Credit\_Card

Chinki

June 15, 2017

source : <https://www.kaggle.com/dalpozz/creditcardfraud>

library(readr)  
creditcard <- read\_csv("C:/Computational Statistics/Kaggle/Credit card/creditcard.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## Time = col\_integer(),  
## Class = col\_integer()  
## )

## See spec(...) for full column specifications.

## Warning: 1 parsing failure.  
## row col expected actual file  
## 153759 Time no trailing characters .00E+05 'C:/Computational Statistics/Kaggle/Credit card/creditcard.csv'

creditcard$Class=as.factor(creditcard$Class)

#Structure of the data  
str(creditcard)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 284807 obs. of 31 variables:  
## $ Time : int 0 0 1 1 2 2 4 7 7 9 ...  
## $ V1 : num -1.36 1.192 -1.358 -0.966 -1.158 ...  
## $ V2 : num -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...  
## $ V3 : num 2.536 0.166 1.773 1.793 1.549 ...  
## $ V4 : num 1.378 0.448 0.38 -0.863 0.403 ...  
## $ V5 : num -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...  
## $ V6 : num 0.4624 -0.0824 1.8005 1.2472 0.0959 ...  
## $ V7 : num 0.2396 -0.0788 0.7915 0.2376 0.5929 ...  
## $ V8 : num 0.0987 0.0851 0.2477 0.3774 -0.2705 ...  
## $ V9 : num 0.364 -0.255 -1.515 -1.387 0.818 ...  
## $ V10 : num 0.0908 -0.167 0.2076 -0.055 0.7531 ...  
## $ V11 : num -0.552 1.613 0.625 -0.226 -0.823 ...  
## $ V12 : num -0.6178 1.0652 0.0661 0.1782 0.5382 ...  
## $ V13 : num -0.991 0.489 0.717 0.508 1.346 ...  
## $ V14 : num -0.311 -0.144 -0.166 -0.288 -1.12 ...  
## $ V15 : num 1.468 0.636 2.346 -0.631 0.175 ...  
## $ V16 : num -0.47 0.464 -2.89 -1.06 -0.451 ...  
## $ V17 : num 0.208 -0.115 1.11 -0.684 -0.237 ...  
## $ V18 : num 0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...  
## $ V19 : num 0.404 -0.146 -2.262 -1.233 0.803 ...  
## $ V20 : num 0.2514 -0.0691 0.525 -0.208 0.4085 ...  
## $ V21 : num -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...  
## $ V22 : num 0.27784 -0.63867 0.77168 0.00527 0.79828 ...  
## $ V23 : num -0.11 0.101 0.909 -0.19 -0.137 ...  
## $ V24 : num 0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...  
## $ V25 : num 0.129 0.167 -0.328 0.647 -0.206 ...  
## $ V26 : num -0.189 0.126 -0.139 -0.222 0.502 ...  
## $ V27 : num 0.13356 -0.00898 -0.05535 0.06272 0.21942 ...  
## $ V28 : num -0.0211 0.0147 -0.0598 0.0615 0.2152 ...  
## $ Amount: num 149.62 2.69 378.66 123.5 69.99 ...  
## $ Class : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## - attr(\*, "problems")=Classes 'tbl\_df', 'tbl' and 'data.frame': 1 obs. of 5 variables:  
## ..$ row : int 153759  
## ..$ col : chr "Time"  
## ..$ expected: chr "no trailing characters"  
## ..$ actual : chr ".00E+05"  
## ..$ file : chr "'C:/Computational Statistics/Kaggle/Credit card/creditcard.csv'"  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 31  
## .. ..$ Time : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ V1 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V2 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V3 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V4 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V5 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V6 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V7 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V8 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V9 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V10 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V11 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V12 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V13 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V14 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V15 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V16 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V17 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V18 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V19 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V20 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V21 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V22 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V23 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V24 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V25 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V26 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V27 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ V28 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Amount: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Class : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

#Table of class   
table(creditcard$Class)

##   
## 0 1   
## 284315 492

#Praportion of fraud   
prop.table(table(creditcard$Class))

##   
## 0 1   
## 0.998272514 0.001727486

set.seed(12)  
training=sample(284807,200000)  
credit\_training=creditcard[training,]  
credit\_test=creditcard[-training,]

Training a model on the data

#C50 model  
library(C50)  
model=C5.0(credit\_training[-31],credit\_training$Class)  
model

