Data Mining Final Project Report - Team Galois

Eugene Han (eugeneh), Judy Kong (junhank)

12/7/2018

Introduction

Experiencing a delayed flight can be an especially frustrating experience. Since airlines do not reimburse for flight delays, the average delay can cost travelers hundreds of dollars in lost time and out of pocket expense. Using the Airline On-Time Performance Data made available through the Bureau of Transportation Statistics of the U.S. Department of Transportation, we analyze flight departures from the Pittsburgh International Airport in 2015 and 2016 to build a model to predict flight delays.

Exploration

Data Preprocessing

We first filtered the data such that we only viewed departing flights from Pittsburgh International Airport (PIT). We chose to combine the flight data from 2015 and 2016 since they both had very similar base rates. We used this combined dataset for training and tested on the visible 2017 dataset. In total, we had 50,918 observations for trianing and 13,588 observations for testing.

Table 1: Base rates.

	Percent No Delay	Percent Delay
training	87.0	13.0
testing	90.3	9.7

From Table 1, we see that the samples are highly imbalanced. Using a combination of manual inclusion, logistic regression, and stepwise regression with AIC penalty, the initial variables chosen to be important include: `QUARTER`, `MONTH`, `DAY_OF_MONTH`, `DAY_OF_WEEK`, `CARRIER`, `DEST`, `DISTANCE`, `CRS_DEP_TIME`.

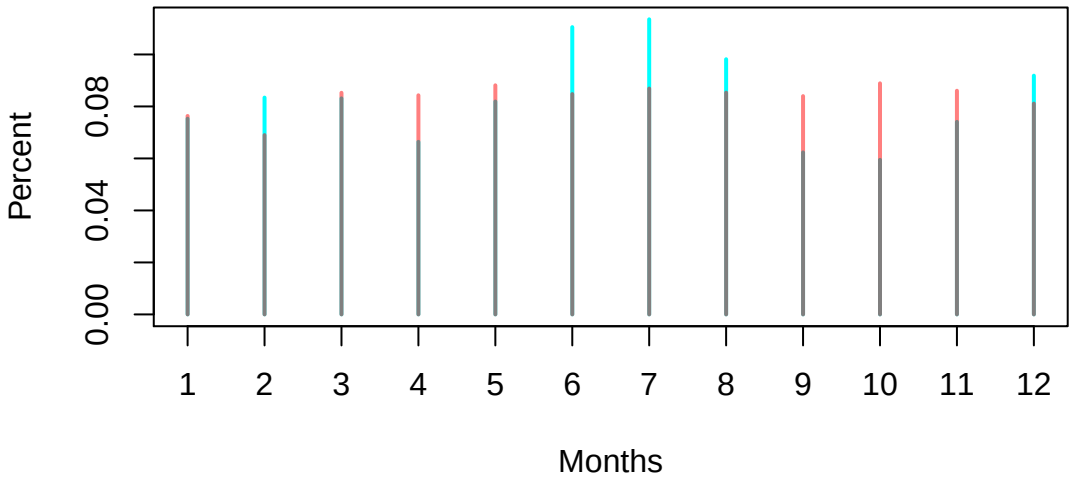


Figure 1: Distribution of the MONTHS variable for delayed and non-delayed flights. The cyan represents delayed flights and red represents non-delayed flights; the colors are overlayed.

Table 2: Summary of numeric variables.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
DISTANCE, no delay	182	402	526	750.5465	1060	2254
DISTANCE, delay	182	402	526	727.3774	994	2254
CRS_DEP_TIME, no delay	505	720	1120	1166.8489	1615	2245
CRS_DEP_TIME, delay	511	1135	1542	1444.9816	1756	2230

From Figure 1, we see that flight delays are more likely to occur in the summer months of June, July, and August. We also note December which is most likely for the winter holidays. In Table 2, we see that the `DISTANCE` variable does not actually differ much between the two labels, however `CRS_DEP_TIME` does appear to have a significant difference.

Supervised Analysis

After acquiring the above variables, we further ran a lasso regression to aid in variable selection. This was primarily because we had many categorical variables and we deemed them to not all be useful. We used the `model.matrix` function to create a design matrix to account for each categorical variable.

To acquire an adequate tuning parameter, we ran a 10-fold cross validation for lasso regression. We then used the 1-standard error λ rather the minimum due to the large imbalance in labels. Through the lasso regression, we further removed MONTH6, MONTH9, MONTH12, DISTANCE. These were the final variables used for our model, excluding some specific months in `MONTHS`.

