UNIVERSITATEA BABEŞ-BOLYAI

Facultatea de Ştiinţe Economice şi Gestiunea Afacerilor

Informatică Economică

Proiect Big Data

Predicția părăsirii serviciilor bancare

Student Ilieș Dragoș

Informatică economică – IF

An 3

2024

Introducere

Sistemul bancar reprezintă o componentă importantă a sistemului financiar al unei țări. În zilele noastre, mediul bancar a devenit unul din ce în ce mai competitiv, datorită apariției pe piață a multor jucători. Cei care au de câștigat de pe urma acestei competitivități ridicate sunt agenții economici și clienții, care dispun de o gamă variată de servicii bancare. Toate instituțiile bancare doresc să atragă cât mai mulți clienți. Uneori se practică și "furtul" clienților din alte bănci, prin propunerea unor oferte sau servicii mai avantajoase. Cu toate acestea, prioritatea băncilor ar trebui sa fie aceea de a-și menține loiali clienții existenți, mai ales într-un mediu în care opțiunile bancare sunt tot mai numeroase.

Fenomenul de churn (părăsirea serviciilor de către clienți) reprezintă o provocare cu care se confruntă majoritatea competitorilor de pe piață. Atragerea de noi clienți necesită, de multe ori, costuri mai mari decât păstrarea clienților existenți. Anticiparea și detectarea factorilor care influențează decizia unui client de a renunța la serviciile curente ajută băncile să implementeze strategii mai eficiente de retenție.

Scopul proiectului este de a evalua performanța diferitor metode de învățare automată în vederea prezicerii churn-ului în cadrul unei entități bancare. Analiza curentă va include mai multe metode de clasificare pentru a putea stabili factorii determinanți care contribuie la acest fenomen. De asemenea, prin crearea unor modele predictive se va urmări identificarea clienților cu o probabilitate mai mare de a renunța la serviciile bancare.

Prin acest studiu, ne propunem să răspundem la următoarele întrebări de cercetare:

1. Care sunt factorii cei mai importanți care influențează decizia unui client de a renunța la serviciile băncii
2. Ce metode de machine learning funcționează cel mai bine în vederea predicției churn-ului clienților, pentru un set de date dezechilibrat?
3. Cum poate balansarea setului de antrenament influența specificitatea și sensibilitatea modelelor de clasificare?

Obținerea unor răspunsuri concludente la aceste întrebări de cercetare este esențială pentru instituțiile bancare care se confruntă cu fenomenul de pierdere a clienților. În plus, prin depistarea factorilor care contribuie la decizia unui client de a renunța la serviciile bancare, se pot lua măsuri pentru a fideliza clientela.

În era digitală, tehnicile și algoritmii de învățare automată au devenit instrumente cruciale pentru analiza și predicția comportamentului clienților. Aceste metode vin în ajutorul instituțiilor bancare care pot procesa și analiza volume mari de date. Implementarea tehnologii menționate anterior oferă băncilor un avantaj competitiv deoarece pot anticipa nevoile clienților și pot răspunde cerințelor de pe piață.

Prin această analiză, băncile pot aplica strategii eficiente de retenție a clienților care contribuie la o creștere a economiei. Pe de altă parte, clienții beneficiază la rândul lor de servicii personalizate și o experiență bancară îmbunătățită. Nu în ultimul rând, pe termen lung, gestionarea relației client-bancă contribuie la stabilitatea și creșterea sectorului bancar.

Descrierea setului de date

În vederea realizării acestui proiect, am utilizat un set de date care conține informații legate de clienții unei bănci din U.S. Sursa setului de date poate fi accesată la acest link: https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction/data.

Din punct de vedere al structurii, setul de date este împărțit în 14 coloane, și anume: RowNumber (numărul înregistrării), CustomerId (Id-ul clientului), Surname (numele de familie), CreditScore, Geography (locația clientului), Gender (sexul clientului), Age (vârsta clientului), Tenure (de câți ani persoana respectivă este un client al băncii), Balance (balanța medie a clientului), NumOfProducts (numărul de produse sau servicii ale băncii pe care clientul le utilizează), HasCrCard (valoare 1 daca clientul are card la bancă sau 0 în caz contrar), IsActiveMember (valoarea 1 daca clientul este membru activ sau 0 în caz contrat) EstimatedSlary (salariul estimat la clientului) și Exited (are valoarea 1 daca clientul a părăsit serviciile băncii sau 0 în cazul în care nu a părăsit).

