Universitatea „Alexandru Ioan Cuza” din Iaşi
Facultatea de Economie și Administrarea Afacerilor
Master Data Mining

Proiect Regresia logistică

Student: Dancă Alexandra-Simona
310440105001SM211018
DM21

IAȘI 2023

FEAA
FACULTATEA DE ECONOMIE
ȘI ADMINISTRAREA AFACERILOR
UNIVERSITATEA "AL. I. CUZA" IAȘI

Cuprins

1. Prezentarea setului de date 3

1.1. Operații preliminare 3

1.2. Analiza descriptivă a variabilelor numerice și nenumerice 4

1.2.1. Analiza descriptiva a variabilelor numerice 4

1.2.2. Analiza descriptiva a variabilelor nenumerice 9

1.2.2. Identificarea outlierilor și tratarea acestora 11

2. Selectarea variabilelor prin aplicarea procedeului Purposeful 14

2.1. Regresie logistică simplă (univariată) pentru variabilele numerice 14

2.2. Tabel de contingență pentru variabilele nenumerice 17

2.3. Regresie logistică cu variabilele independente selectate 18

2.4. Testul raportul de verosimilitate 20

3. Selectarea variabilelor prin aplicarea procedeului Stepwise 22

4. Evaluarea ajustării modelului cu ajutorul Testului Omnibus 24

5. Evaluarea clasificării prin matricea de clasificare 25

6. Compararea celor două procedee de selectare a variabilelor- Coeficientul Mallows' Cq 27

7. Interpretarea modelului final 28

1. Prezentarea setului de date

Baza de date utilizată în această lucrare este obținută de pe https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset. Aceasta conține 12 variabile, atât numerice, cât și nenumerice pentru 10000 de clienți a unei bănci.

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
1	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
6	15574012	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	15592531	822	France	Male	50	7	0.00	2	1	1	10062.80	0
8	15656148	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	15792365	501	France	Male	44	4	142051.07	2	0	1	74940.50	0
10	15592389	684	France	Male	27	2	134603.88	1	1	1	71725.73	0

Figura 1. Baza de date Bank-Customer-Churn

Variabilele au următoarea semnificație:

- ***customer_id*** → ID client - ID unic dat pentru a identifica clientul;
- ***credit_score*** → Scorul de credit - Este scorul care determină solvabilitatea unui client;
- ***country*** → Țara - Țara în care locuiește clientul;
- ***gender*** → Gen - Sexul clientului;
- ***age*** → Vârsta - Vârsta clientului;
- ***tenure*** → Mandat - Numărul de ani de când clientul are un cont bancar la banca respectivă;
- ***balance*** → Sold - Suma de bani din contul clientului;
- ***products_number*** → Număr de produse - Numărul de produse/servicii deținute;
- ***credit_card*** → Card de credit – Dacă clientul deține un card de credit;
- ***active_member*** → Membru activ - Dacă clientul este membru activ;
- ***estimated_salary*** → Salariu estimat - Venitul total al clientului;
- ***churn*** → Pierderea clienților existenți ai băncii;

1.1. Operații preliminare

Primul pas efectuat pentru operațiile peliminare a fost verificarea dacă în baza de date există valori lipsă.

```
> colSums(is.na(Bank_Customer_Churn))
    customer_id     credit_score          country           gender              age           tenure
              0                0                0                0                0                0
        balance  products_number      credit_card  active_member estimated_salary            churn
              0                0                0                0                0                0
```

Figura 2. Verificare valori lipsă

Rezultatul afișează lipsa valorilor nule în baza de date inițiale, deci în continuare aceasta nu reprezintă o problemă pentru analiza dorită.

Mai departe verificăm dacă clasa variabilelor este cea potrivită.

```
> # clasele variabilelor
> sapply(Bank_Customer_Churn,class)
    customer_id     credit_score          country           gender              age           tenure
      "numeric"        "numeric"      "character"      "character"        "numeric"        "numeric"
        balance  products_number      credit_card    active_member estimated_salary            churn
      "numeric"        "numeric"        "numeric"        "numeric"        "numeric"        "numeric"
```

Figura 3. Clasele din care fac parte variabilele

Se poate observa că variabile precum ”*gender*”, ”*products_number*”, ”*credit_card”,* ”*active_member*” și ”*churn*” fac parte din tipul de date numeric sau character. În următorul cod acestea sunt transformate în varibile de tip factor.

```
> # transformam din char/numeric in factor
> Bank_Customer_Churn$gender<-as.factor(Bank_Customer_Churn$gender)
> Bank_Customer_Churn$products_number<-as.factor(Bank_Customer_Churn$products_number)
> Bank_Customer_Churn$credit_card<-as.factor(Bank_Customer_Churn$credit_card)
> Bank_Customer_Churn$active_member<-as.factor(Bank_Customer_Churn$active_member)
> Bank_Customer_Churn$churn<-as.factor(Bank_Customer_Churn$churn)
> # verificam daca s-au transformat
> sapply(Bank_Customer_Churn,class)
    customer_id     credit_score          country           gender              age           tenure
      "numeric"        "numeric"      "character"         "factor"        "numeric"        "numeric"
        balance  products_number      credit_card    active_member estimated_salary            churn
      "numeric"         "factor"         "factor"         "factor"        "numeric"         "factor"
```

Figura 4. Transformarea și verificarea modificărilor pentru clasele variabilelor

Transformarea lor a fost realizată cu succes.

```
levels(Bank_Customer_Churn$churn)<-c("nu","da")
```

Figura 5. Transformarea categoriilor variabilei churn

Mai departe s-a realizat transformarea categoriilor variabilei *churn* astfel în loc de numărul 0 avem ”nu” și în locul lui 1 avem ”da”. Această transformare ne ajută pentru diversificarea rezultatelor în analizele care urmează.

1.2. Analiza descriptivă a variabilelor numerice și nenumerice

Analiza descriptivă a variabilelor numerice analizează media, mediana, minimul, maximul cât și quartilele, iar pentru variabilele nenumerice o analiza descriptivă a grupurilor.

1.2.1. Analiza descriptiva a variabilelor numerice

Pentru această analiză voi crea un subset de date, în care voi include doar cele 5 variabile numerice pentru a putea aplica aceeași funcție asupra tuturor variabilelor numerice .

```
> #### 1.2. Analiza descripriva a variabilelor numerice si nenumerice ####
> ## 1.2.1. Pentru variabile numerice
> df_numeric <- Bank_Customer_Churn %>%
+   select (credit_score, age, tenure, balance, estimated_salary)
> summary(df_numeric)
  credit_score        age             tenure          balance       estimated_salary
 Min.   :350.0   Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :    11.58
 1st Qu.:584.0   1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.: 51002.11
 Median :652.0   Median :37.00   Median : 5.000   Median : 97199   Median :100193.91
 Mean   :650.5   Mean   :38.92   Mean   : 5.013   Mean   : 76486   Mean   :100090.24
 3rd Qu.:718.0   3rd Qu.:44.00   3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:149388.25
 Max.   :850.0   Max.   :92.00   Max.   :10.000   Max.   :250898   Max.   :199992.48
```

