Primer uporabe podatkovnega rudarjenja v skupini NLB

Peter Konda, Jure Peljhan
NLB, Ljubljana
peter.konda@nlb.si; jure.peljhan@nlb.si

Izvleček

V zadnjem času smo priča velikim premikom na področju poslovnega obveščanja, kamor uvrščamo tudi podatkovno rudarjenje, ki ga v skupini NLB uporabljamo za napoved naklonjenosti strank k sklenitvi depozita. Algoritmi, ki rešujejo takšne klasifikacijske probleme, so odločitvena drevesa, nevronske mreže, naivni Bayes in logistična regresija. Rezultat uporabe algoritma je model – repozitorij pravil za obravnavani poslovni problem.
Uspešna implementacija zahteva sodelovanje na več organizacijskih ravneh, kot so poslovna mreža, trženje in informacijska tehnologija. Uporaba preizkušenih metodologij zagotavlja kakovostno izvedbo vseh stopenj razvoja, od določitve poslovnega cilja do uporabe rezultatov. Modeli, ki jih uporabljamo v skupini NLB, dosegajo višjo natančnost od segmentno usmerjenih metod, kar predstavlja nov korak v smeri uporabe prodajnih poti po meri posameznika.
Ključne besede: podatkovno rudarjenje, bančni informacijski sistem, poslovno obveščanje, metodologija CRISP-DM, algoritem.

Abstract

A CASE OF DATA MINING IN NLB BANK

Recently we have witnessed major developments in the field of business intelligence, which also includes data mining. NLB group uses the latter for predicting deposits propensity score. Algorithms used for solving such classification problems are decision trees, neural networks, naive Bayes and logistic regression. Data mining algorithm outputs a model – repository of rules for a certain problem.
Successful implementation requires cooperation at various organizational levels: branch network, marketing and ICT. Using proper methodologies ensures a high quality of all phases of development, from the establishment of business objectives to the end results. Resulting models in NLB group are achieving a higher precision than segment-oriented methods. This represents a new step in individualizing our customers' needs.
Key words: data mining, banking information system, business intelligence, CRISP-DM methodology, algorithm.

1 UVOD

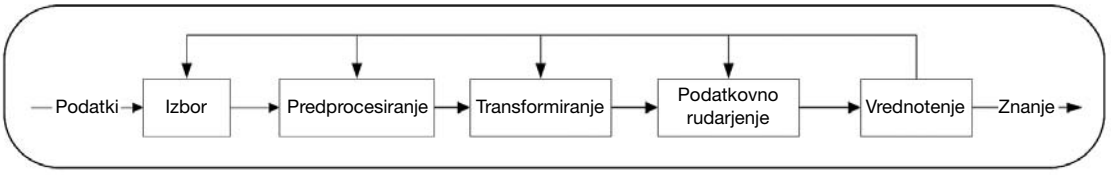
Posledice svetovne krize so najpogostejši razlog za slabše poslovanje bank v preteklem letu. Vzrok za to pa je predvsem v nezadostni uporabi podatkov za poslovno odločanje pri sprejemanju poslovnih odločitev. Te so sedaj prisiljene racionalizirati poslovanje, za kar obstaja več možnosti. Ena izmed njih je uporaba inovativnih rešitev na področju informacijske tehnologije. V NLB smo v ta namen uporabili tehnike poslovnega obveščanja (angl. business intelligence). Poslovno obveščanje je novo področje, ki je v zadnjih letih vse bolj pomembno in ga tudi po ocenah Gartnerja čaka svetla prihodnost (Gartner, 2009).

Podatkovno rudarjenje predstavlja samo en del poslovnega obveščanja, vendar z njegovo uporabo hitro dosežemo znižanje oz. boljše obvladovanje stroškov. Vsi poznamo danes že legendarni primer ameriške trgovske verige Walmart z zlaganjem otroških plenic in piva na skupno polico. V bančništvu uporabljamo podatkovno rudarjenje npr. za razvrščanje komitentov v skupine (segmentacija), napoved plačilne sposobnosti, analizo prebegov ipd. V NLB po določenih kriterijih mesečno izbiramo tiste komitente, ki bodo z veliko verjenostjo sklenili depozit. S podatkovnim rudarjenjem smo poskušali izboljšati natančnost obstoječih metod, pri tem pa proces napovedi sklenitve depozita avtomatizirati po metodologiji CRISP-DM.

