STA 380 Homework 2

Alexandria Nguyen

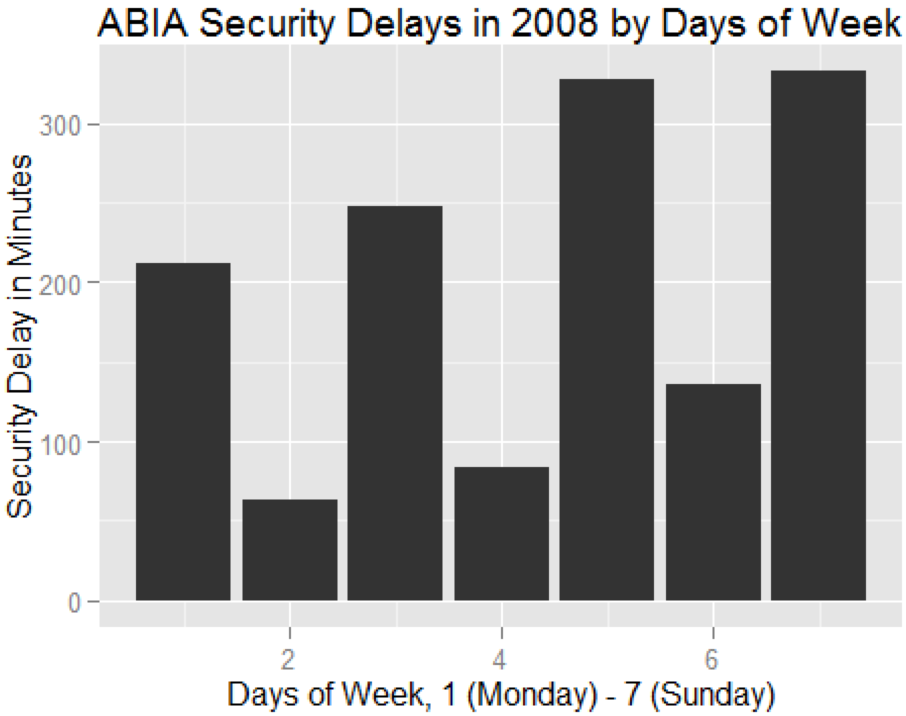
August 19, 2015

Flights at ABIA

Required Files:

- ABIA.csv

Q: What is the best day of the week to fly to minimize delays?



A: Tuesday

Author Attribution

Required Files:

- c50train folder (50 .txt files)
- c50test folder (50 .txt files)

Model 1: Naive Bayes

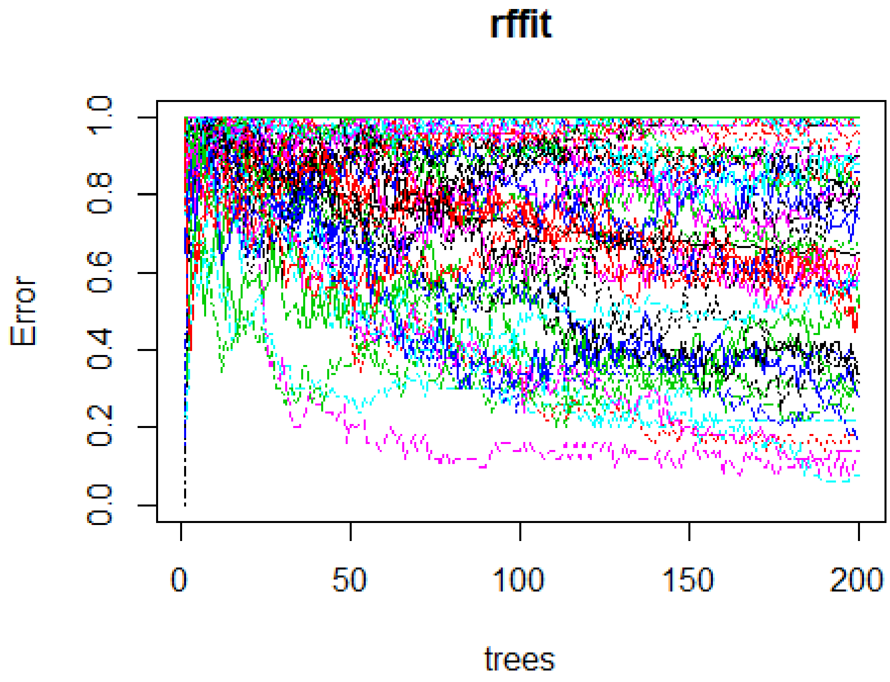
```
## Loading required package: NLP
## 
## Attaching package: 'NLP'
## 
## The following object is masked from 'package:ggplot2':
## 
##     annotate
```

