

Bank fraud machine learning

simi

2024-08-17

R Markdown

```
#load packages  
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
#Load the train and test dataset
```

```
train<-read.csv("/Users/user/Music/machine learning R/fraud_train.csv")
```

```
test<-read.csv("/Users/user/Music/machine learning R/fraud_test.csv")
```

```
#create column to differentiate train and test dataset
```

```
train$isTrain<-"yes"
```

```
test$isTrain<-"no"
```

```
#combine train and test datasets
```

```
comb<-rbind(train,test)
```

```
missing_values<-colSums(is.na(comb))
```

```
missing_values
```

```
##   Sector_score  LOCATION_ID  PARA_A  Score_A  Risk_A  
##           9           0        20        29        12  
##   PARA_B      Score_B      Risk_B      TOTAL      numbers  
##        26        36        17        27        26  
##   Risk_C  Money_Value  Score_MV      Risk_D  District_Loss  
##        38        24        21         8        27  
##   PROB      RiSk_E      History      Prob      Risk_F  
##        30        11        49        22        45  
##   Score  Inherent_Risk  CONTROL_RISK  Detection_Risk  Audit_Risk  
##        21        32        12        52        11  
##   Risk      isTrain  
##        0           0
```

```
names(comb)
```

```
## [1] "Sector_score" "LOCATION_ID" "PARA_A" "Score_A"  
## [5] "Risk_A" "PARA_B" "Score_B" "Risk_B"  
## [9] "TOTAL" "numbers" "Risk_C" "Money_Value"  
## [13] "Score_MV" "Risk_D" "District_Loss" "PROB"  
## [17] "RiSk_E" "History" "Prob" "Risk_F"  
## [21] "Score" "Inherent_Risk" "CONTROL_RISK" "Detection_Risk"  
## [25] "Audit_Risk" "Risk" "isTrain"
```

```
features<-c("Sector_score", "PARA_A", "Score_A", "Risk_A", "PARA_B", "Score_B", "Risk_B", "TOTAL", "numbers")
```

```
comb<-comb[,features]  
str(comb)
```

```
## 'data.frame': 775 obs. of 26 variables:  
## $ Sector_score : num 3.89 3.89 3.89 3.89 3.89 3.89 3.89 3.89 3.89 3.89 ...  
## $ PARA_A : num 0 0 0 NA 1.1 8.4 NA 5.47 0 1.95 ...  
## $ Score_A : num NA 0.2 0.2 0.2 0.4 0.6 0.6 0.6 0.2 0.4 ...  
## $ Risk_A : num 0 0 0 0 0.44 ...  
## $ PARA_B : num 4.83 10.8 0.08 0.83 7.41 ...  
## $ Score_B : num 0.2 0.6 0.2 0.2 0.4 0.6 0.2 0.4 0.2 0.4 ...  
## $ Risk_B : num 0.966 6.48 0.016 0.166 2.964 ...  
## $ TOTAL : num 4.83 10.8 0.08 0.83 8.51 ...  
## $ numbers : num 5 6 5 5 5 5.5 5 5 5 5 ...  
## $ Risk_C : num 1 3.6 1 1 1 2.2 1 1 1 1 ...  
## $ Money_Value : num 0.94 11.75 0 2.95 44.95 ...  
## $ Score_MV : num 0.2 0.6 0.2 0.2 0.6 0.4 0.2 0.6 0.2 0.4 ...  
## $ Risk_D : num 0.188 7.05 0 0.59 26.97 ...  
## $ District_Loss : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ PROB : num 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 ...  
## $ RiSk_E : num 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 ...  
## $ History : int 0 0 0 0 0 NA 0 1 0 0 ...  
## $ Prob : num 0.2 NA 0.2 0.2 0.2 0.2 0.2 0.4 NA 0.2 ...  
## $ Risk_F : num 0 0 0 0 0 0 0 0.4 0 0 ...  
## $ Score : num 2 4.4 NA 2 NA 4.2 2.4 3.6 2 3 ...  
## $ Inherent_Risk : num 2.55 17.53 1.42 2.16 31.77 ...  
## $ CONTROL_RISK : num 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.8 0.4 0.4 ...  
## $ Detection_Risk : num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 NA ...  
## $ Audit_Risk : num 0.511 3.506 0.283 0.431 6.355 ...  
## $ Risk : chr "not.fraudulent" "fraudulent" "not.fraudulent" "not.fraudulent" ...  
## $ isTrain : chr "yes" "yes" "yes" "yes" ...
```

```
# Impute missing values  
pre.process<-preProcess(comb,method = "bagImpute")  
imputed.data<-predict(pre.process,comb)
```

```
comb<-imputed.data  
comb$Risk<-as.factor(comb$Risk)
```

```
#split dataset back to train and test set
train<-comb[comb$isTrain=="yes",]

test<-comb[comb$isTrain=="no",]
```

```
#remove variable isTrain from both train and test set

train$isTrain<-NULL

test$isTrain<-NULL
```

```
##modeling #Logistic regression
```

```
set.seed(8819)
control <- trainControl(
  method = "cv",           # Cross-validation
  number = 10,             # 10-fold CV
  classProbs = TRUE,       # Compute class probabilities
  summaryFunction = twoClassSummary, # Calculate ROC, Sensitivity, etc.
  savePredictions = TRUE
)

model.logit <- train(
  Risk~.,
  data = train,
  method = "glm",
  family = "binomial",
  trControl = control,
  preProcess = c("zv", "center", "scale"),
  tuneLength = 16,
  metric="ROC"
)
```