În realizarea modelelor predictive de clasificare vom folosi doar datele relevante pentru a analiza dacă un client decide să renunțe sau nu la serviciile băncii. Astfel, pentru a avea o privire de ansamblu asupra datelor, vom analiza întregul set de date.

```
> str(data)
'data.frame':   10000 obs. of  14 variables:
 $ RowNumber      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CustomerId     : int  15634602 15647311 15619304 15701354 15737888 15574012 15592531 15656148 15792365
5592389 ...
 $ Surname        : chr  "Hargrave" "Hill" "Onio" "Boni" ...
 $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
 $ Geography      : chr  "France" "Spain" "France" "France" ...
 $ Gender         : chr  "Female" "Female" "Female" "Female" ...
 $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
 $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
 $ Balance        : num  0 83808 159661 0 125511 ...
 $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
 $ HasCrCard      : int  1 0 1 0 1 1 1 1 0 1 ...
 $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
 $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
 $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

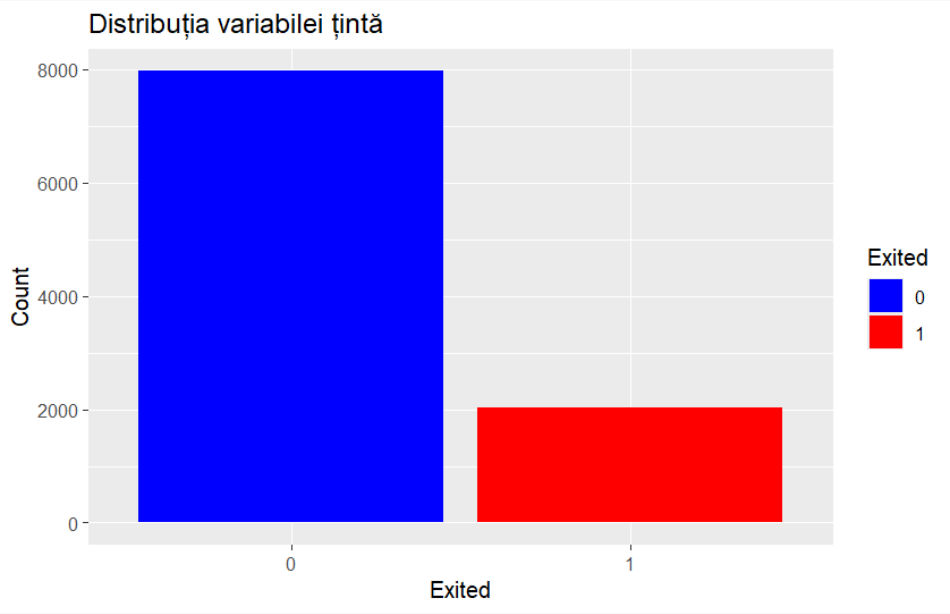
Din poza de mai sus, se poate desprinde faptul că setul de date conține 10.000 de observații cu 14 coloane diferite, atât cu date numerice, cât și cu date categorice. Având în vedere că ne confruntăm cu o problemă de clasificare, vom elimina coloanele irelevante și vom factoriza coloanele categorice pentru a simplifica procesul de analiză.

```
> str(data)
'data.frame':   10000 obs. of  11 variables:
 $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
 $ Geography      : Factor w/ 3 levels "France","Germany",..: 1 3 1 1 3 3 1 2 1 1 ...
 $ Gender         : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 2 1 2 2 ...
 $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
 $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
 $ Balance        : num  0 83808 159661 0 125511 ...
 $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
 $ HasCrCard      : int  1 0 1 0 1 1 1 1 0 1 ...
 $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
 $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
 $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

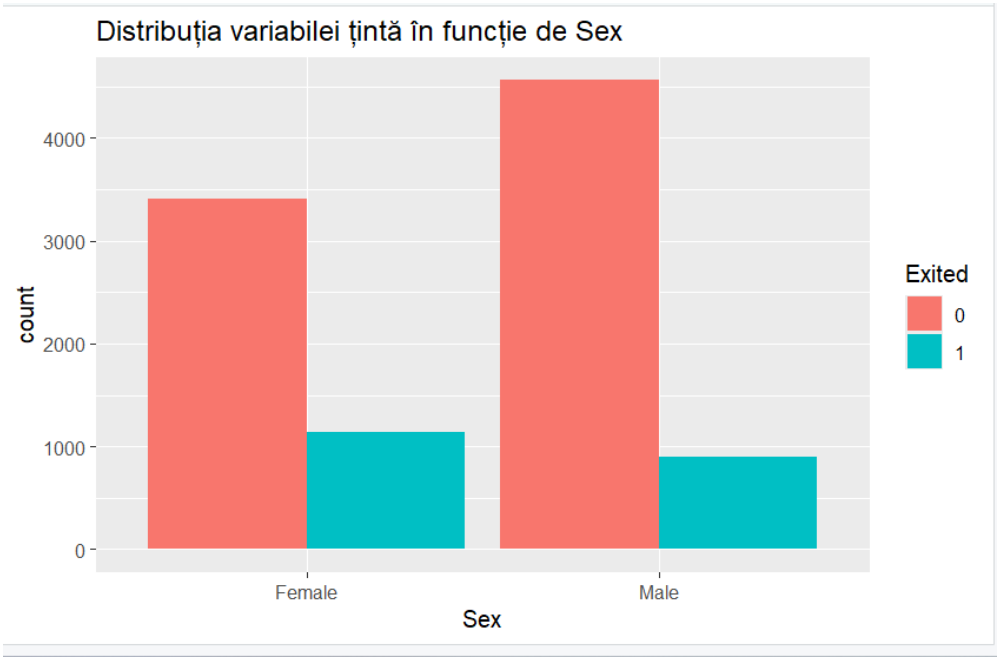
Prin procesul de curățare și procesare a datelor, am eliminat coloanele: RowNumber, CustomerId și Surname. De asemenea, am factorizat coloanele Geography și Gender, pentru care avem 3 nivele (France, Germany și Spain), respectiv 2 nivele (Male și Female).

