

# 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

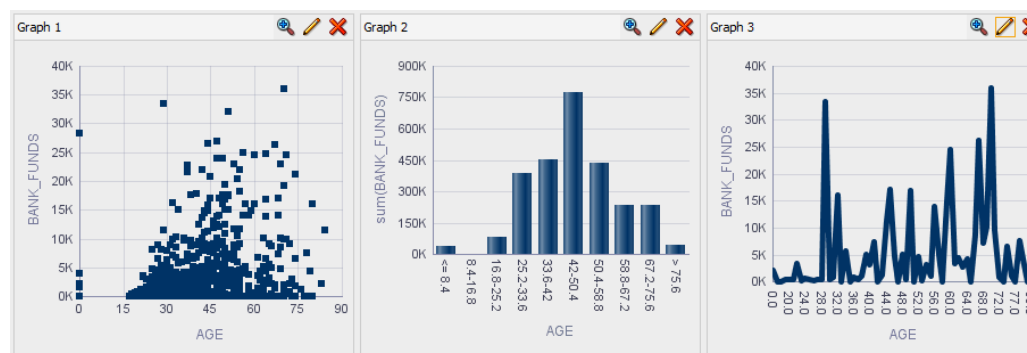
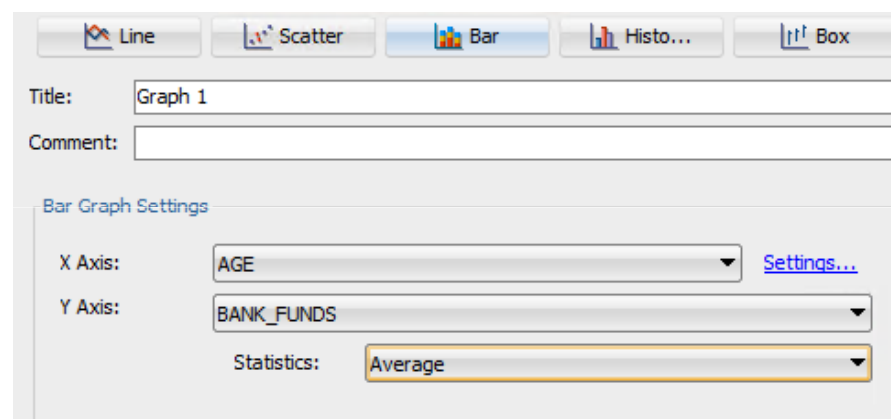
# Povijest Oracle Data Mininga

- 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 Enterprise (2012.)

# NOVO U ORACLE DEVELOPERU 4.0

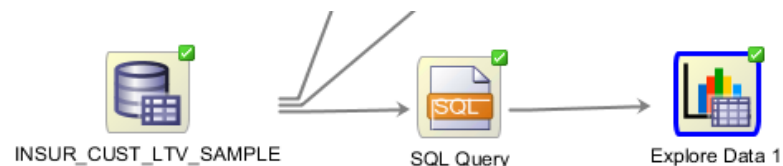
# Graph Node

- 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

Advanced Model Settings

Model Settings

Name	Algorithm	Date	Data Usage	Columns Excluded by Rules
CLAS_GLM_1_1	Generalized Linear Model			3
CLAS_SVM_1_1	Support Vector Machine	9/26/13 3:05 PM		3
CLAS_DT_1_1	Decision Tree	9/26/13 3:05 PM		3
CLAS_NB_1_1	Naive Bayes	9/26/13 3:05 PM		3

Data Usage | Algorithm Settings | Performance Settings

Use default settings for CLAS\_GLM\_1\_1 [Show](#)

