Nove mogućnosti rudarenja

Oracle Data Mining – novo u Oracle Developeru 4.0 i Oracle Database 12c

Branko Radovanović

Krešimir Bokulić

Sadržaj

- Uvod
  - Kratka povijest Oracle Data Mininga
- Novo u Oracle Developeru 4.0
- Novo u Oracle Database 12c
- Zaključak

Povijest Oracle Data Mininga

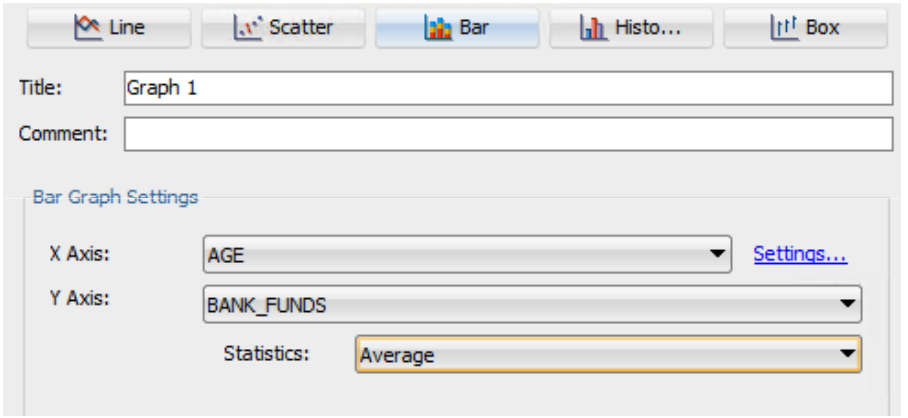
- Preteča ODM-a je analitički alat Darwin (Thinking Machines, u vlasništvu Oraclea od 1999.)
- ODM razvijen od nule po uzoru na Darwin i uveden u Oracle 9i R2 (2002.)
- Novi workflow GUI u 11gR2 (2009.) – integracija s SQL Developerom
- Dio Advanced Analytics opcije Oracle baze, uz Oracle R Enteprise (2012.)

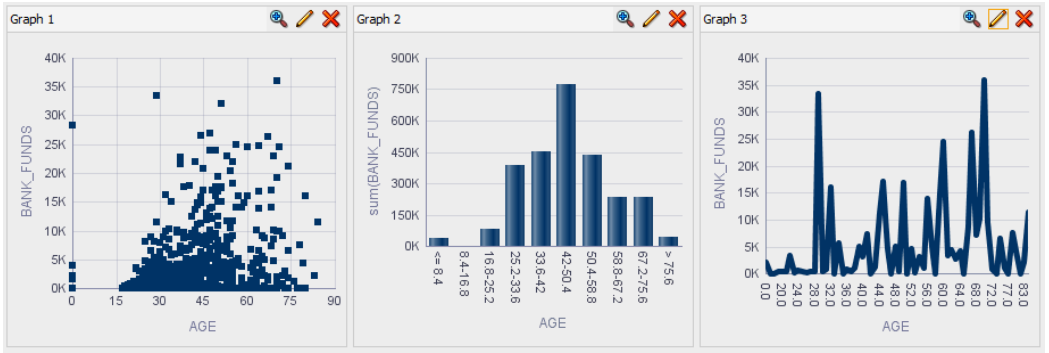
NOVO U ORACLE DEVELOPERU 4.0

multicom
INFORMACIJSKI SUSTAVI

Graph Node

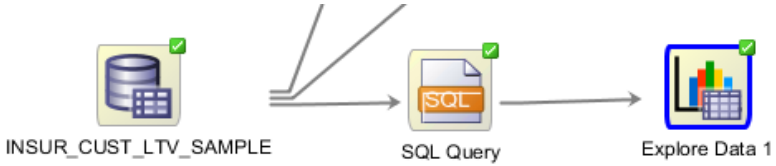
- Nova vrsta WF čvora
- Grafički prikazi: Line, Scatter, Bar, Box, Histogram
- Bira se X i Y os i agregacija odnosno binning





