

Podatkovno rudarjenje v Oracle okolju

Žiga Vaupot

ziga.vaupot@bron.si

16. strokovno srečanje
SIOUG 2011

Kongresni center Bernardin
Portorož, 18. - 21. September 2011



BRON

- 16 redno zaposlenih, 2 redna zunanja sodelavca (polni delovni čas), stalno izobraževanje v tujini in pridobivanje certifikatov
- združujemo poslovna, tehnična in praktična znanja pri uvajanju zahtevnih informacijskih sistemov
- sodelujemo v vseh fazah projektov, od zasnove, planiranja, implementacije do upravljanja
- področja dela:
 - Oracle Business Intelligence (OBI, ODI, OWB)
 - Oracle Hyperion Planning
 - Razvoj aplikativne programske opreme (Oracle, Java)
 - načrtovanje in implementacija zahtevnih storitveno orientiranih sistemov (SOA)
- zastopamo

ORACLE

sgi[®]
INNOVATION
FOR RESULTS™

Agenda

- Uvod v podatkovno rudarjenje
- Oracle Data Mining
 - Funkcije podatkovnega rudarjenja
 - Algoritmi podatkovnega rudarjenja
 - Priprava podatkov
- Demo



Kaj je podatkovno rudarjenje?

Podatkovno rudarjenje je proces odkrivanja smiselnih povezav, vzorcev in trendov v velikih bazah podatkov, z uporabo tehnologij za razpoznavo vzorcev ter statističnih in matematičnih tehnik.

--- Gartner Group

Definicija po Oraclu

- Podatkovno rudarjenje je proces preiskovanja velikih baz podatkov s ciljem odkrivanja vzorcev in trendov, ki presegajo okvire enostavne analize.
- Podatkovno rudarjenje uporablja matematične algoritme za segmentacijo podatkov in ocenjevanje verjetnosti dogodkov v prihodnosti.
- Oracle ponuja „all-in-database“ podatkovno rudarjenje

Ključne lastnosti podatkovnega rudarjenja

- Avtomatično razpoznavanje vzorcev
- Predvidevanje najverjetnejših dogodkov
- Kreiranje informacij potrebnih za odločitve
- Osredotočenje na velike zbirke podatkov
- Podatkovno rudarjenje odgovarja na vprašanja, na katera z enostavnimi poizvedbami ne moremo odgovoriti

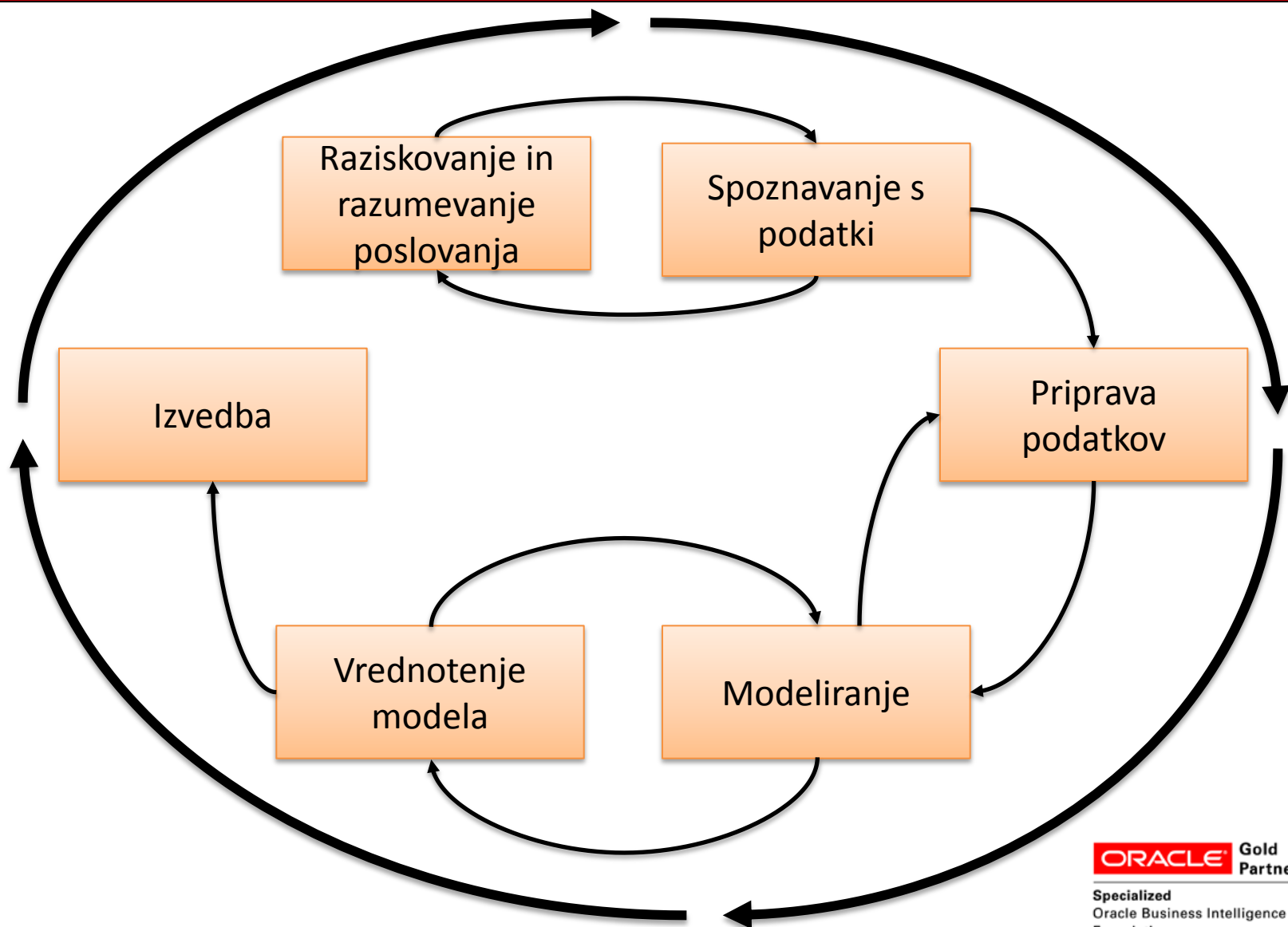
Tipični primeri aplikativne uporabe podatkovnega rudarjenja

- Klasifikacija komitentov, ki zaprošajo za kredit glede na tveganja: nizko, srednje in visoko.
- Ugotavljanje prijavljenih škodnih primerov z namenom goljufije.
- Napovedovanje, kateri od obstoječih kupcev bodo v naslednjih šestih mesecih nadgradili svojo storitev.
- Kupci, ki kupujejo izdelek X, najpogosteje kupijo še izdelka Y in Z.