##   
## Call:  
## C5.0.default(x = credit\_training[-31], y = credit\_training$Class)  
##   
## Classification Tree  
## Number of samples: 200000   
## Number of predictors: 30   
##   
## Tree size: 7   
##   
## Non-standard options: attempt to group attributes

summary(model)

##   
## Call:  
## C5.0.default(x = credit\_training[-31], y = credit\_training$Class)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Tue Jun 20 08:07:17 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 200000 cases (31 attributes) from undefined.data  
##   
## Decision tree:  
##   
## V17 > -2.515183: 0 (199660/99)  
## V17 <= -2.515183:  
## :...V10 > 1.291319: 0 (33)  
## V10 <= 1.291319:  
## :...V14 > -3.182065:  
## :...V8 <= -0.3873986: 1 (6)  
## : V8 > -0.3873986: 0 (28/5)  
## V14 <= -3.182065:  
## :...V27 <= 0.8593842: 1 (157/3)  
## V27 > 0.8593842:  
## :...V26 <= -0.2712055: 0 (30/3)  
## V26 > -0.2712055: 1 (86/10)  
##   
##   
## Evaluation on training data (200000 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 7 120( 0.1%) <<  
##   
##   
## (a) (b) <-classified as  
## ----- -----  
## 199644 13 (a): class 0  
## 107 236 (b): class 1  
##   
##   
## Attribute usage:  
##   
## 100.00% V17  
## 0.17% V10  
## 0.15% V14  
## 0.14% V27  
## 0.06% V26  
## 0.02% V8  
##   
##   
## Time: 13.4 secs

pred=predict(model,credit\_test)  
library(gmodels)  
CrossTable(pred,credit\_test$Class)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 84807   
##   
##   
## | credit\_test$Class   
## pred | 0 | 1 | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 84654 | 53 | 84707 |   
## | 0.109 | 61.699 | |   
## | 0.999 | 0.001 | 0.999 |   
## | 1.000 | 0.356 | |   
## | 0.998 | 0.001 | |   
## -------------|-----------|-----------|-----------|  
## 1 | 4 | 96 | 100 |   
## | 91.985 | 52263.297 | |   
## | 0.040 | 0.960 | 0.001 |   
## | 0.000 | 0.644 | |   
## | 0.000 | 0.001 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 84658 | 149 | 84807 |   
## | 0.998 | 0.002 | |   
## -------------|-----------|-----------|-----------|  
##   
##

model1=C5.0(credit\_training[-31],credit\_training$Class,trails=100)  
model1

##   
## Call:  
## C5.0.default(x = credit\_training[-31], y = credit\_training$Class, trails  
## = 100)  
##   
## Classification Tree  
## Number of samples: 200000   
## Number of predictors: 30   
##   
## Tree size: 7   
##   
## Non-standard options: attempt to group attributes

summary(model1)

##   
## Call:  
## C5.0.default(x = credit\_training[-31], y = credit\_training$Class, trails  
## = 100)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Tue Jun 20 08:08:54 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 200000 cases (31 attributes) from undefined.data  
##   
## Decision tree:  
##   
## V17 > -2.515183: 0 (199660/99)  
## V17 <= -2.515183:  
## :...V10 > 1.291319: 0 (33)  
## V10 <= 1.291319:  
## :...V14 > -3.182065:  
## :...V8 <= -0.3873986: 1 (6)  
## : V8 > -0.3873986: 0 (28/5)  
## V14 <= -3.182065:  
## :...V27 <= 0.8593842: 1 (157/3)  
## V27 > 0.8593842:  
## :...V26 <= -0.2712055: 0 (30/3)  
## V26 > -0.2712055: 1 (86/10)  
##   
##   
## Evaluation on training data (200000 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 7 120( 0.1%) <<  
##   
##   
## (a) (b) <-classified as  
## ----- -----  
## 199644 13 (a): class 0  
## 107 236 (b): class 1  
##   
##   
## Attribute usage:  
##   
## 100.00% V17  
## 0.17% V10  
## 0.15% V14  
## 0.14% V27  
## 0.06% V26  
## 0.02% V8  
##   
##   
## Time: 15.1 secs

