Zadatak: German Credit

Nola Čumlievski, Ivan Šimičić

31/10/2020

## Uvod

Banka je jedna od temeljnih financijskih institucija na tržištu koja posreduje u novčanom prometu i ostalim poslovima vezanim uz novac. Takvu gospodarsku djelatnost nazivamo bankarstvom. Jedna od najčešćih djelatnosti unutar bankarstva je uzimanje i davanje zajmova, iliti kredita.

Kredit je novčani dužničko-vjerovnički odnos u kojem pružatelj kredita ustupa pravo na određeni iznos novčanih sredstava dužniku, na dogovoreno vrijeme i uz dogovorene uvjete povrata.

Banka kao pružatelj kredita mora razmotriti sve potencijalne dužnike kako bi bila sigurna da će im dužnik biti u mogućnosti vratiti kredit.

Budući da banke davanjem kredita ostvaruju profit (kroz kamate), važno je da svakom potencijalnom dužniku s dobrim financijskim uvjetima odobre kredit te isto tako, da svakog rizičnog potencijalnog dužnika odbiju, budući da bi im on stvarao financijski gubitak.

U ovom zadatku ćemo kreirati modele na temelju demografskih i financijskih informacija o potencijalnim dužnicima te otkriti koji od potencijalnih dužnika predstavlja dobrog klijenta za ugovaranje kredita.

## Skup i istraživanje podataka

U ovom zadatku koristit ćemo skup podataka GermanCredit u .csv formatu koji se sastoji od 32 varijable te 1000 opažanja, a svako od opažanja predstavlja jednog potencijalnog dužnika.

### Tablica varijabli

|  |  |  |  |
| --- | --- | --- | --- |
| Naziv varijable | Opis | Tip varijable | Opis vrijednosti varijable |
| OBS# | broj opažanja | kategorička | redni broj u skupu podataka |
| CHK\_ACCT | status transakcijskog računa | kategorička | 0: < 0 DM, 1: 0-200 DM,2: > 200 DM, 3: nema transakcijski račun |
| DURATION | trajanje kredita u mjesecima | numerička | - |
| HISTORY | povijest kredita | kategorička | 0: nema povijesti kredita, 1: svi krediti otplaćeni u roku, 2: postojeći krediti otplaćeni u roku, 3: odgode u prijašnjim otplatama, 4: kritičan račun |
| NEW\_CAR | svrha podizanja kredita | binarna | novi automobil, 0: ne, 1: da |
| USED\_CAR | svrha podizanja kredita | binarna | korišteni automobil, 0: ne, 1: da |
| FURNITURE | svrha podizanja kredita | binarna | namještaj/oprema, 0: ne, 1: da |
| RADIO/TV | svrha podizanja kredita | binarna | radio/televizija, 0: ne, 1: da |
| EDUCATION | svrha podizanja kredita | binarna | obrazovanje, 0: ne, 1: da |
| RETRAINING | svrha podizanja kredita | binarna | prekvalifikacija, 0: ne, 1: da |
| AMOUNT | iznos kredita | numerička | - |
| SAV\_ACCT | prosječni iznos na štednom računu | kategorička | 0: < 100 DM, 1: 101-500 DM, 2: 501-1000 DM, 3: > 1000 DM, 4: nema štedni račun |
| EMPLOYMENT | trenutni status zaposlenja | kategorička | 0: nije u radnom odnosu, 1: < 1 godina radnog odnosa, 2: 1-3 godine radnog odnosa, 3: 4-6 godina radnog odnosa, 4: ≥ 7 godina radnog odnosa |
| INSTALL\_RATE | rata kao postotak prihoda | numerička | - |
| MALE\_DIV | potencijalni dužnik je muškog spola i rastavljen | binarna | 0: ne, 1: da |
| MALE\_SINGLE | potencijalni dužnik je muškog spola i slobodan | binarna | 0: ne, 1: da |
| MALE\_MAR\_WID | potencijalni dužnik je muškog spola i oženjen ili udovac | binarna | 0: ne, 1: da |
| CO-APPLICANT | potencijalni dužnik ima sudužnika | binarna | 0: ne, 1: da |
| GUARANTOR | potencijalni dužnik ima jamca | binarna | 0: ne, 1: da |
| PRESENT\_RESIDENT | potencijalni dužnik je trenutni stanovnik | kategorička | 0: ≤ 1 godine, 1: 1-2 godine, 2: 2-3 godine, 3: ≥ 3 godine |
| REAL\_ESTATE | potencijalni dužnik ima u vlasništvu neku nekretninu | binarna | 0: ne, 1: da |
| PROP\_UNKN\_NONE | nije vlasnik imovine | binarna | 0: ne, 1: da |
| AGE | starost | numerička | - |
| RENT | živi u najmu | binarna | 0: ne, 1: da |
| OWN\_RES | posjeduje stambenu lokaciju | binarna | 0: ne, 1: da |
| NUM\_CREDITS | broj postojećih kredita u banci | numerička | - |
| JOB | priroda posla | kategorička | 0: nezaposlen/nekvalificiran, 1: nekvalificiran, ali zaposlen, 2: kvalificiran zaposlenik, 3: poduzetnik/menadžer/visoko kvalificiran zaposlenik |
| NUM\_DEPENDENTS | broj ljudi koji ovise o dužniku | numerička | - |
| TELEPHONE | dužnik ima telefon u svoje ime | binarna | 0: ne, 1: da |
| FOREIGN | dužnik je strani radnik | binarna | 0: ne, 1: da |
| RESPONSE | dužnik ima dobar rating za kredit | binarna | 0: ne, 1: da |