SQL Query Node

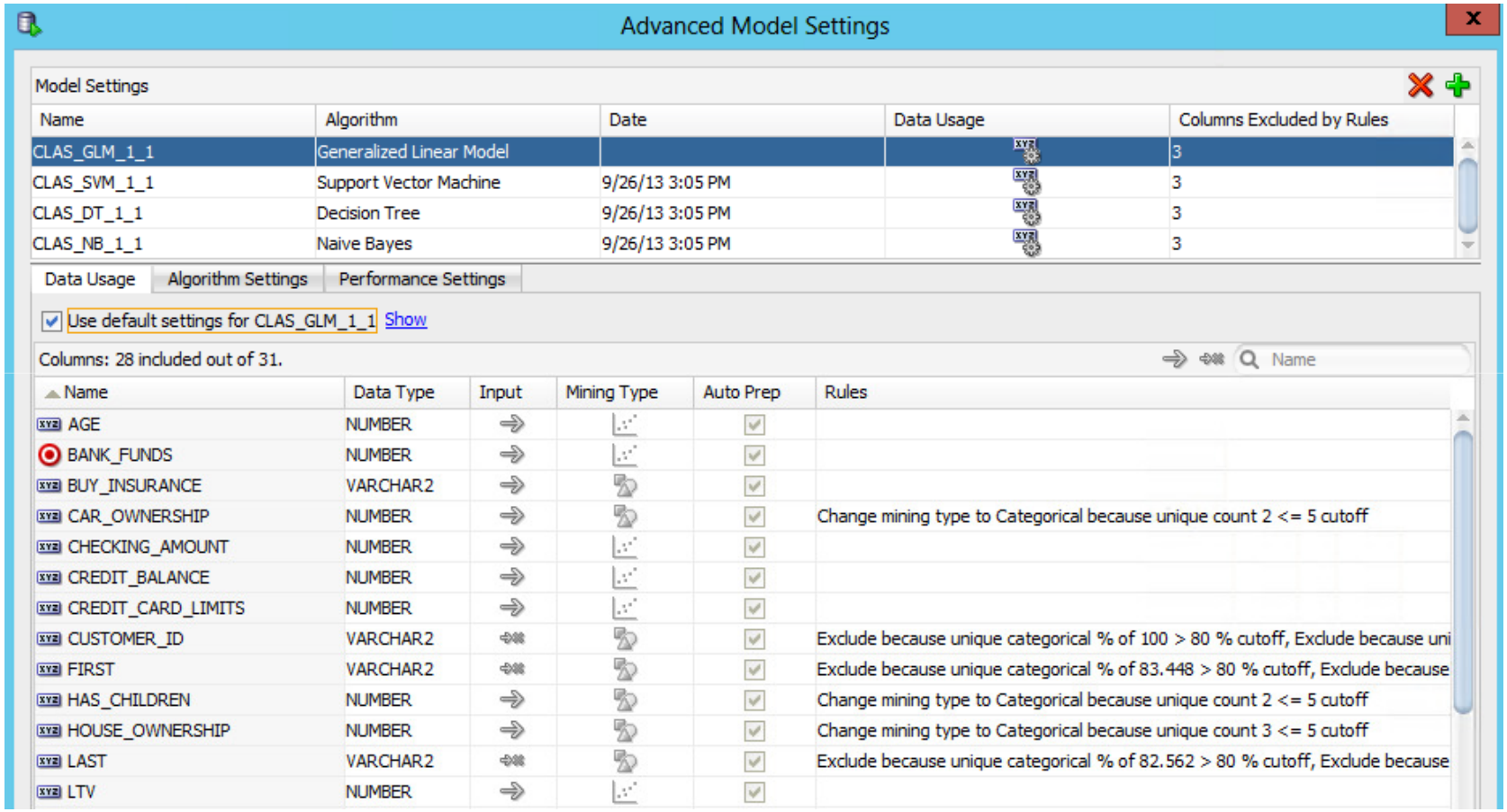
- Nova vrsta WF čvora
- Podržava SQL transformacije
- Podržava R skripte
  - Data Mining i drugi algoritmi iz R libraryja
  - Gotove R skripte
  - Razne transformacije
- Prikaz rezultata unutar samog ODM-a




```
select "INSUR_CUST_LTV_SAMPLE_N$10001"."AGE",
       "INSUR_CUST_LTV_SAMPLE_N$10001"."BUY_INSURANCE",
       sum("INSUR_CUST_LTV_SAMPLE_N$10001"."T_AMOUNT_AUTOM_PAYMENTS") as T_AMOUNT_AUTOM_PAYMENTS
       from INSUR_CUST_LTV_SAMPLE_N$10001
       group by "INSUR_CUST_LTV_SAMPLE_N$10001"."AGE",
       "INSUR_CUST_LTV_SAMPLE_N$10001"."BUY_INSURANCE"
```