2 PODATKOVNO RUDARJENJE

Pojem podatkovno rudarjenje (angl. data mining) se je pojavil v devetdesetih letih prejšnjega stoletja. Temelji te tehnične stroke so bili postavljeni v petdesetih letih s pojavom strojnega učenja (angl. machine learning). Takrat so razvili prve algoritme za iskanje znanja, ki se v izboljšanih različicah uporabljajo še danes. Če smo zelo natančni, so prave temelje postavili že prvi statistiki z opredelitvijo osnovnih pojmov, kot so enota, populacija, vzorec in spremenljivka.

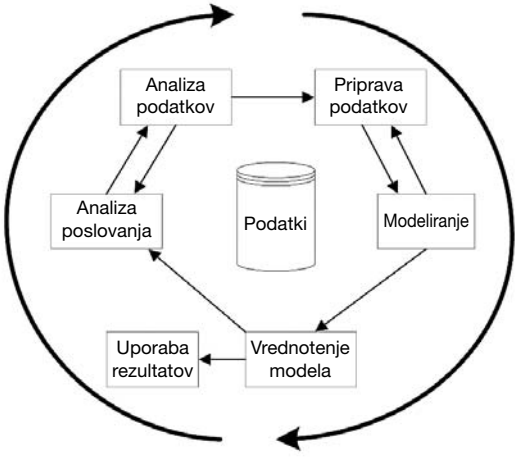
Podatkovno rudarjenje je del procesa KDD (angl. Knowledge Discovery in Database), ki se osredinja na iskanje znanja v poljubnih vhodnih podatkih. Prikazan je na sliki 1.



Slika 1: **Podatkovno rudarjenje kot del procesa KDD (Kononenko in Kukar, 2007)**

Proces KDD se opira na več strok, kot sta strojno učenje in statistika. Podatkovno rudarjenje predstavlja ključni korak procesa KDD. Prva odprta metodologija, ki pokriva vse faze razvoja podatkovnega rudarjenja, je CRISP-DM (angl. Cross-Industry Standard Process for Data Mining). Temelje za njen nastanek so postavila podjetja Daimler Chrysler, SPSS in NCR. Leta 1997 so oblikovala konzorcij s ciljem razviti standardni industrijski proces za podatkovno rudarjenje. Namenjen naj bi bil za uporabo v katerem koli okolju neodvisno od programskega orodja in gospodarskega področja. Izkušnje iz prakse ter mnenja o tem, kako izboljšati proces podatkovnega rudarjenja, so pridobili na odprtih delavnicah (Shearer, 2000). Rezultat dela konzorcija je standard CRISP-DM 1.0, ki je nastal leta 2000 (Chapman in drugi, 1999).

CRISP-DM je splošno razumljiva metodologija za podatkovno rudarjenje. Razčlenjena je na šest razvojnih stopenj. Stopnje se naprej razčlenijo na več opravil.



Slika 2: **Razvojne stopnje metodologije CRISP-DM (Kononenko in Kukar, 2007)**

Notranje puščice prikazujejo povezanost med stopnjami. Zunanji krog simbolizira iterativno naravo samega podatkovnega rudarjenja.

