UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

DIPLOMSKO DELO

PODATKOVNO RUDARJENJE IN KXEN ANALITIČNO OGRODJE

Ljubljana, januar 2005 **MARJAN ZIDAR**

IZJAVA

Študent Marjan Zidar izjavljam, da sem avtor tega diplomskega dela, ki sem ga napisal pod mentorstvom dr. Jurija Jakliča, in skladno s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah dovolim objavo diplomskega dela na fakultetnih spletnih straneh.

Podpis:_____

V Ljubljani, dne 13.01.2005

KAZALO

1. UVOD

1.1. OPIS PROBLEMATIKE

Poslovno okolje podjetij postaja čedalje bolj kompleksno. Povečuje se količina podatkov, na podlagi katerih podjetja poslujejo, povečuje se število spremenljivk, ki vplivajo na poslovne odločitve, postopki odločanja so čedalje bolj zapleteni in nepredvidljivi. Metode, ki so jih podjetja uporabljala za uspešno poslovanje v preteklosti, so se izkazale za neuporabne, saj niso več sposobne zagotavljati zadovoljivih rezultatov. Poslovanje s pomočjo teh metod je postalo neobvladljivo in nekonkurenčno.

Podjetja že nekaj desetletij iščejo načine, ki bi omogočili kvalitetno prilagajanje spremembam, ki se pojavljajo v vsakodnevnem poslovanju. Odkrila so načine, ki omogočajo izrabo prednosti kompleksnejšega sveta. S pridom so podjetja začela izrabljati vse podatke, ki so jih beležila ob svojem poslovanju. Zavedati so se začela, da so zbrani podatki lahko prednost in ne nujno problem, ki spremlja vsakodnevno poslovanje. Ugotovila so namreč, da se lahko na podlagi preteklih podatkov naučijo marsikaj, kar jim v prihodnjem poslovanju prinaša konkurenčne prednosti. Rodilo se je t.i. podatkovno rudarjenje, kateremu je tudi namenjeno to diplomsko delo.

1.2. NAMEN IN CILJ DIPLOMSKE NALOGE

Moj namen v tem diplomskem delu je predvsem predstaviti bralcu tehniko podatkovnega rudarjenja kot veliko priložnost, ki se ponuja podjetjem na najrazličnejših gospodarskih področjih. V delu nisem želel pretiranega poudarka dajati teoretičnemu delu, saj je gradiva na temo podatkovnega rudarjenja dovolj.

Veliko pozornosti sem namenil analizi enega izmed boljših orodij za podatkovno rudarjenje. Skušal sem predstaviti KXEN analitično ogrodje, ki je v boljših podjetjih po svetu že zelo znano in ta podjetja dosegajo odlične rezultate s pomočjo tega orodja, v slovenskem prostoru pa še ni dovolj znano, da bi podjetja lahko posegala po priložnostih, ki jih ponuja. V posameznih točkah sem omenjeno orodje primerjal tudi z drugimi orodji, ki se pojavljajo kot konkurenti oz. bolje rečeno kot alternativne rešitve. Veliko pozornosti namenjam predstavitvi lahkotnosti uporabe sodobnih orodij za podatkovno rudarjenje. Zaradi možnosti, ki sem jih imel, lahkotnost uporabe predstavljam z vidika orodja KXEN, postopek pa je podoben tudi v ostalih orodjih za podatkovno rudarjenje. Namen diplomskega dela je zagotovo tudi opogumiti bralce oz. podjetja, ki uporabljajo »klasične[1] metode« podatkovnega rudarjenja, da

[1]Mišljene so metode oz. postopki, ko v podjetjih s pomočjo statističnih znanj skušajo ugotoviti določene zakonitosti, ki vladajo na področju poslovanja. Kot alternativna možnost klasični metodi podatkovnega rudarjenja se pojavljajo programska orodja, ki proces podatkovnega rudarjenja avtomatizirajo in poenostavijo, hkrati pa nudijo boljše rezultate.

se pustijo prepričati o kvaliteti sodobnih orodij za izkoriščanje uporabnih informacij, ki se skrivajo v podatkovnih bazah, in jih do sedaj še niso odkrili.

1.3. METODE DELA

Prvi del diplomskega dela temelji na proučevanju literature, ki že obstaja na temo podatkovnega rudarjenja. V tem delu sem skušal zbrati nekaj teoretičnih pogledov različnih avtorjev na temo podatkovnega rudarjenja. V drugem delu, ko pišem konkretno o KXEN-u, pa gre za predstavitev orodja, ki sem ga v praksi imel možnost tudi sam preizkusiti in uporabljati. V tem delu sem se opiral predvsem na interne vire podjetja KXEN in pa tudi na lastne izkušnje, v veliki meri pa sem si lahko pomagal z orodjem samim.

1.4. STRUKTURA DIPLOMSKE NALOGE

Diplomska naloga je razdeljena na štiri dele. V uvodnem delu je opredeljena problematika diplomskega dela, namen in cilj tega dela, predstavljene so metode dela, podana pa je tudi struktura diplomske naloge.

Drugi del diplomske naloge je namenjen teoretičnim spoznanjem s področja podatkovnega rudarjenja. Podana je opredelitev samega pojma, poleg tega pa so na kratko predstavljene tudi vse najpogosteje omenjene tehnike podatkovnega rudarjenja. Predstavljene so tako tehnike najbližjega soseda, razvrščanja v skupine, inducirana pravila, odločitvena drevesa in tudi nevronske mreže. Na koncu drugega dela sem podal še nek logičen sklep, ki se nanaša na prihodnost podatkovnega rudarjenja.

Tretji del je v celoti namenjen spoznavanju KXEN analitičnega ogrodja, ki je trenutno eno boljših orodij za podatkovno rudarjenje. V začetku tretjega dela je najprej predstavljen KXEN na splošno. Predstavljene so bistvene razlike med klasičnim in KXEN pristopom k podatkovnem rudarjenju. Predstavljene so bistvene prednosti, ki jih ponuja KXEN analitično ogrodje. V nadaljevanju je predstavljena struktura samega analitičnega ogrodja, saj ne gre zgolj za eno komponento, pač pa za kar osem posameznih komponent, ki so vsaka zase nepogrešljiv del orodja. Po razlagi posameznih komponent orodja so predstavljeni osnovni koraki, ki jih moramo prehoditi, da zgradimo kvaliteten in zanesljiv napovedovalni model. Postopek izgradnje napovedovalnega modela sem vključil v diplomsko delo tudi zaradi dejstva, da si ogromno ljudi ne zna niti predstavljati, kako poteka v praksi podatkovno rudarjenje, posledično pa se neznane novosti tudi bojijo in je ne sprejmejo. Naslednji del je namenjen predstavitvi dveh primerov uporabe KXEN-a v praksi. Prvi primer prikazuje uporabo KXEN-a na področju telekomunikacij, drugi primer pa je s področja bančništva.

Četrti del diplomske naloge predstavlja zaključek, v katerem sem skušal diplomsko nalogo zaokrožiti v smiselno celoto.

2. POPODATKOVNO RUDARJENJE

2.1. OPREDELITEV POJMA PODATKOVNO RUDARJENJE

Preden se lotim razlage pojma podatkovno rudarjenje, je potrebno opredeliti še nekatere druge pojme, ki bodo kasneje služili lažjemu razumevanju. V mislih imam predvsem pojme, kot so model, vzorec in vzorčenje.

Model je v Slovarju slovenskega knjižnega jezika opredeljen kot: »predmet, izdelan za ponazoritev, prikaz načrtovanega ali obstoječega predmeta« (Slovar slovenskega knjižnega jezika, 1997). V smislu teme diplomskega dela pa je kot model mišljen opis zgodovinskih podatkov, ki ga je mogoče ponovno uporabiti za nove podatke z namenom napovedovanja manjkajočih oz pričakovanih vrednosti (Berson, Smith, Thearling, 2000, str.110).

Vzorec je dogodek oz. kombinacija dogodkov, ki se pojavlja v analiziranih podatkih bolj pogosto, kot je bilo mogoče pričakovati v primeru naključij. Vzorec je odraz podatkov samih. Običajno jih ponazarjamo grafično, saj si jih tako lažje predstavljamo (Berson, Smith, Thearling, 2000, str.110).

Vzorčenje je postopek, ko za analiziranje ne uporabljamo celotnega nabora podatkov, ki jih imamo v podatkovnih bazah. Običajno je podatkov enostavno preveč, da bi lahko analizirali vse, obdelava bi bila prezahtevna, zato v analizo vključimo le majhen del vseh podatkov, ki so na razpolago. Za oblikovanje nabora vzorčnih podatkov lahko uporabimo različne pristope, kot npr. naključno vzorčenje (ang. random sampling). Poenostavljeno povedano je vzorčenje reduciranje podatkov iz podatkovne baze, na podlagi teh podatkov pa iščemo napovedovalne modele (Berson, Smith, Thearling, 2000, str.110).

V teoriji ni enotne definicije pojma **podatkovno rudarjenje**, saj posamezni avtorji podajajo različne definicije pojma. Najenostavneje bi lahko podatkovno rudarjenje opredelili kot avtomatizirano iskanje pomembnih vzorcev v bazi podatkov (Berson, Smith, Thearling, 2000, str.110). Berry in Linoff (2000, str. 7) opredeljujeta pojem kot proces avtomatskega ali polavtomatskega analiziranja velikih količin podatkov, pri čemer je namen odkriti nove, zanimive uporabne vzorce in pravila. Pri podjetju SAS Institute[2] štejejo v okrilje pojma podatkovno rudarjenje »napredne metode za odkrivanje in modeliranje povezav na velikih količinah podatkov«. Definicija, ki jo podaja Pirc, je: »proces zbiranja, preučevanja in modeliranja velikih količin podatkov, da bi odkrili prej neznane vzorce in pravila v podatkih, kar je konkurenčna prednost«. S svojimi besedami pa lahko podatkovno rudarjenje opredelim kot postopek pridobivanja koristnih informacij iz velike količine podatkov, ki so shranjeni v podatkovnih bazah. Informacije pridobimo s postopkom, ki se ga je v teoriji in praksi prijel

[2] SAS Institute, Inc. – ameriško podjetje, ki se ukvarja z razvojem programske opreme.

termin »rudarjenje« oz. kopanje. V slovenskem jeziku sta se uveljavila predvsem pojma podatkovno rudarjenje in izkopavanje podatkov, v obeh primerih pa je mišljena ista stvar.

V preteklosti so za podatkovno rudarjenje podjetja imela vrsto statistikov, ki so »ročno« pregledovali podatke in s pomočjo matematičnih, statističnih in drugih metod odkrivali določene vzorce oz. pravila, na podlagi katerih so kasneje zgradili napovedovalne modele. Danes orodja za podatkovno rudarjenje statistikov niso izpodrinila, čeprav se je zaradi najrazličnejših dejavnikov (trend zbiranja podatkov v podjetjih, priprava podatkovnih skladišč v podjetjih, močni konkurenčni pritiski, pocenitev računalniške opreme,…) na trgu pojavilo kar nekaj orodij, ki so v določenem delu nadomestila delo statistikov. Statistikom se danes skoraj ni več potrebno ukvarjati s pripravo podatkov in z »ročnim« odkrivanjem vzorcev, saj zato poskrbijo orodja, nepogrešljivo pa je njihovo znanje na področju interpretiranja dobljenih rezultatov. Največja slabost orodij za podatkovno rudarjenje je namreč ta, da ne poznajo izkušenj in ne poznajo intuicije za odkrivanje nepomembnih in pomembnih povezav. S tega vidika bodo ljudje v postopkih podatkovnega rudarjenja ostali nepogrešljivi.

Omeniti je potrebno, da poznamo **usmerjeno** (natančno vemo, katero spremenljivko postaviti kot ciljno) **in neusmerjeno** (ciljna spremenljivka ni določena oz. znana) **podatkovno rudarjenje**. Aktivnosti, ki jih lahko štejemo med usmerjeno podatkovno rudarjenje, so: klasifikacija, ocenjevanje in napovedovanje. Opisovanje in vizualizacija, asociacije in razvrščanje pa so aktivnosti neusmerjenega podatkovnega rudarjenja. Kot **klasifikacijo** je potrebno razumeti postopek, s katerim opazovane enote analiziramo in jih razporejamo v vnaprej oblikovane razrede. **Ocenjevanje** je postopek, s katerim na podlagi vhodnih podatkov ocenimo vrednost neke določene spremenljivke[3], npr. višino prometa v različnih poslovalnicah. **Napovedovanje** je aktivnost, pri kateri se opazovane enote klasificirajo na podlagi pričakovanega dogajanja v prihodnosti oz. na podlagi vrednosti spremenljivk v prihodnosti. Pri **opisovanju in vizualizaciji** gre za pojasnjevanje določenih zakonitosti, ki so značilne za podatke v podatkovnih bazah. Vizualizacija je le grafična predstavitev opisov. Pod pojmom **asociacija** pa razumemo ugotavljanje dejstev, na podlagi katerih določene stvari spadajo skupaj (Konič, 2003, str. 5). Primer je npr. pozicioniranje posameznih izdelkov v prodajalnah. Na Petrolu so npr. ugotovili na podlagi proučevanja obnašanja potrošnikov, da je potrebno čipse postaviti v prodajalnah blizu pijačam, saj je nekako logično, da k slanemu prigrizku sodi osvežilna pijača. Z analizo preteklih podatkov obnašanja potrošnikov so tako prišli do vzorca obnašanja, ki se med kupci slanih prigrizkov ponavlja. Zadnja aktivnost neusmerjenega podatkovnega rudarjenja pa je **razvrščanje v skupine**. Mišljeno je razvrščanje opazovanih metod v skupine oz. segmente. V isto skupino se razvrstijo enote, ki so si med seboj čim bolj podobne, same skupine pa naj bi bile čim bolj različne (Konič, 2003, str. 15).

[3] Spremenljivka predstavlja posamezno lastnost opazovanje enote. Pri avtomobilu bi lahko omenili naslednje spremenljivke: barva, moč motorja, vrsta motorja, število vrat,… (Blejec et al., 2003, str. 3).

2.2. TEHNIKE PODATKOVNEGA RUDARJENJA

Teorija omenja najpogosteje pet različnih tehnik, ki se uporabljajo v praksi. V nadaljevanju bom skušal prestaviti tehniko najbližjega soseda, tehniko razvrščanja v skupine, tehniko induciranih pravil, tehniko odločitvenih dreves in tehniko nevronskih mrež. Najbolj znana so verjetno drevesa odločanja, saj je uporaba enostavna. Tehnike so se v omenjenem zaporedju tudi razvijale, razvoj pa se je začel že pred mnogimi desetletji. Nevronske mreže so se začele razvijati šele proti koncu prejšnjega stoletja, ponujajo pa še precej možnosti razširitve in izboljšav.

2.2.1. NAJBLIŽJI SOSED

Najbližji sosed (ang. nearest neighbor) je ena izmed najstarejših tehnik, ki se uporablja na področju podatkovnega rudarjenja. Tehnika je enostavna in ljudem razumljiva, saj deluje na podoben način kot razmišljamo ljudje – iščemo podobnosti na podlagi katerih lahko podajamo določene sklepe. Za uporabo te tehnike je potrebno imeti podatke za preteklost, saj skuša tehnik na podlagi proučevanih enot iz preteklosti napovedati manjkajočo vrednost oz. določeno lastnost za enoto, za katero želimo napoved. Za določeno proučevano enoto želimo izdelati konkretno napoved, ki jo dobimo s pomočjo iskanja podobnih enot v preteklih podatkih. V bazi preteklih podatkov poiščemo enote, ki so glede na posamezne spremenljivke (spremenljivke, ki so znane za enoto, za katero izdelujemo napoved) najbolj podobne proučevani enoti in s pomočjo podatkov za najdene enote napovemo vrednost spremenljivki, ki jo iščemo za proučevano enoto.

Tehniko bi lahko ponazorili s praktičnim primerom iz prakse. Ko se odločamo za prodajo starega avtomobila, se največkrat pozanimamo pri podjetjih, ki se ukvarjajo s preprodajo avtomobilov, kakšna je bila realizirana prodajna cena podobnih avtomobilov (enaka znamka, isti letnik, podobno število prevoženih kilometrov, ohranjenost, dodatna oprema,…). Če so se podobni avtomobili v bližnji preteklosti prodajali za milijon tolarjev, obstaja verjetnost, da je tudi naše vozilo vredno približno milijon tolarjev. Če je iz preteklosti znanih več realiziranih prodaj s podobno ceno, je verjetnost, da se bo naš avtomobil prodal za podoben znesek, še toliko večja. Večja verjetnost nastopa določenega rezultata je namreč povezana z večjim številom podobnih realizacij iz preteklosti.

Tehnika najbližji sosed izgleda na primeru prodaje starega avtomobila povsem enostavna, saj je do zaključkov mogoče priti na enostaven način. Zavedati pa se je potrebno, da se v podjetjih srečujemo s kompleksnejšimi problemi oz. napovedmi, ki jih moramo rešiti. V teh primerih pa rešitve niso več tako trivialne, saj je »bližina« sestavljena iz množice dejavnikov, ki jih ni več mogoče razlagati oz. primerjati z enostavnim premislekom oz. »na pamet«.

2.2.2. RAZVRŠČANJE V SKUPINE

Pri tej tehniki (ang. clustering) gre za razvrščanje opazovanih enot v skupine, ki naj bi se med seboj čimbolj razlikovale, meje med posameznimi skupinami naj bi bile očitne, enote, ki bi se nahajale znotraj posamezne skupine, pa naj bi bile čimbolj homogene oz. čimbolj podobne. Pri tej tehniki v začetku analize ni znano, koliko skupin naj bi se oblikovalo, oz. na podlagi katerih spremenljivk naj bi se skupine oblikovale. Teoretično je mogoče, da so vse enote podobne med seboj in se ustvari zgolj ena skupina, možen pa je tudi primer, da vsaka preučevana enota tvori svojo skupino. V obeh skrajnih primerih uporabne vrednosti rezultata ni.

Orodja, ki uporabljajo tehniko razvrščanja v skupine, največkrat ponujajo možnost, da uporabnik sam izbere število skupin, ki jih želi tvoriti. Ta možnost je uporabna, saj uporabniki največkrat vedo, kakšne rezultate želijo (malo skupin s heterogenimi enotami oz. večje število skupin s homogenimi enotami). Običajno se glede na posamezne skupine enot oblikuje določene ukrepe, posameznik pa ve, koliko skupinam se je mogoče prilagoditi.

Omeniti velja opozorilo, ki ga podaja Edelstein (1999, str. 6). Tehnike razvrščanja namreč ne smemo mešati s segmentiranjem[4], saj pri segmentiranju načeloma že v začetku vemo, na kakšen način bodo oblikovane skupine, oz. na podlagi katerih spremenljivk se bodo oblikovali posamezni segmenti. Pri tehniki razvrščanja v skupine pa ne poznamo »sistema« razvrščanja. Segmente naj bi bil sposoben razlagati oz. interpretirati skorajda vsak, skupin, dobljenih z razvrščanjem, pa naj ne bi interpretiral nihče drug kot oseba, ki se na preučevano področje spozna.