Construirea unor grafice relevante pentru variabilele din setul de date, ne poate ajuta să realizăm câteva observații preliminare.

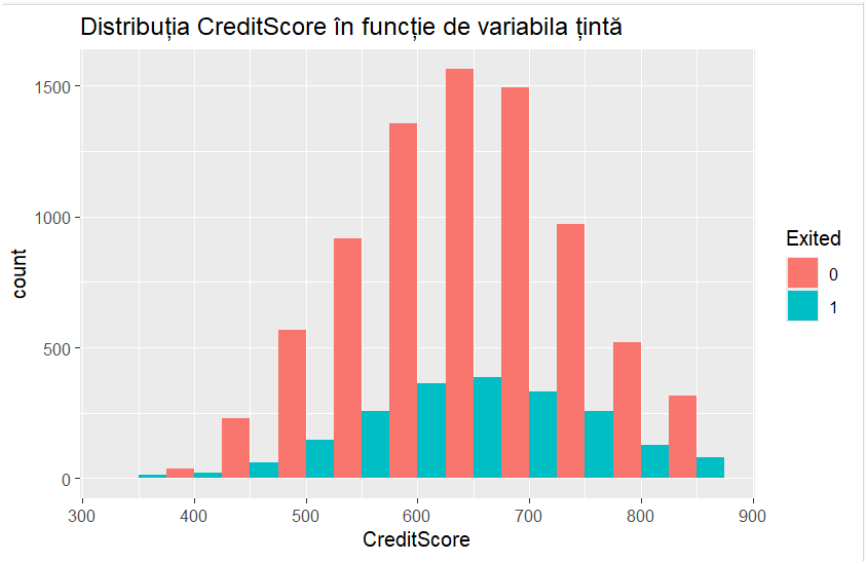
Din graficul de mai jos se poate identifica distribuția valorilor pentru variabila țintă (Exited). Din numărul total de observații (10.000), 8.000 de clienți nu au renunțat la serviciile bancare, în timp ce 2.000 de clienți au părăsit aceste servicii. Setul de date este destul de dezechilibrate, clienții care renunță la serviciile bancare reprezintă 20% din numărul total.



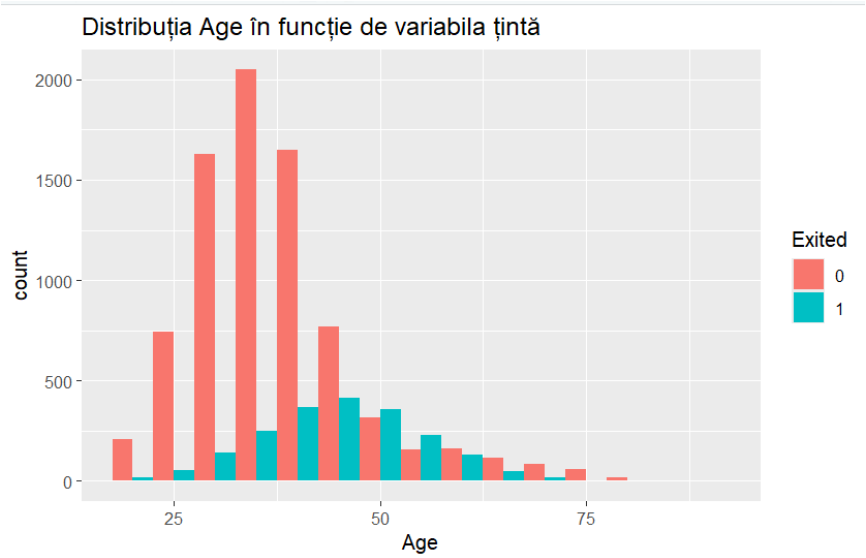
Următorul grafic pune în lumină distribuția variabilei dependente în funcție de sexul clienților. Cu toate că procentajul persoanelor de sex feminin este mai mic (45%) comparativ cu cel al bărbaților, se poate observa că femeile au o probabilitate puțin mai mare de a părăsi serviciile bancare.



Graficul care evidențiază distribuția variabilei CreditScore în funcție de variabila țintă sugerează anumite informații relevante analizei viitoare. Se poate observa că între valorile 550 și 750 a coloanei CreditScore există o creștere a ambelor clase de clienți. Înafara acestui interval, proporțiile clienților care părăsesc serviciile bancare încep să scadă. Prin urmare, instituțiile ar putea să se concentreze pe dezvoltarea unor strategii de retenție pentru persoanele care se încadrează în intervalul specificat mai sus, deoarece aceștia par cei mai predispuși fenomenului churn.



