**Prognozarea și analiza pierderii clienților din cadrul companiei bazate pe comerț electronic**

**Inculeț Iulian**

Faculatae Calculatoara, Informatică și Microelectronică

Universitatea Tehnică a Moldovei

Chișinău,

Email: iulian.inculet@gmial.com

**Munteanu Viorel**

Faculatae Calculatoara, Informatică și Microelectronică

Universitatea Tehnică a Moldovei

**Abstract**

Acest articol descrie studiul efectuat în domeniul comerțului electronic pentru a analiza problema de pierdere a clineților (churn) care este esențială din cauza unei concurnețe în permanentă creștere și a chieltuielilor enorme pentru atragerea clienților loiali. Acestă cercetare include un studiu de analiză efectuat asupra datelor ce caracterizează clienții și interacțiunea acestora cu domeniul de comerț electronic în urma căruia a fost detectat că bărbații sunt mai predispuși să renunțe, având în vedere că avem 64% de bărbați care au renunțat la aplicație, clienții căsătoriți sunt cel mai mare segment de clienți în comapnie, iar cei mai mulți clienți care pleacă sunt predominant din categoria celor ce sunt singuri, etc. Frcvența datelor categoriale au indicat unele tendințe de bază ce predomină în setul de date, prin care cel mai des folosit dispozitiv este telefonul mobil aproximativ 74%, pentru 50% din clienți cele ma preferate categorii de comenzi sunt Telefoanele mobile și Laptopuri și accesorii. Prezența unei plîngeri indică o rată de pierdere mai mare a clienților. Scorul de satisfacere mai mare a clientului nu indică o loialitate a acestuia, nici nu influențeză la numărul de comezi. Clientul cu o vechime mai mare și o distanță mai mică de la depozit spre casă este mai loial. De asemnea pentru a prezice pierderea clienților a fost necesar de preprocesarea datelor ce include imputarea valorilor lipsă folosind mediana, curățarea datelor aberante și transformarea datelor categoriale, urmate de scalarea acestora. Pentru datele primite au fost implementate unele modele de învățare automată precum regresia logistică cu o precizie de 0.77, Support Vector Machine cu precizia de 0.80, Rețele neuronale cu precizia de 0.79. Cel mai bun model evaluat după metrica de precizie este Random Fores care are 0.90.

**Introducere**

Clienții sunt cele mai importante active pentru orice afacere. Păstrarea unui client fidel este de șase ori mai ieftină decât recrutarea de noi clienți.[1] Predicția pierderii clienților poate ajuta companiile să își rafineze obiectivul strategiei de retenție pentru a minimiza pierderile și de a îmbunătăți deciziile de marketing.

Companiile se concentrează pe predicția ratei de dezabonare pentru a delibera care sunt punctele de dezabonare a clienților și să taie cauzele în mod preventiv. În special, industriile care sunt avantajoase pentru a stabili relații pe termen lung cu clienții, cum ar fi companiile aeriene, sectorul bancar, comerțul electronic.[2]

Ca răspuns la problema de mai sus, cercetările existente privind pierderea clienților includ în principal predicții legate de statisticile tradiționale, inteligența artificială, predicțiile bazate pe învățarea statistică. Această lucrare propune cercetări privind comerțul electronic, predicția ratei de dezabonare a clienților pe baza segmentării clienților, utilizând SMOTE îmbunătățit pentru echilibrul datelor, iar apoi utilizarea a patru algoritmi diferiți de învățare automată pentru predicție.

Proiectul este important pentru prezicerea și analiza pierderii clienților ce va permite companiilor de comerț electronic să anticipeze și să determine ce clienți sunt susceptibili de a migra.

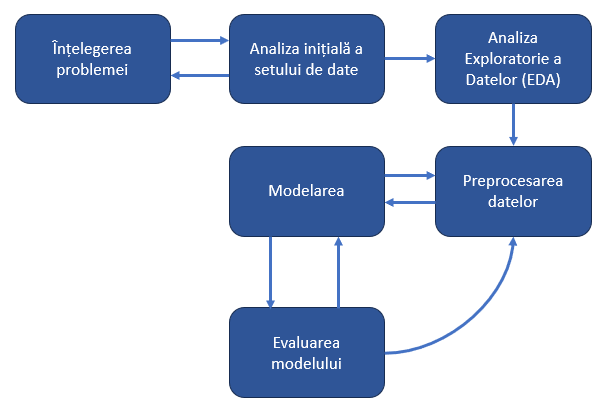
Pentru aceasta, mai întîi am efectuat o pre-procesare a datelor, inclusiv, sumarizarea, verificarea calității și eliminarea valorilor liplă. Ulterior, au fost analizate distribuțiile și anomaliile din setul de date. Analiza exploratorie a datelor (AED) efectuată în R [3] a arătat că pierderea clienților este influențată de unele caraterestii precum Genul, Vechimea clientului, depunerea plîngerii și altele. În final, pentru prezicerea probabilității pierderii clienților am testat modele precum Regresia Logistică, Random Forest. Rezultatele evaluării modleului au arătat că modelul Random Forest este cel mai potrivit având o precizie de 95% în comparație cu regresia logisitcă care are 89%. Pentru acest studiu au fost folosite limbajul R[4] și python[5].

**Date și metode**

În cadrul cercetării vor fi efectuate etapele ciclice care încep cu înțelegerea afacerii în care identificăm problema și necesitatea de soluționare a problemei. Apoi, trecem la înțelegerea datelor care sunt structurate într-un format Excel care conține 5630 de observații a clienților unei companii de comerț electronic și 20 de variabile.

După aceea, trebuie să pregătim setul în care datele colectate vor trece prin diferite etape inclusiv curățarea datelor prin eliminarea valorilor lipsă, a valorilor aberante și a coloanelor inutile.

În cele din urmă, vom încheia cu modelarea, evaluarea și implementarea, unde vom implementa modelele și le vom testa pentru a le măsura acuratețea. În figura 1 este prezentat procesul de lucru care a fost urmat în acest articol.



**Fig1.** Fluxul de lucru

**Descrierea setului de date studiat**

Datele sunt colectate pentru acest proiect din Kaggle pentru un site de comerț electronic. Datele de consum ale clienților care au achiziționat bunuri de pe site-ul web sunt selectate pentru analiză și predicție. Setul de date conține înregistrările de consum ale clienților precum și interacțiunile comportamentale a clienților ce utilizează platforma.

Datele sunt colectate de pe o platformă de comerț electronic de top, sunt date istorice care conțin detalii și experiențe ale clienților, iar rezultatul lor este un indicator de pierdere a clienților (pierdut = 1, loial = 0). Setul de date prezintă peste 5 000 de clienți și interacțiunea și preferințele acestora în cadrul platformei. Eficiența acestor date este că ele conțin anumite atribute detaliate specifice care vor ajuta la segmentarea clienților, cum ar fi: dispozitivul de conectare preferat, scorul de satisfacție și alte attribute (vedeți Tabelul 1).

Aceste atribute ne vor ajuta la studierea cauzelor de dezabonare în segmentul fiecărui client pentru a identifica factorii declanșatori care duc la dezabonarea clienților.

Tabelul 1. Descrierea setului de date utilizat în acest studio.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descrierea** | **Tipul de date** |
| CustomerID | Id-ul utilizatorului | Numeric |
| Churn (variabila dependenta) | Churn Flag | Numeric |
| Tenure | Vechimea clientului în companie | Numeric |
| PreferredLoginDevice | Dispozitivul de conectare preferat al clientului | Caracter |
| CityTier | Nivelul orașului | Numeric |
| WarehouseToHome | Distanța dintre depozit și domiciliul clientului | Numeric |
| PreferredPaymentMode | Metoda de plată preferată de client | Caracter |
| Gender | Genul clientului | Caracter |
| HourSpendOnApp | Numărul de ore petrecute pe aplicația mobilă sau pe site-ul web | Numeric |
| NumberOfDeviceRegistered | Numărul total de dispozitive este înregistrat pe un anumit client | Numeric |
| PreferedOrderCat | Categoria de comandă preferată de client în ultima lună | Caracter |
| SatisfactionScore | Punctaj satisfăcător al clientului privind serviciul | Numeric |
| MaritalStatus | Starea civilă a clientului | Caracter |
| NumberOfAddress | Numărul total de adaosuri adăugate la un anumit client | Numeric |
| Complain | A fost depusă vreo plângere în ultima lună | Numeric |
| OrderAmountHikeFromlastYear | Creșterea procentuală a comenzilor față de anul trecut | Numeric |
| CouponUsed | Numărul total de cupoane a fost folosit în ultima lună | Numeric |
| OrderCount | Numărul total de comenzi a fost plasat în ultima lună | Numeric |
| DaySinceLastOrder | Ziua de la ultima comandă a clientului | Numeric |
| CashbackAmount | Cashback mediu în ultima lună | Numeric |

**Analiza descriptivă a setului de date**

Pentru acest articol am utilizat analiza descriptivă în scopul înțelegerii variabilelor și valorilor din setul de date studiat. Analiza descriptivă permite rezumarea și descrierea proprietăților primare ale unui set de date. Oferă informații vitale despre distribuția frecvenței datelor, tendința centrală, dispersia. [6]

Această metodologie permite să gestionăm cu valorile lipsă, pentru prezentul set de date acestea cuprind 1.6 % din toate valorile. Introducerea valorilor lipsă, folsoind media, sau înserarea valorilor din observațiile apropiate. De asemena analiza descriptivă ne-a permis să găsim date ironate cum ar fi valorile duplicante din variabila PreferredPaymentMode ce conținea date aberante, ca exemplu ”CC” are aceeași semnificare ca ”Credit Card”.

