

DESIGNUL UNEI APLICAȚII DE DATA MINING PENTRU ACHIZIȚIONAREA NOILOR CLIENȚI

Alina POPA^{1*}

¹Academia de Studii Economice, Școala Doctorală Marketing, București, România

*Autorul corespondent: Alina Popa, popa.alina.alexiei@gmail.com

Rezumat. *Clienții reprezintă cel mai important activ al unei organizații. Astfel, o companie trebuie să-și planifice și să utilizeze o strategie clară pentru gestionarea acestora. Pentru identificarea de noi clienți și extinderea bazei de clienți cu clienți noi și rentabili, companiile pot să utilizeze aplicații de Data Mining.*

În această lucrare, se prezintă atât designul unei astfel de aplicații, cât și etapele și activitățile care trebuie efectuate. De asemenea, se oferă sugestii cu privire la algoritmi de inteligență artificială potriviți pentru fiecare pas. În cele din urmă, se menționează o suită de metrici care pot fi utilizate în evaluarea sistemului dezvoltat.

Cuvinte cheie: *Inteligență Artificială, Machine Learning, Marketing, Comunicare*

Introducere

Pentru marea majoritatea a companiilor, mijlocul principal de creștere implică achiziționarea de noi clienți. Aceștia s-ar putea să nu fi fost conștienți de produsul sau serviciile companiei, să nu se fi calificat sau avut nevoie de produsele companiei sau chiar să fie clienți pierduți. În oricare dintre aceste situații, metoda tradițională era să se dezvolte o combinație dintre campanii de marketing în masa și campanii de marketing direct pe baza cunoștințelor pe care le dețin persoanele de comunicare din cadrul companiei despre clienții vizați și bazele de date disponibile [1]. O persoană de marketing cu experiența poate adesea înțelege și alege cele mai relevante caracteristici socio-demografice, însă o dată cu creșterea cantității de date și schimbării rapide a tiparelor comportamentale, procesul devine nefezabil. În același timp, se dorește o înmagazinare a cunoștințelor care în trecut erau deținute cu precădere de un grup de persoane specializate în cadrul proceselor operaționale ale companiei. În acest demers, aplicațiile de Data Mining (DM) ce folosesc algoritmi de Machine Learning (ML) pot fi utilizate pentru a identifica potențiali clienți rentabili în o bază de date, prin găsirea de „clone” ale clienților valoroși existenți în bazele de date ale companiei [2, 3].

De cele mai multe ori, o campanie de achiziție de clienți începe cu o listă de prospecți. O listă de prospecți reprezintă o listă de entități (persoane fizice sau juridice) care au fost selectate pe baza unor criterii socio-demografice destul de generale. Dintre aceste entități, unele ar putea fi interesate de produsul/serviciile companiei [1]. Uneori, aceste liste pot să existe deja în companie, exemplu fiind lista clienților pierduți sau bazele de date rezultate după fuziunea a două companii. Alteori, acestea trebuie cumpărate de la companiile specializate în domeniu.

În cadrul acestui articol se va prezenta designul unei care are ca scop generarea de liste optimizate pentru achiziționarea de noi clienți.

Prezentarea obiectivelor aplicației

Aplicația are următoarele obiective:

- Sa se extragă din lista generală de prospecți numai acele entități care au probabilitatea ridicată de a deveni consumatori.
- În același timp, pentru fiecare dintre acești potențiali consumatori, este nevoie să se decidă mesajul de comunicare astfel încât acesta să aibă cel mai mare succes în atragerea clientului.

- Un alt obiectiv este să se abordeze numai prospekți care vor deveni consumatori valoroși. Acest lucru poate fi condiționat de faptul că resursele companiei care se ocupă de menținerea relației cu clientul sunt limitate sau însăși menținerea relației cu consumatorul este costisitoare, astfel că se dorește maximizarea clienților profitabili.

Designul aplicației

În mod schematic, o aplicație de generare de liste optimizate este prezentată în Figura 1. Orice aplicație de DM începe prin stabilirea datelor necesare și surselor de date utilizate. Pentru a optimiza atragerea clienților potențiali, este nevoie de date interne pe care le are compania despre clienții actuali în termeni de 1) informații socio-demografice și 2) comportament de cumpărare. Aceste două tabele vor fi principalele date utilizate în antrenarea modelului. O altă sursă de date pe care se va utiliza modelul dezvoltat este după cum s-a menționat și înainte, lista de prospekți care conține și informații socio-demografice despre aceștia.