... in še cela vrsta drugih

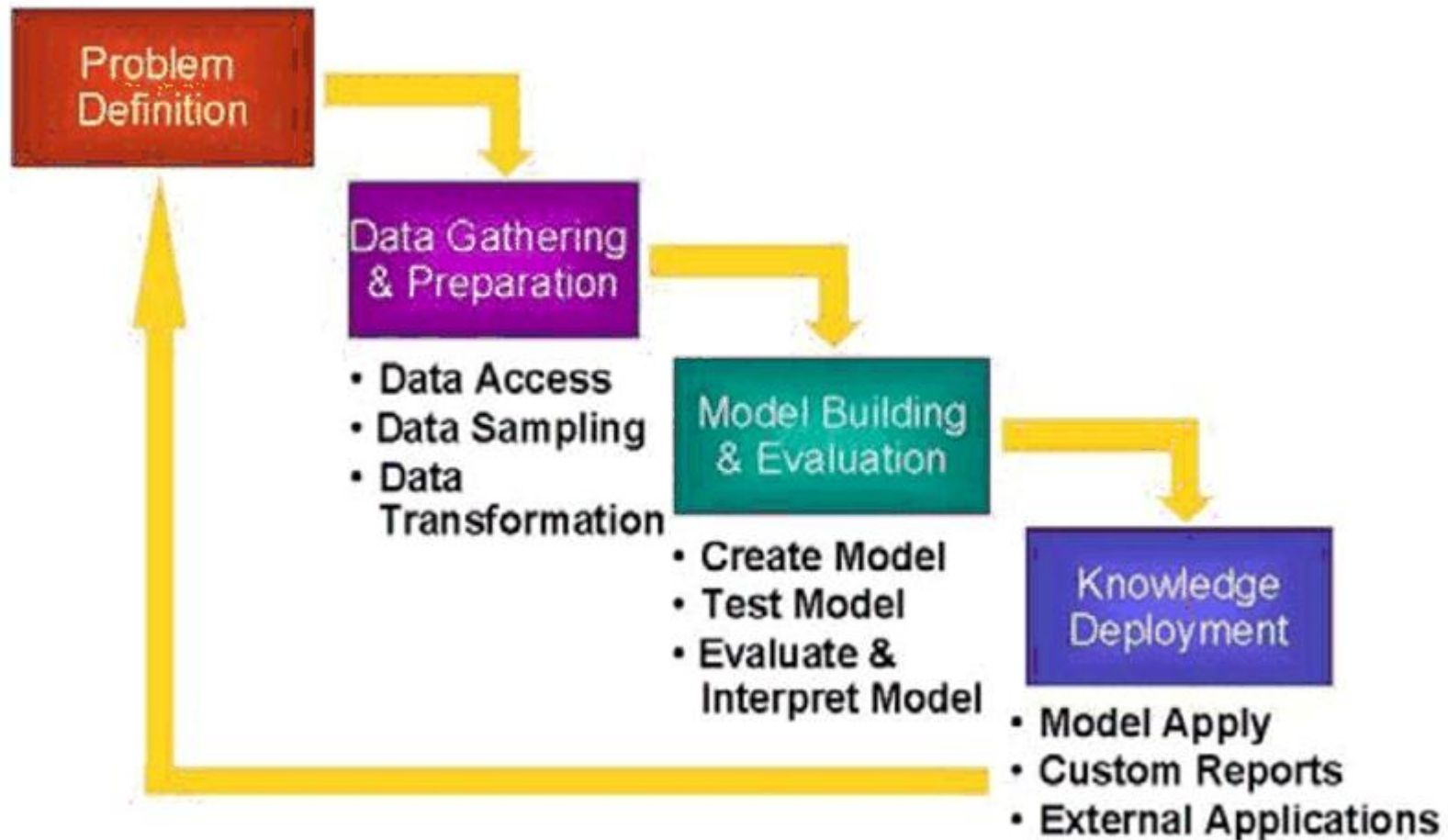
- **Retail**
 - Customer segmentation
 - Response modeling
 - Recommend next likely product
 - Profile high value customers
- **Banking**
 - Credit scoring
 - Probability of default
 - Customer profitability
 - Customer targeting
- **Insurance**
 - Risk factor identification
 - Claims fraud
 - Policy bundling
 - Employee retention
- **Higher Education**
 - Alumni donations
 - Student acquisition
 - Student retention
 - At-risk student identification
- **Healthcare**
 - Patient procedure recommendation
 - Patient outcome prediction
 - Fraud detection
 - Doctor & nurse note analysis
- **Life Sciences**
 - Drug discovery & interaction
 - Common factors in (un)healthy patients
 - Cancer cell classification
 - Drug safety surveillance
- **Telecommunications**
 - Customer churn
 - Identify cross-sell opportunities
 - Network intrusion detection
- **Public Sector**
 - Taxation fraud & anomalies
 - Crime analysis
 - Pattern recognition in military surveillance
- **Manufacturing**
 - Root cause analysis of defects
 - Warranty analysis
 - Reliability analysis
 - Yield analysis
- **Automotive**
 - Feature bundling for customer segments
 - Supplier quality analysis
 - Problem diagnosis
- **Chemical**
 - New compound discovery
 - Molecule clustering
 - Product yield analysis
- **Utilities**
 - Predict power line / equipment failure
 - Product bundling
 - Consumer fraud detection

Podatkovno rudarjenje je proces*



*CRISP-DM: Cross-Industry Standard Process for Data Mining

Data Mining proces po Oraclu



Oracle Data Mining

Funkcije podatkovnega rudarjenja

Funkcije podatkovnega rudarjenja

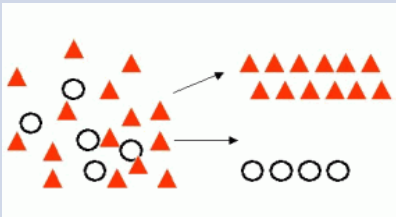
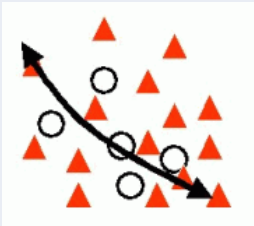
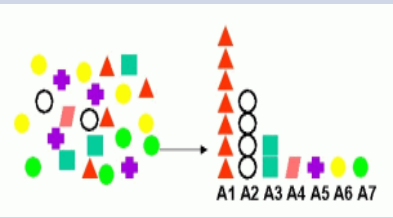
- Nadzorovano učenje*
 - strojno učenje iz učne množice primerov, ki imajo pripadajoče izhodne vrednosti; rezultat nadzorovanega učenja je napovedni model, ki za poljuben vhodni objekt napove pripadajočo izhodno vrednost
- Nenadzorovano učenje*
 - strojno učenje iz učne množice primerov, ki nimajo dodeljenih izhodnih vrednosti; rezultat nenadzorovanega učenja je povzetek ali pojasnjevanje množice učnih primerov

