**Preciza clienți care ar putea părăsi seriviciul card de credit.**

**Capațîna Cristian, IA-211,III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

**Absract**

În contextul preocupării unui director bancar cu privire la creșterea întreruperilor serviciilor de card de credit, acest studiu utilizează datele furnizate de https://leaps.analyttica.com/home pentru a analiza și anticipa nemulțumirile clienților. Setul de date, compus din informații despre 10.000 de clienți, inclusiv vârsta, salariul, starea civilă și limita cardului de credit, a fost analizat pentru a identifica factorii care ar putea contribui la deciziile de retragere ale clienților.

Cu un procent actual de retragere de 16,07%, dificultatea antrenării unui model de previziune este recunoscută. Scopul final al acestei analize este să permită directorului bancar să identifice clienții nemulțumiți în mod proactiv, oferindu-le servicii îmbunătățite și, astfel, contracarând deciziile de retragere. Această abordare proactivă ar putea contribui la menținerea și fidelizarea clienților, consolidând astfel relația cu aceștia în cadrul serviciilor de card de credit.

**Introducere**

Într-un peisaj financiar dinamic, dominat de o preocupare crescândă legată de întreruperile serviciilor de card de credit, un director bancar proactiv a căutat să implementeze o abordare de analiză avansată pentru a anticipa și gestiona nemulțumirile clienților. Prin exploatarea datelor furnizate de https://leaps.analyttica.com/home, acest studiu intră într-un teritoriu complex de analiză, având ca obiectiv identificarea subtilităților și a factorilor determinant în comportamentul clienților pentru a dezvolta strategii de intervenție eficiente.

Setul de date cuprinzător, documentând detaliat profilurile a 10.000 de clienți, aduce laolaltă informații privind vârsta, salariul, starea civilă, limita cardului de credit și alți parametri critici, punând bazele pentru o analiză detaliată și cuprinzătoare. În ciuda unui procent actual relativ scăzut de retragere, de 16,07%, directorul bancar recunoaște complexitatea în antrenarea unui model predictiv robust, fapt ce impune o abordare avansată și sofisticată a analizei datelor.

Această inițiativă nu se rezumă doar la identificarea clienților cu potențial de retragere, ci vizează și înțelegerea profundă a motivelor care stau la baza deciziilor acestora. Prin intermediul tehnologiilor avansate de analiză de date, se propune dezvoltarea unui model predictiv care să ofere o înțelegere contextuală a schimbărilor în comportamentul clienților, astfel încât să permită abordarea proactivă a problemelor și oferirea de soluții personalizate.

Această abordare avansată nu doar răspunde cerințelor actuale, ci deschide noi orizonturi în managementul relațiilor cu clienții în industria financiară. Într-o eră a analizei de date și a inteligenței artificiale, această inițiativă ilustrează angajamentul ferm în inovație și adaptabilitate în abordarea problemelor complexe, conturând direcții strategice care pot transforma nu doar modul în care se gestionează retragerile clienților, ci și modul în care se construiește loialitatea și se optimizează experiența acestora.

**Materiale & Metode**

**Setul de date:**

Pentru analiza privind comportamentul clienților în domeniul serviciilor de card de credit, am explorat un set de date detaliat disponibil pe Kaggle, intitulat "Credit Card Customers". Acest set de date furnizează o perspectivă comprehensivă asupra caracteristicilor clienților și a factorilor care pot influența deciziile acestora de a părăsi sau de a rămâne în serviciile de card de credit. Mai jos se regăsește o prezentare structurată a informațiilor relevante:

**Structura Setului de Date:**

* Număr de Înregistrări (rânduri): 10.127
* Număr de Coloane (variabile): 21

Acest set de date extins reprezintă o resursă valoroasă pentru analiza comportamentului clienților în domeniul serviciilor de card de credit. Datele furnizează o imagine detaliată a variabilelor relevante, permițând profesioniștilor din domeniu, cercetătorilor de piață și decidenților din industrie să obțină insights semnificative privind dinamica și factorii cheie care influențează deciziile clienților în ceea ce privește părăsirea sau menținerea serviciilor de card de credit. Acest set de date poate servi drept instrument esențial pentru dezvoltarea de strategii informate în politica de retentie a clienților și optimizarea experienței acestora.

**Preprocesarea Datelor**

În cadrul procesului de preprocesare a datelor pentru setul de date "Credit Card Customers", au fost întreprinse acțiuni pentru gestionarea valorilor lipsă. Iată o scurtă prezentare a pașilor efectuați în procesul de preprocesare, cu evidențierea gestionării valorilor lipsă:

* S-au identificat valorile lipsă în setul de date prin examinarea fiecărei variabile.
* S-a efectuat o analiză a distribuției valorilor lipsă pentru a determina impactul acestora asupra variabilității datelor și interpretării rezultatelor.

Gestionarea valorilor lipsă a fost esențială pentru asigurarea calității datelor și pentru a evita influențarea rezultatelor analizei de către informații incomplete sau eronate. Procesul de preprocesare urmărește să ofere o bază solidă pentru analize ulterioare și interpretări precise ale comportamentului clienților în domeniul serviciilor de card de credit. Este important să se menționeze că această abordare a fost adoptată pentru a asigura coerența și fiabilitatea setului de date în contextul obiectivului analizei.