pred1=predict(model1,credit\_test)  
CrossTable(pred1,credit\_test$Class)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 84807   
##   
##   
## | credit\_test$Class   
## pred1 | 0 | 1 | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 84654 | 53 | 84707 |   
## | 0.109 | 61.699 | |   
## | 0.999 | 0.001 | 0.999 |   
## | 1.000 | 0.356 | |   
## | 0.998 | 0.001 | |   
## -------------|-----------|-----------|-----------|  
## 1 | 4 | 96 | 100 |   
## | 91.985 | 52263.297 | |   
## | 0.040 | 0.960 | 0.001 |   
## | 0.000 | 0.644 | |   
## | 0.000 | 0.001 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 84658 | 149 | 84807 |   
## | 0.998 | 0.002 | |   
## -------------|-----------|-----------|-----------|  
##   
##

Logistic Regression model

model2=glm(Class~. ,family = binomial(link = 'logit'),data = credit\_training)  
model2

##   
## Call: glm(formula = Class ~ ., family = binomial(link = "logit"), data = credit\_training)  
##   
## Coefficients:  
## (Intercept) Time V1 V2 V3   
## -8.494e+00 -4.260e-06 1.228e-02 9.865e-02 -3.283e-02   
## V4 V5 V6 V7 V8   
## 7.268e-01 2.200e-01 -1.179e-01 -7.007e-03 -1.522e-01   
## V9 V10 V11 V12 V13   
## -1.683e-01 -8.489e-01 -7.415e-02 1.339e-01 -3.232e-01   
## V14 V15 V16 V17 V18   
## -4.818e-01 -1.176e-01 -2.450e-01 1.166e-02 -1.884e-02   
## V19 V20 V21 V22 V23   
## -1.990e-02 -4.420e-01 4.426e-01 8.330e-01 -1.349e-01   
## V24 V25 V26 V27 V28   
## 1.872e-01 -1.838e-01 3.279e-01 -7.541e-01 -2.606e-01   
## Amount   
## 8.964e-04   
##   
## Degrees of Freedom: 199998 Total (i.e. Null); 199968 Residual  
## (1 observation deleted due to missingness)  
## Null Deviance: 5054   
## Residual Deviance: 1576 AIC: 1638

summary(model2)

##   
## Call:  
## glm(formula = Class ~ ., family = binomial(link = "logit"), data = credit\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.4246 -0.0290 -0.0178 -0.0109 4.6004   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.494e+00 3.108e-01 -27.332 < 2e-16 \*\*\*  
## Time -4.260e-06 2.722e-06 -1.565 0.11749   
## V1 1.228e-02 5.054e-02 0.243 0.80798   
## V2 9.865e-02 7.936e-02 1.243 0.21388   
## V3 -3.283e-02 6.222e-02 -0.528 0.59773   
## V4 7.268e-01 1.048e-01 6.934 4.10e-12 \*\*\*  
## V5 2.200e-01 8.734e-02 2.519 0.01177 \*   
## V6 -1.179e-01 8.460e-02 -1.394 0.16346   
## V7 -7.007e-03 8.979e-02 -0.078 0.93780   
## V8 -1.522e-01 3.621e-02 -4.201 2.65e-05 \*\*\*  
## V9 -1.683e-01 1.513e-01 -1.112 0.26623   
## V10 -8.489e-01 1.249e-01 -6.795 1.08e-11 \*\*\*  
## V11 -7.415e-02 9.662e-02 -0.767 0.44282   
## V12 1.339e-01 1.130e-01 1.184 0.23628   
## V13 -3.232e-01 9.883e-02 -3.270 0.00108 \*\*   
## V14 -4.818e-01 7.524e-02 -6.404 1.51e-10 \*\*\*  
## V15 -1.176e-01 1.032e-01 -1.140 0.25415   
## V16 -2.450e-01 1.704e-01 -1.438 0.15034   
## V17 1.166e-02 9.111e-02 0.128 0.89820   
## V18 -1.884e-02 1.742e-01 -0.108 0.91390   
## V19 -1.990e-02 1.237e-01 -0.161 0.87224   
## V20 -4.420e-01 1.017e-01 -4.347 1.38e-05 \*\*\*  
## V21 4.426e-01 7.702e-02 5.747 9.07e-09 \*\*\*  
## V22 8.330e-01 1.714e-01 4.859 1.18e-06 \*\*\*  
## V23 -1.349e-01 6.927e-02 -1.947 0.05150 .   
## V24 1.872e-01 1.710e-01 1.095 0.27370   
## V25 -1.838e-01 1.494e-01 -1.230 0.21865   
## V26 3.279e-01 2.305e-01 1.423 0.15481   
## V27 -7.541e-01 1.348e-01 -5.594 2.22e-08 \*\*\*  
## V28 -2.606e-01 9.955e-02 -2.617 0.00886 \*\*   
## Amount 8.964e-04 4.941e-04 1.814 0.06966 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5054.1 on 199998 degrees of freedom  
## Residual deviance: 1575.6 on 199968 degrees of freedom  
## (1 observation deleted due to missingness)  
## AIC: 1637.6  
##   
## Number of Fisher Scoring iterations: 12