- `QUARTER`: Earlier in our EDA we saw that the summer months typically saw more delays.
- `MONTH`: For the same reason as `QUARTER` (there might be some correlation with `QUARTER` but we chose to ignore it in this case).
- `DAY_OF_MONTH`: Not to much intuition behind this other than perhaps more people travel in the beginnings and ends of the month rather than in the middle.
- `DAY_OF_WEEK`: While for the most part the day of the week was uniformly distributed, there tended to be less delays over the weekend.
- `CARRIER`: In our EDA we saw that different carriers have larger rates of delay
- `DEST`: Certain destinations might have higher chances of extreme weather conditions or high traffice (mostly NY airports)
- `CRS_DEP_TIME`: In our EDA we saw that delays usually occurred later in the day

We continued further by using a random forest with 500 trees. At first we used the naive random forest, and got a ROC curve only slightly better than the previous models. However, as mentioned above, the training data set is extremely imbalanced. For that reason we decided to use a balanced random forest. From this, we were able to achieve an AUC of 0.82 on training and 0.62 on testing. We reported an approximation of 0.59 due to an accidental error while coding that was resolved for this report.

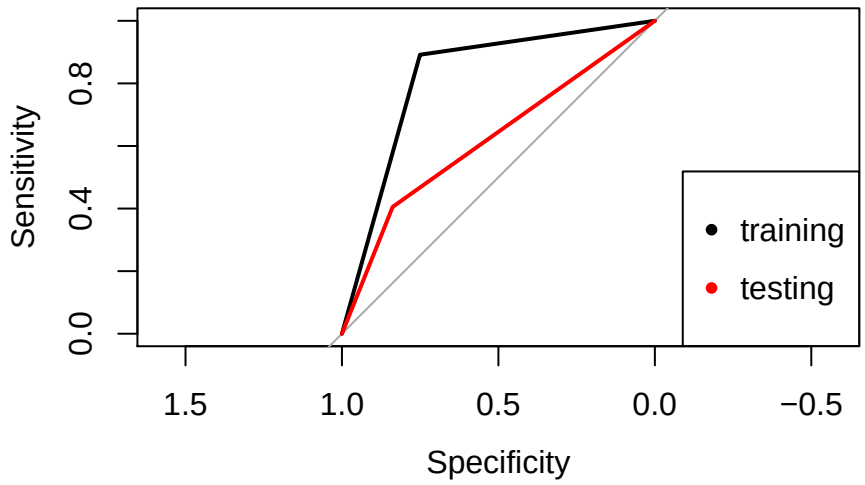


Table 3: Sensitivity and Specificity of the balanced random forest.

	Sensitivity	Specificity	Accuracy
Training	0.75	0.89	0.77
Testing	0.84	0.40	0.80

Analysis of Results

In looking at the variable importance plot, we see that `CRS_DEP_TIME` seems to be the dominating variable.