Figura 6. Analiza descriptiva a variabilelor numerice

A. Credit_score

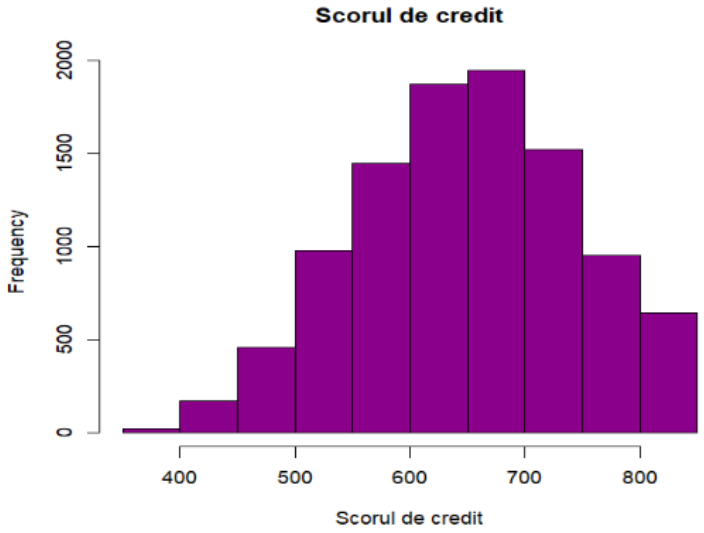

Figura 7. Histogramă pentru credit_score

Scorul de credite este în **medie** egal cu 650,5 de unități. Scorul de credit **minim** este egal cu 350 de unități. Scorul de credit **maxim** este egal cu 850 de unități. Acest scor indică gradul de risc cel mai ridicat care este asociat cu emiterea unui anumit împrumut unui client.

Q1 (quartila 1) - ne arată faptul că 25% dintre scorurile de credit au scorul de până la 584 de unități, iar restul de 75% peste 584 de unități.

Q2 (quartila2/mediana) - ne arată faptul că 50% dintre scorurile de credit au scorul de până la 652 de unități, iar restul de 50% peste 652 de unități.

Q3 (quartila3) - ne arată faptul că 75% dintre scorurile de credit au scorul de până la 718 de unități, iar restul de 25% peste 718 de unități.

B. Age

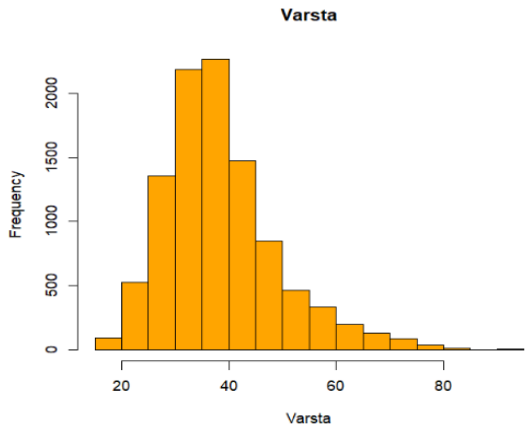

Figura 8. Histogramă pentru age

Vârsta **medie** este egală cu aproape 39 de ani. Vârsta **minimă** este de 18 ani iar cea **maximă** de 92 de ani.

Q1 (quartila 1) - ne arată faptul că 25% dintre clienți au vârsta de până la 32 de ani, iar restul de 75% peste 32 de ani.

Q2 (quartila2/mediana) - ne arată faptul că 50% dintre clienți au vârsta de până la 37 de ani, iar restul de 50% peste 37 de ani.

Q3 (quartila3) - ne arată faptul că 75% dintre clienți au vârsta de până la 44 de ani, iar restul de 25% peste 44 ani.

C. Tenure

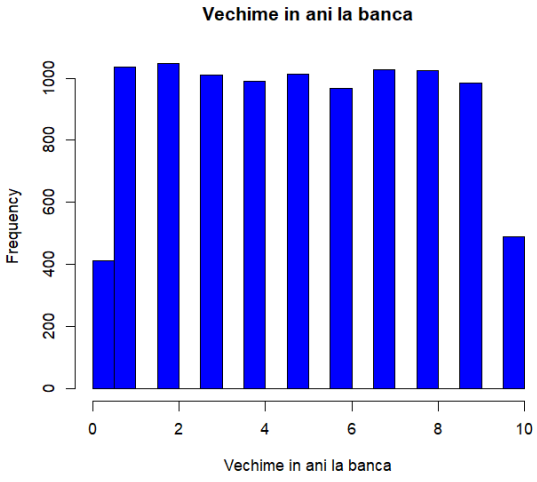

Figura 9. Histogramă pentru tenure

Numărul de ani de când un client este la bancă este în **medie** egal cu 5 ani. Numărul **minim** de ani de când un client este la bancă este de 0 ani și **maxim** de 10 ani.

Q1 (quartila 1) - ne arată faptul că 25% dintre clienți sunt la banca respectivă de mai puțin de 3 ani, iar restul de 75% peste 3 ani.

Q2 (quartila2/mediana) - ne arată faptul că 50% dintre clienți sunt la banca respectivă de mai puțin de 5 ani, iar restul de 50% peste 5 ani.

Q3 (quartila3) - ne arată faptul că 75% dintre clienți sunt la banca respectivă de mai puțin de 7 ani, iar restul de 25% peste 7 ani.

D. Balance

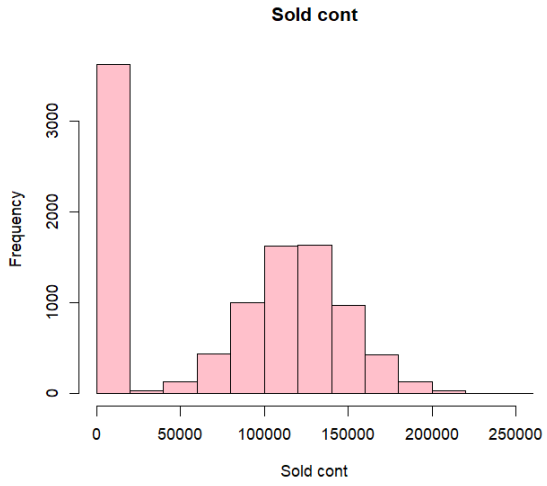

Figura 10 Histogramă pentru balance

Soldul clientului la banca este în **medie** egal cu 76486$. Soldul **maxim** este de 250898$ iar cel **minim** este de 0$.

Q1 (quartila 1) - ne arată faptul că 25% dintre clienți au soldul de mai puțin de 0$, iar restul de 75% de peste 0$.

Q2 (quartila2/mediana) - ne arată faptul că 50% dintre clienți au soldul de mai puțin de 97199$, iar restul de 50% de peste 97199$.

Q3 (quartila3) - ne arată faptul că 75% dintre clienți au soldul de mai puțin de 127644$, iar restul de 25% de peste 127644$.

E. Estimated_salary

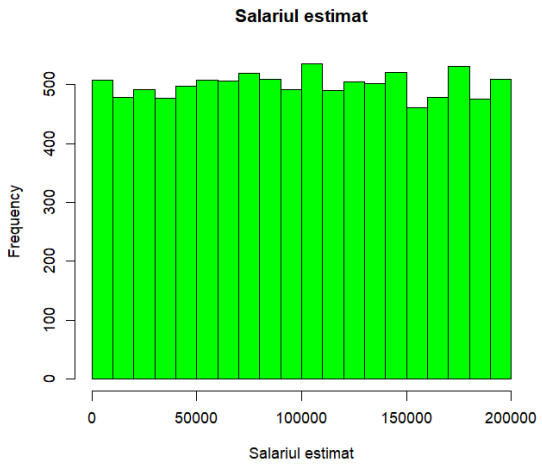

Figura 11. Histogramă pentru estimated_salary

Salariul estimat al clientului este în **medie** egal cu 100090,24$. Salariul **maxim** estimat al clientului este de 199992,48$. Salariul **minim** estimat al clientului este de 11,58$.

