Startup Investments - Predictive Analysis

**Technical Report – Part 2**

**Predictive Analysis**

**This technical report includes technical details of the project. It includes comments, interpretations and results.**

Group 6

Vineeth Penugonda, Sai Rachana Bandi, Chandra Vardhan

04 May 2020

library(caret)  
library(doParallel)  
library(pROC)

#memory.limit(24000)  
memory.limit(10 \* 10^10)

## Warning: 'memory.limit()' is Windows-specific

## [1] Inf

# Parallelization

We paralleled the operations for faster results.

cl <- makeCluster(detectCores())  
registerDoParallel(cl)

#dat <- readRDS("../Dataset/Data\_Cleansed.rds")  
dat <- readRDS("../Dataset/Data\_CE\_Filtered.rds")  
datCopy <- dat  
#datCopy$name <- NULL

# Splitting Data

We split the data into train data and test data. We split the data into 70% train data and 30% test data.

# post\_success  
'%ni%' <- Negate('%in%') # define 'not in' func  
options(scipen=999) # prevents printing scientific notations.  
  
# Prep Training and Test data.  
set.seed(100)  
trainDataIndex <- createDataPartition(datCopy$post\_success, p=0.7, list = F) # 70% training data  
trainData <- datCopy[trainDataIndex, ]  
testData <- datCopy[-trainDataIndex, ]  
table(trainData$post\_success)

##   
## 0 1   
## 11060 1389

# Logistic Regression

# Run algorithms using 10-fold cross validation  
fitControl <- trainControl(method="cv", number= 10, allowParallel = TRUE, verboseIter = TRUE)

## Train a logistic regression model with 10-fold cross-validation  
set.seed(100)  
logit\_fit <- train(post\_success ~ ., data = trainData,  
 trControl = fitControl,  
 method="glm", family=binomial(link='logit'))

## Aggregating results  
## Fitting final model on full training set

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

print(logit\_fit)

## Generalized Linear Model   
##   
## 12449 samples  
## 58 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 11204, 11204, 11204, 11204, 11204, 11204, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9934135 0.9665532

# In-sample performance  
confusionMatrix(logit\_fit)

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 88.6 0.4  
## 1 0.3 10.8  
##   
## Accuracy (average) : 0.9934

# Out-of-sample performance  
confusionMatrix(predict(logit\_fit, newdata = testData),  
 testData$post\_success, positive = '1')

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 4738 6  
## 1 2 589  
##   
## Accuracy : 0.9985   
## 95% CI : (0.997, 0.9994)   
## No Information Rate : 0.8885   
## P-Value [Acc > NIR] : <0.0000000000000002  
##   
## Kappa : 0.9924   
##   
## Mcnemar's Test P-Value : 0.2888   
##   
## Sensitivity : 0.9899   
## Specificity : 0.9996   
## Pos Pred Value : 0.9966   
## Neg Pred Value : 0.9987   
## Prevalence : 0.1115   
## Detection Rate : 0.1104   
## Detection Prevalence : 0.1108   
## Balanced Accuracy : 0.9947   
##   
## 'Positive' Class : 1   
##

# Gradient Boosting Machine

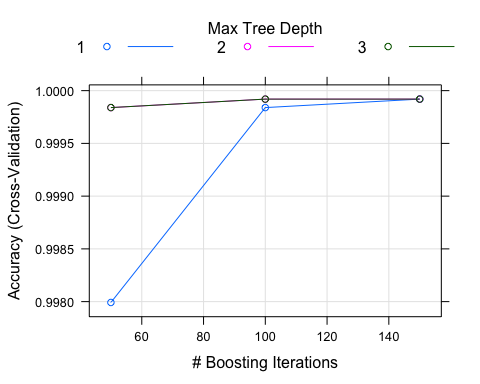
## Train a GBM model with 10-fold cross-validation  
set.seed(100)  
gbm\_fit <- train(post\_success ~ ., data = trainData,  
 trControl = fitControl, method = "gbm",  
 verbose=TRUE)

