Week 4 Rundown

James Rundle, Bradley Odac, Avi

9/17/2017

Types of Purchases for Masked Id

Sorted the purchases made with credit cards by descriptor 2. We find Retail and Entertainment are top purchases

```
Top_Cat2<- (head(sort(summary(Cust12_Des2),decreasing = TRUE))[1:3])
print(Top_Cat2)
```

```
## RETAIL / DEPARTMENT STORES              ENTERTAINMENT
##                         41                         22
##  AIRLINES / TRANSPORTATION
##                          9
```

Most descriptive purchases from sorted descriptor 2

Used the top categories, Entertainment and Retail, to show every purchase Maybe some regex to find similar qualites

```
Entertainment_Cat <- print(subset(Cust12, subset = Des2 == "ENTERTAINMENT")$Des3)
```

```
##  [1] "VIDEO GAME ARCADES/ESTABLISHMENTS"
##  [2] "DRINKING PLACES (ALCOHOLIC BEVERAGES) BARS, TAVERNS,NIGHTCLUBS, COCKTAIL LOUNGES, AND DISCOTHE
##  [3] "MOTION PICTURE THEATERS"
##  [4] "VIDEO TAPE RENTAL STORES"
##  [5] "BANDS, ORCHESTRAS, AND MISCELLANEOUS ENTERTAINERS NOT ELSEWHERE CLASSIFIED"
##  [6] "Government-Licensed Horse/Dog Racing"
##  [7] "Digital Goods - Games"
##  [8] "ARTISTS SUPPLY AND CRAFT SHOPS"
##  [9] "COMMERCIAL PHOTOGRAPHY, ART, AND GRAPHICS"
## [10] "MOTION PICTURE AND VIDEO TAPE PRODUCTION AND DISTRIBUTION"
## [11] "BICYCLE SHOPS - SALES AND SERVICE"
## [12] "BOAT RENTALS AND LEASING"
## [13] "MOTION PICTURE THEATERS (BMG INFERRED DEFINITION)"
## [14] "MOTORCYCLE SHOPS AND DEALERS"
## [15] "BILLIARD AND POOL ESTABLISHMENTS"
## [16] "MOTION PICTURE THEATERS"
## [17] "CAMPER, RECREATIONAL AND UTILITY TRAILER DEALERS"
## [18] "PHOTOGRAPHIC STUDIOS"
## [19] "SNOWMOBILE DEALERS"
## [20] "VIDEO GAME ARCADES/ESTABLISHMENTS"
## [21] "BILLIARD AND POOL ESTABLISHMENTS"
## [22] "THEATRICAL PRODUCERS (EXCEPT MOTION PICTURES) AND TICKET AGENCIES"
```

```
Retail_Cat <- print(subset(Cust12, subset = Des2 == "RETAIL / DEPARTMENT STORES")$Des3)
```

```
##  [1] "HOUSEHOLD APPLIANCES"
##  [2] "MISCELLANEOUS APPAREL AND ACCESSORY SHOPS"
```

```
##  [3] "CIGAR STORES AND STANDS"
##  [4] "BOOKS, PERIODICALS AND NEWSPAPERS"
##  [5] "SHOE STORES"
##  [6] "LUGGAGE AND LEATHER GOODS STORES"
##  [7] "RECORD STORES"
##  [8] "TYPEWRITER STORES, SALES, RENTALS, SERVICE"
##  [9] "PRECIOUS STONES, METALS, WATCHES, JEWELRY"
## [10] "HOUSEHOLD APPLIANCES"
## [11] "PAWN SHOPS"
## [12] "MISCELLANEOUS FOOD STORES CONVENIENCE STORES AND SPECIALTY MARKETS"
## [13] "FLORISTS SUPPLIES, NURSERY STOCK AND FLOWERS"
## [14] "GIFT, CARD, NOVELTY AND SOUVENIR SHOPS"
## [15] "JEWELRY, PRECIOUS METAL"
## [16] "JEWELRY, SILVERWARE AND PLATED WARE"
## [17] "TYPEWRITER STORES, SALES, RENTALS, SERVICE"
## [18] "GLASSWARE/CRYSTAL STORES"
## [19] "NEWS DEALERS AND NEWSSTANDS"
## [20] "FLORISTS"
## [21] "BOOK STORES"
## [22] "COMMERCIAL FOOTWEAR"
## [23] "SPORTS AND RIDING APPAREL STORES"
## [24] "PRECIOUS STONES, METALS, WATCHES, JEWELRY"
## [25] "COMPUTER MAINTENANCE, REPAIR, SERVICES (NOT ELSEWHERE CLASSIFIED)"
## [26] "FURRIERS AND FUR SHOPS"
## [27] "MENS AND BOYS CLOTHING AND ACCESSORIES STORES"
## [28] "GLASSWARE/CRYSTAL STORES"
## [29] "DURABLE GOODS (NOT ELSEWHERE CLASSIFIED)"
## [30] "FLORISTS"
## [31] "COMPUTERS, COMPUTER PERIPHERAL EQUIPMENT, SOFTWARE"
## [32] "MUSIC STORES MUSICAL INSTRUMENTS, PIANOS, AND SHEET MUSIC"
## [33] "PAWN SHOPS"
## [34] "FAMILY CLOTHING STORES"
## [35] "MENS AND BOYS CLOTHING AND ACCESSORIES STORES"
## [36] "BOOK STORES"
## [37] "COMPUTER PROGRAMMING, DATA PROCESSING, AND INTEGRATED SYSTEMS DESIGN SERVICES"
## [38] "FLORISTS"
## [39] "COSMETIC STORES"
## [40] "MISCELLANEOUS APPAREL AND ACCESSORY SHOPS"
## [41] "PRECIOUS STONES, METALS, WATCHES, JEWELRY"
```

