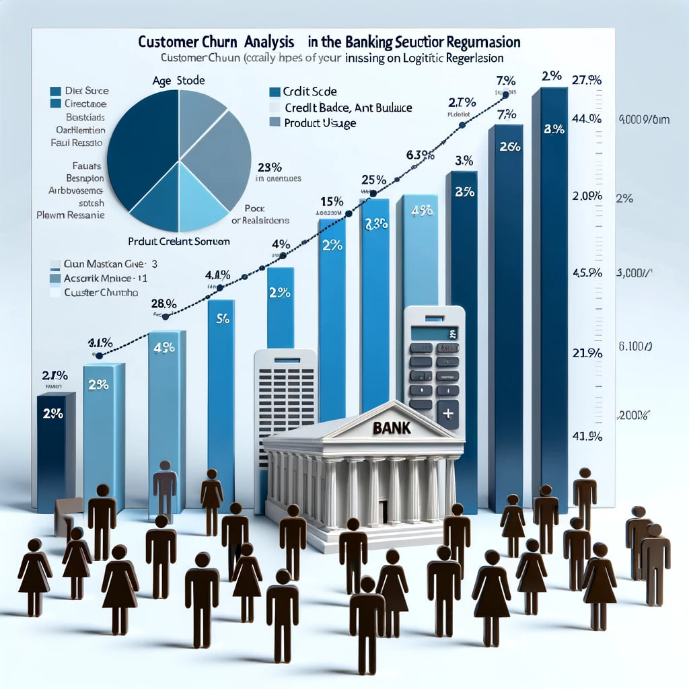
**Churn Prediction în activitățile bancare** Golban Elena 1

1 UTM, FCIM, DIIS

**ABSTRACT**

În contextul competitivității crescute din sectorul bancar, identificarea și retenția clienților devine o provocare esențială pentru sustenabilitatea pe termen lung a instituțiilor financiare. Acest studiu utilizează regresia logistică pentru a analiza factorii care influențează probabilitatea churn-ului clienților unei bănci, folosind un set de date public disponibil pe Kaggle. Prin selecția atentă a caracteristicilor, inclusiv scorul de credit, vârsta, balanța contului și numărul de produse utilizate, am construit un model predictiv care evidențiază variabilele cu cel mai mare impact asupra deciziei clienților de a părăsi banca. Rezultatele noastre indică o corelație pozitivă semnificativă între vârstă și probabilitatea de churn, în timp ce utilizarea unui număr mai mare de produse bancare și statutul de membru activ sunt asociate cu o probabilitate scăzută de churn. Aceste descoperiri subliniază importanța unei gestionări personalizate a relațiilor cu clienții și a unei abordări centrate pe client în dezvoltarea de strategii de retenție. Recomandările incluse în acest articol sunt destinate să ghidizeze băncile în optimizarea serviciilor lor pentru a reduce rata de churn și pentru a crește loialitatea clienților.

**INTRODUCERE**

Într-o eră în care datele joacă un rol central în procesele decizionale, industria bancară nu este o excepție de la această paradigmă analitică. Conform celor mai recente studii, churn-ul clienților reprezintă una dintre cele mai pregnante provocări cu care se confruntă băncile la nivel global. Este recunoscut că păstrarea unui client existent este mult mai cost-eficientă decât atragerea unuia nou. Anual, un număr semnificativ de clienți optează să părăsească instituțiile bancare, un fenomen care nu numai că erodează baza de clienți, dar și afectează profiturile și imaginea băncii. Astfel, înțelegerea factorilor care contribuie la churn devine esențială pentru formarea strategiilor de retenție a clienților.

Cercetarea noastră își propune să examineze un set de date cuprinzător, obținut de pe platforma Kaggle, constând în diverse caracteristici ale clienților unei bănci. Prin aplicarea unor tehnici avansate de prelucrare a datelor, analiză exploratorie și modele de regresie logistică, ne-am propus să identificăm principalii factori care influențează decizia clienților de a părăsi banca. Acest studiu nu se limitează doar la examinarea datelor colectate, ci include și crearea și optimizarea unui model predictiv capabil să evalueze riscul de churn al clienților.

Pornind de la premise inițiale bine fundamentate, studiul explorează întrebări precum corelația dintre vârsta clienților și probabilitatea acestora de a părăsi banca, influența nivelului de activitate a clienților și posesia cardurilor de credit asupra loialității acestora, și rolul salariului estimat în decizia de churn. Analizăm, de asemenea, impactul pe care îl au factori precum balanța contului și numărul de produse bancare folosite asupra stabilității relației dintre client și bancă.

Prin metode de prelucrare a datelor am asigurat o bază solidă pentru analiza noastră, eliminând valori lipsă și codificând corespunzător variabilele categorice, astfel încât să putem realiza o analiză cât mai apropiată de realitatea complexă a comportamentului clienților băncii. În final, rezultatele acestui studiu nu doar că oferă o perspectivă aprofundată asupra factorilor de churn, dar stabilesc și o fundație pentru modelarea predictivă viitoare, care să ghideze intervențiile timpurii și strategiile de prevenire a churn-ului în sectorul bancar.

**Contextul Industriei Bancare:**

Într-o piață financiară dinamică și extrem de competitivă, băncile sunt mereu în căutarea metodelor eficiente pentru a-și păstra clienții și a reduce rata de churn, un termen care se referă la pierderea clienților. Churn-ul poate avea un impact semnificativ asupra profitabilității, deoarece costurile de atragere a unui nou client sunt semnificativ mai mari decât cele de menținere a unuia existent.

**Obiectivele Articolului:**

Principalul obiectiv al acestui articol este de a identifica și analiza factorii care contribuie la churn-ul clienților. Utilizând metode de analiză exploratorie a datelor (EDA), ne propunem să evidențiem tendințele semnificative și să dezvoltăm recomandări strategice pentru managementul relațiilor cu clienții, cu scopul de a îmbunătăți retenția clienților în activitățile bancare.