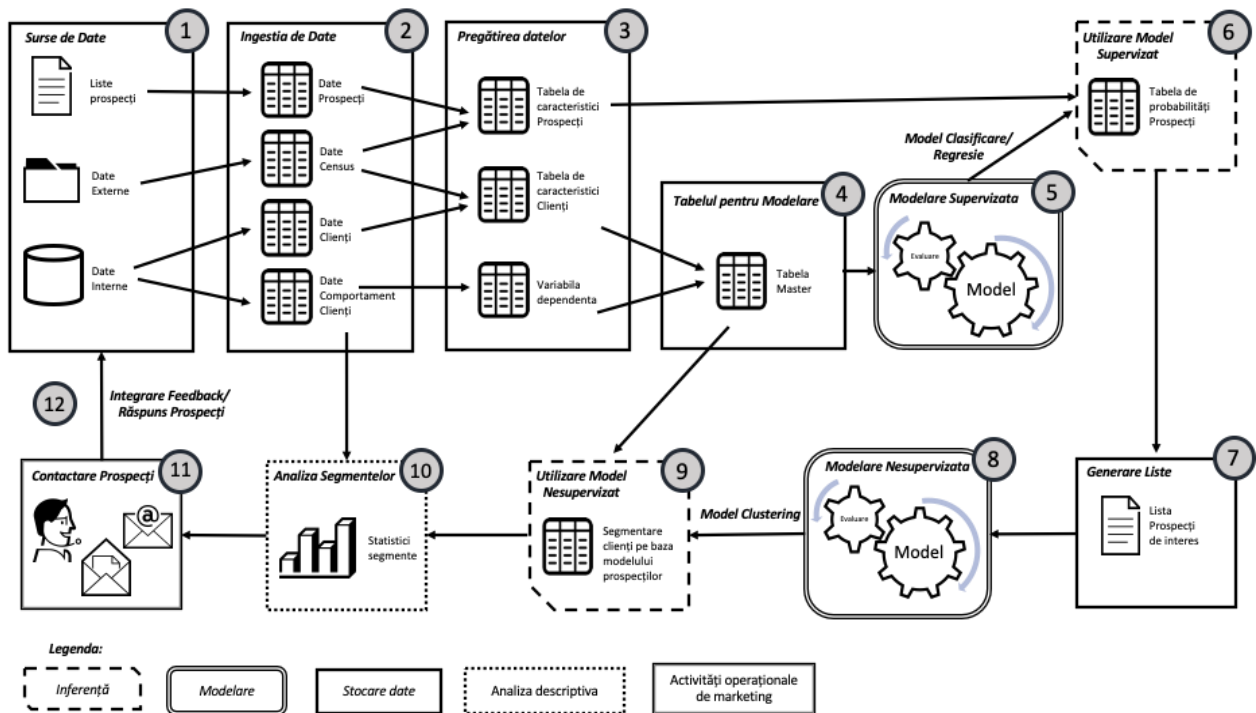


Figura 1. Schema sistemului de Data Mining pentru generarea de liste optimizate

Pentru a îmbogăți datele existente despre clienți și prospekți, este necesară utilizarea unor surse de date externe, cum ar fi date census agregate la nivel de orașe, zone ale țării sau chiar cartiere ale unui oraș. După ce au fost stabilite sursele de date necesare (Figura 1., 1), tabele corespunzătoare sunt extrase și încărcate în baza de date a aplicației (Figura 1, 2).

Următorul pas este pregătirea datelor (Figura 1., 3). Această etapă ia în jur de 80% din timpul necesar dezvoltării aplicației.

O dată ce tabelul de caracteristici și cel de variabile dependente sunt pregătite, acestea se unesc pe baza identificatorului de client într-o singură tabelă, mai numit și tabela master (Figura 1., 4). Tabela reprezintă forma finală a datelor pe baza cărora va fi antrenat modelul.

În partea de modelare (Figura 1., 5), tabela master se împarte în două seturi: de antrenament și de test. Pe setul de antrenament, pentru fiecare variabilă dependentă considerată de interes se aplică diferiți algoritmi de Machine Learning. Se încearcă configurații diverse pentru fiecare tip de algoritm. Pe setul de test se validează performanțele diferitor modele știindu-se valorile variabilei dependente, adică “răspunsurile corecte”.

Dintre algoritmi de ML potriviți pentru modelarea variabilelor numerice se poate alege GLM [4], Arborii de decizie CART [5], Păduri Aleatoare [6] sau Rețelele Neurale [7].

Pentru variabila binară, setul de metode se extinde la SVM [8] și KNN [9]. În cazul KNN, este nevoie să se menționeze că modelul nu încearcă să găsească o relație dintre datele de intrare, caracteristicile prospectilor și probabilitatea de a deveni client al companiei, ci mai degrabă identifică clienții potențiali în o bază de date, prin găsirea de „clone” ale clienților existenți în bazele de date ale companiei.

Deoarece înțelegerea profilului unui client vs. non-client al companiei este de interes pentru dezvoltarea mesajului comunicării, se pune accent pe utilizarea de algoritmi care pot fi înțeleși de oameni. De exemplu un arbore de decizie poate fi descompus într-o serie de reguli “dacă, atunci”.

