HW2_Tyler_Nicholas.Rmd

Tyler Nicholas

August 9, 2016

Flights at ABIA

We are trying to answer the question of which month is the best time to fly to minimize delays. For the purpose of this question we define delays to be leaving over 30 minutes after scheduled departure time. The data was masked to show only flights leaving Austin so we could determine the best time to fly out of Austin.

In Figure 1 we see the number of delays that are longer than half an hour each month. There are a distinct top 3 and a distinct bottom 3. Here we see that June, December, and March are the months with the most flights delayed over half an hour and November, October, and September are the months with the least. This gives us an idea of the best months to fly to reduce massive delays. However this may not be telling us the full picture. We really want to see what percentage of flights are delayed each month to see what the likelihood is of having our flight delayed. We see this in Figure 2. We get the same top 3 and bottom 3 months.

With this we can say that November, October, and September are the best times to fly out of Austin to reduce delays. If you are flying out of Austin in June, December, or March your flight is likely to be delayed. This means that if you have a connecting flight, you should schedule a longer layover to make sure that you catch your next flight.

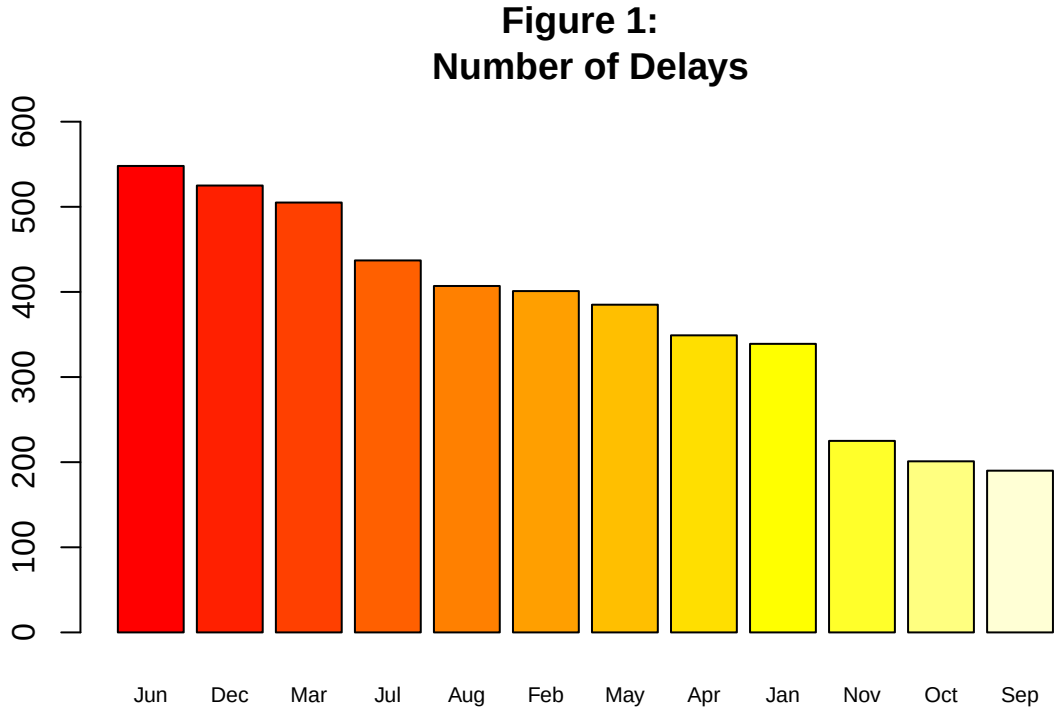
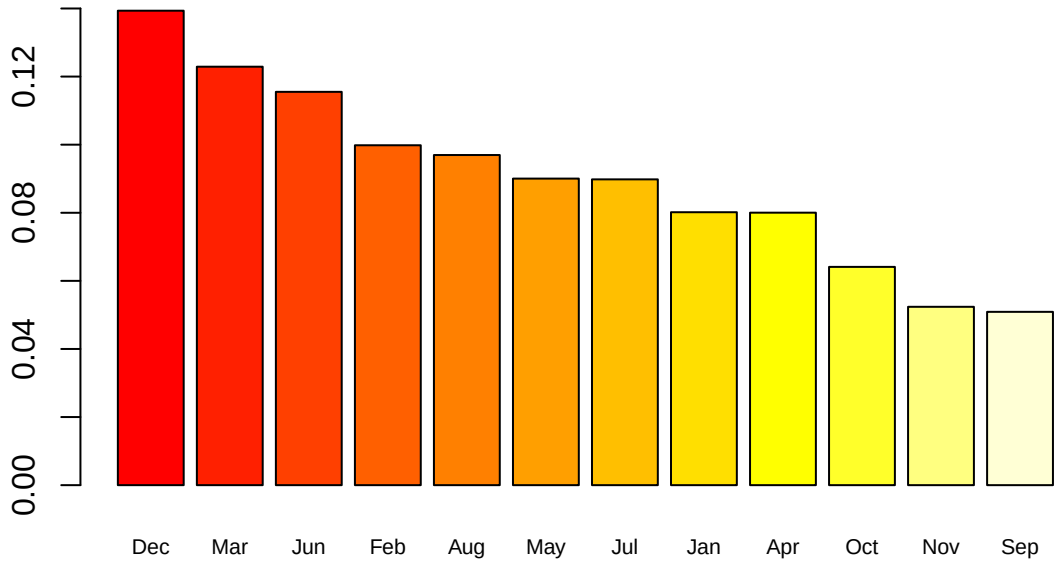


