Lag 7 day Counts Increasing and Decreasing

Janis Corona

3/6/2020

This R markdown file shows all the work to gather statistical information on counts and other behind the scenes stock information on 52 hand picked stocks with time series information from Jan 2007- Feb 2020 compared to the 17 stock information of the same in the ROI\_HandPickedStocks.Rmd file.

Lets use the data set we created in the ROI-HandPickedStocks.Rmd file, ALL52 data set, in the ‘ALL\_52.csv’ file to see how well the machine learning does on this data frame.

ALL\_52 <- read.csv('ALL\_52.csv', sep=',', header=TRUE, na.strings=c('',' '))

Lets first get the median value for these stocks’ ROI.

colnames(ALL\_52)

## [1] "stock" "stockInfo" "businessType"   
## [4] "medianStockValue" "avgStockValue" "startValue"   
## [7] "finalValue" "stock\_ROI" "medn\_cSum\_decr\_L7"  
## [10] "Q3\_cSum\_decr\_L7" "max\_cSum\_decr\_L7" "medn\_cSum\_incr\_L7"  
## [13] "Q3\_cSum\_incr\_L7" "max\_cSum\_incr\_L7"

Lets add in some columns features for classifying this data. One to show if the stock has a low or high ROI based on the median ROI for these 52 stocks, one to show if the stock decreases more or less than the median number of times the stock has decreased from 2007-2020, and one to show if the stock increases more or less than the median number of times all stocks increased.

med52ROI <- median(ALL\_52$stock\_ROI)  
med52ROI

## [1] 1.374265

ALL\_52$ROI\_Low\_High <- ifelse(ALL\_52$stock\_ROI > med52ROI,'High',  
 'Low')  
med52Decr <- median(ALL\_52$medn\_cSum\_decr\_L7)  
med52Decr

## [1] 4

ALL\_52$MedCountsDecreasing <- ifelse(ALL\_52$medn\_cSum\_decr\_L7 > med52Decr,  
 'High Decreasing Counts',  
 'Low Decreasing Counts')  
  
med52Incr <- median(ALL\_52$medn\_cSum\_incr\_L7)  
med52Incr

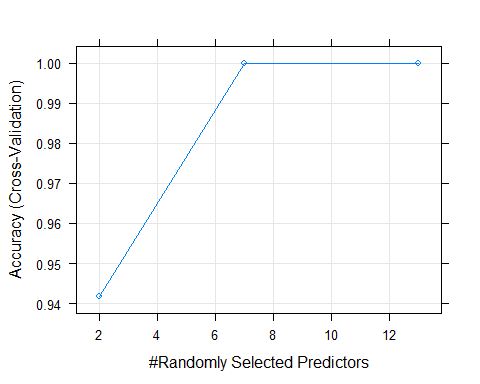
## [1] 4

ALL\_52$MedCountsIncreasing <- ifelse(ALL\_52$medn\_cSum\_incr\_L7 > med52Incr,   
 'High Increasing Counts',   
 'Low Increase Counts')

row.names(ALL\_52) <- ALL\_52$stock  
ALL\_52\_ML <- ALL\_52[,-c(1:3)]  
  
  
write.csv(ALL\_52, 'ALL\_52\_ml', row.names=TRUE)  
write.csv(ALL\_52\_ML, 'ALL\_52\_ML', row.names=TRUE)

set.seed(12356789)  
  
inTrain <- createDataPartition(y=ALL\_52\_ML$ROI\_Low\_High, p=0.7, list=FALSE)  
  
trainingSet <- ALL\_52\_ML[inTrain,]  
testingSet <- ALL\_52\_ML[-inTrain,]

rfMod <- train(ROI\_Low\_High~., method='rf', data=(trainingSet),   
 trControl=trainControl(method='cv'), number=5)  
plot(rfMod)



predRF <- predict(rfMod, testingSet)  
  
predDF <- data.frame(predRF, type=testingSet$ROI\_Low\_High)  
predDF

## predRF type  
## 1 High High  
## 2 High High  
## 3 Low Low  
## 4 Low Low  
## 5 High High  
## 6 High High  
## 7 High High  
## 8 Low Low  
## 9 Low Low  
## 10 High High  
## 11 Low Low  
## 12 Low Low  
## 13 High High  
## 14 High Low

