Journal of Information Technologies, Vol. 8, No. 1, 2015, ISSN: 1337-7467

# Logistická regresia ako nástroj pre klasifikáciu zákazníkov

# Logistic regression as a Tool for Customer Classification

***Andrej Trnka***

**Abstract:** The article describes possibility of using Data Mining's method in telecomuunication area. The telecomunication's provider can segments its customer base by service usage patterns, categorizing the customers into different groups. In our article we use only demographic data to predict group membership. As a result of Data Mining analysis can be the customization of offers for individual prospective customers. As a Data Mining method we use multinomial logistic regression.The dataset contains a lot of information about each customer (variables). The goal is to identify most important variables as a predictors and to make the classification table.

**Key words:** customer, classification, data mining, multinomial logistic regression

**Abstrakt:** Článok popisuje možnosti použitia metód dolovania dát v oblasti telekomunikácií. Poskytovateľ telekomunikačných služieb môže segmentovať svojich zákazníkov podľa používania jednotlivých služieb a kategorizovať ich do viacerých skupín. Pre vytvorenie predikcie zaradenia zákazníkov do jednotlivých skupín používame v článku iba demografické dáta. Výsledkom analýzy dolovania dát môže byť prispôsobenie ponuky pre budúcich zákazníkov. Ako použitú metódu dolovania dát sme zvolili multinomickú logistickú regresiu. Dátový súbor obsahuje veľké množstvo informácií o každom zákazníkovi (premenné). Cieľom je identifikácia dôležitých premenných (prediktorov) a vytvorenie klasifikačnej tabuľky.

**Kľúčové slová:** zákazník, klasifikácia, dolovanie dát, multinomická logistická regresia

# 1. Úvod

Logistická regresia je štatistická technika používaná ku klasifikácii záznamov založená na hodnotách vstupných polí (premenných). Je analógiou k lineárnej regresii, ale ako cieľovú premennú vyžaduje kategorickú (lineárna regresia pracuje s číselnými premennými). Jej výstupom sú modely, kde cieľová premenná je súbor premenných s viac ako dvomi možnými hodnotami.

V prípade, ktorý popisuje článok, má poskytovateľ telekomunikačných služieb segmentovaných zákazníkov podľa vzorov používania služieb. Zákazníci sú kategorizovaní do štyroch skupín. Názvy skupín sú pre potreby tohto článku len všeobecné. Ak môžu byť použité demografické dáta na zaradenie do jednotlivých skupín, je možné prispôsobiť ponuku pre každého individuálneho zákazníka.(1), (2)

Príklad je zameraný na použitie len demografických dát na predikciu vzorov správania sa zákazníkov. Cieľová premenná *kat_zakaznika* obsahuje štyri možno hodnoty, ktoré zodpovedajú štyrom zákazníckym skupinám (tabuľka 1).

***Tabuľka 1: Zákaznícke skupiny***

| Hodnota premennej | Názov |
|---|---|
| 1 | Skupina 1 |
| 2 | Skupina 2 |
| 3 | Skupina 3 |
| 4 | Skupina 4 |

Zdroj: vlastné spracovanie

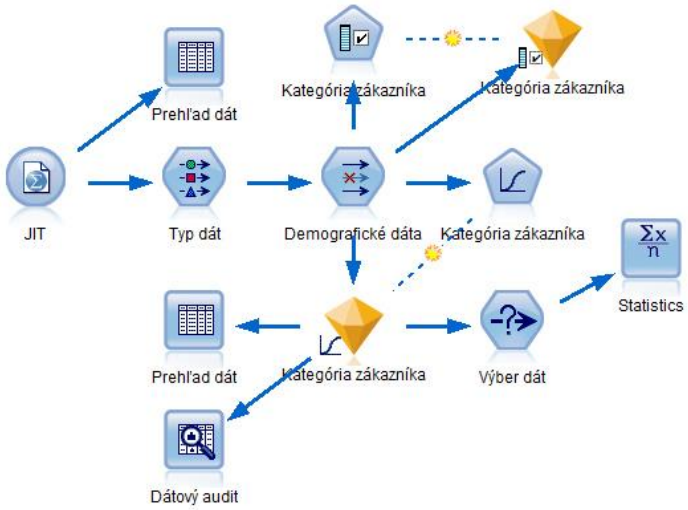
Pre analýzu pomocou dolovania dát využijeme softvérový produkt IBM SPSS Modeler 14.1. Keďže cieľová premenná má viacero kategórií, použijeme multinomický model. V prípade, že by cieľová premenná mala len dve odlišné kategórie, (čiže bola by dichotomická), mohli by sme použiť binomický model. (1), (2)