Tehnika razvrščanja podatkov v skupine je namenjena predvsem ugotavljanju določenih zakonitosti, ki veljajo za posamezno podatkovno bazo. Z uporabo te metode dobimo boljši vpogled na strukturo in povezanost podatkov, ki jih imamo shranjene v bazah. To znanje o shranjenih podatkih omogoča izboljševanje sistemov za podporo odločanju, saj z lahkoto ugotavljamo, katere spremenljivke sploh niso relevantne za reševanje določenih poslovnih problemov, oz. ugotavljamo, katerim spremenljivkam je potrebno namenjati dodatno pozornost in katere je potrebno izpuščati iz analize.

2.2.3. INDUCIRANA PRAVILA

Tehnika induciranih (ang. induction rules) pravil je izmed opisanih tehnik verjetno najboljša (z vidika enostavnosti, predstavljivosti) za odkrivanje vzorcev, ki se pojavljajo v bazi podatkov. Ob uporabi te tehnike dobimo kot odgovor široko množico najrazličnejših povezav med različnimi neodvisnimi spremenljivkami. Pravila so definirana v obliki **če-potem** (if-then). Če nadaljujemo primer Petrolovih prodajaln, bi lahko bil primer induciranega pravila;

[4] Po SSKJ pomeni segmentirati razčlenjevati, segment pa pomeni del, odsek.

npr.: če kupec kupi slan čips, bo verjetno kupil tudi osvežilno pijačo. Verjetnost oz. veljavnost induciranih pravil je običajno enostavno izračunati, saj potrebujemo trivialen izračun. Izračunati je potrebno, kolikokrat se določen vzorec ponovi v celotnem naboru podatkov iz podatkovne baze. Človek, ki uporablja tovrstna pravila, pa se mora potem na podlagi dobljenih rezultatov odločati o ukrepih, ki bi dajali boljše poslovne rezultate. Osnovni namen podatkovnega rudarjenja je namreč izboljšanje poslovnih rezultatov.

Uporaba tehnike induciranih pravil je običajno zahtevna z vidike uporabe, saj se zahteva sistematično analiziranje vseh vzorcev, ki se pojavljajo v bazi podatkov. Podatkovne baze so velike, prav tako pa je običajno veliko število različnih vzorcev, ki se pojavljajo. Za vsak vzorec je potrebno določiti tudi stopnjo zaupanja oz. verjetnost ponovitve posameznega vzorca. Izračun verjetnosti je sicer enostaven, saj je potrebno zgolj sešteti vse ponovitve posameznega vzorca in jih deliti s celotnim številom dogodkov. Dobljeni rezultat se imenuje stopnja pokritja oz. podpora.

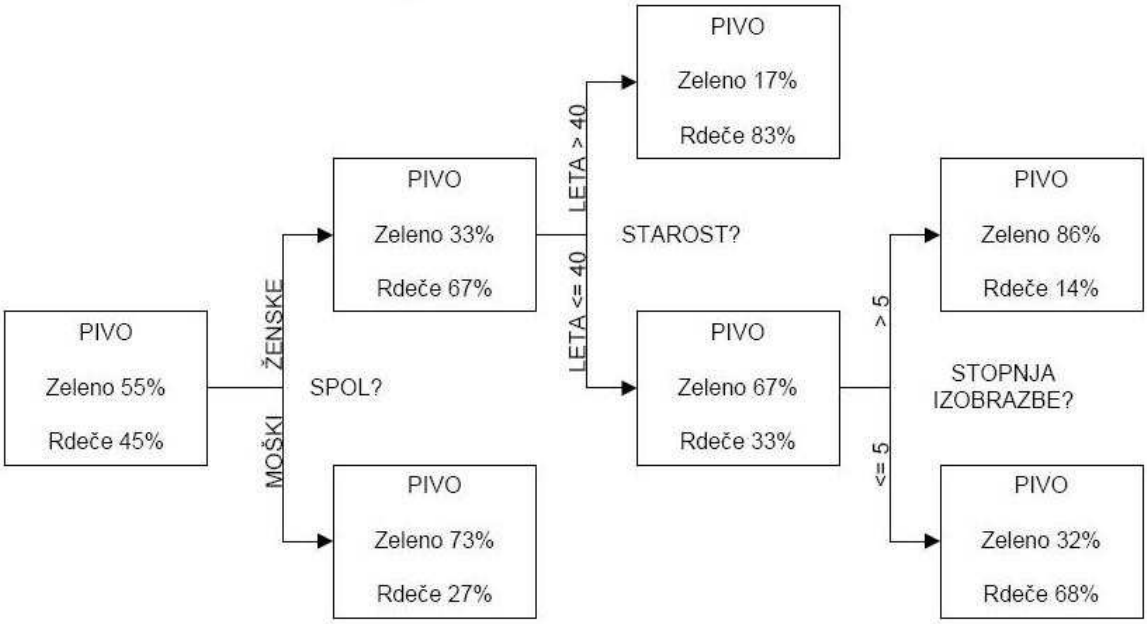
Velika prednost tehnike indukcijskih pravil je zagotovo možnost visoke avtomatičnosti in pa sistematične urejenosti pravil, ki jih običajno dobimo kot rezultat. Seveda je tehniko mogoče uporabljati tudi v primeru »ročnega« podatkovnega rudarjenja, ki pa v praksi zaradi kakovostnih orodij izginja. Tehnika omogoča podrobno spoznavanje poslovnih problemov.

2.2.4. ODLOČITVENA DREVESA

Že samo ime »odločitvena drevesa« (ang. decision trees) pove, da gre za napovedovalne modele, ki so zgrajeni v obliki dreves. Tehnika je bila prvotno razvita za potrebe statistike, danes pa je uporabljena za avtomatizacijo procesov odločanja. Ime po drevesu je tehnika dobila zaradi prikaza v obliki diagrama, ta diagram pa izgleda kot drevo. Diagram je sestavljen iz treh bistvenih komponent, in sicer vozlišč odločanja, vej in listov (Edelstein, 1999, str. 11). Berson podaja svojo definicijo drevesa, po kateri naj bi veje predstavljale določena vprašanja, ki služijo odločanju oz. razvrščanju podatkov. Listi drevesa naj bi po njegovem predstavljali posamezen del podatkov, ki je razvrščen na podlagi »vej« (Berson, Smith, Thearling, 2000, str.110).

Izgradnja napovedovalnega modela s pomočjo dreves odločanja je enostavna, saj je rezultat enostaven za razumevanje, poleg tega pa je velika prednost te tehnike tudi zmožnost grafične ponazoritve. Običajen diagram drevesa odločanja izgleda običajno kot drevo, ki raste od zgoraj navzdol, saj običajno iz začetne (izhodiščne) točke z vsakim vozliščem postaja čedalje bolj košato oz. razvejano. Naš primer odločitvenega drevesa pa je obrnjen za 90 stopinj, kot lahko vidimo na sliki 1. Raste torej od leve proti desni.

Slika 1: Primer odločitvenega drevesa



Vir: Konič, 2003, str. 18.

Potrebno je opozoriti na pojav prevelikega prilagajanja. Pri tem pojavu gre za pregloboko segmentiranje. S poglabljanjem segmentiranja lahko pridemo do pojava, ko je en segment tudi ena opazovana enota. Tovrstni segmenti so sicer popolnoma homogeni, vendar pa ne dajejo nobene dodane vrednosti h kvaliteti napovedanih rezultatov. Pred prevelikim prilagajanjem se lahko zavarujemo predvsem na dva načina, in sicer s t.i. tehniko »bonsai«, ko že v začetku določimo maksimalno število segmentov, ki jih želimo dobiti; druga tehnika pa se imenuje »obrezovanje«, ko v že zgrajenem drevesu odstranimo neuporabne veje (Konič, 2003, str. 19). Teoretično je skrajna meja rasti drevesa odločanja primer, ko je ena opazovana enota en segment, običajno pa je skrajna meja rasti situacija, ko so vse opazovane enote v segmentu dovolj homogene, da nadaljnje odločanje oz. razdeljevanje ni več smiselno.

2.2.5. NEVRONSKE MREŽE

Nevronske mreže (ang. neural networks) so na področju podatkovnega rudarjenja dokaj nove, saj se je razvoj začel proti koncu prejšnjega stoletja. Ideja pa je bila omogočiti, da bi računalniki posnemali delovanje človeških možganov. Ideja je bila naučiti računalnike razmišljati na način, kot to počnemo ljudje. V začetku so bile nevronske mreže domena umetne inteligence, danes pa je to tehniko mogoče uporabljati tudi v podatkovnem rudarjenju.

Tehnika nevronskih mrež se uporablja predvsem na področju regresije in na področju razvrščanja. Pri regresiji in tudi razvrščanju gre za zapletene postopke, tehnika nevronskih mrež pa se izkaže kot uspešna. Podobno kot drevesa odločanja je nevronske reže mogoče prikazati v grafičnem načinu, vendar je taka mreža težje predstavljiva in razumljiva, saj gre za kompleksnejše povezave. Danes je ta tehnika poleg dreves odločanja najpogosteje uporabljena, saj daje odlične rezultate na področju napovedovalnih modelov. V praksi je

znano, da se tehnika uporablja predvsem na področjih, kjer ni potrebno v podrobnosti poznati zgrajenega modela, saj je model zgrajen s pomočjo tehnike nevronskih mrež izredno težko razumljiv in s tega vidika neprimeren za poslovne uporabnike. Tehnika se uporablja pogosto, saj daje ob modeliranju izredno kompleksnih problemov z večjim številom spremenljivk dobre rezultate. Zavedati se je potrebno, da v današnjem poslovnem svetu skorajda ni več enostavnih problemov, saj se poslovno okolje spreminja in zapleta. Na podlagi tega dejstva pa je v prihodnosti mogoče pričakovati še dodaten razmah uporabnosti nevronskih mrež.

2.3. PODATKOVNO RUDARJENJE V PRIHODNOSTI

Že v preteklosti so se podjetja zavedala, da se je na podlagi preteklih izkušenj oz. podatkov mogoče bolje odločati o prihodnjem poslovanju. Zavedala so se možnosti sprejemanja boljših odločitev s pomočjo analiz preteklih podatkov. Na podlagi dejstev iz preteklosti so »izkopavali« znanje, ki so ga lahko uporabila v prihodnosti. Rudarjenje je bilo sicer »ročno« ob pomoči statistikov, ki so analizirali določene pojave iz preteklosti.

Danes skorajda ni podjetja, ki ne bi vsaj za določen del svojega poslovanja uporabljalo določenih elementov podatkovnega rudarjenja, saj že človeški možgani (torej tudi direktorjev) delujejo v smeri odkrivanja vzorcev, ki bi jih bilo mogoče uporabiti v dobro podjetja. Edina velika težava, ki jo je mogoče ugotoviti v praksi, je nedojemljivost za hiter razvoj tehnologije tudi na področju podatkovnega rudarjenja. V veliki večini slovenskih, evropskih in tudi svetovnih podjetij še vedno uporabljajo klasične metode s pomočjo statistikov, ki so se v preteklosti izkazale za dovolj dobre, danes v kompleksnejšem poslovnem svetu pa ne dajejo več dovolj dobrih rezultatov. Težava je predvsem v hitrosti izdelave modelov in s tem v hitrosti prilagajanja spreminjajočemu se poslovnemu svetu.

V prihodnosti se bodo podjetja morala še posebej potruditi obdržati stik s svojimi konkurenti. Zagotovo bo pomembno vlogo v tem razmerju imelo tudi podatkovno rudarjenje, saj se obveščenost podjetij o novih tehnologijah povečuje. Zanimivo je stanje na slovenskem trgu, kjer so sodobna orodja za podatkovno rudarjenje dobro znana in tudi dobro sprejeta. Manjka pa še tista usodna odločitev, ko bodo podjetja prešla na avtomatizirano podatkovno rudarjenje. Slovenska podjetja se zaenkrat še premalo zavedajo konkurenčnih prednosti, ki jih tovrstna orodja ponujajo. V prihodnosti pa bo prehod na nove tehnologije zagotovo neizbežen, saj ga narekuje globalna konkurenca, ki se ji bo potrebno prilagoditi.

3. RUDARJENJE PO PODATKIH IN ORODJE KXEN

3.1. ORODJA ZA PODATKOVNO RUDARJENJE

Podatkovno rudarjenje se ni pojavilo pred nekaj leti, ampak že pred nekaj desetletji, zato je logično, da je danes, ko živimo v dobi informacijske tehnologije, za podatkovno rudarjenje na voljo precej orodij, ki pa se razlikujejo po nekaterih ključnih dejavnikih. Orodja se razlikujejo glede na tehniko podatkovnega rudarjenja, ki jo uporabljajo (odločitvena drevesa, nevronske mreže, inducirana pravila,…), razlikujejo se glede izhodišč, na podlagi katerih so se razvila (statistične potrebe, dodatek programskim paketom,…), razlikujejo se tudi s finančnega vidika (izredno draga orodja, orodja s povprečno ceno, orodja, ki so na voljo brezplačno,…), orodja se razlikujejo tudi z vidika usmeritve (rezultatsko usmerjena orodja, metodološko usmerjena orodja,…). Kriterijev razlikovanja orodij je še veliko. Sam bom prikazal razdelitev glede na tehniko, ki jo posamezna orodja uporabljajo, prav tako pa bom skušal prikazati tudi razdelitev nekaterih bolj znanih orodij z vidika kvalitete in cene. Razdelitev bo služila predvsem kasnejšemu lažjemu primerjanju KXEN analitičnega ogrodja z ostalimi orodji.

Prva razdelitev orodij za podatkovno rudarjenje ponuja možnost uvrščanja posameznega orodja prisotnega na trgu v skupine, ki uporabljajo podobne tehnike podatkovnega rudarjenja. Na trgu ni prisotnih veliko orodij, ki bi bila brezplačna, kar prikazuje, da je podatkovno rudarjenje zapleten postopek. V nasprotnem primeru bi se na trgu zagotovo pojavljalo ogromno različic programov, ki bi omogočala kvalitetne »izkope« znanja. V tabeli 1 je naštetih ogromno programskih orodij, za katere pa velika večina ljudi ni slišala še nikoli. Razlog za to je predvsem na strani prevlade in moči proizvajalcev nekaterih programskih orodij. Prevlada je seveda povezana tudi s kvaliteto in uporabnostjo. Celotna razdelitev programskih orodij je predstavljena v tabeli 1, na str. 11. Opozoriti je potrebno, da nekatera orodja podpirajo več tehnik podatkovnega rudarjenja.

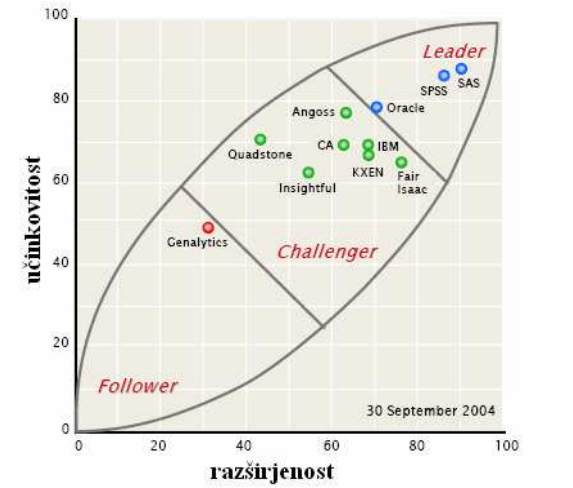
Tabela 1: Razdelitev orodij, ki se pojavljajo na trgu, glede na tehniko in plačljivost

tehnika	plačljiva orodja	brezplačna orodja
tehnika odločitvenih dreves	AC2, Alice d'Isoft 6.0, Business Miner, CART 4.0, Cognos Scenario, Decisionhouse, KnowledgeSeeker, PolyAnalyst, SPSS AnswerTree, XpertRule Miner	LMDT, IND, EC4.5, C4.5, OC1, PLUS
verjetnostni mrežni pristop	Analytica, AT-Sigma Data Chopper, Bayesware Discoverer 1.0, Data Digest Business Navigator 5, DXpress, HUGIN, KnowledgeMiner, Netica, PrecisionTree	BAYDA 1.0, Bayesian belief network software, FDEP, GeNIe, JavaBayes, MSBN Microsoft Belief Network Tools, Pulcinella, SPI, RoC
tehnika pravil	AIRA, Datamite, DataDowser, PolyAnalyst, SuperQuery, WizWhy, XpertRule Miner	CBA, CLaudien, CN2, DBPredictor, KINOsuite-PR, RIPPER
tehnika nevronskih mrež	Neural Network FAQ, BioComp iModel, COGNOS 4Thought, BrainMaker, DB Prophet, KINOsuite, MATLAB Neural Net Toolbox, Neural Innovation Proforma, NeuralWare, NeuroSolutions, SPSS Neural Connection 2, STATISTICA Neural networks	Tiberius, PNL Neural Shareware list, SNNS
hibridne tehnike	Affinium Model Suite, SPSS Clementine, Oracle Darwin, KINOsuite PR, Knowledge Studio, MarketMiner, Polyanalyst, PredictionWorks, Previa Classpad, KXEN, Datalogic, K-DYS, Discipulus, Evolver, MARS, WINROSA	BSVM, LIBSVM, Kernel Machines, Grobian, Rough Enough, PEBLS, TiMBL 2.0 MLC++, JAM SIPINA-W, ROC Convex Hull Program

Vir: AAAI Spring Symposium on Information Refinement and Revision for Decision Making, 2004.

Z vidika poslovnih uporabnikov je zagotovo boljša razdelitev orodij po kriterijih razširjenosti in učinkovitosti. Na sliki 2 lahko vidimo razporeditev nekaterih najbolj znanih programskih orodij za podatkovno rudarjenje glede na omenjena kriterija. Potrebno je opozoriti na tri različne skupine orodij. To so vodilna orodja na trgu, orodja, ki jih lahko imenujemo »izzivalci«, in pa skupina, v katero spadajo orodja, ki zgolj sledijo ostalima dvema skupinama orodij. Med vodilna orodja se uvrščajo SAS, SPSS in Oracle. Glede na raziskavo, ki jo je opravilo podjetje MetaGroup, Inc (Metaspectrum 60.1 - Data Mining Tools, 2003, str. 6), so omenjena tri orodja najboljša tako z vidika razširjenosti kot tudi učinkovitosti, tesno pa jim sledijo tudi Angoss, IBM, KXEN, CA,… Na tem mestu je potrebno opozoriti tudi na cenovne razrede omenjenih orodij, saj se cenovni razredi tesno prilegajo omenjenim trem skupinam orodij.

Slika 2: Razdelitev orodij za podatkovno rudarjenje glede na kriterija učinkovitosti in razširjenosti.



Vir: Metaspectrum 60.1 - Data Mining Tools, 2003, str. 5.

Potrebno je poudariti, da je v Sloveniji prisotnih kar nekaj zgoraj omenjenih orodij, zato ima druga razdelitev orodij verjetno tudi večjo uporabno vrednost. V Sloveniji je na področju podatkovnega rudarjenja tako mogoče najti predvsem SAS, SPSS, Oracle, Angoss in KXEN. Prisotna so še nekatera druga orodja, ki pa so v veliki meri prisotna, ker so priložena kot dodatek večjim informacijskim sistemom oz. programom za upravljanje s podatkovnimi bazami (Interno gradivo podjetja UT informacijski sistemi d.o.o.).