**Analiza Exploratorie a Datelor**

Am efectuat o analiză exploratorie detaliată a setului de date "Credit Card Customers" pentru a investiga comportamentul clienților în domeniul serviciilor de card de credit. Iată un rezumat structurat al acestei analize:

* Am examinat distribuția variabilelor categorice cum ar fi genul, nivelul de educație, starea civilă și categoria cardului.
* Am utilizat diagrame de bare pentru a ilustra proporțiile relative ale diferitelor categorii.
* Am creat un histogramă pentru a evalua distribuția vârstei clienților.
* Am identificat eventuale segmentări sau tendințe în funcție de grupurile de vârstă.
* Am examinat distribuția veniturilor în funcție de categorii precum nivelul de educație și categoria cardului.
* Am utilizat diagrame de cutie pentru a evidenția variabilitatea veniturilor în diferite grupuri.
* Am calculat matricea de corelație pentru variabilele numerice, evaluând relațiile dintre acestea.
* Am creat diagrame de dispersie pentru a vizualiza relațiile între venituri, limita de credit și alte variabile relevante.
* Am investigat distribuția variabilei țintă "Attrition\_Flag" pentru a evalua procentul de clienți care au părăsit serviciile de card de credit.
* Am utilizat diverse vizualizări, inclusiv diagrame de bare, histogramă, diagrame de cutie, pentru a evidenția caracteristicile cheie ale setului de date.
* Am furnizat informații statistice relevante pentru variabilele cheie prin intermediul funcției summary.

Această analiză exploratorie furnizează o perspectivă detaliată asupra comportamentului clienților în domeniul serviciilor de card de credit, evidențiind tendințele, variabilitatea și relațiile dintre diferitele variabile. Evidențiază, de asemenea, informații semnificative pentru a ghida analizele ulterioare și deciziile strategice în cadrul industriei financiare.

**Modelarea și Evaluare Performanței**

În această etapă crucială a analizei comportamentului clienților în domeniul serviciilor de card de credit, am abordat modelarea și evaluarea performanței pentru a dezvălui relațiile complexe dintre variabilele-cheie. Obiectivul nostru a fost să identificăm influențele diferitelor aspecte, precum vârsta clienților, limitele de credit, istoricul tranzacțiilor și altele, asupra deciziilor de părăsire a serviciilor de card de credit.

* Am împărțit setul de date în set de antrenare și set de testare pentru a evalua performanța modelelor pe date independente.
* Am identificat și selectat caracteristicile relevante care pot influența comportamentul clienților.
* Am selectat algoritmi de clasificare potriviți pentru problema dată, cum ar fi logistic regression, random forest sau support vector machines.
* Am antrenat modelele utilizând setul de antrenare și caracteristicile selectate.
* Am ajustat hiperparametrii modelelor pentru a maximiza performanța acestora.
* Am folosit modelele pentru a face predicții pe setul de testare.
* Am evaluat performanța modelelor utilizând matricea de confuzie pentru a evalua acuratețea predicțiilor.
* Am calculat metrici precum acuratețea, precizia, revocarea și F1-score pentru a evalua performanța modelelor.
* Am vizualizat curba ROC pentru a evalua capacitatea modelelor de a face distincție între clienții care rămân și cei care părăsesc serviciile de card de credit.
* Am gestionat variabile complexe, cum ar fi istoricul tranzacțiilor, aplicând transformări specifice pentru a asigura acuratețea și interpretabilitatea rezultatelor.
* Am explorat impactul variabilelor cheie asupra deciziilor clienților, ajustând parametrii modelelor pentru o potrivire optimă cu datele disponibile.

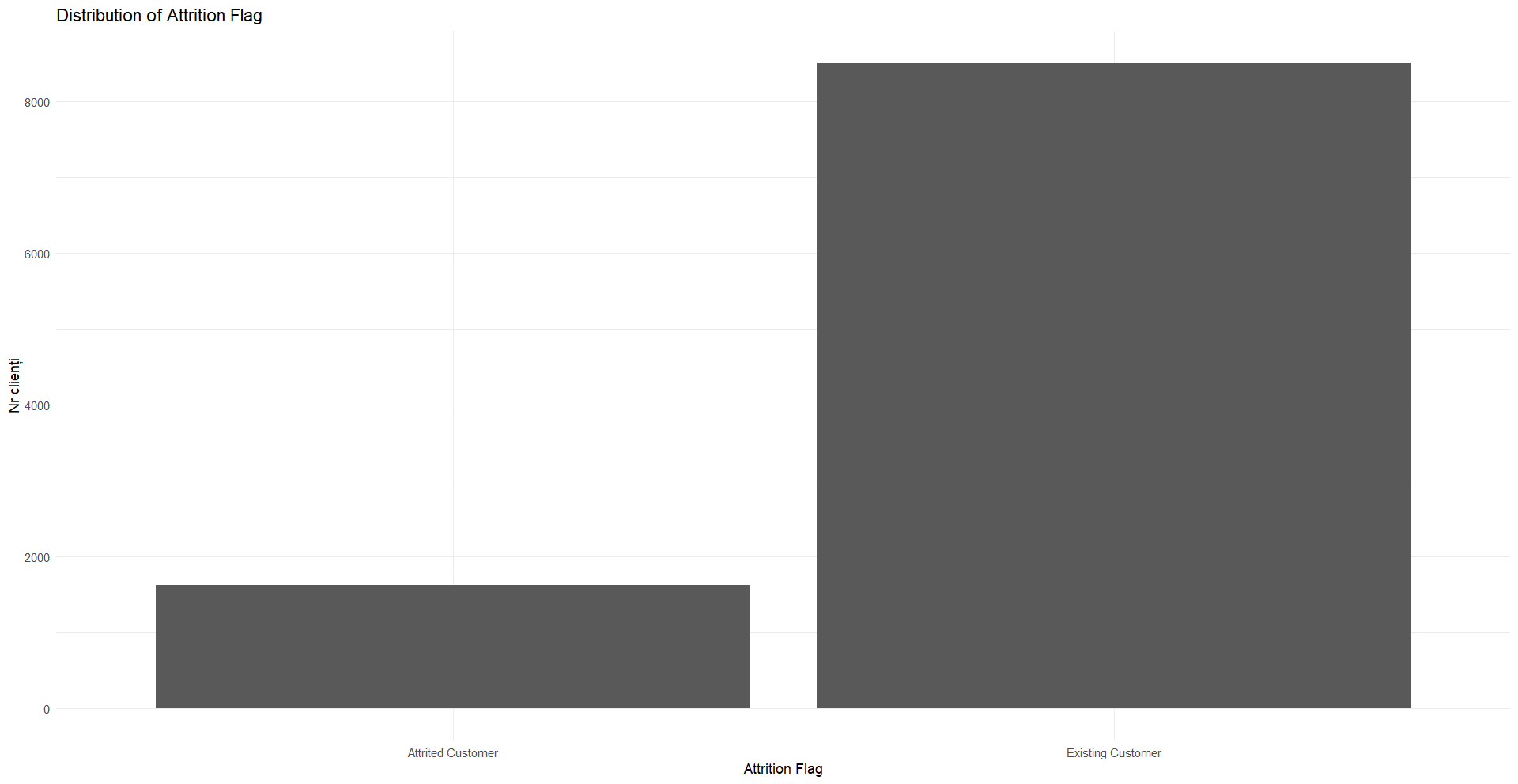
Această abordare sistematică a modelării și evaluării performanței contribuie la dezvoltarea unui model predictiv robust și la obținerea de insights semnificative privind comportamentul clienților în domeniul serviciilor de card de credit. Este esențial să adaptăm procesul la specificul setului de date pentru a asigura rezultate precise și relevante.

