Retenția Clienților priect învățare autonoma

Ionuț Marchiș

30 mai 2024

Cuprins

1 Introducere

Retenția Clienților (RC) sau rata de retenție clienților este una dintre cele mai importante caracteristici pentru orice companie care își dorește sa își evolueze și sa evalueze afacerea.

RC se referă la abilitatea unei companii sau a unui produs de a își păstra clienții pe o perioadă anume de timp. În cazul unei retenții mari, clienții unui produs sau ai unei afaceri tind să se întoarcă, să continue să cumpere, cu alte cuvinte, să nu se orienteze spre alt produs sau afacere, sau să nu înceteze să mai utilizeze produsul/afacerea respectivă. În general, organizațiile din domeniul vânzărilor încearcă să reducă numărul clienților care pleacă la concurență. Retenția clienților începe cu primul contact dintre organizație și client și continuă pe parcursul întregii relații, strategiile de retenție de succes luând în calcul acest ciclu de viață. Abilitatea unei companii de a atrage și de a păstra clienți noi este legată nu numai de produsul sau serviciile pe care le oferă, ci și de modul în care își tratează clienții pe care îi are deja, valoarea pe care clienții o generează ca rezultat al utilizării soluțiilor și reputația pe care o creează în general, în cadrul pieței.

Motivația din spatele creării unei astfel de aplicații este să ofere companiilor instrumentele necesare pentru a identificarea și abordarea factoriilor care contribuie la părăsirea clienților. Prin colectarea și analiza datelor relevante despre comportamentul clienților, interacțiunile cu produsele sau serviciile, se poate dezvolta o înțelegere mai profundă a motivelor care stau la baza plecării clientului. Această înțelegere poate servi la implementarea strategiilor proactive pentru reținerea clienților și îmbunătățirea experienței acestora.

2 Context

RC este importanta deoarece este mai costisitor să se obțină clienți noi decât să se mențina colaborarea cu clienți care deja folosesc serviciile sau produsele oferite. De fapt, o creștere a retenției clienților cu doar 5% poate genera o creștere de cel puțin 25% a profitului. Acest lucru se datorează faptului că utilizatorii care se întorc vor cheltui probabil cu 67% mai mult pentru produsele și serviciile companiei. Prin urmare, compania poate cheltui mai puțin pe costurile de operare pe care le implică achiziționarea de noi clienți.

Pentru acest proiect am ales un set de date ale unei companii de telecomunicații. Datasetul este preluat de pe site-ul Kaggle (https://www.kaggle.com/datasets/blastchar/telco-customer-churn). Coloanele din baza de date sunt: clienții care au plecat în ultima lună - coloana se numește Churn, serviciile la care fiecare client s-a înscris - telefon, linii multiple, internet, securitate online, backup online, protecție dispozitive, asistență tehnică și streaming TV/filme. Informații despre contul clientului - cât timp au fost clienți, contract, metodă de plată, facturare electronică, taxe lunare/totale și informații demografice despre clienți - gen, vârstă, parteneri.

Baza de date conține anumite informații despre clienți care ajuta sa se poate identifica RC. Aplicația trebuie să folosească un algoritm de învățare automata pentru a detecta persoanale care trebuie să fie prioritizate pentu reducerea costurilor companiei. Trebuie să se gasească un model care să dea ranament bun și să clasifice corect persoanale din baza de date. RC dintr-o companie de telecomunicații este o problemă de clasificare doarece rezultatul este unul de tip boolean (da sau nu).

3 Aspecte teoretice

Pentru a calcula rata de retenție a clienților, se utilizează această formulă: raportul dintre clienții pierduți(LostC) și totalul de clienți(TotalC) înmulțit cu 100.

$$RZ = \frac{LostC}{TotalC} \cdot 100 \tag{1}$$

Rata de retenție poate fi măsurată lunar, trimestrial sau anual. În timp ce rata lunară de retenție este utilă pentru a urmări tendințele pe termen scurt, rata anuală de retenție oferă o imagine mai amplă a clienților pe o perioadă mai lungă.