Name
RQ$FITDISTR
RQ$getRversion
RQ$installed.packages
RQ$packageVersion
RQ$R.Version
RQG$boxplot
RQG$cdplot
RQG$hist
RQG$matplot
RQG$pairs
RQG$plot1d
RQG$plot2d
RQG$smoothScatter

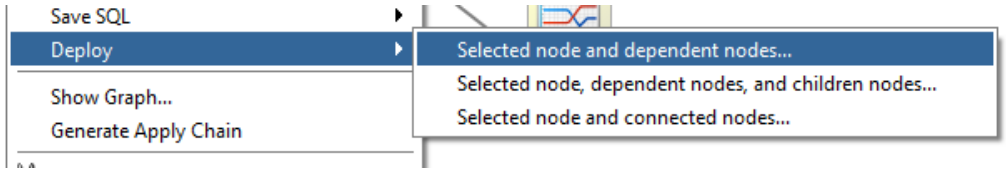
Model Build Node



Dodan prikaz heuristike nad atributima

Workflow SQL Script Deployment

- Workflow ili neki njegov dio se sada mogu dobiti kao PL/SQL skripte



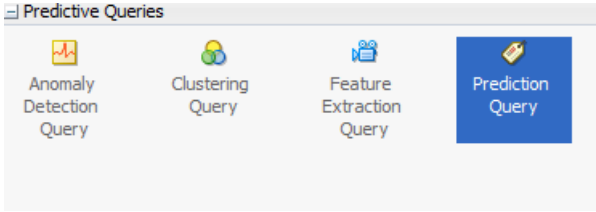
INSUR_CUST_LTV_SAMPLE.sql

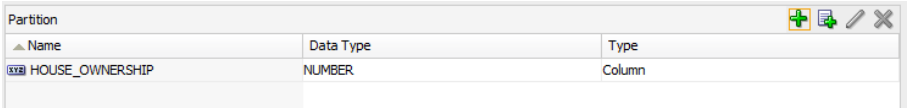
```
33      ignoreid := DBMS_SQL.EXECUTE(curid);
34      DBMS_SQL.CLOSE_CURSOR(curid);
35    EXCEPTION WHEN OTHERS THEN
36      IF DBMS_SQL.IS_OPEN(curid) THEN
37        DBMS_SQL.CLOSE_CURSOR(curid);
38      END IF;
39      RAISE;
40    END;
41
42    FUNCTION formatErrorStack(
43      p_node_name IN VARCHAR2,
44      p_sqlerr        IN VARCHAR2,
45      p_error_stack   IN VARCHAR2 ) RETURN VARCHAR2
46    IS
47    BEGIN
48      RETURN SUBSTR('Error in ' || p_node_name || ': ' || CHR(13) || CHR(10
49                    CHR(13) || CHR(10) || p_error_stack, 1, 4000);
50    END;
51
52  BEGIN
53    v_view_name := generateUniqueName;
54
55    v_sql :=
56      'CREATE VIEW '||v_view_name||' AS SELECT * FROM
57      (
58        select /*+ inline */ "INSUR_CUST_LTV_SAMPLE"."MARITAL_STATUS",
```