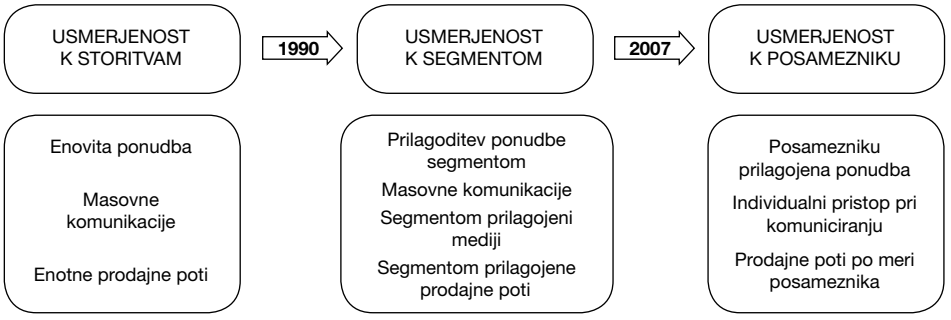
Uspešna izvedba projekta je odvisna od več dejavnikov: primerna podatkovna platforma, raven sodelovanja med organizacijskimi enotami, interes uporabnikov in spremljanje kakovosti. Zaradi slabih izkušenj iz preteklosti uporabniki podatkovnega rudarjenja predlagajo spremembe v metodologiji CRISP-DM, tako da bi ta bolj poudarjala akademsko in gospodarsko sodelovanje (Ghani in Soarez, 2009). Večina proizvajalcev programske opreme teži k večji integraciji svojih produktov, zato je smiselno izbrati čim manjše število različnih orodij in podatkovnih platform. S tem se izognemo visokim stroškom razvoja in vzdrževanja.

3 IMPLEMENTACIJA PODATKOVNEGA RUDARJENJA V NLB

V NLB sledimo sodobnim trendom v informatiki, zato na veliko področjih uvajamo inovativne rešitve že v zgodnjih fazah. Za podatkovno rudarjenje uporabljamo podatkovno platformo SQL Server 2008, ki se je izkazala za zmogljivo in skalabilno. Nekatera orodja (npr. odprtokodna Weka) imajo s skalabilnostjo težave, zato so neprimerna za produkcijsko okolje (Konda, 2009) v velikih sistemih, ki zahtevajo kompleksne analize na velikem številu podatkov.

3.1 Informacijski razvoj trženja v banki

Razvoj tehnik poslovnega obveščanja v NLB poteka že več let. Dosedanji informacijski razvoj trženja v NLB prikazuje slika 3.

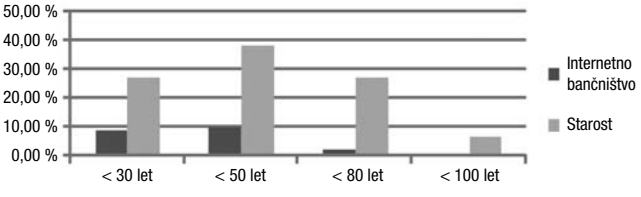


Slika 3: **Prehod iz storitvenega trženja k individualiziranemu**

Podatkovno rudarjenje se uveljavlja v zadnji razvojni fazi, prikazani na sliki 4. Razlogi za njegovo vpeljavo so predvsem v:

- večji natančnosti pri izbiri potencialnih strank,
- hitrosti izvedbe projekta, saj z orodji za hiter razvoj (angl. Rapid Application Development) prej pridemo do boljših rezultatov,
- merljivosti natančnosti, saj z individualnim pristopom dobimo ustrezen odziv stranke, tega pa lahko uporabimo za oceno kakovosti napovedi.

Rezultate lahko neposredno uporabljata kontaktni center in poslovna mreža. Oddelek za raziskave in analize periodično primerja rezultate modelov z dejanskimi podatki iz poslovanja. Spremljanje kakovosti modelov je ključno za uspešno podatkovno rudarjenje.



Slika 4: **Razmerje med starostjo strank in uporabo internetnega bančništva**

3.2 Analiza in priprava podatkov

NLB uporablja podatkovno skladišče, ki je osnovano na IBM-ovih tehnologijah s podatkovno zbirko DB2, zato je treba podatke pred začetkom modeliranja prenesti na platformo SQL Server 2008. Temu postopku pravimo tudi ETL (angl. Extract, Transform, Load). Rezultat transformacije je tabela, ki ima v vrsticah stranke, v stolpcih pa njihove atribute, npr. starost, kraj bivanja, število sklenjenih storitev ipd. Ključen je binarni razredni atribut, ki pove, ali je stranka v določenem obdobju sklenila depozit ali ne.