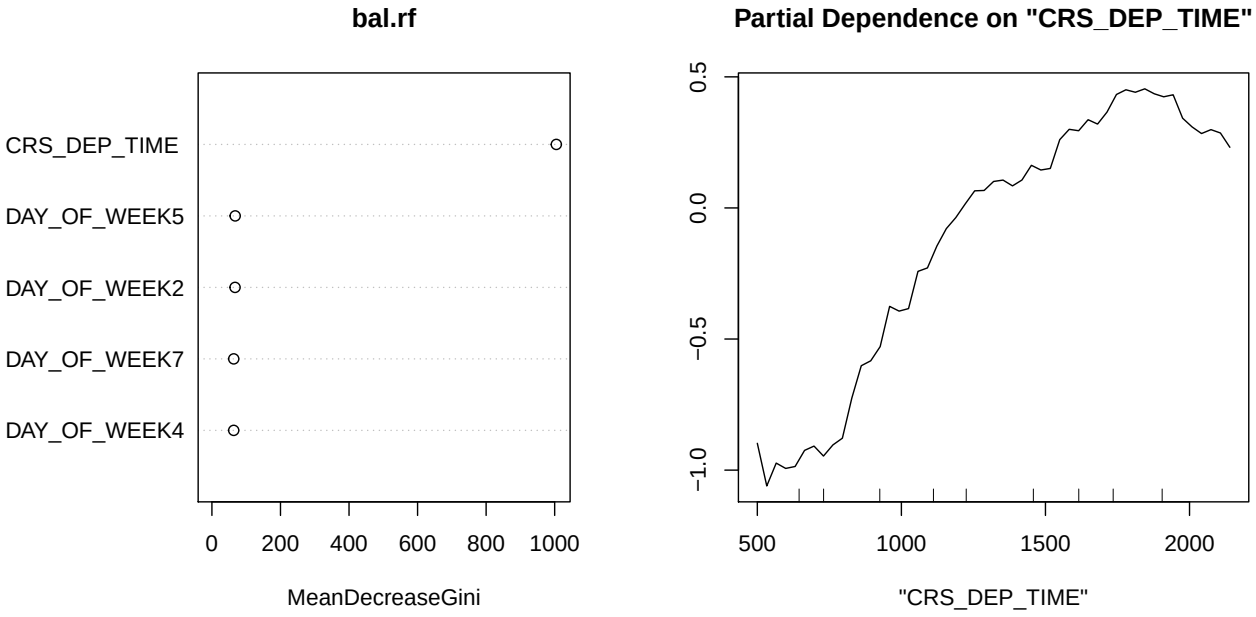
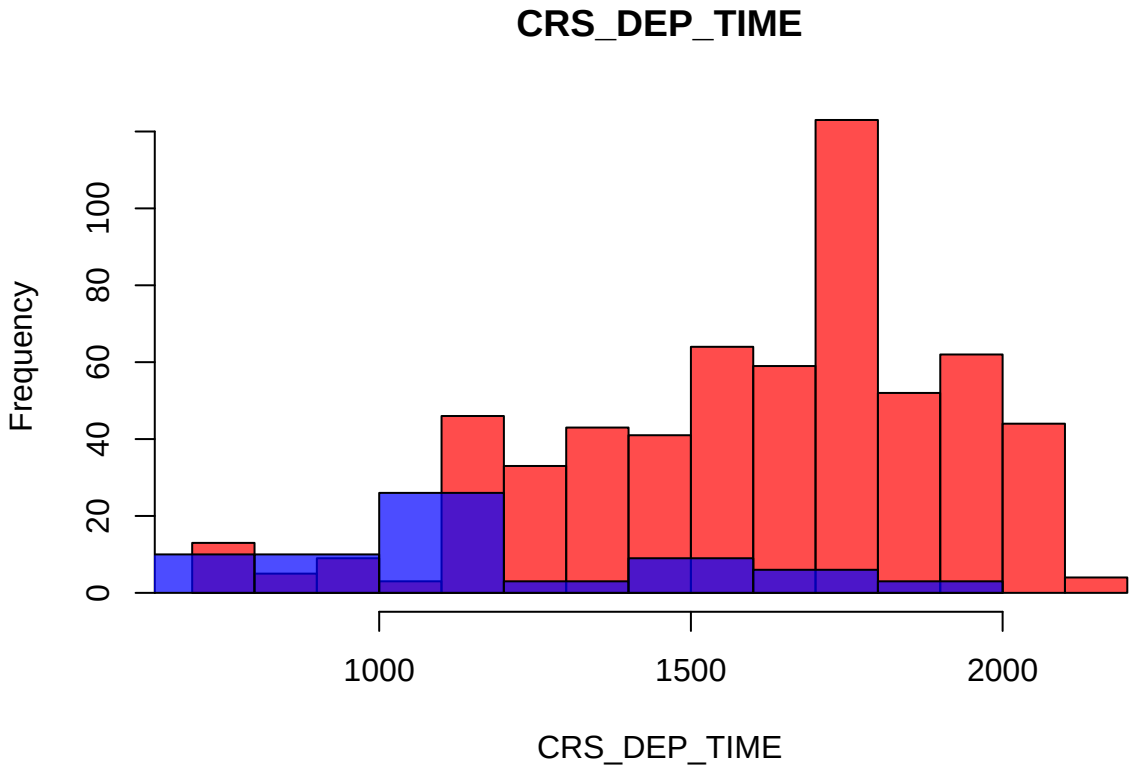


Table 4: Confusion matrix on the predicted results. Rows are predictions and columns are the actual.

	0	1
0	740	601
1	69	150

We see that our model does a good job at capturing the true negatives. It has high sensitivity but very low specificity. We ended up with an actual AUC of 0.66. We see in the histogram below that due to the high bias in departure time, our model automatically assumes that late flights are delayed. If we didn't care for AUC, we would have to adapt for the tradeoff using some type of penalty when training. When the ratio of C to r is large, we would much rather wait for the flight due to the high cost of missing it. Similarly, when C to r is smaller, we would opt to miss the flight rather than wait. Adjusting C to r is essentially determining whether we care about Sensitivity or Specificity more.



Further Remarks

We did not include any hourly weather forecasts for Pittsbugh because we were unable to find a suitable dataset. However, with the prescence of such a dataset, I think it would be quite clear that rain and snow would be strong indicators of late flights.

One thing we noted very early on was that the prescence of a late flight usually delayed succeeding flights. This in itself is actually quite a loaded problem because it involves utilizing the planes arriving into Pittsburgh. Using any distinguisable IDs unique to the planes, one can determine which planes are coming in to Pittsburgh and where they are departing next. One additional piece of information that could be quite useful is which gate the plane is supposed to arrive at. This is because generally the plane that lands at the gate will be the one responsible for that gate's next takeoff. By modeling whether an arriving plane is likely to be delayed or arriving late, one could use this probability of lateness as a prior for a bayesian model that makes use of this fact. This prior can also be equipped with a diminishing return because generally if one plane is delayed then usually the airport will quickly (to some extent of quick) resolve the issue.