Pojem *churn* by sme mohli definovať ako odliv zákazníkov. Je odvodený zo slov change a turn. Jeho redukcia je dôležitá, pretože získavanie nových zákazníkov je vždy nákladnejšie, než udržanie si existujúcich. Za účelom zníženia odlivu zákazníkov a s tým spojeným

zvýšením ziskov, musia spoločnosti predikovať správanie sa zákazníkov, ktorí by chceli ukončiť zmluvu a odísť ku konkurencii. (3), (4)

## 2. Použité metódy

Pre analýzu sme zostavili model, v ktorom sme použili multinomickú regresiu. Tento model je zobrazený na obrázku 1.




***Obrázok 1: Vytvorený dátový model***

Zdroj: vlastné spracovanie

Ako sme už spomenuli, tento článok je zameraný na demografické dáta. Použitý dátový set obsahuje 813 záznamov a je v SPSS formáte. Dáta môžu byť však uložené aj v tzv. dátových skladoch. (5)

Techniky a algoritmy dolovania dát sú použiteľné vo viacerých oblastiach. Novým trendom je ich využitie vo výrobných procesoch. (6)

## 3. Výsledky

Z použitého dátové setu sme vyfiltrovali len tie premenné, ktoré sú relevantné (demografické):

- *region* – región v ktorom zákazník býva
- *vek* – vek zákazníka
- *stav* – rodinný stav

- *adresa* – počet rokov, koľko býva zákazník na poslednej adrese
- *prijem* – celkový príjem v domácnosti
- *vzdelanie* – dosiahnuté vzdelanie
- *zamestnanie* – počet rokov u aktuálneho zamestnávateľa
- *dochodok* – indikátor dôchodku
- *pohlavie* – pohlavie zákazníka
- *osoby* – počet osôb v domácnosti
- *kat_zakaznika* – kategória zákazníka

Ostatné premenné boli z analýzy vylúčené.

Ako referenčnú kategóriu zákazníkov sme zvolili Skupinu 1. Vytvorený model bude porovnávať ostatných zákazníkov so zákazníkmi, ktorí sú v Skupine 1.

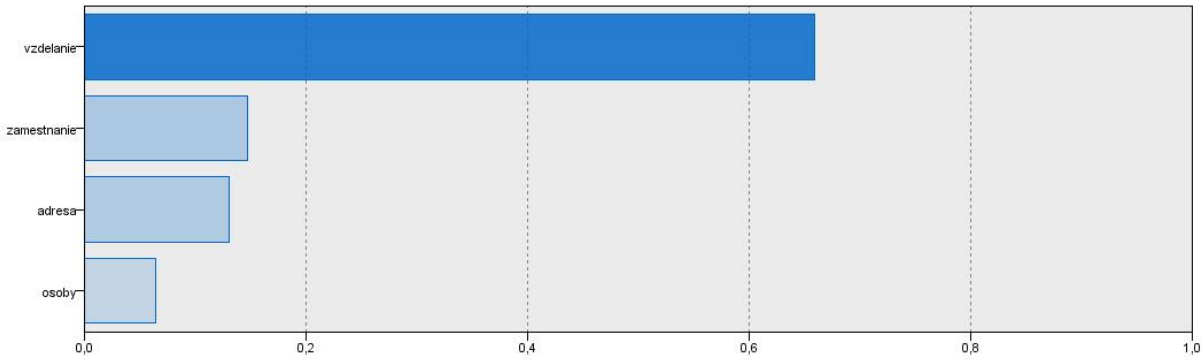
Po spustení modelu a zobrazení výsledkov je vidieť, že model vygeneroval rovnice, ktoré je možné použiť na priradenie záznamov do každej skupiny cieľovej premennej. V našom prípade sú zobrazené štyri možné kategórie a jedna z nich je tzv. základná, pre ktorú nie je zobrazená žiadna rovnica. Tabuľka 2 zobrazuje rovnice, kde kategória 2 reprezentuje Skupinu 2, atď.