Nu în ultimul rând, vârsta clienților jucă un rol important în decizia acestora de a renunța la serviciile bancare de care dispun momentan. După cum se poate observa și în graficul de mai jos, clienții cu vârste între 20 și 30 de ani sunt mai fideli băncii, în timp ce persoanele cu vârsta între 40 și 60 de ani sunt mai sunt mai predispuși să renunțe șa serviciile bancare. O eventuală nemulțumire sau necesitatea de servicii bancare mai personalizate în această etapă a vieții contribuie la decizia acestor persoane de a renunța la serviciile actuale.



Rezultate și discuții

Mediul de lucrul în care s-a efectuat analiza fenomenului churn aspra băncilor a fost RStudio. Înainte de a aplica diferitele metode de clasificare, primul pas a fost reprezentat de împărțirea setului de date într-un set de antrenament cu o proporție de 70% din date și un set de test cu o proporție de 30% din date. De asemenea, prin folosirea construcției “strata = Exited”, am asigurat păstrarea proporțiilor pentru variabila dependentă în ambele seturi create. O mențiune importantă este aceea că pentru matricea de confuzie, clasa pozitivă va fi clasa “No”. Astfel, în procesul de analiză ne vom concentra pe îmbunătățirea valorii pentru specificity. Ne dorim să avem o valoare cât mai bună și pentru curba AUC-ROC. Cu toate că setul de date este destul de dezechilibrat, ne va interesa și obținerea unei valori decente pentru sensitivity

Primul model pe care l-am conceput a fost realizat prin metoda regresie logistice. Deși sensitivity-ul, care arată clienții pe care modelul i-a identificat că nu părăsesc banca, este foarte mare (96%), specificity-ul modelului este doar 22%. Reușim să identificăm corect doar 138 clienți care au părăsit serviciile bancare. Avem un număr destul de mare al testelor de fals pozitive (474) pentru clienții care au părăsit banca. Accuracy-ul este destul de bun, însă nu este foarte relevant pentru un set de date dezechilibrat. Pentru acest model, valoarea lui AUC este de 0.7683.

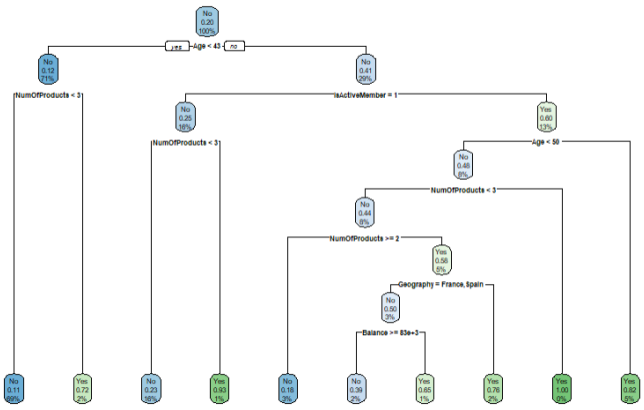
```
          Reference
Prediction   No  Yes
       No  2311  474
       Yes   78  138

               Accuracy : 0.8161
                 95% CI : (0.8017, 0.8298)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : 0.003214

                  Kappa : 0.254

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9674
            Specificity : 0.2255
```

În continuare, am realizat un nou model de predicție prin folosirea metodei Naive Bayes. După cum se poate observa din imaginea de mai jos, atât accuracy-ul (83%) cât și sensitivity-ul (97%) au crescut față de modelul anterior, dar nesemnificativ. Cu toata acestea, indicatorul cel mai important pentru analiza noastră este specificity, iar valoarea acestuia este în continuare foarte mică, doar 28%. Și aici, reușim să identificăm doar 177 de clienți care au părăsit serviciile bancare. Rezultatele sunt în continuare foarte slabe, astfel că vom continua să abordăm alte metode de analiză. Aici AUC are valoarea 0.8141.

```
          Reference
Prediction   No  Yes
       No  2335  435
       Yes   54  177

               Accuracy : 0.8371
                 95% CI : (0.8233, 0.8501)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : 6.068e-09

                  Kappa : 0.3469

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9774
            Specificity : 0.2892
```