Za analizo podatkov uporabljamo statistične metode. Korelacija denimo pokaže moč linearne povezanosti med posameznimi atributi. Z uporabo grafov lahko hitro ugotovimo frekvenčne porazdelitve posameznih atributov, kar je na začetku koristno za spoznavanje s podatki. Na sliki 4 je primer takšne porazdelitve.

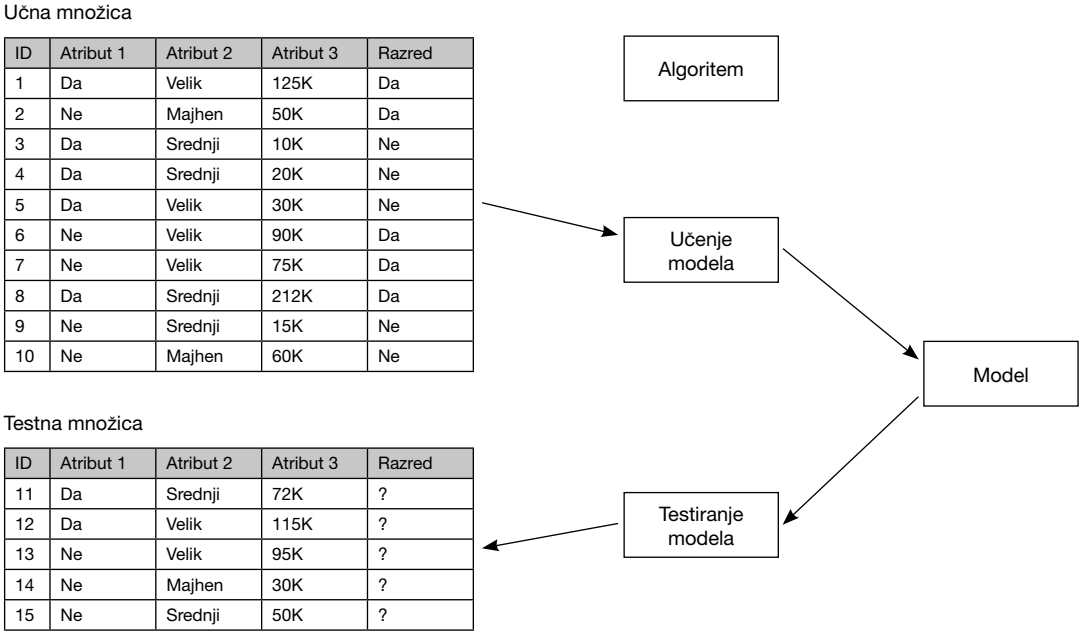
Seznam atributov komitenta za modeliranje pripravimo skupaj z tržnimi analitiki in komercialisti. Prvi poznajo objektivne razloge za sklenitev depozita, drugi pa subjektivne, saj so dnevno v stiku s komitenti.

3.3 Algoritmi

Algoritmi za podatkovno rudarjenje se delijo v dve skupini: nadzorovani (angl. supervised) in nenadzorovani (angl. unsupervised). Pri nadzorovanih se odvisna spremenljivka izračuna na podlagi neodvisnih. Rečemo tudi, da ti algoritmi potrebujejo učitelja (odvisno spremenljivko), da se lahko učijo. Nenadzorovani algoritmi obravnavajo vse spremenljivke neodvisno. Takšni algoritmi se ne učijo na podlagi ciljne spremenljivke, temveč skozi serijo ponovitev konvergirajo proti cilju. Takšen primer je segmentacija, pri kateri cilj predstavlja stabilno ločnico med posameznimi segmenti.