Poate părea o formulă simplă, dar calcularea ratei de retenție are legătură cu modul în care se numară clenții și activările în perioada pe care se anailează. Unele companii folosesc numărul de clienți de la începutul lunii, în timp ce altele vor aștepta până la sfârșit sau vor folosi un număr mediu.

3.1 State of the art

În ultimii ani, domeniul învățării automate pentru predicția părăsirii clienților a evoluat semnificativ:

- Yang, Z., Sun, X., & Wang, J. (2021) propun un sistem de predicție a părăsirii clienților bazat pe tehnici de învățare ensemble, utilizând gruparea caracteristicilor și algoritmi precum Xgboost, regresie logistică, arbore decizional și Naïve Bayes. Rezultatele arată o acuratețe de 96,12% și 98,09% pentru seturile de date analizate [10].
- Gupta, S., Pathak, N., & Agarwal, S. (2021) explorează diferite tehnici de învățare automată pentru predicția părăsirii clienților în sectorul telecomunicațiilor, subliniind importanța retenției clienților pentru reducerea costurilor și creșterea veniturilor [4].
- Huang, Z., & Kechadi, M. T. (2022) investighează utilizarea rețelelor neuronale profunde pentru a îmbunătăți acuratețea predicțiilor de părăsire a clienților în sectorul telecomunicațiilor [5].
- Siddiqa, A., Karim, A., & Jeong, Y. S. (2021) prezintă o abordare hibridă combinând mai multe tehnici de învățare automată pentru a îmbunătăți predicția părăsirii clienților [7].
- Amin, A., Anwar, S., & Adnan, A. (2022) propun un model de predicție a părăsirii clienților bancari utilizând algoritmi de învățare automată și date istorice ale clienților [2].
- Vafeiadis, T., Diamantaras, K. I., & Sarigiannidis, G. (2020) prezintă un model de predicție a părăsirii clienților în comerțul electronic, utilizând tehnici de învățare automată pentru a analiza comportamentul clienților [8].
- Wang, H., & Hong, W. (2023) explorează utilizarea învățării ensemble pentru predicția părăsirii clienților în diverse industrii și compară performanța mai multor algoritmi [9].
- Khan, M. I., & Al-Habsi, S. (2021) examinează metode de predicție a părăsirii clienților bancari și impactul lor asupra strategiei de retenție a clienților [6].
- Ahmed, M. N., & Maheswari, R. (2021) discută aplicarea tehnicilor de învățare automată pentru predicția părăsirii clienților în industria asigurărilor, subliniind beneficiile acestor metode [1].
- Choudhary, A., & Sharma, V. (2020) compară performanța diferitelor algoritmi de învățare automată pentru predicția părăsirii clienților, evidențiind avantajele și dezavantajele fiecărei

metode [3].

4 Implementare

4.1 Colectarea datelor

Datele folosite pentru acest proiect au fost luate de pe site-ul kaggle. Baza de date se afla in formatul csv(comma separated values) și conține următoarele coloane:

- customerID (de tip șir)- un numar de identificare a utilizatorului
- gender (de tip șir) - se reține sexul clientului
- SeniorCitizen (de tip boolean) - această coloană reține dacă persoana este pensionară sau nu
- Partner(bool) - persoana are sau nu partener
- Dependents (de tip boolean) - dacă clientul are persoane în întreținere
- tenure (de tip întreg) - numărul de luni în care clientul a rămas fidel companiei
- PhoneServce (de tip boolean) - coloană pentru a reține dacă persoana are servicii telefonice
- MultipleLines (de tip boolean) - dacă persoana are serviciul de linii multiple
- InternetService (de tip boolean) - dacă persoana a cumpărat pachetul de dinternet (DSL, fibră optică sau deloc)
- OnlineSecuity- daca clientul a achizitionat pachet de securitate in mediul online
- OnlineBackup (de tip boolean) - dacă persoana are pachetul copie de siguranță online
- DeviceProtecion (de tip boolean) - dacă persoana are protecție a dispozitivului utilizat
- TechSuprt (de tip boolean) - dacă clentul are suport tehnic asigurat de companie
- StreamingTV (de tip boolean) - cablu tv existent in abonamet
- StreamingMovies (de tip boolean) - opținue de redare a filmelor în flux
- Contract (de tip șir)- ce tip de contract are(lunar, un an, doi ani)
- PaperlessBilling (de tip boolean) - opțiune de plată online
- PaymentMethod (de tip șir) - modalitatea de plată(automat prin transfer bancar, automat prin card de credit, factură elecrtronică, factură trimisă la adresa clientului)
- MonthlyCharges(float) - platile lunare catre companie
- TotalCharges(float)- Totalul de plata de candd este abonat
- Churn(bool) - Retentia abonatului