* vir: islovar.org (Terminološki slovar informatike)

Nadzorovano oz. usmerjano podatkovno rudarjenje

- Učni proces je usmerjan v odvisnosti od znanega odvisnega oz. ciljnega atributa
- Obnašanje ciljnega atributa je funkcija neodvisnih atributov ali *prediktorjev*
- Nadzorovano podatkovno rudarjenje praviloma rezultira v *prediktivnem* modelu
- Za pripravo modela razdelimo osnovni tj. znani podatkovni set na podatke s katerim zgradimo model, in na podatke s katerimi model testiramo, s čimer preverimo ali je zgrajeni model dovolj dober za splošno uporabo.

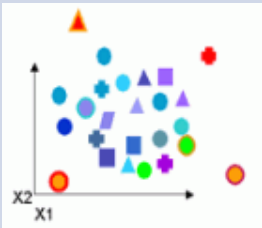
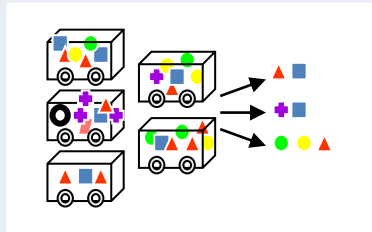
Funkcije nadzorovanega učenja

Funkcija	Opis	Tipični problem
Klasifikacija 	Razvrščanje <i>primerov</i> v diskretne razrede in napovedovanje razredov v katerega posamezen nov primer sodi	Glede na znane podatke o kupcih, napovedovanje njihovega odziva na marketinško akcijo
Regresija 	Ocenjevanje in napovedovanje zveznih vrednost	Na osnovi demografskih podatkov in zgodovine nakupov napovedovanje starosti kupca za določen izdelek.
Pomembnost atributov 	Identifikacija atributov, ki so najpomembnejši za napovedovanje ciljnega atributa	Na osnovi odziva na marketinško akcijo poišči najpomembnejše <i>prediktorje</i> .

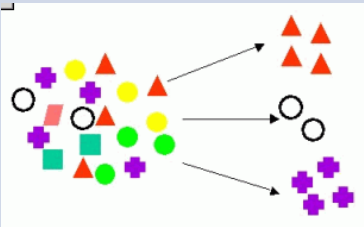
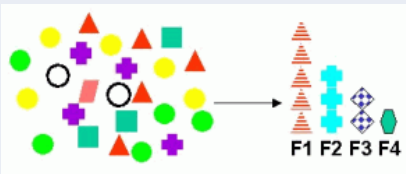
Nenadzorovano oz. neusmerjano podatkovno rudarjenje

- Ni razlike med odvisnimi in neodvisnimi atributi v podatkovnem setu.
- Tudi ni poznanega rezultata s katerim bi zgradili model.
- Z modelom opisujemo in pojasnjujemo odnose med podatki v podatkovnem setu

Funkcije nenadzorovanega učenja (1)

Funkcija	Opis	Tipični problem
Ugotavljanje anomalij v podatkih 	Identifikacija <i>primerov</i> , ki ne ustrezajo <i>normalnim</i> karakteristikam podatkov	Na osnovi demografskih podatkov identificirati kupce z nenavadnimi nakupnimi navadami.
Asociativna pravila 	Identifikacija <i>primerov</i> , ki se skupaj pojavljajo v podatkih in specifikacija pravil, ki veljajo za takšno skupno pojavljanje.	Poišči <i>primere</i> izdelkov, ki se pogosteje nakupujejo skupaj in opredeli relacije med njimi.

Funkcije nenadzorovanega učenja (2)

Funkcija	Opis	Tipični problem
Razvrščanje v gruče 	Identifikacija naravnih gruč v podatkih	Segmentacija demografskih podatkov v gruče in rangiranje verjetnosti, da posamezen <i>primer</i> sodi v posamezno grupo.
Izločanje lastnosti 	Kreiranje novega atributa (<i>lastnosti</i>) z uporabo linearne kombinacije originalnih atributov.	Na osnovi demografskih podatkov določenega nabora kupcev povežemo skupino atributov z visoko korelacijo v novo splošno <i>lastnost</i> kupca.

Oracle Data Mining

Algoritmi

Algoritmi nadzorovanega učenja

Algoritem	Funkcija	Opis
Decision Tree (DT)	Klasifikacija	Odločitvena drevesa omogočajo izločitev informacij, ki so ključne za napovedovanje v obliki razumljivih pravil. Le ta so <i>if-then-else</i> izrazi, ki pojasnjujejo odločitve, ki vodijo do napovedi.
Generalized Linear Models (GLM)	Klasifikacija Regresija	GLM implementira logistično regresijo za klasifikacijo binarnih ciljnih vrednosti in linearno regresijo za zvezne ciljne vrednosti. Pri klasifikaciji podpira meje zaupanja za verjetnostno porazdelitev napovedi, oz. pri regresiji meje zaupanja za napovedi.
Minimum Description Length (MDL)	Pomembnost atributov	MDL predpostavlja, da je za najboljšo in najbolj verjetno razlago podatkov dovolj enostavna in najbolj kompaktna predstavitev podatkov.
Näive Bayes (NB)	Klasifikacija	NB za svoje napovedi uporablja Bayesov teorem, ki izračuna verjetnost napovedi na osnovi dejstev, ki so prisotna v podatkih.
Support Vector Machine (SVM)	Klasifikacija Regresija	SVM omogoča uporabo različnih funkcij za različne nabore podatkov. Podprti sta linearna in gausova (nelinearna) funkcija. V primeru klasifikacije poskuša SVM ločiti ciljne razrede s kar se da večjo razdaljo. Pri regresiji, pa poskuša poiskati zvezno funkcijo, za katero velja, da leži večina podatkovnih točk v območju dovoljenega odstopanja.