Ambele metode pe care le-am analizat până acum avut anumite limitări din punct de vedere al rezultatelor. Regresia logistică și Naive Bayes sunt modele simple care nu garantează capturarea complexității relațiilor dintre variabilele independente și variabila țintă. Un alt factor care poate influența aceste rezultate este reprezentat de dezechilibrul setului de date. În încercarea de a obține niște rezultate cât mai concludente vom aplica metoda arborilor de decizie. Spre deosebire de regresia logistică, arborii de decizie pot captura relațiile non-liniare dintre variabile dependentă și celelalte caracteristici. În plus, arborii de decizie selectează automat variabilele independente cele mai semnificative în procesul de împărțire. Primul arbore pe care l-am creat (m1) este un arbore de decizie basic (cu rpart), fără vreo ajustare. Poza din stânga reprezintă afișarea arborelui sub forma de text, iar în dreapta sub forma grafică. Se poate observa faptul că primul criteriu după care s-a făcut împărțirea a fost Age(43), pe când următoarele împărțiri au pus accent pe nodurile: NumOfProducts, IsActiveMember, Geography și Balance.

```
> m1
n= 6999

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 6999 1425 No (0.79639949 0.20360051)
    2) Age< 42.5 4968  596 No (0.88003221 0.11996779)
      4) NumOfProducts< 2.5 4849  510 No (0.89482367 0.10517633) *
      5) NumOfProducts>=2.5 119   33 Yes (0.27731092 0.72268908) *
    3) Age>=42.5 2031  829 No (0.59182669 0.40817331)
      6) IsActiveMember>=0.5 1130  285 No (0.74778761 0.25221239)
       12) NumOfProducts< 2.5 1087  245 No (0.77460902 0.22539098) *
       13) NumOfProducts>=2.5 43    3 Yes (0.06976744 0.93023256) *
      7) IsActiveMember< 0.5 901  357 Yes (0.39622642 0.60377358)
       14) Age< 49.5 565  269 No (0.52389381 0.47610619)
         28) NumOfProducts< 2.5 531  235 No (0.55743879 0.44256121)
           56) NumOfProducts>=1.5 181   32 No (0.82320442 0.17679558) *
           57) NumOfProducts< 1.5 350  147 Yes (0.42000000 0.58000000)
            114) Geography=France,Spain 244  122 No (0.50000000 0.50000000)
              228) Balance>=82841.02 142   56 No (0.60563380 0.39436620) *
              229) Balance< 82841.02 102   36 Yes (0.35294118 0.64705882) *
            115) Geography=Germany 106   25 Yes (0.23584906 0.76415094) *
         29) NumOfProducts>=2.5 34    0 Yes (0.00000000 1.00000000) *
       15) Age>=49.5 336   61 Yes (0.18154762 0.81845238) *
```



Cu toate acestea, rezultate predicției pentru acest arbore de decizie au fost sub așteptări. Am obținut o valoare de doar 41% pentru specificity. Din nou avem un raport mai mare al testelor fals pozitive de cât a celor true pozitive ceea ce este îngrijorător. Valoarea AUC pentru acest arbore este de: 0.7686. Am încercat să aplic metoda de prunning (cu cp = 0.01) pe acest arbore în speranța obținerii unor rezultate mai bune, însă am primit aceleași rezultate.

```
          Reference
Prediction   No  Yes
       No  2330  355
       Yes   59  257

               Accuracy : 0.862
                 95% CI : (0.8492, 0.8742)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4819

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9753
            Specificity : 0.4199
```

Următorul model pe care l-am realizat a fost tot un arbore de decizie cu rpart (m2), dar am setat acest arbore ca fiind unul fără restricții (cp=0). Deși vizualizarea grafică pentru acest arbore este destul de anevoioasă din cauza numărului foarte mare de noduri, rezultatele acestui model din punct de vedere al specificity-ului au fost mai bune decât cele ale arborelui precedent. Aici AUC a avut valoarea: 0.8307. Pentru acest arbore (m2) am încercat și prunning cu cp=0.02, dar rezultatele pentru specificity au fost mai slabe.

```
          Reference
Prediction   No  Yes
       No  2211  283
       Yes  178  329

               Accuracy : 0.8464
                 95% CI : (0.833, 0.8591)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : 8.944e-13

                  Kappa : 0.4946

 Mcnemar's Test P-Value : 1.274e-06

            Sensitivity : 0.9255
            Specificity : 0.5376
```

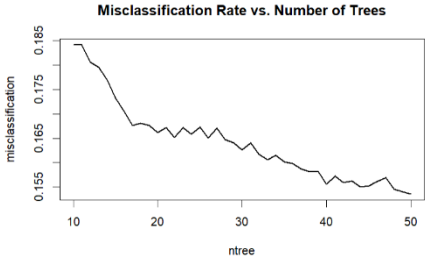
Ca o altă abordare pentru analiza noastră, am realizat doua modele de arbori de decizie puțini mai diferiți de cei cu rpart. Am folosit librăria tree și am creat un arbore de decizie basic și un arbore de decizie cu index Gini. Indexul Gini este o măsura utilizată pentru a evalua împărțirile efectuate pe setul de antrenament. Acesta măsoară "impuritatea" unui set de date, unde 0 reprezintă puritatea perfectă (adică toate elementele aparțin unei singure clase), iar valorile mai mari reprezintă o impuritate mai mare. În poza din stânga se pot observa rezultatele pentru arborele basic, iar în dreapta pentru cel cu indexul Gini. Arborele cu Gini are o valoare mai bună pentru specificity decât celălalt, dar ambele modele de predicție au furnizat rezultate foarte proaste. Pentru aceste cazuri, AUC a avut valoarea 0.8314 (basic), respectiv 0.7577 (Gini).

```
          Reference
Prediction   No  Yes
       No  2354  432
       Yes   35  180

              Accuracy : 0.8444
                95% CI : (0.8309, 0.8572)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : 6.97e-12

                 Kappa : 0.3683

 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.9853
           Specificity : 0.2941
```

```
          Reference
Prediction   No  Yes
       No  2191  330
       Yes  198  282

              Accuracy : 0.8241
                95% CI : (0.81, 0.8375)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : 5.982e-05

                 Kappa : 0.4109

 Mcnemar's Test P-Value : 1.191e-08

           Sensitivity : 0.9171
           Specificity : 0.4608
```