NOVO U ORACLE DATABASE 12C

Predictive Query Nodes

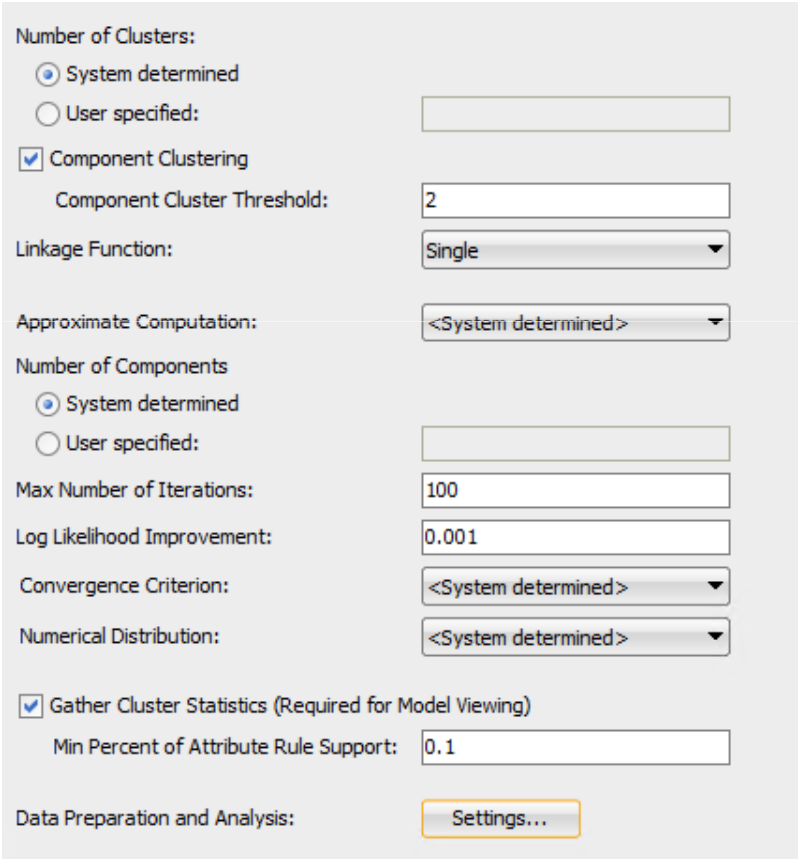
- “Privremeni” rezultati, bez kreiranja modela i bez evaluacije
- Partition opcija – generiranje zasebnih rezultata prema vrijednosti atributa
- Korisno u fazi analize podataka





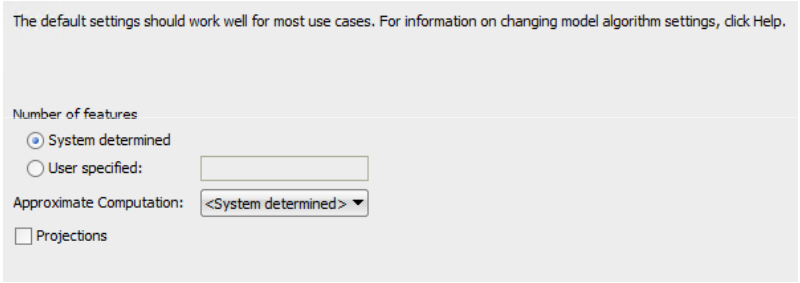
Clustering Node – Expectation Maximization

- Dosadašnji algoritmi za segmentaciju:
  - K-Means
  - O-Cluster
- Novo: Expectation Maximization
- U odnosu na K-Means:
  - Složeniji algoritam s više parametara
  - Kreira i segmente nejednake veličine



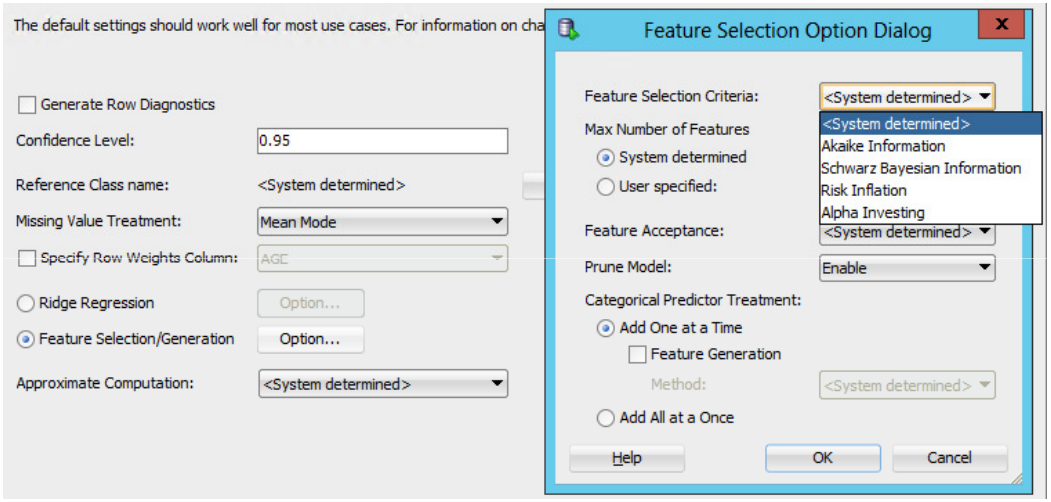
Feature Extraction Node – SVD+PCA

- Dosadašnji algoritmi:
  - Non-negative Matrix Factorization
- Novo:
  - Singular Value Decomposition + Principal Component Analysis
- Robustan algoritam koji zahtijeva razumijevanje podataka