Algoritmi nenadzorovanega učenja

Algoritem	Funkcija	Opis
Apriori (AP)	Asociativna pravila	AP izvaja analizo nakupne košarice z ugotavljanjem ponovljivosti izbranega nabora primerov v zbirki podatkov.
k-Means (KM)	Razvrščanje v gruče	KM je algoritem, ki deluje na osnovi <i>ugotavljanja razdalje</i> , ki razdeli podatke v določeno število gruč. Vsaka gruča ima svoj <i>centroid</i> , v njej pa so <i>primeri</i> , ki so najbližje temu <i>centroidu</i> .
Non-negative Matrix Factorisation (NMF)	Izločanje lastnosti	NMF generira nov atribut z uporabo linearne kombinacije originalnih atributov. Na ta način preslika NMF model orginalne podatke v nov nabor atributov, ki jih je odkril model.
One-class Support Vector Machine (One-class SVM)	Ugotavljanje anomalij v podatkih	One-class SVM zgradi profil na osnovi enega razreda in označi vrednosti, ki odstopajo od samega profila. To omogoča detekcijo redkih primerov, ki niso nujno medsebojno povezani.
Orthogonal Partitioning Clustering (O-Cluster oz. OC)	Razvrščanje v gruče	OC kreira hierarhičen model razvrščanja v gruče. Algoritem kreira gruče, ki določajo področja na osnovi gostote porazdelitve vrednosti atributov. Parameter občutljivosti določa osnovo za stopnjo gostote.

Oracle Data Mining

Priprava podatkov

Priprava podatkov (1)

- Priprava podatkov je ključna faza v procesu podatkovnega rudarjenja
- Podatki morajo biti ustrezno prečiščeni, da se odstranijo morebitne nekonsistentnosti in zadostijo zahteve algoritma podatkovnega rudarjenja
- Praviloma vsi algoritmi zahtevajo določeno stopnjo transformacije podatkov (npr. normalizacija, razvrščanje v diskretne razrede, ipd.)
- Odvisno od funkcije podatkovnega rudarjenja je potrebno pripraviti več naborov podatkov, ki se uporabljajo za izgradnjo in testiranje modela. Vsak od teh naborov mora biti pripravljen na enak način.

Priprava podatkov (2)

- Podatkovno rudarjenje zahteva uporabo le ene „flat“ tabele ali viewja.
- Oracle Data Mining podpira dimenzijske podatkovne modele (npr. zvezdna shema) z uporabo vgnezdenih podatkov (Nested Data)
 - Algoritmi, ki podpirajo vgnezdene podatke: Apriori, GLM, *k*-Means, MDL, Naive Bayes, NMF, SVM
- Oracle podpira podatkovno rudarjenje v nestrukturiranih podatkih (npr. tekst)

Priprava podatkov (3)

- Avtomatična transformacija podatkov
 - V primeru uporabe ADP (Automatic Data Preparation) model sam poskrbi za transformacijo podatkov glede na zahteve algoritma.
- Vključena transformacija podatkov
 - Namesto avtomatične transformacije podatkov lahko izberemo možnost, da vključimo lastne transformacije podatkov, ki se kasneje lahko ponovno uporabijo (npr. transformacijo specificiramo v fazi izgradnje modela, kasneje pa se ista transformacija uporabi ostalih fazah procesa podatkovnega rudarjenja)
- Možnost uporabe tretjih orodij

Kaj pa manjkajoči in neobstoječi podatki?

Podatki	SVM, NMF, <i>k</i> -Means, GLM	NB, MDL, DT, OC	Apriori
Numerični (naključno manjkajoči)	Manjkajoče vrednosti se nadomestijo s povprečnimi vrednostmi	Mankajoče vrednosti se obravajo kot naključno manjkajoče	Vse manjkajoče vrednosti se obravnavajo kot neobstoječe.
Kategorični (naključno manjkajoči)	Manjkajoče vrednosti se nadomestijo s srednjimi vrednostmi	Mankajoče vrednosti se obravajo kot naključno manjkajoče	Vse manjkajoče vrednosti se obravnavajo kot neobstoječe.
Numerični (neobstoječi)	Neobstoječe vrednosti se nadomestijo z vrednostjo 0	DT in OC ne podpirata vgnezenih podatkov in ne podpirata neobstoječih podatkov, NB in MDL nadomestita neobstoječe vrednosti z 0	Algoritem zna obravnavati neobstoječe podatke
Kategorični (neobstoječi)	Neobstoječe vrednosti se nadomestijo z ničelnim vektorjem	DT in OC ne podpirata vgnezenih podatkov in ne podpirata neobstoječih podatkov, NB in MDL nadomestita neobstoječe vrednosti s posebno vrednostjo DM\$SPARSE	Algoritem zna obravnavati neobstoječe podatke

Kaj potrebujemo v Oracle okolju za podatkovno rudarjenje?

- Oracle Database 11g Enterprise Edition + Oracle Data Mining (ODM) option
- Oracle Data Miner
 - vključen v SQL Developer 3.0
- opcija: Oracle Spreadsheet Add-in for Predictive Analytics

V bazi so PL/SQL package-i ...

- DBMS_DATA_MINING
- DBMS_DATA_MINING_TRANSFORM
- DBMS_PREDICTIVE_ANALYTICS

... vse je PL/SQL ...