4.2 Curatare si eliminare anomalii

In aceasta parte a proiectului am pregatit baza de date intr-o forma in care ii este calculatorului mai usor sa o inteleaga si sa o invate pentru a realiza scopul proiectului. Toate coloanele care sunt de tip bool au fost transformate in int cu valorile 1 pentru adevarat si 0 pentru fals. Coloanele ded tip strin au fost la randul lor impartite in 2 sau mai multe coloane pentru a putea fi manipulate mai usor, un exemplu este contract care singura nu ptea fi utilizata asadar s-au creeat 3 noi coloane anume: contract_oneyear,contract_twoyear,contract_month-to-month. Pentu modelul machine learning coloana cu id-urile utilizatorilor nu este necesara daca nu are nici un folos in detectaea retentiei clientilor.

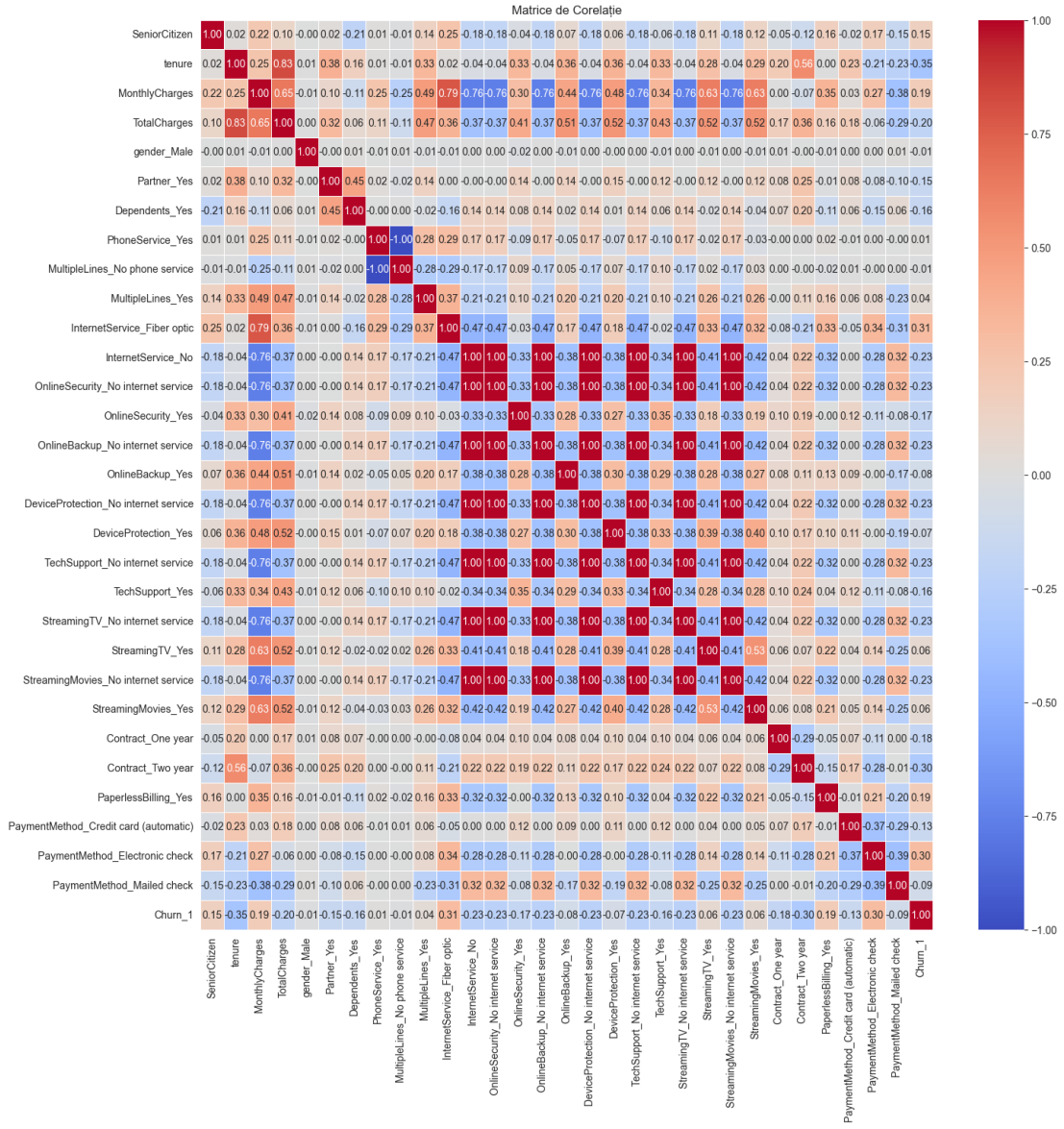

Figura de e mai sus se pote observa matricea de corelatii

Asa cum ziceam mai sus coloana total charges reprezinta suma totala a cheltuielilor efectuate e catre clienti in abonament iar tenure este durata abonamentului iar in figura sa observa corelatia puternica si pozitiva dintre cele doua coloane. Acest lucru reprezinta ca clientii care raman mai mult timp in companie au tendinta de a cheltuii mai mult, deci a ramane sa utilizeze serviciile companiei.

Între Internetservice_no si monthlycharges se observă o corelație negativă puternică. Așadar se observă cum cei care nu au serviciul internet plătesc foarte putin, deci aduc putini bani in companie, dedci nu platesc mult intr-o luna.

Tot din acea figura se observa o corelatie foarte puternica intre coloanele InternetService_FiberOptic si MonthlyCharges au o corelatie foarte puternica deoarece acest serviciu de internet este cel mai scump serviciu pe care compania il ofera explicandu-se corelatia puternica dintre cele doua.

