Modelovanje ponašanja klijenata u banci

eng. Churn Modelling

Aleksandra Zdravković Ognjen Lazić Kosta Ljujić Mihajlo Srbakoski

Uvod

Banke i osiguravajuće kompanije često koriste analizu odliva kupaca (*eng. churn analysis*) i stope odliva klijenata kao jednu od svojih ključnih poslovnih pokazatelja, jer su troškovi zadržavanja postojećih kupaca daleko manji od sticanja novog.

Ova analiza se fokusira na ponašanje bankarskih klijenata za koje je veća verovatnoća da će napustiti banku (tj. zatvoriti svoj bankovni račun). Cilj je otkrivanje najupečatljivijih ponašanja kupaca kroz istraživačku analizu podataka, kao i upotreba tehnika prediktivne analize kako bi se utvrdili kupci koji će najverovatnije napustiti banku.

Pretprocesiranje podataka (*eng. Data Preprocessing*)

```
data <- read.csv("data.csv")
head(data)
```

```
##   RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311     Hill         608     Spain Female  41      1
## 3         3   15619304     Onio         502    France Female  42      8
## 4         4   15701354     Boni         699    France Female  39      1
## 5         5   15737888 Mitchell         850     Spain Female  43      2
## 6         6   15574012      Chu         645     Spain   Male  44      8
##     Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00             1         1              1       101348.88      1
## 2  83807.86             1         0              1       112542.58      0
## 3 159660.80             3         1              0       113931.57      1
## 4      0.00             2         0              0        93826.63      0
## 5 125510.82             1         1              1        79084.10      0
## 6 113755.78             2         1              0       149756.71      1
```

```
glimpse(data)
```

```
## Rows: 10,000
## Columns: 14
## $ RowNumber       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ CustomerId      <int> 15634602, 15647311, 15619304, 15701354, 15737888, 1...
## $ Surname         <chr> "Hargrave", "Hill", "Onio", "Boni", "Mitchell", "Ch...
```

```
## $ CreditScore     <int> 619, 608, 502, 699, 850, 645, 822, 376, 501, 684, 5...
## $ Geography       <chr> "France", "Spain", "France", "France", "Spain", "Sp...
## $ Gender          <chr> "Female", "Female", "Female", "Female", "Female", "...
## $ Age             <int> 42, 41, 42, 39, 43, 44, 50, 29, 44, 27, 31, 24, 34,...
## $ Tenure          <int> 2, 1, 8, 1, 2, 8, 7, 4, 4, 2, 6, 3, 10, 5, 7, 3, 1,...
## $ Balance         <dbl> 0.00, 83807.86, 159660.80, 0.00, 125510.82, 113755....
## $ NumOfProducts   <int> 1, 1, 3, 2, 1, 2, 2, 4, 2, 1, 2, 2, 2, 2, 2, 2, 1, ...
## $ HasCrCard       <int> 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, ...
## $ IsActiveMember  <int> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, ...
## $ EstimatedSalary <dbl> 101348.88, 112542.58, 113931.57, 93826.63, 79084.10...
## $ Exited          <int> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...
```

RowNumber	Redni broj reda (od 1 do 10 000)
CustomerId	Jedinstveni identifikacioni broj klijenta banke
Surname	Prezime klijenta
CreditScore	Kreditni skor klijenta
Geography	Zemlja porekla klijenta
Gender	Pol klijenta (muško ili žensko)
Age	Godine klijenta
Tenure	Broj godina koliko je dugo klijent u banci
Balance	Stanje na racunu
NumOfProducts	Broj proizvoda banke koje klijent koristi
HasCrCard	Indikator da li klijent poseduje kreditnu karticu banke
IsActiveMember	Indikator da li je klijent aktivan u banci
EstimatedSalary	Procenjena plata klijenta (u dolarima)
Exited	Indikator da li je klijent napustio banku

