

APLICAREA TEHNOLOGIILOR DATA MINING PENTRU DEPOZITARE, ANALIZĂ ȘI RAPORTARE A DATELOR. PUBLICAREA RAPOARTELOR

Viorica MOȘOI, Facultatea Matematică și Informatică

*Motto: "Knowledge is the source of the highest quality power".
Alvin Toffler*

The objective of this study is to describe the applications of Data Mining techniques for better business decision making. A churn prediction model for Internet Services providers and Digital Television market is trained using Decision Tree algorithm in Microsoft SQL Server Analysis. There are several reports designed in Microsoft SQL Server Reporting Services which allow organizational managers to decide their future action regarding customer retention.

Creșterea rapidă și continuă a volumului de date, care a depășit capacitatea umană de a le interpreta, a condus la apariția necesității extragerii automate a cunoștințelor din bazele de date de mari dimensiuni. Începând cu anii '90, apare un nou domeniu de cercetare denumit *mineritul datelor* (DM – *Data Mining*).

Aflat la confluența mai multor discipline, statistica, sisteme de baze de date, inteligență artificială, termenul a fost utilizat în special de către statisticieni și analiști în cadrul comunităților de management al sistemelor informatice [1]. Conform Gartner Glossary „Inteligența afacerilor este un termen generic care include aplicațiile, infrastructura și instrumentele, cele mai bune practici, care să permită accesul la analiza informațiilor în vederea îmbunătățirii și optimizării deciziilor și a performanțelor”, iar rolul mineritului de date este extragerea de cunoștințe noi, implicite și cu acțiune directă din colecții mari de date, descoperirea lucrurilor care nu sunt evidente din date, care nu pot fi extrase manual, reprezentând informații folositoare care pot îmbunătăți procesul curent de acțiune [2, 3].

În urma efectuării unei analize detaliate a metodelor și tehnicilor mineritului de date (Bayes naiv, algoritmul celor mai apropiați k vecini, metoda rețelelor neuronale, tehnica arborilor de decizie), se

evidențiază caracteristicile de bază, avantajele, dezavantajele, precum și aplicabilitatea acestora în domeniul inteligenței afacerilor.

Ca rezultat al tendințelor care au marcat în ultimii anii piața telecomunicațiilor, furnizorii de servicii au realizat importanța reținerii clienților existenți, previziunea și prevenirea migrării acestora devenind cea mai îngrijorătoare problemă în activitatea antreprenorială la moment.

Literatura de specialitate definește migrarea clienților din sfera telecomunicațiilor ca schimbarea prestatorului de servicii, iar managementul migrărilor de clienți – procesul și modalitatea de reținere a clienților profitabili [4]. În practică acest tip de management se axează pe segmentul celor mai profitabili clienți. Când compania adoptă o strategie reactivă, aceasta așteaptă ca clientul să solicite rezilierea contractului după care îl poate tenta cu o ofertă atractivă.

Din alt punct de vedere, strategia proactivă încearcă să identifice clienții predispuși să migreze înainte ca cei din urmă să o facă. Programele reactive sunt mai puțin costisitoare, comparativ cu cele proactive, care pot genera costuri extrem de mari în cazul unor predicții superficiale contactând clienți care nu ar avea intenția de migrare [5]. O soluție de prevenire a intenției de plecare a consumatorilor ar fi dezvoltarea unor modele predictive, care presupune un proces foarte complex, desfășurat în mai multe etape: definirea problemei, pregătirea datelor, explorarea datelor, construirea modelului, examinarea și validarea modelului, lansarea și actualizarea modelului (Fig.1).