Dupa toate aceste observatii si modificari s-a verificat baza de date pentu a le curate de inconsistente sau chiar a le sterge complet. S-a observat ca in dataset se gaseau niste valori necomplete asa ca s-au completat cu NaN(vid) si au fost sterse din dataset pentu a nu avea probleme calulatorul mai tarziu la modelele ml. Numarul acestor inconsistente a fost de 11 inregistrari.

4.3 Entropia

După curățarea datelor a fost calculată entropia datelor din dataset.În învățarea automată, entropia măsoară nivelul de dezordine sau incertitudine dintr-un set de date sau sistem dat. Este un parametru care cuantifică cantitatea de informații dintr-un set de date și este utilizat în mod obișnuit pentru a evalua calitatea unui model și capacitatea acestuia de a face predicții exacte. Pentru datele prelucrate în acest proiect s-au obținut următoarele valori:

```
customerID 12.779719355143406
gender 0.9999364550464405
SeniorCitizen 0.6400214027212845
Partner 0.9991170306066908
Dependents 0.8794404586827863
tenure 5.914257891990094
PhoneService 0.45844925718753027
MultipleLines 0.9823408405367391
OnlineSecurity 0.8642191874399958
OnlineBackup 0.9293863281716901
DeviceProtection 0.9284615332282073
TechSupport 0.8688534723572245
StreamingTV 0.9610806778229317
StreamingMovies 0.9637379297728217
PaperlessBilling 0.9750506859651258
MonthlyCharges 10.03978145590052
TotalCharges 12.612577727899449
Churn 0.835351115333023
InternetService_DSL 0.9281961426694367
InternetService_Fiber optic 0.9896822991083725
InternetService_No 0.7530837185380153
Contract_Month-to-month 0.9924665688282985
Contract_One year 0.7401977722036489
Contract_Two year 0.7944059240036077
PaymentMethod_Bank transfer (automatic) 0.7588566257578604
PaymentMethod_Credit card (automatic) 0.7533479249119055
PaymentMethod_Electronic check 0.9212532758241578
PaymentMethod_Mailed check 0.7746846444711583
```

Codul și rezultatele

4.4 Modelele ML

Dupa ce s-au terminat toti pașii anteriori, urma aplicarea câtrorva modele pentru a vedea cel mai optim model care sa clasifice cu o acuratețe mare.