Opozoril bi rad tudi na dejstvo, da je težje primerjati orodja s področja izzivalcev z orodji s področja vodij. Težava je namreč podobna primerjavi avtomobilov iz različnih kakovostnih razredov. Težko je primerjati npr. Golfa z Mercedesom serije S. Golf je namreč izredno dober avto, še vedno pa je veliko slabši od Mercedesa serije S. Podobno je tudi z DM orodji, saj je težje primerjati orodje KXEN z orodjem SAS, zato velikokrat primerjava temelji na primerjavi orodja KXEN z bližnjimi konkurenti, saj je potrebno za primerjavo upoštevati več dejavnikov, kot npr. cena, velikost, razširjenost, lahkotnost uporabe, velikost podjetij, zgodovina podjetij,…

Zanimiva je še razdelitev orodij na metodološko orientirana orodja in pa orodja, orientirana na rezultat. Tovrstno razdelitev orodij so izdelali pri podjetju KXEN za kategoriziranje svojih največjih konkurentov. Orodja, ki spadajo v skupino metodološko orientiranih DM orodij, so: samostojna orodja (SPSS Clementine, SAS E.M, Quadstone, Think Analytics, Unica, Insightful) in dodatki podatkovnim bazam (Oracle DataMining, TeradataWarehouse Miner), med orodja, orientirana na rezultat, pa štejejo svoje orodje, ki se s tem tudi bistveno razlikuje od konkurenčnih orodij.

3.2. PREDSTAVITEV ORODJA KXEN

KXEN analitično ogrodje (Knowledge Extraction Engines) je močan nabor analitičnih komponent, ki omogočajo vsem, profesionalnim uporabnikom tehnologije za rudarjenje po podatkih in poslovnim uporabnikom, pretvarjanje podatkov v znanje v izredno kratkem času. Patentirana, napovedovalno in opisovalno modeliranje, sta zasnovana na podlagi Teorije statističnega učenja avtorja Vladimirja Vapnika[5]. Obe analizi (napovedovalno in opisovalno modeliranje) sta lahko uporabljeni za točkovanje (ang. scoring), klasifikacijo, segmentiranje in raziskovanje poslovnih informacij, kot npr. prispevek posamezne spremenljivke k proučevanemu problemu.

V preteklosti je bil proces priprave podatkov, kodiranja in modeliranja izredno dolgotrajen in je zahteval ogromno količino znanja. Podjetju KXEN pa je z integracijo najsodobnejših tehnologij in matematičnih spoznanj uspelo avtomatizirati pripravo podatkov, kodiranje in modeliranje. Prav tako KXEN omogoča dostop do veliko različnih virov podatkov. Omogoča, da se uporabniki preko procesa odločanja s pomočjo orodja lahko koncentrirajo na dodajanje vrednosti podatkom skozi zahtevne analize, kar pa omogoča izboljšane poslovne odločitve.

3.2.1. KXEN NA SPLOŠNO

KXEN analitično ogrodje je programski paket, ki skuša podatkovno rudarjenje približati poslovnim uporabnikom. S svojo enostavnostjo uporabe se zagotovo močno razlikuje od klasičnih metod[6] izdelovanja modelov za podatkovno rudarjenje. Potrebno je opozoriti predvsem na (z vidika povprečnega poslovnega uporabnika) dokaj težavno uporabo orodij SPSS in SAS. Uporabniki omenjenih orodij tudi v Sloveniji iščejo orodja, ki bi zagotavljala enako dobre rezultate, omogočala pa bi lažjo uporabo[7]. Pri klasičnih pristopih je bil postopek izdelave napovedovalnega modela precej dolgotrajen, težaven in tudi drag proces. Danes pa rudarjenje postaja konkurenčna prednost, ki je dosegljiva vsem.

---

[5] Vladimir Vapnik je oče Teorije statističnega učenja, ki je osnova za vse komponente KXEN-a. S pomočjo VC koncepta (Vapnik-Chervonenkis) dimenzij je Vladimir Vapnik predstavil novo paradigmo za podatkovno modeliranje. Namesto potrebe po predpostavljanju (spreminjanju in prilagajanju) osnovnih podatkov, ki vplivajo na določen poslovni problem, Vapnikova teorija pušča podatke take kot so, jih ne spreminja in omogoča modeliranje podatkovnih modelov brez omejitev števila spremenljivk, ki so vključene v model. Uporabljeni algoritmi izdelajo robusten in visoko kvaliteten model v neprimerno krajšem času, kot je potreben ob uporabi klasičnih metod in orodij.
Vladimir Vapnik je pokazal, da je veliko tehnik modeliranja, kot npr. kontrolirane nevronske mreže, del njegove teorije. Vapnik je prvi, ki je dokazal, »zakaj« nevronske mreže zagotavljajo izredno kvalitetne napovedi.

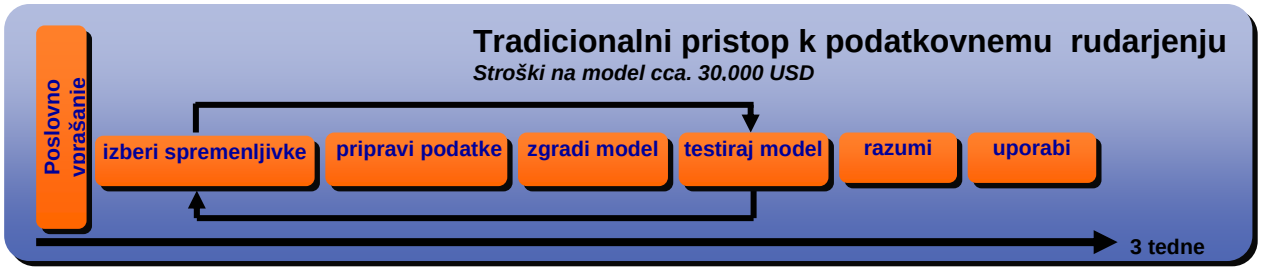
[6] Zopet so mišljene metode oz. postopki, ko v podjetjih s pomočjo statistikov skušajo ugotoviti določene zakonitosti, ki vladajo na področju poslovanja. Proces klasične metode ustreza procesu, ki je prikazan na sliki 3. Kot alternativna možnost klasični metodi podatkovnega rudarjenja se pojavljajo programska orodja, ki proces podatkovnega rudarjenja avtomatizirajo in poenostavijo, hkrati pa nudijo boljše rezultate.

[7] Argument temelji na osebnih izkušnjah, ko sem KXEN analitično ogrodje predstavil zaposlenim v enem večjih slovenskih podjetij, ki trenutno uporabljajo SAS. Uporabniki omenjenega orodja so se pritoževali predvsem nad težavnostjo orodja. Veliko vlogo naj bi pri delovanju obstoječega sistema igral oddelek informatike, kar pa je z določenih vidikov (preobremenjenost, nesamostojnost uporabnikov,…) včasih velika slabost.

Izdelava napovedovalnih modelov po tradicionalnih postopkih

Vedno je za izgradnjo napovedovalnega modela moralo obstajati poslovno vprašanje, ki so ga skušali rešiti. Statistiki so morali v naslednji fazi izbrati spremenljivke, ki so se jim zdele pomembne za odgovor na zastavljeno vprašanje. Naslednji korak je bil pripravljanje podatkov, na podlagi katerih se je zgradil model. Na koncu pa je bilo potrebno model še testirati. V primeru, da model ni dajal zadovoljivih odgovorov na zastavljeno vprašanje, je bilo potrebno postopek od priprave spremenljivk dalje ponavljati. To rotiranje omenjenih štirih faz izgradnje modela je bilo izredno dolgotrajno, s tem pa je izgradnja modela predstavljala ogromne stroške. Potrebno je poudariti, da tudi po testiranju modela nikoli niso mogli vedeti, ali je model uporaben, oz. ali je kvaliteten (med pripravo modela so se lahko vzorci spremenili, vprašanje mogoče ni bilo več pomembno,…). Povrhu vsega pa po končanem postopku izgradnje modela poslovno vprašanje morda ni bilo več relevantno in so bili vsi napori in stroški zaman. Stroški izdelave povprečno zahtevnega modela je v podjetju KXEN ocenjeno na cca. 30.000 ameriških dolarjev, v povprečju pa so bili za izdelavo modela potrebni trije tedni (KXEN, 2003, str. 6). V primeru, da je model na testu ponudil zadovoljive rezultate, je bilo potrebno model najprej proučiti in ga razumeti ter ga nato začeti uporabljati. Potrebno je poudariti, da veliko zgrajenih modelov ni nikoli bilo uporabljenih. To pa je podatkovno rudarjenje delalo neprivlačno med vodilnimi v podjetjih. Shematični prikaz postopka izgradnje napovedovalnega modela si lahko ogledamo na sliki 3.

Slika 3: Shematični prikaz klasičnega postopka izdelave napovedovalnega modela s pomočjo statistikov ekspertov[8]



Vir: 1- short presentation, 2003, str. 6.

Izdelava napovedovalnih modelov z orodjem KXEN

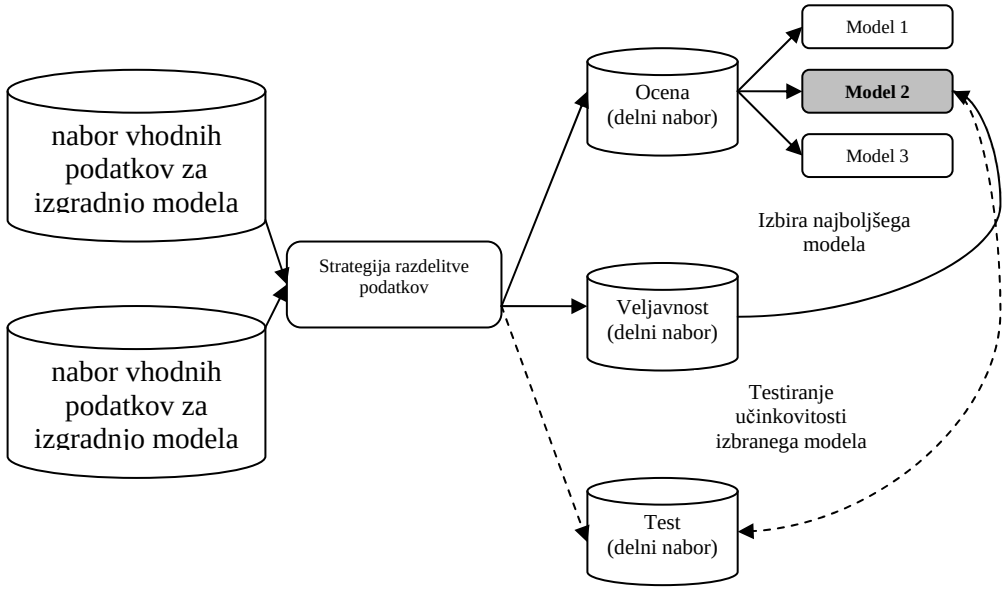
Izdelava napovedovalnega modela z orodjem KXEN je veliko enostavnejša, ponuja pa tudi veliko boljše in hitrejše rezultate. Kvaliteta zgrajenega modela ni več odvisna od kakovosti statistika, ki analizira spremenljivke, ampak je konstantna, saj je v ozadju posebna matematična logika.

[8] Zaradi kompleksnosti problemov, ki so jih v podjetjih želeli reševati, je bilo potrebno ekspertno statistično znanje. V praksi je veljalo, da so boljši statistiki izdelovali boljše in natančnejše modele. Za podjetja je to pomenilo, da več kot so namenila za to področje, boljše statistike so imeli in posledično boljše modele.

Izgradnjo modela je shematično mogoče prikazati z zgolj štirimi koraki, ki jih lahko vidimo tudi na sliki 5. Potrebno se je zavedati, da je shematični prikaz na tej sliki precej poenostavljen, vendar zgolj z namenom poudarka racionalnejše izrabe časa. Še vedno je za izgradnjo potrebno imeti poslovno vprašanje, ki bi ga radi rešili s pomočjo podatkovnega rudarjenja. Naslednja faza je izgradnja modela, v kateri so združene faze priprave spremenljivk, podatkov, izgradnje modela in testiranja. Ena faza iz tradicionalnih štirih nastane predvsem zaradi dejstva, da je sedaj celoten postopek izgradnje modela avtomatiziran. V fazi izgradnje modela je še vedno v ozadju skupek štirih faz, ki smo jih prikazali na sliki 3. Še vedno je potrebno izbrati spremenljivke, ki jih bomo uporabili, še vedno je potrebno pripraviti podatke, potrebno je zgraditi model in na koncu model še testirati. Omenjene štiri faze sem združil v eno zaradi enostavnosti, ki jo ponuja orodje KXEN. Zavedati se je potrebno, da brez shranjenih podatkov ne moremo zgraditi napovedovalnega modela, saj potrebujemo podatke iz preteklosti. Izbor spremenljivk je v orodju KXEN enostaven, saj dodatna spremenljivka, ki jo želimo vključiti v model, ne predstavlja večje obremenitve. Podaljša se čas izdelave modela, kvaliteta pa se ne poslabša. Če spremenljivka, ki jo dodatno vključimo v model, ne prinaša dodatne vrednosti h kvaliteti rezultata, je edina slaba stvar nepotreben čas, ki je porabljen ob generiranju modela. Mogoče je dobro opozoriti, da je zaradi te dodatne obremenitve potem potrebno izključiti spremenljivko, ki ne prinaša dodatne vrednosti, saj si je potrebno predstavljati situacijo, ko se model na novo generira na ogromni bazi podatkov in je optimalna izraba časa namenjenega za izgradnjo modela ključnega pomena (predstavljajmo si izgradnjo modela na bazi podjetja Telekom oz. Mobitel, kjer vsakodnevno pridobijo ogromno količino novih podatkov). Ali nam določena spremenljivka prinaša h kvaliteti ali ne, vidimo iz statističnih in grafičnih prikazov, ki nam jih omogoča KXEN. Prispevek spremenljivke seveda lahko ugotovimo šele potem, ko je model zgrajen, vendar je bolje na novo zgraditi model brez spremenljivk, ki ne prinašajo dodane vrednosti rezultatu, kot pa vsakodnevno dodatno obremenjevati sistem za upravljanje s podatkovno bazo.

Priprava podatkov je v orodju KXEN precej poenostavljena, saj ni potrebno posebej graditi nove tabele, ki bi vsebovala samo potrebne podatke oz. samo potrebne spremenljivke. Orodje je samo sposobno prilagoditi vhodne podatke v obliko, ki omogoča kvalitetno in hitro modeliranje. Potrebno je vedeti, da KXEN avtomatsko vhodne podatke razdeli na tri dele, in sicer na podatke namenjene oblikovanju modela (cenitvi, pripravi), na podatke, namenjene izboru veljavnega modela (najboljšega modela) in na podatke, namenjene testiranju modela. Pri tem razdeljevanju podatkov KXEN pozna osem različnih strategij za razdelitev podatkov. Vsaka pa je primerna za določeno obliko poslovnega vprašanja. Razlike pa je mogoče opaziti pri finalnih »brušenjih« modela. Shematično razdelitev vhodnih podatkov oz. proces priprave podatkov, izgradnje modela in testiranja modela vidimo na sliki 4. Potrebno je poudariti, da tukaj nastane ključna razlika od klasičnega podatkovnega rudarjenja (delo statistikov), saj se z avtomatizacijo teh faz v podatkovnem rudarjenju pridobi ogromno dragocenega časa, prav tako pa pridobimo tudi neposredno primerjavo med modeli, ki jih KXEN zavrne, in modelom, ki je izbran kot najboljši model.

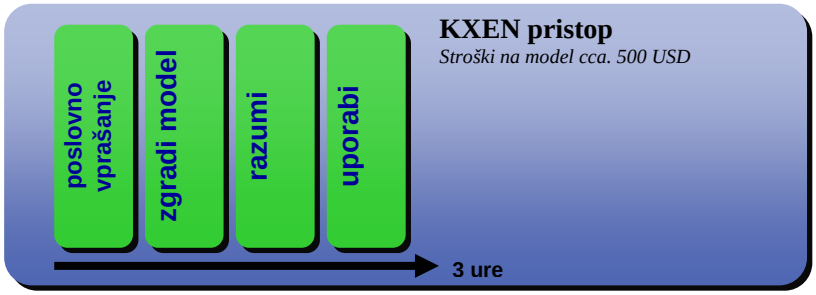
Slika 4: Shematični prikaz postopka izgradnje modela



Vir: KXEN Analytic Framework User Guide 3.1.0., 2004, str. 54.

Naslednja faza v KXEN pristopu je razumevanje modela, ki je močno olajšano, saj imamo v orodju KXEN na razpolago kar nekaj možnosti, ki omogočajo enostavnejše razumevanje zgrajenega modela (različni grafi, izračuni, vrtanje v globino,…). Zadnja faza pa je uporaba modela. Za enako kvaliteten oz. še boljši model, za katerega bi po tradicionalnem postopku potrebovali tri tedne, z orodjem KXEN potrebujemo 3 ure[9]. Takoj nam je lahko jasno, da se s tem znižajo povprečni stroški posameznega modela. Istočasno pa je velika prednost na strani hitrosti, saj je v poslovnem svetu potrebno čedalje hitreje reševati določene probleme, saj je poslovno okolje čedalje bolj zahtevno in hitro se spreminjajoče.

Slika 5: Shematični prikaz KXEN postopka izdelave napovedovalnega modela



Vir: 1- short presentation, 2003, str. 6.

[9] Ocena treh ur za izgradnjo povprečnega modela z orodjem KXEN se nanaša na raziskavo, ki so jo opravili pri podjetju KXEN. Primerjali so namreč čas, potreben za generiranje napovedovalnega modela v podjetjih, ki so se odločila za prehod na KXEN analitično ogrodje (banke in telekomunikacijska podjetja). Vse modele, ki so jih v podjetjih imeli narejene s pomočjo statistikov oz. drugih orodij, so ponovno zgradili z orodjem KXEN in potem izračunali povprečen čas izdelave modela. Izračunani čas je bil približno 3 ure, ta čas pa ne vključuje vgradnje zgrajenega modela v obstoječe sisteme v podjetjih, ampak zgolj čas, potreben za izgradnjo modela. Enako so izračunali tudi čas, ki so ga v povprečju v podjetjih porabili ob »ročni« izgradnji modelov (statistiki). Povprečen čas za izgradnjo modela se je gibal okrog treh tednov (Interno gradivo podjetja KXEN).

Poleg zgoraj omenjenih prednosti KXEN pristopa k napovedovalnim analizam v primerjavi s tradicionalnimi metodami pa KXEN omogoča še kar nekaj izredno pomembnih prednosti. V nadaljevanju bom skušal predstaviti zgolj najbolj pomembne in očitne.

Veliko prednost predstavlja sama programska zasnova KXEN analitičnega ogrodja. KXEN je namreč programiran v C++ programskem jeziku. Zaradi splošne uporabe tega programskega jezika rešitev deluje na vseh operacijskih sistemih oz. platformah. Pomembno pa je tudi dejstvo, da je programska koda maksimalno optimizirana, saj je verzija 3.1.0 velika zgolj 3 MB. Potrebno je tudi poudariti, da KXEN pretirano ne obremenjuje sistema za upravljanje z bazo podatkov, na katerem deluje, saj pri delovanju ni podvajanja podatkov. Večina ostalih orodij, ki obstajajo na trgu, za podatkovno rudarjenje za izgradnjo vsakega modela najprej prekopira podatke in šele nato zgradijo model. Pri KXEN-u tega ni, saj se modeli gradijo brez podvajanja. To pa predstavlja veliko manjšo obremenitev za bazo[10]. Pri sami prednosti z vidika programiranja je potrebno predstaviti tudi dejstvo, da je za izdelavo modelov v KXEN-u, brez kakršnihkoli težav podatke mogoče zajemati iz najrazličnejših virov. Podatke lahko tako zajemamo iz poljubne baze, podatki se lahko nahajajo v tekstovnih dokumentih, v Excelovi preglednici oz. v poljubnem formatu. Za KXEN to ne predstavlja večjih težav, saj je zajemanje podatkov iz različnih virov ena izmed velikih prednosti. Po drugi strani pa je potrebno poudariti, da KXEN omogoča izvoz zgrajenega modela v poljubne programske jezike, s tem pa je olajšana vgradnja izdelanega modela v že obstoječe sisteme v podjetjih.