anova(model2,test="Chisq")

## 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

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Class  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 199998 5054.1   
## Time 1 31.54 199997 5022.6 1.959e-08 \*\*\*  
## V1 1 533.46 199996 4489.1 < 2.2e-16 \*\*\*  
## V2 1 490.42 199995 3998.7 < 2.2e-16 \*\*\*  
## V3 1 728.07 199994 3270.6 < 2.2e-16 \*\*\*  
## V4 1 722.73 199993 2547.9 < 2.2e-16 \*\*\*  
## V5 1 51.60 199992 2496.3 6.803e-13 \*\*\*  
## V6 1 26.16 199991 2470.1 3.150e-07 \*\*\*  
## V7 1 17.68 199990 2452.4 2.610e-05 \*\*\*  
## V8 1 69.43 199989 2383.0 < 2.2e-16 \*\*\*  
## V9 1 50.37 199988 2332.6 1.276e-12 \*\*\*  
## V10 1 462.06 199987 1870.6 < 2.2e-16 \*\*\*  
## V11 1 57.18 199986 1813.4 3.980e-14 \*\*\*  
## V12 1 25.48 199985 1787.9 4.468e-07 \*\*\*  
## V13 1 19.99 199984 1767.9 7.804e-06 \*\*\*  
## V14 1 82.94 199983 1685.0 < 2.2e-16 \*\*\*  
## V15 1 0.04 199982 1685.0 0.8442390   
## V16 1 36.41 199981 1648.6 1.601e-09 \*\*\*  
## V17 1 0.23 199980 1648.3 0.6301951   
## V18 1 0.05 199979 1648.3 0.8281811   
## V19 1 0.52 199978 1647.7 0.4687282   
## V20 1 3.33 199977 1644.4 0.0680803 .   
## V21 1 7.34 199976 1637.1 0.0067412 \*\*   
## V22 1 22.85 199975 1614.2 1.755e-06 \*\*\*  
## V23 1 12.03 199974 1602.2 0.0005229 \*\*\*  
## V24 1 1.05 199973 1601.1 0.3046789   
## V25 1 0.95 199972 1600.2 0.3304804   
## V26 1 1.38 199971 1598.8 0.2403892   
## V27 1 13.38 199970 1585.4 0.0002549 \*\*\*  
## V28 1 5.98 199969 1579.5 0.0144890 \*   
## Amount 1 3.85 199968 1575.6 0.0498105 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pred=predict(model2,newdata = credit\_test,type='response')  
pred=ifelse(pred>0.5,1,0)  
misclass=mean(pred!=credit\_test$Class)  
print(paste('Accuracy',1-misclass))

## [1] "Accuracy 0.999127430518707"

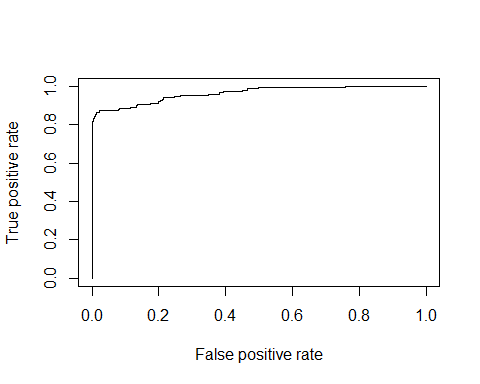
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

p=predict(model2,newdata = credit\_test,type="response")  
pr=prediction(p,credit\_test$Class)  
prf=performance(pr,measure = "tpr",x.measure = "fpr")  
plot(prf)