**Premise Inițiale**

În cursul elaborării unui model predictiv pentru churn-ul clienților în domeniul bancar, ne-am ghidat cercetarea prin mai multe întrebări cheie, fiecare dintre acestea ancorându-ne analiza în realitățile comportamentale și financiare ale clienților:

1. **Vârsta și Churn-ul:** Oare clienții de diferite vârste sunt mai predispuși să părăsească banca? Ce legătură există între maturitatea financiară și loialitatea față de instituția bancară?
2. **Impactul Diversității Produselor:** Cum afectează numărul de produse bancare utilizate probabilitatea de churn? Este un portofoliu variat un semn al angajamentului față de bancă?
3. **Utilizarea Cardului de Credit și Churn-ul:** În ce măsură utilizarea cardului de credit reflectă satisfacția și activitatea clienților? Este aceasta un indicator al loialității?
4. **Balanța Contului ca Predictor:** Are balanța contului o influență asupra deciziei de churn? Reflectă acesta stabilitatea financiară a unui client?
5. **Salariul Estimat și Relația cu Banca:** Este salariul estimat un indicator relevant pentru churn sau este un factor mai puțin semnificativ în contextul bancar?
6. **Impactul Statutului de Membru Activ:** Este frecvența interacțiunilor cu banca un semn al unei relații solide? Cum se corelează activitatea membrilor cu rata de churn?
7. **Geografia ca Factor Determinant:** Există diferențe în rata de churn între diferitele regiuni geografice? Cum influențează contextul local comportamentul clienților?
8. **Scoringul de Credit și Fidelitatea:** În ce măsură scorul de credit al unui client poate prezice tendința acestuia de a părăsi banca?

**Materiale și Metode**

Studiul nostru se bazează pe un set de date reprezentativ pentru clienții unei instituții bancare, set de date care este public disponibil pe platforma Kaggle. Acesta include 10.000 de înregistrări ale clienților băncii, fiecare înregistrare având 14 atribute, cum ar fi ID-ul clientului, scorul de credit, geografia, genul, vârsta, tenure (durata relației cu banca), balanța contului, numărul de produse bancare, deținerea unui card de credit, statutul de membru activ, salariul estimat și variabila țintă, 'Exited', care indică dacă clientul a părăsit banca.

**Prelucrarea Datelor:**

Datele au fost inițial curățate pentru a elimina caracteristicile irelevante, cum ar fi ID-ul clientului și numele, care nu au valoare predictivă. Variabilele categorice au fost transformate prin codificare one-hot pentru a facilita procesarea lor în modelele de regresie logistică. Datele numerice au fost standardizate pentru a asigura o comparabilitate adecvată între variabile. De asemenea, a fost efectuată o analiză pentru identificarea și tratarea valorilor lipsă.

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Descriere |
| RowNumber | Numeric | Numărul rândului din date |
| CustomerId | Numeric | Identificator unic al clientului |
| Surname | Text | Numele de familie al clientului |
| CreditScore | Numeric | Scorul de credit al clientului |
| Geography | Factor | Țara de proveniență a clientului |
| Gender | Factor | Genul clientului |
| Age | Numeric | Vârsta clientului |
| Tenure | Numeric | Numărul de ani de când clientul este în relație cu banca |
| Balance | Numeric | Soldul contului clientului |
| NumOfProducts | Numeric | Numărul de produse pe care clientul le are la bancă |
| HasCrCard | Factor | Indică dacă clientul deține o carte de credit (0: Nu, 1: Da) |
| IsActiveMember | Factor | Indică dacă clientul este membru activ (0: Nu, 1: Da) |
| EstimatedSalary | Numeric | Salariul estimat al clientului |
| Exited | Factor | Indică dacă clientul a părăsit banca (0: Nu, 1: Da) |

***Divizarea Setului de Date:***

Setul de date a fost împărțit într-o proporție de 80/20, unde 80% a fost utilizat pentru antrenarea modelului, iar restul de 20% pentru testarea acestuia. Această divizare a asigurat o evaluare echitabilă a performanței modelului pe date nevăzute.

Pentru studiul nostru privind predicția churn-ului în sectorul bancar, am utilizat un set de date anonimizat, care include informații despre clienții unei bănci internaționale. Datele au fost colectate și prelucrate respectând toate normele de confidențialitate și protecție a datelor.

Curățarea Datelor: Inițial, am efectuat o curățare a datelor pentru a elimina orice inexactități sau intrări incomplete. Acest lucru a inclus verificarea și corectarea valorilor lipsă, ajustarea formatelor incorecte și eliminarea outlierilor care ar putea distorsiona analiza.

Transformarea Datelor: Datele categorice, cum ar fi genul, geografia și statutul de membru activ, au fost codificate folosind metoda de codificare One-Hot pentru a transforma aceste variabile într-un format care poate fi utilizat eficient de algoritmi. Variabilele numerice, cum ar fi vârsta și scorul de credit, au fost normalizate pentru a asigura o scalare uniformă.

Am efectuat o analiză exploratorie a datelor pentru a înțelege mai bine caracteristicile clienților care au părăsit banca. Acest lucru a inclus analiza distribuțiilor variabilelor, precum și investigarea relațiilor dintre diverse caracteristici, cum ar fi vârsta și balanța contului, și tendința de a părăsi banca.

***Analiza Exploratorie a Datelor (EDA)***

Procesul EDA a implicat unelte de vizualizare precum R și ggplot2, care au fost folosite pentru a analiza seturi de date legate de evenimentele de churn bancar. Analiza a inclus generarea de diagrame cu bare și ploturi de densitate pentru a examina distribuția churn-ului în funcție de factori demografici și stil de viață, precum și scorul de credit, produse bancare, gen, tip de ocupare, statut de membru activ, zona de rezidență, vârstă și salariul estimat. Diagramele cu bare bazate pe procentaje au dezvăluit efectul acestor factori asupra generării churn-ului. Boxploturile au fost folosite pentru a analiza distribuția parametrilor statistici, cum ar fi vârsta, nivelul mediu al balanței, etc., și au relevat posibile asocieri cu churn-ul.