### Pregled statističkih podataka po stupcu

Uz pomoć funkcije *summary* možemo vidjeti pregled statističkih podataka za svaku od varijabli skupa podataka.

summary(GermanCredit)

## OBS. CHK\_ACCT DURATION HISTORY   
## Min. : 1.0 Min. :0.000 Min. : 4.0 Min. :0.000   
## 1st Qu.: 250.8 1st Qu.:0.000 1st Qu.:12.0 1st Qu.:2.000   
## Median : 500.5 Median :1.000 Median :18.0 Median :2.000   
## Mean : 500.5 Mean :1.577 Mean :20.9 Mean :2.545   
## 3rd Qu.: 750.2 3rd Qu.:3.000 3rd Qu.:24.0 3rd Qu.:4.000   
## Max. :1000.0 Max. :3.000 Max. :72.0 Max. :4.000   
## NEW\_CAR USED\_CAR FURNITURE RADIO.TV EDUCATION   
## Min. :0.000 Min. :0.000 Min. :0.000 Min. :0.00 Min. :0.00   
## 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.00 1st Qu.:0.00   
## Median :0.000 Median :0.000 Median :0.000 Median :0.00 Median :0.00   
## Mean :0.234 Mean :0.103 Mean :0.181 Mean :0.28 Mean :0.05   
## 3rd Qu.:0.000 3rd Qu.:0.000 3rd Qu.:0.000 3rd Qu.:1.00 3rd Qu.:0.00   
## Max. :1.000 Max. :1.000 Max. :1.000 Max. :1.00 Max. :1.00   
## RETRAINING AMOUNT SAV\_ACCT EMPLOYMENT   
## Min. :0.000 Min. : 250 Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.: 1366 1st Qu.:0.000 1st Qu.:2.000   
## Median :0.000 Median : 2320 Median :0.000 Median :2.000   
## Mean :0.097 Mean : 3271 Mean :1.105 Mean :2.384   
## 3rd Qu.:0.000 3rd Qu.: 3972 3rd Qu.:2.000 3rd Qu.:4.000   
## Max. :1.000 Max. :18424 Max. :4.000 Max. :4.000   
## INSTALL\_RATE MALE\_DIV MALE\_SINGLE MALE\_MAR\_or\_WID CO.APPLICANT   
## Min. :1.000 Min. :0.00 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:2.000 1st Qu.:0.00 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000   
## Median :3.000 Median :0.00 Median :1.000 Median :0.000 Median :0.000   
## Mean :2.973 Mean :0.05 Mean :0.548 Mean :0.092 Mean :0.041   
## 3rd Qu.:4.000 3rd Qu.:0.00 3rd Qu.:1.000 3rd Qu.:0.000 3rd Qu.:0.000   
## Max. :4.000 Max. :1.00 Max. :1.000 Max. :1.000 Max. :1.000   
## GUARANTOR PRESENT\_RESIDENT REAL\_ESTATE PROP\_UNKN\_NONE   
## Min. :0.000 Min. :1.000 Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:2.000 1st Qu.:0.000 1st Qu.:0.000   
## Median :0.000 Median :3.000 Median :0.000 Median :0.000   
## Mean :0.052 Mean :2.845 Mean :0.282 Mean :0.154   
## 3rd Qu.:0.000 3rd Qu.:4.000 3rd Qu.:1.000 3rd Qu.:0.000   
## Max. :1.000 Max. :4.000 Max. :1.000 Max. :1.000   
## AGE OTHER\_INSTALL RENT OWN\_RES   
## Min. :19.00 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:27.00 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000   
## Median :33.00 Median :0.000 Median :0.000 Median :1.000   
## Mean :35.55 Mean :0.186 Mean :0.179 Mean :0.713   
## 3rd Qu.:42.00 3rd Qu.:0.000 3rd Qu.:0.000 3rd Qu.:1.000   
## Max. :75.00 Max. :1.000 Max. :1.000 Max. :1.000   
## NUM\_CREDITS JOB NUM\_DEPENDENTS TELEPHONE   
## Min. :1.000 Min. :0.000 Min. :1.000 Min. :0.000   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:0.000   
## Median :1.000 Median :2.000 Median :1.000 Median :0.000   
## Mean :1.407 Mean :1.904 Mean :1.155 Mean :0.404   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :4.000 Max. :3.000 Max. :2.000 Max. :1.000   
## FOREIGN RESPONSE   
## Min. :0.000 Min. :0.0   
## 1st Qu.:0.000 1st Qu.:0.0   
## Median :0.000 Median :1.0   
## Mean :0.037 Mean :0.7   
## 3rd Qu.:0.000 3rd Qu.:1.0   
## Max. :1.000 Max. :1.0