Za komunikacijo z uporabnikom skrbi vmesnik, narejen v programskem jeziku Java. Vmesnik je mogoče z enostavnimi postopki prilagoditi potrebam posameznega podjetja. Spreminjamo lahko barve in oblike vmesnika, dodajamo logotip podjetja, skratka vse, kar je potrebno, da KXEN dobi podobo ostalih sistemov v podjetju. Vse te olepšave vmesnika so izredno enostavne, saj je v dokumentaciji lepo razloženo, na kakšen način je mogoče spreminjati določene parametre vmesnika. Omenjena dokumentacija je še ena izmed prednosti, saj je narejena na način, ki omogoča enostavno in hitro iskanje odgovorov na vprašanja, ki se zastavljajo. Vsebina dokumentacije je razdeljena na različna poglavja, ki omogočajo hitro brskanje.

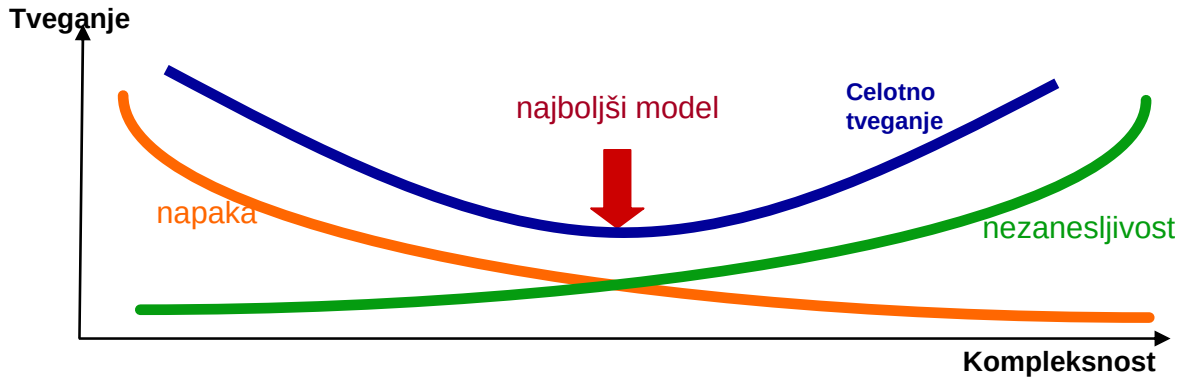
Največjo prednost KXEN analitičnega ogrodja zagotovo predstavlja uporabljena matematična logika, ki je v ozadju. V programskih orodjih, ki se pojavljajo na trgu, je namreč uporabljena matematična logika, ki je stara že kar nekaj desetletij in ne ponuja zadovoljivih rezultatov. Nekatera orodja dajejo celo slabše rezultate, kot bi jih dobili z naključnim izbiranjem.[11] Pri

[10] Potrebno se je zavedati, da so sistemi za upravljanje z bazami v večjih podjetjih izredno obremenjeni in dodatno obremenjevanje npr. s kopiranjem podatkov enostavno ne pride v poštev. Predstavljati si je potrebno npr. sistem za upravljanje z bazo podatkov v podjetjih, kot je npr. Telekom Slovenije d.d.. Vsakodnevno se v podatkovni bazi shranjujejo enormne količine podatkov.

[11] Na tekmovanju orodij za podatkovno rudarjenje (The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000) dosegajo nekatera orodja slabše rezultate, kot bi jih dobili z naključnim izbiranjem. Ena izmed razlag je tudi neustrezna matematična logika, na podlagi katere deluje orodje. Potrebno je vedeti, da ta orodja ne dosegajo slabih rezultatov na vseh področjih, saj je rezultat odvisen tudi od postavljenega poslovnega vprašanja.

KXEN-u pa so uporabljena najsodobnejša matematična spoznanja, ki omogočajo boljše rezultate (Teorija statističnega učenja avtorja Vladimirja Vapnika). Celotno orodje temelji na principu SRM (Structured Risk Minimization). Z orodjem KXEN skušamo dobiti modele z najmanjšim tveganjem ob ustrezni kompleksnosti modela (ob ustreznem številu uporabljenih spremenljivk). S povečevanjem kompleksnosti modela narašča verjetnost, da je model nezanesljiv, saj je težje obvladljiv in težje nadzorovan. Z večanjem kompleksnosti pa se možnost pojavljanja napak tudi zmanjšuje. Najboljši model je torej tisti, ki pri določeni kompleksnosti omogoča najmanjše skupno tveganje, to je v točki, ki je na sliki 6 označena s puščico.

Slika 6: Najboljši model



Vir: 1- short presentation, 2003, str. 7.

Izrednega pomena so glavni kazalniki poslovanja (KPI - Key Performance Indicators), ki omogočajo neposreden nadzor kvalitete in zanesljivosti zgrajenih modelov. Takoj vemo, ali lahko nek model uporabimo ali ne. V klasičnem pristopu je bilo potrebno model testirati in šele takrat smo izvedeli, ali je model uporaben ali ne. Pri KXEN-u je potrebno omeniti predvsem dva kazalca. To sta kazalca KI (KXEN Information Indicator) in KR (KXEN Robustness Indicator). Kazalec KI je kazalec kvalitete in nam pove, ali naš model ponuja dobre ali slabe rezultate. Če je kazalec pozitiven, to pomeni, da z modelom dobimo boljše rezultate, kot bi jih dobili z naključnim izbiranjem. Potrebno je poudariti, da je model uporaben tudi, če je izredno nizek, npr. 0,1, saj lahko uporaba takega modela pripomore k ogromnim prihrankom oz. k povečanju konkurenčnih prednosti. Za razumevanje ustreznosti modela s pomočjo tega kazalca je potrebno poznati tudi poslovno vprašanje. Če je KI enak 1, to pomeni idealen model, če pa je enak 0, pa to pomeni, da je model enak, kot če bi naključno izbirali. Teoretično je sicer mogoče tudi, da bi bil ta kazalec negativen, vendar ne za modele narejene z orodjem KXEN. Drugi pomemben kazalec pa je kazalec zanesljivosti (KR), ki nam pove stopnjo robustnosti modela. Pove nam, kako učinkovita bo uporaba modela na novih podatkih. Pove nam torej, kako zanesljive bodo napovedi za nov nabor podatkov. Za zanesljive modele se štejejo tisti modeli, ki dosegajo KR višji od 0,95. V teh primerih gre za odklone, manjše od 5 odstotkov. Drugače povedano, napovedi z modelom, ki ima KR 0,95, bodo v 95 odstotkih pravilne. Kazalec KR je mogoče izboljšati s povečanjem količine

podatkov, na podlagi katerih se model gradi, nikakor pa ne z odstranjevanjem določenih spremenljivk iz modela.

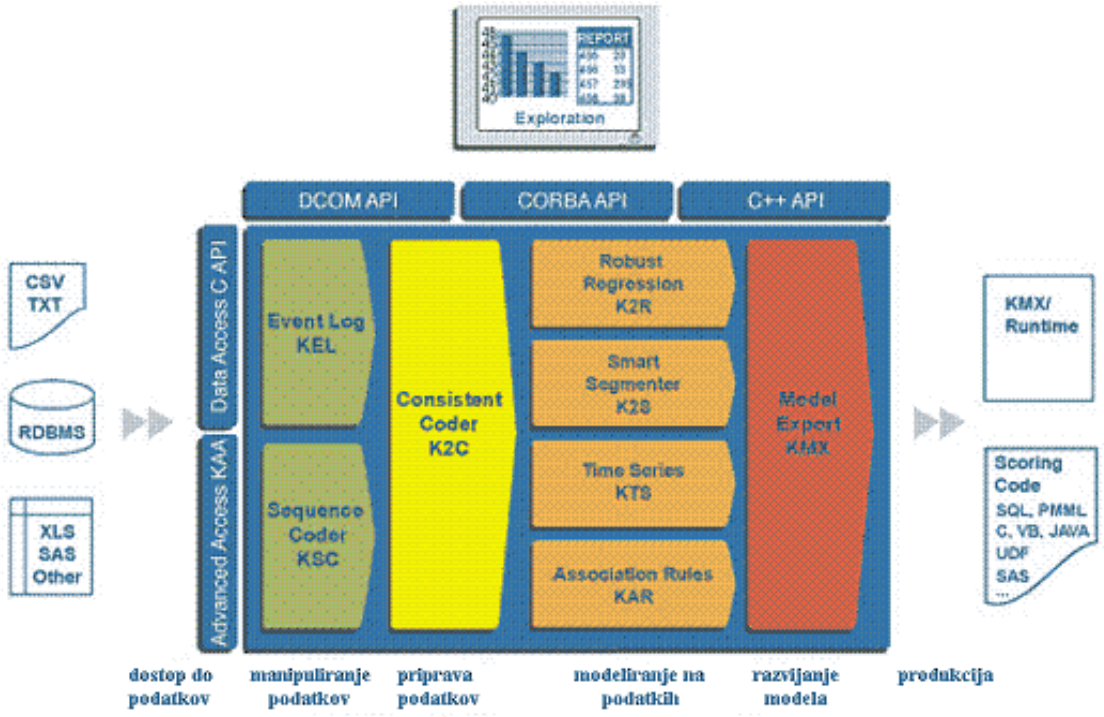
Podobne kazalce ponujajo seveda tudi druga orodja, vendar predvsem orodja, kot so npr. SAS, SPSS, Oracle. Poudariti je potrebno, da z vidika teh kazalcev KXEN ni revolucionaren, je pa eden redkih orodij, ki niso raziskovalna orodja, ki ponuja tovrstne kazalce. Bližnji konkurenti kot npr. Angoss, Quadstone, CA, tovrstnih kazalcev ne poznajo.

Nenazadnje je potrebno pri prednostih KXEN analitičnega ogrodja omeniti tudi izredno kakovostno podporo, ki jo nudijo tako neposredno iz podjetja, kot tudi podporo, ki jo ponujajo zastopniki po celem svetu. Pozitivno dejstvo pa je tudi to, da ima KXEN pod svojim okriljem nekaj največjih strokovnjakov na svetu z različnih področij (Vladimir Vapnik, Leon Bottou, Olivier Chapelle, Lee Giles,…[12]).

3.2.2. POSAMEZNE KOMPONENTE PROGRAMSKEGA PAKETA KXEN

Platforma programskega paketa KXEN, verzija 3.1.1., je sestavljena iz osmih osnovnih komponent. Osnovni gradniki platforme so: Event log (KEL), Sequence coder (KSC), Consistence coder (K2C), Robust regression (K2R), Smart segmenter (K2S), Time series (K2S), Association rules (KAR) in Model export (KMX). Shematični prikaz strukture platforme je prikazan na sliki 7.

Slika 7: KXEN analitično ogrodje, verzija 3.1.1.



Vir: http://www.kxen.com/products, 2004.

[12] Celotno ekipo svetovnih strokovnjakov, ki sodelujejo pri razvoju KXEN-a, je mogoče najti na spletni strani z naslovom http://www.kxen.com/about/scientific_board.php.

Posamezne komponente KXEN analitičnega ogrodja oz. platforme so predstavljene v nadaljevanju.

3.2.2.1. KXEN Robust Regression (K2R)[13]

K2R – KXEN Robust regression komponenta je regresijski algoritem, ki omogoča izdelovanje modelov za napovedovanje določenih kategorij oziroma zveznih spremenljivk.

K2R predstavlja mape nabora opisnih atributov (vhodov v model) in ciljnih atributov (rezultatov modelov). V K2R komponenti so uporabljeni algoritmi, ki temeljijo na načelih Vladimirja Vapnika o strukturiranem minimiziranju tveganj[14]. Namesto iskanja najboljšega rezultata na znanih podatkih K2R zasleduje cilj najti najboljši kompromis med kvaliteto in robustnostjo. Modeli oz. rezultati so izraženi v obliki polinomov vstopnih spremenljivk in dejanskih vrednosti. Edino, kar mora uporabnik ob tem narediti, je, da programu pove, katere stopnje polinom naj ustvari. Višja ko je stopnja zahtevanega polinoma, boljši naj bi bili rezultati, posledično pa je potrebnega več časa za izdelavo modela. Da bi izboljšali hitrost modeliranja, K2R omogoča tudi izgradnjo modelov z večjim številom ciljnih spremenljivk. Kvaliteta dobljenih rezultatov z dodajanjem ciljnih spremenljiv ne upada, poveča se le čas, potreben za generiranje modela.

K2R omogoča grafično prikazovanje prispevka posamezne spremenljivke k modelu, ki pomaga izbrati najbolj pomembne spremenljivke, ki vplivajo na poslovno vprašanje. Istočasno pa omogoča tudi izogibanje spremenljivkam, ki na določen problem nimajo nikakršnega vpliva in zgolj upočasnjujejo procesni čas, potreben za generiranje modela.

Modeli omogočajo direktno uporabo v simulacijskem načinu za posamezen set vhodnih podatkov. Rezultate dobimo v realnem času, s tem pa lahko hitro oblikujemo poslovno odločitev, ki je primerna za določen primer.

Velika prednost K2R je tudi sproščanje delovnega časa za eksperte rudarjenja po podatkih, s tem pa omogoča preusmeritev znanja teh ekspertov na področja, kjer je njihovo znanje resnično potrebno. Predvsem se sprosti veliko časa, ki je bil prej namenjen pripravi podatkov in izgradnji modela. K2R izdela model v nekaj minutah (za izgradnjo modela je na povprečnemu prenosnemu računalniku potrebnih le slabih 10 sekund, da izgradi model z dvajsetimi spremenljivkami in nekaj več kot 50.000-imi zapisi).

Ker so KXEN modeli izredno kvalitetni in robustni, omogočajo spreminjanje nabora podatkov in števila spremenljivk, s tem početjem pa lahko določen problem raziščemo bolj podrobno in bolj kvalitetno.

[13] Vir: KXEN Product Overview, 2004, str. 2.
[14] V originalu Structured Risk Minimization.

K2R je samostojen algoritem, ki ne zahteva dodatnega nastavljanja, omogoča pa konstantno zelo dobre rezultate. Komponenta predstavlja univerzalno rešitev za probleme klasificiranja in regresijske probleme, hkrati pa omogoča vgradnjo izdelanih modelov v različne aplikacije, ki se jih v podjetju že uporablja. Za vgradnjo izdelanega modela v obstoječe aplikacije ni potrebnega veliko dodatnega dela, zaradi tega pa se sprosti kar nekaj časa programerjev in informatikov, ki v pridobljenem času lahko rešujejo pomembnejše probleme.

Primer uporabe: Klasifikacija – uporaba določene trženjske akcije

Potrebno je zbrati podatke o rezultatih trženjske akcije. V ta namen lahko pošljemo npr. 5.000 elektronskih sporočil strankam in jim ponudimo nov izdelek. Zberemo podatke o odzivih testne populacije in na podlagi teh zgradimo osnovni model. Ko je model zgrajen, preverimo kazalca kvalitete in robustnosti in v primeru, da sta oba na dovolj visoki ravni (če sta kazalca preslaba, je potrebno testno populacijo povečati), lahko model uporabimo na vseh podatkih o naših strankah, ki jih imamo. Za vsako stranko dobimo številsko oceno o odzivu na nov izdelek. Marketinški oddelek se potem na podlagi podatkov odloči, koliko oz. katerim izmed strank bomo ponudili nov izdelek. Model seveda lahko v vsakem trenutku nadgrajujemo z novo pridobljenimi podatki (npr. po tednu dni uporabe določene trženjske akcije lahko model ponovno zgradimo na večjem številu zapisov), s tem pa zagotovo še dodatno izboljšamo kvaliteto modela.

Z modelom smo privarčevali tiste stroške, ki bi nastali, če bi marketinško kampanjo izvajali na vseh strankah, ki jih imamo v bazi podatkov. V našem primeru smo usmerili marketinške dejavnosti zgolj na tiste osebe, ki so imele najvišje število točk oz. največjo verjetnost, da bodo nov izdelek pozitivno sprejele. Izvedli smo usmerjeno trženjsko akcijo.

3.2.2.2. KXEN Smart Segmenter (K2S)[15]

Komponenta K2S je namenjena izgradnji modelov, ki omogočajo grupiranje oz. segmentiranje. Omogoča pridobivanje opisnih podatkov o skupinah podatkov. Pridobimo lahko podatke o tem, zakaj spadajo določeni zapisi podatkov v določeno skupino. Najbolj običajno segmentiranje je oblikovanje segmentov strank. Komponenta omogoča poglobljeno raziskavo spremenljivk, ki določen zapis (osebo) uvrščajo v določeno skupino s podobnimi lastnostmi.

Segmentiranje je v podjetjih izredno pomembno, saj na podlagi različnih ciljnih skupin oblikujemo različne ukrepe, ki so za posamezno skupino najbolj ustrezni.

K2S gradi modele z uporabo preslikave med naborom opisnih atributov, ki predstavljajo vhod v model, in segmenti na drugi strani, ki predstavljajo rezultat modela. Za analizo vstopnih

[15] Vir: KXEN Product Overview, 2004, str. 3.

podatkov skrbi komponenta K2C, ki K2S-ju omogoča določanje »centrov«[16] za posamezne segmente. Oblikovanje segmentov poteka na podlagi osnovnih statističnih metodologij, kot npr. frekvenčna porazdelitev, povprečja, regresija, navzkrižna statistika za primerjavo segmentov,…

Uporabna vrednost komponente za segmentiranje je predvsem v lahkotnosti oblikovanja različnih segmentov, ki so pomembni za trženjske ukrepe. Vsak segment, pridobljen z orodjem KXEN, predstavlja homogeno skupino, ki se oblikuje glede na vstopne podatke. Najbolj pomembno je dejstvo, da vsak, ki kreira segmente s KXEN-om, dobi enake rezultate. Ob uporabi klasičnih »ročnih metod« segmentiranja je bilo namreč oblikovanje segmentov odvisno od posameznika, saj je bil velik del segmentiranja odvisen od različnih predstav oz. mišljenj posameznikov. Na podlagi določenega nabora podatkov sta lahko dva posameznika popolnoma različno oblikovala segmente, ki bi jih uporabili v marketingu. S KXEN-om tega ni. Vedno dobimo enake rezultate, ki so objektivni. Poleg tega pa veliko uporabno vrednost predstavlja tudi ocenjevanje, ki ga K2S ponuja. Za vsak segment je namreč mogoče pridobiti podatke o tem, do kolikšne mere ustrezajo osebki v posameznem segmentu pogojem, ki smo jih oblikovali na začetku oblikovanja segmentov. Velika prednost avtomatiziranega oblikovanja segmentov s K2S-jem pa je zagotovo tudi dejstvo, da lahko oblikujemo segmente na podlagi ogromne količine podatkov. Edina posledica, ki bo nastala z večjo količino podatkov, je ta, da bodo rezultati segmentiranja boljši. Z naraščanjem količine uporabljenih podatkov narašča namreč tudi kvaliteta segmentiranja.

