FRA_Logistic_NO_SMote

Sudipta

June 18, 2016

```
#Setup working Directory

setwd("/Users/sudiptamondal/Documents/BABI Program/Financial Analytics/FRA")
getwd()
```

```
## [1] "/Users/sudiptamondal/Documents/BABI Program/Financial Analytics/FRA"
```

```
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
```

```
##
```

```
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
```

```
##
## Attaching package: 'gdata'
```

```
## The following object is masked from 'package:stats':
##
##     nobs
```

```
## The following object is masked from 'package:utils':
##
##     object.size
```

```
Train=read.xls("train.xlsx", sheet=2, header = TRUE)
Test=read.xls("test.xlsx",sheet=2 , header = TRUE)
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.2.5
```

```
Train <- rename(Train,
                  c("SeriousDlqin2yrs" = "Default",
                    "RevolvingUtilizationOfUnsecuredLines" = "RevUtiUL",
                    "NumberOfOpenCreditLinesAndLoans" = "NoOpenCCLoans",
                    "NumberOfDependents" = "NoOfDep"
                    ))

Test <- rename(Test,
                c("SeriousDlqin2yrs" = "Default",
                  "RevolvingUtilizationOfUnsecuredLines" = "RevUtiUL",
                  "NumberOfOpenCreditLinesAndLoans" = "NoOpenCCLoans",
                  "NumberOfDependents" = "NoOfDep"
                ))

library(mice)
```

```
## Loading required package: Rcpp
```

```
## Warning: package 'Rcpp' was built under R version 3.2.5
```

```
## mice 2.25 2015-11-09
```

```
md.pattern(Train)
```

```
##      Casenum Default RevUtiUL DebtRatio NoOpenCCLoans NoOfDep
## 4858       1       1        1         1             1       1   0
##  142       1       1        1         1             1       0   1
##            0       0        0         0             0     142 142
```

```
md.pattern(Test)
```

```
##     Casenum Default RevUtiUL DebtRatio NoOpenCCLoans NoOfDep
## 980       1       1        1         1             1       1  0
##  20       1       1        1         1             1       0  1
##           0       0        0         0             0      20 20
```

```
Train_A = Train[c(1,3:6)]
Train_B = Train[c(1:2)]
Test_A = Test[c(1,3:6)]
Test_B = Test[c(1:2)]

Total_A <- rbind(Train_A, Test_A)

summary(Total_A$RevUtiUL)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.031    0.169    3.260    0.566 6324.000
```

```
qn1 = quantile(Total_A$RevUtiUL, c(0.05, 0.95), na.rm = TRUE)
#qn1
Total_A = within(Total_A,
                 { RevUtiUL = ifelse(RevUtiUL < qn1[1], qn1[1], RevUtiUL)
                   RevUtiUL = ifelse(RevUtiUL > qn1[2], qn1[2], RevUtiUL)})
#summary(Total_A$RevUtiUL)

#summary(Total_A$DebtRatio)
qn2 = quantile(Total_A$DebtRatio, c(0.05, 0.95), na.rm = TRUE)
#qn2
Total_A = within(Total_A,
                 { DebtRatio = ifelse(DebtRatio < qn2[1], qn2[1], DebtRatio)
                 DebtRatio = ifelse(DebtRatio > qn2[2], qn2[2], DebtRatio)})
summary(Total_A$DebtRatio)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0047    0.1763    0.3672  254.4000    0.8285 2441.0000
```

```
summary(Total_A$NoOpenCCLoans)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.000   8.000   8.418  11.000  46.000
```

```
qn3 = quantile(Total_A$NoOpenCCLoans, c(0.05, 0.95), na.rm = TRUE)
#qn3
Total_A = within(Total_A,
                 { NoOpenCCLoans = ifelse(NoOpenCCLoans < qn3[1], qn3[1], NoOpenCCLoa
ns)
                 NoOpenCCLoans = ifelse(NoOpenCCLoans > qn3[2], qn3[2], NoOpenCCLoans
)})

library(mice)
imp_Total_A <- mice(Total_A,print=FALSE)
imp_Total_A$predictorMatrix
```

```
##                Casenum RevUtiUL DebtRatio NoOpenCCLoans NoOfDep
## Casenum              0        0         0              0       0
## RevUtiUL             0        0         0              0       0
## DebtRatio            0        0         0              0       0
## NoOpenCCLoans        0        0         0              0       0
## NoOfDep              1        1         1              1       0
```

```
pred <- imp_Total_A$predictorMatrix
pred[,"Casenum"] <- 0
imp_total <- mice(Total_A, pred=pred, print=FALSE)
Total_A = complete(imp_Total_A)

Train_F = merge(Total_A,Train_B, by='Casenum')
Test_F = merge(Total_A, Test_B, by= 'Casenum')


# Verify the % Defaulter from Train and test file

dim(Train_F);dim(Test_F)
```

```
## [1] 5000    6
```

```
## [1] 1000    6
```

```
table(Train_F$Default)
```

```
##
##    0    1
## 4695  305
```

```
table(Test_F$Default)
```

```
##
##   0   1
## 937  63
```

```
Train_X <- Train_F
Test_X <- Test_F
# Information Value Calculation
#install_github("riv","tomasgreif")
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 3.2.5
```

```
library(woe)
row.names(Train_X) <- 1:nrow(Train_X)
iv.mult(Train_X,"Default",TRUE)
```

```
## Loading required package: DBI
```

```
## Warning: package 'DBI' was built under R version 3.2.5
```

```
## Loading required package: tcltk
```

```
## Information Value 0
## Information Value 1.04
## Information Value 0.15
## Information Value 0.03
## Information Value 0.03
```

```
##        Variable InformationValue Bins ZeroBins   Strength
## 1      RevUtiUL       1.03829308    3        0 Suspicious
## 2     DebtRatio       0.15265407    3        0    Average
## 3       NoOfDep       0.02862935    2        0       Weak
## 4 NoOpenCCLoans       0.02820952    2        0       Weak
## 5       Casenum       0.00000000    1        0  Wery weak
```

```
modeltrain = glm(Default ~
                   RevUtiUL
#                 + DebtRatio
                  + NoOpenCCLoans
#                  + NoOfDep
                 , family=binomial,data=Train_X)
summary(modeltrain)
```

```
##
## Call:
## glm(formula = Default ~ RevUtiUL + NoOpenCCLoans, family = binomial,
##     data = Train_X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8539  -0.3503  -0.2170  -0.1790   2.9783
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.50650    0.19552  -23.05  < 2e-16 ***
## RevUtiUL       2.93429    0.18110   16.20  < 2e-16 ***
## NoOpenCCLoans  0.04172    0.01381    3.02  0.00253 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2297.1  on 4999  degrees of freedom
## Residual deviance: 1983.5  on 4997  degrees of freedom
## AIC: 1989.5
##
## Number of Fisher Scoring iterations: 6
```

```
#Detecting multicolinearity and removing the variables
#install.packages("car")
library('car')
```

```
## Warning: package 'car' was built under R version 3.2.4
```

```
vif(modeltrain)
```

```
##      RevUtiUL NoOpenCCLoans
##      1.118518      1.118518
```

```
##Train Data
predictTrain=predict(modeltrain,data=Train_X,type="response")
predictDecisionTr = predictTrain > 0.5
confusion_Train = table(Train_X$Default,predictDecisionTr)
confusion_Train
```

```
##    predictDecisionTr
##     FALSE
##   0  4695
##   1   305
```