NA vrednosti

Proverava se da li postoje NA vrednosti:

```
sapply(data, function(x) sum(is.na(x)))
```

```
##       RowNumber      CustomerId         Surname     CreditScore       Geography
##               0               0               0               0               0
##          Gender             Age          Tenure         Balance   NumOfProducts
##               0               0               0               0               0
##       HasCrCard  IsActiveMember EstimatedSalary          Exited
##               0               0               0               0
```

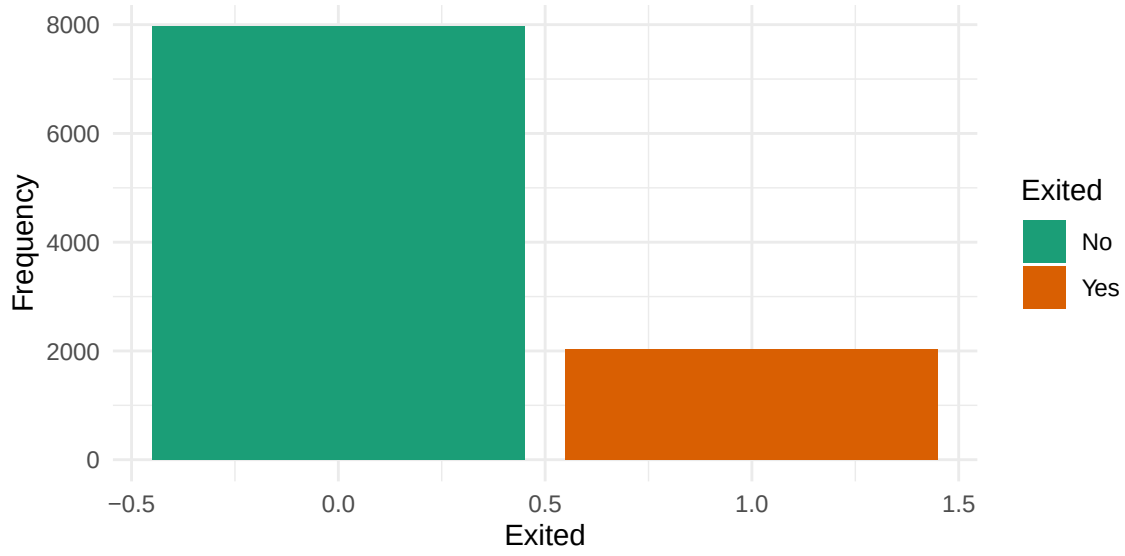
Nema nedostajućih vrednosti.

```
summary(data)
```

```
##    RowNumber       CustomerId         Surname           CreditScore
##  Min.   :    1   Min.   :15565701   Length:10000       Min.   :350.0
##  1st Qu.: 2501   1st Qu.:15628528   Class :character   1st Qu.:584.0
##  Median : 5000   Median :15690738   Mode  :character   Median :652.0
##  Mean   : 5000   Mean   :15690941                      Mean   :650.5
```

```
##  3rd Qu.: 7500   3rd Qu.:15753234                     3rd Qu.:718.0
##  Max.   :10000   Max.   :15815690                     Max.   :850.0
##   Geography            Gender               Age            Tenure
##  Length:10000       Length:10000       Min.   :18.00   Min.   : 0.000
##  Class :character   Class :character   1st Qu.:32.00   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Median :37.00   Median : 5.000
##                                        Mean   :38.92   Mean   : 5.013
##                                        3rd Qu.:44.00   3rd Qu.: 7.000
##                                        Max.   :92.00   Max.   :10.000
##     Balance       NumOfProducts    HasCrCard      IsActiveMember
##  Min.   :     0   Min.   :1.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 97199   Median :1.00   Median :1.0000   Median :1.0000
##  Mean   : 76486   Mean   :1.53   Mean   :0.7055   Mean   :0.5151
##  3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :250898   Max.   :4.00   Max.   :1.0000   Max.   :1.0000
##  EstimatedSalary         Exited
##  Min.   :    11.58   Min.   :0.0000
##  1st Qu.: 51002.11   1st Qu.:0.0000
##  Median :100193.91   Median :0.0000
##  Mean   :100090.24   Mean   :0.2037
##  3rd Qu.:149388.25   3rd Qu.:0.0000
##  Max.   :199992.48   Max.   :1.0000
```