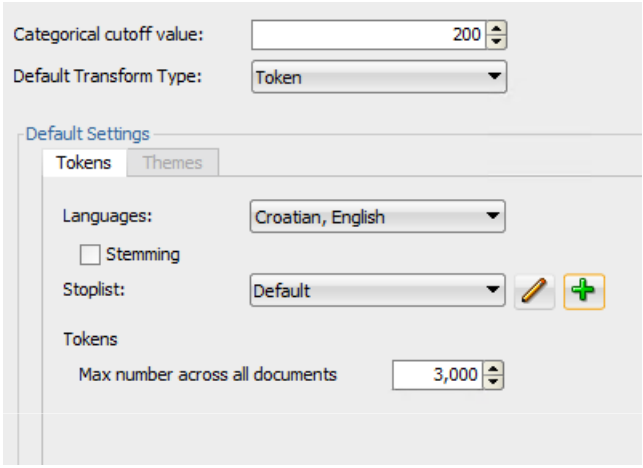
GLM i Feature Selection

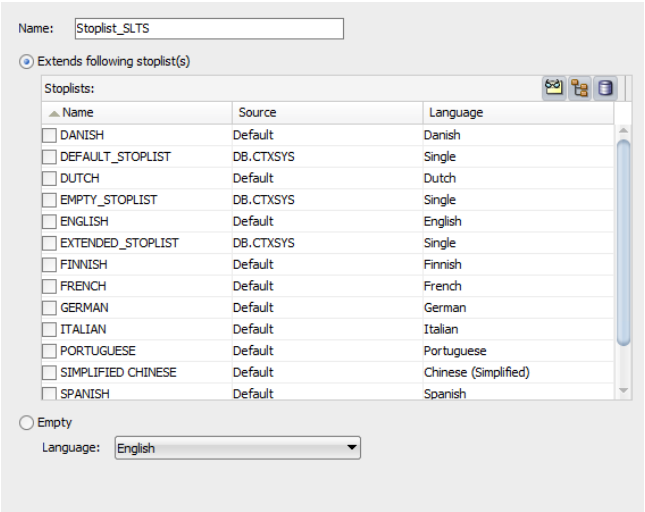
- Generalized Linear Model – Nove postavke algoritma vezane za Feature Selection



Text Mining

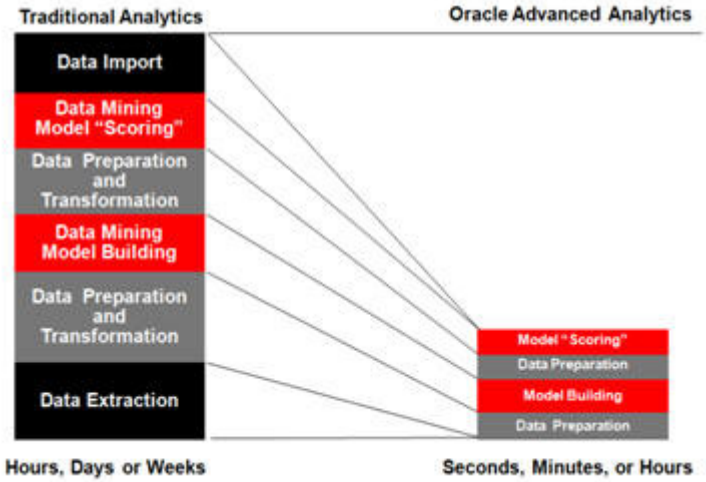
- Tekst se, kao i dosad, može koristiti kao ulaz za ODM algoritme (klasifikacija, segmentacija, itd.)
- Transformacije teksta su sada dio automatske pripreme podataka





Zaključak

- Olakšavanje i ubrzanje procesa rudarenja – uz integraciju daje ključnu prednost nad konkurentskim alatima
- Novi algoritmi i opcije
- R kao nadgradnja
- Jedan od smjerova u budućnosti: Big Data



Pitanja?

Branko.Radovanovic@multicom.hr
Kresimir.Bokulic@multicom.hr