Un pas essential ce conține această metodologie este detectarea valorilor aberante, folosind diagramele de distribuție și scatterplot-urile. Identificarea și eliminarea acestora va conduce la omiterea erorilor posibile în prezicerea pierderii clinților folosin modele antrenate pe acest set de date.

**Exploratory data analysis (EDA)**

Am folosit EDA pentru a identifica relațiile dintre variabilele prezente cu variabila dependentă (Chunr). Analiza exploratorie a datelor se referă la procesul critic de efectuare a investigațiilor inițiale asupra datelor, astfel încât să se descopere modele, să se testeze ipotezele și să se verifice ipotezele cu ajutorul statisticilor sumare și a reprezentărilor grafice.[7] Prin urmare EDA mi-a permis să formez unele ipoteze precum:

1. Pierderea clienților este influențată de gen. Numărul de comenzi depinde de genul Clientului.
2. MaritalStatus influențează aspura pierderii clienților.
3. CityTier mai mare are un număr mai mare de comenzi, și clienți cu o vechime mai mare.
4. Prezența unei plângeri influențează la DaySinceLastOrder.
5. Pierderea clienților este influențată de dispozitiul de logare preferat.
6. Un scor de satisfacere este corelat cu numărul de comenzi efectuate.
7. Creșterea procentuală a valorii comenzilor față de anul trecut afectează rata de dezabonare.
8. Există o relație între categoria de comandă preferată și rata de dezabonare.

Testarea și verificare ipotezelor folsosind reprezentările grafice au oferit unele informații referitor la legătura dintre variabilele setului de date cu variabila ce descrie pierderea clientului.

În urma follosirii EDA am verificat toate ipotezele anterioare, și am primit o înțelegere mai completă asupra setului de date, legăturilor între aceste variabile și factorii de influență asupra plecării clienților.

**OneHotEncoder**

Datele folosite în acest studiu necesită modificări pentru a îmbunătăți acuratețea și calitatea de antrenare a modelelor. Prin umrae asupra datelor categoriare au fost efectuate transformări prin folosirea metodie OneHotEncodercare convertește datele pentru a le pregăti pentru un algoritm și a obține o predicție mai bună. Cu one-hot, convertim fiecare valoare categorică într-o nouă coloană categorică și atribuim o valoare binară de 1 sau 0 acestor coloane.

**MinMaxScaler**

Centrarea și scalarea se porduc independent pe fiecare caracterisitcă. Din motivul că setul de date conține multe variabile a căror valori este 0 și 1 iar alte variabile pot ajunge pînă la sute precum CashbackAmount, acestea pot distorsiona unii algoritmi de învățare automată, din motiv că aceștea acordă atenție valorilor mari. O modalitate de a normaliza caracteristicile/variabilele de intrare este scalerul Min-Max. În acest fel, toate caracteristicile vor fi transformate în intervalul [0,1], ceea ce înseamnă că valoarea minimă și maximă a unei caracteristici/variabile va între 0 și, respectiv, 1.

**Regresia logisitcă binară**

Regresia logistică binară modelează relația dintre un set de variabile independente și o variabilă dependentă binară. Este utilă atunci când variabila dependentă este de natură dihotomică, cum ar fi moartea sau supraviețuirea, pierderea sau loialitatea clientului.[8] Variabilele independente pot fi categorice sau continue, de exemplu, sexul, modul de plată, numărul de comenzi sau vechimea clientului. Regresia logistică binară modelează o variabilă dependentă ca un logit de p, unde p este probabilitatea ca variabilele dependente să ia valoarea 1. Acest model a fost folosit pentru studiu pierderii cleinților din domeniului telecomunicării care a primit un rezultat de 85.25%[9].

Funcția standard de regresie logistică, pentru prezicerea rezultatului unei observații în funcție de o variabilă predictivă (x), este o curbă în formă de s definită ca p = exp(y) / [1 + exp(y)] (James et al. 2014). Aceasta poate fi, de asemenea, scrisă simplu ca p = 1/[1 + exp(-y)], unde:

* y = b0 + b1\*x,
* exp() este exponențialul
* p este probabilitatea ca evenimentul să se producă (1) având în vedere x. Matematic, aceasta se scrie ca p(eveniment=1|x) și se prescurtează ca p(x), deci px = 1/[1 + exp(-(b0 + b1\*x))]`

**Random forest**

Acest algoritm a fost introdus de un cercetător numit Tin Kam Ho. Conceptul algoritmului Random Forest este de a crea o predicție cu privire la pădurile necorelate de arbori, iar predicția este cunoscută ca fiind mai precisă în comparație cu orice tip de arbore individual. În articolul ce se referă la prezicerea pierderii clienților folosind Random Forest au menționat că Random Forest utilizează caracterul aleatoriu al caracteristicilor și bagging-ul, ori de câte ori există o cerință de creare a unui arbore individual din păduri de arbori necorelate.

În plus, cercetătorii au propus să utilizeze clasificatorul Random Forest împreună cu Support Vector Machine pentru calcularea ratei de pierdere a clienților. În cadrul studiului s-a arătat că algoritmi menționați au sporit acuratețea la 95%.[10] Prin urmare, făcându-i un instrument mai potrivit și mai eficient pentru calcularea pierderilor de clienți.

**Rețele neuronale**

Rețelele neuronale artificiale constau din diferite straturi de neuroni artificiali interconectați, alimentați de funcții de activare care ajută la activarea/dezactivarea acestora. La fel ca algoritmii mașinilor tradiționale, și aici există anumite valori pe care rețelele neuronale le învață în faza de instruire. Fiecare neuron primește o versiune multiplicată a intrărilor și a ponderilor aleatorii, la care se adaugă apoi o valoare de polarizare statică (unică pentru fiecare strat de neuroni); aceasta este apoi transmisă unei funcții de activare corespunzătoare care decide valoarea finală care va fi dată de neuron. [11]

În cadrul acestei cercetări va fi folosită o rețea neuronală de tip sigmoid care va avea 4 nivele, primul nivel va conține 30 neuroni, urmat de 12, 10 și la nivelul final este doar un neuron care va returna rezultatul y prezis. Toate straturile în afară de ultimul folosește funcția de activarea Rectified linear activation function (ReLU).

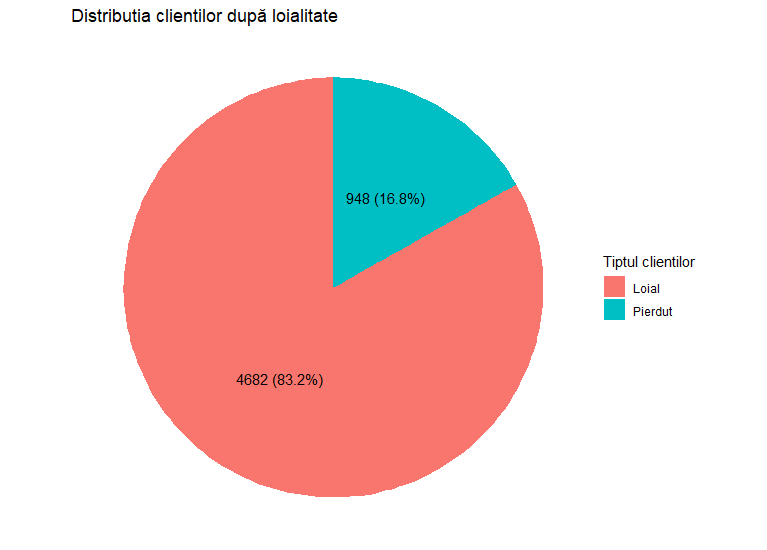
**Support vector machine**

O mașină vectorială de suport (SVM) este un algoritm de învățare automatizată supravegheată care clasifică datele prin găsirea hiperplanului care separă în mod maxim clasele diferite într-un spațiu cu dimensiuni mari. Acesta realizează acest lucru prin identificarea vectorilor de suport, care sunt punctele de date aflate cel mai aproape de limita de decizie, și prin optimizarea hiperplanului pentru a maximiza marja dintre clase, sporind capacitatea modelului de a generaliza la date noi. [12]

