Week 4 Rundown

James Rundle, Bradley Odac, Avi

9/17/2017

```
##Hope this works to salvage Data
#CardData <- WF_CardData2
bills <- list("INSURANCE / FINANCIAL SERVICES","TAX","UTILITIES / TELECOM")
entertainment <- list("ENTERTAINMENT","RESTAURANTS")
personal<- list("PET / VETERINARY","PERSONAL SERVICES","EDUCATION","HEALTHCARE / PHARMACY","LODGING")
transport<- list("AIRLINES / TRANSPORTATION","AUTO / GAS")
work<- list("BUILDING SUPPLY / WHOLESALE","OFFICE SUPPLY / STATIONERY","HOUSEHOLD","CONTRACTOR / BUSINES
common<- list("GROCERIES","RETAIL / DEPARTMENT STORES")
charity <- list("CHARITY / COMMUNITY ORG.")
other <- list("OTHER", NA)


TrimmedDesList <- list("BILLS"= bills,"ENTERTAINMENT"=entertainment,"PERSONAL CARE"=personal,"TRANSPORTA
```

Condensing Des2 Catagories down to 7

Running randomForests last week I felt like there were too many catagories to accurately use. Des 2 had the least descriptors coming in at 28. I felt it was a good middle ground to classify each "purchase" to a condensed list of 7 catacories. After wasting way too much time trying to use for loops to propagate a new column I just bruteforced it using nested lists of the 28 descriptors in the Des2 column

```
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$BILLS] <- "BILLS"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$ENTERTAINMENT] <- "ENTERTAINMENT"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$`PERSONAL CARE`] <- "PERSONAL CARE"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$TRANSPORTATION] <- "TRANSPORTATION"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$`WORK RELATED`] <- "WORK RELATED"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$`COMMON PURCHASES`] <- "COMMON PURCHASES"
CardData$MiniDes2[CardData$Des2 %in% TrimmedDesList$CHARITY] <- "CHARITY"
CardData$MiniDes2[CardData$Des2 == "OTHER" ] <- "OTHER"
CardData$MiniDes2[ is.na(CardData$Des2)] <- "OTHER"
```

Using the new MiniDes2 column to run randomForests

Well... It's different. I think Im going in the right direction but need some work. Right now the catagorical Random forest appears to only predict "COMMON PURCHASES". Thats not super great.

The regression forest is telling me *"The response has five or fewer unique values. Are you sure you want to do regression?"* Still coming up with no culprits after some intense googling.

```
MergedWMini <- merge.data.frame(month_end_balances,CardData,by="masked_id", all.x = TRUE )

MergedWMini$MiniDes2 = as.factor(MergedWMini$MiniDes2)

library(randomForest)
```