Regression Tree

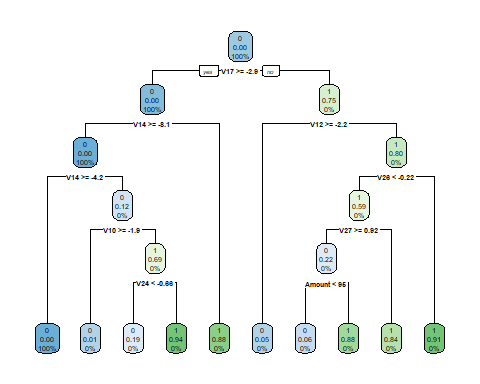
library(rpart)  
model3=rpart(Class~. ,data=credit\_training)  
model3

## n= 200000   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 200000 343 0 (0.9982850000 0.0017150000)   
## 2) V17>=-2.864836 199684 106 0 (0.9994691613 0.0005308387)   
## 4) V14>=-8.09768 199658 83 0 (0.9995842891 0.0004157109)   
## 8) V14>=-4.209589 199373 48 0 (0.9997592452 0.0002407548) \*  
## 9) V14< -4.209589 285 35 0 (0.8771929825 0.1228070175)   
## 18) V10>=-1.858638 237 2 0 (0.9915611814 0.0084388186) \*  
## 19) V10< -1.858638 48 15 1 (0.3125000000 0.6875000000)   
## 38) V24< -0.662693 16 3 0 (0.8125000000 0.1875000000) \*  
## 39) V24>=-0.662693 32 2 1 (0.0625000000 0.9375000000) \*  
## 5) V14< -8.09768 26 3 1 (0.1153846154 0.8846153846) \*  
## 3) V17< -2.864836 316 79 1 (0.2500000000 0.7500000000)   
## 6) V12>=-2.189234 21 1 0 (0.9523809524 0.0476190476) \*  
## 7) V12< -2.189234 295 59 1 (0.2000000000 0.8000000000)   
## 14) V26< -0.2246553 103 42 1 (0.4077669903 0.5922330097)   
## 28) V27>=0.9150463 41 9 0 (0.7804878049 0.2195121951)   
## 56) Amount< 94.99 33 2 0 (0.9393939394 0.0606060606) \*  
## 57) Amount>=94.99 8 1 1 (0.1250000000 0.8750000000) \*  
## 29) V27< 0.9150463 62 10 1 (0.1612903226 0.8387096774) \*  
## 15) V26>=-0.2246553 192 17 1 (0.0885416667 0.9114583333) \*

summary(model3)