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

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

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

```
pred_y = predict(model.logit,test)
```

```

#create confusion matrix
conf_mat = confusionMatrix(pred_y,test$Risk)
print(conf_mat)

## Confusion Matrix and Statistics
##
##               Reference
## Prediction      fraudulent not.fraudulent
##   fraudulent           81           2
##   not.fraudulent       10          139
##
##               Accuracy : 0.9483
##               95% CI : (0.9114, 0.973)
##   No Information Rate : 0.6078
##   P-Value [Acc > NIR] : < 2e-16
##
##               Kappa : 0.8898
##
##   Mcnemar's Test P-Value : 0.04331
##
##               Sensitivity : 0.8901
##               Specificity : 0.9858
##               Pos Pred Value : 0.9759
##               Neg Pred Value : 0.9329
##               Prevalence : 0.3922
##               Detection Rate : 0.3491
##   Detection Prevalence : 0.3578
##               Balanced Accuracy : 0.9380
##
##               'Positive' Class : fraudulent
##

#random forest

set.seed(8819)

control <- trainControl(
  method = "cv",           # Cross-validation
  number = 10,             # 10-fold CV
  classProbs = TRUE,       # Compute class probabilities
  summaryFunction = twoClassSummary, # Calculate ROC, Sensitivity, etc.
  savePredictions = TRUE
)

model.rf <- train(
  Risk~.,
  data = train,
  method = "rf",
  metric="ROC",
  trControl = control,
  tuneLength = 16
)

```

```
pred_y = predict(model.rf,test)
```

```
#create confusion matrix  
conf_mat = confusionMatrix(pred_y,test$Risk)  
print(conf_mat)
```

```
## Confusion Matrix and Statistics  
##  
##              Reference  
## Prediction    fraudulent not.fraudulent  
##   fraudulent           91           0  
## not.fraudulent         0          141  
##  
##              Accuracy : 1  
##              95% CI : (0.9842, 1)  
##   No Information Rate : 0.6078  
##   P-Value [Acc > NIR] : < 2.2e-16  
##  
##              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 : 0.3922  
##              Detection Rate : 0.3922  
##   Detection Prevalence : 0.3922  
##   Balanced Accuracy : 1.0000  
##  
##   'Positive' Class : fraudulent  
##
```

```
#Decision tree
```

```
set.seed(8819)  
  
control <- trainControl(  
  method = "cv",           # Cross-validation  
  number = 10,             # 10-fold CV  
  classProbs = TRUE,       # Compute class probabilities  
  summaryFunction = twoClassSummary, # Calculate ROC, Sensitivity, etc.  
  savePredictions = TRUE  
)  
  
model.dtree <- train(  
  Risk~.,  
  data = train,  
  method = "rpart",  
  trControl = control,
```

```
tuneLength = 16
)
```

```
## Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
## in the result set. ROC will be used instead.
```

```
pred_y = predict(model.dtree,test)

#create confusion matrix
conf_mat = confusionMatrix(pred_y,test$Risk)
print(conf_mat)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    fraudulent not.fraudulent
##  fraudulent           0           0
##  not.fraudulent       91          141
##
##              Accuracy : 0.6078
##              95% CI : (0.5417, 0.671)
##    No Information Rate : 0.6078
##    P-Value [Acc > NIR] : 0.5287
##
##              Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.0000
##              Specificity : 1.0000
##              Pos Pred Value :   NaN
##              Neg Pred Value : 0.6078
##              Prevalence : 0.3922
##              Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##              Balanced Accuracy : 0.5000
##
##              'Positive' Class : fraudulent
##
```

```
library(fastAdaboost)
set.seed(8819)

model.adaboost <- train(
  Risk~.,
  data = train,
  method = "adaboost",
  trControl = control,
  metric="ROC",
  tuneLength = 16
)
```

```

pred_y = predict(model.adaboost,test)

#create confusion matrix
conf_mat = confusionMatrix(pred_y,test$Risk)
print(conf_mat)