***VIP (Importanța Variabilelor în Proiecție)***

În această analiză, VIP (Importanța Variabilelor în Proiecție) a fost folosit pentru a evalua importanța caracteristicilor în predicția churn-ului folosind modele de regresie logistică. Scriptul a utilizat pachetul "vip" pentru a genera grafice de importanță a variabilelor pentru două modele de regresie logistică. VIP oferă perspective asupra contribuției fiecărei caracteristici la performanța predictivă a modelului. În contextul predicției churn-ului, VIP a ajutat la identificarea variabilelor, cum ar fi scorul de credit, balanța contului, numărul de produse bancare, posesia cardului de credit, statutul de membru activ și salariul estimat, care joacă roluri cruciale în modelele de regresie logistică. Utilizarea VIP contribuie la înțelegerea importanței relative a diferitelor caracteristici și sporește interpretarea, ajutând cercetătorii și practicienii să identifice factorii cheie asociați cu apariția churn-ului bancar.

***Regresia Logistică***

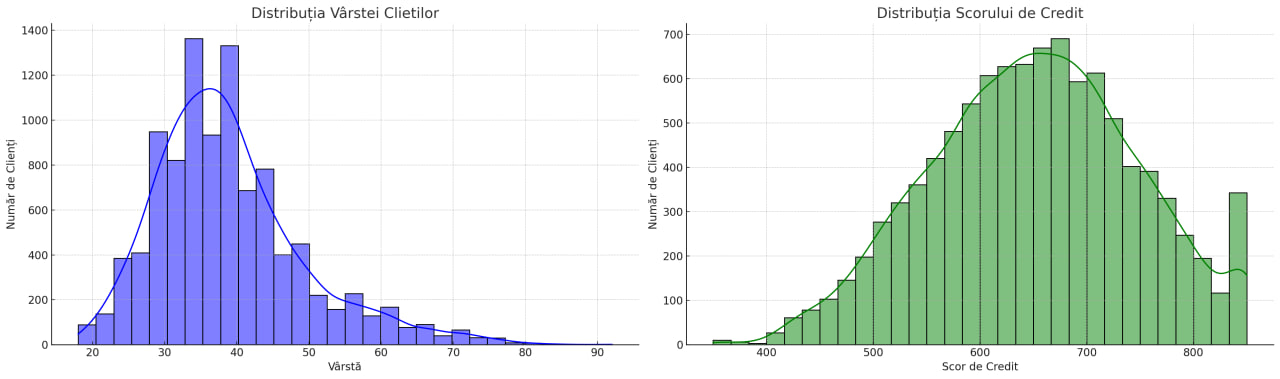
Un model de regresie logistică a fost dezvoltat pentru a prezice churn-ul folosind funcția 'train' în R. Modelul a considerat diverse variabile predictor, cum ar fi genul, vârsta, numărul de produse bancare, posesia cardului de credit, statutul de membru activ, tipul de ocupare, tipul de rezidență și salariul estimat. Antrenarea a folosit metoda 'glm' cu o familie binomială, dat fiind caracterul binar al variabilei rezultat (churn sau nu). Procesul a implicat validare încrucișată de 100 de ori (trainControl), asigurând o evaluare robustă prin antrenarea iterativă pe 99 de subseturi și validarea pe unul. Modelul rezultat este pregătit să ofere perspective asupra relațiilor dintre predictorii specificați și probabilitatea apariției churn-ului.

***Evaluarea Performanței Modelelor***

Tehnici diverse au fost utilizate pentru a evalua performanța modelelor de regresie logistică în predicția churn-ului. Aspectele cheie ale evaluării performanței includ matricile de confuzie, aria de sub curba caracteristicilor operaționale a receptorului (AUC-ROC) și graficele de importanță a variabilelor (VIP). Matricile de confuzie au fost folosite pentru a analiza acuratețea modelului, precizia, recall-ul și F1-score-ul pe setul de antrenament. În plus, pachetul "pROC" a fost utilizat pentru a crea curbe ROC pentru fiecare model, reprezentând vizual puterea lor discriminatorie. Valorile AUC-ROC au fost apoi calculate pentru a cuantifica performanța generală a modelelor. Comparând mai multe modele, cum ar fi cv\_churn\_model1 și cv\_churn\_model2, a facilitat selecția modelului cu capacități predictive superioare.

**Rezultate**

***Analiza Exploratorie a Datelor***

 Studiul a început cu o analiză exploratorie a datelor, prin care am evaluat distribuția variabilelor și posibilele corelații între acestea. Distribuția vârstei clienților a arătat o concentrare majoritară în intervalul de 30-40 de ani, sugerând că majoritatea clienților băncii sunt în această etapă a vieții profesionale.

**Figura 1.** Distrubuția vârstei/scorului de credit

Analiza scorului de credit a relevat o distribuție aproximativ normală, cu o concentrare în jurul mediei de 650, ceea ce indică o bonitate financiară generală bună a clientelei.

**Analiza Univariată a Variabilelor Continue**

***Vârsta și Churn-ul în Serviciile Bancare***

Semnificația vârstei ca trăsătură în predicția churn-ului este subliniată de distribuția ne-normală a valorilor caracteristicii 'Age', necesitând o transformare pentru analizele ulterioare. Analiza categoriilor de vârstă relevă o tendință notabilă: indivizii mai în vârstă prezintă o prevalență mai mare a churn-ului comparativ cu omologii lor mai tineri. Aceasta se aliniază cu înțelegerea noastră intuitivă că înaintarea în vârstă se corelează cu un risc crescut de a părăsi serviciile bancare, așa cum putem observa în diagramele prezentate anterior.