### Korelacija varijabli i “čišćenje” skupa

Prije nego što krenemo koristiti skup podataka, moramo vidjeti koje varijable najbolje koreliraju s ciljnom varijablom kako bi mogli “očistiti” skup.

library(corrr)  
korelacija = cor(GermanCredit)  
korelacija[,"RESPONSE"]

## OBS. CHK\_ACCT DURATION HISTORY   
## -0.034606444 0.350847483 -0.214926665 0.228784733   
## NEW\_CAR USED\_CAR FURNITURE RADIO.TV   
## -0.096900471 0.099790757 -0.020970599 0.106922171   
## EDUCATION RETRAINING AMOUNT SAV\_ACCT   
## -0.070087664 -0.036129062 -0.154738641 0.178942736   
## EMPLOYMENT INSTALL\_RATE MALE\_DIV MALE\_SINGLE   
## 0.116002036 -0.072403937 -0.050062617 0.080676803   
## MALE\_MAR\_or\_WID CO.APPLICANT GUARANTOR PRESENT\_RESIDENT   
## 0.019630297 -0.062728366 0.055039254 -0.002967159   
## REAL\_ESTATE PROP\_UNKN\_NONE AGE OTHER\_INSTALL   
## 0.119299516 -0.125750044 0.091127409 -0.113285167   
## RENT OWN\_RES NUM\_CREDITS JOB   
## -0.092785447 0.134588776 0.045732489 -0.032735001   
## NUM\_DEPENDENTS TELEPHONE FOREIGN RESPONSE   
## 0.003014853 0.036466190 0.082079499 1.000000000

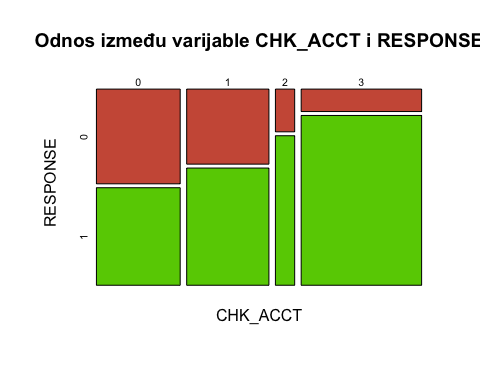
Možemo vidjeti kako najviše koreliraju varijable CHK\_ACC, DURATION, HISTORY, AMOUNT, SAV\_ACCT, EMPLOYMENT, REAL\_ESTATE, PRO\_UNKN\_NONE, OTHER\_INSTALL te OWN\_RES i iz tog razloga te varijable ostavljamo u skupu za daljnji rad budući da ostale varijable nemaju veliki utjecaj na ciljnu varijablu.

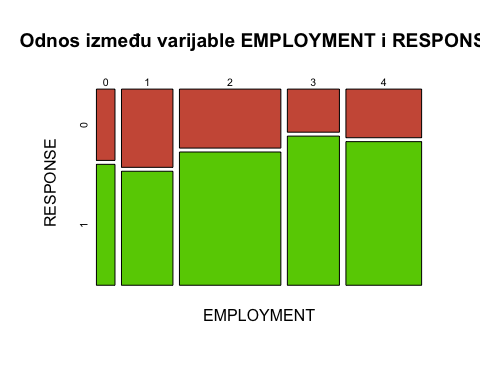
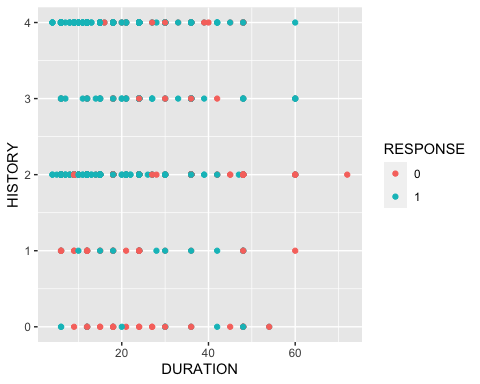
Uz pomoć funkcije head možemo vidjeti kako skup izgleda nakon odabira varijabli:

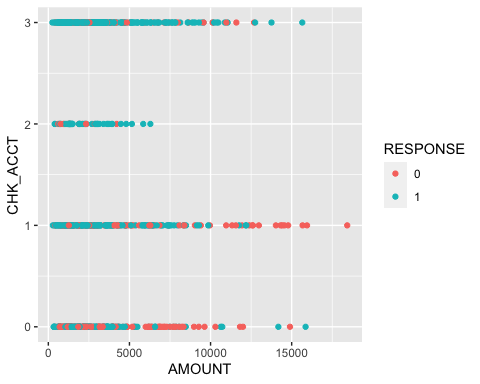
head(GermanCredit)

## CHK\_ACCT DURATION HISTORY AMOUNT SAV\_ACCT EMPLOYMENT REAL\_ESTATE  
## 1 0 6 4 1169 4 4 1  
## 2 1 48 2 5951 0 2 1  
## 3 3 12 4 2096 0 3 1  
## 4 0 42 2 7882 0 3 0  
## 5 0 24 3 4870 0 2 0  
## 6 3 36 2 9055 4 2 0  
## PROP\_UNKN\_NONE OTHER\_INSTALL OWN\_RES RESPONSE  
## 1 0 0 1 1  
## 2 0 0 1 0  
## 3 0 0 1 1  
## 4 0 0 0 1  
## 5 1 0 0 0  
## 6 1 0 0 1

### Vizualizacija

 Na prethodnoj slici možemo vidjeti odnos između varijable CHK\_ACCT (transakcijski račun) i RESPONSE koja predstavlja je li potencijalni dužnik dobar kandidat za kredit (credit rating). Možemo uočiti kako potencijalni dužnici koji nemaju transakcijski račun češće dobiju odobrenje kredita nego dužnici koji imaju manje od 200 DM na postojećem transakcijskom računu.

 Na prethodnoj slici možemo vidjeti odnos između varijable EMPLOYMENT i RESPONSE te možemo zaključiti kako zaposlene osobe imaju veću šansu dobiti odobrenje kredita, iako možemo primjetiti malo odstupanje u slučaju nezaposlenih i osoba koje imaju manje od godinu dana radnog odnosa. Tome možda možemo pripisati mogućnost otkaza nedugo prije odobravanja kredita, ili čak i poslije odobravanja kredita.  Na prethodnoj slici možemo vidjeti odnos između varijabli HISTORY i DURATION te zaključiti kako oni koji nisu imali nikakvih kredita u banci imaju manju šansu dobiti kredit od dužnika koji su se pokazali kao pouzdani, ali također možemo vidjeti i nelogično odstupanje budući da ima i puno osoba s kritičnim stanjem računa, a dobili su kredit.

 Na prethodnoj slici možemo vidjeti odnos između varijabli AMOUNT i CHK\_ACCT te možemo zaključiti kako te dvije varijable nemaju veliki utjecaj na sam ishod odobravanja kredita, ali također možemo uočiti kako su krediti do 10000 DM odobreni u puno većem broju nego oni iznad tog iznosa. Također osobe koje nisu imale transakcijski račun, isto kao i osobe koje su u prekoračenju, imaju veću šansu dobiti kredit u iznosu do 5000 DM.

### Distribucija klasa

Kako bi nastavili dalje, moramo utvrditi distribuciju klase s obzirom na vrijednost varijable RESPONSE. Možemo vidjeti kako 30% opažanja ima vrijednost 0, a 70% vrijednost varijable 1.

## Frekvencija Postotak  
## 0 300 30  
## 1 700 70

#### Kreiranje skupova za treniranje i testiranje

Kako bi mogli preći na izradu modela, prvo moramo kreirati skupove za treniranje i testiranje.

set.seed(1)  
indeksi = sample(1:nrow(GermanCredit), 0.8\*nrow(GermanCredit))  
treniranje = GermanCredit[indeksi,]  
testiranje = GermanCredit[-indeksi,]  
test\_y = GermanCredit$RESPONSE[-indeksi]

## Izrada modela

Nakon što smo vizualizirali podatke i kreirali skupove za treniranje i testiranje, možemo početi s izradom modela.

### Logistička regresija

Logistička regresija je jedan od najkorištenijih algoritama za izradu modela strojnog učenja, a koristi se za rješavanje problema klasifikacije. Budući da u našem slučaju imamo samo dva moguća ishoda varijable RESPONSE, kod nas se radi o binarnoj logističkoj regresiji. Potencijalni dužnik može imati dobar (1) credit rating, ili loš (0), tj. može dobiti kredit ili će ga banka odbiti.