Purchase Statistics

Here are some descriptive statistics for the customers with Masked ID 1 and 12

We can use this data to see in which quartile a majority of their purchases lie and classify them as a "spending type"

```
BigList <- list(PstatsID1, PstatsID12)
## Had Trouble figuring out how to  append to big list
print(BigList)
```

```
## [[1]]
## [[1]]$Stats
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    14.0    67.0   173.0   398.7   258.8  7237.0
##
## [[1]]$Outliers
##  [1] 5137 5505 1300 1079 7237 1945 2043 1054  950 2005 1047 1630 3180 1962
##
##
## [[2]]
## [[2]]$Stats
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    97.5   182.0   386.3   287.5  7298.0
##
## [[2]]$Outliers
##  [1]  876  863 1159 1292 1720  831 7298  927 2094 1921 1357 1146 2081 1440
## [15] 6007  994  789
```

Using data to predict

Here I am using the data from checking Balance, Savings Balance, ATM withdrawls, and Age to predict if the person will make purchases about 1000 dollars.

Now Im not really sure if this is what the function is displaying but we're getting something out of it...

```
merged_data <- merge(month_end_balances,WF_CardData,by="masked_id")
library(randomForest)
```

```
## randomForest 4.6-10
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
fit <- randomForest(as.factor(Payment> 1000) ~ check_bal_altered +
                      sav_bal_altered + atm_withdrawls_cnt + age,
                      data= merged_data,
                      importance=TRUE,
                      ntree=2000)
print(fit)
```

```
##
## Call:
##  randomForest(formula = as.factor(Payment > 1000) ~ check_bal_altered +      sav_bal_altered + atm_w
##                Type of random forest: classification
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 8.45%
## Confusion matrix:
##       FALSE TRUE class.error
## FALSE 19050    0           0
## TRUE   1758    0           1
```

Variable importance Plot

Almost all values do not seem to affect accuracy but savings balance, check balance, and age all seem to decrese the GINI sharply. If i interpret GINI correctly that means that these variables have much more

relevance to the event of a Payment being greater than 1000.

```
varImpPlot(fit)
```