Q1 (quartila 1) - ne arată faptul că 25% dintre clienți au salariul estimat de de până la 51002,11$, iar restul de 75% peste 51002,11$.

Q2 (quartila2/mediana) - ne arată faptul că 50% dintre clienți au salariul estimat de de până la 100193,91$, iar restul de 50% peste 100193,91$.

Q3 (quartila3) - ne arată faptul că 75% dintre clienți au salariul estimat de de până la 149388,25$, iar restul de 25% peste 149388,25$.

Coeficientul de asimetrie (Skewness)

```
> skewness(df_numeric)
   credit_score            age         tenure        balance estimated_salary
   -0.071595867    1.011168559    0.010989809   -0.141087544     0.002085045
```

Figura 12. Coeficientul de asimetrie pentru variabilele numerice

Se observă că distribuția variabilei *credit_score* și *balance* este asimetrică la stânga deoarece -0.071595867 respectiv -0.141087544 sunt mai mici ca 0. Distribuția celorlalte trei variabile *age*, *tenure* și *estimated_salary* este asimetrică la dreapta deoarece 1.011168559, 0.010989809 și 0.002085045 sunt mai mari decât 0.

Coeficientul de boltire (Kurtosis)

```
> kurtosis(df_numeric)
   credit_score            age         tenure        balance estimated_salary
       2.573887        4.394050        1.834757        1.510733        1.818472
```

Figura 13. Coeficientul de boltire pentru variabilele numerice

Din figură se observă că distribuția celor cinci variabile este leptocurtică, deoarece valoarea aferentă boltirii este mai mare ca 0 .

1.2.2. Analiza descriptiva a variabilelor nenumerice

```
> ## 1.2.1. Pentru variabile nenumerice
> table(Bank_Customer_Churn$country)

 France Germany   Spain
   5014    2509    2477
> table(Bank_Customer_Churn$gender)

Female   Male
  4543   5457
> table(Bank_Customer_Churn$products_number)

   1    2    3    4
5084 4590  266   60
> table(Bank_Customer_Churn$credit_card)

   0    1
2945 7055
> table(Bank_Customer_Churn$active_member)

   0    1
4849 5151
> table(Bank_Customer_Churn$churn)

  nu   da
7963 2037
```

Figura 14. 1.2.2. Analiza descriptiva a variabilelor nenumerice

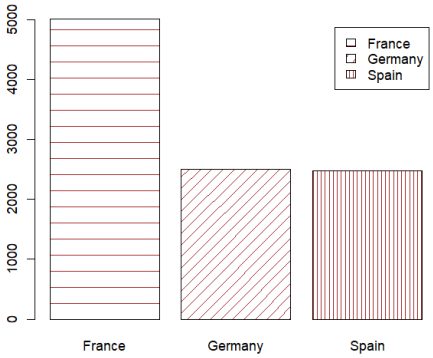

Figura 15. Histogramă variabilă country

Pentru variabila *country* predomină cu 5014 țara ”France”.

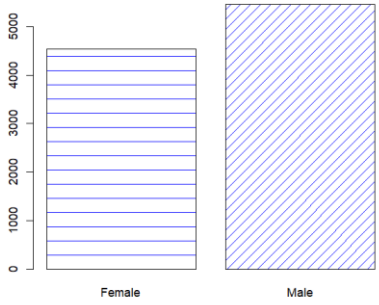

Figura 16. Histogramă variabilă gender

Pentru variabila *gender* predomină cu 5457 de observații categoria ”Male” față de ”Female” care are 4543 de observații.

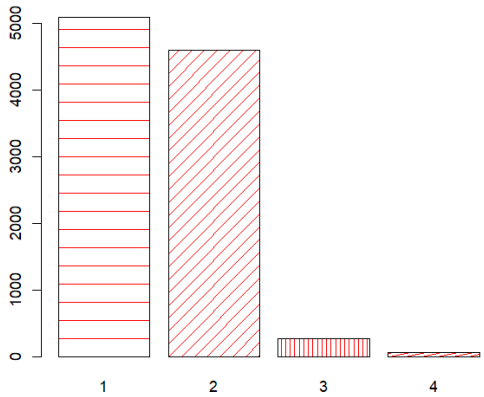

Figura 17. Histogramă variabilă products_number

Majoritatea clienților dețin doar un singur produs de la bancă. Pe locul 2 se află pachetul cu 2 produse urmat de cel cu 3 produse și în cele din urmă pachetul cu 4 produse deținut doar de 60 de persoane.

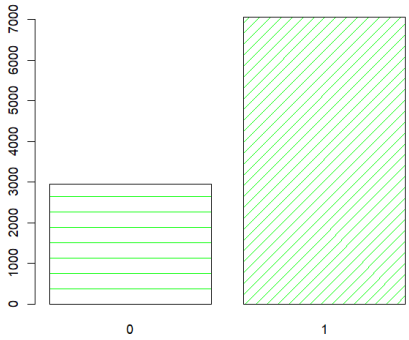

Figura 18. Histogramă variabilă credit_card

Variabila *credit_card* scoate în evidență că 7055 de persoanedețin un card de credit în timp ce 2945 nu.

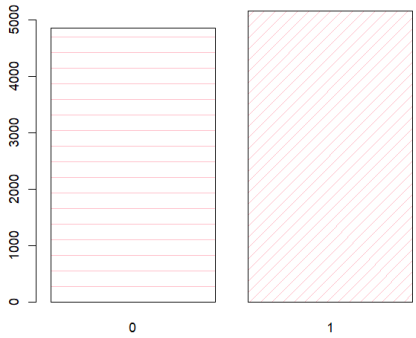

Figura 19. Histogramă variabilă active_member

Variabila *active_member* arată că 5151 de obsevații încă mai sunt membri activi ai băncii.

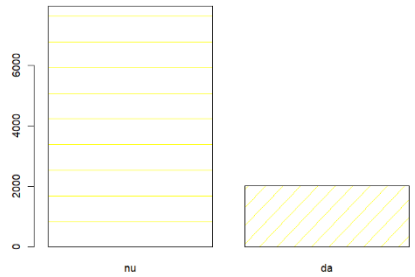

Figura 20. Histogramă variabilă churn

Cei mai mulți clienți ai băncii nu au renunțat înca la serviciile prestate de bancă, în timp ce 2037 au renunțat la aceste servicii.

1.2.2. Identificarea outlierilor și tratarea acestora

Pentru a identifica valorile extreme pentru variabilele studiate se va folosi graficul Box plot.

```
> ### Oulieri
> # vedem ce dimensiune are baza noastra de date inainte de eliminarea outlierilor
> dim(Bank_Customer_Churn)
[1] 10000    12
> # vizualizam boxplot
> boxplot(Bank_Customer_Churn$credit_score)
> boxplot(Bank_Customer_Churn$age)
> boxplot(Bank_Customer_Churn$tenure)
> boxplot(Bank_Customer_Churn$balance)
> boxplot(Bank_Customer_Churn$estimated_salary)
```

Figura 21. Dimensiunea bazei de date înainte de eliminarea outlierilor

Se remarcă faptul că avem 10000 de observații înainte de eliminarea oulierilor.