Rezultatele mult sub așteptări primite în urma implementării, antrenării și testării modelelor de arborilor de decizie prezentați mai sus, sunt cauzate în principal de overfitting (supraînvățare). Aceste modele tind să se potrivească prea bine pe setul de antrenament și învață inclusiv zgomotul datelor de antrenament. Însă, când sunt executate testele, performanța este una scăzută deoarece modelul nu reușește să generalizeze suficient de bine. Pentru a încerca să obținem rezultate mai bune, vom aplica metode avansate a arborilor de decizie precum: Bagging și Random Forest. Astfel, următorul model creat a fost realizat cu bagging. Metoda bagging combină mai multe modele de arbori de decizie și face o medie a acestora cu scopul de a reduce variabilitatea si overfittingul. În mod implicit, bagging folosește 25 de bags. Din nou, rezultatele obținute sunt destul de slabe, avem un specificity puțin peste 50%, iar numărul de teste fals pozitive este în continuare ridicat. Pentru acest model basic de bagging, AUC a avut valoarea: 0.8362. Am analizat graficul pentru rata de misclasificare în raport cu numărul de abori. După cum se poate vedea, rata de misclasificare scade semnificativ cu cât numărul de arbori crește. Nu prea se poate observa un punct exact în care rata s-ar stabiliza, deoarece pare să aibă o scădere continuă. Cu toate acestea, voi încerca să optimizez acest model de bagging prin încercarea mai multor valori pentru numărul de arbori, astfel încât să obțin un model cât mai stabil.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2265  303
       Yes  124  309

              Accuracy : 0.8577
                95% CI : (0.8447, 0.87)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 0.5083

 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.9481
           Specificity : 0.5049
```



După mai multe încercări, am obținut cele mai bune rezultate pentru un bagging cu 32 de arbori. Diferența dintre acești doi arbori pentru care am folosit metoda bagging este nesemnificativă. Toate metricele au aproximativ aceleași valori. Deși metoda de bagging a reușit să aibă valori puțin mai bune pentru specificity față de arborii de decizie anteriori (exceptând arborele m2 fără restricții), tot există un număr mare de predicții fals pozitive. O limitare a acestei abordări ar fi dată de arborii individuali de decizie care pot avea un bias mare. Acest bias se datorează lipsei de profunzime și complexitate, care este transferat apoi în modelul final de bagging. Nici această abordare nu a adus rezultatele așteptate, astfel că următorul pas este aplicarea metodei Random Forest.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2264  302
       Yes  125  310

               Accuracy : 0.8577
                 95% CI : (0.8447, 0.87)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.509

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9477
            Specificity : 0.5065
```

Modelul de predicție realizat prin metoda random forest a avut rezultate mai slabe decât modelele anterioare realizate cu bagging. Valorea specificity-ului este doar 47%, ceea ce înseamnă ca acest model nu reușește să identifice nici măcar jumate din clienții care vor părăsi serviciile bancare. Deși sensitivity-ul are o valoare destul de promițătoare, avem mult prea multe teste fals pozitive ceea ce este îngrijorător pentru analiza noastră. Astfel, pentru modelul basic random forest basic, am încercat să folosesc o metodă de optimizare pentru obținerea unor rezultate mai avantajoase. Prin acestă metodă de tunning random forest, am definit o grilă de hiperparametri și am antrenat modelul cu ajutorul celor mai buni hiperparametri găsiți. Această metodă crește performanța modelului și se asigură că nu este prea simplu sau prea complex. Însă, nici acest model optimizat nu ne-a ajutat să obținem rezultate mai bune, din această cauză am decis să nu mai prezint individual rezultatele, ci doar într-un tabel final pentru a se putea observa mai bine.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2316  324
       Yes   73  288

               Accuracy : 0.8677
                 95% CI : (0.8551, 0.8796)
    No Information Rate : 0.7961
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5192