**Codul disponibil**

Întregul cod pentru executarea operațiilor,și pentru crearea modelului de regresie logistcă, este dispobil pe link-ul: https://github.com/Cristian628/AnalizaDatelor

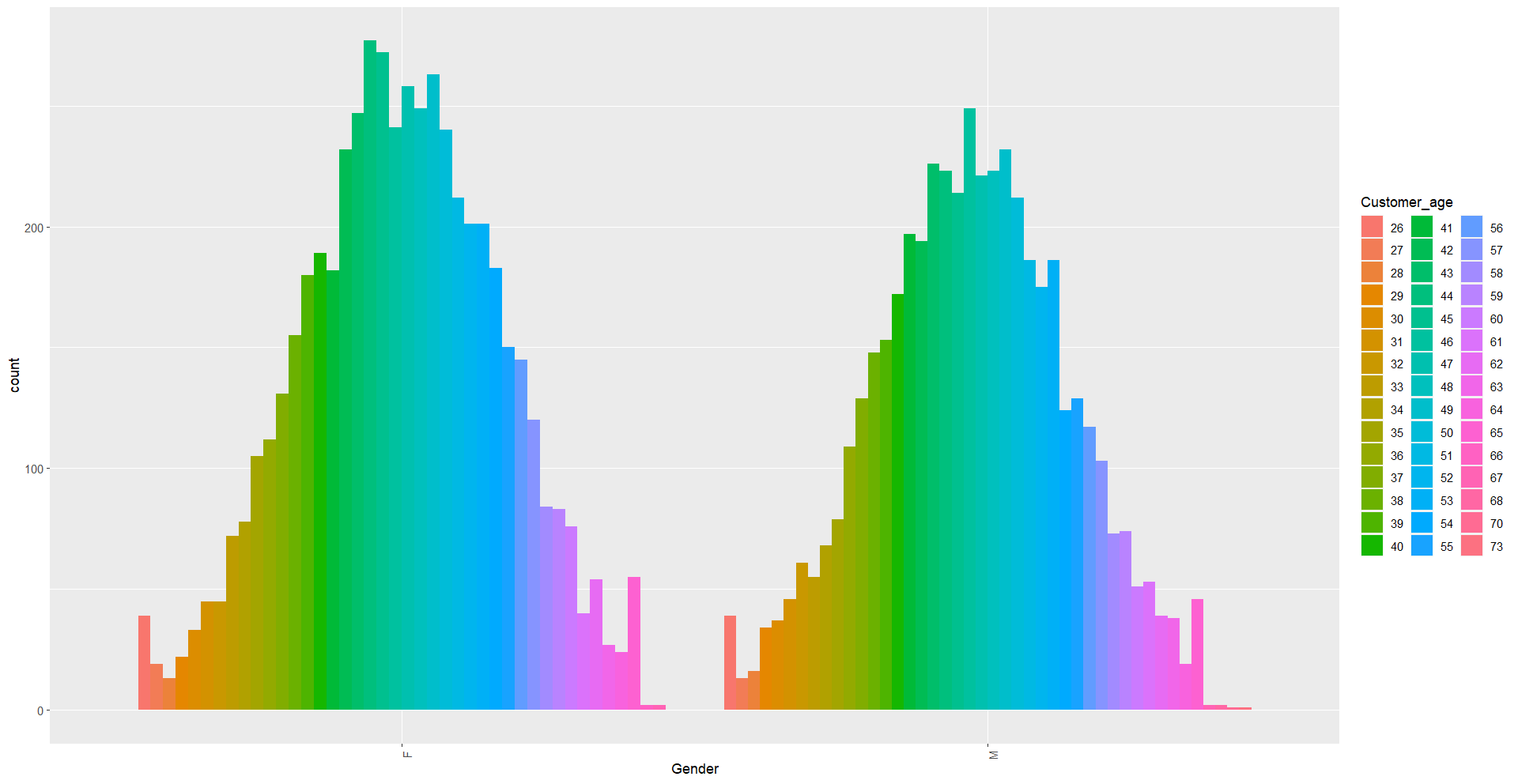
**Rezultate:**

După prelucrarea datelor despre clienții unei bănci și eliminarea unor coloane nerelevante, am inițiat o analiză exploratorie a datelor (Exploratory Data Analysis - EDA) pentru a înțelege mai bine compoziția și comportamentul clienților. Prin utilizarea unor biblioteci precum ggplot2 și plotly în limbajul de programare R, am creat vizualizări pentru diverse aspecte ale setului de date.