**Rezultate și discuții**

**Rezultatele analizei exploratorii a datelor**

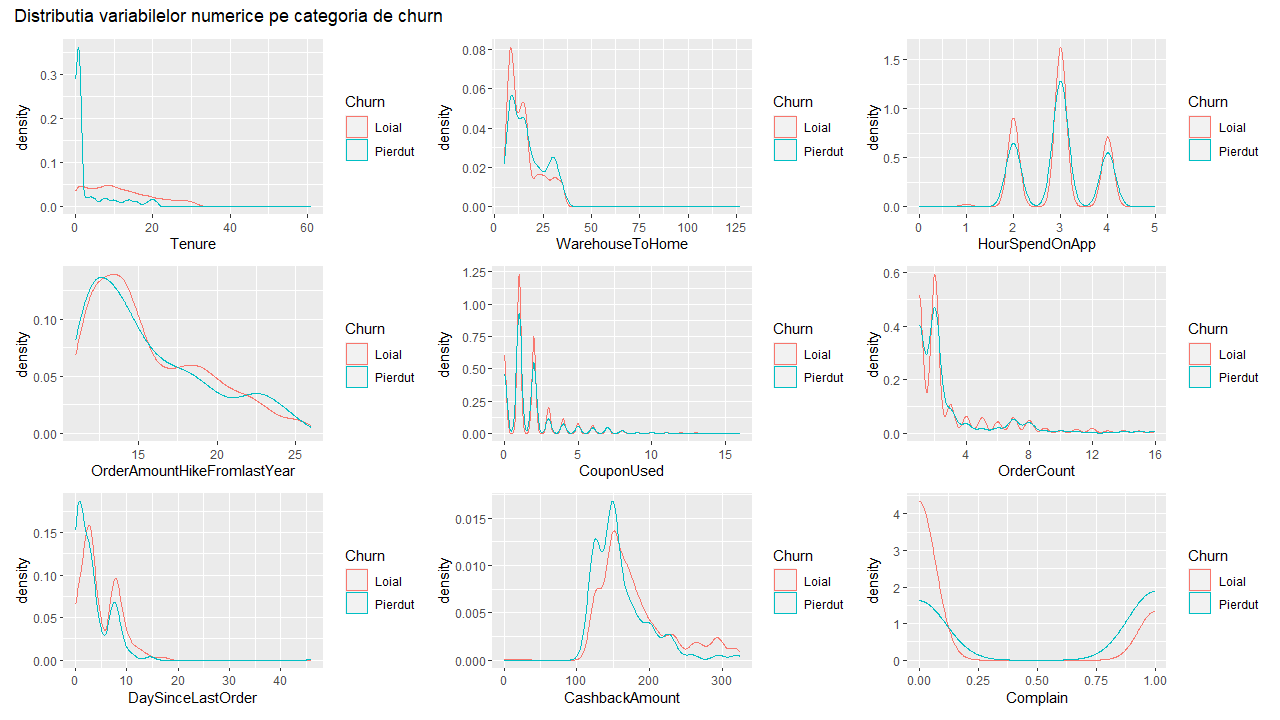
Prezicerea eficentă a pierdererii clienților este direct proporțională cu calitatea setului de date de care este folosit pentru antrenare modelelor de învățare automată. Prin urmare pentru a putea primi un set de date curat este necesar de efectuat o analiză a acestuia și de determinat erorile, datele aberante, existența datelor lipsă și analiza relațiilor între caracteristici.



**Fig 1.1 Clienți loiali vs pierduți**

Cum este descris în figura 1.1 putem observa că majoritate observațiilor din setul de date este compusă din clienți loiali ce va îngreunătăți capacitatea modelelor de a prezice cu cercitudinecare clienții ce sunt susceptibili de a pleca.

Pentru a înțelege mai bine realția dintre caracteristicile clienților și starea acestuia de a pleca sau a fi loial, în continuare o să prezint rezultatele variabilelor numerice în dependență de loialitatea clientului.



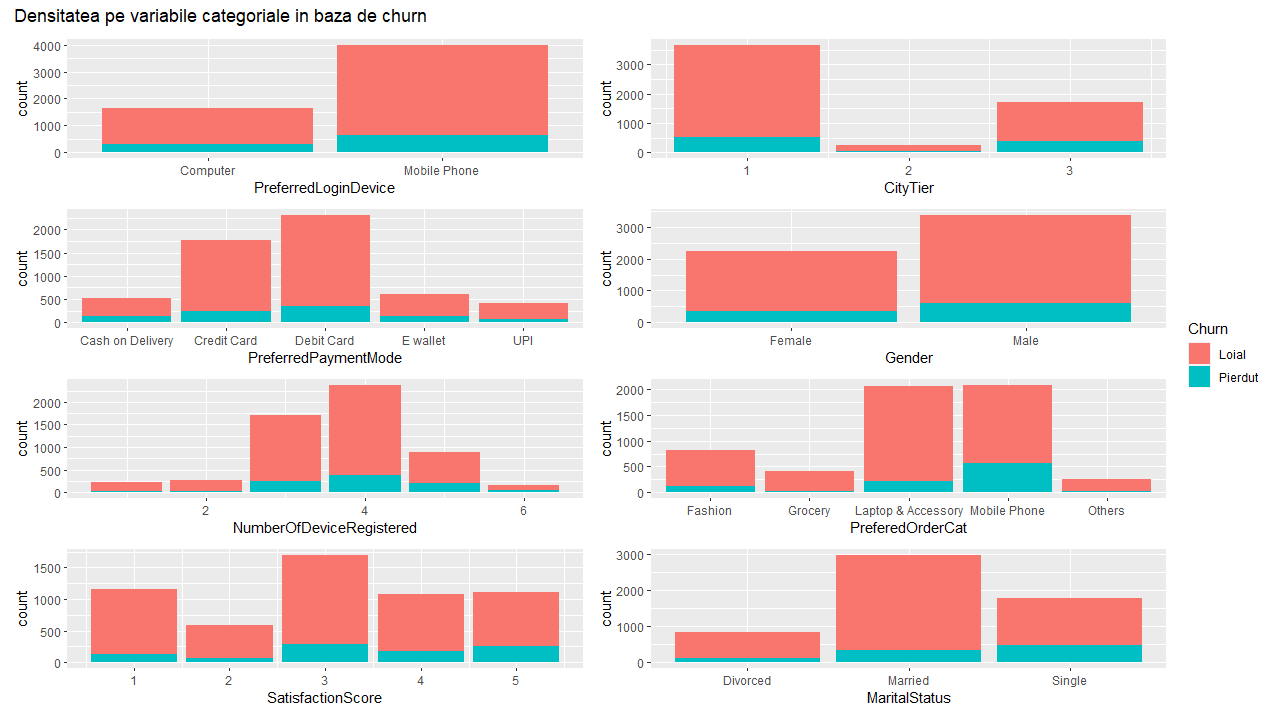
**Fig 1.2 Distribuția variabilelor numerice în dependență de churn**

Prin analiza diagramelor de distribuție din figura 1.2 în care se descrie densitatea valorilor pentr caracteristicile indicate, în raport cu starea clientului, care poate fi loială sau definit ca pierdut. Prin urmare primim următoarele rezultate observaționale pentru caracteristici:

* Tenure(vechimea clientului): Clienții cu o vechime mai mare par mai puțin predispuși la dezabonare. Este logic, deoarece o vechime mai mare indică satisfacție.
* WarehouseToHome (distanța depozit-casă): Distanțele mai scurte între depozit și domiciliu au o rată de renunțare mai mică. Livrările mai rapide pot îmbunătăți satisfacția.
* HourSpendOnApp (utilizarea aplicației ore): Numărul de ore a clienților în aplicație nu implică un impact semnificativ asupra clienților, prin care se poate de observat că clienții loiali și cei pierduți în general folosesc același număr de ore.
* OrderAmountHikeFromLastYear (creșterea valorii comenzilor față de anul trecut): Clienții care au cheltuit în medie mai mult sunt mai loiali, însă de asemnea se poate de observat că clienții pierduții predomină în intervalul de creștere a valorii de 22.5 pînă la 25, ceea ce nu poate fi un predictor bun.
* CouponUsed (cuponul folosit): Utilizarea cupoanelor se corelează cu o rată de renunțare mai mică. Însă se poate de observat că atît clineții loiali cît și cei pierduți pot folosi un număr mare de cupoane.
* OrderCount (număr de comenzi): Un număr mai mare de comenzi se asociază cu o rată de dezabonare mai mică. Utilizarea frecventă creează obiceiuri. Tot odată și unii clienți pierduții au un număr de comenzi care este mai mare de 10.
* DaySinceLastOrder (ziua de la ultima comandă): O perioadă mai lungă de la ultima comandă se corelează cu o rată de dezabonare mai mare. În diagramă însă se observă o diferență slabă deoarece și clienții loiali au un număr de zile care pot fi până la 20 de zile ca și cei loiali.
* CashbackAmount (Suma rambursată): O umă de rambursare mai mare indică o frecvență mai mare de clienți loiali, în timp ce clientii pierduții primesc o sumă de rambursare mai mică decât 200.
* Complaint (Plîngere): Depunearea plîngerii poate fi asociată cu o rată de pierdere a clienților mai mare. Depunere unei plîngeri poate să fie un bun predictor de pierdere a clienților.