Za izračun naklonjenosti stranke k sklenitvi depozita se uporabljajo klasifikacijski (nadzorovani) algoritmi, npr. odločitvena drevesa, nevronske mreže, naivni Bayes in logistična regresija. Klasifikacija pomeni uvrščanje objektov (komitentov) v binarni razred: 1 – sklene depozit, 0 – ne sklene depozita. Vsako stranko opisujejo določeni atributi (lastnosti). Atributi so lahko diskretne ali zvezne neodvisne spremenljivke. Vrednost binarnega razreda se izračuna iz vrednosti neodvisnih spremenljivk. Algori-

tem iz učne množice podatkov inducira pravila za klasifikacijo strank, ki so shranjena v modelu. Pravilnost modela se preveri na testnih podatkih. Postopek učenja in testiranja modela prikazuje slika 5.



Slika 5: **Postopek reševanja klasifikacijskega problema**

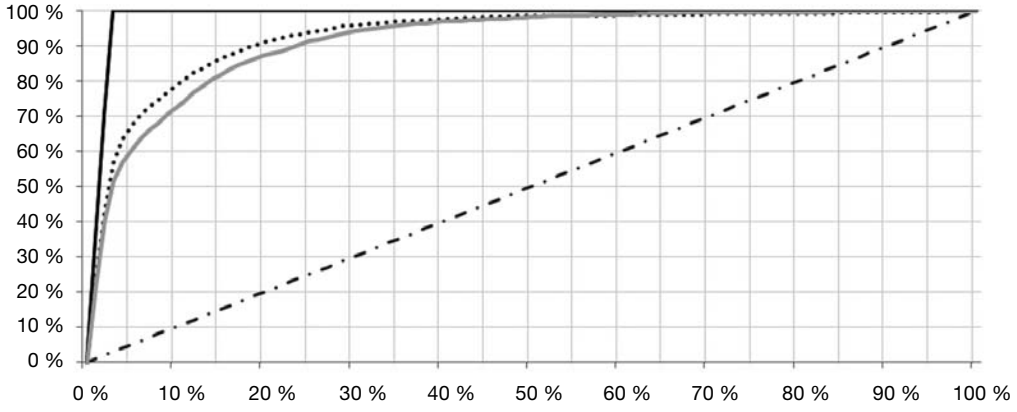
Posebno pozornost je treba posvetiti izbiri učne in testne množice. Običajno imamo pri podatkovnem rudarjenju opravka z veliko količino neenakomerno razporejenih podatkov, zato uporabimo vzorčenje (SAS Institute Inc., 1998). Razmerje odzivnih strank proti neodzivnim je v celotni populaciji majhno. To lahko povzroča težave pri indukciji pravil. Ta so lahko ali preveč prilagojena učni množici ali pa preveč splošna, zaradi česar je model nenatančen. Problem neuravnovešenosti klasifikatorja v podatkih stroka rešuje na različnih delavnicah (Chawla, Japkowitz in Kolcz, 2004).

3.4 Modeliranje in vrednotenje rezultatov

Pripravi podatkov sledi modeliranje, to je uporaba ustreznega algoritma, ki pravila shrani v modelu. Platforma SQL Server 2008 ima za klasifikacijo na voljo naslednje algoritme: odločitvena drevesa, nevronske mreže, naivni Bayes in logistična regresija. Vsi algoritmi z izjemo naivnega Bayesa uporabljajo diskretne in zvezne vhodne atribute. Med vhodne atribute glede na poslovni problem štejemo demografske podatke (kraj bivanja, starost), podatke o poslovanju (uporaba mobilnega bančništva, število sklenjenih storitev) in podatke o materialnem stanju (sredstva, obveznosti). Vrednost izhodnega atributa je binarna in pove, ali je stranka v določenem obdobju sklenila depozit ali ne. Rezultati procesiranja modela se shranijo v podatkovni bazi.

Natančnost napovedi štirih modelov preverimo na testni množici podatkov z uporabo dveh metod; to sta odzivni diagram (angl. Lift Chart) (Vuk in Curk, 2006) in križno preverjanje na podmnožicah (angl. K-fold Cross Validation) (Microsoft, Cross-Validation (Analysis Services - Data Mining)).