****

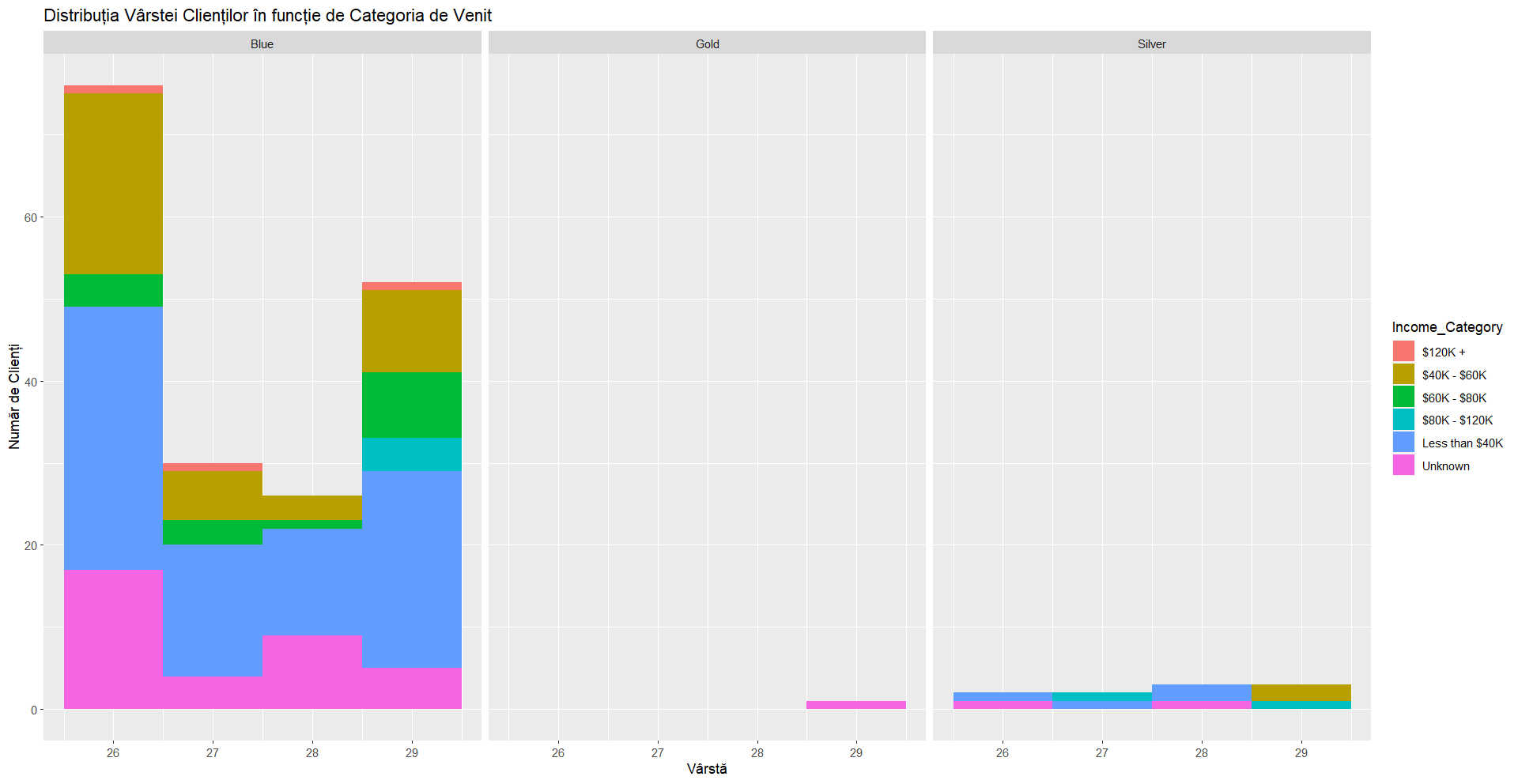
**Figura 1** Distribuția Attrition Flag

Această figură arată distribuția clienților care au părăsit serviciul de card de credit și a celor care încă sunt împreună cu acest serviciu. Pe axa orizontală este reprezentat numărul de clienți, iar pe axa verticală este reprezentată flag-ul de aterizare, care indică dacă clientul a părăsit serviciul sau nu. După cum se poate observa, aproximativ 20% dintre clienți au părăsit deja serviciul de card de credit. Acest lucru reprezintă o pierdere semnificativă pentru companie, deoarece acești clienți nu vor mai genera venituri din dobânzi sau taxe.