În rezultat putem indica o legătură slabă între rata de peirdere a clientului și carateristicile: plîngere, vechima clientului, suma rambursată, numărul de comenzi, numărul de zile de la ultima comanda.

Pentru analiza variabilelor categoriale se va folosi reprezentarea grafică prin barplot-uri.



**Fig 1.3 Densitatea variabilelor categoriale în bază de churn**

Prin analiza diagramelor de densitate din fugura 1.3 se poate de observat care dintre valorile categoriale indică o rată de pierdere mai mare a clienților, după analiza diagramelor putem obține următoarele rezultate:

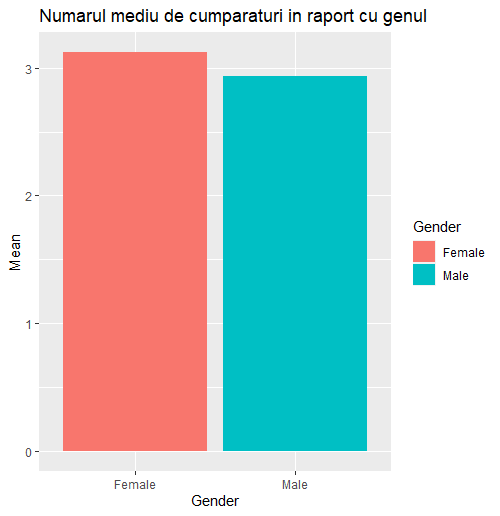
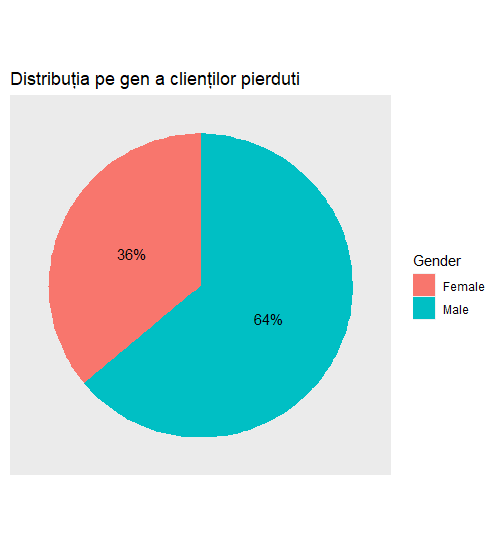
* PreferedLoginDevice (device-ul de logare preferat): Cît clienții loiali atît și cei pierduți folosesc telefonul mobil pentru conectare, însă se poate de asemena de observat că numărul de clienți care folosesc telefonul mobil sunt mai susceptibili pentru a pleca.
* CityTir (tir-ul orașului): Putem observa că cel mai frecvent tir al orașului esste 1. De asemena putem observa că clienții pierduții sunt în orașe cu Tirul 1 și 2.
* PreferedPaymentMode (modul preferat de plata): Clienții pierduți cel mai des folosesc Debit Card, și Credit Card urmat de E-wallet.
* NumberOfDeviceRegistered (numărul de dispozitive înregistrate): Clienții cel mai des au 4 dispozitive conectate , urmate de 4 și respectiv 5 dispozitive. Cleinții care au 4 dispozitive sunt mai predispuși să plece.
* PrefferedOrderCat (Categoria de comenzi preferate): Cele mai populare sunt categoriile de Telefoane mobile și laptop și accesorii. Aceste categorii conțin însă și un număr mare de clienți pierduți.
* Scorul de satisfacere: Cel mai frecvent scor de satisfacere este 3. Un scor de satisfacere mai mare nu indică o asociere srânsă cu loialitatea clienților.

Prin umrare nu se poate de determinat o asociare strînsă între caracteristicile descrise anterior și dorința clientului de a pleca. Se observă doar o tendință a clienților de a pleca în caz că aceștea au ca PreferedLoginDevice – telefonul mobil, Gender- Bărbat, PrefferedOrderCat – Mobile Phone, NumberOfDeviceRegistered – 4 și MaritalStatus – Single.

După explorarea setului de date, am identificat anumite tendințe și modele care indică posibile asocieri între variabile. Îmi propun să formulez și să analizez ipoteze specifice pentru a investiga mai departe aceste relații și a aduce o înțelegere mai detaliată a factorilor care influențează rezultatele observate:

1. Există o relație între gen și rata de dezabonare? Ce gen are mai multe comenzi?
2. Care MartialStatus are cea mai mare rată de dezabonare?
3. Care CityTier are un nivel mai ridicat de Tenure și OrderCount?
4. Este vre-o legătură între scorul de satisfacere și orele în aplicație?
5. Care este relația dintre Complain și DaySinceLinceLastOrder?
6. Există o relație între PreferredLoginDevice și Churn?
7. Există o legătură între scorul de satisfacție și numărul de comenzi?
8. Există o relație între categoria de comandă preferată și rata de dezabonare?

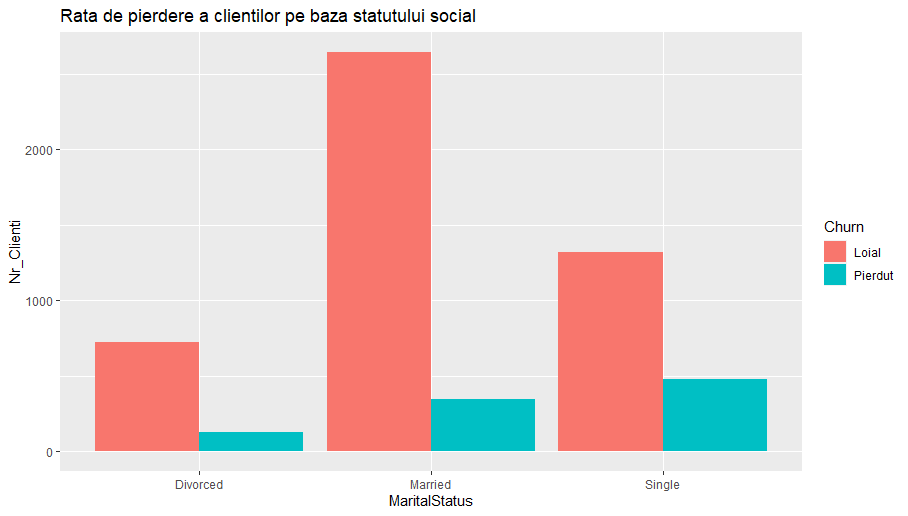
Rezultatele ipotezeilor indicate mai presus sunt următoarele:

**Fig 1.5 Analiza variabilei de gen**

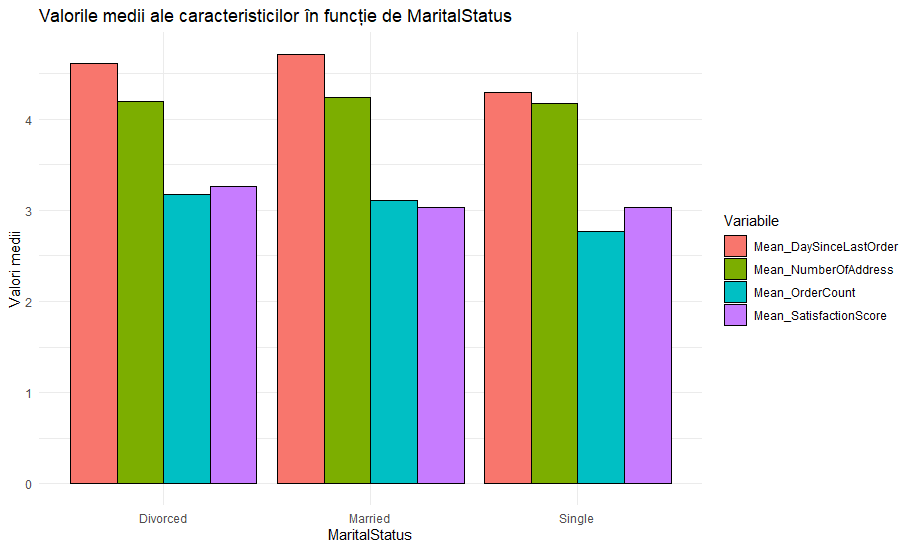
În setul de date sunt 3384 de bărbați și 2246 de femei, prin care se observă că totuși numărul de comenzi a femeilor este de 3.1 iar a bărbaților de 2.9.

Bărbații conform figurii 1.5 sunt mai predispuși să renunțe, având în vedere că avem 64% de bărbați care au renunțat la interacțiunea cu compania. Acest lucri descrie o tendință a bărbaților de a suspenda folosirea serviciilor companiei, deci se poate confirma o asociere între gen și loialitate clienților.



**Fig 1.6 Analiza statutului social**

În diagrama de mai sus se poate de observat asociere între statutul social și rata de plecare a clientului. Clienții căsătoriți sunt cel mai mare segment de clienți în comapnie, iar cei mai mulți clienți care pleacă sunt predominant din categoria celor ce sunt singuri. Prin urmarea clienții singuri, și care sunt în căsătorie sunt mai susceptibili de a pleca în comparație cu cei divorțați.