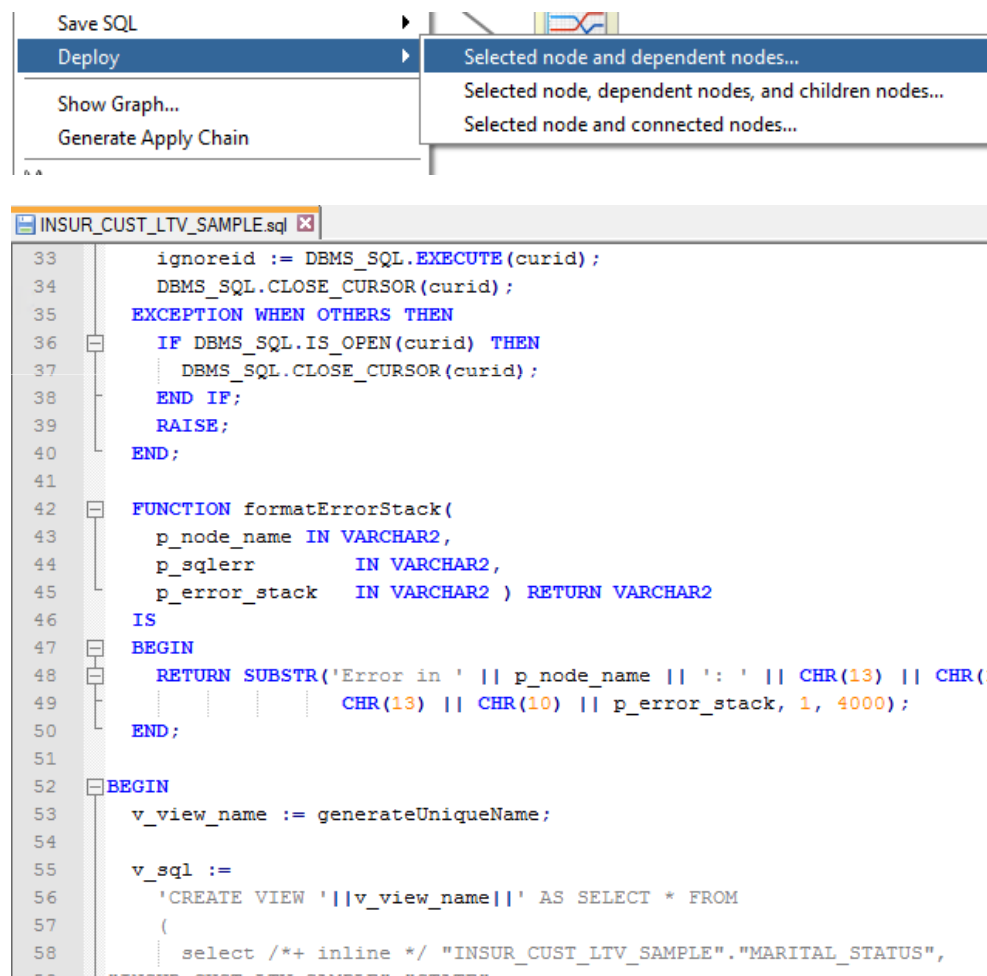
Columns: 28 included out of 31. → 🔍 Name

Name	Data Type	Input	Mining Type	Auto Prep	Rules
AGE	NUMBER	→		<input checked="" type="checkbox"/>	
BANK_FUNDS	NUMBER	→		<input checked="" type="checkbox"/>	
BUY_INSURANCE	VARCHAR2	→		<input checked="" type="checkbox"/>	
CAR_OWNERSHIP	NUMBER	→		<input checked="" type="checkbox"/>	Change mining type to Categorical because unique count 2 <= 5 cutoff
CHECKING_AMOUNT	NUMBER	→		<input checked="" type="checkbox"/>	
CREDIT_BALANCE	NUMBER	→		<input checked="" type="checkbox"/>	
CREDIT_CARD_LIMITS	NUMBER	→		<input checked="" type="checkbox"/>	
CUSTOMER_ID	VARCHAR2	⇄		<input checked="" type="checkbox"/>	Exclude because unique categorical % of 100 > 80 % cutoff, Exclude because uni
FIRST	VARCHAR2	⇄		<input checked="" type="checkbox"/>	Exclude because unique categorical % of 83.448 > 80 % cutoff, Exclude because
HAS_CHILDREN	NUMBER	→		<input checked="" type="checkbox"/>	Change mining type to Categorical because unique count 2 <= 5 cutoff
HOUSE_OWNERSHIP	NUMBER	→		<input checked="" type="checkbox"/>	Change mining type to Categorical because unique count 3 <= 5 cutoff
LAST	VARCHAR2	⇄		<input checked="" type="checkbox"/>	Exclude because unique categorical % of 82.562 > 80 % cutoff, Exclude because
LTV	NUMBER	→		<input checked="" type="checkbox"/>	

Dodan prikaz heuristike nad atributima

# Workflow SQL Script Deployment

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



The screenshot shows a software interface with a context menu open over a 'Deploy' button. The menu options are:

- Selected node and dependent nodes...
- Selected node, dependent nodes, and children nodes...
- Selected node and connected nodes...

Below the menu, a code editor window titled 'INSUR\_CUST\_LTV\_SAMPLE.sql' displays the following PL/SQL code:

```

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",
59

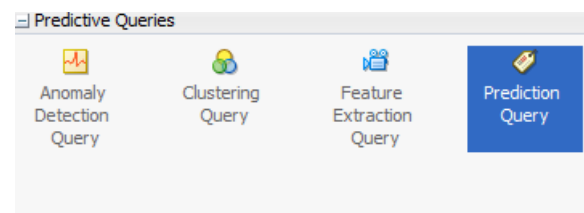
```



# 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



Name	Data Type	Type
HOUSE_OWNERSHIP	NUMBER	Column

# 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

The screenshot shows the configuration interface for the Expectation Maximization clustering node. It includes the following settings:

- Number of Clusters:**
  - System determined
  - User specified:
- Component Clustering
  - Component Cluster Threshold:
- Linkage Function:
- Approximate Computation:
- Number of Components**
  - System determined
  - User specified:
- Max Number of Iterations:
- Log Likelihood Improvement:
- Convergence Criterion:
- Numerical Distribution:
- Gather Cluster Statistics (Required for Model Viewing)
  - Min Percent of Attribute Rule Support:
- Data Preparation and Analysis:

# Feature Extraction Node – SVD+PCA

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

The default settings should work well for most use cases. For information on changing model algorithm settings, click Help.

