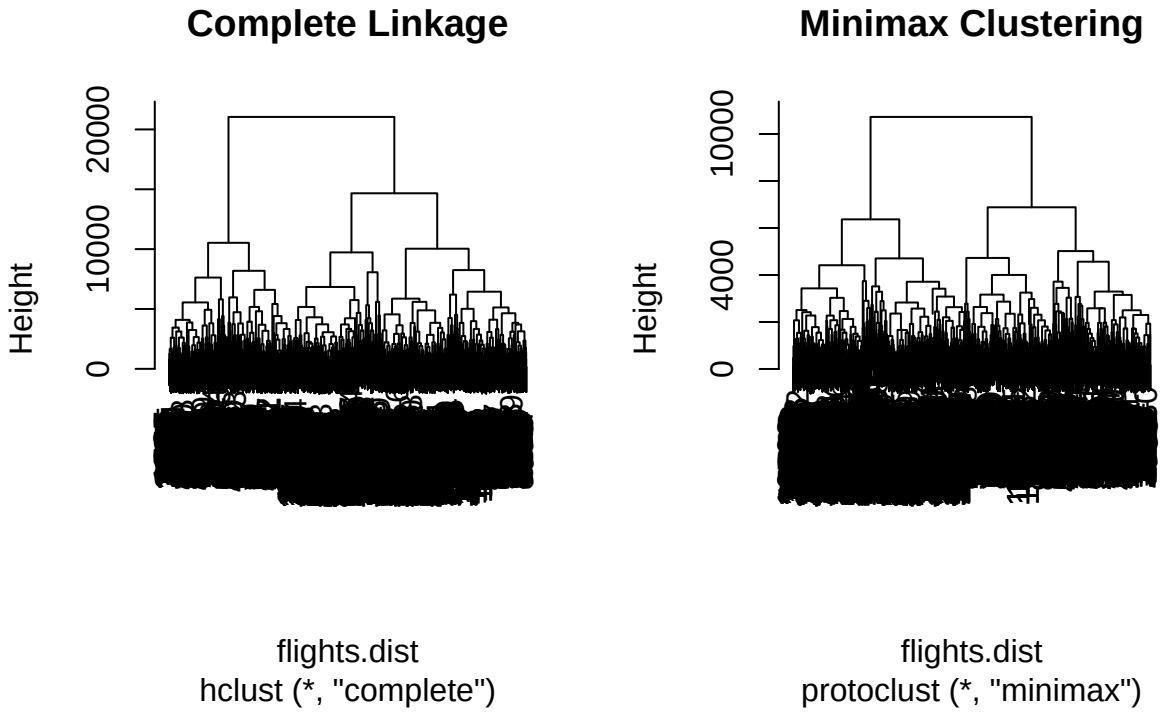
In this report, we will focus on subset *departing* flights and predict whether or not the flight will be delayed. We will first take a general unsupervised learning approach, attempting to ascertain any underlying structure to the dataset. Then we will move towards supervised analysis, attempting to actually predict whether or not a flight will be delayed.

EDA and unsupervised analysis

First, we'll examine delays on each day, in order to reduce *some* of the noise, we'll also plot some moving averages of delays.



Here we can see clear seasonal effects surrounding delays when looking at the moving average. While there is *a lot* of volatility as seen in the red weekly average, we can relatively smooth this with the blue monthly average. This tells us we can gain some insight with date information. Next, we can try clustering to see if the flights break down into meaningful groups. This perhaps could give us some insight on the general types of flights we may see.



Generally, these are doing pretty poorly, garnering misclassification rates of nearly 50%, **much** worse than the base rate of 16.46%, when using the assigned clusters, at almost any level, as a classification. This rarity

of delayed flights makes hierarchical clustering problematic in that small groups of points are difficult. We also tried single linkage clustering because it has the ability to accommodate this behavior, but it performed much worse than either complete or minimax clustering.

Because we have found that flight delays are relatively rare, we hypothesized that the distribution of flights that were not delayed is different from the distribution of flights that were delayed. We utilized latent class regression as implemented in the `poLCA` package to predict the probability of class membership, using indicator variables for weather delay and NAS delay as predictors. In this method, we were able to cluster the data into two separate groups, and evaluate a density on each group.

From our latent class regression, we found that data had a probability of 0.9146 of being in class 1, and a probability of 0.0854 of being in class 2. These clearly correspond to delayed flights being class 2 and non-delayed flights being class 1. When conditioned on class membership, we found that data in class 1 had almost 0 probability of having either type of delay, but data in class 2 had less polar splits in probability.

Table 1: Weather delay class conditional probabilities

	Pr(1)	Pr(2)
class 1:	1.0000	0.0000
class 2:	0.8989	0.1011

Table 2: NAS delay class conditional probabilities

	Pr(1)	Pr(2)
class 1:	1.0000	0.0000
class 2:	0.5342	0.4658

In our evaluation, the overall probability for each data point was calculated by multiplying together the conditional probabilities of each predictor, and weighting by the class probabilities.

$$\sum_{c \in 1,2} P(X_{class} = c) \prod_{d \in weather.delay, NAS.delay} P(d = 1|class = c)^{X_d} \cdot P(d = 0|class = c)^{1-X_d}$$

Supervised analysis

Before we began our supervised analysis, we created new features in the data. After looking at the raw data given to us about the flights we guess on, we determined that there was not much information there. We built features based on previous events in the day, mostly focusing on the percentage of previous flights that have been affected by events that day. For each day, we calculated what percent of flights were delayed before the plane was scheduled to leave, and we did the same for weather delay, national air system delay, and delays on arrival flights. We also added a variable for the time that an outgoing flight was scheduled to leave Pittsburgh, or the time an incoming flight was scheduled to arrive in Pittsburgh.