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9694
            Specificity : 0.4706
```

Tabel pentru compararea rezultatelor pentru modelele de predicție

Model	Specificity	Sensitivity	AUC-ROC	P-Value
Regresie logistică	0,2255	0,9674	0,7683	0.003214
Naive Bayes	0,2892	0,9774	0,8141	6.068e-09
Arbori de decizie (m1) basic (cu rpart)	0,4199	0,9753	0,7686	< 2.2e-16
Arbori de decizie m1, cu cp=0.01	0,4199	0,9753	0,7686	< 2.2e-16
Arbori de decizie (m2) fără restricții, cp=0 (cu rpart)	**0,5376**	**0,9255**	**0,8307**	**8.944e-13**
Arbori de decizie m2, cu cp=0.02	0,4444	0,9607	0,7679	< 2.2e-16
Arbori de decizie basic (cu tree)	0,2941	0,9853	0,8134	6.97e-12
Arbori de decizie cu index Gini	0,4608	0,9171	0,7577	5.982e-05
Model basic bagging	0,5049	0,9481	0,8362	< 2.2e-16
Model bagging cu numărul de optim de baggs	0,5065	0,9477	0,8405	< 2.2e-16
Model basic random forest	0,4706	0,9694	0,8663	< 2.2e-16
Tunning random forest	0,4657	0,9644	0,8582	< 2.2e-16

După cum se poate observa din tabelul de mai sus, sensitivity-ul modelelor pe care le-am realizat este unul destul de bun, avem în general o valoare peste 96% cea ce este foarte bine. Prin acest parametru reușim să identificăm acei clienți care rămân fideli băncii și nu părăsesc serviciile instituției. Cu toate acestea, analiza noastră se concentrează mai mult pe identificarea acelor clienți care decid să renunțe la serviciile bancare actuale, precum și cauzele din spatele deciziei. Din păcate, modelele pe care le-am realizat reușesc să identifice, într-o proporție destul de mică, clienții care părăsesc serviciile bancare. Specificity-ul cel mai mare pentru modelele de mai sus a fost 53%, un procentaj destul de mic. Aceasta valoarea a fost obținută pentru modelul de arbore de decizie m2 fără restricții (cp=0). Sensitivity-ul pentru acest model este puțin mai scăzut decât a altor modele, însă depășește pragul de 90% (92,55%), ceea ce este îmbucurător.

Din cauza faptului că modelele de predicție de mai sus nu au reușit să atingă un procentaj de 60-70% pentru specificity, am încercat să balansăm setul pe care sunt antrenate modelele. Prin această tehnică, ne dorim să avem în setul de antrenament o proporție egală pentru cele două instanțe ale coloanei Exited (No sau Yes) pentru a se evita overfitting-ul (supraînvățarea modelelor). Procedeul de balansare a setului de antrenament ne va ajuta să antrenăm mai bine modelele și să obținem rezultate mai bune pe setul de test. Principalul factor pe care ne axăm este creșterea specificity-ului, adică să reușim să identificăm corect cât mai mulți clienți care părăsesc banca și să reducem numărul de teste fals pozitive.

După ce am folosit libraria ROSE pentru a balansa setul de antrenament, distribuția coloanei Exited în set arată astfel:

```
> # Verificarea distribuției
> table(train_balanced$Exited)