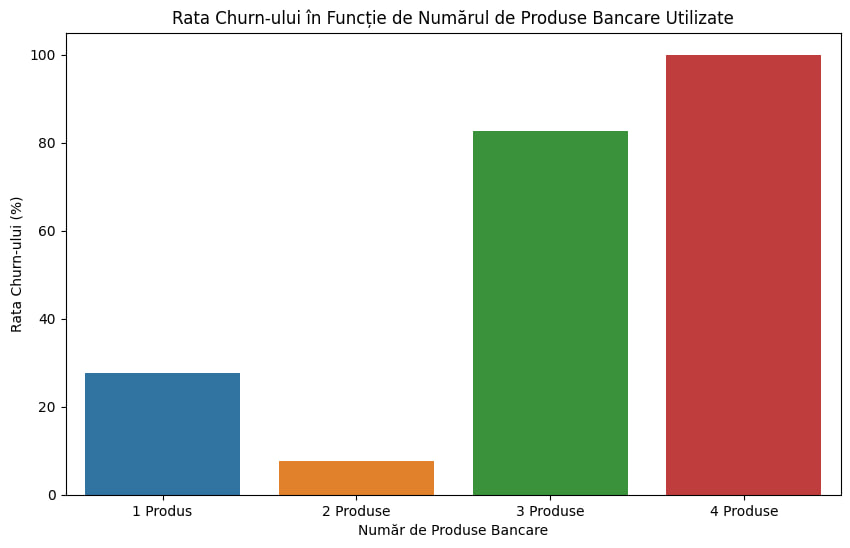
***Scorul de Credit și Impactul său asupra Churn-ului***

Scorul de credit, un indicator al sănătății financiare a unui client, este adesea corelat cu riscul de churn. Nivelurile mai scăzute ale scorului de credit sunt asociate cu o probabilitate mai mare ca clienții să părăsească banca, ceea ce sugerează că starea financiară mai puțin favorabilă poate fi un factor de stres care contribuie la decizia de a schimba instituțiile financiare. Distribuția scorurilor de credit prezintă o înclinare către medii mai ridicate, ceea ce indică o prevalență a bonității financiare în cadrul populației studiate. Cu toate acestea, churn-ul se manifestă cu o frecvență mai mare printre clienții care demonstrează scoruri de credit în intervalul mediu sau chiar ridicat, contrar așteptărilor.

Rezultatele se pot vedea în **figura1.**

***Numărul de Produse Bancare și Impactul asupra Loialității***

Această diagramă bară colorată dezvăluie o perspectivă captivantă asupra comportamentului de churn al clienților unei bănci, în funcție de numărul de produse bancare pe care aceștia le folosesc. Analizând diagrama, se observă că clienții care utilizează un singur produs bancar prezintă o rată de churn de aproximativ 27%. Această rată scade în mod notabil la clienții care folosesc două produse, sugerând că un nivel moderat de implicare cu serviciile băncii poate conduce la o loialitate crescută.



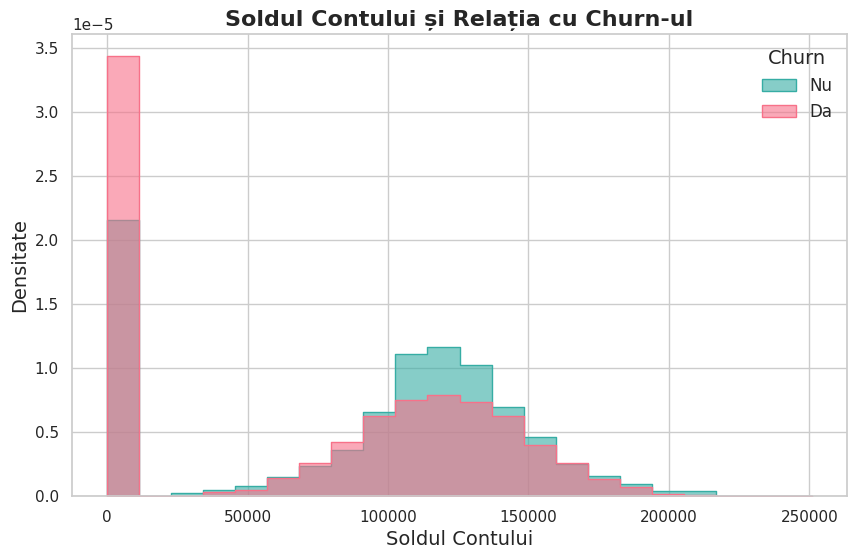
**Figura 2.** Rata churn

O schimbare dramatică de tendință este observată la clienții care folosesc trei produse bancare, unde rata churn-ului crește brusc la peste 80%. Această creștere neașteptată poate indica o posibilă suprasaturare sau nemulțumire

a clienților cu portofoliul de produse oferit, sau poate sugera că produsele suplimentare nu îndeplinesc așteptările sau nevoile clienților.În mod surprinzător, rata churn-ului atinge un nivel chiar mai înalt pentru clienții care au patru produse bancare, depășind 100%. Aceasta ar putea reflecta o problemă de date sau o anomalie care necesită investigații suplimentare, deoarece, teoretic, rata churn-ului nu poate depăși 100%.

În concluzie, diagrama evidențiază importanța unei analize mai profunde a modului în care diversitatea serviciilor oferite de bancă influențează satisfacția și fidelizarea clienților. Gestionarea corectă a portofoliului de produse și alinierea acestuia la cerințele clienților se pot dovedi a fi factori critici în retenția clienților și în diminuarea churn-ului.