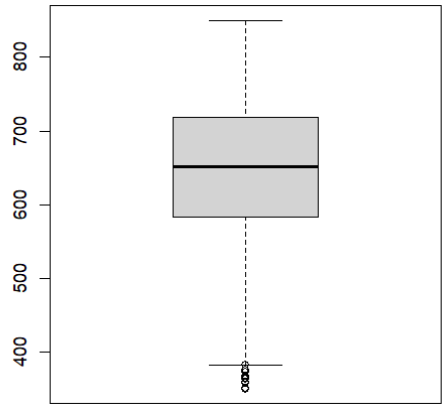

Figura 22. BoxPlot credit score

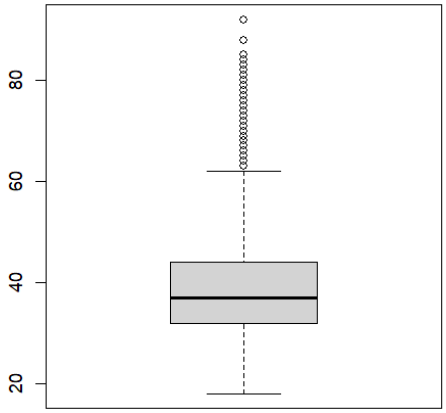

Figura 23. BoxPlot age

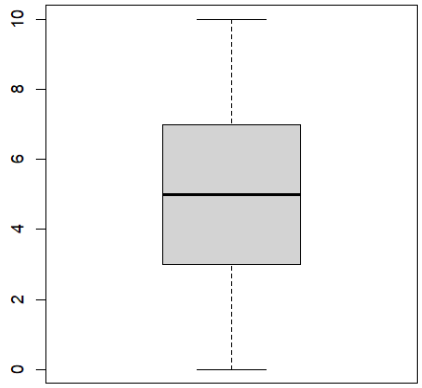

Figura 24. BoxPlot tenure

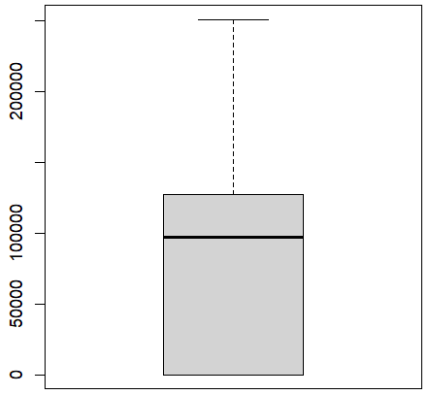

Figura 25. BoxPlot balance

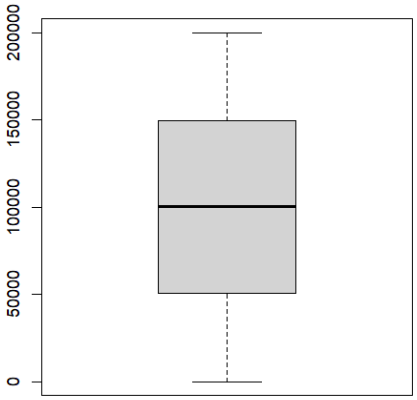

Figura 26. BoxPlot estimated_salary

Putem observa că doar în diagramele box-plot pentru *credit_score* și *age* avem valori extreme. Acestea vor fi eliminate deoarece afecteaza analizele care urmează.

```
> # eliminam outlierii pentru credit_score
> quartiles <- quantile(Bank_Customer_Churn$credit_score, probs=c(.25, .75), na.rm = FALSE)
> IQR <- IQR(Bank_Customer_Churn$credit_score)
> Lower <- quartiles[1] - 1.5*IQR
> Upper <- quartiles[2] + 1.5*IQR
> data_no_outlier <- subset(Bank_Customer_Churn, Bank_Customer_Churn$credit_score > Lower & Bank_Customer_Churn$credit_score <
 Upper)
> # eliminam outlierii pentru age
> quartiles <- quantile(data_no_outlier$age, probs=c(.25, .75), na.rm = FALSE)
> IQR <- IQR(data_no_outlier$age)
> Lower <- quartiles[1] - 1.5*IQR
> Upper <- quartiles[2] + 1.5*IQR
> Bank_Customer_Churn <- subset(data_no_outlier, data_no_outlier$age > Lower & data_no_outlier$age < Upper)
> # vizualizam boxplot pentru cele 2 modificate
> boxplot(Bank_Customer_Churn$credit_score)
> boxplot(Bank_Customer_Churn$age)
> # vedem ce dimensiune are baza noastra de date dupa eliminarea outlierilor
> dim(Bank_Customer_Churn)
[1] 9573   12
```

Figura 27. Dimensiunea bazei de date după de eliminarea outlierilor

De asemenea se poate remarca faptul că din 10000 de observații, cât erau la început, au mai ramas 9573 după eliminarea outlierilor pentru *credit_score* și *age*.

Baza de date este împărțită în 70% training si 30% test.

```
> #### impart baza in training si test
> #70% training si 30% test
> sample <- sample(c(TRUE, FALSE), nrow(Bank_Customer_Churn), replace=TRUE, prob=c(0.7,0.3))
> train  <- Bank_Customer_Churn[sample, ]
> test   <- Bank_Customer_Churn[!sample, ]
> dim(train)
[1] 6719   12
> dim(test)
[1] 2854   12
```

Figura 28. Dimensiune df training și test

Baza de date training este formată din 6719 de observații în timp ce baza de date test are 2854 de observații.

2. Selectarea variabilelor prin aplicarea procedeului Purposeful

2.1. Regresie logistică simplă (univariată) pentru variabilele numerice

Formularea ipotezelor:

- H0 : nu există nicio diferență semnificativă.

- H1: există o diferență semnificativă. Variabila predictivă este semnificativă statistic în model.

Regula de decizie:

- $P_{value} < 0,25$, se respinge H0

- $P_{value} \geq 0,25$, nu se respinge H0

Credit_score

```
> model1 <- glm(churn ~ credit_score, data = train, family = 'binomial')
> summary(model1)

Call:
glm(formula = churn ~ credit_score, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7247  -0.6808  -0.6628  -0.6382   1.8547

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9079081  0.2074959  -4.376 1.21e-05 ***
credit_score -0.0007231  0.0003174  -2.278   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6683.0  on 6649  degrees of freedom
Residual deviance: 6677.8  on 6648  degrees of freedom
AIC: 6681.8

Number of Fisher Scoring iterations: 4
```

Figura 29. Output model

Interpretare: Valoarea P_{value} este 0.0227 deci se respinge ipoteza nulă conform căreia nu exisă nicio diferență semnificativă. În concluzie variabila *credit_score* este semnificativă în model.

Estimate: o creștere de o unitate a variabilei de predicție *credit_score* este asociată cu o modificare medie de -0,0007231 a șanselor logaritmice ale variabilei răspuns că iau o valoare de 1. Aceasta înseamnă că valorile mai mari ale *credit_score* sunt asociate cu o probabilitate mai mică.