Zavisna promenljiva



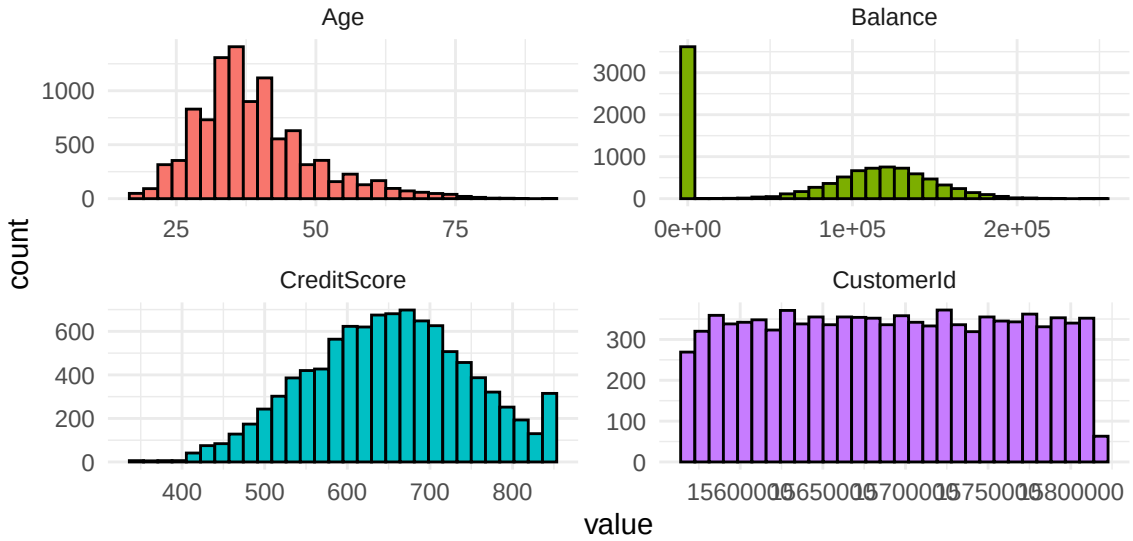

```
table(data$Exited)
```

```
##
##    0    1
## 7963 2037
```

Vidi se da većina korisnika nije napustila banku.

Analiza prediktora

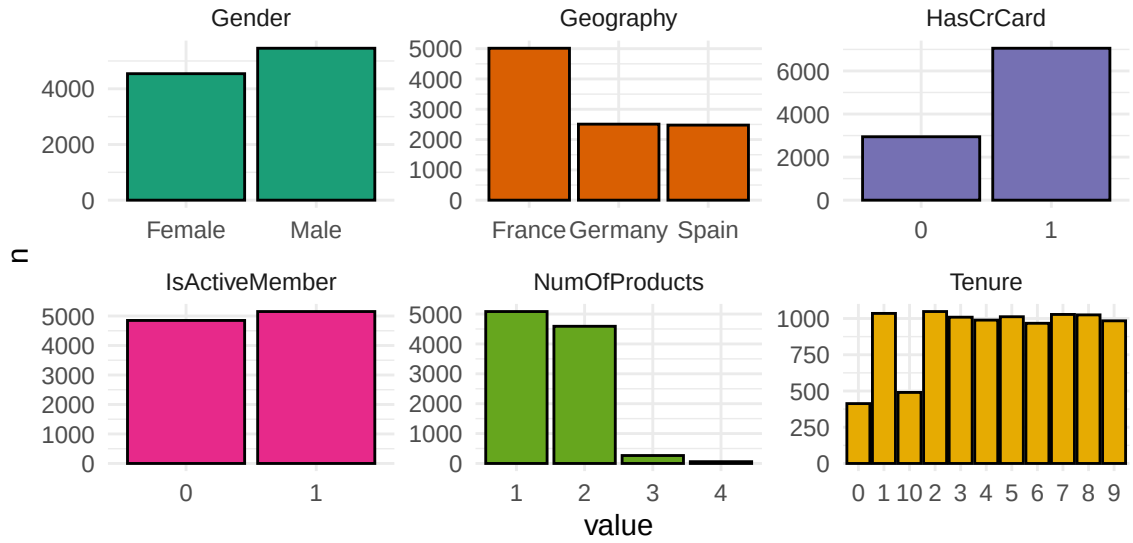
Pogledajmo prvo raspodele neprekidnih prediktora.