****

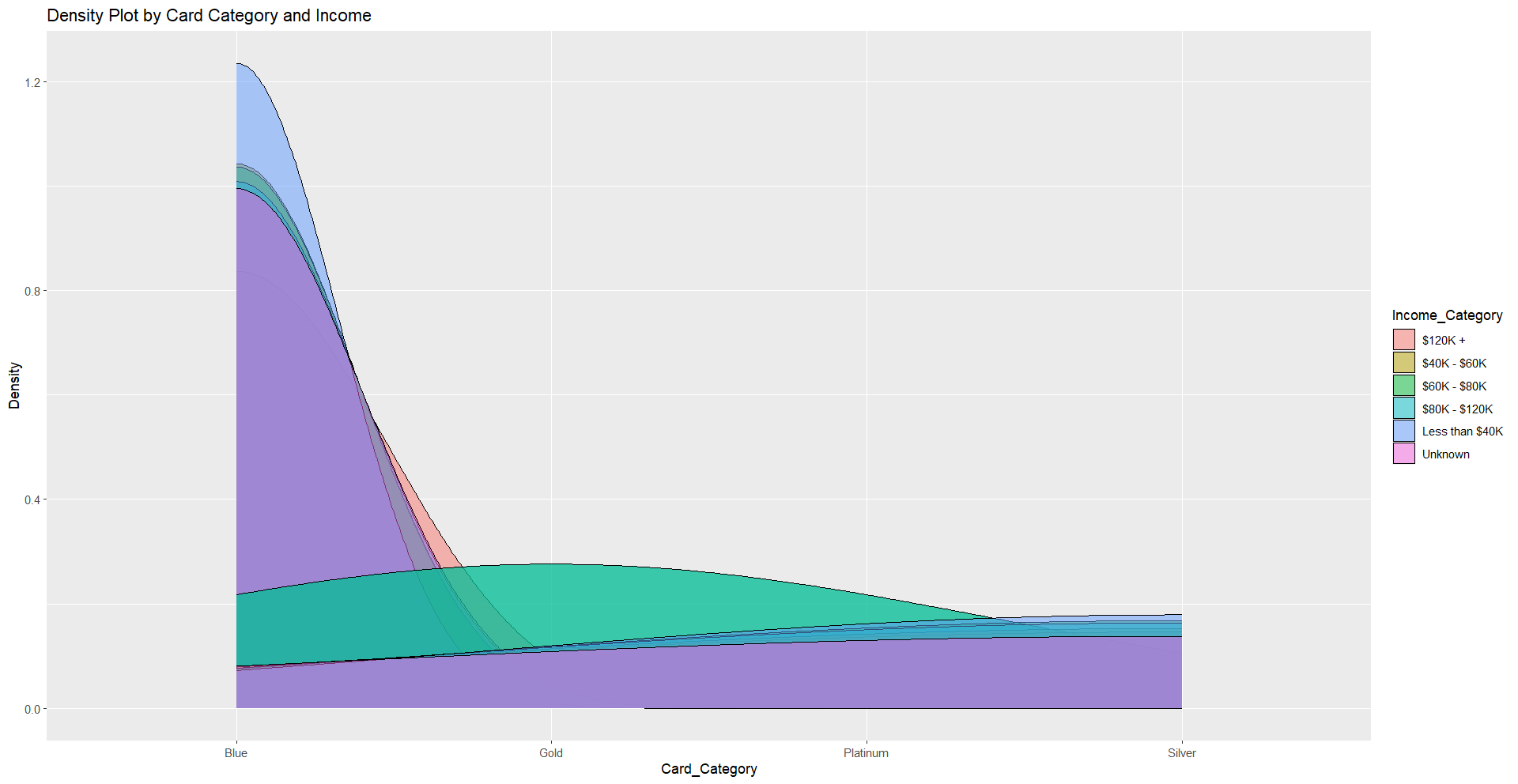
**Figura 2** Distribuția genului(m/f) pe vârste

Pe axa orizontală este reprezentată vârsta, iar pe axa verticală este reprezentat numărul de clienți. După cum se poate observa, cei mai mulți clienți sunt cu vârsta cuprinsă între 43 și 44 de ani, indiferent de gen. La femei, numărul de clienți este mai mare în intervalul de vârstă 43-44 de ani, în timp ce la bărbați, numărul de clienți este mai mare în intervalul de vârstă 46-47 de ani. Această diferență poate fi explicată prin faptul că femeile tind să aibă o speranță de viață mai mare decât bărbații. De asemenea, este posibil ca femeile să fie mai dispuse să utilizeze carduri de credit decât bărbații.