The Naive Bayes model correctly matched the authors to their works 0.6036 (or 60%) of the time.

Model 2: Random Forest

```
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

Random Forest Visualization



```
##     AaronPressman        AlanCrosby    AlexanderSmith   BenjaminKangLim 
##                20                19                20                 4 
##     BernardHickey       BradDorfman  DarrenSchuettler       DavidLawder 
##                23                54                82                67 
##     EdnaFernandes       EricAuchard    FumikoFujisaki    GrahamEarnshaw 
##                 1                 7                92                52 
##  HeatherScoffield     JaneMacartney        JanLopatka      JimGilchrist 
##                 3               116                34                51
##          JoeOrtiz      JohnMastrini      JonathanBirt    JoWinterbottom
##                 0               103                93               108
##       KarlPenhaul         KeithWeir    KevinDrawbaugh     KevinMorrison
##                68                16                41                24
##     KirstinRidley KouroshKarimkhany         LydiaZajc    LynneO'Donnell
##                32               170                39                49
##   LynnleyBrowning   MarcelMichelson      MarkBendeich        MartinWolk
##                50                52                63                16
##      MatthewBunce     MichaelConnor        MureDickie         NickLouth
##                58                 0                59                29
##   PatriciaCommins     PeterHumphrey        PierreTran        RobinSidel
##                49                92                27                 9
##      RogerFillion       SamuelPerry      SarahDavison       ScottHillis
##                43               188                 1                55
##       SimonCowell          TanEeLyn    TheresePoletti        TimFarrand
##               165                65                 0                28
##        ToddNissen      WilliamKazer
##                63                 0
```

The random forest model correctly matched the authors to their works 0.0096 (or 96%) of the time.

Interpretation:

- The random forest model performs surprisingly well at 96% accuracy, but this could be attributable to chasing noise.
- Based solely on the percentage accuracy for predictions, I would use the random forest model.

Association Rule Mining

Required Files:

- groceries.txt

A priori algorithm:

```
## Loading required package: Matrix
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:tm':
##
##     inspect
##
## The following objects are masked from 'package:base':
##
##     %in%, write
```

```
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##         0.6    0.1    1 none FALSE            TRUE   0.005      1     10
##  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), 9835 transaction(s)] done [0.02s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4 done [0.02s].
## writing ... [22 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
```