- Primer kreiranja SVM klasifikacijskega modela

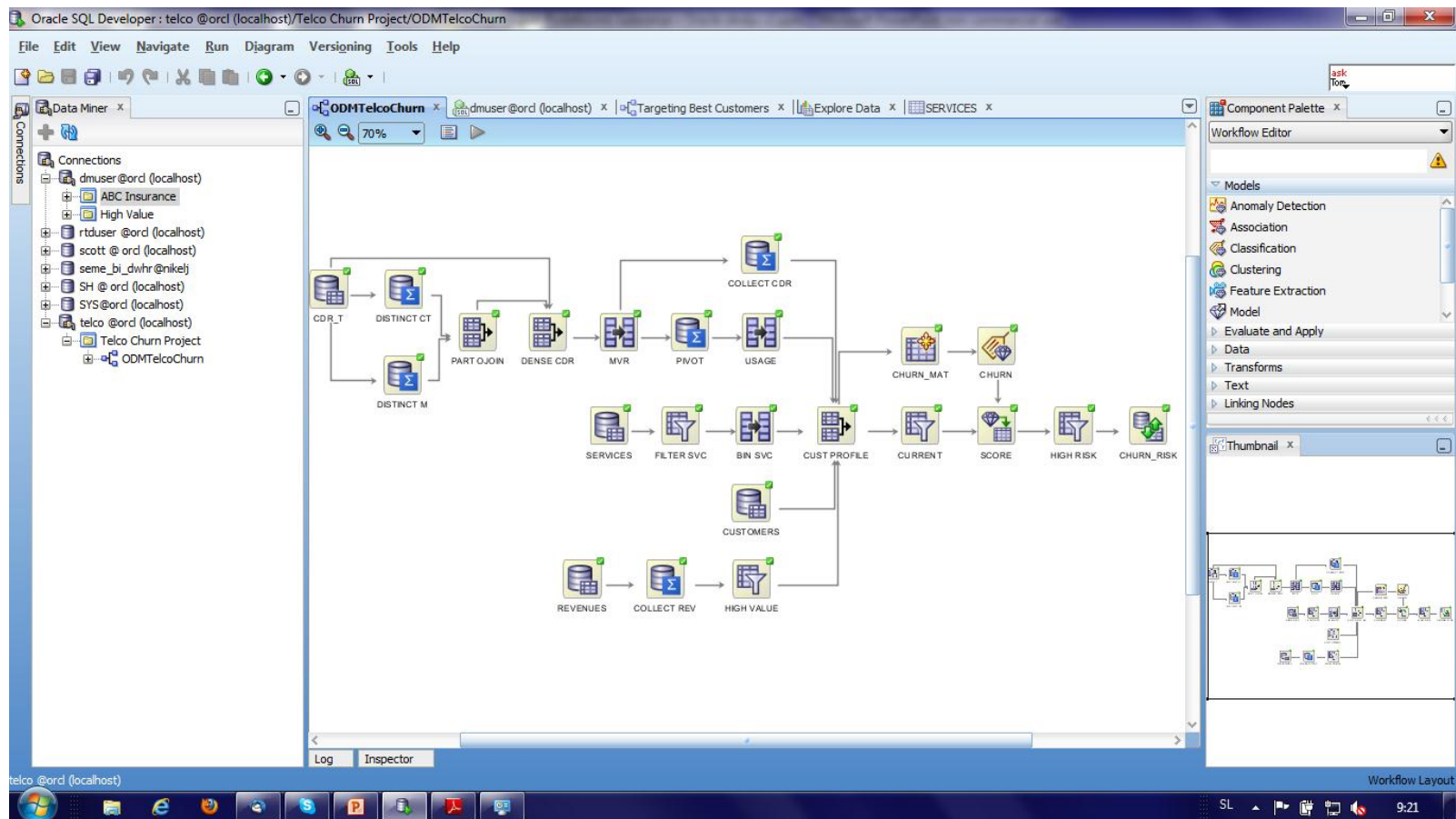
```
CREATE TABLE nastavitev_za_moj_model(  
  setting_name VARCHAR2(30),  
  setting_value VARCHAR2(4000));  
BEGIN  
  INSERT INTO nastavitev_za_moj_model  
    (dbms_data_mining.algo_name,  
     dbms_data_mining.algo_support_vector_machines);  
  COMMIT;  
END;  
  
BEGIN  
  DBMS_DATA_MINING.CREATE_MODEL(  
    model_name => 'moj_model',  
    mining_function => dbms_data_mining.classification,  
    data_table_name => 'mining_data_build',  
    case_id_column_name => 'kupec_id',  
    target_column_name => 'kartica_zvestobe',  
    settings_table_name => 'nastavitev_za_moj_model');  
END;
```

... vse je PL/SQL ...

- Izpis verjetnosti, da bo kupec npr. naročil izdelek z uporabo NB klasifikacije

```
SELECT kupec_id, prob  
FROM (SELECT kupec_id,  
      PREDICTION_PROBABILITY (nb_sh_clas_sample, 1 USING *) prob  
      FROM mining_data_apply_v  
WHERE kupec_id < 100011)  
ORDER BY kupec_id;
```

Oracle Data Miner

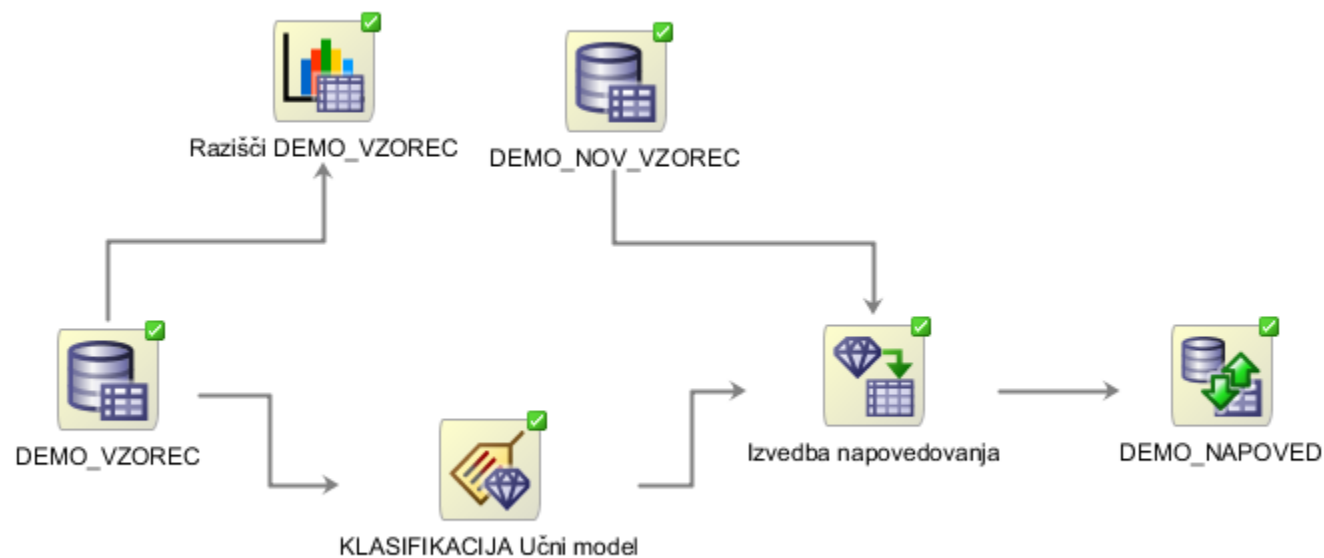


Demo

Poslovni problem

- Katerim potencialnim kupcem naj pošljemo novo ponudbo za zavarovanje, pri čemer imamo na voljo informacije o kupcih, ki so se že oz. se še niso odločili za nakup zavarovanja?