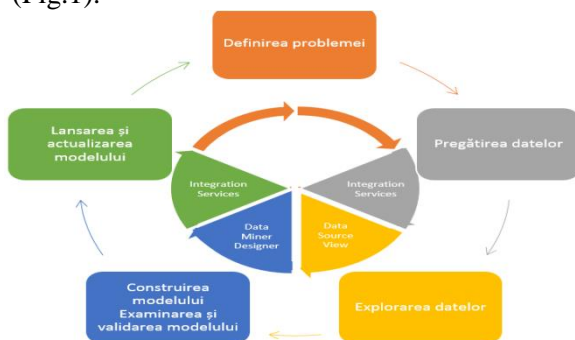


Fig. 1. Relațiile dintre etapele procesului DM și tehnologiile Microsoft SQL Server 2008

În lucrarea dată, au fost utilizate instrumentele Microsoft Data Mining în mediul SQL Server Business Intelligence Development Studio pentru proiectarea unui model de predicție a migrării clienților pe piața serviciilor de Internet prin fibră optică și televiziune digitală cu ajutorul arborilor de decizie. A fost definită problema de business, după care a urmat analiza datelor disponibile, prelucrarea acestora pentru a obține un format acceptat de mediul de analiză. La această etapă, s-au utilizat instrumente ca Microsoft SQL Server 2008 Integration Services și Data Source View Designer în BI Development Studio. Pentru a obține modelul final, au fost grupate variabilele de intrare în mai multe blocuri, realizându-se câte un model separat pentru fiecare cu scopul de a le delimita pe cele mai importante dintre ele. Modelul construit pentru realizarea predicției include doar cei mai semnificativi factori ca urmare a selecției realizate anterior.

Atât în procesul de selectare a variabilelor, cât și în procesul de învățare a modelului, s-au utilizat arborii de decizie prin intermediul Microsoft SQL Server 2008 Analysis Services. Structura arborelui de decizie și relațiile de dependență dintre variabilele independente și variabila dependentă au fost urmărite cu ajutorul Microsoft Tree Viewer și Dependency Network.

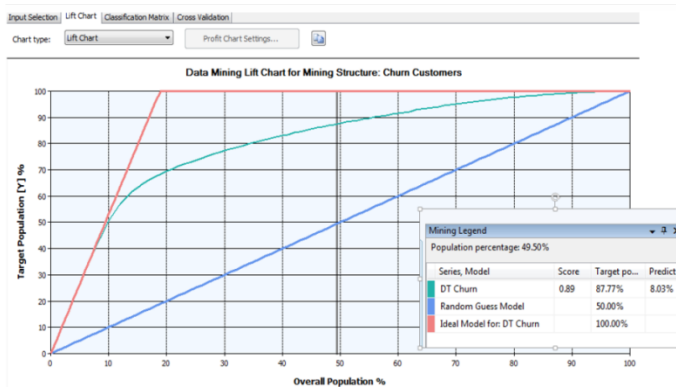


Fig. 2. Lift Chart în Microsoft SQL Server 2008

Aplicat pe setul de testare, Lift Chart descrie modelul proiectat (Fig.2). Astfel s-a realizat o comparație între modelul obținut cu un model de selecție aleatoare și un presupus model perfect, s-a evaluat scorul modelului și probabilitatea de predicție care oferă cele mai bune rezultate.

Pentru a stabili acuratețea modelului, s-a construit și analizat matricea de confuzie care arată în valori absolute, în baza setului de date de testare, în ce proporții modelul face predicții bune sau greșite.

Exportul rezultatelor se efectuează prin intermediul rapoartelor proiectate cu ajutorul Microsoft SQL Server 2008 Reporting Services și publicate pe server, aceasta fiind una dintre cele mai eficiente metode de distribuție a rezultatelor către utilizatorii finali (managementul companiei, persoanele implicate în procesul de luare a deciziilor) care nu sunt utilizatori ai pachetelor software Microsoft SQL Server 2008.

Referințe:

1. ARUN, K. *Data mining techniques*. Universities Press, 2001.
2. *Gartner IT Glossary*. Gartner Inc. Disponibil pe Internet: <http://www.gartner.com/it-glossary/>. [Accesat 11 ianuarie 2016].
3. *An Introduction to Data Mining*. Disponibil pe Internet: <http://www.theartling.com/text/dmwhite/dmwhite.htm>. [Accessed mai 2016].
4. BERSON, A. , SMITH, S. , THEARLING, K. *Building data mining applications for CRM*, New York. NY: McGraw-Hill, 2000.
5. SHARMA, A. , PANIGRAHI, P. K. *A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services*, Information Systems Area Indian Institute of Management, Indore, India, 2011.

*Recomandat
Ion ANDRIEȘ, dr., conf. univ.*