Odzivni diagram za dva algoritma je na sliki 6.



Slika 6: **Nevronska mreža ima pri enaki velikosti vhodne populacije večji odziv kot naivni Bayes.**

Slika prikazuje napovedno moč modela. Ta pri 10 odstotkih populacije pravilno napove kar 80 odstotkov vseh strank, ki so sklenile depozit. V teoriji to pomeni v primerjavi z naključnim izborom osemkratno zmanjšanje stroškov, potrebnih za kontaktiranje strank.

Primerjava med algoritmi kaže, da razlik med algoritmi – z izjemo naivnega Bayesa – tako rekoč ni. Do razlik pride šele z uporabo križnega preverjanja. Ta metoda na podmnožicah izvede klasifikacijski algoritem in izračuna statistične metrike: število pravilno klasificiranih primerov, kvadratni koren povprečne kvadratne napake (RMSE) ipd. Rezultati križnega preverjanja kažejo, da so odločitvena drevesa nekoliko boljša od preostalih algoritmov. V praksi je dobro izbrati algoritem, ki deluje dobro ne glede na vhodno množico. Takšen je denimo algoritem logistična regresija, ki se je v NLB izkazal za uspešnega pri napovedi sklenitve depozitov. Napovedna moč tega modela je boljša od segmentno usmerjenih metod, ki ciljajo na določene značilnosti strank, npr. starostni segment (mladi, zaposleni, upokojenci) ali tip osebnega računa (študentski, srebrni, zlati itn.).

3.5 Uporaba rezultatov

V NLB rezultate uporabljajo komercialisti v poslovalnicah in kontaktni center. Komercialisti za poslovanje s fizičnimi osebami uporabljajo informacijski sistem za podporo poslovanju na bančnem okencu (NBO). Njegovo programsko ogrodje omogoča preprost prikaz rezultatov iz podatkovnega skladišča z uporabo transakcij. Pri tem je treba paziti, da s prenosom rezultatov iz platforme SQL Server 2008 ne zaklenemo zapisov v tabelah podatkovnega skladišča DB2. Temu se izognemo tako, da rezultate najprej zapišemo v klonirano ciljno tabelo, nato pa podatke preklopimo naenkrat.

Komercialisti depozit ponudijo tistim strankam, ki imajo v podatkovnem skladišču zapisano visoko verjetnost za sklenitev depozita, npr. nad 95-odstotno. NBO uporablja sistem za povratno informacijo strank o njihovi odločitvi. Drugi del uporabnikov predstavlja kontaktni center, ki izvaja neposredno trženje. Ti uporabniki prejmejo seznam najbolj potencialnih strank za sklenitev depozita. Seznam je lahko v poljubni obliki, pomembni so le kontaktni podatki osebe in njen odziv na ponudbo. V preskusnem obdobju seznam omejimo na 100–200 strank, kasneje pa število omejimo glede na pričakovani donos. Kontaktni center po koncu akcije pošlje povratno informacijo oddelku za raziskave in analize, ki preveri ujemanje napovedi z dejansko sklenjenimi posli. Povratne informacije uporabljamo pri nadaljnjem razvoju modelov.

4 SKLEP

Opisani primer potruje Gartnerjevo napoved o povečevanju vlaganj v poslovno obveščanje. S projektom napovedi sklenitve depozitov bi v teoriji lahko za osemkrat zmanjšali stroške obveščanja komitentov. Čeprav bi bil ekonomski učinek manjši, podatkovno rudarjenje kaže velik potencial.