## Call:  
## rpart(formula = Class ~ ., data = credit\_training)  
## n= 200000   
##   
## CP nsplit rel error xerror xstd  
## 1 0.46064140 0 1.0000000 1.0000000 0.05394860  
## 2 0.05830904 1 0.5393586 0.5626822 0.04048320  
## 3 0.05539359 2 0.4810496 0.4985423 0.03810818  
## 4 0.03352770 3 0.4256560 0.4781341 0.03732069  
## 5 0.02623907 5 0.3586006 0.4198251 0.03497283  
## 6 0.01749271 8 0.2769679 0.3848397 0.03348494  
## 7 0.01000000 9 0.2594752 0.3352770 0.03125575  
##   
## Variable importance  
## V17 V12 V16 V10 V18 V11 V14 V27 V28 V6   
## 22 17 13 13 10 10 3 1 1 1   
## V7 V9 V5 V8 V26 V20 V22 V24 V4 V3   
## 1 1 1 1 1 1 1 1 1 1   
## Time V19 V21 Amount   
## 1 1 1 1   
##   
## Node number 1: 200000 observations, complexity param=0.4606414  
## predicted class=0 expected loss=0.001715 P(node) =1  
## class counts: 199657 343  
## probabilities: 0.998 0.002   
## left son=2 (199684 obs) right son=3 (316 obs)  
## Primary splits:  
## V17 < -2.864836 to the right, improve=354.4360, (0 missing)  
## V12 < -4.534977 to the right, improve=335.6545, (0 missing)  
## V14 < -6.019218 to the right, improve=278.1957, (0 missing)  
## V10 < -3.890088 to the right, improve=253.7031, (0 missing)  
## V16 < -4.029988 to the right, improve=250.9867, (0 missing)  
## Surrogate splits:  
## V12 < -4.858586 to the right, agree=0.999, adj=0.658, (0 split)  
## V16 < -4.029988 to the right, agree=0.999, adj=0.611, (0 split)  
## V10 < -5.571985 to the right, agree=0.999, adj=0.453, (0 split)  
## V18 < -3.740074 to the right, agree=0.999, adj=0.427, (0 split)  
## V11 < 3.889282 to the left, agree=0.999, adj=0.411, (0 split)  
##   
## Node number 2: 199684 observations, complexity param=0.05830904  
## predicted class=0 expected loss=0.0005308387 P(node) =0.99842  
## class counts: 199578 106  
## probabilities: 0.999 0.001   
## left son=4 (199658 obs) right son=5 (26 obs)  
## Primary splits:  
## V14 < -8.09768 to the right, improve=40.648780, (0 missing)  
## V12 < -4.839175 to the right, improve=36.063280, (0 missing)  
## V11 < 3.604715 to the left, improve=14.909000, (0 missing)  
## V10 < -3.84725 to the right, improve=11.614010, (0 missing)  
## V4 < 5.321502 to the left, improve= 5.381362, (0 missing)  
## Surrogate splits:  
## V12 < -5.374364 to the right, agree=1, adj=0.462, (0 split)  
## V11 < 5.426486 to the left, agree=1, adj=0.231, (0 split)  
##   
## Node number 3: 316 observations, complexity param=0.05539359  
## predicted class=1 expected loss=0.25 P(node) =0.00158  
## class counts: 79 237  
## probabilities: 0.250 0.750   
## left son=6 (21 obs) right son=7 (295 obs)  
## Primary splits:  
## V12 < -2.189234 to the right, improve=22.19524, (0 missing)  
## V3 < -0.9361123 to the right, improve=20.06349, (0 missing)  
## V26 < -0.2246553 to the left, improve=18.99598, (0 missing)  
## V10 < 3.012843 to the right, improve=18.96000, (0 missing)  
## V7 < 2.317753 to the right, improve=18.40977, (0 missing)  
## Surrogate splits:  
## V10 < 6.108595 to the right, agree=0.968, adj=0.524, (0 split)  
## V3 < 0.3881728 to the right, agree=0.965, adj=0.476, (0 split)  
## V4 < -1.983926 to the left, agree=0.965, adj=0.476, (0 split)  
## V28 < -2.183272 to the left, agree=0.965, adj=0.476, (0 split)  
## V5 < 2.369373 to the right, agree=0.962, adj=0.429, (0 split)  
##   
## Node number 4: 199658 observations, complexity param=0.02623907  
## predicted class=0 expected loss=0.0004157109 P(node) =0.99829  
## class counts: 199575 83  
## probabilities: 1.000 0.000   
## left son=8 (199373 obs) right son=9 (285 obs)  
## Primary splits:  
## V14 < -4.209589 to the right, improve=8.5505960, (0 missing)  
## V12 < -4.753613 to the right, improve=4.5649400, (0 missing)  
## V17 < -2.280395 to the right, improve=2.7241190, (0 missing)  
## V10 < -3.07378 to the right, improve=2.1724030, (0 missing)  
## V11 < 3.568853 to the left, improve=0.6891755, (0 missing)  
## Surrogate splits:  
## V9 < 7.154781 to the left, agree=0.999, adj=0.109, (0 split)  
## V17 < -2.372232 to the right, agree=0.999, adj=0.091, (0 split)  
## V10 < 11.49631 to the left, agree=0.999, adj=0.067, (0 split)  
## V18 < 3.200217 to the left, agree=0.999, adj=0.063, (0 split)  
## V11 < 4.001686 to the left, agree=0.999, adj=0.046, (0 split)  
##   
## Node number 5: 26 observations  
## predicted class=1 expected loss=0.1153846 P(node) =0.00013  
## class counts: 3 23  
## probabilities: 0.115 0.885   
##   
## Node number 6: 21 observations  
## predicted class=0 expected loss=0.04761905 P(node) =0.000105  
## class counts: 20 1  
## probabilities: 0.952 0.048   
##   
## Node number 7: 295 observations, complexity param=0.0335277  
## predicted class=1 expected loss=0.2 P(node) =0.001475  
## class counts: 59 236  
## probabilities: 0.200 0.800   
## left son=14 (103 obs) right son=15 (192 obs)  
## Primary splits:  
## V26 < -0.2246553 to the left, improve=13.662840, (0 missing)  
## V14 < -3.429828 to the right, improve=10.038070, (0 missing)  
## V27 < 1.692174 to the right, improve= 8.795633, (0 missing)  
## V22 < -0.1901425 to the left, improve= 8.119524, (0 missing)  
## V9 < -4.031951 to the right, improve= 8.045455, (0 missing)  
## Surrogate splits:  
## V9 < -0.1000254 to the right, agree=0.685, adj=0.097, (0 split)  
## V4 < -0.4657571 to the