Because percent of flights delayed at a given point is not very indicative of the rate of delays at a certain time, we created a feature based on the derivative of these ratios as a function of time. In calculating this, we kept the denominator of the ratio fixed and equal to the total number of flights scheduled for that day.

Mixture Models

We started with making predictions using a mixture model. The mixture model was trained on an indicator variable which was based on current truth, which is not available in the guess data set which is why we predicted that all flights are not delayed. When evaluated on our 2015 training data and our 2016 testing data, we had a training error of 12.02% and a testing error of 6.58% both of which beat our base rate.

However, when we evaluted on our prediction set, we were unable to generate the indicator variables for weather delay and NAS delay directly from the prediction set, so we had to rely on the data in the test set to generate these features. This resulted in predicting that all flights were not delayed.

Unused, but evaluated models

We had many other approaches to making a prediction model. We began with a random forest on all predictors to determine the most important ones by using a variable importance plot. We then put these predictors into mixture models, linear models, complete linkage clustering, prototype clustering, support vector machines, and random forests. These models proved to be at or near the base rate of classifying delayed and not delayed flights.

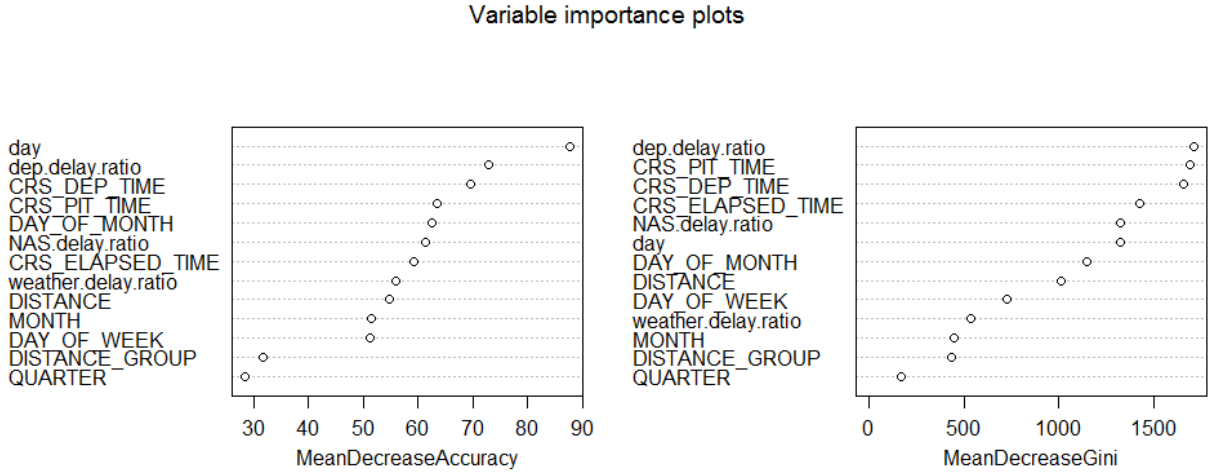
Table 3: Unused model summary

models	training.error	test.error	guess.delayed
base rate	16.46%	8.94%	
mixture models	12.02%	6.57%	0
linear model	15.6%	8.54%	0
complete linkage clustering	62.81%		
prototype clustering	51.25%		
svm	16.46%	8.94%	
random forest	16.99%	9.16%	14

For each model we first generated a training error. If it was close to or better than the base rate, we then generated a test error. If the test error was then better than the base, rate we then saw how many flights were guessed to be delayed in the submission dataset.

Resolution: Random Forests

We chose to use random forest because we are working in high dimensional space. Random forests can perform variable selection by choosing important predictors to split on and can also account for interactions between predictors. We used the ratio features we created, time variables, and distance variables. As can be seen in the variable importance plot, the most important predictors are the day, the percent of flights delayed on that day so far, and Pittsburgh time. These last two predictors are the same for departing flights, so they are correlated. Since random forest automatically accomodates interactions, this is not an issue



The training and test error for the random forest are both slightly greater than the base rate, but it at least makes a prediction that *some* flights are delayed. This model is better than the mixture model because it was not trained on variables that were created from the truth.

Conclusion

For our final model, a random forest, we guess that 14 flights would be delayed out of the 871 provided in the guessing dataset. We chose to use a random forest for these guesses as it was the only model that both provided useful predictions close to the base rate, but also did not guess the *same thing* for every single flight.

If we wanted to improve the dataset provided, there are several approaches we could have pursued. One approach would have been to expand the dataset beyond the city of Pittsburgh, perhaps picking up on trends in airports that could have indirectly affected Pittsburgh. We would then have more information into the underlying causes of flight delays. In the expansion, we could perhaps more easily track individual planes, and find a relationship between specific planes and delays.

Additionally, we could have expanded the breadth of the dataset by including Twitter data. Previous attempts at predicting delays[1] using this dataset have used volume of negative sentiment of the tweets to predict delays in airports. While it didn't predict it perfectly, it is an interesting approach that is worth taking a look at.

[1]http://ddowey.github.io/cs109-Final-Project/#!index.md