Zaključujemo:

- Raspodela prediktora *Age* je pomerena udesno.
- Prediktor *Balance* je blizu normalno raspodeljen.
- Većina predikotra *Credit score* je veća od 600. Moguće je da će baš ovi klijenti napustiti banku.

Pogledajmo sada raspodele kategoričkih prediktora.

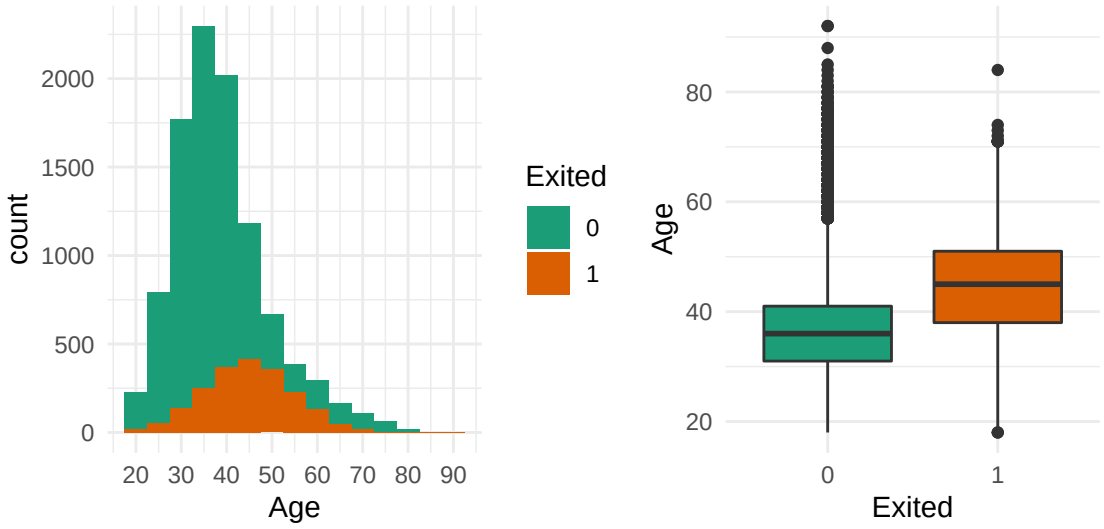


Zaključujemo:

- Veći broj klijenata je muškog pola.
- Klijenti su većinski iz Francuske.
- Većina klijenata ima kreditnu karticu.

- Broj aktivnih i neaktivnih članova je veoma sličan.
- Većina klijenata koristi 1 do 2 proizvoda banke, dok jako malo klijenata koristi 3 i 4 proizvoda.
- Broj klijenata koji su članovi banke $1, 2, ..., 9$ godina je približno isti.