Za izradu modela logističke regresije koristimo biblioteku ISLR i funkciju *glm*.

logisticki = glm(formula = RESPONSE~., family = "binomial",  
 data = treniranje)

Uz pomoć funkcije *summary* možemo vidjeti informacije o modelu:

summary(logisticki)

##   
## Call:  
## glm(formula = RESPONSE ~ ., family = "binomial", data = treniranje)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5637 -0.8375 0.4557 0.7490 1.8761   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.365e-01 3.599e-01 -2.324 0.020111 \*   
## CHK\_ACCT 5.699e-01 7.702e-02 7.399 1.37e-13 \*\*\*  
## DURATION -3.158e-02 9.235e-03 -3.420 0.000626 \*\*\*  
## HISTORY 2.863e-01 8.614e-02 3.323 0.000889 \*\*\*  
## AMOUNT -5.938e-06 3.948e-05 -0.150 0.880424   
## SAV\_ACCT 2.246e-01 6.334e-02 3.546 0.000391 \*\*\*  
## EMPLOYMENT 2.067e-01 7.468e-02 2.768 0.005644 \*\*   
## REAL\_ESTATE 4.235e-01 2.196e-01 1.928 0.053818 .   
## PROP\_UNKN\_NONE -2.322e-01 2.873e-01 -0.808 0.418917   
## OTHER\_INSTALL -5.386e-01 2.163e-01 -2.490 0.012763 \*   
## OWN\_RES 4.272e-01 2.157e-01 1.981 0.047628 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 965.23 on 799 degrees of freedom  
## Residual deviance: 775.53 on 789 degrees of freedom  
## AIC: 797.53  
##   
## Number of Fisher Scoring iterations: 5

Koeficijent varijable nam govori o značaju varijabli. U slučaju kada je on jednak 0, varijabla nema nikakvog značaja. U slučaju kada je veći od nule, značaj varijable raste zajedno s vjerojatnosti za većom vrijednosti varijable RESPONSE (u našem slučaju, 1). U slučaju kada je on veći od nule, vrijedi obrnuto, tj. da će vjerojatnost za manjom vrijednosti (u našem slučaju, 0) varijable RESPONSE biti veća.

Također možemo vidjeti kako najveći značaj imaju varijable CHK\_ACCT, DURATION, HISTORY, SAV\_ACC i EMPLOYMENT.

Uz pomoć funkcije *predict* kreiramo predikciju na temelju ranije kreiranog modela:

predikcija\_log = predict(logisticki, testiranje, type = "response")

Uz pomoć funkcije *table* kreiramo matricu konfuzije kako bi vidjeli rezultat predikcije:

## test\_y  
## predikcija\_log\_y 0 1  
## 0 27 17  
## 1 40 116

Uz pomoć funckije *mean* možemo izračunati stopu pogreške klasifikacije:

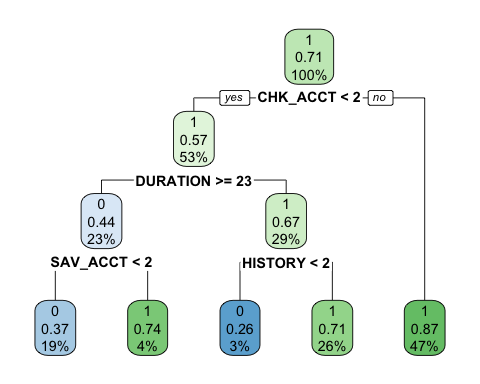
## [1] 0.285

### Stablo odlučivanja

Stablo odlučivanja je jedan od poznatijih algoritama za strojno učenje, a poseban je po tome što ga možemo koristiti za rješavanje problema klasifikacije, ali i za rješavanje problema regresije. Stablo odlučivanja radi na način da kreiramo model za učenje koji radi predikciju na temelju ranije naučenih odluka tijekom treniranja. Algoritam kreće od korijena stabla (korijenski čvor) te sukladno odlukama se grana (čvorovi odluke) sve dok ne dođe do “listova”, iliti krajnjih čvorova.

Model stabla odlučivanja u R programskom jeziku kreiramo uz pomoć funkcije *rpart* koja dolazi uz rpart biblioteku:

stablo = rpart(RESPONSE~CHK\_ACCT+DURATION+HISTORY+SAV\_ACCT+EMPLOYMENT,   
 data = treniranje, method = 'class')

