Zagrebačka banka
UniCredit Group

UEFA CHAMPIONS LEAGUE
UniCredit
official sponsor

Studija slučaja Zagrebačka banka:

Migracija SAS na Oracle Advanced Analytics

Siniša Behin – Zagrebačka banka, BI Development, CRM voditelj tima
Krešimir Bokulić – Multicom, Data Mining voditelj tima

multicom

Listopad 16th, 2015

Agenda

- Razlozi migracije
- Opseg projekta i konačni cilj
- Izazovi tijekom projekta
- Konačni rezultati i sljedeći koraci

Multicom

- Predictive analytics & Data mining
- Master data management
- Data warehouse and business inteligence
- Data Quality Management
- Corporate performance mmanagement
- Customer relationship management
- Billing

ORACLE Gold Partner

Deloitte. 2010 Technology Fast 50 First place in 2009. category "Rising Stars"

Cro Cert ISO 9001:2008 certified

SAP Partner

Ciljevi migracije

- Povećanje efikasnosti prilikom izgradnje i testiranja modela
  - Spori proces pred selekcije
  - Razumijevanje načina razvoja (Base, EG, Miner)
- Integracija modela s podacima
  - Jednostavnija arhitektura
  - Jednostavnije održavanje
- Povezivanje procesa bodovanja sa ostalim analitičkim aktivnostima OLTP sustava
- Povećanje fleksibilnosti za razvoj modela korištenjem ORE i R-a
- Smanjenje TCO

Opseg projekta

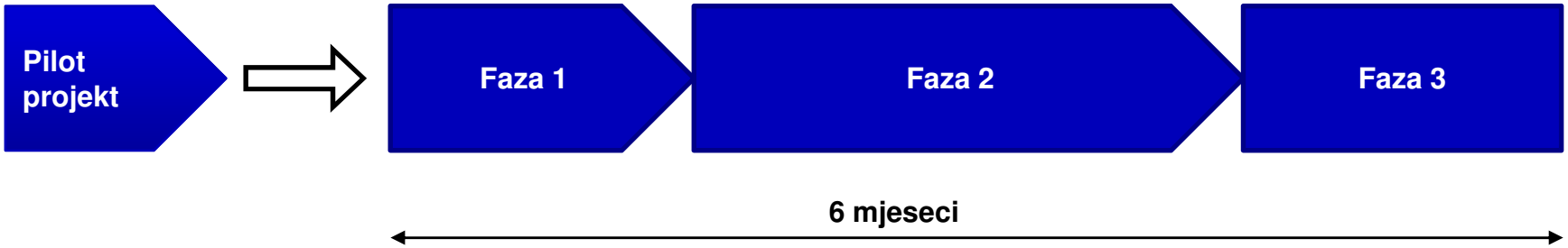
- Analiza opsega
- Obrnuti inženjering procesa razvoja modela u SAS-u
- Gap analiza
- Razvoj komponenti i testiranje
- Implementacija modela rizika
- Testiranje modela rizika
- Primjena modela

Opseg projekta

- Migrirati 12 modela rizika sa SAS-a na Oracle platformu
- Prva migracija takvog tipa u svijetu!

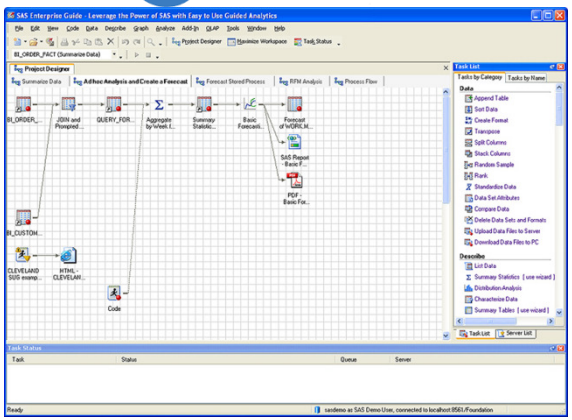
- Glavni zahtjevi:
  - Zadržavanje identičnog procesa razvoja modela kao u SAS-u
  - Dobivanje identičnih modela (brojki) kao i u slučaju modela razvijenih pomoću SAS-a
- Vremenski okvir:
  - 3 faze
  - 6 mjeseci