Regresia logistica este un algoritm de învățare automată supravegheată care îndeplinește sarcini de clasificare binară prin prezicerea probabilității unui rezultat, eveniment sau observație. Modelul oferă un rezultat binar sau dihotomic limitat la două rezultate posibile: da/nu, 0/1 sau adevărat/false. In urma folosirii acestui model s-au obtinut urmatoarele rezultate:

```
model = LogisticRegression(max_iter=10000,penalty="l2",solver='liblinear')
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
cm = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:\n", cm)
model = RandomForestClassifier(n_estimators=500,criterion="entropy",max_depth=10,min_samples_split=10,
                               min_samples_leaf=11,max_features="log2",oob_score=callable(1),max_samples=1.0)
model = XGBClassifier()


Executed at 2024.05.16 04:46:56 in 296ms
```

```
Accuracy: 0.7668246445497631
Precision: 0.5680473372781065
Recall: 0.5133689839572193
Confusion Matrix:
 [[1330  219]
 [ 273  288]]
```

Codul si rezultatele

Random Forest este o metodă de învățare de ansamblu pentru clasificare, regresie și alte sarcini care funcționează prin construirea unei multitudini de arbori de decizie în momentul instruirii. Pentru sarcinile de clasificare, rezultatul pădurii aleatoare este clasa selectată de majoritatea arborilor.

```python
model1 = RandomForestClassifier(n_estimators=500,criterion="entropy",max_depth=10,
                                min_samples_split=10,min_samples_leaf=11,
                                max_features="log2",oob_score=callable(int),max_samples=1.0)
model1.fit(X_train, y_train)

y_pred = model1.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)

print("Evaluation Metrics:")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Confusion Matrix:\n", cm)
```

Executed at 2024.05.16 04:44:35 in 3s 241ms

```
Evaluation Metrics:
Accuracy: 0.7985781990521327
Precision: 0.6642512077294686
Confusion Matrix:
 [[1410  139]
 [ 286  275]]
```

Codul si rezultatele

Algoritmul The k-nearest neighbors (KNN) este un clasificator de învățare supravegheată, neparametric, care utilizează proximitatea pentru a face clasificări sau predicții cu privire la gruparea unui punct de date individual. Este unul dintre cei mai populari și mai simpli clasificatori de clasificare și regresie utilizați.

```python
knn = KNeighborsClassifier(n_neighbors=8,leaf_size=100)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
print("Evaluation Metrics:")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
print("Confusion Matrix:\n", cm)
```

Executed at 2024.05.16 07:15:39 in 219ms

```
Evaluation Metrics:
Accuracy: 0.785781990521327
Precision: 0.6786885245901639
Recall: 0.3689839572192513
F1-score: 0.5393258426966292
Confusion Matrix:
 [[1451   98]
 [ 354  207]]
```

Codul si rezultatele

În urma analizării rezultatelor se observă faptul ca algoritmul Random Forest este cel mai eficient, deci acel model este ales pentru a se folosi mai departe în proiect. landscape