Model - klasifikacija



Določitev *Target* in *Case ID* atributov

Edit Classification

Target: BUY_INSURANCE

Case ID: CUSTOMER_ID

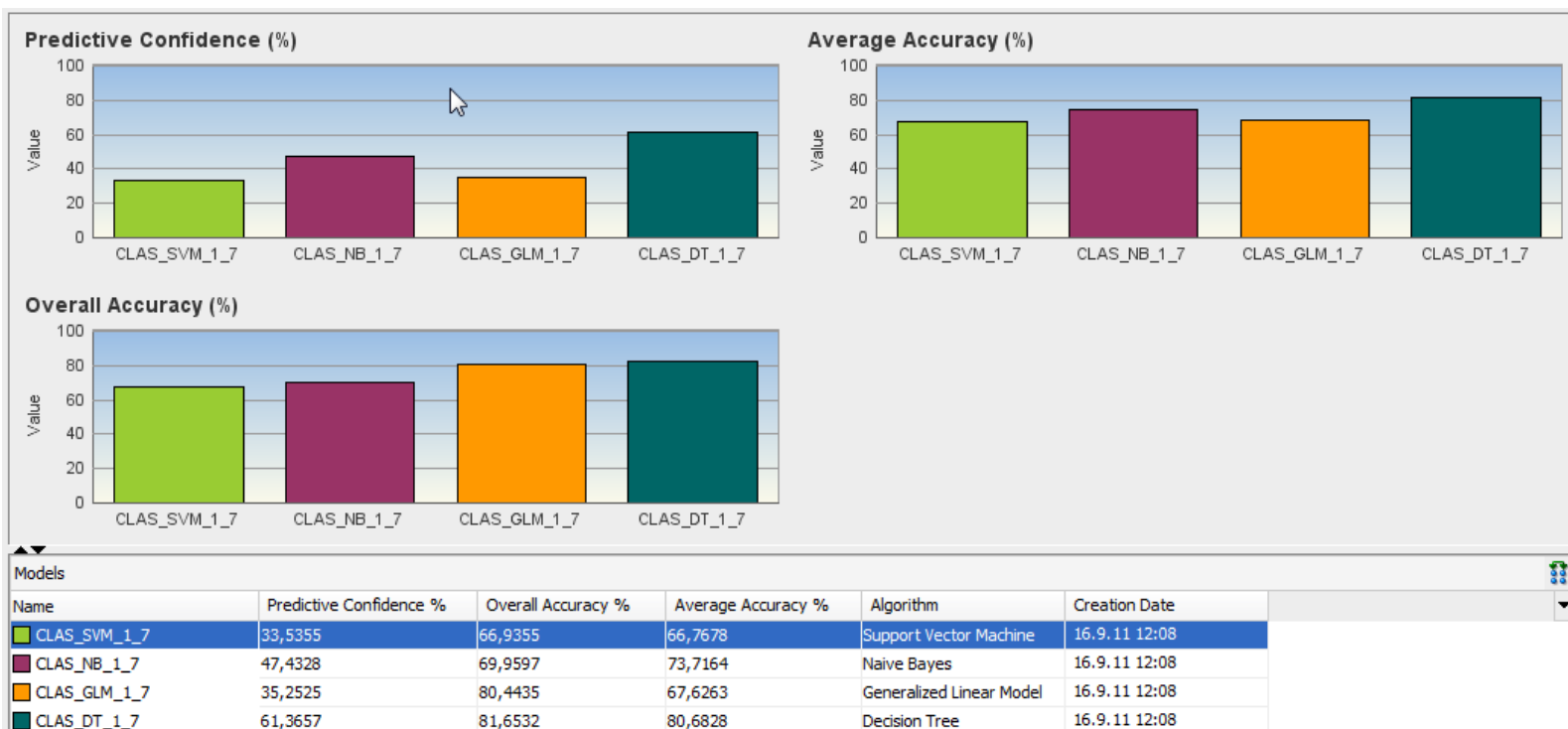
Model Settings

Name	Algorithm	Date	Auto
CLAS_GLM_1_7	Generalized Linear Model	16.9.11 12:08	☒
CLAS_SVM_1_7	Support Vector Machine	16.9.11 12:08	☒
CLAS_DT_1_7	Decision Tree	16.9.11 12:08	☒
CLAS_NB_1_7	Naive Bayes	16.9.11 12:08	☒

Advanced...