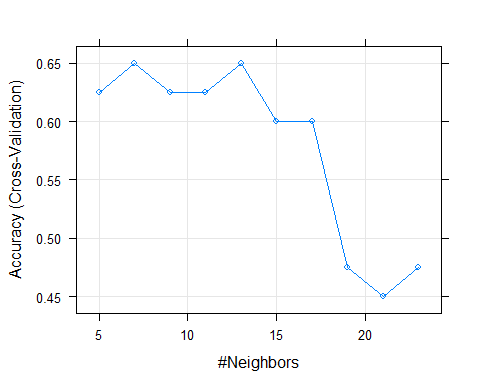
sum <- sum(predRF==testingSet$ROI\_Low\_High)   
length <- length(testingSet$ROI\_Low\_High)  
accuracy\_rfMod <- (sum/length)   
accuracy\_rfMod

## [1] 0.9285714

results <- c(round(accuracy\_rfMod,2), round(100,2))  
results <- as.factor(results)  
results <- t(data.frame(results))  
  
colnames(results) <- colnames(predDF)  
Results <- rbind(predDF, results)   
Results

## predRF type  
## 1 High High  
## 2 High High  
## 3 Low Low  
## 4 Low Low  
## 5 High High  
## 6 High High  
## 7 High High  
## 8 Low Low  
## 9 Low Low  
## 10 High High  
## 11 Low Low  
## 12 Low Low  
## 13 High High  
## 14 High Low  
## results 0.93 100

knnMod <- train(ROI\_Low\_High ~ .,  
 method='knn', preProcess=c('center','scale'),  
 tuneLength=10, trControl=trainControl(method='cv'), data=trainingSet)  
plot(knnMod)



rpartMod <- train(ROI\_Low\_High ~ ., method='rpart', tuneLength=7, data=trainingSet)

glmMod <- train(ROI\_Low\_High ~ .,   
 method='glm', data=trainingSet)

predKNN <- predict(knnMod, testingSet)  
predRPART <- predict(rpartMod, testingSet)  
predGLM <- predict(glmMod, testingSet)

length=length(testingSet$ROI\_Low\_High)  
  
sumKNN <- sum(predKNN==testingSet$ROI\_Low\_High)  
sumRPart <- sum(predRPART==testingSet$ROI\_Low\_High)  
sumGLM <- sum(predGLM==testingSet$ROI\_Low\_High)

accuracy\_KNN <- sumKNN/length   
accuracy\_RPART <- sumRPart/length   
accuracy\_GLM <- sumGLM/length

predDF2 <- data.frame(predRF,predKNN,predRPART,predGLM,   
 TYPE=testingSet$ROI\_Low\_High)  
colnames(predDF2) <- c('RandomForest','KNN','Rpart','GLM','TrueValue')  
  
results <- c(round(accuracy\_rfMod,2),   
 round(accuracy\_KNN,2),   
 round(accuracy\_RPART,2),  
 round(accuracy\_GLM,2),   
 round(100,2))  
  
results <- as.factor(results)  
results <- t(data.frame(results))  
colnames(results) <- c('RandomForest','KNN','Rpart','GLM','TrueValue')  
Results <- rbind(predDF2, results)   
Results

## RandomForest KNN Rpart GLM TrueValue  
## 1 High Low High Low High  
## 2 High Low High High High  
## 3 Low Low Low Low Low  
## 4 Low Low Low Low Low  
## 5 High High High High High  
## 6 High High High High High  
## 7 High High High Low High  
## 8 Low High Low Low Low  
## 9 Low Low Low Low Low  
## 10 High High High High High  
## 11 Low Low Low Low Low  
## 12 Low High Low Low Low  
## 13 High High High High High  
## 14 High Low High Low Low  
## results 0.93 0.71 0.93 0.86 100