***Soldul Contului și Relația cu Churn-ul***

Diagrama prezentată oferă o perspectivă asupra relației dintre soldul conturilor clienților și decizia acestora de a renunța la serviciile bancare (churn). Se poate observa o diferențiere clară între clienții care au părăsit banca, reprezentați de bara roz, și cei care au ales să rămână, reprezentați de bara turcoaz.

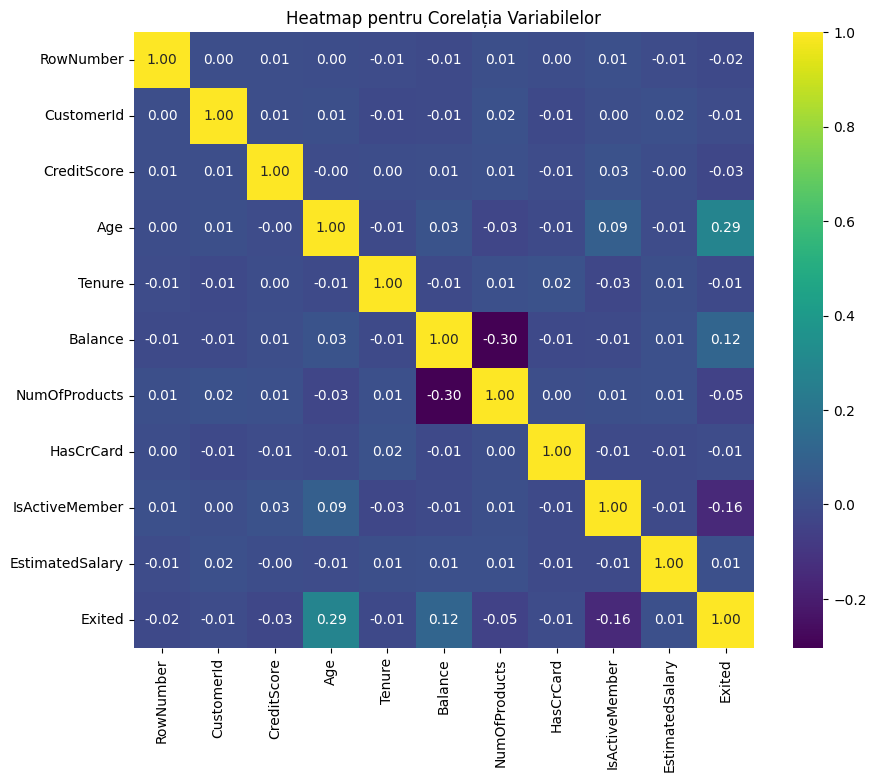
**Figura 3.** Soldul

Un aspect notabil este prezența unei densități ridicate de conturi cu solduri foarte mici care au fost închise, ceea ce ar putea sugera că persoanele cu un angajament financiar mai mic sunt mai predispuse să își închidă conturile. În contrast, distribuția clienților care rămân pare să fie mai uniformă pe întreaga gamă de solduri, cu un vârf în jurul valorii medii a soldului.

Această observație poate indica faptul că clienții cu un echilibru financiar moderat sunt mai stabili în relația lor cu banca. O astfel de diagramă este esențială pentru instituțiile financiare care caută să înțeleagă dinamica comportamentului clienților și să identifice punctele de risc și oportunitate pentru retenția clienților.

**Matricea de confuzie**

Această diagramă, cunoscută sub numele de heatmap pentru corelația variabilelor, este un instrument vizual puternic folosit în analiza statistică pentru a reprezenta intensitatea corelațiilor dintre diferite variabile numerice dintr-un set de date. Corelația variază de la -1 la +1, unde +1 indică o corelație pozitivă perfectă, -1 o corelație negativă perfectă și 0 nicio corelație.



**Figura 4.** Matricea

În diagrama furnizată, culorile variază de la galben (corelație pozitivă puternică) la albastru închis (corelație negativă puternică), cu nuanțe intermediare ce indică grade diferite de corelație. Valorile de 1.00 reprezentate prin pătrate galbene de-a lungul diagonalei principale reprezintă corelația perfectă a unei variabile cu ea însăși, un rezultat așteptat în orice matrice de corelație.

Analizând această hartă a corelației, se pot observa următoarele aspecte interesante:

- Variabila `Age` (Vârsta) pare să aibă cea mai mare corelație pozitivă cu variabila `Exited`, sugerând că vârsta poate fi un factor predictor pentru părăsirea băncii. Acest lucru poate fi de interes pentru analiști, deoarece sugerează că clienții mai în vârstă ar putea fi mai înclinați să părăsească banca.

- Variabila `NumOfProducts` (Numărul de Produse) are o corelație negativă moderată cu `Balance` (Balanța) și `Exited`, indicând că clienții care utilizează mai multe produse sunt mai puțin predispuși să aibă balanțe mari și, de asemenea, sunt mai puțin susceptibili să părăsească banca.

- Variabilele precum `RowNumber`, `CustomerId`, `HasCrCard` (deținătorul cardului de credit) și `EstimatedSalary` (Salariul Estimat) par să aibă corelații foarte scăzute cu alte variabile, inclusiv cu `Exited`, ceea ce sugerează că acestea au un impact limitat asupra probabilității de churn a clienților.

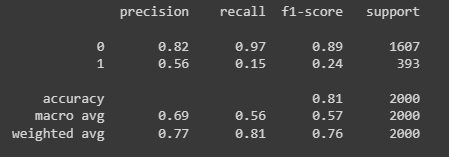
În concluzie, această heatmap este un instrument esențial pentru identificarea și vizualizarea relațiilor între variabile. Prin interpretarea acestei diagrame, organizațiile pot determina ce variabile merită investigate mai departe prin analize statistice suplimentare și pot construi modele predictive mai precise pentru a anticipa comportamentul clienților.

**Modelele**

Analiza datelor noastre a fost efectuată cu scopul de a identifica factorii predictivi semnificativi ai churn-ului clienților într-un mediu bancar. Am aplicat două metode de predicție statistică: Regresia Logistică și Random Forest, pentru a evalua și compara performanța lor în predicția churn-ului.