Nakon što smo kreirali model stabla odlučivanja, uz pomoć funkcije *.plot* možemo isto stablo i vizualizirati.  Na prethodnoj slici možemo vidjeti kako izgleda ranije kreirano stablo odlučivanja. Korijenski čvor nam govori da vjerojatnost da će vrijednost varijable RESPONSE biti jednaka 1 iznosi 71%, a to vrijedi za sva opažanja. Nakon toga se grana na temelju varijable CHK\_ACCT. U nastavku možemo vidjeti kako u slučaju kada varijabla CHK\_ACCT ima vrijednost veću od 2, vjerojatnost da će varijabla RESPONSE imati vrijednost 1 iznosi 87% i to vrijedi za 470 opažanja. Idućih 530 opažanja se grana s obzirom na varijablu DURATION i možemo vidjeti kako 290 opažanja ima vjerojatnost 67% da će vrijednost varijable RESPONSE biti jednaka 1, a nakon toga, samo 30 opažanja (od prethodnih 290) će imati vjerojatnost od 26% da će im vrijednost varijable RESPONSE biti 0 u slučaju kada varijabla HISTORY ima vrijednost manju od 2. Ostalih 260 opažanja će imati vjerojatnost od 71% da će im vrijednost varijable RESPONSE biti jednaka 1. S druge strane imamo 230 opažanja koji se dalje granaju s obzirom na varijablu SAV\_ACCT te 190 od njih ima vjerojatnost od 37% da će im vrijednost varijable RESPONSE biti jednaka 0 u slučaju kada varijabla SAV\_ACCT ima vrijednost manju od 2. Sukladno tome, samo 40 opažanja će imati vjerojatnost od 74% da će im vrijednost varijable RESPONSE biti 1 u slučaju kada varijabla SAV\_ACCT ima vrijednost veću od 2.

Predikciju ćemo kreirati uz pomoć funkcije *predict*:

predikcija\_stablo = predict(stablo, testiranje, type = 'class')

Uz pomoć funkcije *table* kreiramo matricu konfuzije:

## predikcija\_stablo  
## test\_y 0 1  
## 0 26 41  
## 1 15 118

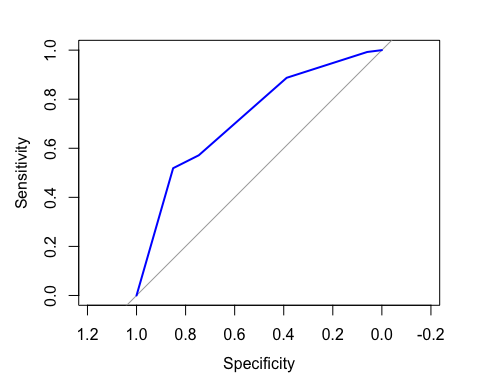
Izračun točnosti modela:

## [1] "Točnost modela iznosi: 0.72"

Izračun vjerojatnosti za opažanja iz skupa za testiranje:

## 0 1  
## 18 0.2571429 0.7428571  
## 23 0.2877358 0.7122642  
## 26 0.2877358 0.7122642  
## 32 0.6315789 0.3684211  
## 38 0.1322751 0.8677249

U nastavku ćemo kreirati ROC krivulju uz pomoć funkcije *roc* te istu vizualizirati:



Izračun prostora “ispod” krivulje ćemo izračunati uz pomoć funkcije *auc*:

## Area under the curve: 0.725

## Slučajne šume

Slučajne šume, kao što im ime govori, sadržajavu više stabala odlučivanja. Stabla zajedno dolaze do rješenja, a zahvaljujući zakonu krda, više stabala (iako nije nužno da koreliraju) dolazi do boljih rješenja nego neko jedinično stablo.

Slučajne šume daju najbolje rezultate kada koristimo metodu “bagging” gdje zapravo kreiramo šumu od stabala koja koriste različite značajke iz cijelokupnog skupa značajki.

Za izradu modela slučajnih šuma koristimo funkciju *train* unutar koje zadajemo vrijednost kontrole varijablom cv\_kontrola koju smo ranije kreirali za naših 10 varijabli te zadajemo metodu “rf” budući da radimo s klasičnim slučajnim šumama:

set.seed(5)  
suma\_model <- train(factor(RESPONSE) ~ ., data=treniranje, method="rf",  
 trControl=trainControl("cv",number=10), importance=TRUE)

U nastavku ćemo vidjeti informacije o kreiranom modelu:

## Random Forest   
##   
## 800 samples  
## 10 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 720, 720, 720, 721, 719, 720, ...   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 2 0.7677035 0.3650113  
## 6 0.7738444 0.4138480  
## 10 0.7650619 0.3976007  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 6.

Možemo vidjeti kako najveću točnost imamo kada je vrijednost mtry jednaka 6, što znači da ćemo koristiti broj 6 kao broj uzoraka koji se koristi za grananje stabala unutar šume.