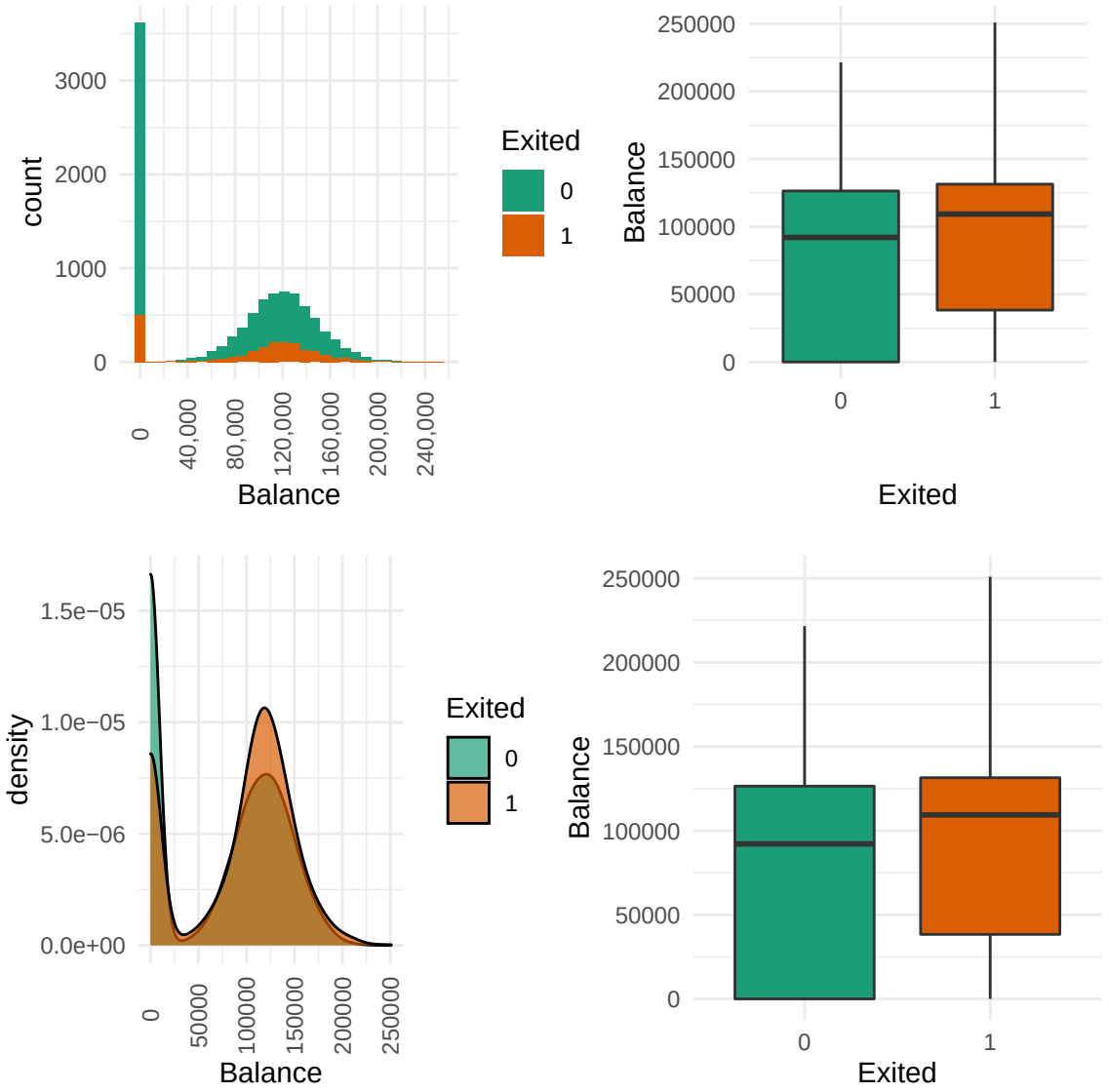
Prediktor *Age*



Zaključujemo:

- Klijenti koji su ostali u banci imaju tendenciju da budu mlađi.
- Veliki broj kllijenata koji su napustili banku ima između 40 i 50 godina.
- Klijenti starosti između 60 i 80 godina imaju tendenciju da ne napuštaju banku.