****

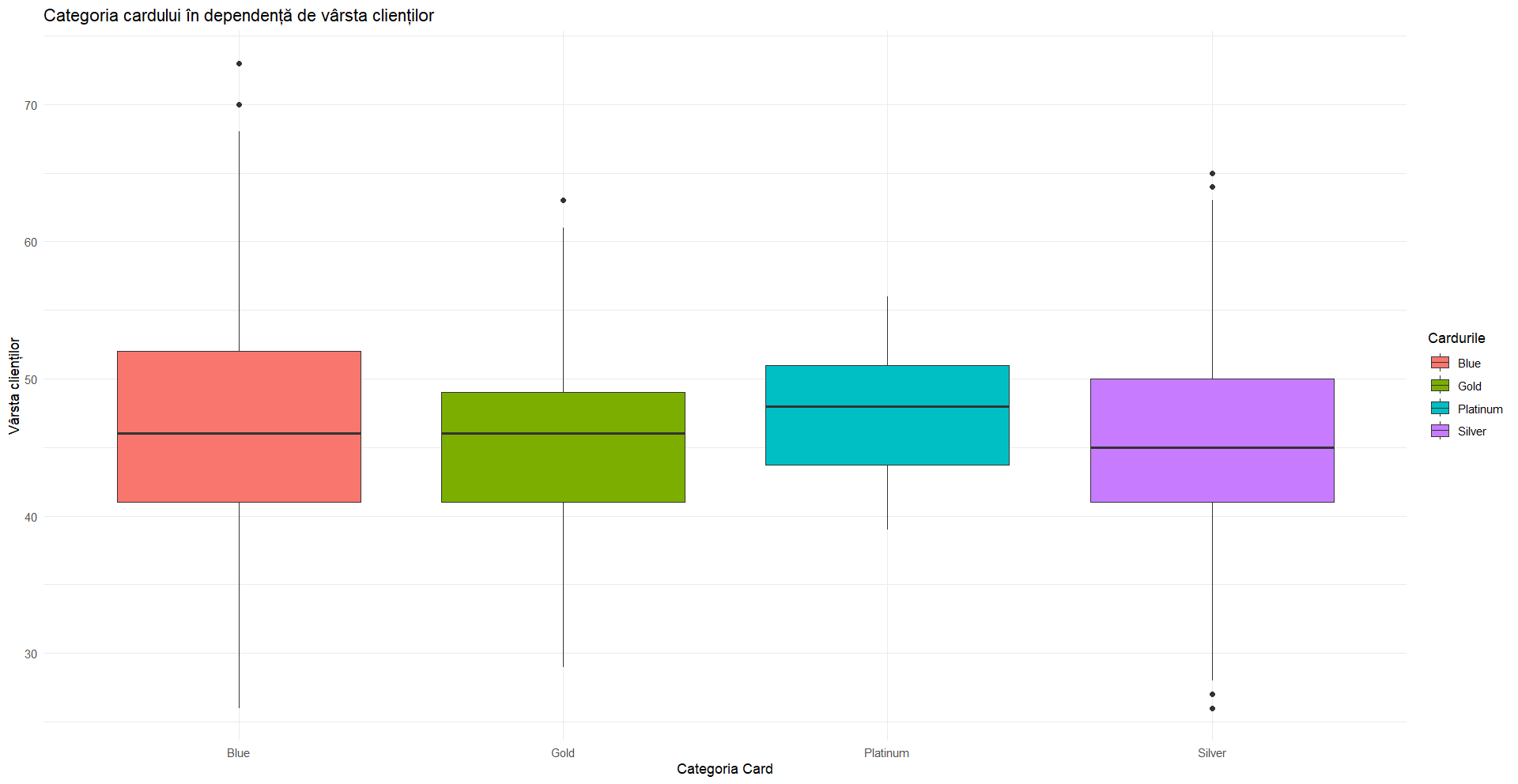
**Figura 3** Distribuția Vârstei Clienților în funcție de Categoria de Venit

Această figură arată distribuția vârstei clienților în funcție de categoriile de venit, precum și de categoria de carduri pe care aceștia le dețin. Pe axa orizontală este reprezentată vârsta, iar pe axa verticală este reprezentat numărul de clienți. După cum se poate observa, clienții cu venituri mai mari tind să fie mai în vârstă decât clienții cu venituri mai mici. De exemplu, în categoria de venit $120K+, majoritatea clienților au vârsta cuprinsă între 28 și 30 de ani. Această tendință se datorează faptului că persoanele cu venituri mai mari au tendința de a câștiga mai multă experiență în carieră și de a obține promoții.

****

**Figura 4** Plotul cu densitatea pentru Categoriile de card și veniturile acestora

În figura 4 putem observa graficul care ne oferă o perspectivă vizuală asupra densității care reprezintă frecvența relativă a categoriilor alese. Precum categoriile de carduri și nivelurile de venit. Acesta ne oferă o imagine vizuală asupra distibuției categoriilor de carduri în rândul clienților,în funcție de nivelurilor lor de venit.

****

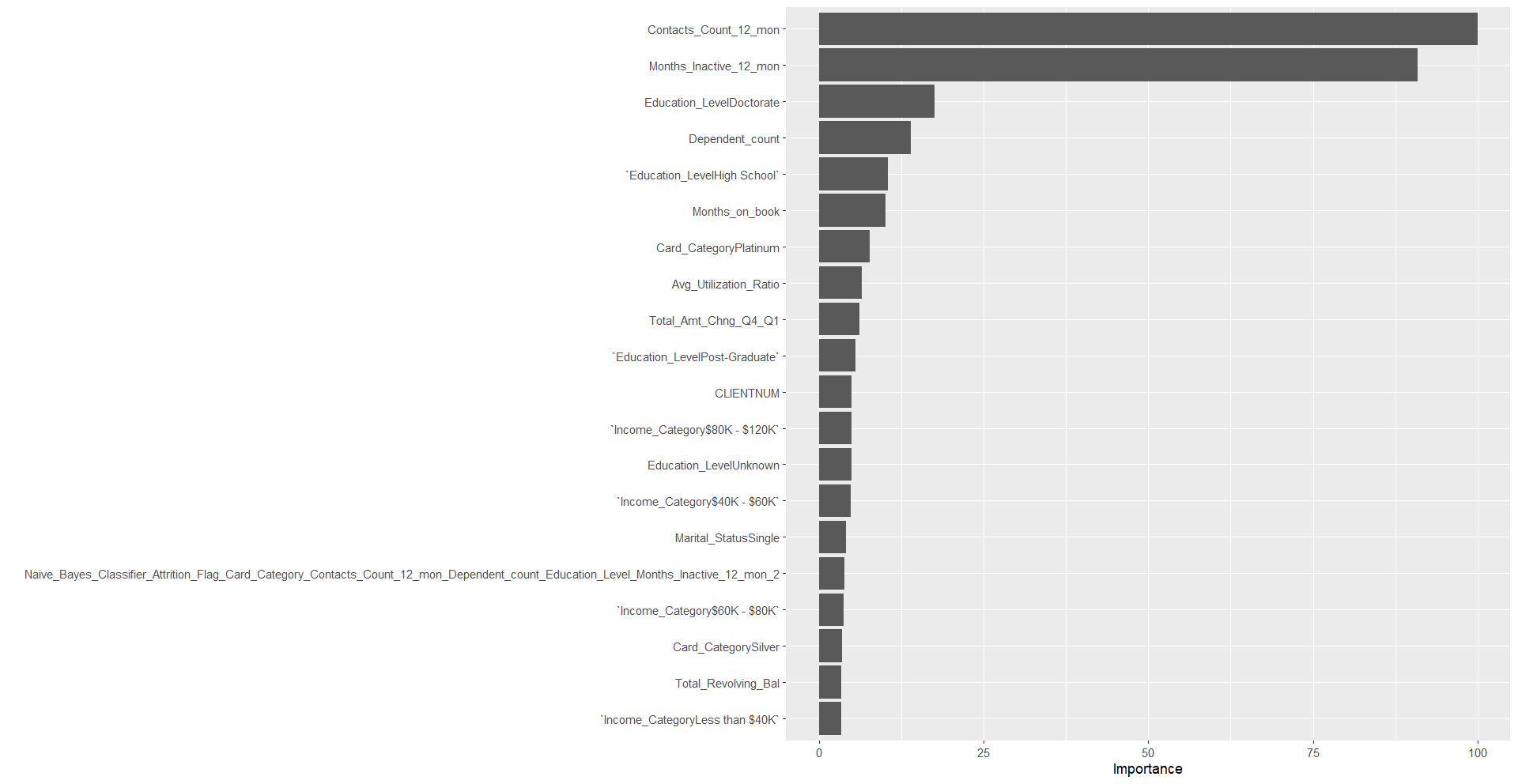
**Figura 5** Categoria cardurilor în dependeță de vârsta clienților

