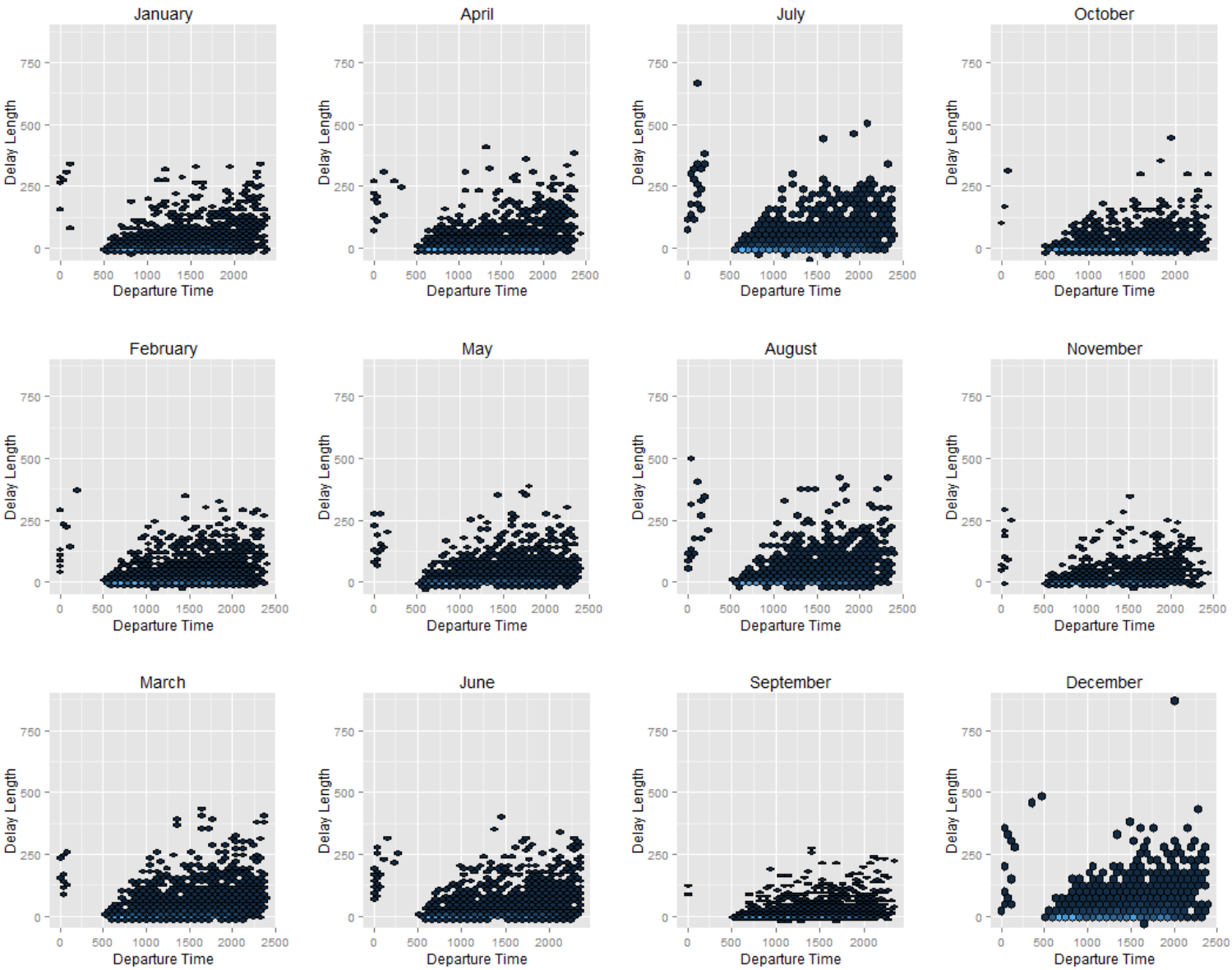
STA380_Homework2_Charles_Z_Aldrich

Charles Z. Aldrich

August 18, 2015

Problem 1: Flights at ABIA



The resulting series of plots show the density in delay length throughout a 24 hr day. All plots are shown in the same scale in order to produce easy comparability across months. Insights from this plot include: the months to fly with the lowest delays throughout each day are September, October, and November. Additionally; June, July, August, and December show the longest delays by month. Finally, as expected, at the start of each day (5:00am), delays are near zero and climb uptil a relatively steady state is reached at rougouy 10:00am each day across each month. Therefore, if an individual wants to fly without having to deal with delays, they have a greater chance of doing so if using flights with departure times prior to 10:00am.

