Problem: Predicting whether Customer Will Default or Not on Loan for Bank Client

Solution: Supervised Classification with Logistic Regression using Lasso for testing with most AUC.

#install.packages("caret") #this may take a while

library(caret)

library(tidyverse)

loan\_data <- read.csv(file = "loan\_default\_dataset.csv", header=T)

loan\_data = read.csv(file=file.choose(),header=TRUE)

table(loan\_data$Default)

# we have 1000 observations and 700 are non-default and 300 are default

#boxplot(loan\_data$~ loan\_data$ )

#install.packages("skimr")

library(skimr)

summaryStats <- skim(loan\_data)

summaryStats

boxplot(loan\_data$Term, ylab= 'Term')

boxplot(loan\_data$Checking\_amount, ylab='Checking\_amount' )

boxplot(loan\_data$Age, ylab= 'Age')

boxplot(loan\_data$Credit\_score, ylab= 'Credit\_score')

boxplot(loan\_data$Amount, ylab= 'Amount')

boxplot(loan\_data$Saving\_amount, ylab= 'Saving\_amount')

boxplot(loan\_data$No\_of\_credit\_acc, ylab= '# Credit Account')

boxplot(loan\_data$Emp\_duration, ylab= 'Emp\_duration')

boxplot(loan\_data$No\_of\_credit\_acc~loan\_data$Default)

boxplot(loan\_data$Amount ~loan\_data$Default)

boxplot(loan\_data$Amount ~loan\_data$Emp\_status)

#Step 1 partition and preprocessing

library("tidyverse")

dummies\_model<-dummyVars(Default~.,data=loan\_data)

Default\_predictors\_dummy<-data.frame(predict(dummies\_model,newdata=loan\_data))

loan\_data<-cbind(Default=loan\_data$Default,Default\_predictors\_dummy)

####Logistic Regression####

####Logistic Regression####

loan\_data$Default<-as.factor(loan\_data$Default)

loan\_data$Default<-fct\_recode(loan\_data$Default, notDefault = "0", Default = "1")

loan\_predictors\_dummy <- model.matrix(Default~ ., data = loan\_data)#create dummy variables expect for the response

loan\_predictors\_dummy<- data.frame(loan\_predictors\_dummy[,-1]) #get rid of intercept

loan\_data <- cbind(Default=loan\_data$Default, loan\_predictors\_dummy)

library(caret)

set.seed(99) #set random seed

index <- createDataPartition(loan\_data$Default, p = .8,list = FALSE)

loan\_train <-loan\_data[index,]

loan\_test <- loan\_data[-index,]

##Step 2: Train or Fit LASSO Logistic Regression Model

# install and load packages for machine learning model

library(e1071)

library(glmnet)

library(Matrix)

set.seed(10)#set the seed again since within the train method the validation set is randomly selected

loan\_model <- train(Default ~ .,

data = loan\_train,

method = "glmnet",

standardize =T,

tuneGrid = expand.grid(alpha =1, #lasso

lambda = seq(0.0001, 1, length = 20)),

trControl =trainControl(method = "cv",

number = 5,

classProbs = TRUE,

summaryFunction = twoClassSummary),

metric="ROC")

loan\_model

#list coefficients selected

coef(loan\_model$finalModel, loan\_model$bestTune$lambda)

##Step 3: Get Predictions using Testing Set Data

#First, get the predicted probabilities of the test data.

predprob\_lasso<-predict(loan\_model , loan\_test, type="prob")

#Step 4: Evaluate Model Performance

#install.packages("ROCR")

library(ROCR)

pred\_lasso <- prediction(predprob\_lasso$Default, loan\_test$Default,label.ordering =c("notDefault","Default") )

perf\_lasso <- performance(pred\_lasso, "tpr", "fpr")

plot(perf\_lasso, colorize=TRUE)

#Get the AUC

auc\_lasso<-unlist(slot(performance(pred\_lasso, "auc"), "y.values"))

auc\_lasso