Această figură arată distribuția categoriilor de carduri în funcție de vârsta clienților. Pe axa orizontală este reprezentată vârsta, iar pe axa verticală este reprezentată procentul de clienți din fiecare categorie de carduri. După cum se poate observa, majoritatea clienților au carduri de categorie Blue. Această tendință se datorează faptului că cardurile Blue sunt cele mai accesibile. Cardurile Blue oferă beneficii minime, dar nu au taxe anuale sau taxe de înlocuire a cardului. Pe măsură ce vârsta clienților crește, procentul de clienți cu carduri de categorie Gold și Silver crește. Această tendință se datorează faptului că persoanele cu vârste mai mari au tendința de a avea venituri mai mari. Cardurile Gold și Silver oferă beneficii mai mari decât cardurile Blue, cum ar fi cashback, puncte de loialitate sau asigurare. Doar vârstnicii au carduri de categorie Platinum. Această tendință se datorează faptului că cardurile Platinum sunt cele mai scumpe.

**Tabel 1** Confusion Matrix and Statistics

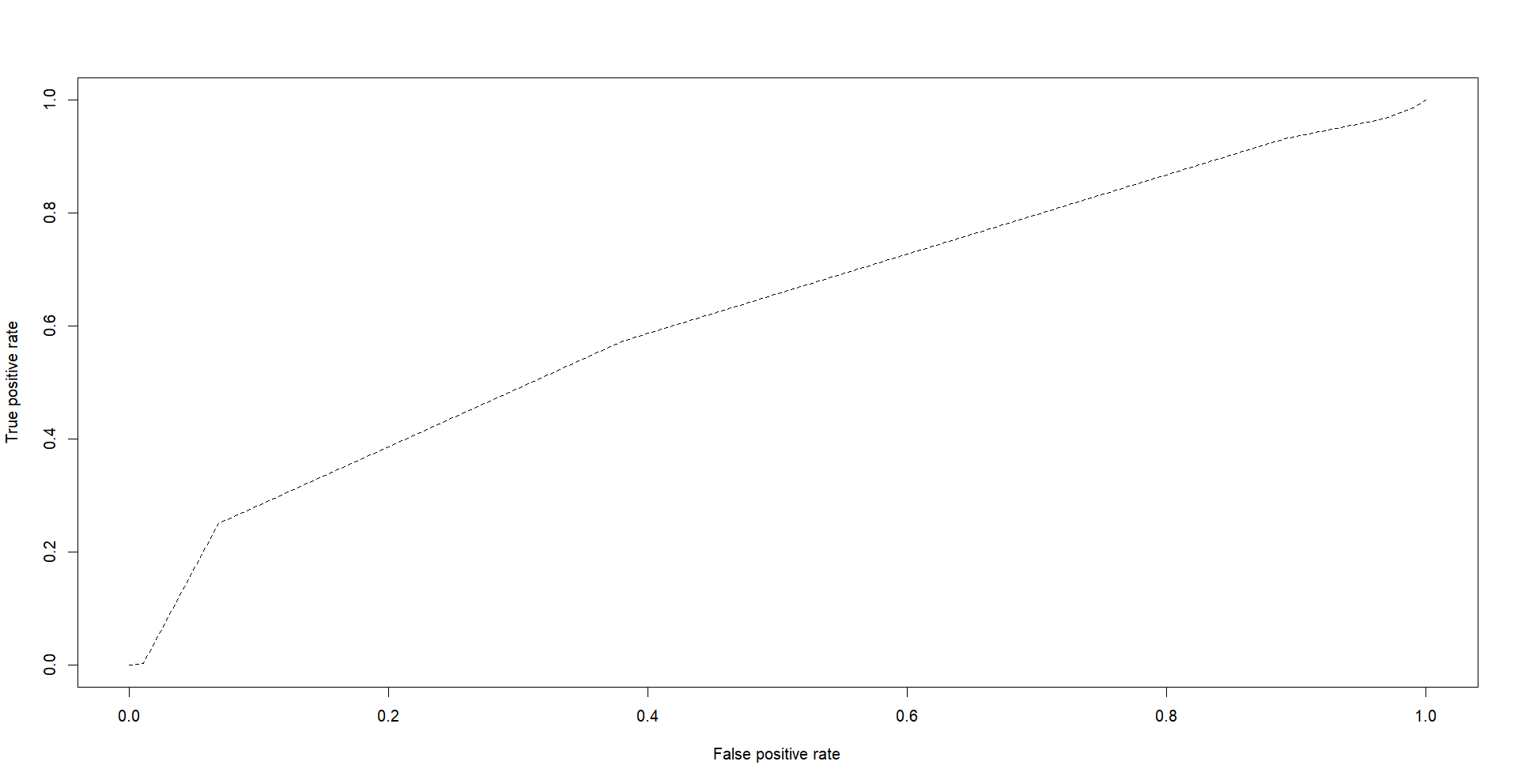
|  |  |  |
| --- | --- | --- |
| Confusion Matrix and Statistics | | |
| Reference | | |
| **Prediction** | **Yes** | **No** |
| **Yes** | 2796 | 151 |
| **No** | 542 | 3599 |
|  | | |
| **Acuracy** | 0.90 |  |
| **95%CI** | (0.8951, 0.909) |
| **No Information Rate** | 0.5291 |
| **P-Value [Acc> NIR]** | <2.2e-16 |
| **Kappa** | 0.8025 |
| **Mcnemar’s Test P-Value** | <2.2e-16 |
| **Sensitivity** | 0.8376 |
| **Specificity** | 0.9597 |
| **Pos Pred Value** | 0.9488 |
| **Neg Pred Value** | 0.8691 |
| **Prevalence** | 0.4709 |
| **Detection Rate** | 0.3945 |
| **Detection Prevalence** | 0.4158 |
| **Balanced Accuracy** | 0.8987 |
| **Positive Class** | yes |