left, agree=0.675, adj=0.068, (0 split)  
## V21 < -2.817149 to the left, agree=0.675, adj=0.068, (0 split)  
## V27 < -4.25345 to the left, agree=0.675, adj=0.068, (0 split)  
## V7 < -21.2971 to the left, agree=0.671, adj=0.058, (0 split)  
##   
## Node number 8: 199373 observations  
## predicted class=0 expected loss=0.0002407548 P(node) =0.996865  
## class counts: 199325 48  
## probabilities: 1.000 0.000   
##   
## Node number 9: 285 observations, complexity param=0.02623907  
## predicted class=0 expected loss=0.122807 P(node) =0.001425  
## class counts: 250 35  
## probabilities: 0.877 0.123   
## left son=18 (237 obs) right son=19 (48 obs)  
## Primary splits:  
## V10 < -1.858638 to the right, improve=36.81226, (0 missing)  
## V7 < -0.9118796 to the right, improve=20.80351, (0 missing)  
## V12 < -2.982949 to the right, improve=17.64382, (0 missing)  
## V21 < 0.2432055 to the left, improve=16.48466, (0 missing)  
## V9 < 0.1438228 to the right, improve=13.28498, (0 missing)  
## Surrogate splits:  
## V7 < -0.9118796 to the right, agree=0.877, adj=0.271, (0 split)  
## V12 < -2.982949 to the right, agree=0.877, adj=0.271, (0 split)  
## V9 < -2.43036 to the right, agree=0.874, adj=0.250, (0 split)  
## V21 < 0.06710666 to the left, agree=0.867, adj=0.208, (0 split)  
## V19 < -3.14212 to the right, agree=0.849, adj=0.104, (0 split)  
##   
## Node number 14: 103 observations, complexity param=0.0335277  
## predicted class=1 expected loss=0.407767 P(node) =0.000515  
## class counts: 42 61  
## probabilities: 0.408 0.592   
## left son=28 (41 obs) right son=29 (62 obs)  
## Primary splits:  
## V27 < 0.9150463 to the right, improve=18.92460, (0 missing)  
## V26 < -0.3338474 to the right, improve=14.06558, (0 missing)  
## V6 < -2.446679 to the left, improve=13.77808, (0 missing)  
## Time < 33000 to the left, improve=12.96399, (0 missing)  
## V28 < 0.2696809 to the right, improve=12.33469, (0 missing)  
## Surrogate splits:  
## V6 < -2.867273 to the left, agree=0.942, adj=0.854, (0 split)  
## V8 < 1.733585 to the right, agree=0.893, adj=0.732, (0 split)  
## V20 < 0.7582707 to the right, agree=0.864, adj=0.659, (0 split)  
## V22 < -0.2765431 to the left, agree=0.864, adj=0.659, (0 split)  
## Time < 29724.5 to the left, agree=0.816, adj=0.537, (0 split)  
##   
## Node number 15: 192 observations  
## predicted class=1 expected loss=0.08854167 P(node) =0.00096  
## class counts: 17 175  
## probabilities: 0.089 0.911   
##   
## Node number 18: 237 observations  
## predicted class=0 expected loss=0.008438819 P(node) =0.001185  
## class counts: 235 2  
## probabilities: 0.992 0.008   
##   
## Node number 19: 48 observations, complexity param=0.02623907  
## predicted class=1 expected loss=0.3125 P(node) =0.00024  
## class counts: 15 33  
## probabilities: 0.312 0.688   
## left son=38 (16 obs) right son=39 (32 obs)  
## Primary splits:  
## V24 < -0.662693 to the left, improve=12.000000, (0 missing)  
## V16 < 2.754354 to the right, improve=10.471150, (0 missing)  
## V18 < 3.081813 to the right, improve= 8.719737, (0 missing)  
## V5 < 1.271888 to the right, improve= 7.932692, (0 missing)  
## V7 < 1.7515 to the right, improve= 6.003947, (0 missing)  
## Surrogate splits:  
## V16 < 2.537539 to the right, agree=0.833, adj=0.500, (0 split)  
## V7 < 0.4087919 to the right, agree=0.812, adj=0.438, (0 split)  
## V18 < 3.007121 to the right, agree=0.812, adj=0.438, (0 split)  
## V19 < -2.271497 to the left, agree=0.812, adj=0.438, (0 split)  
## V5 < 1.271888 to the right, agree=0.792, adj=0.375, (0 split)  
##   
## Node number 28: 41 observations, complexity param=0.01749271  
## predicted class=0 expected loss=0.2195122 P(node) =0.000205  
## class counts: 32 9  
## probabilities: 0.780 0.220   
## left son=56 (33 obs) right son=57 (8 obs)  
## Primary splits:  
## Amount < 94.99 to the left, improve=8.541205, (0 missing)  
## V26 < -0.2712033 to the left, improve=5.224538, (0 missing)  
## Time < 41646 to the left, improve=4.611280, (0 missing)  
## V15 < 0.03454226 to the left, improve=4.575096, (0 missing)  
## V13 < -0.03163388 to the right, improve=4.148780, (0 missing)  
## Surrogate splits:  
## V28 < 0.2696809 to the right, agree=0.951, adj=0.750, (0 split)  
## V11 < 2.996455 to the right, agree=0.927, adj=0.625, (0 split)  
## V17 < -7.718312 to the left, agree=0.927, adj=0.625, (0 split)  
## V18 < -2.602962 to the left, agree=0.927, adj=0.625, (0 split)  
## V9 < -2.541281 to the left, agree=0.902, adj=0.500, (0 split)  
##   
## Node number 29: 62 observations  
## predicted class=1 expected loss=0.1612903 P(node) =0.00031  
## class counts: 10 52  
## probabilities: 0.161 0.839   
##   
## Node number 38: 16 observations  
## predicted class=0 expected loss=0.1875 P(node) =8e-05  
## class counts: 13 3  
## probabilities: 0.812 0.188   
##   
## Node number 39: 32 observations  
## predicted class=1 expected loss=0.0625 P(node) =0.00016  
## class counts: 2 30  
## probabilities: 0.062 0.938   
##   
## Node number 56: 33 observations  
## predicted class=0 expected loss=0.06060606 P(node) =0.000165  
## class counts: 31 2  
## probabilities: 0.939 0.061   
##   
## Node number 57: 8 observations  
## predicted class=1 expected loss=0.125 P(node) =4e-05  
## class counts: 1 7  
## probabilities: 0.125 0.875