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
ROCRTrain = prediction(predictTrain,Train_X$Default)
ROCRTrain_perf = performance(ROCRTrain,"tpr","fpr")
auc_Train <- performance(ROCRTrain,"auc")
auc_Train <- as.numeric(auc_Train@y.values)
KS_Train <- max(attr(ROCRTrain_perf, 'y.values')[[1]]-attr(ROCRTrain_perf, 'x.values'
)[[1]])

#install.packages("ineq")
library('ineq')
gini_Train = ineq(predictTrain, type="Gini")
auc_Train
```

```
## [1] 0.7699495
```

```
KS_Train
```

```
## [1] 0.4778226
```

```
gini_Train
```

```
## [1] 0.5234289
```

```
sum(diag(prop.table(confusion_Train)))
```

```
## [1] 0.939
```

```
##Test Data
predictTest=predict(modeltrain,newdata=Test_X,type="response")
predictDecision = predictTest > 0.5
confusion_Test = table(Test_X$Default,predictDecision)
confusion_Test
```

```
##    predictDecision
##     FALSE
##   0   937
##   1    63
```

```
library(ROCR)
ROCRTest = prediction(predictTest,Test_X$Default)
ROCRTest_perf = performance(ROCRTest,"tpr","fpr")
auc_Test <- performance(ROCRTest,"auc")
auc_Test <- as.numeric(auc_Test@y.values)
KS_Test <- max(attr(ROCRTest_perf, 'y.values')[[1]]-attr(ROCRTest_perf, 'x.values')[[
1]])

library('ineq')
gini_Test = ineq(predictTest, type="Gini")

auc_Test
```

```
## [1] 0.7425166
```

```
KS_Test
```

```
## [1] 0.4408023
```

```
gini_Test
```

```
## [1] 0.5133228
```

```
sum(diag(prop.table(confusion_Test)))
```

```
## [1] 0.937
```