Help OK Cancel

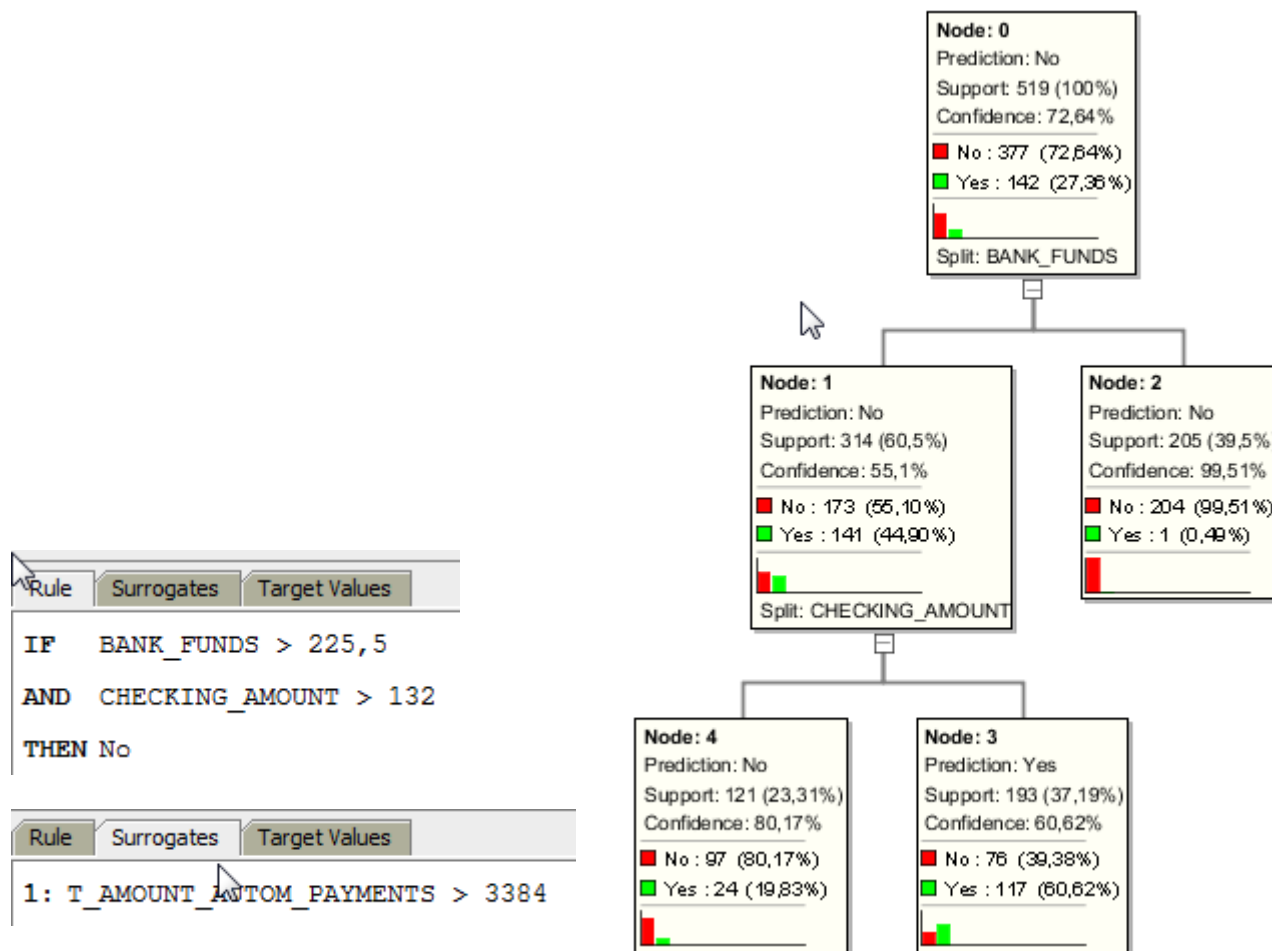
Medsebojna primerjava vseh štirih algoritmov



BRUNO

ORACLE® Gold Partner
Specialized
Oracle Business Intelligence
Foundation

Odločitveno drevo



Rule Surrogates Target Values

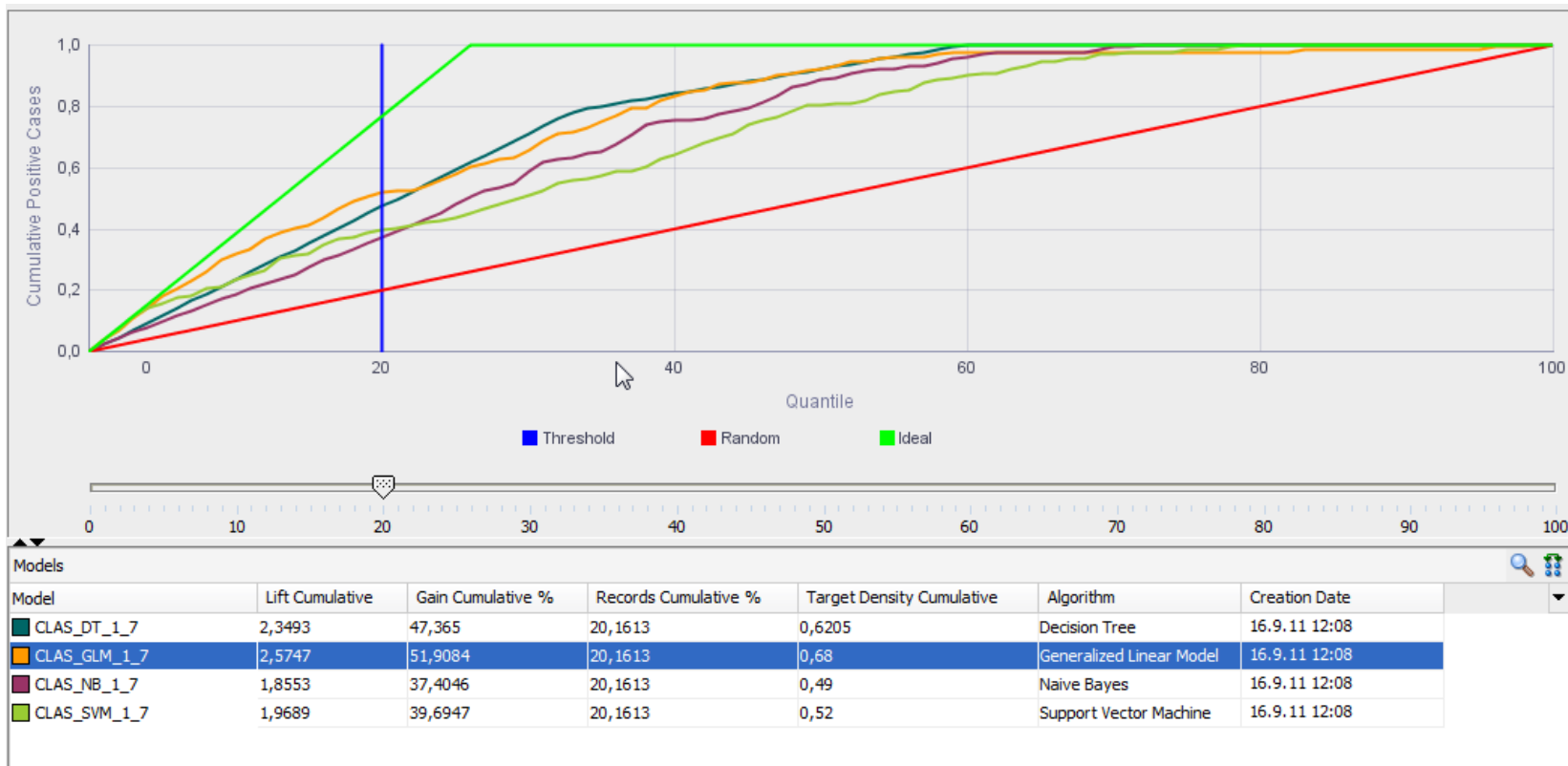
IF BANK_FUNDS > 225,5
AND CHECKING_AMOUNT > 132
THEN No

Rule Surrogates Target Values

1: T_AMOUNT_AUTOM_PAYMENTS > 3384

Value	Record Count	% Percentage
No	97	80,16%
Yes	24	19,83%

Lift



Rezultati v OBIEE nadzorni plošči



* Napoved

☐ (All Column Values)

☒ Yes

☐ No

* Verjetnost napovedi

is greater than or equal to

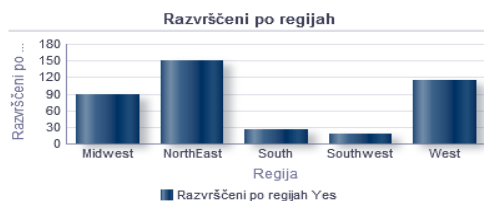
0 100

85.45

Apply Reset

Poročilo o pripravi podatkov za marketinško kampanijo

Kdo bo sprejel ponudbo za zavarovanje?