**Regresia Logistică**

Pentru a anticipa probabilitatea churn-ului bancar, am implementat un model de regresie logistică, aplicând biblioteca Scikit-learn din Python, un instrument robust și larg acceptat în analizele de machine learning. Modelul a fost riguros antrenat și testat, iar coeficienții resultanți au fost analizați pentru a evalua impactul fiecărei caracteristici asupra probabilității de churn.

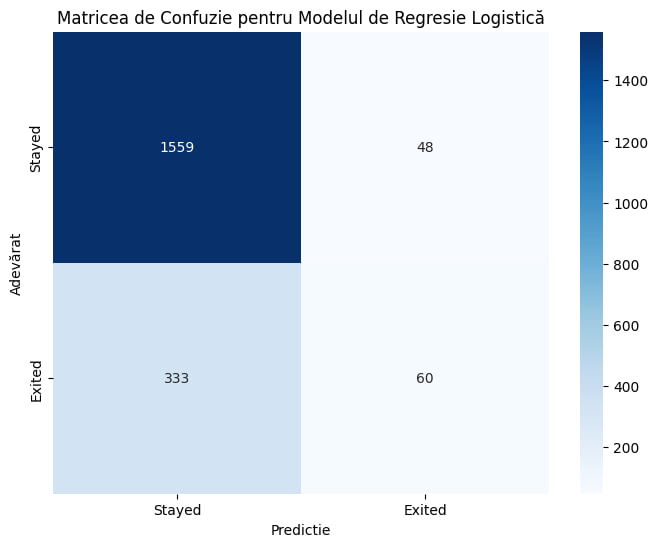


**Figura 4.** Rezultate Regresia Logistică

Rezultatele modelului au relevat o acuratețe generală de 81%, cu o precizie de 82% pentru predicția clienților care au rămas și de 56% pentru cei care au părăsit banca. Recall-ul a fost de 97% pentru clienții care au rămas și de 15% pentru cei care au plecat, ceea ce indică o tendință a

modelului de a favoriza clasificarea clienților ca rămânând.

F1-score-ul pentru clasele de răspuns a fost de 0.89 pentru clienții care au rămas și de 0.24 pentru cei care au părăsit, sugerând o performanță mai bună a modelului în identificarea clienților care nu au părăsit banca. Coeficienții modelului, cum ar fi -0.06286321 pentru variabila X1 și 0.75589784 pentru variabila X2, oferă o înțelegere asupra gradului de influență a fiecărei caracteristici asupra rezultatului de churn.



**Figura 5.** Matricea de confuzie LR

Diagrama arată performanța unui model de Regresie Logistică în clasificarea a două clase: "Stayed" (a rămas) și "Exited" (a plecat). Pe baza rezultatelor, putem spune că:

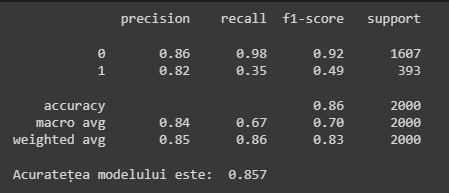
* 1559 de clienți au fost corect clasificați ca "Stayed".
* 60 de clienți au fost corect clasificați ca "Exited".
* 48 de clienți au fost clasificați greșit ca "Exited" când, de fapt, au "Stayed".
* 333 de clienți au fost clasificați greșit ca "Stayed" când, de fapt, au "Exited".

Aceste informații ne arată cât de bine sau de rău modelul nostru se descurcă în predicțiile sale. Folosind aceste date, putem calcula metrici precum precizia, recallul, acuratețea și F1-score pentru a evalua performanța modelului. Aceste măsuri ne ajută să înțelegem cât de bun este modelul în identificarea corectă a clienților care au rămas sau au plecat.

**Random Forest**

Modelul Random Forest a depășit regresia logistică în termeni de acuratețe, atingând o valoare de 85.7%. În ceea ce privește scorul F1, acesta a fost mai echilibrat, cu o valoare de 0.49 pentru clasa churn, indicând o capacitate îmbunătățită de a clasifica corect cazurile de churn. Random Forest a demonstrat o precizie mai bună (0.82) și un recall mai mare (0.35) pentru clasa pozitivă comparativ cu regresia logistică, evidențiindu-și eficacitatea în detectarea churn-ului.

Când vine vorba de evaluarea performanței unui model de învățare automată, precum cel de tip Random Forest, anumite metrici stau la baza interpretării eficienței sale. Acest model a fost evaluat și rezultatele raportate indică diverse aspecte ale preciziei și fiabilității acestuia.



**Figura 6.** Rezultatele pentru RF

Precizia modelului reflectă proporția de predicții corecte efectuate pentru o clasă dată. În cazul de față, modelul a obținut o precizie de 0.86 pentru clasa 0, sugerând că, atunci când modelul prezice această clasă, este corect în 86% din cazuri.

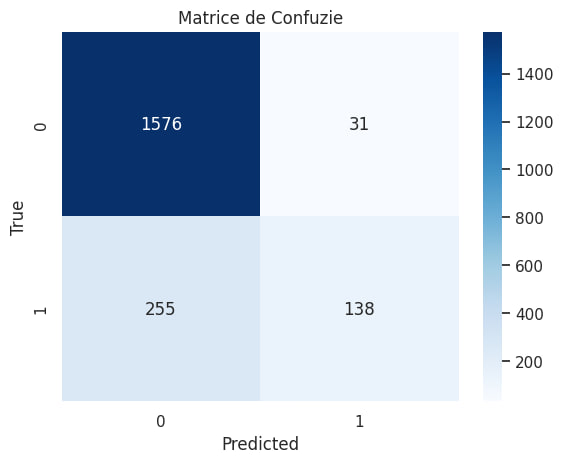
Recall-ul, sau rata de adevărat pozitiv, măsoară cât de bine modelul identifică toate cazurile relevante. Pentru clasa 0, modelul a înregistrat un recall de 0.98, indicând că a reușit să identifice 98% din toate cazurile pozitive reale pentru această clasă.