library(rpart.plot)  
rpart.plot(model3,digits = 2)



pred3=predict(model3,credit\_test)

Random Forest

sum(is.na(credit\_training))

## [1] 1

credit\_training=na.omit(credit\_training)  
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

set.seed(123)  
model4=randomForest(credit\_training[-31],credit\_training$Class ,data = credit\_test,ntree = 25)  
model4

##   
## Call:  
## randomForest(x = credit\_training[-31], y = credit\_training$Class, ntree = 25, data = credit\_test)   
## Type of random forest: classification  
## Number of trees: 25  
## No. of variables tried at each split: 5  
##   
## OOB estimate of error rate: 0.05%  
## Confusion matrix:  
## 0 1 class.error  
## 0 199631 22 0.0001101912  
## 1 75 268 0.2186588921

pred4=predict(model4,credit\_test)  
CrossTable(pred4,credit\_test$Class)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 84807   
##   
##   
## | credit\_test$Class   
## pred4 | 0 | 1 | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 84655 | 39 | 84694 |   
## | 0.143 | 81.023 | |   
## | 1.000 | 0.000 | 0.999 |   
## | 1.000 | 0.262 | |   
## | 0.998 | 0.000 | |   
## -------------|-----------|-----------|-----------|  
## 1 | 3 | 110 | 113 |   
## | 106.881 | 60727.202 | |   
## | 0.027 | 0.973 | 0.001 |   
## | 0.000 | 0.738 | |   
## | 0.000 | 0.001 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 84658 | 149 | 84807 |   
## | 0.998 | 0.002 | |   
## -------------|-----------|-----------|-----------|  
##   
##