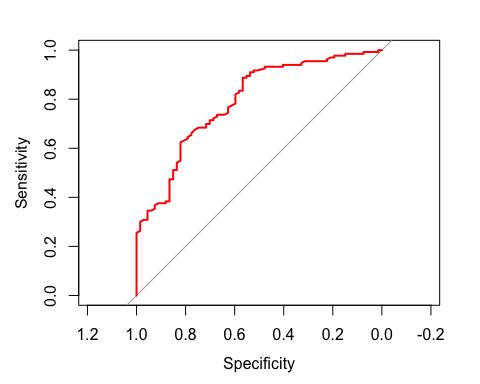
Predikciju kreiramo uz pomoć funkcije *predict*:

predikcija\_rf = predict(suma\_model, newdata = testiranje, type="raw")

Kreiramo konfuzijsku matricu:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 38 29  
## 1 15 118  
##   
## Accuracy : 0.78   
## 95% CI : (0.7161, 0.8354)  
## No Information Rate : 0.735   
## P-Value [Acc > NIR] : 0.08473   
##   
## Kappa : 0.4792   
##   
## Mcnemar's Test P-Value : 0.05002   
##   
## Sensitivity : 0.7170   
## Specificity : 0.8027   
## Pos Pred Value : 0.5672   
## Neg Pred Value : 0.8872   
## Prevalence : 0.2650   
## Detection Rate : 0.1900   
## Detection Prevalence : 0.3350   
## Balanced Accuracy : 0.7599   
##   
## 'Positive' Class : 0   
##

Možemo vidjeti kako točnost iznosi 78%, a Kappa koeficijent iznosi 0.48 što znači da je slaganje između varijabli dobro. (1 je savršeno)

Uz pomoć funkcije *roc* kreiramo ROC krivulju te ju vizualiziramo:  Izračun prostora ispod ROC krivulje:

## Area under the curve: 0.7918

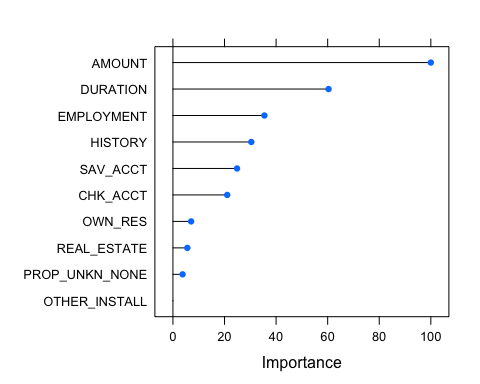
#### Bagging

Za izradu modela slučajnih šuma uz metodu “bagging” moramo unutar funkcije train zadati “treebag” kao metodu, u usporedbi s prethodnim primjerom gdje smo koristili metodu “rf”:

set.seed(5)  
stablo\_bagg = train(factor(RESPONSE) ~ ., data=treniranje, method="treebag",  
 trControl=trainControl("cv",number=10), importance=TRUE)

U nastavku možemo vidjeti informacije o kreiranom modelu:

Možemo uočiti kako nam model ima točnost od 75% što je vrlo dobro, a Kappa koeficijent iznosi 0.36 što nam govori da model ima prolazno (0.2-0.4) slaganje između varijabli.

Uz pomoć funkcije *varImp* možemo saznati važnost varijabli:  Na prethodnom grafu možemo vidjeti kako varijable AMOUNT i DURATION imaju najveću važnost u slučaju kada se koristi algoritam slučajnih šuma s “bagging” metodom.

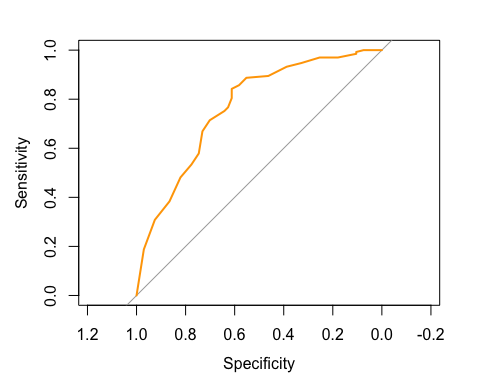
Uz pomoć funkcije *predict* kreiramo model predikcije:

bagg\_predikcija = predict(stablo\_bagg, newdata = testiranje, type="raw")

Kreiramo konfuzijsku matricu:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 39 28  
## 1 19 114  
##   
## Accuracy : 0.765   
## 95% CI : (0.7, 0.8219)  
## No Information Rate : 0.71   
## P-Value [Acc > NIR] : 0.04876   
##   
## Kappa : 0.4544   
##   
## Mcnemar's Test P-Value : 0.24324   
##   
## Sensitivity : 0.6724   
## Specificity : 0.8028   
## Pos Pred Value : 0.5821   
## Neg Pred Value : 0.8571   
## Prevalence : 0.2900   
## Detection Rate : 0.1950   
## Detection Prevalence : 0.3350   
## Balanced Accuracy : 0.7376   
##   
## 'Positive' Class : 0   
##

Možemo uočiti kako nam model predikcije ima točnost od 76.5% što je također dobro.