Age

```
> model2 <- glm(churn ~ age, data = train, family = 'binomial')
> summary(model2)

Call:
glm(formula = churn ~ age, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6008  -0.6430  -0.4732  -0.3093   2.7606

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.774000   0.169375  -34.09   <2e-16 ***
age          0.110328   0.003997   27.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6683.0  on 6649  degrees of freedom
Residual deviance: 5781.7  on 6648  degrees of freedom
AIC: 5785.7

Number of Fisher Scoring iterations: 5
```

Figura 30.Output model

Interpretare: Valoarea P_{value} este 2e-16 deci se respinge ipoteza nulă conform căreia nu există nicio diferență semnificativă. În concluzie variabila *age* este semnificativă în model.

Tenure

```
> model3 <- glm(churn ~ tenure, data = train, family = 'binomial')
> summary(model3)

Call:
glm(formula = churn ~ tenure, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6995  -0.6820  -0.6648  -0.6480   1.8328

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.28302    0.06027 -21.287   <2e-16 ***
tenure      -0.01901    0.01062  -1.791   0.0734 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6683.0  on 6649  degrees of freedom
Residual deviance: 6679.8  on 6648  degrees of freedom
AIC: 6683.8

Number of Fisher Scoring iterations: 4
```

Figura 31. Output model

Interpretare: Valoarea P_{value} este 0.0734 deci nu se respinge ipoteza nulă conform căreia nu există nicio diferență semnificativă. În concluzie variabila *tenure* nu este semnificativă în model.

Balance

```
> summary(model4)

Call:
glm(formula = churn ~ balance, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8830  -0.7231  -0.5665  -0.5665   1.9540

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.749e+00  5.367e-02 -32.582   <2e-16 ***
balance      4.548e-06  5.059e-07   8.991   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6683.0  on 6649  degrees of freedom
Residual deviance: 6599.8  on 6648  degrees of freedom
AIC: 6603.8

Number of Fisher Scoring iterations: 4
```

Figura 32. Output model

Interpretare: Valoarea P_{value} este 2e-16 deci se respinge ipoteza nulă conform căreia nu exisă nicio diferență semnificativă. În concluzie variabila *balance* este semnificativă în model.

Esimated_salary

```
> model5 <- glm(churn ~ estimated_salary , data = train, family = 'binomial')
> summary(model5)

Call:
glm(formula = churn ~ estimated_salary, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6828  -0.6752  -0.6679  -0.6604   1.8077

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.417e+00  6.168e-02 -22.968   <2e-16 ***
estimated_salary  3.965e-07  5.319e-07   0.746    0.456
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6683.0  on 6649  degrees of freedom
Residual deviance: 6682.4  on 6648  degrees of freedom
AIC: 6686.4

Number of Fisher Scoring iterations: 4
```

Figura 33. Output model

Interpretare: Valoarea P_{value} este 0.456 deci nu se respinge ipoteza nulă conform căreia nu exisă nicio diferență semnificativă. În concluzie variabila *estimated_salary* nu este semnificativă în model.

2.2. Tabel de contingență pentru variabilele nenumerice

Să presupunem că pentru aceaste variabile dorim să testăm dacă grupele pe care le dețin sunt la fel distribuite numeric.

Formularea ipotezelor:

- H0 : nu există nicio diferență semnificativă între frecvențele observate și cele așteptate

- H1: există o diferență semnificativă între frecvențele observate și cele așteptate

Regula de decizie:

- $P_{value} < 0,25$, se respinge H0

- $P_{value} \geq 0,25$, nu se respinge H0

Gender

```
> test_gender <- chisq.test(table(train$gender), p = c(1/2, 1/2))
> test_gender

        Chi-squared test for given probabilities

data:  table(train$gender)
X-squared = 60.735, df = 1, p-value = 6.529e-15
```

Figura 34. Output Chi-squared test

Interpretare: valoarea P_{value} este 6.529e-15 deci, la nivelul de semnificație de 25%, se respingem ipoteza nulă conform căreia frecvențele observate și așteptate sunt egale.

Products_number

```
> test_products_number <- chisq.test(table(train$products_number),
+                                    p = c(1/4, 1/4, 1/4, 1/4))
> test_products_number

        Chi-squared test for given probabilities

data:  table(train$products_number)
X-squared = 5899.6, df = 3, p-value < 2.2e-16
```

Figura 35. Output Chi-squared test

Interpretare: valoarea P_{value} este 2.2e-16 deci, la nivelul de semnificație de 25%, se respingem ipoteza nulă conform căreia frecvențele observate și așteptate sunt egale.

Credit_card

```
> test_credit_card <- chisq.test(table(train$credit_card),
+                                p = c(1/2, 1/2))
> test_credit_card

        Chi-squared test for given probabilities

data:  table(train$credit_card)
X-squared = 1142.1, df = 1, p-value < 2.2e-16
```

Figura 36. Output Chi-squared test

Interpretare: valoarea P_{value} este 2.2e-16 deci, la nivelul de semnificație de 25%, se respingem ipoteza nulă conform căreia frecvențele observate și așteptate sunt egale.

Active_member

```
> test_active_member <- chisq.test(table(train$active_member),
+                                  p = c(1/2, 1/2))
> test_active_member

        Chi-squared test for given probabilities

data:  table(train$active_member)
X-squared = 0.021622, df = 1, p-value = 0.8831
```

Figura 37. Output Chi-squared test

Interpretare: valoarea P_{value} este 0.8831 deci nu se respingem ipoteza nulă conform căreia frecvențele observate și așteptate sunt egale.

Country

```
> #Country
> test_country <- chisq.test(table(train$country),
+                                    p = c(1/3, 1/3, 1/3))
> test_country

        Chi-squared test for given probabilities

data:  table(train$country)
X-squared = 909.34, df = 2, p-value < 2.2e-16
```

Figura 38. Output Chi-squared test

Interpretare: valoarea P_{value} este 2.2e-16 deci se respingem ipoteza nulă conform căreia frecvențele observate și așteptate sunt egale.

2.3. Regresie logistică cu variabilele independente selectate

Primul nostru model multivariabil conține toate covariatele care sunt semnificative în analiza univariabilă cu un nivel de semnificațiede 25%.

Formularea ipotezelor:

- H0 : nu există nicio diferență semnificativă

- H1: există o diferență semnificativă.