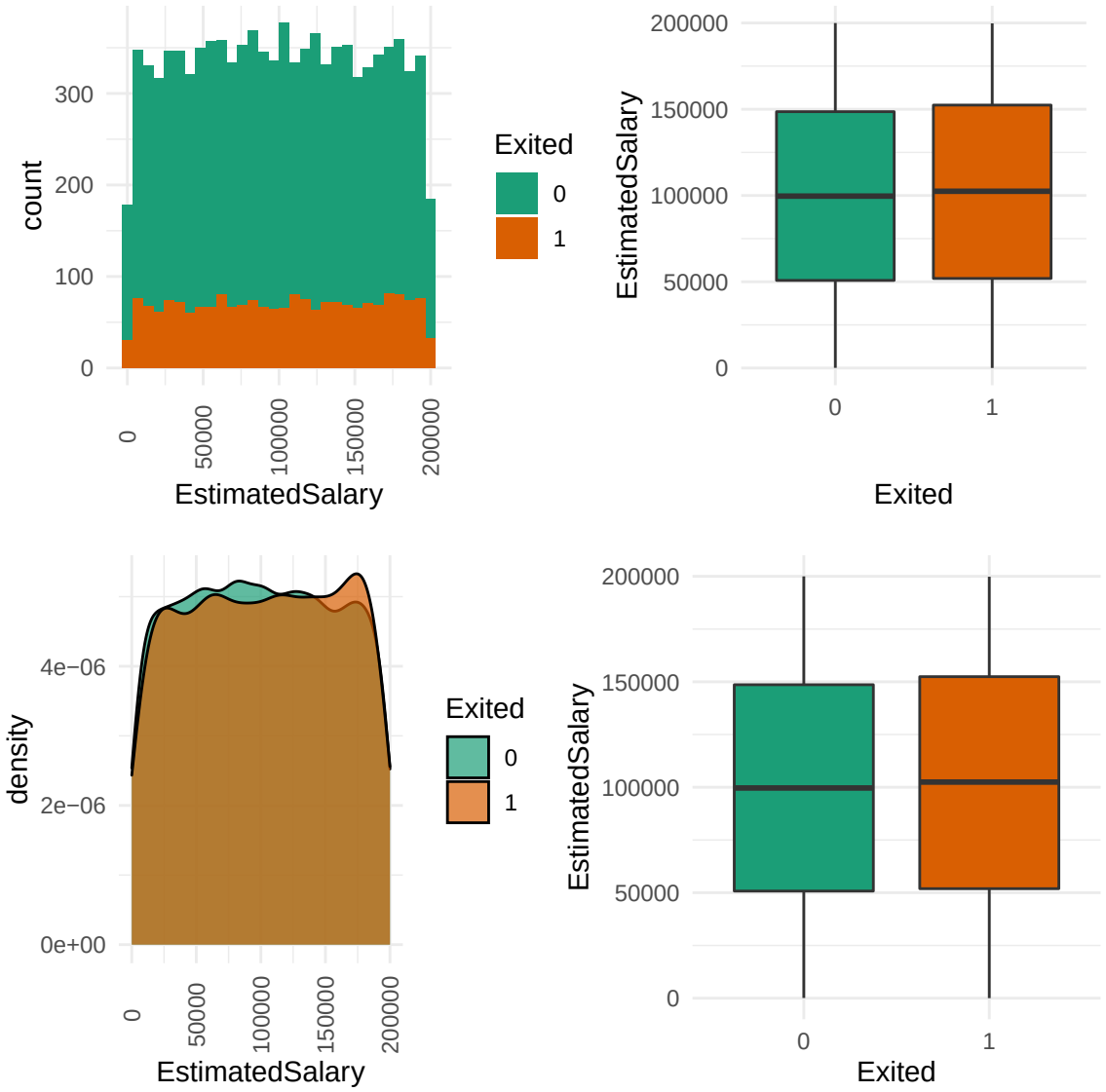
Prediktor *Balance*



Zaključujemo:

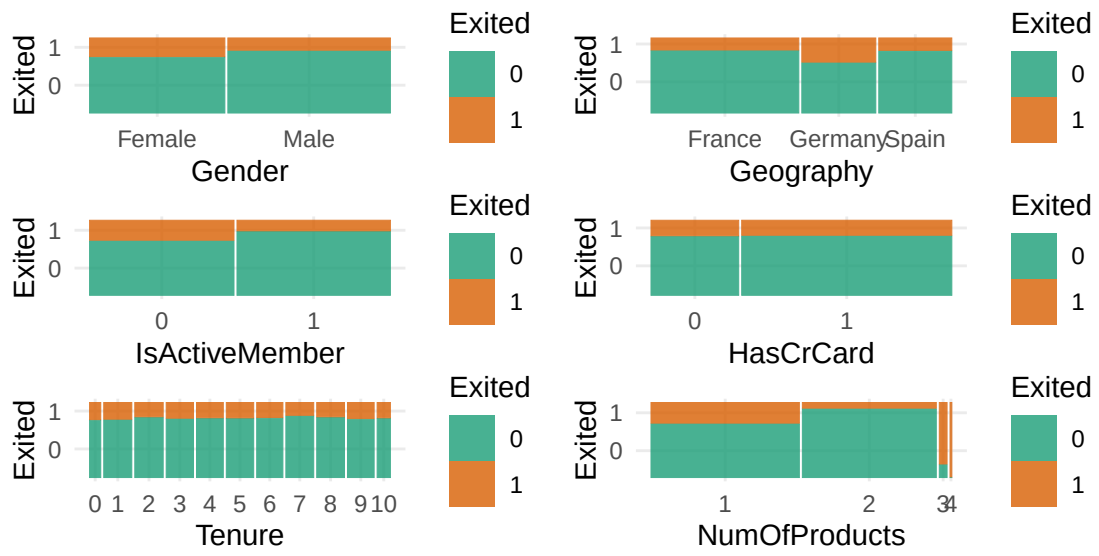
- Klijenti koji ostaju u banci imaju manje sredstava na računu od onih koji napuštaju banku.

Prediktor *Estimated Salary*



Zaključujemo:

- Ne postoji vidna razlika u zaradi između klijenata koji napuštaju/ne napuštaju banku.



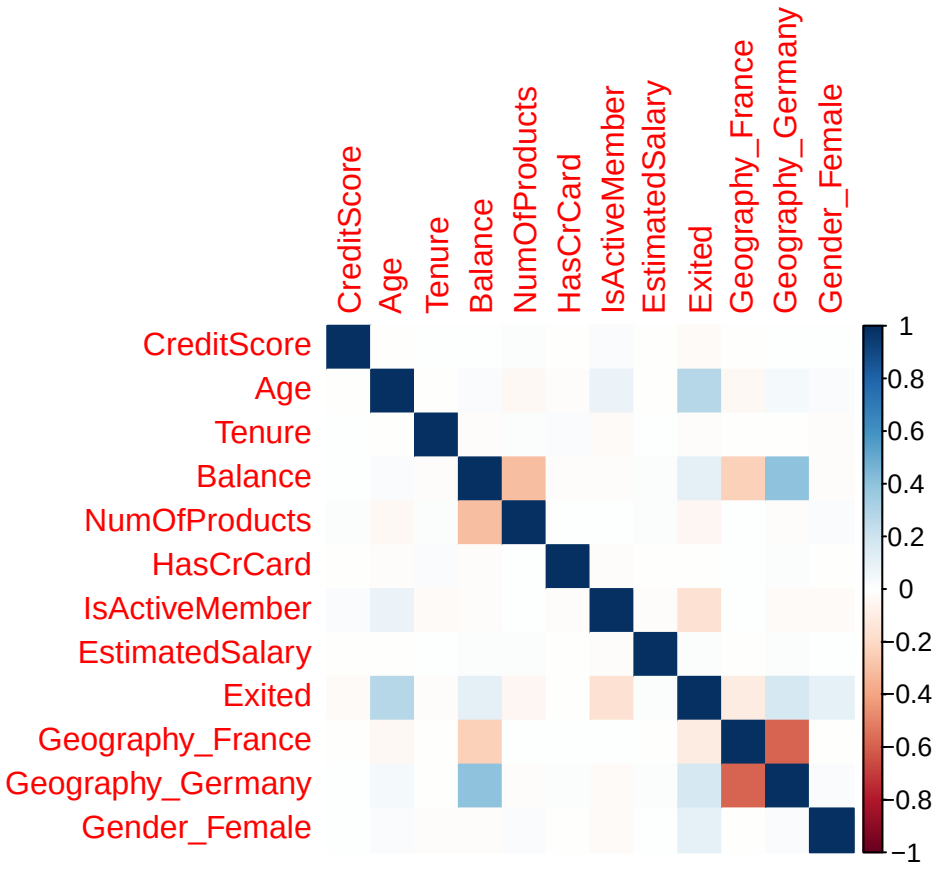
Čišćenje podataka (*eng. Data Cleaning*)

Kako smatramo da *RowNumber, CustomedId, Surname* nisu značajni prediktori, nećemo ih posmatrati u daljem radu. Kategorički prediktor *Geography* ćemo prebaciti u *dummy varijablu*, dok ćemo *Gender* kodirati binarno.