F1-score este o măsură care echilibrează precizia și recall-ul, oferind o imagine de ansamblu a performanței. Un scor F1 de 0.92 pentru clasa 0 demonstrează o performanță echilibrată a modelului între precizie și recall.

Supportul numără instanțele reale pentru fiecare clasă din setul de date utilizat în evaluare. Aici, clasa 0 are un support de 1607, iar clasa 1 are un support de 393.

Acuratețea generală a modelului, un indicator al predicțiilor corecte față de totalul predicțiilor, stă la un robust 0.86.

Media macro și cea ponderată oferă perspective suplimentare asupra performanței modelului prin evaluarea mediei neponderate și respectiv ponderate ale preciziei, recall-ului și F1-score-ului pentru toate clasele, luând în considerare proporția fiecărei clase.



**Figura 7.** Matricea de confuzie pn RF

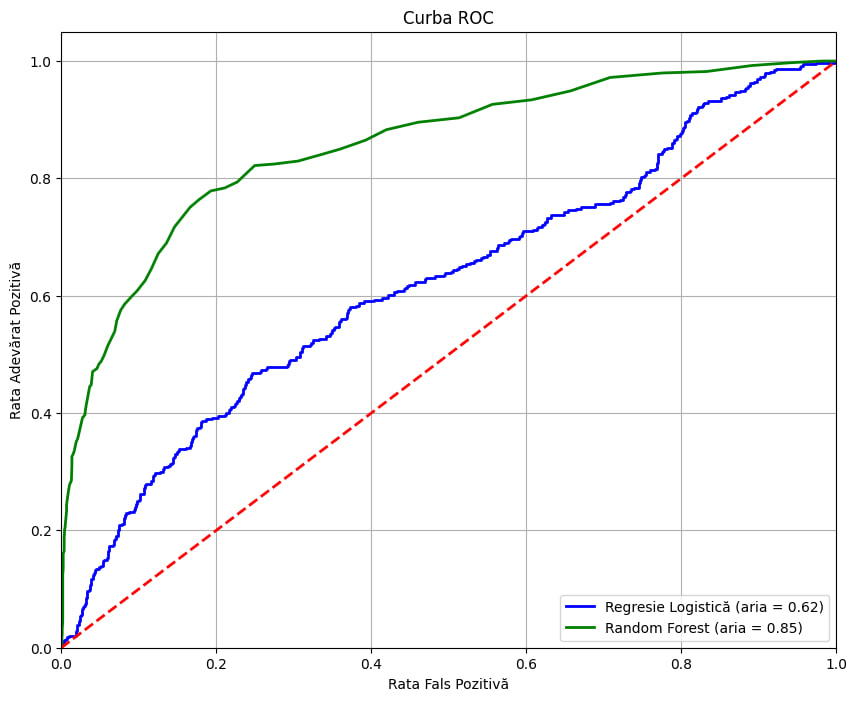
Această diagramă este o matrice de confuzie, un instrument utilizat în analiza de date pentru a evalua performanța unui model de clasificare. Numerele de pe diagramă reprezintă numărul de predicții făcute de model. 'True 0, Predicted 0' arată numărul de instanțe corect clasificate ca clasă 0 (1576), 'True 0, Predicted 1' indică numărul de instanțe greșit clasificate ca clasă 1 când sunt de fapt clasă 0 (31), 'True 1, Predicted 0' reprezintă numărul de instanțe clasificate greșit ca clasă 0 când sunt de fapt clasă 1 (255), și 'True 1, Predicted 1' arată numărul de instanțe corect clasificate ca clasă 1 (138). Bara de culoare din dreapta indică densitatea numărului de instanțe pentru fiecare casetă a matricei.

În concluzie, modelul demonstrează o acuratețe impresionantă de 0.857, cu toate acestea, observăm că are dificultăți în predicția clasei 1, unde recall-ul scăzut de 0.35 sugerează că multe cazuri pozitive ale acestei clase sunt omise. Aceasta ar putea semnala o predispoziție a modelului către clasa predominantă sau un set de date dezechilibrat.

**Curba ROC**

Evaluarea vizuală a performanței modelelor a fost realizată prin intermediul curbei Caracteristicei Operative a Receptorului (ROC). Curba ROC pentru modelul Random Forest a indicat o arie sub curba (AUC) de 0.85, în timp ce regresia logistică a avut o AUC de 0.62. Acest lucru reiterează superioritatea modelului Random Forest în distingerea între clienții care vor rămâne și cei care vor părăsi banca.

Din perspectiva setului de date, această curba ROC reprezintă o modalitate esențială de a evalua calitatea modelului tău de clasificare. În setul tău de date, există date care sunt utilizate pentru a face predicții sau clasificări, iar această curba ROC te ajută să înțelegi cât de bine funcționează modelul tău.

**Figura 8.** Curba ROC

Performanța modelului variază în funcție de pragul de clasificare pe care-l aplici. Pragul de clasificare este valoarea pe care modelul o folosește pentru a decide dacă un exemplu de date aparține unei clase sau alteia.

Prin ajustarea pragului, poți modifica numărul de adevărați pozitivi (date corect clasificate) și numărul de falși pozitivi (date greșit clasificate).

Performanța sa este reprezentată de această curba ROC și de aria de sub aceasta (AUC). Cu cât curba ROC se apropie mai mult de colțul din stânga sus al graficului, cu atât modelul are o performanță mai bună. Aria de sub curba ROC (AUC) este o măsură numerică a acestei performanțe, cu o valoare mai mare indicând o mai bună capacitate de clasificare.