Regula de decizie:

- $P_{value} < 0,05$, se respinge H0, cu un risc asumat de 5%

- $P_{value} \geq 0,05$, nu se respinge H0, cu o probabilitate de 95%

```
> model_all_1 <- glm(churn ~ credit_score+age+tenure+balance+gender+products_number+credit_card+country, data = train,
 family = 'binomial')
> summary(model_all_1)

Call:
glm(formula = churn ~ credit_score + age + tenure + balance +
    gender + products_number + credit_card + country, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3523  -0.5792  -0.3492  -0.1806   3.2357

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.522e+00  3.249e-01 -13.918  < 2e-16 ***
credit_score     -5.875e-04  3.799e-04  -1.547   0.1220
age               1.034e-01  4.297e-03  24.068  < 2e-16 ***
tenure           -5.070e-05  1.243e-02  -0.004   0.9967
balance          -1.423e-06  7.027e-07  -2.025   0.0429 *
genderMale       -5.437e-01  7.269e-02  -7.479 7.47e-14 ***
products_number2 -1.630e+00  8.807e-02 -18.511  < 2e-16 ***
products_number3  2.427e+00  2.142e-01  11.329  < 2e-16 ***
products_number4  1.607e+01  2.067e+02   0.078   0.9380
credit_card1      1.381e-03  7.924e-02   0.017   0.9861
countryGermany    9.168e-01  8.946e-02  10.249  < 2e-16 ***
countrySpain     -9.517e-02  9.514e-02  -1.000   0.3171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6767.8  on 6691  degrees of freedom
Residual deviance: 4902.8  on 6680  degrees of freedom
AIC: 4926.8

Number of Fisher Scoring iterations: 14
```

Figura 39. Output model M1

Pentru *credit_score*, valoarea p = 0.1220, care este > 0,05 arată că, cu o probabilitate de 95% nu se respinge H0, deci nu există nicio diferență semnificativă.

Pentru *age*, valoarea p = 2e-16, care este < 0,05 arată că, cu un risc asumat de 5% se respinge H0, deci există diferență semnificativă.

Pentru *tenure*, valoarea p = 0.9967, care este > 0,05 arată că, cu o probabilitate de 95% nu se respinge H0, deci nu există nicio diferență semnificativă.

Pentru *balance*, valoarea p = 0.0429, care este < 0,05 arată că, cu un risc asumat de 5% se respinge H0, deci există diferență semnificativă.

Pentru *genderMale*, valoarea p = 7.47e-14, care este < 0,05 arată că, cu un risc asumat de 5% se respinge H0, deci există diferență semnificativă.

Pentru *products_number4*, valoarea p = 0.9380, care este > 0,05 arată că, cu o probabilitate de 95% nu se respinge H0, deci nu există nicio diferență semnificativă.

Pentru *credit_card1*, valoarea p = 0.9861, care este > 0,05 arată că, cu o probabilitate de 95% nu se respinge H0, deci nu există nicio diferență semnificativă.

Pentru *countrySpain*, valoarea p = 0.3171, care este > 0,05 arată că, cu o probabilitate de 95% nu se respinge H0, deci nu există nicio diferență semnificativă.

În concluzie variabilele precum *credit_score*, *tenure*, *balance*, *products_number*, *credit_card* și *country* sunt eliminate modelul redus care urmează să fie facut.

```
> model_all_2 <- glm(churn ~ age+balance+gender, data = train, family = 'binomial')
> summary(model_all_2)

Call:
glm(formula = churn ~ age + balance + gender, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9208  -0.6391  -0.4496  -0.2695   3.0094

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.920e+00  1.821e-01 -32.516   <2e-16 ***
age          1.109e-01  4.071e-03  27.241   <2e-16 ***
balance      5.315e-06  5.556e-07   9.568   <2e-16 ***
genderMale  -5.931e-01  6.724e-02  -8.821   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6715.0  on 6659  degrees of freedom
Residual deviance: 5646.5  on 6656  degrees of freedom
AIC: 5654.5

Number of Fisher Scoring iterations: 5
```

Figura 40. Output model M2

Cele 3 variabile *age*, *balance* și *gender* au p-value < 0,05, deci, cu un risc asumat de 5% se respinge H0 și există diferență semnificativă.

2.4. Testul raportul de verosimilitate

În cele din urmă, calculăm statistica noastră de testare. Pentru a face acest lucru, găsim loglikelihood-urile fiecărui model și le conectăm la formula **-2*[loglikelihood(M2)-loglikelihood(M1)]**. Statistica noastră de test urmează o distribuție chi-pătrat cu grade de libertate egale cu diferența dintre numărul de parametri liberi dintre modelul complex (M1) și modelul mai putin complex (M2). Cu aceste informații, putem calcula valoarea p, iar dacă este mai mică decât nivelul nostru de semnificație, respingem ipoteza nulă.

Deoarece funcția logLik() oferă mai multe informații decât valoarea numerică, utilizați funcția as.numeric() pentru a izola valoarea numerică.

Formularea ipotezelor:

- H0 : Atât modelul mai puțin redus (M1), cât și cel mai redus (M2) se potrivesc la fel de bine cu datele. Ca rezultat, ar trebui să utilizăm modelul cel mai redus (M2).

- H1: Modelul mai puțin redus (M1) depășește semnificativ modelul cel mai redus (M2) în ceea ce privește potrivirea datelor. Ca rezultat, ar trebui să utilizăm modelul M1.

Regula de decizie:

- P_{value} < 0,05, se respinge H0, cu un risc asumat de 5%

- $P_{value} \geq 0,05$, nu se respinge H0, cu o probabilitate de 95%

```
> #testul raportul de verosimilitate
> #H0: Modelul M2 este mai bun
> #H1: Modelul M1 este mai bun
> library(lmtest)
> (A <- logLik(model_all_2))
'log Lik.' -2884.691 (df=4)
> (B <- logLik(model_all_1))
'log Lik.' -2451.384 (df=12)
> (teststat <- -2 * (as.numeric(A)-as.numeric(B)))
[1] 866.6156
> #df = 12 - 4 = 8
> (p.val <- pchisq(teststat, df = 8, lower.tail = FALSE))
[1] 8.954757e-182
> lrtest(model_all_2, model_all_1)
Likelihood ratio test

Model 1: churn ~ age + balance + gender
Model 2: churn ~ credit_score + age + tenure + balance + gender + products
_number +
    credit_card + country
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -2884.7
2  12 -2451.4  8 866.62  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 41. Output LRT

Interpretare: Putem vedea din rezultat că testul raportului de probabilitate are o valoare p de 2.2e-16 care este <0,05 de aceea respingem ipoteza nulă deoarece cu un risc asumat de 5%.

Alternativ, putem folosi analiza variației (ANOVA) pentru a explora diferența dintre modele.

```
> anova( model_all_1 , model_all_2 , test="Chisq")
Analysis of Deviance Table

Model 1: churn ~ credit_score + age + tenure + balance + gender + products_number +
    credit_card + country
Model 2: churn ~ age + balance + gender
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      6742     4893.7
2      6750     5753.2 -8  -859.48 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 42. Output ANOVA

Rezultatele sunt exact aceleași. Ca rezultat, putem concluziona că modelul cu 8 predictori depășește modelul cu 3 predictori deoarece crește acuratețea modelului nostru cu o cantitate substanțială.