Primer uporabe – Uporaba scenarijev za različne skupine ljudi, ki kličejo v klicni center

V podjetju zbiramo podatke o obstoječih strankah, ki jih imamo, in jih povežemo z določenimi zahtevami, ki jih imajo posamezne stranke. Na ta način pridobimo podatke, na podlagi katerih lahko generiramo segmente s pomočjo KXEN - K2S-a, ki opredeli skupine strank, ki imajo podobne značilnosti. Za vsako skupino strank raziščemo, kaj je najbolj primeren pristop, oz. kaj je najprimernejša oblika ponudbe za sklenitev posla (kreiramo scenarije). Segmente, pridobljene s KXEN-om, in oblikovane scenarije povežemo in že imamo idealno aplikacijo, ki nam bo ob vnosu podatkov o novi stranki takoj pokazala, kaj bo najverjetneje najučinkovitejši pristop za sklenitev posla, prav tako pa bomo takoj videli statistično verjetnost uspeha, če bomo uporabili določen scenarij.

3.2.2.3. KXEN Consistence Coder (K2C)[17]

Komponenta K2C je ključna za veliko prednosti, ki jih ponuja KXEN analitično ogrodje. K2C namreč omogoča avtomatično pripravo podatkov, ki se jih potem lahko uporablja v KXEN analitičnem ogrodju. V bistvu gre za orodje, ki omogoča pretvarjanje nominalnih,

[16] Mišljen je določen nabor povprečnih vrednosti posameznih spremenljivk, okrog katerih so zbrani posamezni primerki, ki spadajo v določen segment. Opozoriti je potrebno, da je lahko spremenljivk, na podlagi katerih oblikujemo segmente, poljubno mnogo, lahko tudi več tisoč!
[17] Vir: KXEN Product Overview, 2004, str. 5.

vrstnih in zveznih spremenljivk v obliko, ki omogoča obdelavo podatkov za kreiranje modelov. K2C omogoča avtomatično zaznavanje manjkajočih vrednosti in tudi vrednosti, ki niso ustrezne (napačni vnosi, neustrezen format, podatki, ki so zunaj dovoljenega območja,…). V klasičnih metodah so statistiki morali večino priprave podatkov opraviti ročno, kar pa jim je vzelo ogromno časa, saj je bilo potrebno podatke pregledati in jih popraviti. Vse to je sedaj možno zgolj z nekaj kliki v KXEN-u.

K2C oblikuje kodirne sheme na podlagi testnih podatkov, ki jih potrebujemo za oblikovanje modela, ki ponuja rešitev na poslovno vprašanje. Nominalne in opisne spremenljivke se v procesu obdelave podatkov s K2C preoblikuje v številske vrednosti, na podlagi katerih je lažje kreirati modele. Zvezne spremenljivke pa se bodisi normalizirajo ali pa se uporabi postopek transformacije po koščkih zvezne spremenljivke. S tem se ugotovi nelinearne povezave s podatki.

Komponenta je bistvenega pomena predvsem z vidika priprave podatkov za oblikovanje modelov, saj je to najbolj občutljiva faza, od katere je odvisna končna kvaliteta in zanesljivost modela. Poslovnim uporabnikom je pomembna predvsem hitrost, avtomatika in pa enostavnost, saj ni potrebna ročna obdelava. Velika uporabna lastnost je tudi vrtanje v globino (ang. drilling down), ki omogoča raziskavo o tem, katere spremenljivke in kako vplivajo (pozitivno, negativno, močno, rahlo,…) na končni rezultat. V celotnem procesu generiranja modelov za rudarjenje podatkov pa je bistvena prednost, v primerjavi s klasičnimi orodji, hitrost, saj so sedaj modeli generirani v precej krajšem času. Pomembno je, da se ne izgublja veliko časa s pripravo podatkov, ampak se raje ta čas porabi za razumevanje rešitev in pripravo ustreznih razlag in ukrepov.

3.2.2.4. KXEN Event Log (KEL)[18]

KEL je komponenta KXEN analitičnega ogrodja, ki omogoča manipulacijo s podatki. Osnovna naloga komponente je pripravljati oz. zajemati tiste podatke, ki se beležijo v obliki zapisov za vsak dogodek (druga možnost je, da se podatki zajemajo po sekvencah) v obliki, ki omogoča podatkovno rudarjenje. KEL združuje statične podatke z dinamičnimi informacijami, ki se nahajajo v zgodovinskih tabelah. Zgodovinske tabele za posamezne dogodke se generirajo avtomatično, KEL pa omogoča manipuliranje z njimi.

V veliko primerih je v praksi potrebno graditi napovedovalne modele na podlagi statičnih podatkov, ki so razpršeni po različnih tabelah. Primeri tovrstnih podatkov so npr. demografske spremenljivke posameznikov, spiski opreme, dnevniki izvedenih prodaj, … Da lahko na tovrstnih podatkih zgradimo kvalitetne napovedovalne modele, potrebujemo agregirane povezave s podatki, ki se lahko razlikujejo med posamezniki. KEL generira potrebne agregate na podlagi zahtev o periodah, ki jih poda uporabnik. Periode so lahko dan, teden, mesec,… Agregati so preračunani na podlagi dneva referenc, ki so lahko fiksirane oz.

[18] Vir: KXEN Product Overview, 2004, str. 4.

različne od primera do primera. Primer reference je npr. dan prvega nakupa stranke. Komponenta KEL je oblikovana na način, ki omogoča enostavno preprogramiranje, saj omogoča različne specifikacije agregatov. Kot agregate lahko določamo maksimume, minimume, vsote, števce,…

Velika prednost komponente KEL je, da ni potrebno končnemu uporabniku programirati, da bi zagotovil potrebno agregacijo podatkov in s tem pripravo podatkov, da postanejo uporabni za podatkovno rudarjenje. Zaradi velike hitrosti komponente KEL je mogoče preverjati različne oblike agregatov. Vedno lahko poskusimo, na kakšen način dobimo boljše rezultate. Veliko prednost ponuja KEL tudi na področju zajemanja zgodovinskih podatkov, saj v klasičnih metodah zaradi obsežnosti oz. kompleksnosti tovrstno početje skorajda ni bilo mogoče. Z vključevanjem zgodovinskih podatkov pa pridobimo na kakovosti zgrajenih modelov in posledično tudi na kakovosti rezultatov. Na področju informacijske tehnologije pa KEL omogoča enostavno oblikovanje agregatov, pri tem pa ne obremenjuje informacijske infrastrukture, saj je postopek priprave agregatov končan v nekaj minutah in ne v nekaj dneh, kot je bilo to v preteklosti običajno. Pomembno je, da za sprotno izdelavo agregatov ni potrebno spreminjanje osnovnih shem.

Primer uporabe

Za CRM (ang. Customer Relationship Management) oz. upravljanje odnosov s strankami je najpomembnejša informacija kako je stranka sodelovala s podjetjem, oz. kako je bila zadovoljna s storitvami oz. produkti. Tovrstne informacije so običajno shranjene v zgodovini nakupov oz. v dnevniku klicnega centra. V primerih, ko želimo napovedovati odhod strank k našemu konkurentu, je izrednega pomena, da v modele vključimo tudi zgodovinske informacije o tem, kako je podjetje sodelovalo s kupcem do trenutka, ko je odšel. Izredno pomemben je tudi trenutek odhoda. Omenjeni potrebi zahtevata agregiranje podatkov, povezanih s trenutkom prehoda stranke. Stranke prehajajo h konkurentu v različnih trenutkih, zato bi bilo agregiranje na fiksen datum najverjetneje nesmiselno. Za tovrstne primere je pomembno avtomatično agregiranje podatkov za število realiziranih poslov, število pritožb, vsot nakupov,… Ko s KEL pripravimo omenjene agregirane podatke, je s komponento K2R enostavno mogoče zgraditi model, ki omogoča napovedovanje prehodov strank. S tovrstno napovedjo pa lahko z določenimi dejanji preprečimo odhode strank (KXEN, 2004, str. 4).

3.2.2.5. KXEN Sequence Coder (KSC)[19]

KSC je ravno tako kot KEL komponenta za manipuliranje s podatki. Kot nasprotje komponenti KEL ta komponenta skrbi za združevanje statičnih podatkov z dinamičnimi, pri tem se gradi transakcijska tabela agregiranih podatkov, ki se nanašajo na posamezne dogodke. Pomembno je predvsem zajemanje podatkov o transakcijah, ki se navezujejo na različne dogodke.

[19] Vir: KXEN Product Overview, 2004, str. 5.

Podatki o obnašanju nam lahko veliko bolj pomagajo (veliko bolj kot demografski in statični podatki) pri pridobivanju kvalitetnih modelov. Izrednega pomena so lahko podatki o nakupovalnih navadah strank, o odzivih na določene novosti, spremembe, podatki o pritožbah, podatki o načinu plačevanja, o načinu pridobivanja sredstev, med te podatke lahko vključujemo tudi dnevnik obiskov na domači strani, kjer beležimo zaporedje odprtih strani,… Zaporedje določenih dejanj posamezne stranke nam lahko omogoči kvalitetno napoved, kakšen bo naslednji korak stranke. Z vidika manipuliranja s podatki je KSC izrednega pomena, saj omogoča manipuliranje s podatki, ki so bolj zapleteni.

KSC generira agregate določenih potekov določenega dogodka, osnovne podatke pa pridobiva iz transakcijske tabele. KSC zgenerira agregatne podatke o določenih fazah dogodka za vsako stranko posebej. Na podlagi tovrstnih agregatov je mogoče kasneje kreirati napovedovalne modele, ki omogočajo napovedi določenega dejanja v prihodnosti. Na ta način lahko v podjetju sprejmejo ustrezne ukrepe, da bi določeno dejanje preprečili oz. ne.

KSC komponenta omogoča različnim tipom uporabnikov precej prednosti. Ni potrebno programiranje za izvajanje prefinjenih agregatov, uporabniki z lahkoto vključijo podatke iz transakcijskih tabel, že osnovna prednost je ta, da sploh omogoča uporabo tovrstnih podatkov, velika prednost je tudi velika hitrost in izredno majhna občutljivost tudi na velike količine podatkov,… Izgradnja agregatov iz transakcijske tabele je enostavna in izredno hitra (nekaj minut).

Primer uporabe

Dober primer uporabe je npr. uporaba KSC-ja na spletnih straneh. Vzporednice uporabe lahko potegnemo z domačo stranjo www.amazon.com. Spletni vmesnik vsaki stranki ponuja določene artikle iz ponudbe. Ponujeni so artikli, za katere obstaja največja verjetnost, da jih bo obiskovalec naročil. Verjetno se zdi zanimivo, da namigi, ki jih ponuja ta spletna stran, niso naključni, ampak so usmerjeni za vsakega prijavljenega obiskovalca posebej. Podobno storitev omogoča komponenta KSC, saj le-ta omogoča izgradnjo modelov za napovedovanje na podlagi podatkov o »obnašanju« posamezne stranke. Enako deluje tudi amazon.com, saj je v ozadju model, ki na podlagi preteklih podatkov o naročilih oz. obnašanju strank napoveduje prihodne možne prodaje. Namesto naključne ponudbe se uporablja usmerjena ponudba, ki se izkaže za izredno uspešno.

3.2.2.6. KXEN Time Series (KTS)[20]

KTS komponenta je v KXEN analitičnem ogrodju namenjena odkrivanju pomembnih vzorcev in trendov, ki so se pojavljali v poslovanju določenega podjetja. Vzorce in trende komponenta išče na podlagi podatkov, ki jih ima podjetje v bazah podatkov. Komponenta je namenjena napovedovanju določene spremenljivke v prihodnosti, npr. napovedovanju prometa podjetja v

[20] Vir: KXEN Product Overview, 2004, str. 6.

prihodnosti (v naslednjih mesecih, v naslednjem četrtletjih, naslednjem letu,…). KTS omogoča odkrivanje tako periodičnih gibanj kot tudi sezonskih gibanj določene spremenljivke, ki so lahko še tako zelo skriti. Na podlagi periodičnih in sezonskih gibanj omogoča natančne in kvalitetne napovedi preučevane spremenljivke v prihodnosti.

Tudi delovanje komponente KTS temelji na sistemu strukturiranega minimiziranja tveganj (SRM). S komponento KTS je primerno napovedovati zlasti gibanje prodaje v prihodnosti, rast podjetja, rast dobička, rast prodaje,… Mogoče je napovedovati vse tiste spremenljivke, ki jih moramo običajno napovedovati v realnem življenju.

KTS omogoča napovedi na podlagi zgrajenega modela, ki se tesno prilega gibanju preučevane spremenljivke v preteklosti. Na podlagi preteklih podatkov KTS izlušči trend in tudi sezonska gibanja ter ostala periodična gibanja. Na podlagi vsega tega pa na koncu napove gibanje preučevane spremenljivke v prihodnosti.

Uporabnost KTS je predvsem v tem, da lahko podjetja na podlagi napovedi določenih spremenljivk ukrepajo tako, da rezultate izboljšajo, oz. negativne napovedi preprečijo.

3.2.2.7. KXEN Association Rules (KAR)[21]

Komponenta KAR je najnovejša komponenta v KXEN analitičnem ogrodju. Namenjena je generiranju pravil, ki omogočajo odkrivanje zakonitosti skupnega pojavljanja določenih dogodkov. KAR omogoča kreiranje pravil, ki na podlagi preteklih podatkov omogočajo nadzor nad določenimi poteki sklepanja poslov. Primer je npr. prodaja digitalnih kamer. Ponavadi se ob nakupu kamere kupci odločajo tudi za nakup baterij oz. polnilcev. Podani primer je trivialen. KAL pa omogoča kreiranje pravil tudi na bolj kompleksnih primerih, ko povezave niso tako lahko opazne.

Asociacijska pravila se lahko uporabljajo pri analizah uporabe domačih strani, odkrivanju vzorcev telefonskih klicev, odkrivanju določenih vzorcev obnašanja na različnih področjih, pri analizah obnašanja, ki lahko vodi do prevar,… KAR omogoča podjetjem, da optimizirajo svojo ponudbo, tako da oblikujejo promocije z boljšimi rezultati in da odkrivajo tveganja na različnih področjih.
KAR analizira podatke, ki se nanašajo na posamezno stranko (npr. z vidika nakupov - nakupovalna košarica posameznika (kaj kupi)) oz. na posamezne dogodke. Komponenta skuša odkriti potencialno zanimive zakonitosti, ki obstajajo med analiziranimi podatki. KAR omogoča odkrivanje velikega števila pravil, saj omogoča iskanje pravil na izredno širokem področju (trgovine, klicni centri, domače strani, banke,…).

[21] Vir: http://www.kxen.com/products/components/kar.php, 2004.

Prednost uporabe KAL komponente se kaže v možnosti sprejemanja boljših odločitev na področju promocijskih ponudb, širine ponujenih izdelkov,… Poslovnemu uporabniku omogoča odkrivanje pomembnih pravil na enostavnejši in bolj zanesljiv način.

Velika prednost je tudi ta, da KAL tako kot ostale komponente ponuja različne kazalce kakovosti (Performance indicators), ki omogočajo enostavno odločanje, katera pravila uporabiti in katera ne in omogoča izbiranje najbolj relevantnih pravil za posamezne poslovne primere.

Obstaja možnost enostavne vgradnje asociacijskih pravil generiranih s KAL v že obstoječe sisteme v podjetjih.

Prednosti uporabe komponente KAR so predvsem štiri:

- omogoča izdelavo nedvoumnih in razumljivih rezultatov;
- omogoča nenadzorovano podatkovno rudarjenje, kar pomeni, da ni potrebno opredeliti ciljne spremenljivke;
- omogoča raziskovanje izredno velikih naborov podatkov, kar je posledica sposobnosti oblikovanja asociacijskih pravil zgolj na delu raziskovanih podatkov, preden jih je potrebno agregirati (t.i. raziskovanje po kosih);
- omogoča kreiranje koristnih oz. ustreznih pravil.

3.2.2.8. KXEN Model Export (KMX)[22]

Komponenta KMX je namenjena izvažanju zgrajenih modelov v KXEN analitičnem ogrodju v različne računalniške izvorne kode. KMX omogoča predvsem izvoz v SQL, PMML, C++, VB, Java, Java Script, SAS in še mnoge druge manj znane izvorne kode. Izvoz izvorne kode je izredno pomemben, saj to omogoča enostavno vgradnjo zgrajenih modelov v obstoječo računalniško infrastrukturo v podjetjih. Pomembno je poudariti, da se z izvozom modelov kvaliteta ne spremeni. Izvoz modelov je pomemben tudi z vidika testiranja oz. uporabe modelov na novih podatkih. Pomembno je, da modele, zgrajene in primerne za uporabo, ločimo od okolja, kjer izdelujemo nove modele (od KXEN analitičnega ogrodja).

Izvažanje modelov v različne oblike izvorne kode je izredno uporabno, saj se v podjetjih uporablja različne informacijske sisteme. Težave zaradi različnosti sistemov tako izginejo, saj je s pomočjo KMX komponente mogoče izvoziti modele v tako obliko, ki ustreza tehnologiji v podjetju, pri tem pa ne potrebujemo programiranja, spreminjanja,…

KMX je uporaben za različne uporabnike rudarjenja podatkov. Poslovni uporabniki še nikoli prej niso imeli možnosti prenašati zgrajenih modelov za podatkovno rudarjenje. Vedno so potrebovali pomoč informatikov, ki so jim te premike omogočili. Prav tako veščim

[22] Vir: KXEN Product Overview, 2004, str. 6.

uporabnikom tehnologije za rudarjenje po podatkih omogoča enostavno vgradnjo modelov v sisteme, ki so jih uporabljali v preteklem delu, s tem pa enostaven začetek uporabe modela. O prednosti za informatike pa govori že dejstvo, da ni potrebno za generirane modele izdelovati programske kode, saj je ta že izdelana, potreben pa je le klik na miškin gumb. KMX komponenta predvsem omogoča enostavnejšo uporabo zgrajenih modelov.

3.3. POSTOPEK NAPOVEDOVALNE ANALIZE Z ORODJEM KXEN

Uporaba KXEN analitičnega ogrodja je nadvse enostavna, saj je platforma prilagojena znanju povprečnega uporabnika računalniške opreme v podjetju. Seveda govorimo o osnovnem delu z orodjem KXEN. Za poglobljeno raziskovanje rezultatov oz. za pripravljanje modelov, ki so resnično dodelani, je potrebno še vedno statistično znanje, prav tako pa tudi znanje informatikov, saj še vedno v ozadju potekajo določene statistične operacije, deluje pa na podatkovnih zbirkah.

Primer uporabe KXEN analitičnega ogrodja bom prikazal na enostavnem primeru. Gre za napovedovanje osebe, ki ustreza oz. ne ustreza kriteriju, ali je njegova plača višja od 50.000 USD. V bistvu bom iskal osebe, ki imajo potencial, da je njihova letna plača večja od 50.000 USD. Podatke sem pridobil s pomočjo podjetja KXEN od ene izmed univerz v Kaliforniji (University of California, Irvine). Model bomo izdelali na podlagi 48.000 zapisov, ki se nahajajo v bazi podatkov. Zapisi pa so sestavljeni iz petnajstih spremenljivk. V vsakem zapisu imamo tako podatke o starosti, izobrazbi, status, delovnem razredu, spolu, številu delovnih ur na teden, narodnost,… Na podlagi podatkov, ki jih imamo v zbirki podatkov, bomo skušali napovedati vrednost ciljne spremenljivke (plača višja od 50.000 USD). Vrednost ciljne spremenljivke nam lahko izredno pomaga npr. na področju trženja (ljudem, ki imajo višjo plačo od 50.000 USD, npr. lahko ponudimo avto višjega razreda; človeku, ki ne ustreza vzorcu zaposlenega z višjo plačo od 50.00 USD, npr. ne ponudimo luksuznih počitnic, ampak prilagojeno ponudbo,…). Ta podatkovna zbirka služi za učenje oz. oblikovanje modela, ki ga je potem mogoče uporabiti na novih podatkih.