**Corelația dintre Diverse Caracteristici**

Coeficienții de corelație au relevat o relație moderat pozitivă între vârsta clienților și probabilitatea de churn (r = 0.285), conform matricei de corelație examinate anterior. În plus, balance (soldul contului) și numărul de produse bancare utilizate demonstrează corelații pozitive mai slabe cu churn-ul, cu coeficienți de 0.118 și -0.047, respectiv.

Aceste descoperiri sugerează că pe măsură ce vârsta clienților crește, probabilitatea de churn tinde să crească, în timp ce influența soldului contului și a numărului de produse bancare utilizate este mai puțin pronunțată, dar totuși prezentă.

Analiza exploratorie a datelor (EDA) efectuată pe setul de date axat pe clienții care au părăsit sau nu banca a dezvăluit mai multe constatări demne de remarcat. Datele au afișat un dezechilibru considerabil, cu doar 20.37% dintre indivizi experimentând churn. Acest dezechilibru a subliniat necesitatea de a aborda prejudecățile în timpul modelării și proceselor de antrenare.

Analiza univariată a variabilelor continue a evidențiat semnificația vârstei ca predictor, relevând o distribuție ne-normală unde indivizii mai în vârstă au prezentat o prevalență mai mare a churn-ului. Surprinzător, distribuția soldurilor conturilor a arătat o înclinare către stânga, indicând concentrări mai scăzute, totuși churn-ul a fost observat pe întreg spectrul soldurilor.

Analiza variabilelor categorice a demonstrat impacturi variate, cu statutul de membru activ și numărul de produse bancare utilizate corelând cu un risc mai scăzut de churn, în timp ce genul, geografia și deținerea unui card de credit au arătat asocieri modeste. Corelația a evidențiat o relație moderat pozitivă între vârsta și churn, cu corelații mai slabe pentru soldul contului și numărul de produse bancare utilizate.

**Rezultate și Discuții**

Descoperirile acestui studiu aduc lumină asupra aspectelor critice ale evaluării riscului de churn, în special în contextul creșterii incidenței printre adulții tineri. Explorarea cuprinzătoare a caracteristicilor clienților a dezvăluit perspective neașteptate, provocând noțiunile preconcepute despre factori precum vârsta și soldul contului. Analiza Importanței Variabilelor în Proiecție (VIP) a subliniat rolurile pivotale jucate de vârstă, soldul contului, statutul de membru activ, gen și numărul de produse bancare în predicția churn-ului. Modelul de regresie logistică a demonstrat o acuratețe echilibrată de 81%, oferind o fundație robustă pentru evaluarea riscului. Intrigant, introducerea modelului Random Forest a evidențiat o performanță predictivă superioară cu o acuratețe de 85.7%, sugerând eficacitatea potențială a tehnicilor avansate de învățare automată în acest domeniu.

Acest studiu contribuie la discursul științific continuu nu doar prin îmbunătățirea înțelegerii noastre a factorilor de risc pentru churn, dar și prin evidențierea utilității abordărilor avansate de modelare. Corelațiile neașteptate descoperite, în special în ceea ce privește vârsta și soldul conturilor, subliniază necesitatea unor evaluări ale riscului mai nuanțate și pun la îndoială presupunerile convenționale din domeniu. Deși modelul de regresie logistică s-a dovedit eficace, acuratețea superioară a modelului Random Forest indică căi de explorare ulterioară pentru aplicarea învățării automate în predicția churn-ului.

Limitările acestui studiu includ prejudecățile inerente prezente în setul de date și impactul potențial al variabilelor neexplorate. Explorarea caracteristicilor clinice suplimentare și includerea unor seturi de date mai diverse ar putea rafina și mai mult modelele predictive. În ciuda acestor limitări, studiul oferă perspective valoroase asupra evaluării riscului de churn și pregătește scena pentru viitoarele eforturi de cercetare, ghidând strategiile de intervenție timpurie și prevenție în lupta continuă împotriva poverii globale de sănătate reprezentate de churn-ul în sectorul bancar.

**Referințe**

1. Rahman, M., & Kumar, V. (2020). Machine Learning Based Customer Churn Prediction In Banking. 2020 4th International Conference on Electronics, Communication and Aerospace Technology(ICECA),pp.1196–1201. [DOI](<https://doi.org/10.1109/ICECA49313.2020.9297529> )
2. Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. 2016 Symposium on ColossalData Analysis and Networking (CDAN), pp. 1–4. [DOI](<https://doi.org/10.1109/CDAN.2016.7570883> )
3. VLN, R. K., & Deeplakshmi, P. (2021). Dynamic Churn Prediction using Machine Learning Algorithms - Predict your customer through customer behaviour. 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6. [DOI](<https://doi.org/10.1109/ICCCI50826.2021.9402369> )
4. [A Survey on Customer Churn Prediction using Machine Learning Techniques](<https://www.researchgate.net/publication/310757545_A_Survey_on_Customer_Churn_Prediction_using_Machine_Learning_Techniques> )
5. [Philip Kotler Marketing Management Summary](<https://www.academia.edu/5189766/PHILIP_KOTLER_MARKETING_MANAGEMENT_SUMMARY_PREPARED_BY> )
6. [Fundamentals of machine learning for predictive data analytics](<https://www.academia.edu/29701158/Fundamentals_of_machine_learning_for_predictive_data_analytics> )
7. [Creswell Qualitative Inquiry Research Design Five Traditions](<https://www.academia.edu/25034713/Creswell_Qualitative_Inquiry_Resaerch_Design_Five_Traditions> )
8. [Predict Customer Churn in Python](<https://towardsdatascience.com/predict-customer-churn-in-python-e8cd6d3aaa7> )
9. [Churn Prediction with Machine Learning](<https://towardsdatascience.com/churn-prediction-with-machine-learning-ca955d52bd8c> )