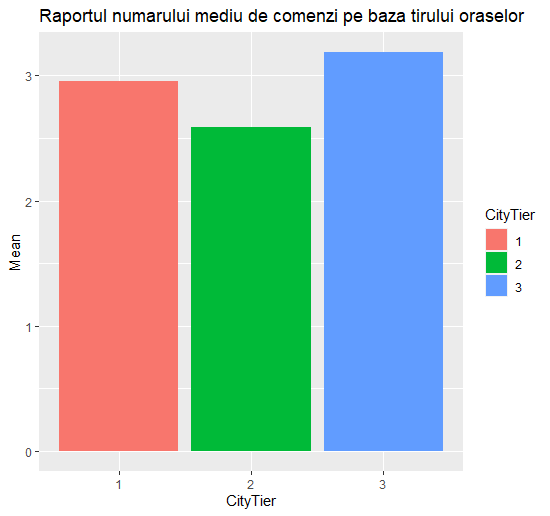
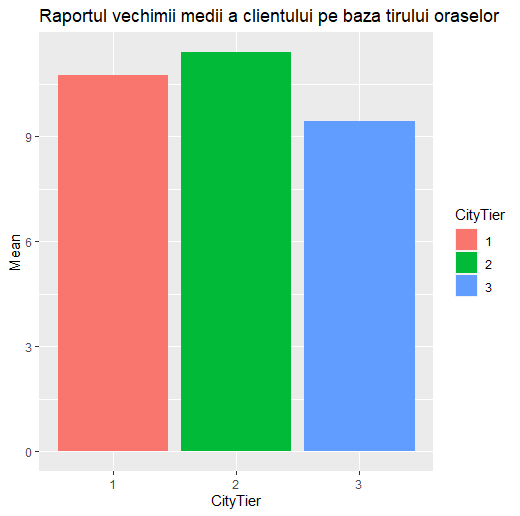


**Fig 1.7 Analiza statutului social**

Din moment ce cel mai mare număr din clienți fac parte din categoria căsătorit, este necesar de analizt ce impact are acestă categorie în raport cu clelelalte. Prin care o să putem să determinăm dacă clienții căsătoriții trebuie să fie condisderați mai valoroși, deoarece aceștea pot duce un impact mai mare supra numărului de comenzi, numărului de aderesării ceea ce este desris în figura 1.7.

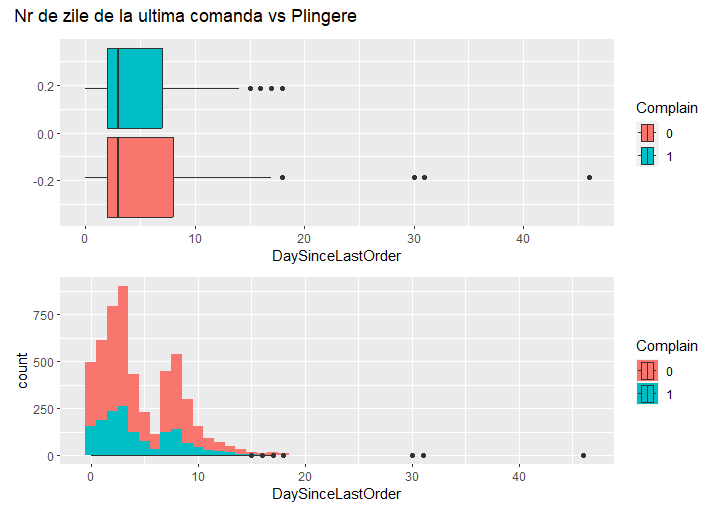
În acestă figură se observă variații în aceste valori în funcție de statutul marital. Diagrama este un grafic cu bare colorate care reprezintă valorile medii ale diferitelor variabile în funcție de statutul marital. Există trei categorii pe axa X: Divorced, Married și Single. Pe axa Y sunt reprezentate valorile medii ale caracteristicilor. Există patru seturi diferite de date reprezentate prin culori distincte: verde pentru “Mean\_DaysSinceLastOrder”, roșu pentru “Mean\_NumberOfAddresses”, albastru pentru “Mean\_OrderCount” și mov pentru “Mean\_SatisfactionScore”.

Prin urmare se poate observa că în medie după numărul de comenzi clienții divorțați și căsătoriți au valori similare, urmat cu o diferență minimă de aproximativ 0.20 de cei singur. De asemena numărul de zile de la ultima comandă în mediu este mai mare pentru clienții căsătoriți urmat de cei divorțați și singuri. Numărul de adresări este aproximativ similar pentru fiecare categorie, ce descrie că acestă caracteristică nu este influențată de statutul marțial. Prin urmare putem deduce că chiar și dacă cei mai mulți clienți sunt căsătoriți acesta nu influențeză semnificativ la numărul de comenzi și de adresări, la scorul de satisfacere și la numărul de zile de la ultima comandă.



**Fig 1.8 Analiza variabilei cityTier**

Clienții care au vochime în mediu mai sus de 9.5 sunt din orașe cu tirul 2 însă numărul de comenzi în mediu efectuate de acești clienți este de 2.5 care este cea mai mică în rapor cu cele late tire a orașelor. Prin umrare nu este vizibilă o dependență între tirul orașelor vechimea clienților în aceste orașe și numărul de comenzi. Clienții din tirul orașelor 1 are în mediu o vechime de 9.6, din 2 9.8 iar din 3 este de 9.1. Iar numărul comenziolor de asemena deviază de la 2.6 pînă la 3.2. Prin care nu poate fi observată o legătură.



**Fig 1.9 Relația plîngere – Nr. Zile de la ultima comandă**

Clienții ce nu au depus plîngeri au în mediu un număr de zile de la ultima comandă mai mică decât cei care au depus o plîngere. Clienții care nu au făcut plingeri au un DaySinceLastOrder mai mare, însă este vorba de un singur client, deci este o valoare aberantă, dacă îl eliminăm, clienții care nu au făcut reclamații vor avea un DaySinceLastOrder mai mic.

Prin urmare în urma efectuării EDA putem deduce următoarele rezultate:

* Nu există diferență între numărul de comenzi mediu pe gen.
* Bărbații sunt mai predispuși să renunțe, având în vedere că avem 64% de bărbați care au renunțat la aplicație.
* Cei căsătoriți reprezintă cel mai mare segment de clienți din cadrul companiei, iar cei singuri sunt cei mai predispuși să renunțe la aplicație.
* Citytier 2 are cea mai mare vechime a clienților, dar aceasta nu pare a fi un factor important.
* Citytier 3 are cea mai mare valoare medie a comenzilor, dar nu pare să fie un factor important în ceea ce privește renunțarea la clienți.
* Există o relație negativă slabă între complainig și numărul de zile de la ultima comandă.
* Utilizatorii de telefoane mobile sunt mai predispuși la dezabonare.
* Atunci când procentul de comenzi din ultimul an crește, rata de dezabonare scade, astfel încât OrderAmountHikeFromlastYear are un efect pozitiv asupra ratei de dezabonare și trebuie să ne concentrăm asupra clienților cu procente între 12% și 15%.
* Reclamațiile nu afectează rata de dezabonare, clienții care au făcut reclamații în proporție de 68% nu fac dezabonare.

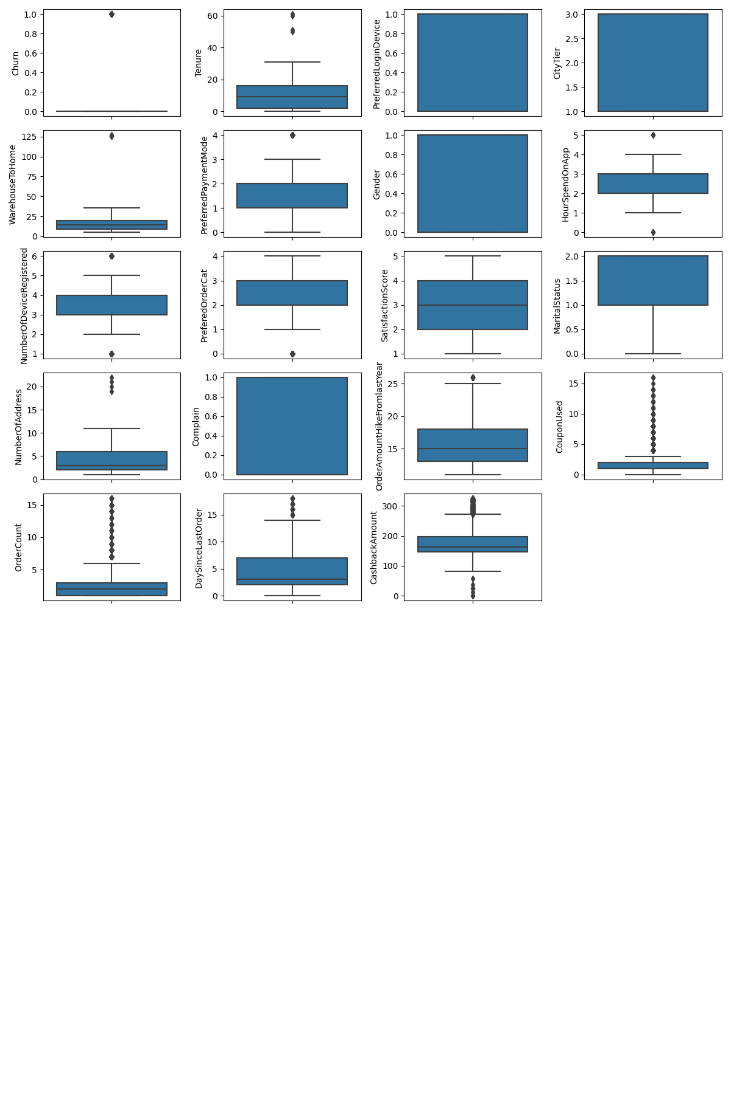
Obiectivul principal a fost de a face distincția între clienții ratați și cei reținuți, în plus față de găsirea atributelor asociate care conduc la dezabonare. La început, s-a observat că bărbații singuri au o probabilitate ușor mai mare de a renunța. În plus, categoria preferată de telefoane mobile este, de asemenea, legată de dezabonarea clienților. În plus, clienții care au renunțat la serviciile de client sunt ușor mai mari în cazul dispozitivului de conectare preferat de telefon mobil. De asemenea, s-a constatat că churned au o medie mai mare în ceea ce privește reclamația, nivelul orașului Numărul de adrese și numărul de dispozitive înregistrate. Cu toate acestea, studiul nostru arată că scorul de satisfacție este mai mare la clienții pieruți ceea ce nu era de așteptat. Pe de altă parte, Tenure, și numărul de comenzi este mai mic pentru clienții churned, ceea ce este rezonabil.