3. Selectarea variabilelor prin aplicarea procedeului Stepwise

Procedeului Stepwise este o combinație de selecții *forward* și *backward*. Se începe modelul fără predictori, apoi adăugăm secvențial cei mai contributivi predictori (ca *forward selection*). După adăugarea fiecărei variabile noi, eliminăm orice variabilă care nu mai oferă o îmbunătățire a potrivirii modelului (ca *backward selection*).

```
> train$churn<-as.numeric(train$churn)
> both <-step(lm(churn~.,data=train),direction="both")
Start:  AIC=-14426.68
churn ~ customer_id + credit_score + country + gender + age +
    tenure + balance + products_number + credit_card + active_member +
    estimated_salary

                  Df Sum of Sq    RSS    AIC
- tenure           1     0.001 771.53 -14429
- credit_card      1     0.001 771.53 -14429
- estimated_salary 1     0.102 771.63 -14428
- credit_score     1     0.131 771.66 -14428
<none>                       771.53 -14427
- customer_id      1     0.283 771.81 -14426
- balance          1     1.652 773.18 -14414
- gender           1     6.455 777.98 -14373
- country          2    18.324 789.85 -14274
- active_member    1    19.601 791.13 -14261
- age              1    93.352 864.88 -13664
- products_number  3   115.530 887.06 -13499
```

Figura 43. Output

În primul rând, ne-am potrivit modelul numai cu variabile intercept. Apoi, am adăugat predictori la model secvențial. Cu toate acestea, după adăugarea fiecărui predictor, am eliminat și orice predictori care nu mai oferă o îmbunătățire a potrivirii modelului.

```
Step:  AIC=-14428.68
churn ~ customer_id + credit_score + country + gender + age +
    balance + products_number + credit_card + active_member +
    estimated_salary

                  Df Sum of Sq    RSS    AIC
- credit_card      1     0.001 771.53 -14431
- estimated_salary 1     0.102 771.63 -14430
- credit_score     1     0.131 771.66 -14430
<none>                       771.53 -14429
- customer_id      1     0.283 771.81 -14428
+ tenure           1     0.001 771.53 -14427
- balance          1     1.652 773.18 -14416
- gender           1     6.464 777.99 -14375
- country          2    18.324 789.85 -14276
- active_member    1    19.603 791.13 -14263
- age              1    93.352 864.88 -13666
- products_number  3   115.538 887.07 -13501
```

Figura 44. Output

```
Step:  AIC=-14430.67
churn ~ customer_id + credit_score + country + gender + age +
    balance + products_number + active_member + estimated_salary

                  Df Sum of Sq    RSS    AIC
- estimated_salary  1     0.102 771.63 -14432
- credit_score      1     0.131 771.66 -14432
<none>                         771.53 -14431
- customer_id       1     0.282 771.81 -14430
+ credit_card       1     0.001 771.53 -14429
+ tenure            1     0.001 771.53 -14429
- balance           1     1.651 773.18 -14418
- gender            1     6.466 778.00 -14377
- country           2    18.333 789.86 -14278
- active_member     1    19.602 791.13 -14265
- age               1    93.399 864.93 -13668
- products_number   3   115.538 887.07 -13503
```

Figura 45. Output

```
Step:  AIC=-14431.78
churn ~ customer_id + credit_score + country + gender + age +
    balance + products_number + active_member

                  Df Sum of Sq    RSS    AIC
- credit_score      1     0.130 771.76 -14433
<none>                         771.63 -14432
- customer_id       1     0.277 771.91 -14431
+ estimated_salary  1     0.102 771.53 -14431
+ credit_card       1     0.001 771.63 -14430
+ tenure            1     0.001 771.63 -14430
- balance           1     1.651 773.28 -14420
- gender            1     6.469 778.10 -14378
- country           2    18.391 790.02 -14278
- active_member     1    19.626 791.26 -14266
- age               1    93.347 864.98 -13670
- products_number   3   115.572 887.20 -13504
```

Figura 46. Output

```
Step:  AIC=-14432.65
churn ~ customer_id + country + gender + age + balance + products_number +
    active_member

                  Df Sum of Sq    RSS    AIC
<none>                         771.76 -14433
- customer_id       1     0.282 772.05 -14432
+ credit_score      1     0.130 771.63 -14432
+ estimated_salary  1     0.102 771.66 -14432
+ credit_card       1     0.001 771.76 -14431
+ tenure            1     0.001 771.76 -14431
- balance           1     1.667 773.43 -14420
- gender            1     6.458 778.22 -14379
- country           2    18.432 790.19 -14279
- active_member     1    19.673 791.44 -14266
- age               1    93.443 865.21 -13670
- products_number   3   115.638 887.40 -13504
```

Figura 47. Output

Am repetat acest proces până am ajuns la un model final. În total s-au realizat 5 modele.

```
> both$anova
               Step Df     Deviance Resid. Df Resid. Dev       AIC
1                   NA           NA      6677   771.5294 -14426.68
2          - tenure  1 0.0006465545      6678   771.5301 -14428.68
3     - credit_card  1 0.0007346908      6679   771.5308 -14430.67
4 - estimated_salary  1 0.1021960667      6680   771.6330 -14431.78
5    - credit_score  1 0.1303098977      6681   771.7633 -14432.65
```

Figura 48. Output

4. Evaluarea ajustării modelului cu ajutorul Testului Omnibus

Se testează dacă modelul cu predictori este semnificativ diferit de modelul fără predictori (modelul nul). Acest test poate fi interpretat ca un test al capacității tuturor predictorilor din model de a prezice variabila răspuns.

Formularea ipotezelor:

- H0 : nu există nicio diferență semnificativă

- H1: există diferență semnificativă. Modelul curent este mai bun decât modelul nul.

Regula de decizie:

- $P_{value} < 0,05$, se respinge H0, cu un risc asumat de 5%

- $P_{value} \geq 0,05$, nu se respinge H0, cu o probabilitate de 95%

```
> logit2.res <- lrm(churn ~ credit_score+country+gender+age+tenure+
+                     balance+products_number+credit_card+active_member+
+                     estimated_salary, data = train, y = TRUE, x = TRUE)
> residuals(logit2.res, type = "gof")
Sum of squared errors     Expected value|H0                    SD                     Z                     P
         726.81088041          733.26668233            2.95228145           -2.18671628            0.02876324
```

Figura 49. Output

Interpretare: Deoarece valoarea lui p este 0.02876324 < 0,05, testul este semnificativ și corespunde ipotezei de cercetare ceea ce înseamnă că cel puțin unul dintre predictori este semnificativ legat de variabila răspuns.

5. Evaluarea clasificării prin matricea de clasificare

Matricea de confuzie clasifică datele reale în funcție de datele prezise. Aceasta evaluează predicțiile făcute pe datele de testare, adică numărul de predicții corecte făcute, precum și predicțiile greșite făcute pe date.

```
> model <- glm(churn ~ credit_score+country+gender+age+tenure+
+               balance+products_number+credit_card+active_member+
+               estimated_salary, family="binomial", data=train)
> #use model to predict probability of default
> glm.probs <- predict(model, test, type="response")
> test$pred_glm = ifelse(glm.probs > 0.5, "1", "0")
> test$churn <- ifelse(test$churn=="da", 1, 0)
> test$pred_glm = as.factor(test$pred_glm)
> test$churn = as.factor(test$churn)
> levels(test$pred_glm)
[1] "0" "1"
> levels(test$churn)
[1] "0" "1"
> confusionMatrix(test$churn,test$pred_glm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2168  101
         1  320  256

               Accuracy : 0.852
                 95% CI : (0.8384, 0.8649)
    No Information Rate : 0.8745
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.466

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8714
            Specificity : 0.7171
         Pos Pred Value : 0.9555
         Neg Pred Value : 0.4444
             Prevalence : 0.8745
         Detection Rate : 0.7620
   Detection Prevalence : 0.7975
      Balanced Accuracy : 0.7942

       'Positive' Class : 0
```

Figura 50. Output matricea de confuzie

- True Positive (TP) – 2168 au fost clasificate în clasa corecta 0 (adică churn ”nu”)

- True Negative (TN) – 256 au fost clasificate în clasa corecta 1 (adică churn ”da”)

- Fals Pozitiv (FP) – 101 au fost clasificate în clasa greșita de 1 (adică churn ”da”)

- Fals Negativ (FN) – 320 au fost clasificate în clasa greșita de 0 (adică churn ”nu”)

Mai departe calculăm sensibilitatea, spusă și „rata adevărată pozitivă” – procentul de indivizi pe care modelul a prezis corect ar fi implicit.

```
> sensitivity(test$churn, test$pred_glm)
[1] 0.8713826
```

Figura 51. Output sensibilitate

Se observă ca 87% din date au fost clasificate adevărat pozitive (în clasa 0).