5 Testare și Rezultate

S-a încercat testarea în mai multe istanțe, dar fără succes. În prima faza s-au intodus numere manual dupa date generate de inteligența arificială dar nu s-a ajuns la nici un rezultat valid, accuratețea fiid de 50-60% în ambele cazuri. Acea valoare nu este cu nimic mai mare decât metoda aruncării unei monede. Singura metodă prin care s-a putut testa algoritmul eficient a fost realizarea acelorași pași realizați în Jupyter Notebooks și în programul Rapidminer. În Rapidminer au fost adăugațe urmatoarele valori(cu atributele initiale, adicâ Contract, Dependenți, Protecție Dispozitiv, Serviciu Internet, Cheltuieli Lunare, Linii Multiple, Backup Online, Securitate Online, Facturare Electronică, Metodă de Plată, Serviciu Telefonic, Cetățean Senior, Filme Online, TV Online, Suport Tehnic, Cheltuieli Totale, Gen și Perioadă de utilizare, deoarece în program se poate lucra direct cu mai multe valori, nu doar 1 și 0):

1. *Abonament lunar, Da, Da, DSL, 0.5, Nu, Nu, Da, Da, Factură electronică, Da, Nu, Nu, Nu, Da, 1.2, Bărbat, 0.3*
2. *Abonament lunar, Da, Nu, Fibra optică, 0.72, Da, Nu, Da, Da, Transfer bancar, Da, Da, Da, Nu, Da, 0.2, Bărbat, 0.5*
3. *Un an, Da, Nu, DSL, 85.45, Nu, Nu, Da, Nu, Transfer bancar, Nu, Nu, Da, Nu, Da, 1.0, Femeie, 1.1*
4. *Un an, Nu, NU, Fibra optică, 1.46, Da, Nu, Nu, Da, Factură fizică, Da, Da, Da, Da, Da, -0.5, Bărbat, 0.2*

Penrtu exemplul 1 precizia programului a fost de 76.22% cu răspunusl nu. La exemplul 2 precizia este de 76.44% cu răspunsul nu. Pentru exemplul 3 răspunsul este tot nu cu precizia de 87.10%. În exemplul 4 răspusul este da cu precizia 73.17%. Procentul general de precizie calculat prin media procentelor de mai sus este 78.23%.

6 Concluzii

În concluzie algoritmul realizat este eficient, dar prin realizarea de ensamble de algoritmi consider ca s-ar putea înbunatății eficiența algoritmului, ca în cazul cercetătorilor Yang, Z., Sun, X., & Wang, J. care au obținut între 96-98% accuratețe, printr-o alegere bună a sistemului de votare.

Referințe

[1] M. N. Ahmed și R. Maheswari, "Machine learning techniques for customer churn prediction in the insurance industry", în *Journal of Big Data* 8 (2021), p. 110.

[2] A. Amin, S. Anwar și A. Adnan, "Customer churn prediction in banking using machine learning: A case study", în *IEEE Access* 10 (2022), pp. 116420–116432.

[3] A. Choudhary și V. Sharma, "A comparative analysis of machine learning algorithms for customer churn prediction", în *Procedia Computer Science* 173 (2020), pp. 162–169.

[4] S. Gupta, N. Pathak și S. Agarwal, "Churn prediction in the telecom sector using machine learning techniques", în *Journal of Emerging Technologies and Innovative Research* 8.5 (2021), pp. 242–248.

[5] Z. Huang și M. T. Kechadi, "Customer churn prediction in telecommunications using deep learning", în *IEEE Access* 10 (2022), pp. 12504–12515.
[6] M. I. Khan și S. Al-Habsi, "Predicting customer churn using machine learning techniques in the banking sector", în *Journal of Banking & Finance* 127 (2021), p. 106116.
[7] A. Siddiqa, A. Karim și Y. S. Jeong, "A hybrid approach to customer churn prediction using machine learning techniques", în *Computers & Electrical Engineering* 93 (2021), p. 107273.
[8] T. Vafeiadis, K. I. Diamantaras și G. Sarigiannidis, "A machine learning model for customer churn prediction in e-commerce", în *Expert Systems with Applications* 139 (2020), p. 112847.
[9] H. Wang și W. Hong, "Ensemble learning for customer churn prediction: A comprehensive study", în *Journal of Business Research* 145 (2023), pp. 731–742.
[10] Z. Yang, X. Sun și J. Wang, "A telecom churn prediction system based on ensemble learning using feature grouping", în *Applied Sciences* 11.11 (2021), p. 4742.