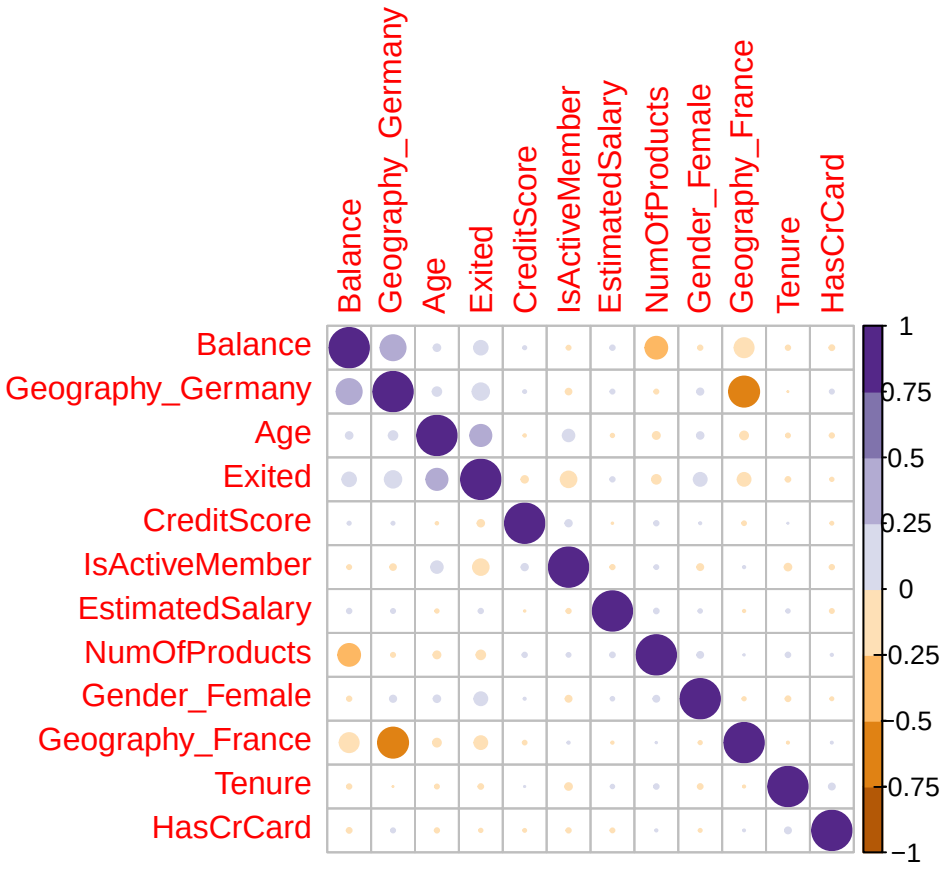
```
# odbacujemo navedene kolone
data <- data[, -which(names(data) %in% c("RowNumber", "CustomerId", "Surname"))]

# kolona Geography u dummy
data <- fastDummies::dummy_cols(data)

# odbacujemo Geography_Spain i Gender_Male, kako su zavisne od ostalih, kao i
# Geography i Gender
data <- data[, -which(names(data) %in% c("Geography_Spain",
                                         "Gender_Male",
                                         "Geography",
                                         "Gender"))]
```

Korelacija

```
corrplot(cor(data), method = "color")
```




```
corrplot(cor(data), order = "hclust",
         col = brewer.pal(n = 8, name = "PuOr"))
```



Kreiranje modela

```
index_train <- sample(nrow(data), 0.8 * nrow(data))
train <- data[index_train, ]
test <- data[-index_train, ]
```