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

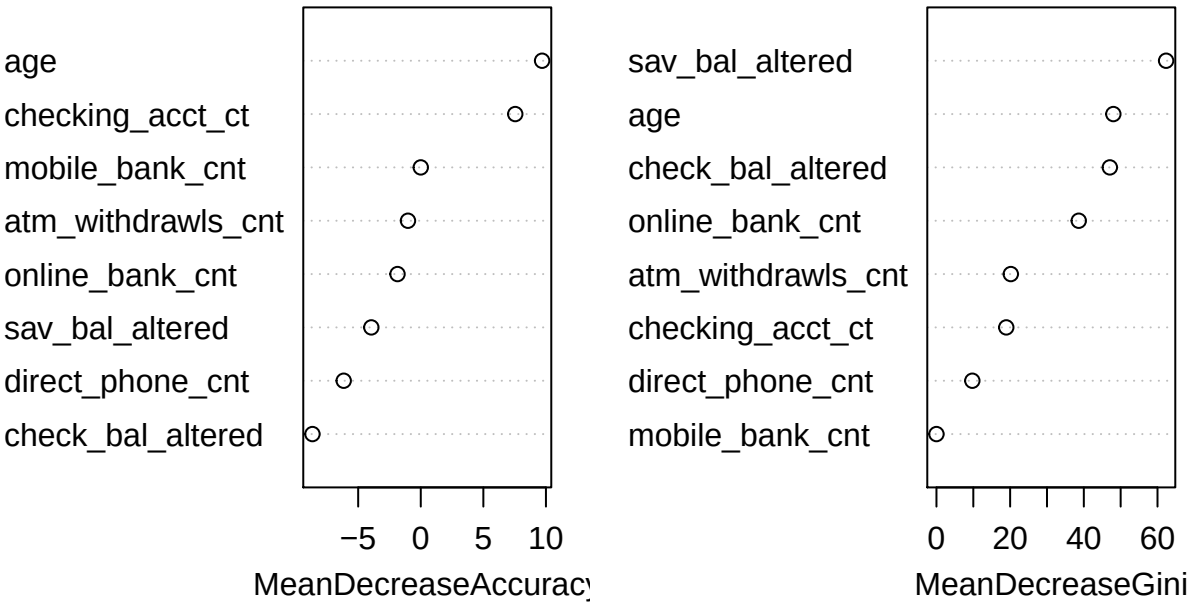
```
## Type rfNews() to see new features/changes/bug fixes.
```

```
CatPredict <- randomForest(as.factor(MiniDes2) ~  checking_acct_ct + online_bank_cnt + direct_phone_cnt
                      data=MergedWMini, mtry = 3 ,localImp = TRUE,
                      importance=TRUE,na.action = na.omit,
                      ntree=2000)
print(CatPredict)
```

```
##
## Call:
##  randomForest(formula = as.factor(MiniDes2) ~ checking_acct_ct +      online_bank_cnt + direct_phone
##                Type of random forest: classification
##                      Number of trees: 2000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 68.34%
## Confusion matrix:
##                   BILLS CHARITY COMMON PURCHASES ENTERTAINMENT OTHER
## BILLS                 0       0             2046             0     0
## CHARITY               0       0              342             0     0
## COMMON PURCHASES      0       0             6588             0     0
## ENTERTAINMENT         0       0             2958             0     0
## OTHER                 0       0             1698             0     0
## PERSONAL CARE         0       0             2448             0     0
## TRANSPORTATION        0       0             2358             0     0
## WORK RELATED          0       0             2370             0     0
##                   PERSONAL CARE TRANSPORTATION WORK RELATED class.error
## BILLS                         0              0            0           1
## CHARITY                       0              0            0           1
## COMMON PURCHASES              0              0            0           0
## ENTERTAINMENT                 0              0            0           1
## OTHER                         0              0            0           1
## PERSONAL CARE                 0              0            0           1
## TRANSPORTATION                0              0            0           1
## WORK RELATED                  0              0            0           1
```

```
varImpPlot(CatPredict)
```

CatPredict



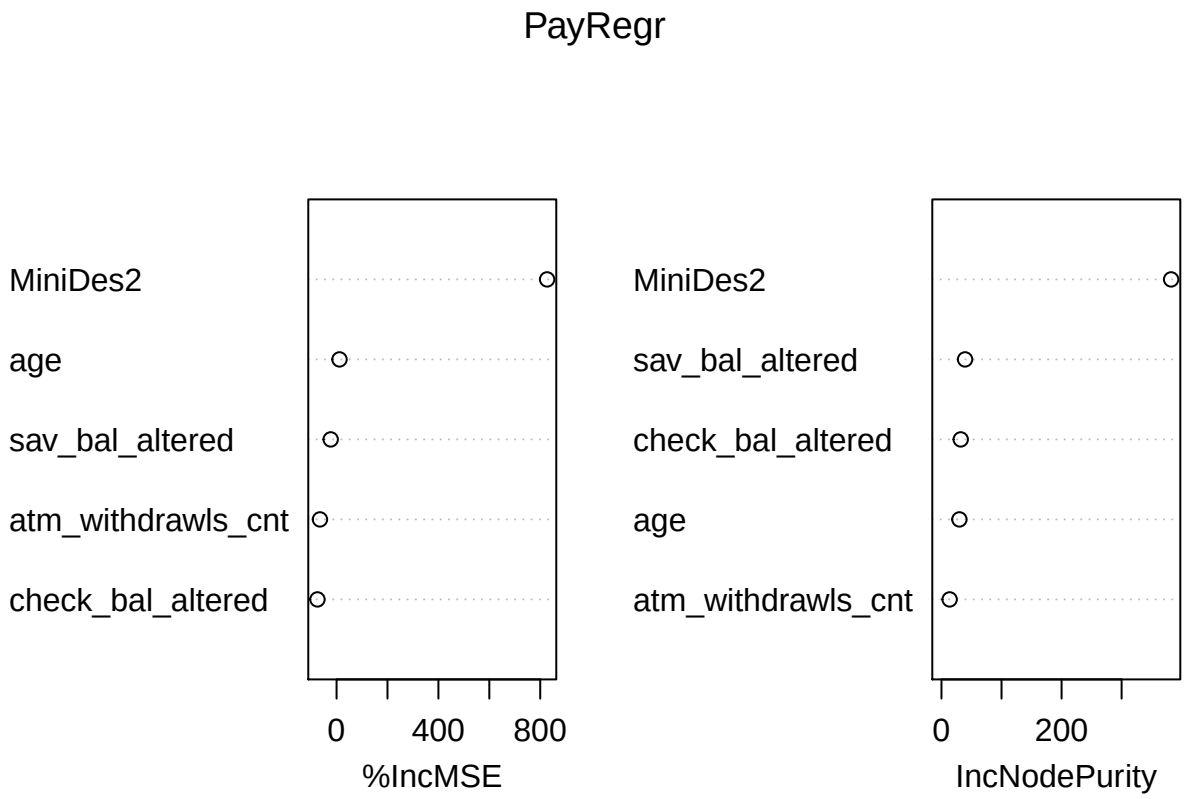

```
PayRegr <- randomForest((Payment > 1000) ~ MiniDes2 + check_bal_altered + sav_bal_altered +atm_withdrawl
                        data=MergedWMini, mtry = 2 ,localImp = TRUE,
                        importance=TRUE,na.action = na.omit,
                        ntree=2000)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
print(PayRegr)
```

```
##
## Call:
##  randomForest(formula = (Payment > 1000) ~ MiniDes2 + check_bal_altered +      sav_bal_altered + atm
##                Type of random forest: regression
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.06324698
##                     % Var explained: 18.23
```

```
varImpPlot(PayRegr)
```