**Imputarea valorilor lipsă**

În setul de date au fost indetificate date lipsă pentru caracteristicile: Tenure, WarehauseToHome, HoursSpenOnApp, OrderAmountHikeFromlastYear, CouponUsed, OrderCount, DaysSinceLastOrde. Totalitatea datelor lisă din aceste caracteristi constituia într 4.6% și 5.5% prin umrare aceste valori au fost imputate folosind mediana deoarece valoarile mediei și medianei sunt similare pentru fiecare variabilă, prin urmare datele introduse vor fi mai corectecte, din motiv că unele variabile numerice sunt întregi, predcum CouponUsed, OrderCount, Complain, etc. Prin urmare toate valorile lipsă au fost completate, fără a influența la distribuția caracteristicilor.

**Eleiminarea dateleor aberante**

În umrma efectuării EDA au fost observate prezența datelor aberante ce pot infulența la acuratețea antrenării modelelor. Prin urmare acestea au fost omise prin folosire interavlului intercuanilelor care a lăsat doar datele ce sunt poziționate între cuantila q1 – IQR \* 1.5 și cuantila 3 q3 + IQR \* 1.5.



**Fig 1.10 Idnetificarea datelor aberante**

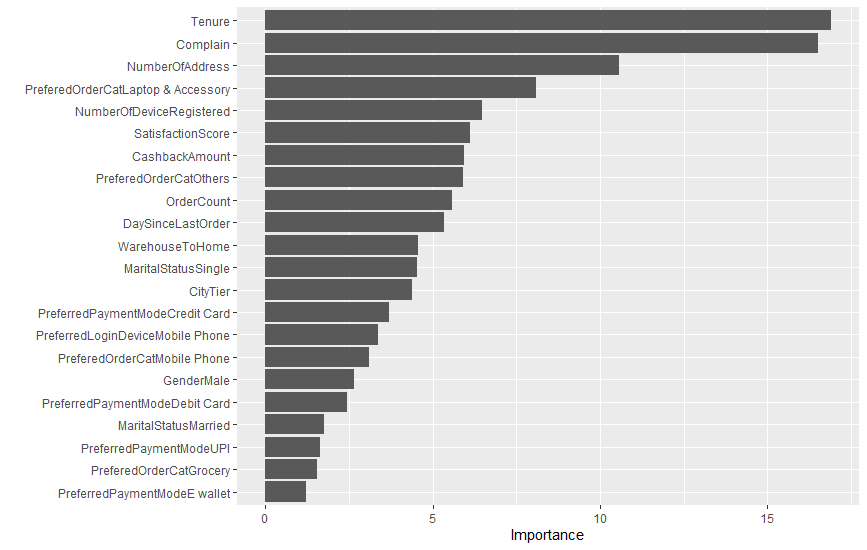
În fiugra 1.10 pot fi observate datele aberante prezene în caracteristicile precum: Tenure în care sunt prezente 2 valori care sun mai mari decît 40, de asemena se poate observa o distanță depozit-casă care este mai mare decât 125 de kilometri, dinte care restul obseervațiilor ajung maxim pînă la 50. De asemena au fost eleminate datele din caracteristica DaysSinceLastOrder prin prezența datelor mia mari de 40 de zile, ceea ce a fost detectat din analiza EDA. Prin urmare datele au devenit mai curate și pot fi folosite mai eficent în antrenarea modelelor.

**Transformare datelor**

Datele categoriale au fost transformate prin metoda OneHotEncoding prin care 19 variabile inițiale au ajuns la 30, din motivul creării pentur fiecare variabilă precum: PreferredPaymentMode, PrefferedOrderCategory, PrefferedDeviceUsed cîte o colană adițională pentru viecare valoare. Prin urmare acesta a crescut numărul de variabile, dar a permis folosire acestor variabile categoriale în antrenarea modelelo. De asemena am folosit SacalareaMinMax pentru a transforma valorile fiecării variabile într-un interval între 0 și 1.

**Selectarea caracteristiclor**

Prin utilizarea metodie stepAIC care este una dintre cele mai frecvent utilizate metode de căutare pentru selectarea caracteristicilor. Încercăm să continuăm să minimizăm valoarea stepAIC pentru a obține setul final de caracteristici. "stepAIC" nu înseamnă neapărat îmbunătățirea performanței modelului, însă este utilizat pentru a simplifica modelul fără a avea un impact semnificativ asupra performanței. Astfel, AIC cuantifică valoarea pierderii de informații datorată acestei simplificări. AIC înseamnă "Akaike Information Criteria". Cu ajutorul biblioteilor bootStepAIC și MASS au fost eliminate variabilel: HourSpendOnApp, CouponUsed, OrderAmountHikeFromlastYear.



**Fig 1.11 Importanța caracteristicilor în antrenarea modelului**

Pentru a antrena un model eficient, este nevoie de selectat cele mai importatne caracteristici prin care varaibila dependentă (Churn) este descrisă cel mai bine. Prin rezultatel primite (vezi fugura 1.11) putem observa că ce mai mare infulență o are vechimea clientului, prin care cu cît acesta are mai puține luni interacționate cu compania, acest este mai susceptibil să plece. Plângerea de asemnea este un factor esențial, prezența acesteia mărește șansa clientului de a pleca. Numărul de adresări este a treia careacteristică după importanță, prin care adresările mai dese indică un factor de loialitate a clientului. Caracteristicile care de asemena au o importanță majoră sunt: categoriea preferată, numărul de dispozitive logate, scorul de satisfacere, caschback-ul primit etc.

