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
```

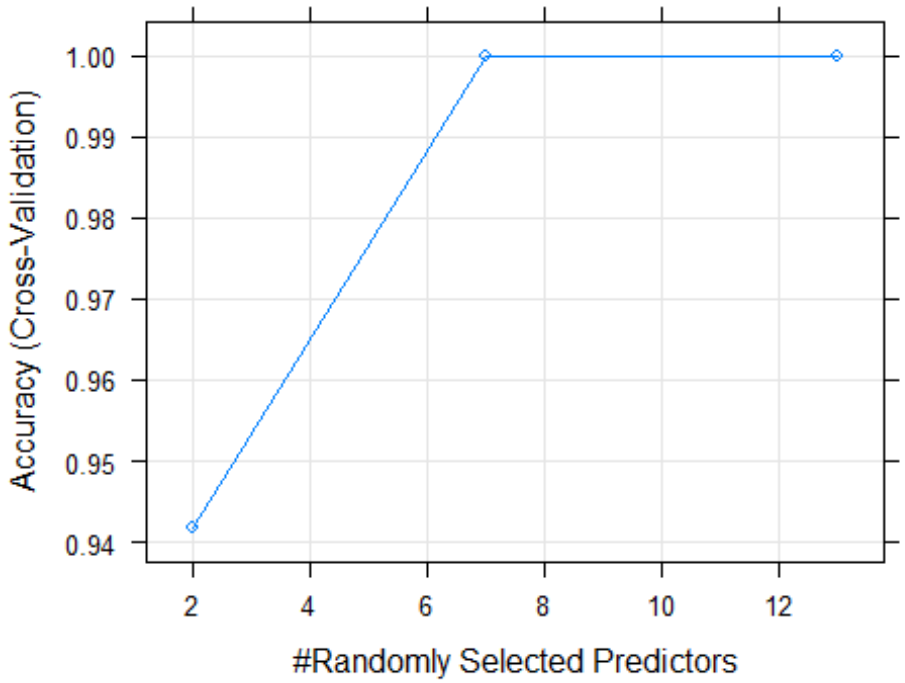
```
## [1] 1.374265
```

```
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')
```

```r
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)
```



```r
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
```

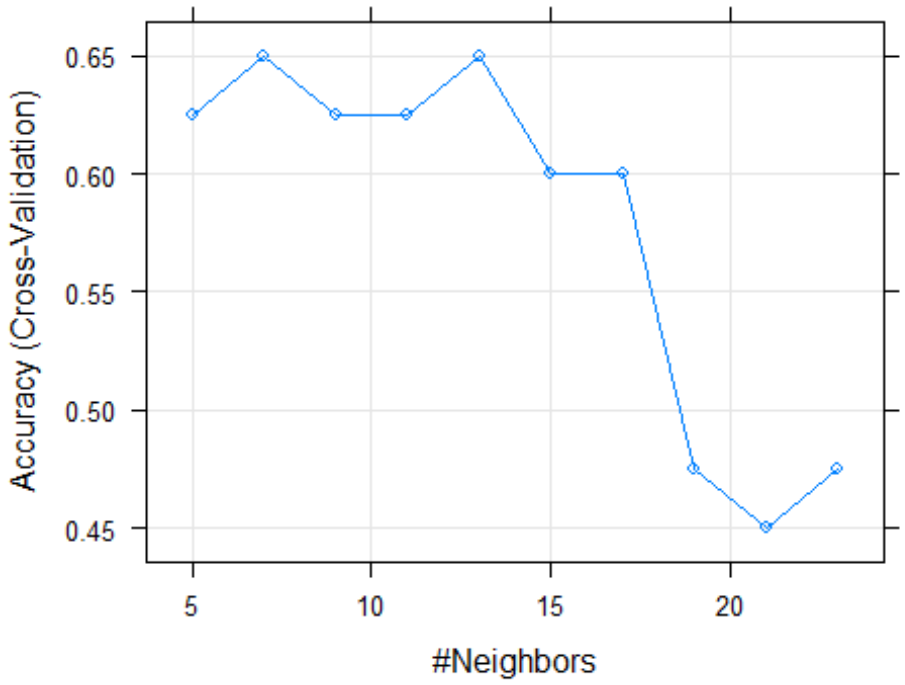
```r
sum <- sum(predRF==testingSet$ROI_Low_High)
length <- length(testingSet$ROI_Low_High)
accuracy_rfMod <- (sum/length)
accuracy_rfMod
```

```
## [1] 0.9285714
```

```r
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
```

```r
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
```