Mai departe calculăm specificitatea, spusă și „rata negativă adevărată” – procentul de indivizi pe care modelul a prezis corect nu ar fi implicit.

```
> specificity(test$churn, test$pred_glm)
[1] 0.7170868
```

Figura 52. Output specificitate

Se observă ca 71,70% din date au fost clasificate adevărat negative (în clasa 1).

```
> Accuracy = (2168 + 265) / (2168 + 320 + 265 + 101)
> Accuracy
[1] 0.8524877
> Error_rate = (101 + 320) / (2168 + 320 + 265 + 101)
> Error_rate
[1] 0.1475123
```

Figura 53. Output acurateţe şi rata de eroare

85,24% din date au fost clasificate corect în timp ce modelul a prezis greșit 14,75% din valori. Rata de eroare este de 14,75%. În general, cu cât această rată este mai mică, cu atât modelul este mai capabil să prezică rezultatele, astfel încât acest model particular se dovedește a fi foarte bun în a prezice dacă un individ va renunța la serviciile bancare sau nu.

6. Compararea celor două procedee de selectare a variabilelor- Coeficientul Mallows' Cq

Coeficientul Mallows' Cq este o măsurătoare care este utilizată pentru a alege cel mai bun model de regresie dintre mai multe modele potențiale.

Putem identifica „cel mai bun" model de regresie prin identificarea modelului cu cea mai mică valoare Cp care este aproape de p +1, unde p este numărul de variabile predictoare din model.

Cel mai simplu mod de a calcula Cp-ul lui Mallows în R este să utilizați funcția *ols_mallows_cp()* din pachetul *olsrr*.

```
> #### 6.Compararea celor două procedee de selectare a variabilelor- Coeficientul Mallows' Cq ########
> library(olsrr)
> #fit full model
> full_model <- lm(churn ~ ., data = train)
> # modelul din procedeului Purposeful
> model1 <- lm(churn ~ credit_score + age + tenure + balance + gender + products_number
+              + credit_card + country, data = train)
> # modelul din procedeului Stepwise
> model2 <- lm(churn ~ credit_score + estimated_salary + credit_card + tenure, data = train)
> #calculate Mallows' Cp for each model
> ols_mallows_cp(model1, full_model)
[1] 137.7331
> ols_mallows_cp(model2, full_model)
[1] 2666.216
```

Figura 54. Output Coeficientul Mallows' Cq

Model 1: p + 1 = 9, Mallows' Cp = 137.7331

Model 2: p + 1 = 5, Mallows' Cp = 2666.216

Putem vedea că modelul 1, cel din procedeul de selecție Purposeful, are o valoare pentru coeficientul Mallows' Cq care este cea mai apropiată de p + 1, ceea ce indică faptul că este cel mai bun model care duce la cea mai mică cantitate de părtinire dintre cele 2 modele potențiale.

7. Interpretarea modelului final

Pe baza coeficienților de regresie din modelul 1, cel din procedeul de selecție Purposeful, vedem că șansele ca un client să renunțe la serviciile bancare cresc odată cu vârsta clientului, numărul de produse de 3 și 4, țara de origine Germania și Spania, iar scad odată cu scorul de credite, vechimea la bancă, soldul, genul masculin al clientului, numărul de produse sa fie 2 și prezența unui credit card. Aceste detalii se pot observa pe baza semnului pozitiv sau negativ din fiecare coeficient de regresie.

În concluzie, am putea spune că cu cât clientul înaintează în vârstă, deține mai mult de 2 produse la bancă și țara de origine este Germania sau Spania, cu atât mai probabil clientul va renunța la serviciile bancare. Pe de altă parte, cu cât crește scorul de credite, anii de vechime la bancă, soldul, numărul de 2 produse, prezența unui credit card și genul să fie masculin, cu atât este mai puțin probabil să renunțe la serviciile bancare.

```
> ########## 7. Interpretarea modelului final ######################
> coef <- coef(model1)
> coef
     (Intercept)     credit_score              age           tenure          balance       genderMale
   -1.807627e-01    -7.098956e-05     1.385257e-02    -6.957535e-04    -2.224105e-07    -7.001915e-02
products_number2 products_number3 products_number4     credit_card1   countryGermany     countrySpain
   -1.814883e-01     4.861406e-01     6.100514e-01    -8.750577e-04     1.342311e-01     3.674906e-04
```

Figura 55. Output coeficienții de regresie pentru M1

În continuare, dorim să cunoaștem valoarea impactului fiecăreia dintre aceste variabile asupra variabilei *churn*, adică asupra pierderii clienților băncii. În primul rând, trebuie să ne amintim că regresia logistică a modelat variabila răspuns la log(odds) care Y = 1. Aceasta implică că coeficienții de regresie permit modificarea log(odds) în randamentul unei schimbări de unitate în variabila predictor, ținând toate alte variabile predictoare constante.

Deoarece log(odds) sunt greu de interpretat, îl vom transforma exponențiând rezultatul după cum urmează.

```
> exp(coef(model1))
     (Intercept)     credit_score              age           tenure          balance       genderMale
       0.8346334        0.9999290        1.0139490        0.9993045        0.9999998        0.9323760
products_number2 products_number3 products_number4     credit_card1   countryGermany     countrySpain
       0.8340280        1.6260286        1.8405260        0.9991253        1.1436571        1.0003676
```

Figura 56. Output exponenții coeficienților de regresie pentru M1

Observăm că șansele ca un client să renunțe la serviciile bancare sunt crescute cu un factor de 1,013 pentru o creștere de un an a vârstei (în timp ce celelalte variabile rămân constante). De asemenea, șansele ca un client să renunțe la serviciile bancare sunt crescute cu un factor de 1,626

pentru deținerea a 3 produse bancare (în timp ce celelalte variabile rămân constante), de 1,840 pentru deținerea a 4 produse bancare (în timp ce celelalte variabile rămân constante), de 1,143 pentru țara de origine Germania (în timp ce celelalte variabile rămân constante) și de 1,00 pentru țara de origine Spania (în timp ce celelalte variabile rămân constante).

Dimpotrivă, șansele ca un client să renunțe la serviciile bancare sunt înmulțite cu un factor de 0,99 pentru fiecare creștere a scorului de credit. Înseamnă că șansa ca un client să renunțe la serviciile bancare scade cu -1% de fiecare dată când cineva acumulează un scor de credit mai mare, cu -1% de fiecare dată când cineva înaintează în vechime ca client al băncii, cu -6,77% de fiecare dată când clientul este de genul masculin, cu -16,6% de fiecare dată când clientul deține 2 produse bancare, cu -1% de fiecare dată când cineva deține un credit bancar.