***Tabuľka 2: Vygenerované rovnice pre kategórie zákazníkov***

| | |
|---|---|
| Equation For 1<br>+ 0,00000000000000000000 | Equation For 3<br>0,01962 * adresa +<br>0,05073 * [vzdelanie=1] +<br>0,4344 * [vzdelanie=2] +<br>0,1699 * [vzdelanie=3] +<br>0,1224 * [vzdelanie=4] +<br>0,05761 * zamestnanie +<br>0,07063 * osoby +<br>+ -1,141 |
| Equation For 2<br>0,03751 * adresa +<br>-2,62 * [vzdelanie=1] +<br>-1,803 * [vzdelanie=2] +<br>-1,326 * [vzdelanie=3] +<br>-0,9107 * [vzdelanie=4] +<br>0,03585 * zamestnanie +<br>0,1484 * osoby +<br>+ 0,2762 | Equation For 4<br>0,01979 * adresa +<br>-3,959 * [vzdelanie=1] +<br>-2,102 * [vzdelanie=2] +<br>-1,729 * [vzdelanie=3] +<br>-0,738 * [vzdelanie=4] +<br>0,05316 * zamestnanie +<br>0,2555 * osoby +<br>+ 0,3471 |

Zdroj: vlastné spracovanie

Ďalší výsledok analýzy je graf, ktorý zobrazuje dôležitosť jednotlivých prediktorov. Ako je zobrazené na obrázku2, najdôležitejším prediktorom pre klasifikáciu je premenná *vzdelanie*.




***Obrázok 2: Najdôležitejšie prediktory***

Zdroj: vlastné spracovanie

Dôležitú časť výsledkov tvorí tabuľka s názvom Case Processing Sumary, ktorá zobrazuje percentuálny počet záznamov, ktoré spadajú do každej kategórie cieľovej premennej (tabuľka 3). To nám dáva nulový model, ktorý môžeme použiť ako základ pre porovnanie.

Bez vytvorenia modelu, ktorý by používal prediktory, by najlepšou voľbou bolo priradenie všetkých zákazníkov do najčastejšie sa vyskytujúcej skupiny - do Skupiny 3. Ak by sme však na základe trénovacích dát priradili všetkých zákazníkov do nulového modelu, boli by sme presní len 27,7% (225/813). V ďalšom kroku porovnáme prediktory s výsledkami nulového modelu, aby sme videli, ako model pracuje s dátami.

Ďalšia tabuľka vo výsledku je tabuľka s názvom Classification table (tabuľka 4), ktorá zobrazuje výsledky vytvoreného modelu. Tento model je presný na 40,3%.

***Tabuľka 3: Súhrn spracovania dát***

| Case Processing Summary | | | |
|---|---|---|---|
| | | N | Marginal Percentage |
| kat_zakaznika | Skupina 1 | 213 | 26,2% |
| | Skupina 2 | 177 | 21,8% |
| | Skupina 3 | 225 | 27,7% |
| | Skupina 4 | 198 | 24,4% |
| region | Región 1 | 258 | 31,7% |
| | Región 2 | 279 | 34,3% |
| | Región 3 | 276 | 33,9% |
| stav | Slobodný | 409 | 50,3% |
| | Ženatý | 404 | 49,7% |
| vzdelanie | Základná škola | 159 | 19,6% |
| | Stredná škola | 236 | 29,0% |
| | Bc. | 162 | 19,9% |
| | Mgr./Ing. | 200 | 24,6% |
| | PhD. | 56 | 6,9% |
| dochodok | Nie | 772 | 95,0% |
| | Áno | 41 | 5,0% |
| pohlavie | Muž | 393 | 48,3% |
| | Žena | 420 | 51,7% |
| Valid | | 813 | 100,0% |
| Missing | | 0 | |
| Total | | 813 | |
| Subpopulation | | 813(a) | |
| a. The dependent variable has only one value observed in 813 (100,0%) subpopulations. | | | |