Podatki žal niso s področja bančništva oz. s področja telekomunikacijskih podjetij, kjer KXEN dosega najboljše rezultate z vidika uporabnosti. Razlog pa je predvsem na strani varovanja podatkov. Osnovno delovanje KXEN analitičnega ogrodja pa bomo lahko spoznali tudi na primeru, ki ga lahko uporabim.

Ko poženemo KXEN analitično ogrodje, se nam odpre okno, kjer se odločimo, kaj bomo sploh z orodjem skušali narediti. V osnovnem pristopu se odločamo ali bomo KXEN uporabljali za regresijske napovedovalne modele (K2R), mogoče za oblikovanje segmentov (K2S), za oblikovanje asociacijskih pravil (KAR) oz. za namene časovnih napovedi določene spremenljivke (KTS). Ostale možnosti pa so namenjene podrobnejši analizi podatkov, saj ponujajo še nekaj dodatnih možnosti, ki jih določamo na strani zajemanja podatkov.

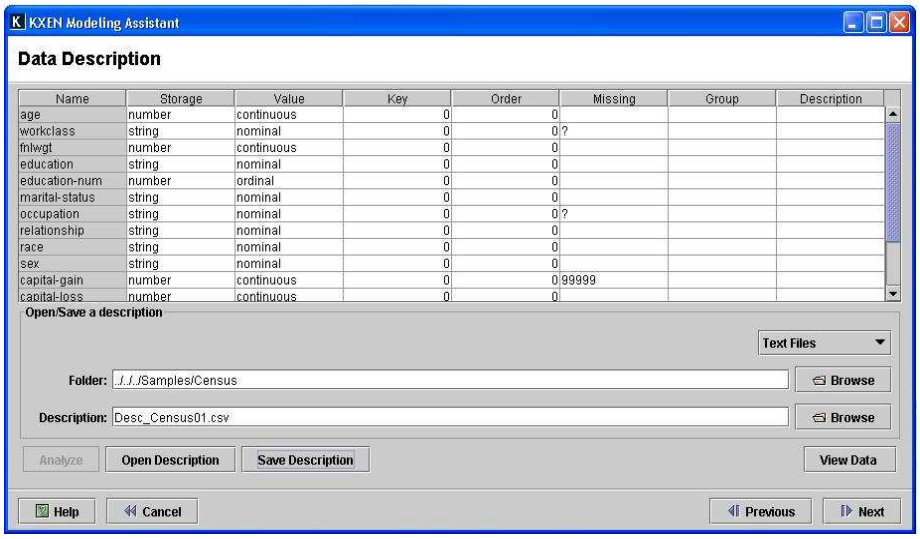
Mi se bomo usmerili na področje napovedovalnih modelov, ki so v praksi tudi najpogosteje uporabljeni modeli, ki so izdelani s pomočjo KXEN-a. Namen je torej odkriti osebo, ki ima plačo višjo od 50.000 USD. Izberemo torej prvo možnost »Classification/Regression (K2R)«.

V naslednjem koraku je potrebno orodju pokazati oz. povedati, na podlagi katerih podatkov skušamo zgraditi določen model. Določimo mesto, kjer se nahajajo podatki (direktoriji, podatkovne baze,…) in podatke, na katerih želimo graditi model (izberemo natančno določeno datoteko oz. tabelo). Opozoriti velja, da brez podatkov iz preteklosti ni napovedovalnih modelov. Poudariti velja, da KXEN omogoča zajemanje podatkov iz najrazličnejših virov. Podprti so vsi bolj znani sistemi za upravljanje z bazami podatkov, prav tako tekstovni dokumenti, možno pa je zajemanje podatkov tudi iz virov, specifičnih za določeno podjetje.

Na tem mestu je potrebno določiti tudi način zajemanja podatkov. Izbiramo lahko med najrazličnejšimi načini: od naključnega izbora podatkov, do točno določenih zapisov v viru podatkov (npr. prvih 20.000 zapisov),… KXEN se namreč na določeni količini podatkov, ki mu jih damo kot vhodno spremenljivko, uči, na določeni količini oblikuje zakonitosti in na določeni količini podatkov testira zgrajene zakonitosti. Običajna je naključna uporaba podatkov, mogoče pa je izbrati tudi drug način (npr. prvih 30 % vstopnih podatkov za učenje, drugih 30 % vstopnih podatkov za oblikovanje zakonitosti in ostali vstopni podatki za namen testiranja oblikovanih zahtev).

V naslednjem koraku je potrebno KXEN analitičnemu ogrodju »pokazati« določene zakonitosti, ki vladajo v naši zbirki podatkov. Prednost KXEN-a za poslovne uporabnike se pokaže na področju avtomatske prepoznave teh zakonitosti (gumb »analyze«), ki omogoča avtomatsko prepoznavanje zakonitosti podatkov (spremenljivke, vrste spremenljivk,…). Za osnovne modele se torej ne zahteva poznavanje strukture podatkov v bazah podatkov, saj KXEN sam analizira vstopne podatke. Za zahtevnejše modele oz. za bolj kvalitetne modele pa je potrebno KXEN-u ročno določiti določene zakonitosti, ki vladajo v zbirki podatkov. Predvsem gre za način zapisovanja manjkajočih vrednosti, saj so v vsaki bazi manjkajoče vrednosti zabeležene drugače. Prednost je tudi ta, da enkrat, ko imamo pripravljen natančen opis strukture podatkov, lahko ta opis shranimo za prihodnjo uporabo. Prav tako lahko na tem mestu odpremo obstoječe opise zakonitosti podatkov.

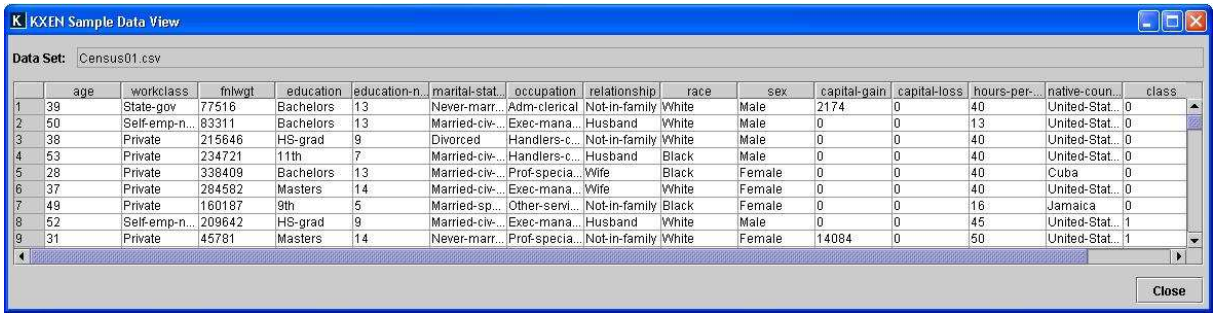
Slika 8: Analiziranje oblike podatkov, na katerih bomo zgradili model



Vir: KXEN Analytic Framework 3.1.0, 2004

Za lažje razumevanje našega primera si na tem mestu oglejmo strukturo podatkov, na katerih bomo izvedli analizo. V tekstovnem dokumentu smo hranili 15 spremenljivk, kot so starost osebe, delovni razred, izobrazba, status (poročen, neporočen, ločen,…), podatke o rasi, podatke o poklicu,… Že prej smo za podatke, na podlagi katerih se učimo, pripravili dodatno spremenljivko, ki se imenuje »Class«. Ta spremenljivka nam služi, da na podatkih iz preteklosti (dejstev) določimo osebo, ki ustreza našemu kriteriju (plača večja od 50.000 USD). Vrednosti spremenljivke Class sta samo dve. 1 pomeni, da oseba ima plačo višjo od 50.000 USD, 0 pa pomeni, da oseba ne ustreza kriteriju. Ta spremenljivka je pomembna, saj se na podlagi te KXEN »uči« zakonitosti, ki jih bo lahko potem uporabil za napoved za novo osebo. Struktura in vrednosti posameznih spremenljivk so prikazane na sliki 9.

Slika 9: Pregled podatkov, na katerih bomo zgradili model

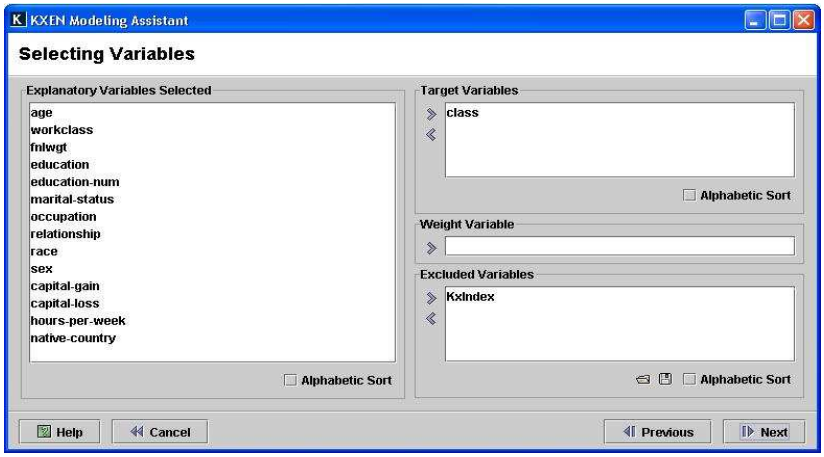


Vir: KXEN Analytic Framework 3.1.0, 2004

V naslednjem koraku moramo KXEN-u povedati, katere spremenljivke bomo vključili v našo analizo in katerih ne bomo. Lepota obravnavanega orodja je v tem, da ni potrebno reduciranje spremenljivk. Analizo lahko izvajamo na praktično neomejenem številu spremenljivk. Pri klasični metodi to enostavno ni bilo mogoče, saj je za »ročno« analiziranje že petnajst spremenljivk težko obvladati.

Določiti je potrebno ciljno spremenljivko. V bistvu se tukaj izbere pogoj, ki ga želimo analizirati. Spodnje desno okno pa je namenjeno spremenljivkam, ki jih ne želimo vključiti v analizo. V našem primeru smo izbrali zgolj spremenljivko »KxIndex«, ki zagotovo ne vpliva na rezultat, saj gre zgolj za »številke« posameznih zapisov, ki pa nimajo vpliva na rezultat, ali oseba ustreza kriteriju ali ne. Izključili smo ga zgolj zaradi hitrosti, saj na ta način prihranimo dodatne stotinke predragocenega časa. Vmesnik za manipuliranje s spremenljivkami je prikazan na sliki 10.

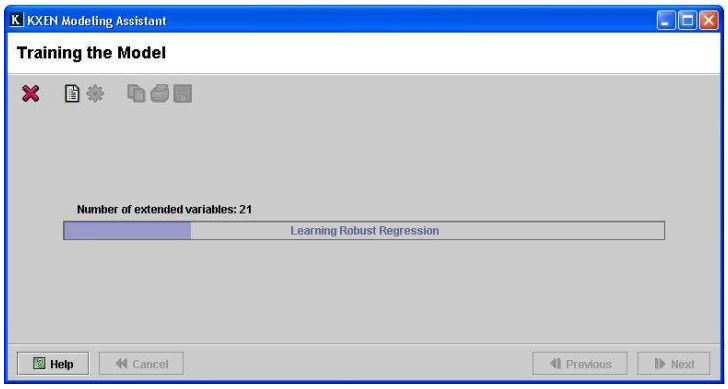
Slika 10: Izbira spremenljivk, ki jih bomo vključili model



Vir: KXEN Analytic Framework 3.1.0, 2004

Na vrsti je čas, ko se generira model, ki ga bo mogoče uporabiti za nove podatke. Postopek oblikovanja modela je podoben procesu učenja pri človeku, saj se tudi KXEN v tem času »uči« določenih zakonitosti, ki jih bo potem uporabil za napovedi rezultata ciljne spremenljivke.

Slika 11: Proces »učenja« na izbranih podatkih oz. gradnja modela



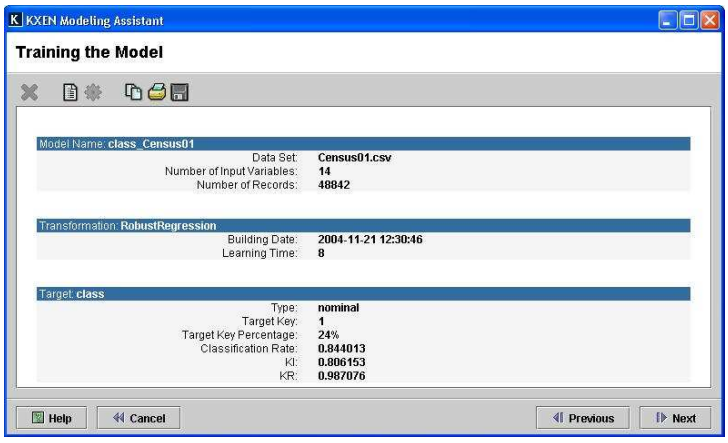
Vir: KXEN Analytic Framework 3.1.0, 2004

Pomembno je poudariti, da je generiranje modela v klasičnih metodah vzelo ogromno časa, saj se v tej fazi oblikujejo zakonitosti oz. model, ki se ga testira, pripravljajo se podatki (vključuje oz. izključuje spremenljivke),… V klasičnih metodah (ročno delo statistikov) bi ta proces za petnajst spremenljivk in 50.000 zapisov trajal kar nekaj dni oz. celo tednov, rezultat

pa bi bil tudi na koncu vprašljiv. KXEN pa s to fazo opravi v nekaj sekundah na povprečno močnem osebnem računalniku. Proces učenja oz. generiranja modela lahko spremljamo na monitorju, določen delček tega procesa pa je v tem delu prikazan na sliki 11.

Ko je postopek kreiranja modela končan, se nam pokaže na zaslonu poročilo o postopku oblikovanja modela. Poročilo je podobno poročilu, ki ga vidimo na sliki 12. V poročilu lahko tako vidimo, na podlagi katerih podatkov se je generiral model, koliko spremenljivk je bilo vključenih v analizo, na kakšni količini podatkov smo izvajali analizo. Prav tako je nekaj splošnih podatkov o tem, kdaj je bil model generiran in koliko časa je KXEN potreboval za »učenje« oz. kreiranje modela. Zanimivo je, da je za naš primer KXEN potreboval zgolj osem sekund, pri klasičnem načinu izdelave podobnega modela pa bi postopek trajal kar nekaj časa. Naslednji razdelek na poročilu pa se nanaša na našo ciljno spremenljivko. Razberemo lahko, da smo izbrali za oblikovanje modela polinom prve stopnje (model je zapisan v obliki matematične formule, ki pa je zgolj polinom prve stopnje). Vidimo lahko, da je na podlagi vstopnih podatkov bilo 24 odstotkov ljudi, ki so ustrezali našemu kriteriju (plača višja od 50.000 USD na letni ravni). Izredno pomembna sta kazalca KI (kazalec kvalitete) in KR (kazalec robustnosti) saj nam ta dva kazalca pripovedujeta, ali je zgrajen model dober in zanesljiv. Pri klasičnih metodah oz. orodjih za rudarjenje po podatkih nismo nikoli vedeli, ali je zgrajene modele mogoče uporabiti in pričakovati dobre rezultate. To je ena izmed bistvenih prednosti KXEN analitičnega ogrodja, saj vedno vemo, ali je model dovolj kvaliteten ali ne. Za naš primer sta oba kazalca na izredno visokem nivoju, kar pomeni, da lahko pričakujemo zanesljive in kvalitetne napovedi za primere, ko bomo zgrajeni model uporabili.

Slika 12: Osnovni podatki o zgrajenem modelu

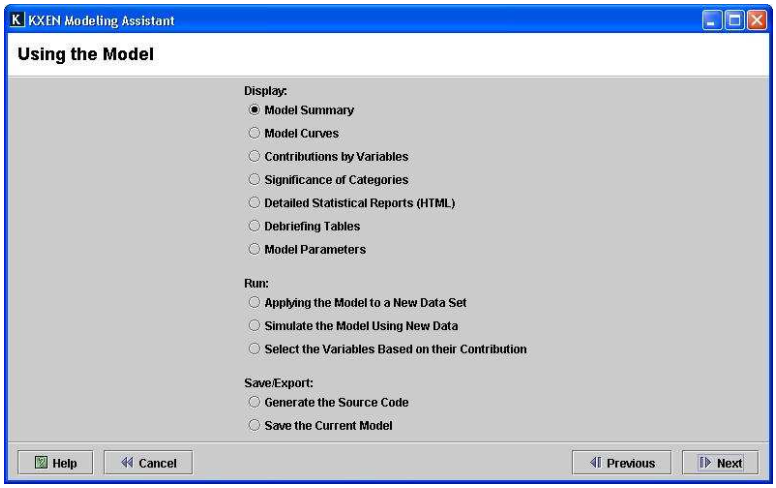


Vir: KXEN Analytic Framework 3.1.0, 2004

V naslednjem koraku lahko izbiramo, kaj bomo z zgrajenim modelom počeli. Na izbiro imamo najrazličnejše možnosti, ki pa so razdeljene v tri skupine, in sicer »preglede«, uporaba modela in možnosti shranjevanja in izvažanja modela. Lahko ponovno pregledamo poročilo o generiranju modela, lahko si pogledamo krivulje modela (kje se nahaja naš model med idealnim modelom in med modelom, ki ga dobimo z naključnim izbiranjem), lahko pogledamo, kako vplivajo posamezne spremenljivke na rezultat, pregledamo lahko podrobna

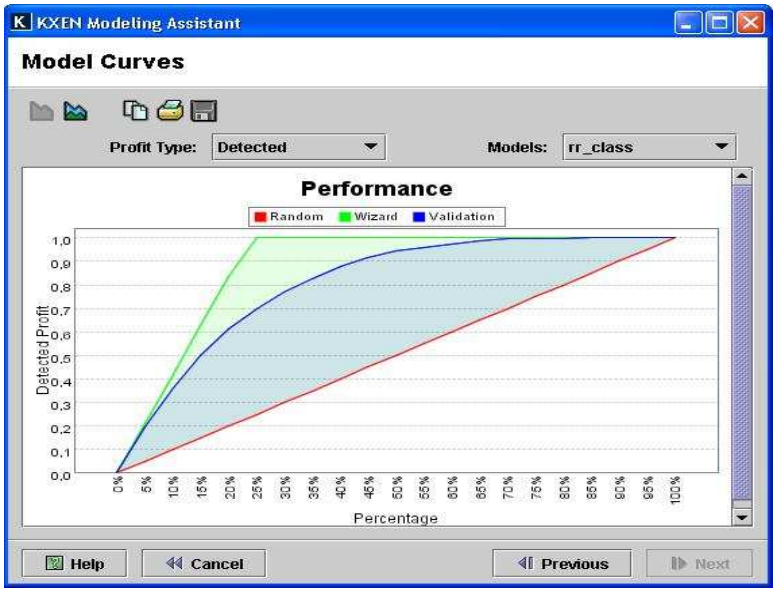
statistična poročila in rezultate določenih izračunov, pregledujemo lahko posamezne parametre modela (spremenljivke, podatke,…). V razdelku o uporabi modela lahko izberemo enega izmed treh načinov uporabe zgrajenega modela. Lahko simuliramo uporabo modela zgolj na enem novem posamezniku, ki ga želimo analizirati, lahko uporabimo model na novem naboru podatkov in analiziramo oz. napovemo, kateri posamezniki ustrezajo kriteriju. Tretja možnost, ki nam je ponujena, pa je, da na najhitrejši možni način spremenimo model z vidika vključevanja in izključevanja posameznih spremenljivk. Tretji način uporabe modela se skriva pod naslednjim razdelkom, saj lahko model izvozimo v različne izvorne kode, ki pa jih potem vgradimo v obstoječe sisteme v podjetju in na ta način uporabimo napovedovalni model. V zadnjem razdelku imamo tudi možnost shranjevanja modela, ki ga lahko naslednjič enostavno odpremo in popravljamo.