## Aggregating results  
## Selecting tuning parameters  
## Fitting n.trees = 100, interaction.depth = 2, shrinkage = 0.1, n.minobsinnode = 10 on full training set

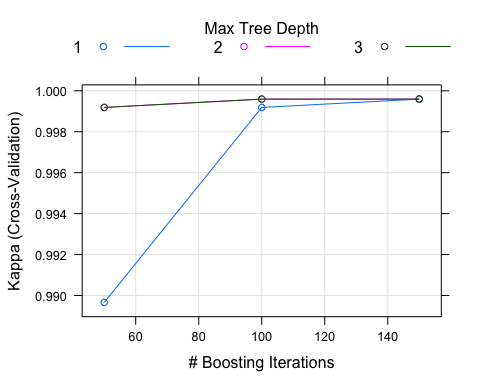
## Warning in (function (x, y, offset = NULL, misc = NULL, distribution =  
## "bernoulli", : variable 93: round\_H has no variation.

## Iter TrainDeviance ValidDeviance StepSize Improve  
## 1 0.5214 nan 0.1000 0.0889  
## 2 0.4387 nan 0.1000 0.0415  
## 3 0.3805 nan 0.1000 0.0288  
## 4 0.3356 nan 0.1000 0.0228  
## 5 0.2991 nan 0.1000 0.0182  
## 6 0.2688 nan 0.1000 0.0155  
## 7 0.2433 nan 0.1000 0.0128  
## 8 0.2213 nan 0.1000 0.0109  
## 9 0.2021 nan 0.1000 0.0094  
## 10 0.1689 nan 0.1000 0.0158  
## 20 0.0694 nan 0.1000 0.0029  
## 40 0.0213 nan 0.1000 0.0009  
## 60 0.0079 nan 0.1000 0.0000  
## 80 0.0033 nan 0.1000 0.0001  
## 100 0.0014 nan 0.1000 0.0001

# Plot resampling profile by accuracy   
plot(gbm\_fit)



# Plot resampling profile by kappa statistic   
plot(gbm\_fit, metric = "Kappa")



# In-sample performance  
confusionMatrix(gbm\_fit)

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 88.8 0.0  
## 1 0.0 11.1  
##   
## Accuracy (average) : 0.9999

# Out-of-sample performance  
confusionMatrix(predict(gbm\_fit, newdata = testData),  
 testData$post\_success, positive = '1')

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 4740 0  
## 1 0 595  
##   
## Accuracy : 1   
## 95% CI : (0.9993, 1)   
## No Information Rate : 0.8885   
## P-Value [Acc > NIR] : < 0.00000000000000022  
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0000   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 1.0000   
## Prevalence : 0.1115   
## Detection Rate : 0.1115   
## Detection Prevalence : 0.1115   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : 1   
##

# KNN

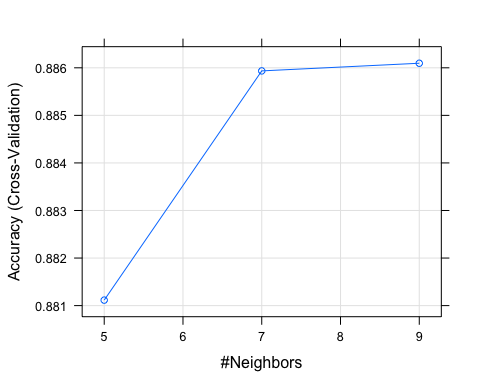
## Train a KNN model with 10-fold cross-validation  
set.seed(100)  
knn\_fit <- train(post\_success ~ ., data = trainData, method="knn", trControl=fitControl)

## Aggregating results  
## Selecting tuning parameters  
## Fitting k = 9 on full training set

# In-sample performance  
confusionMatrix(knn\_fit)

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 87.9 10.5  
## 1 0.9 0.7  
##   
## Accuracy (average) : 0.8861

# Plot accuracy across different n values  
plot(knn\_fit)



# Out-of-sample performance  
confusionMatrix(predict(knn\_fit, newdata = testData),  
 testData$post\_success, positive = '1')