```

```

## Confusion Matrix and Statistics
##
##              Reference
## Prediction    fraudulent not.fraudulent
##  fraudulent           91             0
## not.fraudulent         0            141
##
##              Accuracy : 1
##              95% CI : (0.9842, 1)
##    No Information Rate : 0.6078
##    P-Value [Acc > NIR] : < 2.2e-16
##
##              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 : 0.3922
##              Detection Rate : 0.3922
##    Detection Prevalence : 0.3922
##              Balanced Accuracy : 1.0000
##
##              'Positive' Class : fraudulent
##

```

#SVM

```

set.seed(8819)

model.SVM <- train(
  Risk~.,
  data = train,
  method = "svmRadial",
  trControl = control,
  metric="ROC",
  preProcess = c("center","scale"),
  tuneLength = 16
)

```

```

## Warning in preProcess.default(method = c("center", "scale"), x =
## structure(c(3.89, : These variables have zero variances: Detection_Risk

## Warning in .local(x, ...): Variable(s) ' ' constant. Cannot scale data.

```


[illegible]

[illegible]

```

pred_y = predict(model.SVM,test)

#create confusion matrix
conf_mat = confusionMatrix(pred_y,test$Risk)
print(conf_mat)

## Confusion Matrix and Statistics
##
##              Reference
## Prediction    fraudulent not.fraudulent
##   fraudulent           86           4
##   not.fraudulent        5          137
##
##              Accuracy : 0.9612
##              95% CI : (0.9276, 0.9821)
##   No Information Rate : 0.6078
##   P-Value [Acc > NIR] : <2e-16
##
##              Kappa : 0.9185
##
##  Mcnemar's Test P-Value : 1
##
##              Sensitivity : 0.9451
##              Specificity : 0.9716
##              Pos Pred Value : 0.9556
##              Neg Pred Value : 0.9648
##              Prevalence : 0.3922
##              Detection Rate : 0.3707
##              Detection Prevalence : 0.3879
##              Balanced Accuracy : 0.9583
##
##              'Positive' Class : fraudulent
##

```

#Naive Bayes

```

set.seed(8819)
library(naivebayes)

```

```
## naivebayes 1.0.0 loaded
```

```
## For more information please visit:
```

```
## https://majkamichal.github.io/naivebayes/
```

```

model.Naive_bayes <- train(
  Risk~.,
  data = train,
  method = "naive_bayes",
  trControl = control,
  metric="ROC",

```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
## extra argument 'PreProcess' will be disregarded
```

```
pred_y = predict(model.Naive_bayes,test)

#create confusion matrix
conf_mat = confusionMatrix(pred_y,test$Risk)
print(conf_mat)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    fraudulent not.fraudulent
##  fraudulent           90           5
## not.fraudulent         1          136
##
##              Accuracy : 0.9741
##              95% CI : (0.9446, 0.9905)
##    No Information Rate : 0.6078
##    P-Value [Acc > NIR] : <2e-16
##
##              Kappa : 0.9462
##
##  McNemar's Test P-Value : 0.2207
##
##              Sensitivity : 0.9890
##              Specificity : 0.9645
##              Pos Pred Value : 0.9474
##              Neg Pred Value : 0.9927
##              Prevalence : 0.3922
##              Detection Rate : 0.3879
##    Detection Prevalence : 0.4095
##              Balanced Accuracy : 0.9768
##
##              'Positive' Class : fraudulent
##
```