În matricea de confuzie, numărul de instanțe corect clasificate este indicat de diagonala principală. În acest caz, diagonala principală este formată din 100 de celule, ceea ce înseamnă că toate instanțele au fost clasificate corect. Acuratețea este o măsură simplă, dar eficientă, care ne poate oferi o idee generală despre performanța modelului. În acest caz, acuratețea de 90% ne arată că modelul este foarte bun la clasificarea instanțelor.

****

**Figura 6** Importanța variabilelor ce influențează rata de părăsire

După procesare noi am obținut o frecvență destul de bună pentru a putea crea modelul de regresie logistică.

****

**Figura 7** Regresia Logistică pentru rata de părăsire

Analiza de regresie logistcă ne-a demonstrat că Contacts\_Count\_12\_mon și Months\_Inactive\_12\_mon sunt factorii cei mai importanși care influențează pentru rata de părăsire a clienților pentru serviciu de card de credit. Putem observa desigur că și în Matricia de confuzie(vezi tabelul 1) am obținut o bună performanță, cu o acurateție de 90%. Astfel putem spune că rata de părăsire a clienților în privința seriviciului card de credit este foarte mare.

**Bibliografie**

1. Documenție pentru limbajul R - [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)
2. Dataset Credit Card Customers - <https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers/data> – accesat (20.09.2023)
3. Regresia logistică - <https://ro.education-wiki.com/7539563-logistic-regression-in-r> – accesat (15.11.2023)

**Materiale Suplimentare**

1. **CLIENTNUM:** Număr unic de identificare al clientului.
2. **Attrition\_Flag:** Variabilă țintă; indică dacă un client a părăsit sau nu serviciile de card de credit (valori posibile: 'Existing Customer' sau 'Attrited Customer').
3. **Customer\_Age:** Vârsta clientului.
4. **Gender:** Genul clientului ('M' pentru masculin, 'F' pentru feminin).
5. **Dependent\_count:** Numărul de persoane dependente.
6. **Education\_Level:** Nivelul de educație al clientului.
7. **Marital\_Status:** Starea civilă a clientului.
8. **Income\_Category:** Categoria de venituri a clientului.
9. **Card\_Category:** Categoria cardului (Blue, Silver, Gold, Platinum).
10. **Months\_on\_book:** Numărul de luni de când clientul este client.
11. **Total\_Relationship\_Count**: Numărul total de produse de banking deținute de client.
12. **Months\_Inactive\_12\_mon:** Numărul de luni în care clientul a fost inactiv în ultimele 12 luni.
13. **Contacts\_Count\_12\_mon:** Numărul de contacte ale băncii cu clientul în ultimele 12 luni.
14. **Credit\_Limit:** Limita de credit pe card.
15. **Total\_Revolving\_Bal:** Soldul total al creditului rotativ.
16. **Avg\_Open\_To\_Buy:** Media disponibilă pentru cumpărături.
17. **Total\_Amt\_Chng\_Q4\_Q1:** Modificarea procentuală a sumelor totale de tranzacționare.
18. **Total\_Trans\_Amt:** Suma totală a tranzacțiilor în ultimele 12 luni.
19. **Total\_Trans\_Ct:** Numărul total de tranzacții în ultimele 12 luni.
20. **Total\_Ct\_Chng\_Q4\_Q1:** Modificarea procentuală a numărului total de tranzacții.
21. **Avg\_Utilization\_Ratio:** Raportul mediu de utilizare a creditului.