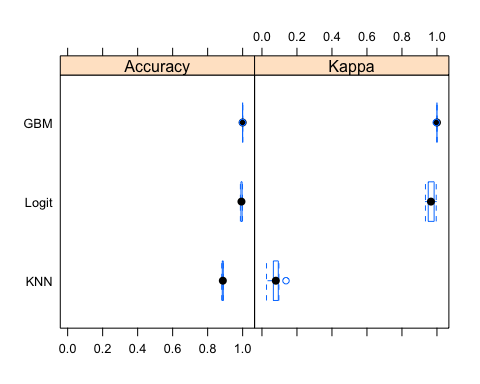
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 4693 551  
## 1 47 44  
##   
## Accuracy : 0.8879   
## 95% CI : (0.8791, 0.8963)   
## No Information Rate : 0.8885   
## P-Value [Acc > NIR] : 0.5627   
##   
## Kappa : 0.1017   
##   
## Mcnemar's Test P-Value : <0.0000000000000002  
##   
## Sensitivity : 0.073950   
## Specificity : 0.990084   
## Pos Pred Value : 0.483516   
## Neg Pred Value : 0.894928   
## Prevalence : 0.111528   
## Detection Rate : 0.008247   
## Detection Prevalence : 0.017057   
## Balanced Accuracy : 0.532017   
##   
## 'Positive' Class : 1   
##

# Comparing the models using performance metrics

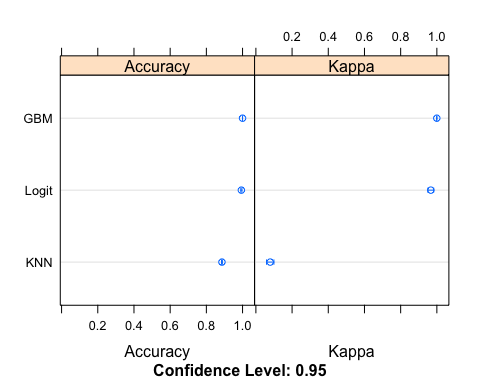
resamps <- resamples(list(Logit=logit\_fit, GBM=gbm\_fit, KNN=knn\_fit))  
  
# Summarize the resamples  
summary(resamps)

##   
## Call:  
## summary.resamples(object = resamps)  
##   
## Models: Logit, GBM, KNN   
## Number of resamples: 10   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## Logit 0.9871486 0.9903614 0.9931727 0.9934135 0.9967871 0.9991968 0  
## GBM 0.9991968 1.0000000 1.0000000 0.9999197 1.0000000 1.0000000 0  
## KNN 0.8803213 0.8837349 0.8867470 0.8860956 0.8891566 0.8898714 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## Logit 0.93438735 0.95023536 0.9657275 0.96655323 0.98380319 0.9959380 0  
## GBM 0.99593802 1.00000000 1.0000000 0.99959380 1.00000000 1.0000000 0  
## KNN 0.02602108 0.06613452 0.0793047 0.07917129 0.09098494 0.1373441 0

# Boxplots of resamples  
bwplot(resamps)



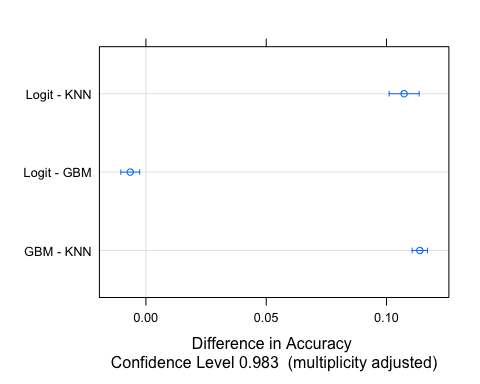
# Dot plots of resamples  
dotplot(resamps)



difValues <- diff(resamps)  
  
summary(difValues)