Pentru acest tip de aplicație, indicatorii de performanță importanți sunt RMSE (Ec. 1) și MSE (Ec. 2) [10]:

- *Eroarea Medie Pătratică* (eng. Mean Square Error, MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2 \quad (1)$$

- *Rădăcina Pătrată din Eroarea Medie Pătratică* (eng. Root Mean Square Error, RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} = \sqrt{MSE} \quad (2)$$

O dată ce modelele ating performanța dorită, acestea se pot utiliza în evaluarea listei de prospecti (Figura 1., 6). Pentru fiecare prospect se va estima o valoare viitoare a tranzacțiilor lunare, număr mediu tranzacții per lună și o variabilă binară care exprimă probabilitatea să devină client al companiei.

Lista de prospecti cu noile informații se sortează descrescător pe baza produsului dintre probabilitatea de a deveni client și valoarea totală medie tranzacționată pe lună. Se aleg prospectii cu perspectiva ridicată pentru lista finală optimizată (Figura 1., 7) [11].

Următorul obiectiv care trebuie abordat în cadrul aplicației este determinarea produselor de interes sau mesajului potrivit pentru fiecare prospect din listă. Acest lucru este dificil având în vedere că nu există un istoric pentru entitățile din lista optimizată. O soluție ar fi segmentarea prospectilor pentru extragerea segmentelor principale existente în lista optimizată (Figura 1., 8), după care pentru fiecare segment se va dezvolta o alternativă a mesajului de marketing.

Segmentarea se poate efectua utilizând orice algoritm de grupare, totuși, deoarece în pasul următor (Figura 1., 9) modelul se aplică pe baza de clienți a companiei cu scopul de a extrage clienții care se află în aceleași segmente ca și prospectii, se va prioritiza utilizarea unui algoritm de grupare care are noțiunea de zgomot, adică DBSCAN [12].

Rezultatul utilizării modelului de grupare (Figura 1., 9) este reprezentat de o listă de clienți actuali ai companiei care sunt asemănători cu prospectii. Pe baza acestor clienți extrași și alocați fiecărui segment, pentru fiecare segment se extrag statistici cu privire la produsele sau categoriile de produse cumpărate, promoțiile atractive (Figura 1., 10).

În etapele ce urmează se trimit mesajele dezvoltate prin instrumentele de marketing direct (Figura 1., 11) și se integrează feedback-ul obținut de la client. Aceste răspunsuri sunt utilizate în reantrenarea și rafinarea modelului de ML, dar și în evaluarea campaniei de achiziție de clienți.

Evaluarea performantei aplicației

Indicatorii de business care exprimă eficiența aplicației sunt Rata de Achiziție (eng. Acquisition Rate) și ROI (eng. Return of Investment) al aplicației:

- $Rata\ de\ Achiziție = 100 \cdot \frac{Numărul\ de\ prospecti\ achizitionati}{Numărul\ de\ prospecti\ targetati}$ (3)

- $ROI\ aplicație = \frac{Venituri\ clienți\ convertiti - Cheltuielile\ cu\ aplicația}{Cheltuielile\ cu\ aplicația}$ (4)

unde:

- Venituri clienți converțiți reprezintă creșterea generată de prospecții deveniți clienți într-o perioadă t.
- Cheltuielile cu aplicația reflectă atât costurile fixe ce țin de cumpărarea listelor prospecți, tehnologiei, serviciilor de consultanță IT, dar și cheltuielile operaționale pentru menținerea sistemului aplicației în perioada t.

O altă modalitate de a estima eficiența campaniei de achiziție este să se realizeze în paralel o campanie de test sau control sub forma unui experiment controlat.

Concluzii

În această lucrare, a fost prezentat un cadru conceptual care poate fi adaptat la o gamă largă de utilizări. De asemenea, s-a explicat în mod simplu și compresiv care sunt componentele, etapele și modul în care se poate dezvolta un astfel de sistem. Nu în ultimul rând, au fost oferite niște metrice care pot fi utilizate în evaluarea sistemului dezvoltat.

Cercetări ulterioare sunt necesare pentru a implementa această abordare și pentru a o utiliza în situații din lumea reală cu scopul de a identifica limitările și avantajele practice.

Referințe:

1. Maimon O., Rokach L. (2010), „Data Mining and Knowledge Discovery Handbook”, Springer
2. Chorianopoulos, A. (2016). Effective CRM using predictive analytics. John Wiley & Sons.
3. Kadiyala, S. S., & Srivastava, A. (2002). Data Mining For Customer Relationship Management. International Business & Economics Research Journal (IBER), 1(6).
4. Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.
5. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). CART. Classification and Regression Trees, Wadsworth and Brooks/Cole, Monterey, CA.
6. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
7. Hinton, G. E. (1992). How neural networks learn from experience. Scientific American, 267(3), 144-151.
8. Vapnik, V., Guyon, I., & Hastie, T. (1995). Support vector machines. Mach. Learn, 20(3), 273-297.
9. Fix, E. (1985). Discriminatory analysis: nonparametric discrimination, consistency properties (Vol. 1). USAF school of Aviation Medicine.
10. Kantardzic, M. (2019). Data mining: concepts, models, methods, and algorithms, 3rd Edition, John Wiley & Sons.
11. Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques în CRM: inside customer segmentation. John Wiley & Sons.
12. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).