```
inspect(groceriesrules)
```

```
##    lhs                        rhs                    support confidence
lift
## 1  {onions,
##     root vegetables}       => {other vegetables} 0.005693950  0.6021505
3.112008
## 2  {curd,
##     tropical fruit}        => {whole milk}       0.006507372  0.6336634
2.479936
## 3  {domestic eggs,
##     margarine}             => {whole milk}       0.005185562  0.6219512
2.434099
## 4  {butter,
##     domestic eggs}         => {whole milk}       0.005998983  0.6210526
2.430582
## 5  {butter,
##     whipped/sour cream}    => {whole milk}       0.006710727  0.6600000
2.583008
## 6  {bottled water,
##     butter}                => {whole milk}       0.005388917  0.6022727
2.357084
## 7  {butter,
##     tropical fruit}        => {whole milk}       0.006202339  0.6224490
2.436047
## 8  {butter,
##     root vegetables}       => {whole milk}       0.008235892  0.6377953
2.496107
## 9  {butter,
##     yogurt}                => {whole milk}       0.009354347  0.6388889
2.500387
## 10 {domestic eggs,
##     pip fruit}             => {whole milk}       0.005388917  0.6235294
2.440275
## 11 {domestic eggs,
##     tropical fruit}        => {whole milk}       0.006914082  0.6071429
2.376144
## 12 {pip fruit,
##     whipped/sour cream}    => {other vegetables} 0.005592272  0.6043956
3.123610
## 13 {pip fruit,
##     whipped/sour cream}    => {whole milk}       0.005998983  0.6483516
2.537421
## 14 {fruit/vegetable juice,
##     other vegetables,
##     yogurt}                => {whole milk}       0.005083884  0.6172840
2.415833
## 15 {other vegetables,
##     root vegetables,
##     whipped/sour cream}    => {whole milk}       0.005185562  0.6071429
2.376144
## 16 {other vegetables,
##     pip fruit,
##     root vegetables}       => {whole milk}       0.005490595  0.6750000
2.641713
## 17 {pip fruit,
##     root vegetables,
##     whole milk}            => {other vegetables} 0.005490595  0.6136364
3.171368
## 18 {other vegetables,
##     pip fruit,
##     yogurt}                => {whole milk}       0.005083884  0.6250000
2.446031
## 19 {citrus fruit,
##     root vegetables,
##     whole milk}            => {other vegetables} 0.005795628  0.6333333
3.273165
## 20 {root vegetables,
##     tropical fruit,
##     yogurt}                => {whole milk}       0.005693950  0.7000000
2.739554
## 21 {other vegetables,
##     tropical fruit,
##     yogurt}                => {whole milk}       0.007625826  0.6198347
2.425816
## 22 {other vegetables,
##     root vegetables,
##     yogurt}                => {whole milk}       0.007829181  0.6062992
2.372842
```

```
inspect(subset(groceriesrules, subset=lift > 3))

##   lhs                     rhs                    support confidence
lift
## 1 {onions,
##    root vegetables}    => {other vegetables} 0.005693950  0.6021505
3.112008
## 2 {pip fruit,
##    whipped/sour cream} => {other vegetables} 0.005592272  0.6043956
3.123610
## 3 {pip fruit,
##    root vegetables,
##    whole milk}         => {other vegetables} 0.005490595  0.6136364
3.171368
## 4 {citrus fruit,
##    root vegetables,
##    whole milk}         => {other vegetables} 0.005795628  0.6333333
3.273165
```

```
inspect(subset(groceriesrules, subset=confidence > 0.65))

##   lhs                     rhs             support confidence     lift
## 1 {butter,
##    whipped/sour cream} => {whole milk} 0.006710727      0.660 2.583008
## 2 {other vegetables,
##    pip fruit,
##    root vegetables}    => {whole milk} 0.005490595      0.675 2.641713
## 3 {root vegetables,
##    tropical fruit,
##    yogurt}             => {whole milk} 0.005693950      0.700 2.739554
```

```
inspect(subset(groceriesrules, subset=support > .005 & confidence > 0.65))

##   lhs                     rhs             support confidence     lift
## 1 {butter,
##    whipped/sour cream} => {whole milk} 0.006710727      0.660 2.583008
## 2 {other vegetables,
##    pip fruit,
##    root vegetables}    => {whole milk} 0.005490595      0.675 2.641713
## 3 {root vegetables,
##    tropical fruit,
##    yogurt}             => {whole milk} 0.005693950      0.700 2.739554
```

Interpretation:

I set the confidence to 0.6 (which means that whole milk or other vegetables occur when the previous sets occur 6 out of 10 times) to try to bring out significant correlations that occur frequently. Any time you buy items on the left, there's a 60% chance the item on the right will be purchased. I set the lift as high as possible without getting a null value to make the relationship between the right hand side as relevant as possible in comparison to the probability of buying items in the left basket.