Problem 2: Author Attribution

Naive Bayes Accuracy Prediction

```
## [1] 0.604

##             authors  x percent
## 8       DavidLawder  5    0.10
## 4   BenjaminKangLim  8    0.16
## 44      ScottHillis 11    0.22
## 7  DarrenSchuettler 14    0.28
## 13 HeatherScoffield 16    0.32
## 35       MureDickie 16    0.32

##            authors  x percent
## 36       NickLouth 42    0.84
## 6      BradDorfman 45    0.90
## 21     KarlPenhaul 46    0.92
## 11  FumikoFujisaki 49    0.98
## 16    JimGilchrist 49    0.98
## 29 LynnleyBrowning 49    0.98
```

From the output above it can be seen that Naive Bayes as a classifer model was able to correctly predict 60.36% of all articles in the test dataset. This is based on the term frequency matrices created from all of the articles in each dataset. Overall, the result is not terrible but could be better. Part of the reason for the relatively low overall accuracy of the model is the poor performance on some particular authors. The second readout above shows the 6 worst authors in terms of the percentage of their work in the test dataset that was correctly classified. It is likely that if there were more articles to train the model on then the word frequency for each author would be improved and more words unique to that author would be included. In doing so, Naive Bayes would be able to do a better job of classifying articles for these authors. Alternatively, the final readout above shows the best author classification from Naive Bayes. In the case of the top 3 authors, all but 1 of their works in the test set was accurately classified. This would indicate that the term frequency of works by these authors are both unique in nature and consistant for the particular author. Given these characteristics, Naive Bayes can more accurately classify their works. Thus, as noted above, the overall accuracy of Naive Bayes on this data set performs reasonably well.

Random Forests Prediction

```
## [1] 0.7716

##          authors  x percent
## 44   ScottHillis 12    0.24
## 50  WilliamKazer 16    0.32
## 35    MureDickie 18    0.36
## 14 JaneMacartney 20    0.40
## 46      TanEeLyn 25    0.50
## 6    BradDorfman 27    0.54

##            authors  x percent
## 30 MarcelMichelson 48    0.96
## 28  LynneO'Donnell 49    0.98
## 41    RogerFillion 49    0.98
## 11  FumikoFujisaki 50    1.00
## 16    JimGilchrist 50    1.00
## 29 LynnleyBrowning 50    1.00
```

I chose to run random forests as the second supervised classifier model because it proved to be one of the stronger predictive models in the first half of this course. After training the model on the train dataset created for use in the Naive Bayes portion the problem and running it against the test dataset, the resulting overall accuracy of the model was 77.16%. This is far better than Naive Bayes. Part of the reason for the improvement in the overall accuracy is that random forests runs a multitude of trees and returns the optimal output as opposed to simply running one model and using its output as the overall output. The second output above shows the bottom 6 authors, meaning these were the hardest authors for the random forests model to classify. Right off the bat it can be seen that the worst performances are far better from the random forests model than from the naive bayes model. Alternatively, when looking at the authors that were best classified, the random forest model was able to classify 3 authors with 100% accuracy. Thus, random forests does a better overall job of classifying an article based on the term frequencies assoicated with each author. This accuracy level was found using an mtry value of 5 and ntrees of 100. These criteria were chosen for the sake of runtime in the model. If both criteria were increased the accuracy of the model will also increase but the runtime also increases. Therefore, from a relative base level, the random forest model still outperformes the Naive Bayes model for classification on this dataset.

Problem 3: Practice with Association Rule Mining

```
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##        0.05    0.1    1 none FALSE            TRUE   0.001      1      4
##  target   ext
##   rules FALSE
##
## Algorithmic control:
```

```
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## apriori - find association rules with the apriori algorithm
## version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.00s].
## sorting and recoding items ... [151 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [1044 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

After playing around with a variety of levels for support and confidence I elected to use the following: support = 0.001 and confidence = 0.05. The reason for the extremely small support level is that it provided results in which paired basket items in the lhs led to items in the rhs. With larger support levels there were only one to one relationships; no many to one relationships were present.

Readout with Lift > 5:

```
##    lhs                   rhs                   support confidence      lif
t
## 1  {popcorn}          => {salty snack}     0.001176778 0.25352113 10.42435
3
## 2  {liquor}           => {red/blush wine}  0.001111402 0.15596330 12.62230
0
## 3  {red/blush wine}   => {liquor}          0.001111402 0.08994709 12.62230
0
## 4  {liquor}           => {bottled beer}    0.002811192 0.39449541  7.61894
2
## 5  {bottled beer}     => {liquor}          0.002811192 0.05429293  7.61894
2
## 6  {condensed milk}   => {coffee}          0.001242155 0.18811881  5.03934
4
## 7  {salt}             => {sugar}           0.001111402 0.16037736  7.36676
3
## 8  {sugar}            => {salt}            0.001111402 0.05105105  7.36676
3
## 9  {herbs}            => {root vegetables} 0.003726464 0.35625000  5.08320
9
## 10 {root vegetables}  => {herbs}           0.003726464 0.05317164  5.08320
9
## 11 {baking powder}    => {sugar}           0.001634414 0.14367816  6.59970
3
## 12 {sugar}            => {baking powder}   0.001634414 0.07507508  6.59970
3
## 13 {flour}            => {sugar}           0.001895921 0.16959064  7.78996
5
## 14 {sugar}            => {flour}           0.001895921 0.08708709  7.78996
```

```
5
## 15 {liquor,
##     red/blush wine}   => {bottled beer}    0.001046025 0.94117647 18.17706
5
## 16 {bottled beer,
##     liquor}           => {red/blush wine}  0.001046025 0.37209302 30.11394
1
## 17 {bottled beer,
##     red/blush wine}   => {liquor}          0.001046025 0.41025641 57.57139
5
## 18 {other vegetables,
##     root vegetables}  => {herbs}           0.001438285 0.05670103  5.42061
9
## 19 {herbs,
##     whole milk}       => {root vegetables} 0.001176778 0.35294118  5.03599
6
## 20 {citrus fruit,
##     pip fruit}        => {tropical fruit}  0.003072699 0.37600000  5.57296
1
## 21 {citrus fruit,
##     tropical fruit}   => {pip fruit}       0.003072699 0.24607330  5.05905
5
```

After playing around with lift thresholds, I chose to go with > 5 because it returned 21 association rules that I felt made logical sense. For example, if a basket contains citrus fruit and pip fruit then it is likely to also contain tropical fruit. Another example is all combinations of liquor, red/blush wine, and bottled beer. A third example is condensed milk and coffee and a forth example is herbs and root vegetables. When I think of my own grocery shopping trips, these associations of items are things I could/would easily purchase together. They make logical sense.

Readout with Confidence > 0.4

```
##    lhs                   rhs                    support confidence      li
ft
## 1  {herbs}            => {other vegetables} 0.004314854  0.4125000  3.3156
07
## 2  {butter}           => {whole milk}       0.014382845  0.4036697  2.4570
36
## 3  {liquor,
##     red/blush wine}   => {bottled beer}     0.001046025  0.9411765 18.1770
65
## 4  {bottled beer,
##     red/blush wine}   => {liquor}           0.001046025  0.4102564 57.5713
95
## 5  {herbs,
##     whole milk}       => {other vegetables} 0.001372908  0.4117647  3.3096
97
## 6  {citrus fruit,
##     onions}           => {other vegetables} 0.001111402  0.4594595  3.6930
```

```
59
## 7  {onions,
##     whole milk}       => {other vegetables} 0.002288180  0.4268293  3.4307
83
## 8  {hamburger meat,
##     root vegetables}  => {other vegetables} 0.001307531  0.4255319  3.4203
55
## 9  {butter,
##     curd}             => {whole milk}       0.001242155  0.4318182  2.6283
69
## 10 {curd,
##     tropical fruit}   => {whole milk}       0.001503661  0.5227273  3.1817
10
## 11 {curd,
##     root vegetables}  => {whole milk}       0.001242155  0.4222222  2.5699
61
## 12 {curd,
##     other vegetables} => {whole milk}       0.003399582  0.4814815  2.9306
57
## 13 {butter,
##     root vegetables}  => {whole milk}       0.002092050  0.5161290  3.1415
48
## 14 {butter,
##     yogurt}           => {whole milk}       0.002615063  0.4210526  2.5628
42
## 15 {butter,
##     other vegetables} => {whole milk}       0.003791841  0.4603175  2.8018
37
## 16 {sausage,
##     soda}             => {rolls/buns}       0.002288180  0.4605263  3.8939
80
```

Similar to what was outlined above; after playing around with various levels for confidence I elected to use > 0.4 because it returned 16 association rules that again made logical sense to me. For example, one association rule was sausage, soda --> rolls/buns. This association rule could be indicative of an upcoming bbq or grill out. Additionally, this level of confidence also returned the association rule between beer, wine, and liquor.

Therefore, as noted by the example in the two pervious paragraphs, and by analyzing the association rules returned from each threshold, I believe the association rules generated in this problem make sense. Given a series of shopping baskets containing a variety of items, I believe that association rule mining found various relationships between basket purchases that make sense for a typical consumer. When looking at each assocation rule, I could see myself purchasing that series of items together on my next grocery trip.