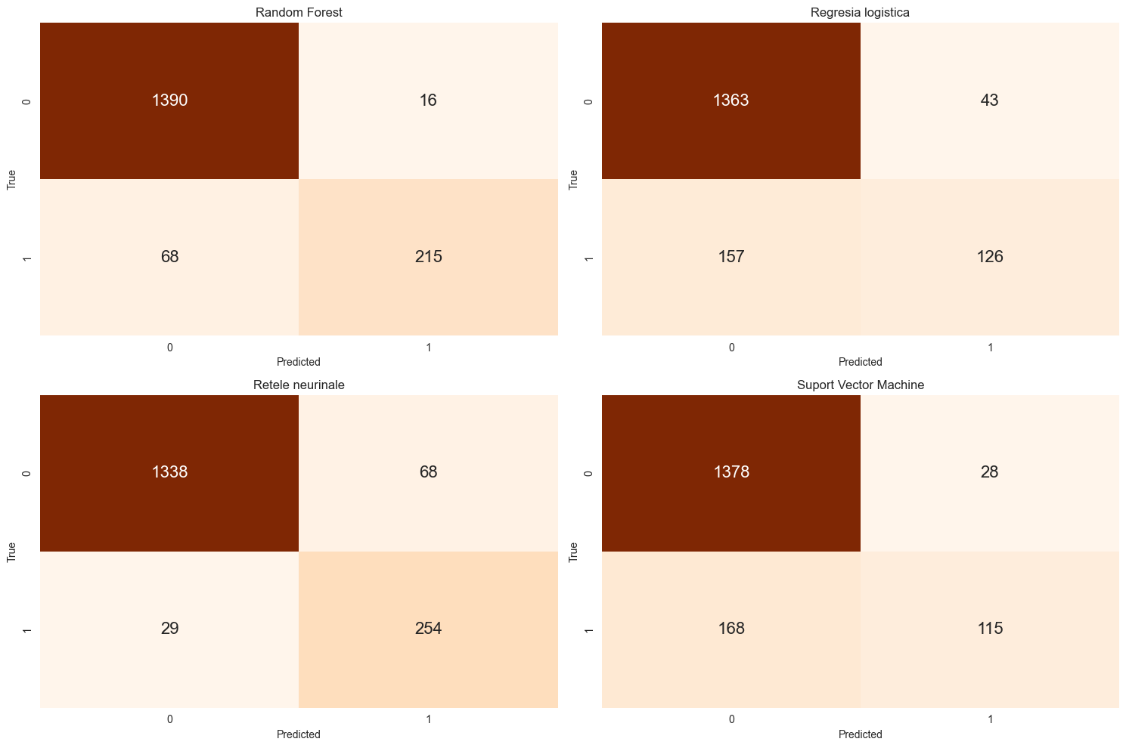
**Crearea Modelelor**

Acest studiu a folosit 4 modele diferite pentru prezicerea pierderii clienților. RandomForest, Regresia logisitcă, rețelele neuronale și support vector machine sunt modele ce au fost folosinte în acest studiu. Pentru antrenarea și testarea acestora, setul de date a fost împărți în set de antrenare care constituie 70% din tot setul de date inițial. Stului de testare îi revine respectiv 30%. Pentru foloisrea modelelor au fost folosite biblioteci python precum sklearn, dar pentru rețeaua neuronală creată individual a fpst folosită bibliotecile keras cu tensurelfow, prin care a fost creat o rețea neuronală multistrat.

**Evaluare modelelor**

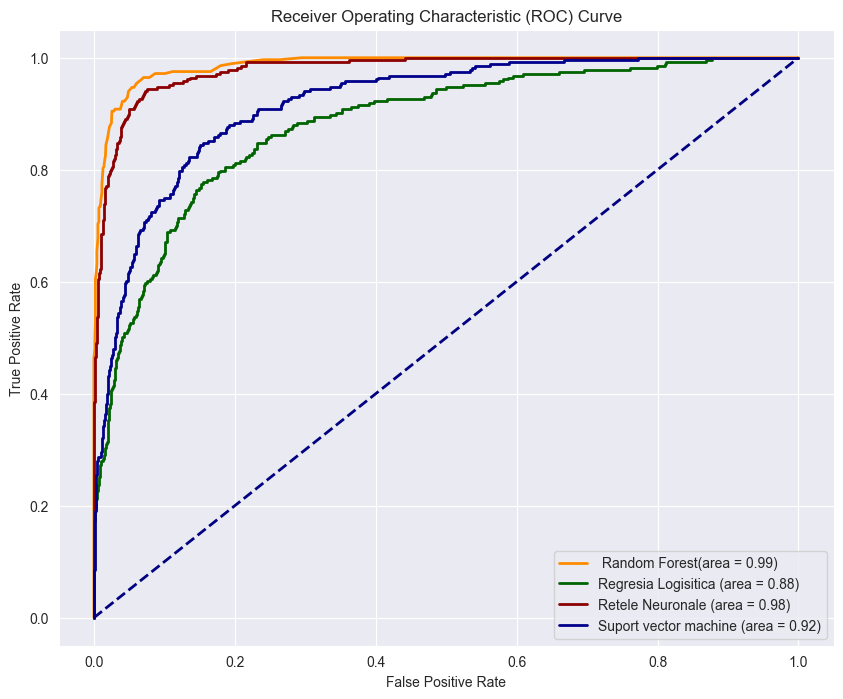
În această secțiune se va compara performanța diferitelor clasificatoare pentru a selecta cea mai bună metodă de predicție a ratei de dezabonare a clienților. Rezultatul predicției modelului este analizat utilizând valorile preciziei, rechemării, scorului F și AUC, care sunt descrise în paragrafele următoare. Aceste valori vor fi comparate cu cele 4 modele.

Prin urmare una pentru evaluare modelelor inițial se vor folosi matricile de confuzie în care se va descrie acuratețea de prezicere a fiecărui model.



**Fig 1.12 Matricile de confuzie**

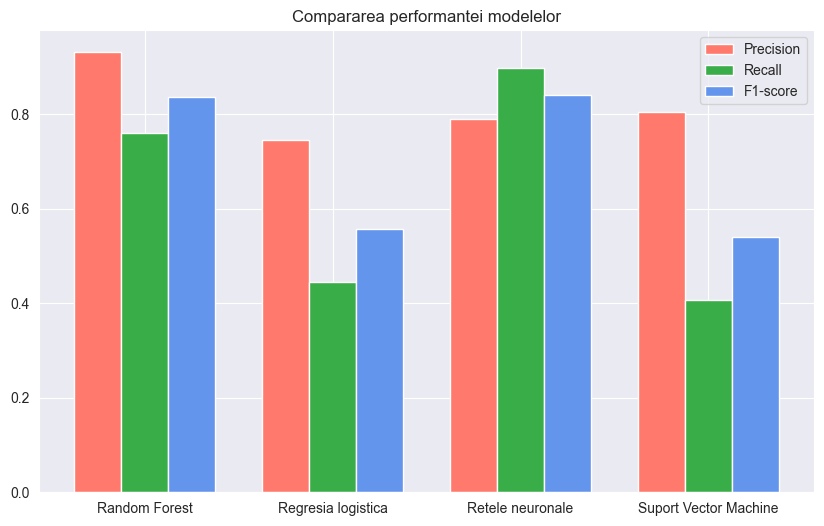
Analizând matricea de confuzie din fiugra 1.12 putem observa că Rețelele Neuronale[13] au detectat corect 254 de clienți pierduțu în raport cu 29 care au fost indicați greșit, însă acest model are o rată de validitate mai mică asupra clienților loiali. Următorul model ce a prezis eficent clienții pierduți a fost Random Forest ce a detctat corect 215 de clienți pierduți însă numărul de detectarea Fals pozitiv este de 68 ce determină o tendință a modelului de obține anumite errori în detectarea clienților pieruți. Acest model însă prezice cu o acuratețe mai mare cleinții loiali, Regresia logistică și Support vecto machine au o rată de false negativ mai mare decît true negative ce descrie incapacitatea acestora de a prezice eficent clienții ce sunt susceptibili să plece. Aceste modele nu satisfac obiectivelor propuse.



**Fig 1.13 Curbe ROC**

Curba ROC arată performanța unui clasificator binar cu praguri de decizie diferite. Acesta trasează rata reală pozitivă (TPR) față de rata fals pozitivă (FPR). Scorul ASC ROC este aria de sub curba ROC. Acesta rezumă cât de bine un model poate produce scoruri relative pentru a distinge între instanțele pozitive și negative în toate pragurile de clasificare. Prin urmare putem observa o perfomanță majoră a modelelor Random Forest cu aria de 0,99 și Rețele neuronale cu area de 0.98, prin urmare aceste model au o distingere mare a claselor pozitive și negative. Prin urmare ptin evaluarea ratelor pozitive și negative a acestor 2 model, putem vizualia o distingere puternică efectuată de model pentru aceste 2 clase idiferent de neegalitate de date. Modelele de SVM și Regresia Logistică au o performanță bună însă oricum mai slabă relativ cu modele descrise mai sus, avîn aria de 0.92 și 0.88.

Din motiv că principalul obiectiv a acestui articol a fost prezicerea clienților pierduți vom evalua modele după precizia, recall și f1-score pentru clasa 1 (client pierdut) pentru a evalua performanța modeului de identificare a acestora.