Uporabo podatkovnega rudarjenja bomo v NLB razširili tudi na druge storitve, kot so sklepanje kreditov, kartično poslovanje, obvladovanje kreditnega tveganja ipd. Dodatno bomo gradili tudi na predvidevanju potreb strank z uporabo naprednih analitik. S premišljenim načrtovanjem lahko obstoječo struk-

turo podatkov uporabimo pri vseh modelih, pri katerih so komitenti v središču opazovanja. Razvoj poslovnega obveščanja usmerjamo v združevanje sorodnih družin izdelkov, kot so poročila, OLAP-kocke in podatkovno rudarjenje. Rezultate tovrstnega razvoja prikazujemo na sodobnih portalih, kot je npr. SharePoint. S tem se izognemo visokim stroškom lastnega razvoja, še posebno ker gre za tehnologije istega proizvajalca. Tovrstno združevanje na skupnem portalu ima za posledico večjo povezanost sodelavcev iz različnih oddelkov, kar je ključnega pomena pri učinkovitem razvoju informacijske tehnologije. Z uporabo podatkovnega rudarjenja smo v NLB dosegli pomembne poslovne učinke z uporabno najsodobnejših tehnologij in pripravili temelje za doseganje večje učinkovitosti ter širitve znanja znotraj celotne skupine NLB.

5 VIRI IN LITERATURA

[1] Gartner. (2009). *Gartner Reveals Five Business Intelligence Predictions for 2009 and Beyond.* Pridobljeno 8. 9. 2010 s http://www.gartner.com/it/page.jsp?id=856714.

[2] Kononenko, I. in Kukar, M. (2007). *Machine learning and data mining.* Chichester: Horwood Publishing, Ltd.

[3] Ghani, R. in Soarez, C. (2009). Editorial: Data Mining for Business Applications. *Special Interest Group on Knowledge Discovery and Data Mining*.

[4] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, V* (4), str. 13–21.

[5] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. idr. (1999). *CRISP-DM 1.0*

[6] *Step-by-step data mining guide*. Pridobljeno 8. 3. 2010 s http://www.crisp-dm.org/.

[7] SAS Institute Inc. (1998). *Data Mining and the Case for Sampling.* SAS Institute Inc.

[8] Chawla, N. V., Japkowitz, N. in Kolcz, A. (2004). *Editorial: Special Issue on Learning from Imbalanced Data.* Pridobljeno 8. 3. 2010 s http://www.sigkdd.org/explorations/issues/6-1-2004-06/edit_intro.pdf.

[9] Konda, P. (2009). Izvedba podatkovnega rudarjenja v bančništvu z uporabo metodologije CRISP-DM. Diplomska naloga.

[10] Vuk, M. in Curk, T. (2006). ROC Curve, Lift Chart and Calibration Plot. *Metodološki zvezki, 3* (1).

[11] Microsoft. *Cross-Validation (Analysis Services - Data Mining)*. Pridobljeno 8. 3. 2010 s: http://msdn.microsoft.com/en-us/library/bb895174.aspx.

▪

Peter Konda je diplomiral na Fakulteti za računalništvo in informatiko Univerze v Ljubljani na temo podatkovnega rudarjenja. Leta 2009 se je zaposlil v upravljalskem centru za informacijsko tehnologijo v NLB. Trenutno se ukvarja z novimi področji uporabe podatkovnega rudarjenja in integracijo storitev poslovnega obveščanja na SharePoint portalu.

▪

Jure Peljhan je leta 1998 diplomiral na Fakulteti za organizacijske vede v Kranju Univerze v Mariboru. Na Ekonomski fakulteti Univerze v Ljubljani je leta 2002 je končal specialistični študij, leta 2003 pa še magistriral s področja celovitega obvladovanja kakovosti. Prispevki s tega področja so bili predstavljeni tudi na mednarodnih konferencah v Benetkah in Zagrebu. Poklicno pot je začel v skupini NLB kot sistemski analitik programer na področju plačilnega prometa v okolju centralnega računalnika ter nadaljeval na področju spletnih tehnologij. Vodil je projekte s področij kadrovskega, plačnega, izobraževalnega informacijskega sistema, skrbniških storitev vzajemnih in pokojninskih skladov ter menedžerskega informacijskega sistema. Bil je svetovalec člana uprave, odgovornega za področje informacijske tehnologije, in direktor sektorja za strateško načrtovanje in upravljanje informacijskega sistema banke. Od leta 2008 dela kot direktor centra za informacijsko tehnologijo v skupini NLB.