- Jednomjesečni pilot projekt - prije glavnog projekta

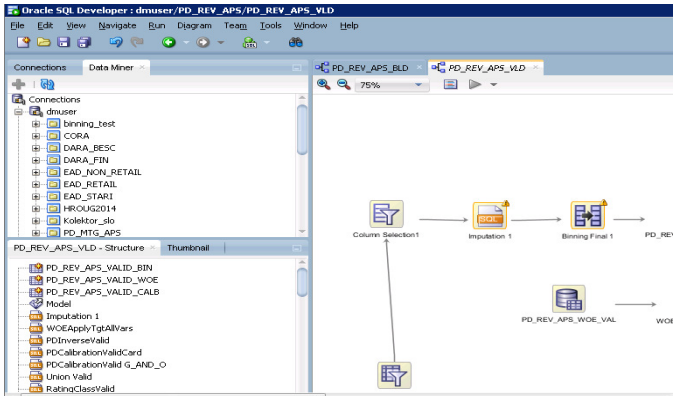


Opseg projekta

- Migrirati = rekreirati modele koristeći iste podate, iste modelarske tehnike i istu metodologiju razvoja modela kao u SAS-u
- Brojevi i predikcije moraju biti identične kao i u slučaju modela kreiranih pomoću SAS-a
- Isti brojevi – zahtjev regulatora

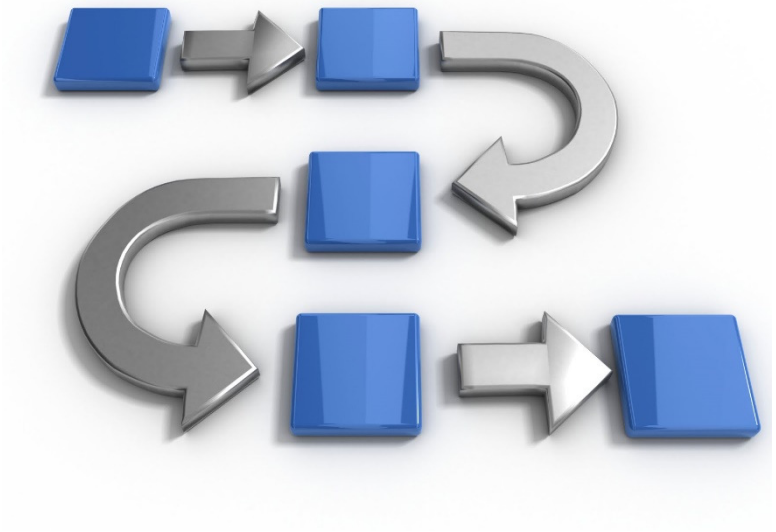


MIGRACIJA



Metodologija razvoja modela

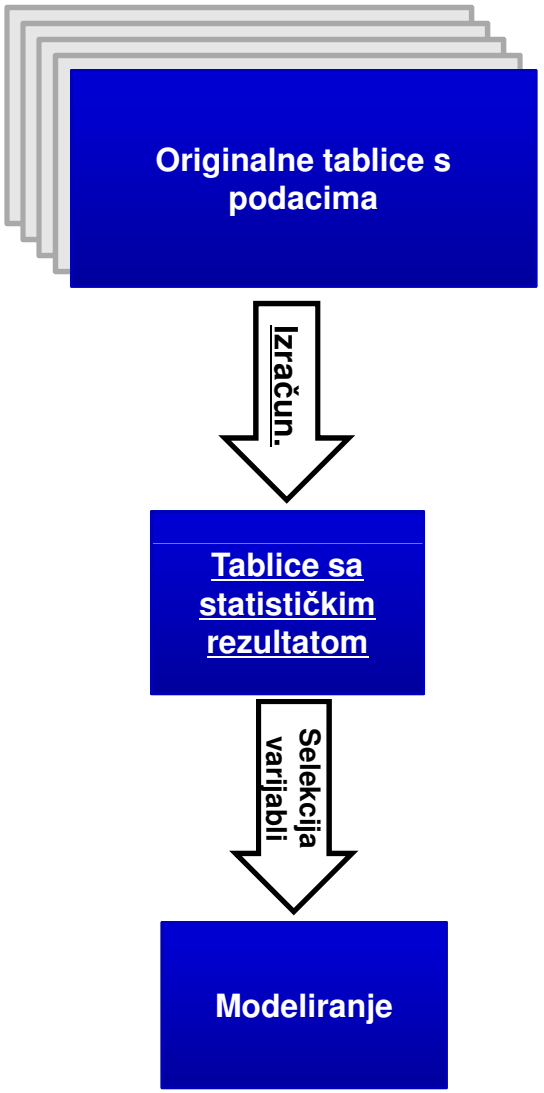
- Predselekcija varijabli
- Klasteriranje varijabli
- Izračun Information value-a (IV)
- Izračun korelacijske matrice
- Kategorizacija (Binning) i izračun weights of evidence vrijednosti (WOE)
- Zamjena vrijednosti sa WOE
- Izgradnja regresije
- Testiranje