Slika 13: Nabor opcij za uporabo modela



Vir: KXEN Analytic Framework 3.1.0, 2004

Slika 14: Pregled krivulj zgrajenega modela



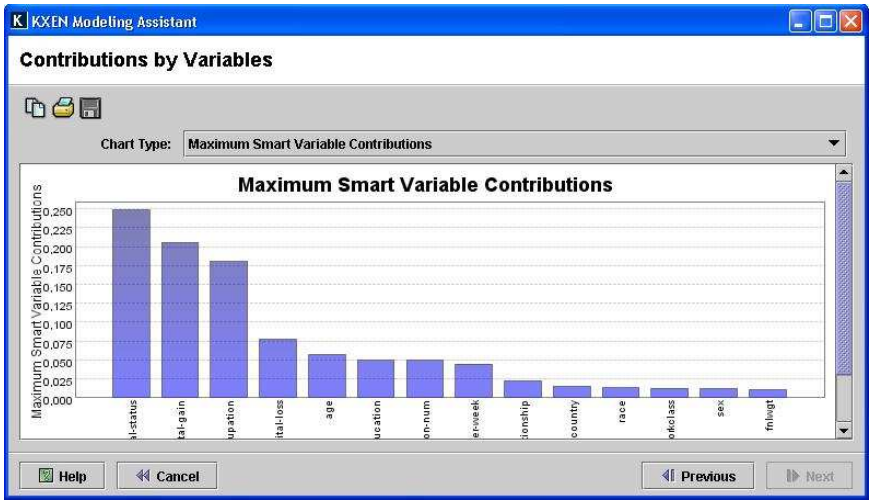
Vir: KXEN Analytic Framework 3.1.0, 2004

Če želimo pregledovati krivulje našega modela, se nam odpre okno, ki ga vidimo na sliki 14. Zelena črta predstavlja idealen model, rdeča prikazuje model, ki bi ga dobili z naključnim izbiranjem posameznikov in ugotavljali, ali ustrezajo ali ne, modra črta pa prikazuje naš model, ki smo ga izdelali. Idealen model pomeni, da bi skušali izmed vseh ljudi izbrati tiste, ki ustrezajo postavljenem kriteriju, pri tem pa ne bi nikoli izbrali napačnega. Izbrali bi 24 odstotkov ljudi iz nabora podatkov, na katerih bi deloval ta model, in nikoli ne bi izbrali napačnega. »Naključni model« pa pomeni, da bi izmed množice odkrili sorazmerno število ljudi, ki bi ustrezalo kriteriju. Če bi izbrali 50 odstotkov vseh ljudi, ki bi jih analizirali, bi odkrili tudi 50 odstotkov tistih, ki ustrezajo postavljenemu kriteriju. Modra črta nam pokaže naš model. Bolj ko se črta približa zeleni črt, boljši je model.

Če želimo pregledovati prispevke posameznih spremenljivk k rezultatu, se nam odpre podobno okno, kot ga vidimo na sliki 15. Spremenljivke so razporejene od tiste, ki na rezultat najmočneje vpliva, do tiste, ki na rezultat vpliva najmanj.

Za naš primer lahko ugotovimo, da je najpomembnejša spremenljivka, ki smo jo vključili v analizo, status, ali je oseba poročena ali ne. Rezultat je zanimiv, ker »na pamet« skoraj nihče ne bi pomislil, da lahko ta spremenljivka tako zelo močno vpliva na končni rezultat (ali oseba ima plačo višjo od 50.000$). Prednost, ki jo ponuja KXEN analitično ogrodje, je tudi v tem, da lahko podrobneje raziskujemo spremenljivke glede na vpliv. Potreben je zgolj dvojni klik na posamezno spremenljivko, saj KXEN omogoča vrtanje v globino (drilling down). Ta možnost pomeni, da lahko natančno raziščemo, katera skupina oz. na kakšen način posamezna skupina vpliva na končni rezultat. Pri vrtanju v globino za spremenljivko starost tako dobimo pregled nad določenimi posameznimi starostnimi skupinami. Za vsako skupino lahko tudi vidimo, ali pozitivno oz. negativno vpliva na končni rezultat, prav tako pa vidimo tudi moč vpliva. Funkcija »drill down« omogoča torej podrobnejšo analizo.

Slika 15: Prispevek spremenljivk, ki smo jih vključili v model

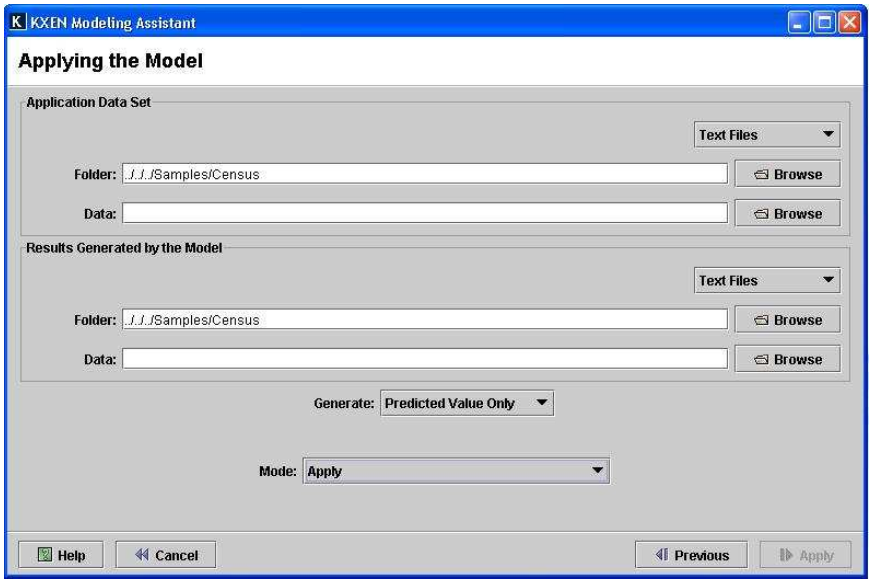


Vir: KXEN Analytic Framework 3.1.0, 2004

V razdelku delnih tabel modela lahko pregledujemo najrazličnejše stvari, povezane z modelom. Lahko pregledujemo vpliv posameznih spremenljivk na končni rezultat, lahko pregledujemo, na kakšen način posamezna spremenljivka vpliva na kazalca kvalitete in robustnosti, pregledujemo lahko posamezne izračune, ki so bili potrebni za izgradnjo modela,…

Če želimo uporabiti zgrajeni model na novih podatkih, imamo tri možnosti. Ena od teh je prikazana na sliki 16. Gre za možnost, da zgrajeni model uporabimo na novem naboru podatkov. Običajno so podatkovne baze v podjetjih precej velike in bi bilo težko zgraditi oz. se »naučiti« model na podlagi vseh podatkov v bazi. Rezultati bi bili sicer v tem primeru zagotovo boljši, saj se z večanjem količine podatkov, na katerih se gradi model, izboljšuje tudi kvaliteta. Zato je v praksi velikokrat potrebno zgraditi model le na določenem deležu vseh podatkov. Npr. na podlagi 10 odstotkov vseh podatkov, ki jih imamo v bazi (recimo, da imamo podatke o plači zgolj za teh 10 odstotkov zapisov). Zgrajeni model pa bi potem radi uporabili še za ostalih 90 odstotkov podatkov v bazi in seveda tudi za vse nove podatke ter z njim napovedali verjetnost, da posamezniki ustrezajo zastavljenemu kriteriju. Za uporabo modela za ostalih 90 odstotkov podatkov je primeren način, ki je prikazan na sliki 16. KXEN-u je potrebno zgolj povedati, na katerih podatkih naj uporabi zgrajeni model.

Slika 16: KXEN analitično ogrodje – možnost uporabe modela na novem naboru podatkov

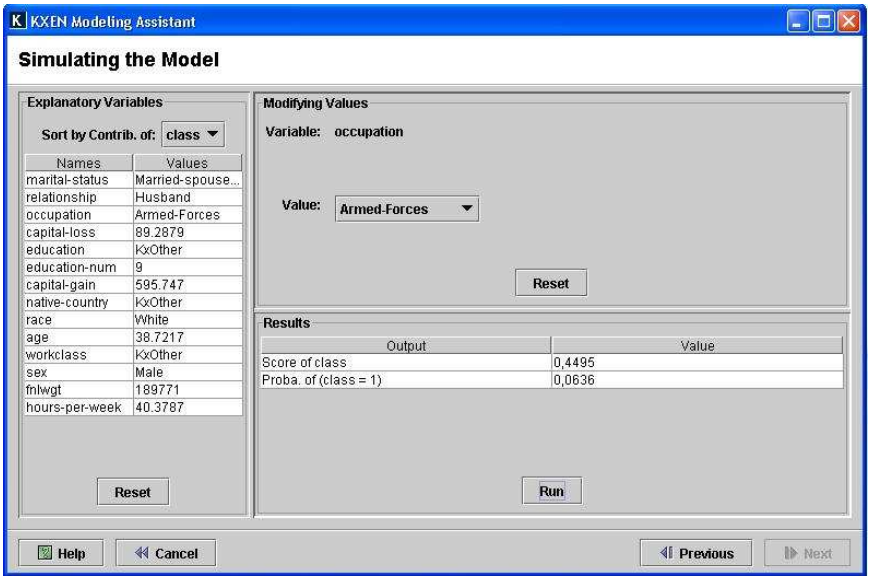


Vir: KXEN Analytic Framework 3.1.0, 2004

Eden izmed načinov uporabe zgrajenega napovedovalnega modela je tudi simulacija delovanja modela za posamezen zapis oz. za posamezno osebo. Simulacijo se izvede neposredno v KXEN analitičnem ogrodju, kjer se nam za te primere odpre okno, podobno oknu na sliki 17. V tem oknu zgolj izberemo oz. vpišemo podatke o osebi, ki jih želimo analizirati, in kliknemo na gumb »run«. Izračuna se nam verjetnost, da vnesena oseba ustreza podanemu pogoju (plača višja od 50.000 USD). Poudariti je potrebno, da za izračun verjetnosti ni potrebno vnašati vseh podatkov. To je še posebej uporabno npr. v klicnih

centrih, ko stranke ne moremo spraševati po vseh podatkih. Model poženemo zgolj na podatkih, ki jih od osebe uspemo pridobiti. Ta način je načeloma uporaben zgolj za tiste primere, ko ni predvidena uporaba modela na velikem številu zapisov oz. oseb.

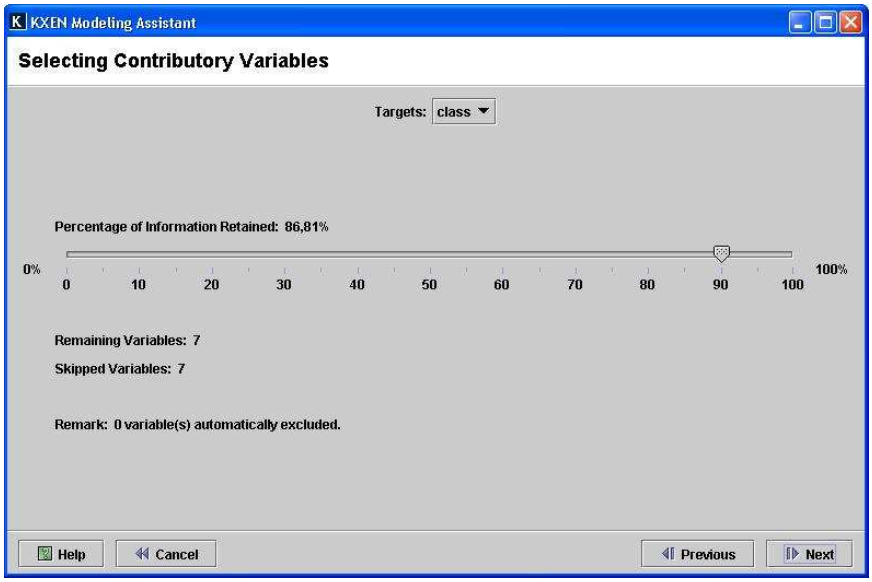
Slika 17: KXEN analitično ogrodje – simulacija uporabe zgrajenega modela na novem zapisu



Vir: KXEN Analytic Framework 3.1.0, 2004

Ena izmed prednosti orodja KXEN je tudi ta, da lahko poljubno izgrajujemo enostavnejše in kompleksnejše modele. Gre za odstranjevanje in dodajanje spremenljivk. Če želimo npr. izločiti nekaj spremenljivk, ki imajo minimalne učinke na rezultat, upočasnjujejo pa delovanje modela, je v KXEN-u dovolj zgolj vpis števila spremenljivk, ki jih želimo uporabiti. Ni potrebno zopet graditi novega modela in s tem zgubljati časa, kot je to bilo potrebno v klasičnih pristopih. Druga možnost pa je, da z drsnikom v oknu, ki je prikazano na sliki 18, KXEN-u določimo zgolj, koliko odstotkov vseh spremenljivk želimo uporabiti. Z zmanjševanjem teh odstotkov se v bistvu odstranjujejo iz modela spremenljivke, ki najmanj vplivajo na rezultat. Z zgrajenim modelom se je torej mogoče »igrati« in ga na ta način še dodatno izpopolniti. Nesmiselno je namreč izvajati analizo nad spremenljivkami, ki ne dodajajo nobene dodatne vrednosti oz. kvalitete h končnemu rezultatu.

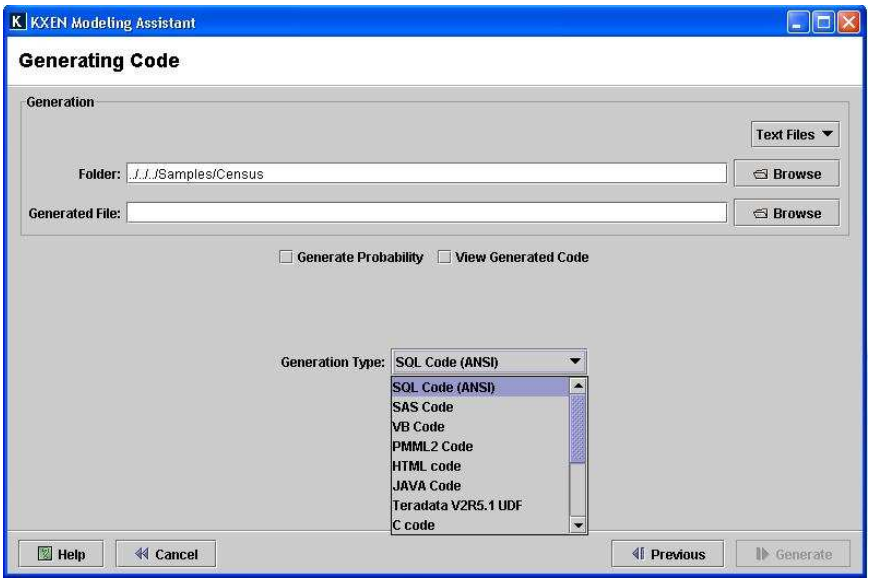
Slika 18: KXEN analitično ogrodje – možnost zmanjševanja števila spremenljivk



Vir: KXEN Analytic Framework 3.1.0, 2004

Zadnja možnost uporabe modelov, zgrajenih s KXEN analitičnem ogrodjem, pa je možnost izvoza izvorne kode v najrazličnejših programskih jezikih. Izvoz izvorne kode je nadvse enostaven, saj je v KXEN-u temu namenjeno podokno, ki ga vidimo na sliki 19. KXEN-u moramo zgolj povedati ime datoteke, ki jo želimo ustvariti, in pa izbrati moramo obliko izvorne kode. KXEN omogoča izvoz modela v vse bolj razširjene programske jezike. Najpomembnejši so zagotovo SQL, VB, C++, Java, HTML, SAS programska koda,… Ta način uporabe modelov je tudi najbolj razširjen, saj lahko zgrajene modele vgradimo v že obstoječe sisteme v podjetju, analiza podatkov oz. napovedovanje pa se izvaja v ozadju, uporabnik pa vidi zgolj rezultat.

Slika 19: KXEN analitično ogrodje – generiranje izvorne kode



Vir: KXEN Analytic Framework 3.1.0, 2004

V tem delu diplomskega dela sem prikazal zgolj osnove postopka izdelave napovedovalnega modela, saj KXEN omogoča še ogromno možnosti, ki so pomembne za zagotavljanje kvalitetnih napovedovalnih modelov.

Bolj kot znanje na področju informatike oz. statistike je pomembna sposobnost postavljanja kvalitetnih poslovnih vprašanj, ki bi jih želeli rešiti s tem orodjem. Običaj v praksi je, da vodstveni delavci v podjetju zgolj povedo, kakšno poslovno vprašanje bi želeli rešiti, naloga informatikov in statistikov pa je, da poslovodstvu zgradijo modele, ki omogočajo odgovore na postavljena vprašanja. Zagotovo je vodstvo veliko bolj zadovoljno, če je rešitev implementirana v izredno kratkem času oz. še bolje, da ima podjetje rešitev pred konkurenco, ki je vsak dan večja, s tem pa postajajo konkurenti tudi čedalje pomembnejši dejavnik uspeha določenega podjetja.

3.4. UPORABA KXEN ANALITIČNEGA OGRODJA V POSLOVNEM SVETU

3.4.1. PRIMER UPORABE V TELEKOMUNIKACIJSKEM PODJETJU[23]

Primer uporabe orodja KXEN se nanaša na veliko evropsko telekomunikacijsko podjetje, ki ponuja telekomunikacijske rešitve na globalnem trgu. Imena podjetja zaradi varovanja poslovnih podatkov ne morem razkriti, lahko pa orišem ozadje poslovanja.

Omenjeno podjetje je vodilni ponudnik telekomunikacijskih storitev v Evropi. Ponujajo lokalno, medkrajevno in mednarodno glasovno in podatkovno storitev, klasično telefonijo, GSM telefonijo, satelitsko telefonijo in vse storitve povezane z internetom. Podjetje ima nekaj milijonov uporabnikov, strankam pa ponujajo svoje storitve v več kot dvestotih državah. Podjetje upravlja z enormnimi količinami podatkov.