Number of features

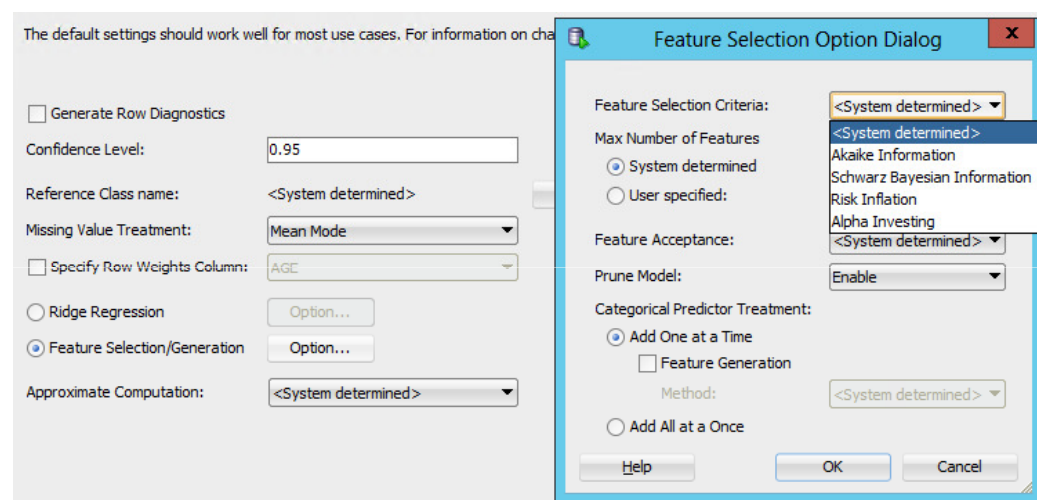
System determined  
 User specified:

Approximate Computation: <System determined> ▼

Projections

# 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

Categorical cutoff value:



Default Transform Type:

Default Settings

Tokens Themes

Languages:

Stemming

Stoplist:   

Tokens

Max number across all documents

Name:

Extends following stoplist(s)

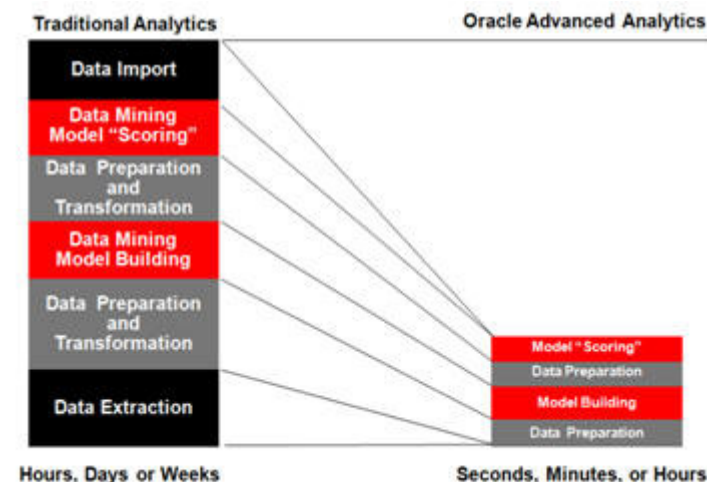
Name	Source	Language
<input type="checkbox"/> DANISH	Default	Danish
<input type="checkbox"/> DEFAULT_STOPLIST	DB.CTXSYS	Single
<input type="checkbox"/> DUTCH	Default	Dutch
<input type="checkbox"/> EMPTY_STOPLIST	DB.CTXSYS	Single
<input type="checkbox"/> ENGLISH	Default	English
<input type="checkbox"/> EXTENDED_STOPLIST	DB.CTXSYS	Single
<input type="checkbox"/> FINNISH	Default	Finnish
<input type="checkbox"/> FRENCH	Default	French
<input type="checkbox"/> GERMAN	Default	German
<input type="checkbox"/> ITALIAN	Default	Italian
<input type="checkbox"/> PORTUGUESE	Default	Portuguese
<input type="checkbox"/> SIMPLIFIED CHINESE	Default	Chinese (Simplified)
<input type="checkbox"/> SPANISH	Default	Spanish

Empty

Language:

# 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](mailto:Branko.Radovanovic@multicom.hr)

[Kresimir.Bokulic@multicom.hr](mailto:Kresimir.Bokulic@multicom.hr)