Predselekcija varijabli

- Korišteno za predselekciju 7000 varijabli iz ABT-a
- Koristi različite statistike poput:
  - Broja nepostojećih vrijednosti
  - Broja identičnih vrijednosti
  - Pearson and Spearman korelacije
  - Kvantila (1,5, 95,99)
  - Auroc (C-statistike)
  - KS testa
  - T- testa
  - Information value (IV)

- Smanjuje 7000 varijabli na 1000-2000



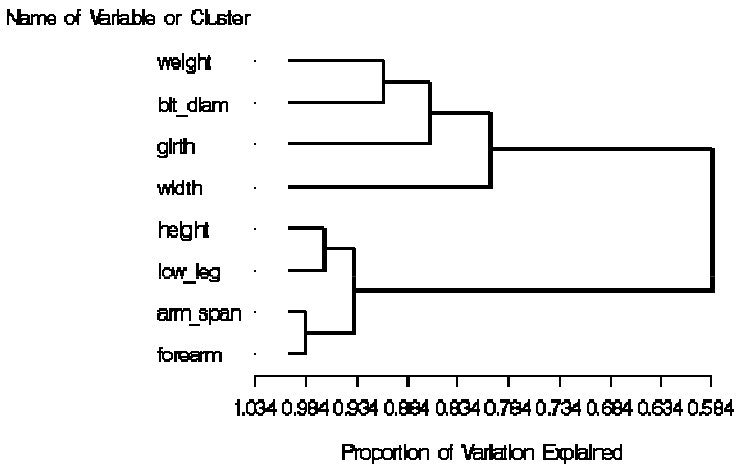
Predselekcija varijabli

- Razvijeno i migrirano na Oracle R Enterprise (ORE)
- Oracle maksimum 1000 varijabli po tablici
- Iteriranje kroz tablice
- Korišten je ORE transperency sloj za ubrzanje
- Rezultati izračuna se spremaju u tablicu na Oracle bazi
- Otpornost na greške obzirom da su rezultati spremani nakon svake varijable
- Modularno – jednostavno je nadodati novu R statistiku za izračun
- Vrijeme potrebno za izračun smanjeno sa 20h na 5h

	TABLE_NAME	VAR	ORE_CLASS	NUM_ROWS	NUM_MISSING	PCT_MISSING	MISSING_THRESHOLD	UNIQUE	MIN	MAX	Q1	Q5	Q50	Q95	Q99	
1	INSUR_CUST_LTV_SAMPLE_1	AGE	ore.numeric	1015.0	21.0	0.020689655172413793	0.1	65.0	(null)	(null)	19.0	21.0	36.0	67.0	77.0	-0
2	INSUR_CUST_LTV_SAMPLE_1	BUY_INSURANCE	ore.factor	1015.0	0.0	0.0	0.1	2.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
3	INSUR_CUST_LTV_SAMPLE_1	CAR_OWNERSHIP	ore.factor	1015.0	0.0	0.0	0.1	2.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
4	INSUR_CUST_LTV_SAMPLE_1	CREDIT_BALANCE	ore.numeric	1015.0	0.0	0.0	0.1	93.0	0.0	170498.0	0.0	0.0	0.0	7095.0	64674.0	-0
5	INSUR_CUST_LTV_SAMPLE_1	CUSTOMER_ID	ore.factor	1015.0	0.0	0.0	0.1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
6	INSUR_CUST_LTV_SAMPLE_1	FIRST	ore.factor	1015.0	0.0	0.0	0.1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
7	INSUR_CUST_LTV_SAMPLE_1	HAS_CHILDREN	ore.factor	1015.0	0.0	0.0	0.1	2.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
8	INSUR_CUST_LTV_SAMPLE_1	HOUSE_OWNERSHIP	ore.factor	1015.0	0.0	0.0	0.1	2.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
9	INSUR_CUST_LTV_SAMPLE_1	LAST	ore.factor	1015.0	0.0	0.0	0.1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
10	INSUR_CUST_LTV_SAMPLE_1	MARITAL_STATUS	ore.factor	1015.0	0.0	0.0	0.1	5.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
11	INSUR_CUST_LTV_SAMPLE_1	N_OF_DEPENDENTS	ore.numeric	1015.0	0.0	0.0	0.1	7.0	0.0	6.0	0.0	0.0	1.0	5.0	6.0	0.
12	INSUR_CUST_LTV_SAMPLE_1	PROFESSION	ore.factor	1015.0	0.0	0.0	0.1	95.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
13	INSUR_CUST_LTV_SAMPLE_1	REGION	ore.factor	1015.0	0.0	0.0	0.1	5.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
14	INSUR_CUST_LTV_SAMPLE_1	SALARY	ore.numeric	1015.0	0.0	0.0	0.1	998.0	37572.0	109943.0	50175.0	56929.0	64173.0	75869.0	90725.0	-0
15	INSUR_CUST_LTV_SAMPLE_1	SEX	ore.factor	1015.0	0.0	0.0	0.1	2.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
16	INSUR_CUST_LTV_SAMPLE_1	STATE	ore.factor	1015.0	0.0	0.0	0.1	22.0	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(n
17	INSUR_CUST_LTV_SAMPLE_1	TIME_AS_CUSTOMER	ore.numeric	1015.0	0.0	0.0	0.1	5.0	1.0	5.0	1.0	1.0	2.0	5.0	5.0	0.
18	INSUR_CUST_LTV_SAMPLE_2	BANK_FUNDS	ore.numeric	1015.0	0.0	0.0	0.1	280.0	0.0	36000.0	0.0	0.0	500.0	14001.0	24500.0	0.