Uz pomoć funkcije *roc* kreiramo ROC krivulju te ju vizualiziramo:  Izračun prostora ispod ROC krivulje:

## Area under the curve: 0.7667

#### Boosting

Za izradu modela slučajnih šuma uz metodu “boosting” moramo unutar funkcije train zadati “gbm” kao metodu, u usporedbi s prethodnim primjerom gdje smo koristili metodu “treebag”:

set.seed(5)  
boost\_model = train(factor(RESPONSE) ~ .,data=treniranje,method="gbm",  
 verbose=F,trControl=trainControl("cv",number=10))

U nastavku možemo vidjeti informacije o kreiranom modelu:

Možemo uočiti kako najbolju točnost imamo u slučaju kada se koristi dubina interakcije 2 s 150 stabala kada točnost iznosi 78%, a Kappa koeficijent 0.40 što je zadovoljavajuće.

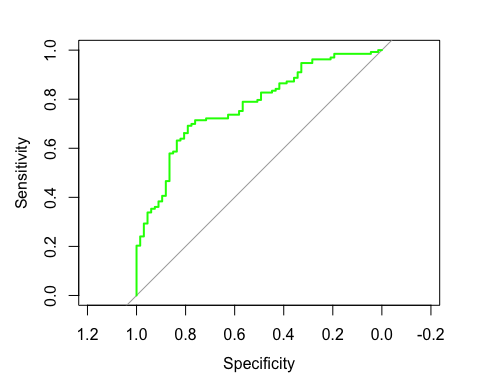
Uz pomoć funkcije *predict* kreiramo model predikcije:

gbm\_predikcija = predict(boost\_model, newdata = testiranje, type="raw")

Kreiramo konfuzijsku matricu:

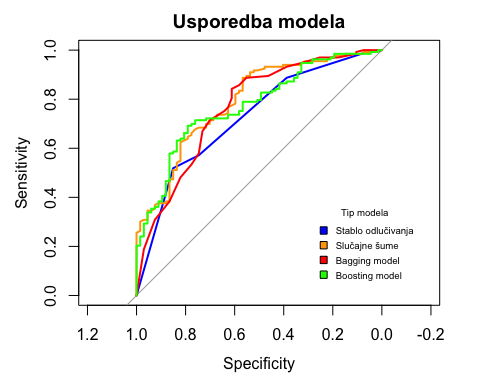
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 29 38  
## 1 21 112  
##   
## Accuracy : 0.705   
## 95% CI : (0.6366, 0.7672)  
## No Information Rate : 0.75   
## P-Value [Acc > NIR] : 0.93753   
##   
## Kappa : 0.2934   
##   
## Mcnemar's Test P-Value : 0.03725   
##   
## Sensitivity : 0.5800   
## Specificity : 0.7467   
## Pos Pred Value : 0.4328   
## Neg Pred Value : 0.8421   
## Prevalence : 0.2500   
## Detection Rate : 0.1450   
## Detection Prevalence : 0.3350   
## Balanced Accuracy : 0.6633   
##   
## 'Positive' Class : 0   
##

Možemo vidjeti kako je točnost predikcije jednaka 70%, a Kappa koeficijent 0.29 što je znatno lošije od vrijednosti samog modela.

Uz pomoć funkcije *roc* kreiramo ROC krivulju te ju vizualiziramo:  Izračun prostora ispod ROC krivulje:

## Area under the curve: 0.7667

#### Usporedba modela slučajnih šuma

 Možemo vidjeti kako model slučajnih šuma daje najbolje rezultate budući da on ima i najveći prostor ispod ROC krivulje, a on iznosi 0.7918. Model stabla odlučivanja je najlošiji što je bilo i za očekivati budući da on koristi samo jedno stablo. Modeli s metodama bagging i boosting imaju jednak prostor ispod krivulje, a on iznosi 0.7667, ali važno za napomenuti da krivulje nisu jednake.

## Zaključak

U ovom zadatku smo prošli kroz skup podataka GermanCredit te na temelju njega kreirali modele logističke regresije, stabla odlučivanja te slučajnih šuma s različitim metodama.

Možemo zaključiti da model slučajnih šuma ima najviši postotak točnosti, a on iznosi 78% što je sukladno i s najboljim Kappa koeficijentom u usporedbi s ostalim modelima. Model slučajnih šuma također ima i najveći prostor ispod ROC krivulje.

Odabir potencijalnih dužnika je vrlo kompliciran proces, ali uz pomoć tehnika strojnog učenja (neke od tehnika smo demonstrirali u ovom zadatku) možemo doći do puno boljih rezultata koji su ključni za odabir stabilnih dužnika koji neće donijeti rizik poslovanju neke banke.