  No  Yes
3542 3457
```

Acum că setul de antrenament a fost balansat, voi prezenta într-un tabel rezultatele modelelor de predicție pe care le-am realizat pe noul set de antrenament, iar mai apoi le-am testat.

Model	Specificity	Sensitivity	AUC-ROC	P-Value
Regresie logistică	0,6912	0,7380	0,7744	1
Naive Bayes	0,6520	0,8054	0,81	0.9986
Arbori de decizie (m1) basic (cu rpart)	0,7190	0,81	0,8165	0.7452
Arbori de decizie m1, cp=0.01	0,7190	0,81	0,8165	0.7452
Arbori de decizie (m2) fără restricții, cp=0 (cu rpart)	0,6503	0,7823	0,8307	1
Arbori de decizie m2, cp=0.02	**0,7565**	0,7417	0,7910	**1**
Model basic bagging	0,6095	0,8907	0,8430	1.158e-07
Model basic random forest	**0,6291**	**0,9041**	**0,8587**	**1.508e-13**

Modele de predicție prezentate mai sus au fost antrenate pe noul set de antrenament în care am balansat valorile variabilei dependente. Din punct de vedere al specificity-ului, aceste modele au obținut rezultate foarte bune. Cea mai mare valoare a specitifity-ului, mai exact 75%, am obținut-o prin metoda arborilor de decizie, în care am aplicat prunning cu cp=0.02. Acest parametru cp (complexity parameter) este utilizat pentru a realiza tăierea arborilor de decizie. Practic, cp elimină ramurile care nu aduc îmbunătățiri semnificative în performanța

modelului. Cu toate acestea, parametru p-value este o metrică foarte importantă care ne arată existența unei diferențe semnificative între acuratețea modelului și NIR. În acest caz, avem o valoare foarte mare pentru p-value, ce indică faptul că modelul nostru nu este semnificativ mai bun decât un model aleatoriu din punct de vedere al acurateței. Deși, modelul de arbori de decizie prunned cu cp=0.02 are cea mai mare valoare a specificity-ului, semnificația statistică redusă indicată de p-value sugerează că modelul nu este suficient de robust.

Astfel, modelul ales ca fiind cel mai bun pentru analiza noastră este modelul random forest. Acest model are un p-value foarte mic ceea ce indică performanța statistică ridicată. Deși modelul random forest nu a obținut cea mai mare valoare pentru specificity, oferă un echilibru între specitifity și sensitivity. De asemenea, avem și cea mai mare valoare pentru AUC-ROC. Este important ca modelele să fie semnificativ mai bune decât un model aleatoriu în termeni de acuratețe.

În general se poate observa o scădere a sensitivity-ului pentru modelele antrenate pe setul balansat, față de ce antrenate pe setul inițial. Motivul pentru care modelele antrenate pe setul inițial au valori așa mari pentru sensitivity este bias-ul pentru clasa majoritară. Din cauza faptului că în setul inițial majoritatea clienților nu părăsesc serviciile bancare, modelele sunt predispuse să fie foarte bune în a identifica acei clienți. Astfel, deoarece clasa majoritară este formată din clienții care nu părăsesc serviciile bancare, vom avea valori mari pentru sensitivity.

Prin balansarea setului de antrenament, am reușit să formăm niște modele mai echilibrate din punct de vedere al predicțiilor pentru ambele clase. Se poate observa că specificity-ul a crescut considerabil pentru aceste modele. Cu ajutorul setului de date balansat am identificat mai bine valorile pentru clasa minoritara (clienții care părăsesc serviciile bancare). Acest lucru implică evident și un compromis, deoarece am îmbunătățit modul de identificare al clienților care părăsesc serviciile bancare, însă s-a redus procentul de identificare corectă al clienților care nu părăsesc serviciile bancare.

După cum am menționat încă de la început, scopul analizei noastre a fost să identificăm corect, prin metodele predictive, cât mai mulți clienți care vor renunța la serviciile bancare. În urma acestei analize a modelelor de predicție, antrenate atât pe setul inițial, cât și pe setul balansat, putem oferi răspunsuri elocvente întrebărilor de cercetare. Astfel:

1. Arborii de decizie (m1 și m2 la care i-am făcut prunning cu cp=0.2) ne-au ajutat să identificăm factorii care influențează decizia unui client de a părăsi serviciile bancare. Acești arbori i-am antrenat atât pe setul de antrenament inițial, dar și pe cel balansat. Principalii factori care contribuie la decizia de renunțare a serviciile bancare sunt: vârsta (Age), numărul de produse utilizate (NumOfProducts), calitatea de membru activ (IsActiveMember), locația (Geography) și balanța medie (Balance). În toți arborii pe care i-am analizat, vârsta este primul criteriu după care se face împărțirea arborilor. Am putut observa că, în general, clienții cu vârsta sub 43 de ani rămân fideli serviciilor bancare, în timp ce clienții trecuți de această vârstă sunt mai predispuși să renunțe. De asemenea, clienții care dețin mai puțin de 3 produse au probabilitate mai mare să părăsească serviciile, iar clienții din Germania sunt mai predispuși decât cei din Spania sau Franța.

2. În prima fază, am făcut o analiză în care am păstrat proporția valorilor pentru variabila dependentă, atât în setul de antrenament, cât și în cel de test. Din cauza faptului că setul este unul dezechilibrat din punct de vedere al valorilor pentru variabila dependentă, modelele de predicție au funcționat mai bine pentru identificarea corectă a valorilor din clasa majoritară. Totuși, analiza noastră s-a bazat pe identificarea corectă valorilor din clasa minoritară. În acest sens, metoda care a funcționat cel mai bine a fost metoda arborilor de decizie fără restricții (cp=0). Acest model a avut cea mai mare valoare pentru parametrul specificity, mai exact 53%. Cu ajutorul arborelui fără restricții a obținut cele mai bune rezultate deoarece este flexibil și a permis modelului să învețe toate caracteristicile din setul de antrenament, inclusiv cele mai subtile diferențe dintre clase. Deși acest tip de arbori este predispus overfitting-ului dacă setul de date este unul dezechilibrat, pentru acest caz overfitting-ul e posibil să aibă avut un efect pozitiv. Dat fiind faptul că exemplele negative erau mult în majoritare, overfitting-ul a permis modelului să învețe foarte bine acest exemple negative și să reducă erorile de clasificare pentru clienții care nu părăsesc serviciile bancare.

3. După cum se poate observa din al doilea, balansarea setului de antrenament pentru variabila dependentă are un impact semnificativ asupra metricelor studiate. Prin intermediul acestui procedeu, am reușit să creștem considerabil valoarea specificity-ului. Aceasta a fost și scopul principal al analizei, de a identica corect cât mai mulți clienți care părăsesc serviciile bancare. Însă, odată cu balansarea setului și creșterea specificității, sensibilitatea modelelor a scăzut față de valorile anterioare. Acest fapt se datorează reducerii bias-ului pentru clasa majoritară (clienții care nu părăsesc serviciile). Ideea de a avea valori mai mici pentru sensitivity decât cele anterioare în detrimentul valorilor mai mari pentru specificity este un compromis asumat și comun în probleme de clasificare.

Concluzii

În urma analizei prezentate, am reușit să explorez în detaliu fenomenul de churn în cadrul unei bănci. Am folosit diverse tehnici de machine learning. Dintre acestea cele mai bune rezultate le-am obținut prin utilizarea arborilor de decizie. Mai exact, pentru cazul în care setul de antrenament nu era balansat, rezultatele cele mai bune pentru predicția dorită le-a avut arborele m2 (fără restricții). Pentru cazul în care setul de antrenament a fost balasat, rezultatele cele mai bune le-a obținut modelul basic random forest.

În final, am demonstrat importanța balansării setului de date și impactul acesteia asupra performanței modelelor. De asemenea, am identificat factorii cheie care influențează decizia clienților de a părăsi banca. Acești factori oferă băncilor informații valoroase pentru dezvoltarea unor strategii eficiente de retenție.