PREDICTION	Razvrščeni
Yes	399
Grand Total	399

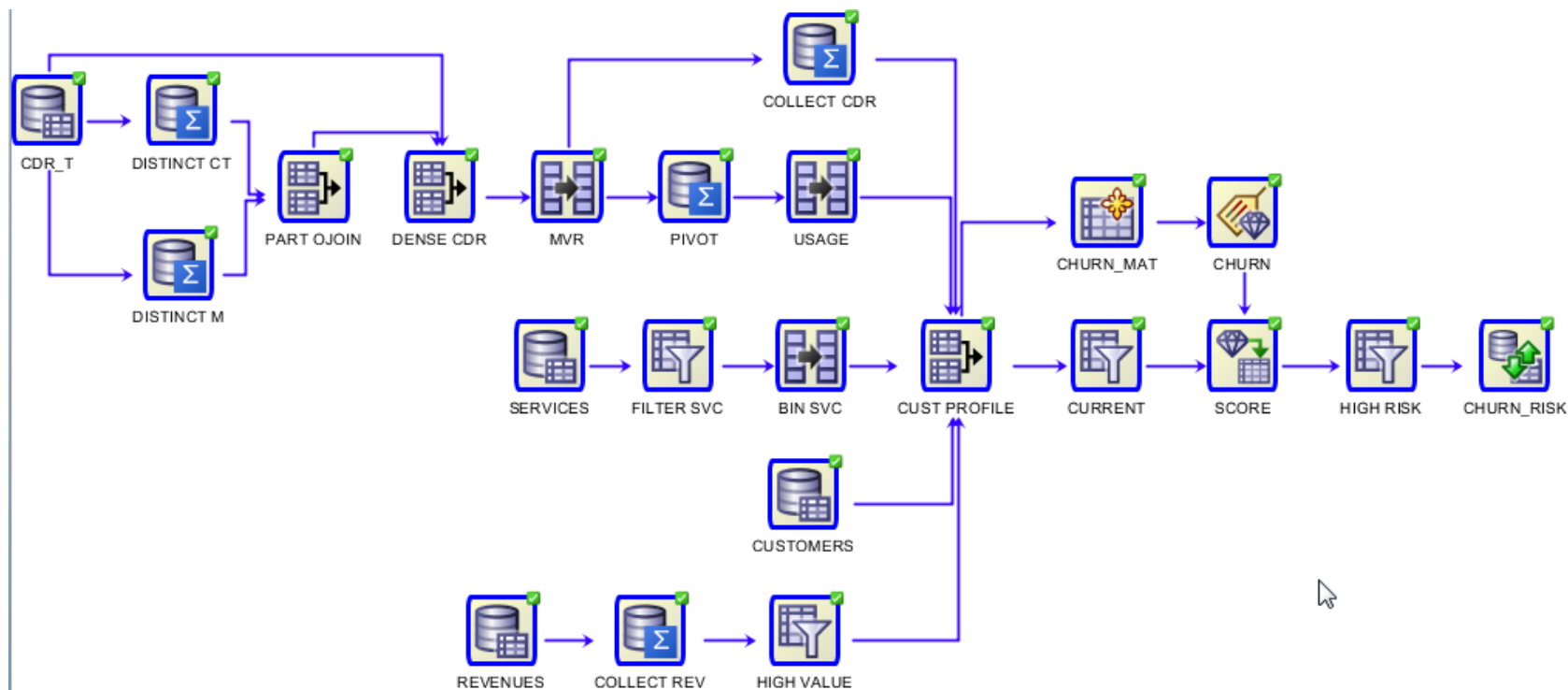


Razvrščeni

Yes

PREDICTION Yes				
CUSTOMER_ID	FIRST	LAST	STATE	PROBABILITY ~ 100
NC0361	MCLAIN	FRANSISCA	CA	99.7937%
NC0816	COMBS	RODOLFO	TX	99.7937%
NC0862	CARON	SEAN	NY	99.7937%
NC0002	ZIEGLER	ABE	OK	99.7639%
NC0779	CROW	RAY	CA	99.7639%
NC0803	CORDOVA	RISA	MI	99.7639%
NC0826	CLOUTIER	ROSARIO	UT	99.7639%
NC0915	BRANCH	TEDDY	CA	99.7639%
NC0027	WHEELER	ALIZA	CA	99.5700%
NC0061	VICK	ARLA	MI	99.5700%
NC0084	TRIPLETT	BECKI	CA	99.5700%
NC0111	STEVENSON	BRET	NY	99.5700%
NC0114	STEPHENSON	BRICE	MI	99.5700%
NC0188	ROMERO	CORDELIA	IL	99.5700%
NC0200	RITCHIE	DAN	CA	99.5700%
NC0235	PRYOR	DEBBRA	NY	99.5700%
NC0242	PRADO	DEON	CA	99.5700%
NC0256	PERKINS	DIRK	MI	99.5700%
NC0280	PADGETT	DREW	MI	99.5700%
NC0323	MOSHER	EMMALINE	CA	99.5700%
NC0342	METZGER	FAYE	MI	99.5700%
NC0352	MCRAE	FLOYD	NY	99.5700%
NC0366	MCKEE	FREDRIC	NY	99.5700%
NC0374	MCCRAY	GAYLORD	NY	99.5700%
NC0396	MARINO	GRANVILLE	NY	99.5700%

Pa še malo bolj kompleksen primer



Smo našli zlato zrno?



Več informacij

- Oracle Data Mining Concepts
- Oracle Data Mining Application Developer Guide
- Oracle spletne strani
 - <http://www.oracle.com/us/products/database/options/data-mining/index.html>
 - <http://www.oracle.com/technetwork/database/options/odm/index.html>
- Mining a Star Schema - Telco Churn Case Study

BRUN

družba za informacijsko tehnologijo