Zdroj: vlastné spracovanie

***Tabuľka 4: Klasifikačná tabuľka***

| Classification | | | | | |
|---|---|---|---|---|---|
| | Predicted | | | | |
| Observed | Skupina 1 | Skupina 2 | Skupina 3 | Skupina 4 | Percent Correct |
| Skupina 1 | 105 | 6 | 53 | 49 | 49,3% |
| Skupina 2 | 42 | 11 | 53 | 71 | 6,2% |
| Skupina 3 | 76 | 7 | 101 | 41 | 44,9% |
| Skupina 4 | 34 | 13 | 40 | 111 | 56,1% |
| Overall Percentage | 31,6% | 4,6% | 30,4% | 33,5% | 40,3% |

Zdroj: vlastné spracovanie

Model vyniká identifikáciou zákazníkov zo Skupiny 4 (56,1%), ale Skupinu 2 identifikuje veľmi slabo (6,2 %). Pokiaľ by sme chceli väčšiu presnosť zákazníkov zo Skupiny 2, budeme musieť nájsť iné prediktory, ktoré ju lepšie identifikujú.

## 4. Záver

V závislosti od toho, čo chceme predikovať, môže byť navrhnutý model postačujúci. Napríklad, ak by sme chceli zamerať na zákazníkov v Skupine 2, môže byť model aj napriek nízkej identifikácie presný. Môže to byť v prípade, že zákazníci zo Skupiny 2 neprinášajú žiadne veľké zisky. V prípade, že vysoká návratnosť investícií pochádza od zákazníkov zo Skupiny 1 alebo 4, môže model poskytnúť informácie, ktoré potrebujeme.

## 5. Poďakovanie

Tento príspevok je čiastkovým výstupom projektu VEGA č. 1/0283/15 „Aspekty marketingovej komunikácie v oblasti procesu tvorby hodnoty zákazníka na trhu B2C v kontexte s maximalizáciou trhového podielu v nákupnom spáde maloobchodu.“

## 6. Zoznam bibliografických odkazov

(1) IBM SPSS Modeler 14.2 Algorithms Guide, IBM 2011

(2) IBM SPSS Modeler 14.2 Applications Guide, IBM 2011

(3) AHN, J. - HAN, S. – LEE, Y. : Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. In. Telecommunications Policy 30(10-11), pp. 552–568, 2006. ISSN 0308-5961

(4) LEE, K. Ch. - JO, N. Y.: Bayesian Network Approach to Predict Mobile Churn Motivations: Emphasis on General Bayesian Network, Markov Blanket, and What-If Simulation. In. Second International Conference, FGIT 2010, Jeju Island, Korea, December 13-15, Springer, 2010. 978-3-642-17568-8

(5) HALENAR, R: Matlab Routines Used for Real Time ETL Method. In: Applied Mechanics and Materials, 229-231 (2012), s. 2125-2129. - ISSN 1660-9336

(6) VAŽAN, P. - TANUŠKA, P.- JUROVATÁ, D.- KEBÍSEK, M.: Analysis of Production Process Parameters by Using Data Mining Methods. In: Applied Mechanics and Materials. Vol. 309, 2013 ISSN 1660-9336

## 7. Adresa autora:

Andrej Trnka, Ing. PhD.
Univerzita sv. Cyrila a Metoda v Trnave
Fakulta masmediálnej komunikácie
Nám. J. Herdu 2
917 01 Trnava
andrej.trnka@ucm.sk