Ker je poslovanje na področju mobilne tehnologije izredno tekmovalno, se vsa podjetja v tem segmentu borijo obdržati obstoječe stranke, saj so ugotovila, da so stroški s pridobivanjem nove stranke nekajkrat večji kot stroški, ki nastanejo, če hoče podjetje stranko ohraniti. Obravnavano podjetje je iskalo način, kako bi lahko napovedovali možno odhajanje strank k drugemu ponudniku, saj bi na ta način lahko naredili marsikaj, da bi stranko obdržali. Iskali so način, kako bi lahko preprečili take prehode strank. Težava v podjetju je nastala, ker zaposleni v oddelku za informatiko niso poznali zanesljivega odgovora oz. rešitve na zahteve managementa. V primeru, ko pa se jim je ponujala rešitev, so ugotovili, da zaposleni nimajo dovolj znanja na področju informatike. Poiskati so morali rešitev, ki bo primerna za poslovne uporabnike, poleg tega pa bi zagotavljala dobre rezultate. V začetku so se v trženju s pomočjo določenih statističnih metod usmerjali na skoraj vse uporabnike. Nekaj takega se dogaja z mobilnimi operaterji v Sloveniji, saj v trženju segmentirajo uporabnike po nekem statističnem ključu, ki je splošno znan. V obravnavanem podjetju je ta pristop sicer ponujal neke »zadovoljive« rezultate, po drugi strani pa je bil izredno stroškovno neučinkovit. Cilj vodstva

[23] Vir: http://www.kxen.com/infocenter/downloads/20031119-US-TelcoEuropeII.pdf, 2004.

podjetja je bil odkriti metodo, ki bo zagotavljala dobre rezultate na področju preprečevanja prehajanja strank k drugim operaterjem, hkrati pa bi metoda morala biti visoko stroškovno učinkovita.

Kot rešitev so v podjetju izbrali KXEN, saj jih je prepričala zanesljivost, natančnost, hitrost in primernost uporabe za poslovne uporabnike. Zagotovo pa je bil odločilen tudi dejavnik, da KXEN ponuja hitro povračilo investicije (ROI[24]).

KXEN je podjetju omogočil odkriti različne »skrite« skupine uporabnikov, z različnimi verjetnostmi prehoda k drugim ponudnikom. Različni segmenti so osnova oddelku za trženje, da oblikuje različne pristope k različnim skupinam strank. Na ta način se poveča zadovoljstvo ciljne skupine, hkrati pa so stroški veliko nižji, saj so akcije potrebne za ohranjanje strank bolj usmerjene, ljudem bolj pisane na kožo in tudi veliko cenejše, saj niso vse stranke deležne najugodnejših ponudb.

Orodje pa je svoje kvalitete pokazalo tudi na drugih področjih poslovanja, saj podjetje sedaj nagrade zvestim strankam oz. bonuse za različne stranke oblikuje na podlagi rezultatov modelov, ki so generirani s KXEN-om. KXEN je za podjetje nepogrešljiv na področju trženja oz. organiziranja trženjskih akcij.

3.4.2. PRIMER ODKRIVANJA PREVAR (BANKE, ZAVAROVALNICE)[25]

V naslednjem primeru bom skušal prikazati uporabno vrednost za banke in zavarovalnice.

Primer temelji na izkušnjah Dişbank, ene vodilnih bank v Turčiji. Banka ima približno milijon strank, ki uporabljajo kreditne kartice, poslovanje pa poteka preko 65.000 POS terminalov. V zadnjih dveh letih so v banki zasledovali strategijo povečevanja števila komitentov, s tem pa so se soočali s težavami, ki jim ni bilo videti konca. Število prevar s kreditnimi karticami je naraščalo hitreje kot prihodki od novih strank. Ljudje so preko kreditnih kartic kupovali najrazličnejše stvari, za katere pa banka ni nikoli dobila plačila. Težavo so najprej skušali rešiti z določenimi pravili, ki so jih generirali za vsako transakcijo, ki se je zgodila preko POS terminalov. Sistem pravil je zahtevo po odobritvi obdelal in v primeru, da je bila transakcija sumljiva, so zaposleni ročno preverjali, ali je zahteva lahko odobrena ali ne. Sistem je zagotovil kratkoročen uspeh, saj so bili zaposleni preobremenjeni, da bi preverili vsako zahtevo in tako je banka vsak dan zgubljala velike vsote denarja. Sistem preprečevanja prevar s pomočjo rešitve na podlagi pravil je bil pomanjkljiv. Potrebno je bilo ukrepati.

V položaju reševanja težav banke se je znašla projektna ekipa, ki je trgu ponujala rešitve orodja KXEN. Banka je zahtevala enostavno rešitev, ki jo je mogoče čez noč implementirati v

[24] ROI – Return of investment – kazalec povratka investicije
[25] Vir: http://www.kxen.com/infocenter/downloads/DisbankFraudDetection.pdf, 2004.

obstoječo informacijsko tehnologijo, za uporabo pa ni potrebno dodatno izobraževanje oz. bremenitev zaposlenih. Ekipa je banki na podlagi preteklih zabeleženih podatkov prevar v enem dnevu izdelala model, ki naj bi prepričal vodilne v banki za nakup. Na podlagi testnih podatkov, ki so jih imeli na voljo, jim je uspelo ugotoviti oz. določiti 92 % prevar preko kreditne kartice (model so preverili na določeni količini zapisov v bazi, ki so temeljili na realnih podatkih). Banka se je na podlagi rezultatov testnega modela odločila za nakup orodja. Vgradnja sistema v obstoječo infrastrukturo je bila enostavna, saj je KXEN kompatibilen z najpogostejšimi sistemi za upravljanje z bazami podatkov. V banki se je oblikovala ekipa tehnikov, ki so sodelovali z ekipo, ki je KXEN zastopala. V dveh tednih so izdelali kar nekaj modelov, ki naj bi reševali težave prevar. Najboljša stvar pri vsem je bila, da so lahko modele vgradili v njihov sistem, pri tem pa se je preverjanje vsake zahteve za odobritev plačila preko kreditne kartice preverila v realnem času in brez pomoči zaposlenih. Zaposleni so morali ročno preverjati zgolj nekaj primerov zahtev po odobritvi plačila, saj so bili določeni primeri zapleteni in je odločitev temeljila na podlagi občutka zaposlenih. Iz 300.000 alarmov, ki naj bi jih zaposleni reševali, se je število alarmov zmanjšalo na 30.000 na četrtletje (ni bil vsak alarm kasneje tudi resnična prevara). S pomočjo KXEN-a so modele nadgrajevali, saj so vsak dan zgradili nov model, ki je bil zgrajen na večjem številu podatkov, s tem pa se je kvaliteta izboljšala, saj se je odkrilo nova pravila prevar, oz. se je odkrilo do tedaj še neznane vzorce prevar.

Rezultati, ki so jih dosegli, so bili naravnost presenetljivi. Število prevar s kreditno kartico se je zmanjšalo za 67 odstotkov, iz 21-ih prevar dnevno, se je število zmanjšalo na 7 prevar dnevno. Banka z uporabo novih modelov zgrajenih s KXEN-om, vsak dan privarčuje približno 25.000 dolarjev, investicija v orodje in izobraževanje zaposlenih pa se jim povrne vsak teden poslovanja. Pozitiven vpliv ima uporaba orodja tudi na zaposlene v podjetju, saj sedaj niso več tako obremenjeni, zaradi tega pa so postali tudi veliko bolj učinkoviti. Težava se je pojavila zgolj na strani prevarantov, saj je od dneva začetke uporabe sistema banki uspelo za zapahe spraviti že več kot petdeset prevarantov, ki so skušali izkoristiti neobvladljivo situacijo v banki. Pozitivne vplive uporabe sistema je moč zaznati tudi na strani komitentov, saj je z zmanjšanjem stroškov prevar banki uspelo komitentom kreirati ugodnejše pogoje poslovanja, s tem pa se število komitentov veča.

4. SKLEP

V diplomski nalogi sem definiral pojem podatkovnega rudarjenja in predstavil nekaj najbolj pogosto omenjenih tehnik za podatkovno rudarjenje. Predstavil sem eno izmed boljših orodij za podatkovno rudarjenje in tudi pokazal, kako enostavno je delo s tem orodjem. Natančno sem opredelili posamezne komponente KXEN analitičnega ogrodja in predstavil, kakšno funkcijo opravljajo pri podatkovnem rudarjenju. Na enostavnem primeru sem prikazal postopek izdelave napovedovalnega modela oz. postopek podatkovnega rudarjenja. Na koncu sem predstavil še dve zgodbi o uspehu s pomočjo podatkovnega rudarjenja in KXEN analitičnega ogrodja.

Poslovno okolje podjetij postaja čedalje bolj kompleksno, temu pa je potrebno prilagoditi tudi postopke in tehnologije, ki jih podjetja uporabljajo v svojem vsakodnevnem poslovanju. S povečevanjem kompleksnosti poslovanja posameznih podjetij se povečuje količina podatkov, ki jih podjetja shranjujejo, temu pa je potrebno prilagajati tudi informacijsko tehnologijo v podjetjih. Pomembno vlogo pri oblikovanju uspešnih strategij podjetij igrajo tudi izkušnje oz. vzorci, s katerimi se je podjetje srečevalo že v preteklem poslovanju. Ob povečani količini podatkov je podatkovno rudarjenje nepogrešljivo, saj podjetjem omogoča, da ogromne količine podatkov uporabljajo sebi v prid. Uporaba podatkovnega rudarjenja je zagotovo v porastu, saj se čedalje več podjetij zaveda prednosti, ki jih podatkovno rudarjenje prinaša v primerjavi s konkurenti. Kvalitetna orodja za podatkovno rudarjenje, kot npr. KXEN, so v vsakodnevnem poslovanju nepogrešljiva, prihodnosti brez tovrstnih orodij pa si enostavno ni mogoče predstavljati.

Iz lastnih izkušenj vem, da tudi v slovenskih podjetjih čedalje bolj pogosto vodstvo razmišlja o kvalitetnem izkoriščanju podatkov, ki jih shranjujejo v svojih podatkovnih bazah. Vodstvo se zaveda pomembnih vzorcev, ki se pojavljajo v zbranih podatkih in čedalje več podjetij se odloča za orodja, ki bi jim omogočala kvalitetno in zanesljivo podatkovno rudarjenje ter s tem povezano bolj konkurenčno poslovanje. Tudi slovenska podjetja torej sledijo svetovnemu trendu povečevanja shranjene količine podatkov in izkoriščanja podatkov iz preteklosti za konkurenčnejše poslovanje v prihodnosti. Prihodnost podatkovnega rudarjenja je torej tudi v slovenskih podjetjih svetla.

LITERATURA

1. Berry Michael J. A., Linoff Gordon: Data Mining Techniques. Indianapolis : Wiley Pub, 2004. 643 str.
2. Berry Michael J.A., Linoff Gordon: Mastering Data Mining: The Art and Science of Customer Relationship Management. New York : John Wiley & Sons, Inc., 2002. 494 str.
3. Berson Alex, Smith Stephen, Thearling Kurt: Building Data Mining Applications for CRM. New York (etc.) : McGraw-Hill, 2000. 510 str.
4. Blejec Matjaž et al.: Statistika. Piran : Gea College, Visoka šola za podjetništvo, 2003. 150 str.
5. Edelstein Herbert A.: Introduction to Data Mining and Knowledge Discovery. B.k. : Two Crows Corp. 1999. 36 str.
6. Greening Dan R.: Data Mining on the Web - There's Gold in that Mountain of Data. B.k. 1999.
7. Hand David J., Mannila Heikki, Smyth Padhraic: Principles of Data Mining. Cambridge : The MIT Press, 2001. 546 str.
8. Hastie Trevor, Tibshirani Robert J., Friedman Jerome H.: The Elements of Statistical Learning. New York : Springer, 2001. 533 str.
9. Klaves Gregor: Uporaba poslovne inteligence v telekomunikacijskih podjetjih. Magistrsko delo. Ljubljana : Ekonomska fakulteta, 2003. 87 str.
10. Kneževič Snežana: Uporaba izkopavanja podatkov: primer podjetja Merkur, d.d.. Diplomsko delo. Ljubljana : Ekonomska fakulteta, 2002. 49 str.
11. Konič Blaž: Uporaba podatkovnega rudarjenja pri odkrivanju nezaželene elektronske pošte. Diplomsko delo. Ljubljana : Ekonomska fakulteta, 2003. 39 str.
12. Metaspectrum 60.1 - Data Mining Tools. Stamford :Metagroup, Inc., 2003, 94 str.
13. Nisbet Robert: Data Mining Tools Review 2003 - How to Choose a Data Mining Suite, 2003. 10 str.
14. Pirc Mitja: Analitični CRM oziroma kako iz podatkov dobiti informacije, iz informacij pa znanje. B.k., B.l.
15. Thearling Kurt: An Introduction to Data Mining. [http://www.thearling.com/dmintro/dmintro_frame.htm], 18.12.2004.
16. Thearling Kurt: Campaign Optimization: Maximizing the Value of Interacting with Your Customers. [http://www.thearling.com/text/optimization/optimization.htm], 18.12.2004.

VIRI

1. 1- short presentation. Paris : KXEN, 2003. 10 str.
2. AAAI Spring Symposium on Information Refinement and Revision for Decision Making:Modeling for Diagnostics, Prognostics, and Prediction Software and Data, 2002. [http://www.cs.rpi.edu/~goebel/ss02/software-and-data.html], 17.12.2004.
3. Angoss domača stran. [http://www.angoss.com/], 17.12.2004.
4. Baza Whatis.com. [http://www.whatis.com], 22.11.2004.
5. Case Study: Fraud Detection. [http://www.kxen.com/infocenter/downloads/DisbankFraudDetection.pdf], 14.11.2004.
6. Case Sudy: Telecommunications. [http://www.kxen.com/infocenter/downloads/20031119-US-TelcoEuropeII.pdf], 14.11.2004.
7. Data Mining Tools. [http://compstat.chonbuk.ac.kr/jskang/mystudy/2000SecondTerm/soft_cls.htm], 17.12.2004.
8. Data Mining Tools. [http://www.metagroup.com/webhost/ONLINE/477658/60.1marketsummary.pdf], 18.12.2004.
9. Interna gradiva podjetja KXEN.
10. Interno gradivo podjetja UT informacijski sistemi d.o.o..
11. KXEN Analytic Framework. [http://www.kxen.com/products/], 14.11.2004.
12. KXEN Analytic Framework 3.1.0, 2004.
13. KXEN Analytic Framework. User Guide 3.1.0.. Paris : KXEN, 2004. 252 str.
14. KXEN Event Log (KEL). [http://www.kxen.com/products/components/kel.php], 15.11.2004.
15. KXEN Model Export (KMX). [http://www.kxen.com/products/components/kar.php], 15.11.2004.
16. KXEN Model Export (KMX). [http://www.kxen.com/products/components/kmx.php], 14.11.2004.
17. KXEN Product Overwiev. Paris : KXEN, 2004. 7 str.
18. KXEN Robust Regression (K2R). [http://www.kxen.com/products/components/k2r.php], 14.11.2004.
19. KXEN Sequence Coder (KSC). [http://www.kxen.com/products/components/ksc.php], 15.11.2004.
20. KXEN Smart Segmenter (K2S). [http://www.kxen.com/products/components/k2s.php], 14.11.2004.
21. KXEN Time Series (K2C). [http://www.kxen.com/products/components/k2c.php], 15.11.2004.
22. KXEN Time Series (KTS). [http://www.kxen.com/products/components/kts.php], 14.11.2004.
23. Raziskovalni odbor. [http://www.kxen.com/about/scientific_board.php], 23.11.2004.

24. SAS: Analytic Intelligence – Data and Text Mining. [http://www.sas.com/technologies/analytics/datamining/index.html], 18.12.2004.
25. SAS: Analytic Intelligence. [http://www.sas.com/technologies/analytics/datamining/miner/], 18.12.2004.
26. SPSS – Predictive Analytics. [http://www.spss.com/predictive_analytics/index.htm?source=homepage&hpzone=pa_text_link], 18.12.2004.
27. SPSS Clementine. [http://www.spss.com/clementine/], 18.12.2004.
28. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000. [http://www.acm.org/sigs/sigkdd/kdd2000/], 14.11.2004.
29. Thearling.com. [http://www.thearling.com/index.htm#wps], 14.11.2004.
30. Using Darwin – Release 3.0.1. [http://download-uk.oracle.com/otn_hosted_doc/darwin/misc/27627_275039.pdf], 18.12.2004.

PRILOGE

PRILOGA 1: Razlaga uporabljenih kratic
PRILOGA 2: Slovarček slovenskih prevodov tujih izrazov

PRILOGA 1: Razlaga uporabljenih kratic

KRATICA	CELOTEN POMEN KRATICE
BI	Business Intelligence
CRM	Customer Relationship Management
DM	Data Mining
HTML	HyperText Markup Language
K2C	KXEN Consistent Coder
K2R	KXEN Robust Regression
K2S	KXEN Smart Segmenter
KAR	KXEN Association Rules
KEL	KXEN Event Log
KI	KXEN Information Indicator
KMX	KXEN Model Export
KPI	Key Performance Indicators
KR	KXEN Robustness Indicator
KSC	KXEN Sequence Coder
KTS	KXEN Time Series
KXEN	Knowledge Extraction Engines
MB	megabyte
PMML	Predictive Model Markup Language
ROI	Return on Investment
SQL	Structured Query Language
SRM	Structured Risk Minimization
SRM	Structured Risk Minimisation

PRILOGA 2: Slovarček slovenskih prevodov tujih izrazov

TUJ IZRAZ	**SLOVENSKI PREVOD**
C++	objektno orientiran programski jezik
Customer Relationship Management	upravljanje odnosov s strankami
Data Mining	podatkovno rudarjenje
HyperText Markup Language	hipertekstni označevalni jezik
Java	programski jezik, ki je bil prioritetno ustvarjen za uporabo v internet okolju
Key Performance Indicators	glavni kazalniki učinka
Knowledge Extraction Engines	Orodje za izločevanje znanja
KXEN Analytic Framework	KXEN analitično ogrodje
KXEN Association Rules	komponenta KXEN analitičnega ogrodja namenjena kreiranju pravil na podlagi »skritih« vzorcev
KXEN Consistent Coder	komponenta KXEN analitičnega ogrodja namenjena avtomatični pripravi podatkov
KXEN Event Log	komponenta KXEN analitičnega ogrodja namenjena zajemanju podatkov po posameznih dogodkih
KXEN Information Indicator	kazalec kvalitete
KXEN Model Export	komponenta KXEN analitičnega ogrodja namenjena izvozi modelov v različnih izvornih kodah
KXEN Robust Regression	komponenta KXEN analitičnega ogrodja namenjena napovedovalni analizi
KXEN Robustness Indicator	kazalec zanesljivosti
KXEN Sequence Coder	komponenta KXEN analitičnega ogrodja namenjena zajemanju podatkov po sekvencah
KXEN Smart Segmenter	komponenta KXEN analitičnega ogrodja namenjena segmentiranju
KXEN Time Series	komponenta KXEN analitičnega ogrodja namenjena odkrivanju vzorcev in trendov
megabyte	enota za merjenje količine podatkov
Predictive Model Markup Language	programski jezik, ki je zasnovan na podlagi XML-a in omogoča definicijo in deljivost napovedovalnih modelov med aplikacijami
Return on Investment	povračilo na investicije
SAS Institute, Inc.	ameriško podjetje, ki se ukvarja z razvojem programske opreme
Structured Query Language	strukturirani poizvedovalni jezik
Structured Risk Minimisation	strukturirano minimiziranje tveganja
Structured Risk Minimization	strukturirano minimiziranje tveganj