Figure 2:
Percentage of Flights Delayed Each Month



Author Attribution

We made two different models, one using multinomial regression and the other using Naive Bayes to predict the author of the test set documents. The regression model only gave us 55.28 % accuracy on the test set so we will focus most of our discussion on the Naive Bayes model which gave us an accuracy of 91.16% on the test set.

Our Naive Bayes model was run with the top 782 components of the PCA and was by far the better of the two models we created. It was also preferable since it took much less time to run. In this particular scenario it was preferable in performance and runtime.

Since the Naive Bayes was our preferred model, lets look at some of the authors that we had difficulty distinguishing from one another in the Naive Bayes Model:

Edna Fernandes: Tim Farrand
Eric Auchard : Therese Poletti
Joe Ortiz : Tim Farrand
Martin Wolk : Therese Poletti
Sarah Davison : Peter Humphrey
Todd Nissen : David Lawder

The left hand side is the actual writer and the right hand side is the writer that we incorrectly predicted. We are only showing instances that occurred at least 4 times in our data set. With 91% accuracy, 4 incorrect instances of the same writers is notable. We can see that there are multiple writers that are repeatedly confused with both Tim Farrand and Therese Poletti.

Though there were a few instances where the Naive Bayes model was incorrect, it was more accurate and quicker to run than our regression model and is overall the best model we created.

Practice with association rule mining

We take the grocery baskets for over 9000 customers to examine what relationships exist between items in those baskets. We pick thresholds for support, confidence and lift. Here we chose .005 for support to get item combinations that appear fairly often in baskets. Then we did a confidence threshold of .5 so that we only predict items that are in half of the baskets containing the left hand side items. Lastly we use a lift threshold of 3.2 to only get predicted relationships that make you much more likely to by the item on the right hand side. These thresholds were chosen to find unique relationships with very high lift. We can see the two resulting relationships here:

```
##   lhs                     rhs                    support confidence     lift
## 1 {curd,
##    tropical fruit}  => {yogurt}           0.005287239  0.5148515 3.690645
## 2 {citrus fruit,
##    root vegetables,
##    whole milk}      => {other vegetables} 0.005795628  0.6333333 3.273165
```

From this we see that if you have curd and tropical fruit in your basket, you are 3.69 times more likely to buy yogurt than the average person. Similarly if you have citrus fruit, root vegetables, and whole milk in your basket, you are 3.27 times more likely to buy other vegetables. These discoveries can lead us to create coupons encouraging purchasing tropical fruit and curd in order to bolster yogurt sales or similarly giving coupons for citrus fruit, root vegetables, and whole milk to encourage purchasing of other vegetables.