trying to make mindes2 applicable for regression

HALP

```
OrdMiniDes <- MergedWMini$MiniDes2
OrdMiniDes <- as.numeric(OrdMiniDes)-7
```

Purchase Statistics

Here created dataframe collecting data on each of the 50 masked ids. The top two "purchase types" for each customer.

As we can see some accounts do not have credit cards, but most are "Common purchases" and "Entertainment"/ "Transportation". I think this is a good start but the classifying needs to be tweaked to supply some really valuable/workable information. Suggestions?

Each masked id also contains the descriptive statistics for the Payments(charges) on their creditcard. I would like to find out a percentage of purchases made within each range. If most of their purchases are high dollar, they need to be marketed towards differently than someone who a majority of their purchases are <50 from grocery stores or something.

```
### Fill dataframe with #1 des, #2 des, descriptive stats,

CardStats <- as.data.frame(matrix(ncol = 9, nrow = 50))
colnames(CardStats) <- c('masked_id',"1stDes",'2ndDes','MinofCharge','Q1ofCharge','MedofCharge','Q3ofCha

for(i in 1:50){
    Top2 <- summary(as.factor(MergedWMini[MergedWMini$masked_id == i,"MiniDes2"]))
```

```
    Top2 <- sort(Top2, decreasing = TRUE)
    CardStats$masked_id[i] <- i
    CardStats$'1stDes'[i] <- names(Top2[1])
    CardStats$'2ndDes'[i] <- names(Top2[2])
    CardStats$MinofCharge[i]<- summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['Min.']
    CardStats$Q1ofCharge[i]<-summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['1st Qu.']
    CardStats$MedofCharge[i]<-summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['Median']
    CardStats$Q3ofCharge[i]<-summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['3rd Qu.']
    CardStats$MaxofCharge[i]<-summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['Max.']
    CardStats$MeanofCharge[i]<-summary((MergedWMini[MergedWMini$masked_id == i,"Payment"]))['Mean']

}

print(head(CardStats,10))
```

```
##    masked_id           1stDes         2ndDes MinofCharge Q1ofCharge
## 1          1 COMMON PURCHASES   WORK RELATED          14         67
## 2          2 COMMON PURCHASES  ENTERTAINMENT          25         77
## 3          3 COMMON PURCHASES TRANSPORTATION          29         91
## 4          4             NA's          BILLS          NA         NA
## 5          5             NA's          BILLS          NA         NA
## 6          6 COMMON PURCHASES TRANSPORTATION          12         62
## 7          7 COMMON PURCHASES  ENTERTAINMENT          16         70
## 8          8             NA's          BILLS          NA         NA
## 9          9 COMMON PURCHASES   WORK RELATED          21         70
## 10        10             NA's          BILLS          NA         NA
##    MedofCharge Q3ofCharge MaxofCharge MeanofCharge
## 1          173        260        7237        398.7
## 2          178        284        7338        369.1
## 3          163        298        6266        403.0
## 4           NA         NA          NA          NaN
## 5           NA         NA          NA          NaN
## 6          157        277        3245        295.1
## 7          173        275        2621        291.8
## 8           NA         NA          NA          NaN
## 9          142        268        6307        360.9
## 10          NA         NA          NA          NaN
```