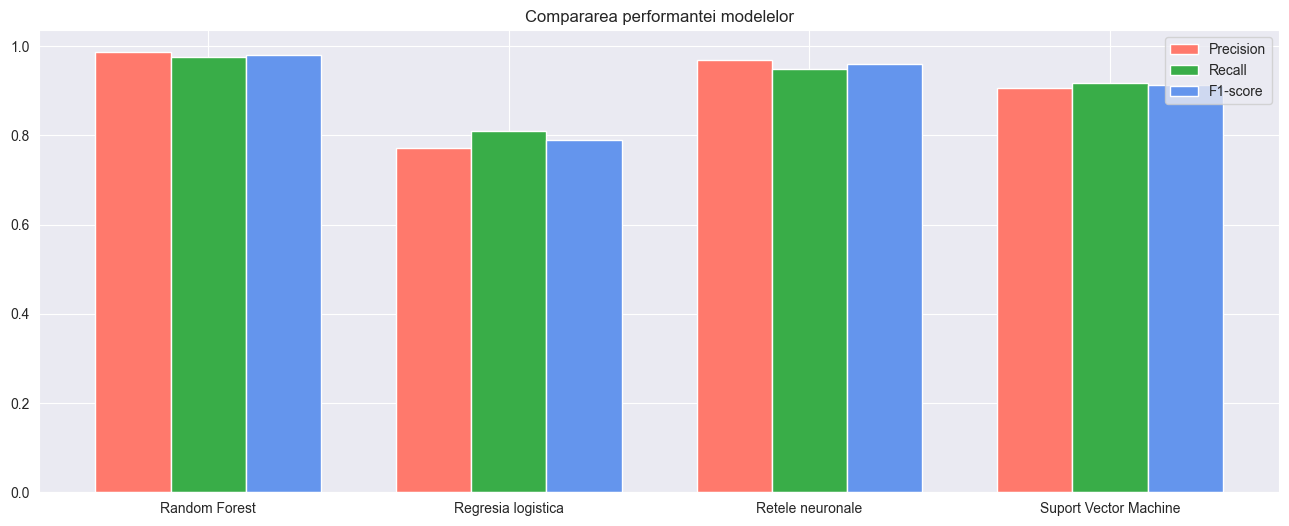
****

**Fig 1.14 Compararea performanței modelului**

Random Forest este un model ce are o precizie de 0.9 ce indică o capacitate mare a modelui de a efectua predicțiile pozitive și de a determina clienții ce sunt susceptibili pentru a pleca. Rețele neurinale de asemena au o precizie de 0.79, însă SVM are dobîndit o precizie mai mare, având 0.8. Regresia logistică are precizia doar de 0.77. Recall esteu o măsură ce ne permite să determinăm performanța modelui de detectarea a clienților ce pot fi susceptibili de a pleca chiar dacă aceștea nu pleacă, din motivul că pierderea unui client loial și înlocuirea cu atragerea alutuia nou este foarte costitsitor. Rețele neuronale au un recall de aproximativ 0.9, urmate de Random forest cu 0.78. Pentru regresia logisitcă și SVM valorile recall sunt de 0.44 și respectiv 0.4. F1 socr este o medie armonică între aceste 2 măsuir și indică aproximativ aceleași valori de 0.81 pentru Random Forest și Rețele neuronale. Prin urmare în cazul în care suntem orientați pentru determinare cît mai precisă a clienților susceptibili pentru plecare este cel mai eficent implementarea modelului de random forest, în caz că dorim detectarea cît mai mare a clienților pierduți chiar și dacă aceștea pot să fie loiali pentru un timp poate fi folosit modelul de Rețele Neuronale.

**Îmbunătățirea modelului prin SMOTETOmek**

Pentru rezolvarea inegalității în setul de date în care predomină observațiile clienților loiali, am implmentat metoda Smote care permite supraeșantionarea cu observații pentru clienții pierduți prin crearea unor observații sintetice noi, iar motoda tomek permite subeșantionarea prin eliminarea unor observații din clasa majoritară care sunt foarte similare cu cele din clasa minoritară prn care poate fi determinată o ambiguitate în capacitățile modelului de a efectua o calificare corectă.



**Fig 1.15 Performanța modelor după somtetomek**

Impementarea acestei metode a oferit posibilitatea de a mări dimensiunea setului de date pînă la 9000 de observații ce a stimulat perfomața modelelor. Prin care se poate de observat că modelul Random Forest are o precizie, recall și f1-scor aproximativ de 0.99 urmat de Rețele neuronale care are de asemnea toate măsurile aproximativ de 0.98-0.97. Pentur regresia logistică și SVM această tehnică de asemnea a îmbunățățit performanța pînă la 0.9 pentru toate metricile precum precizia, recall și f1-score pentru SVM și aproximative 0.8 pentru toate metricile pentru regresia logisitică.

Acestă metodă poate însă provoca o supraantrenre a modelelor pe baza unor date artificiale ce poat duce la unele greșeli în detectarea clienților pierduți bazate pe date adevărate. Din acest motiv este necesar de folosit cu acuratețe de acestă metodă.

**Concluzie și recomandări**

Întreprinderile de comerț electronic alocă sume uriașe de bani pentru a dobândi noi clienți. Cu toate acestea, durata de viață a clienților depinde de o mulțime de variabile, iar acest studiu a fost despre construirea modelelor de predicție a pierderii clienților pentru întreprinderile de comerț electronic. setul de date utilizat pentru acest proiect este pentru una dintre cele mai importante platforme de comerț electronic, care a fost preluat de pe Kaggle. Studiul a început cu analiză exploratorie și vizualizări de date pentru a ne spori înțelegerea clienților dezabonați.

S-a observat că clienții dezabonați sunt asociați cu sexul masculin, starea civilă, dispozitivul de logare preferat: telefonul mobil, și cateogira preferată: teleofoane mobile, laptop și accesorii. Apoi, au fost aplicați patru algoritmi diferiți pentru a prezice rata de dezabonare a clienților, care sunt Random Frorest, Rețele neuronale, regresia logisitică și support vector machine. S-a constatat că Random forest oferă o precizie de 0.90 și un recall de 0.78, iar aria AUC este de 0.99.

Pe baza analizei, s-a utilizat Random Forest pentru a afla semnificația fiecărei variabile. Prin urmare, cei 30 de predictori au fost sortați în ordine descrescătoare în funcție de importanța lor. S-a constatat că primii 10 predictori sunt Tenure, Complain, NumberOfAddress, MaritalStatus, CashBackAmount, DaySinceLastOrder, WarehouseToHome, PreferedOrderCat, OrderAmountHikeFromlastYear, NumberOfAddress care sunt asociați cu 90% din impactul asupra rezultatului. Astfel, acești atribute s-au dovedit a avea cea mai mare influență asupra ratei de dezabonare a clienților.

Câteva recomandări pentru proprietarii de întreprinderi rezultate din analiză sunt enumerate mai jos:

1. Afacerea trebuie să crească vechimea clienților lor, ceea ce se poate face prin inițierea unor programe de fidelizare sau prețuri speciale pentru clienții fideli.
2. Deoarece reclamațiile de asemnea au o importanță majoră, acestea trebuie să fie tratate cu atenție, iar organizația trebuie să se asigure că serviciul său de relații cu clienții este calificat în mod profesionist.
3. Compania trebuie să promoveze produsele care au un interes pentru genul masculin pentru a menține acestă categoire de clineți.
4. Chiar și dacă categoria preferată este telefoane mobile, urmată de laptop și accesorii în acest categorii predomină și o rată de pierdere a clienților mai mare prin urmare este necesar de stimult clienții care folosesc aceste categorii.
5. Scorul de satisfacere nu este un evaluator bun pentru rata de pierdere a clienților, din motivă că rata de pierdere a clienților cu un scor de satisfacere maximal este mai mare decît cea a clienților cu scorul de satisfacer 1. Prin umrare este necesar de implementat altă metodă de evaluare a interacțiunii clienților cu compania.
6. Distanța de la depozit spre casă are de asemena o influență majoră asupra loialității clienților, prin urmare pentru clienții care au o distnță mai mare este mecesar o tratare mai specială, în caz că companie are o intenție de menținere a acestora.

**Links**

**Disponibiliatea datelor**

Datele din acest studiu le găsiți la următorul link: **<https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/>**

**[Disponibilitatea codului](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/)**

**Codurile utilizate în acestă analiză le găsiți la următroul link:**

[**https://github.com/iulic7/ChurnEComerce/tree/main**](https://github.com/iulic7/ChurnEComerce/tree/main)

**Bibliografia**

[1] „(13) THE IMPORTANCE OF CUSTOMERS TO BUSINESSES TODAY | LinkedIn”. Data accesării: 26 noiembrie 2023. [Online]. Disponibil la: https://www.linkedin.com/pulse/importance-customers-businesses-today-oyindamola-olaniyan/

[2] B. Musumali, „An Analysis why customers are so important and how marketers go about in understanding their decisions.”, ian. 2019.

[3] F. Sohil, M. U. Sohali, și J. Shabbir, „An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7”, *Statistical Theory and Related Fields*, vol. 6, nr. 1, pp. 87–87, ian. 2022, doi: 10.1080/24754269.2021.1980261.

[4] „R Core Team (2023). \_R: A Language and Environment for Statistical Computing\_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.”

[5] „Van Rossum, G., & Drake Jr, F. L. (1995). Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.”.

[6] R. W. Cooksey, „Descriptive Statistics for Summarising Data”, *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, pp. 61–139, mai 2020, doi: 10.1007/978-981-15-2537-7\_5.

[7] „What is Exploratory Data Analysis? | IBM”. Data accesării: 27 noiembrie 2023. [Online]. Disponibil la: https://www.ibm.com/topics/exploratory-data-analysis

[8] P. Penman, „Binary Logistic Regression – An introduction”, Paul Penman. Data accesării: 27 noiembrie 2023. [Online]. Disponibil la: https://www.datascienceinstitute.net/blog/binary-logistic-regression-an-introduction

[9] H. Jain, A. Khunteta, și S. Srivastava, „Churn Prediction in Telecommunication using Logistic Regression and Logit Boost”, *Procedia Computer Science*, vol. 167, pp. 101–112, ian. 2020, doi: 10.1016/j.procs.2020.03.187.

[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, și S. W. Kim, „A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector”, *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[11] „What are Neural Networks? | IBM”. Data accesării: 13 decembrie 2023. [Online]. Disponibil la: https://www.ibm.com/topics/neural-networks

[12] R. Berwick, „An Idiot’s guide to Support vector machines (SVMs)”.

[13] „A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.”