##   
## Call:  
## summary.diff.resamples(object = difValues)  
##   
## p-value adjustment: bonferroni   
## Upper diagonal: estimates of the difference  
## Lower diagonal: p-value for H0: difference = 0  
##   
## Accuracy   
## Logit GBM KNN   
## Logit -0.006506 0.107318  
## GBM 0.002576 0.113824  
## KNN 0.000000000006984244 0.000000000000009585   
##   
## Kappa   
## Logit GBM KNN   
## Logit -0.03304 0.88738  
## GBM 0.002654 0.92042  
## KNN 0.00000000000089689 0.00000000000001427

#bwplot(difValues, layout = c(3, 1))  
dotplot(difValues)



accu <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success)$overall['Accuracy'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success)$overall['Accuracy'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success)$overall['Accuracy'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

kappa <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success)$overall['Kappa'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success)$overall['Kappa'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success)$overall['Kappa'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

sensi <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Sensitivity'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Sensitivity'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Sensitivity'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

speci <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Specificity'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Specificity'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Specificity'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

preci <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Precision'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Precision'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['Precision'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

f1 <- c(confusionMatrix(predict(logit\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['F1'],  
 confusionMatrix(predict(gbm\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['F1'],  
 confusionMatrix(predict(knn\_fit, newdata=testData),testData$post\_success,positive='1')$byClass['F1'])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

auc <- c(roc(testData$post\_success, as.numeric(predict(logit\_fit, newdata=testData)))$auc,  
 roc(testData$post\_success, as.numeric(predict(gbm\_fit, newdata=testData)))$auc,  
 roc(testData$post\_success, as.numeric(predict(knn\_fit, newdata=testData)))$auc)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

data.frame(Accuracy = round(accu,3),  
 Kappa = round(kappa,3),  
 Precision = round(preci,3),  
 Recall = round(sensi,3),   
 Specificity = round(speci,3),  
 F1 = round(f1,3),  
 AUC = round(auc,3),  
 Balanced.Accuracy = round((sensi + speci)/2,3),  
 row.names = c('Logit','GBM','KNN'))

## Accuracy Kappa Precision Recall Specificity F1 AUC Balanced.Accuracy  
## Logit 0.999 0.992 0.997 0.990 1.000 0.993 0.995 0.995  
## GBM 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
## KNN 0.888 0.107 0.489 0.074 0.991 0.126 0.532 0.532

logit\_fit\_roc <- roc(testData$post\_success, as.numeric(predict(logit\_fit, newdata=testData)))

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

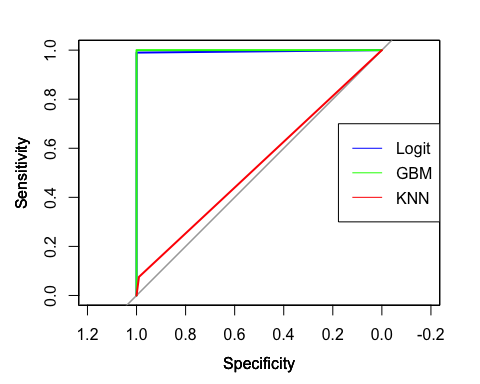
gbm\_fit\_roc <- roc(testData$post\_success, as.numeric(predict(gbm\_fit, newdata=testData)))

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

knn\_fit\_roc <- roc(testData$post\_success, as.numeric(predict(knn\_fit, newdata=testData)))

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(logit\_fit\_roc, col = "blue")  
par(new = TRUE)  
plot(gbm\_fit\_roc, col = "green", xaxt = "n", yaxt = "n")  
par(new = TRUE)  
plot(knn\_fit\_roc, col = "red", xaxt = "n", yaxt = "n")  
  
legend("right", legend = c("Logit", "GBM", "KNN"), col = c("blue", "green", "red"), lty = 1)



# RESULTS

We used 3 algorithms, namely the KNN , logistic regression and the gradient boosting model (GBM) to train 3 different machine learning models. The features obtained during the data preprocessing phase were used to fit the 3 different models.

The AUC score for each of the machine learning models were compared and the model with the maximum AUC score is considered for the given dataset.

Based on AUC score, we can observe that our best model is Gradient Boosting Machine (GBM). Logistic Regression (GLM) is the second best model.

stopCluster(cl)