Klasteriranje varijabli

- Korišteno za klasteriranje varijabli obzirom na njihovu sličnost
- Zamjena za matricu korelacija
- VARCLUS procedura u SAS BASE
- Varclass za ORE
- Razvijeno od strane Oracle razvojnog tima specijalno za Zaba projekt
- Rezultati se spremaju u ORE data-store i lokalne datoteke
- Prva verzija u samo 5 dana od zahtijeva
- Korišteno kao zamjena za SAS VARCLUS

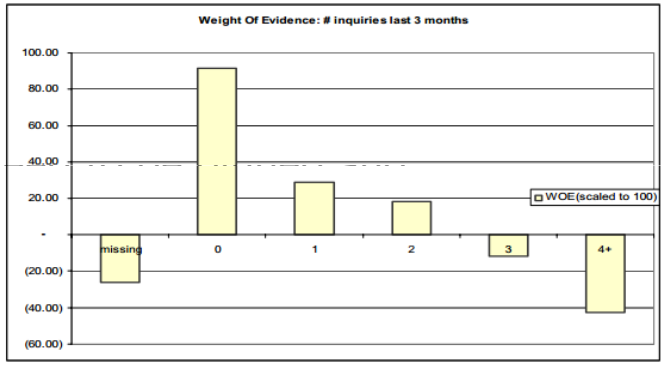


Izračun Information Value-a (IV)

- Korišteno u SAS-u kao mjera prediktivnosti varijable
- Standard u industriji
- Sličnost sa attribute importance-u OAA

$$IV = \sum (DistributionGood_i - DistributionBad_i) \times ln(\frac{DistributionGood_i}{DistributionBad_i})$$

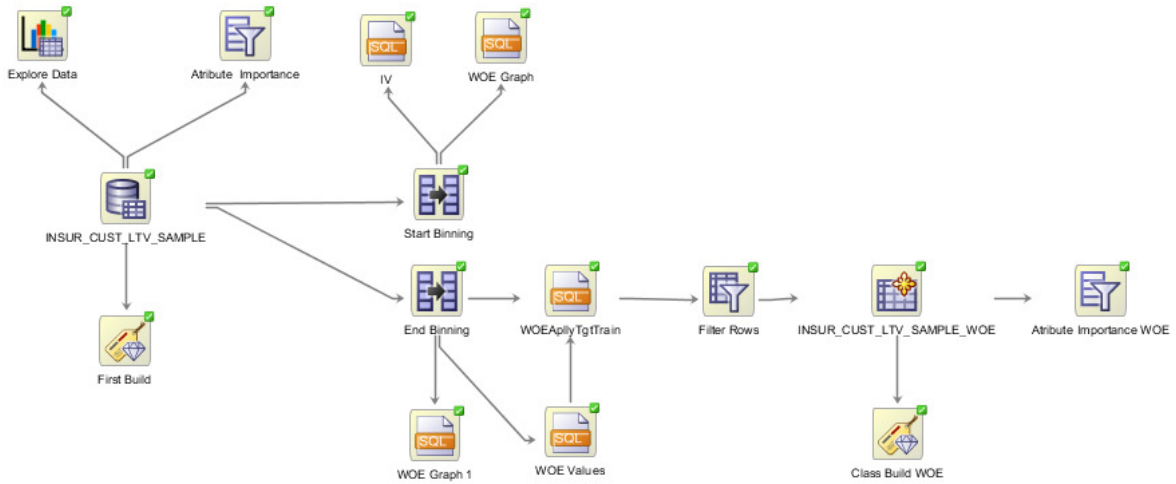
- Za izračun IV-a potrebno je
  - Varijablu (numeričku) potrebno je kategorizirati te izračunati Weight of evidence (WOE) za svaku kategoriju
  - IV se izračunava iz WOE



- Postoji u klaR paketu na CRAN-u, međutim – presporo
- Napravljena verzija koja koristi ORE Transparency sloj za brže računanje IV i WOE vrijednosti
- Spremljeno u Oracle DB te se koristi u ORE-u i Oracle Data miner (ODM)

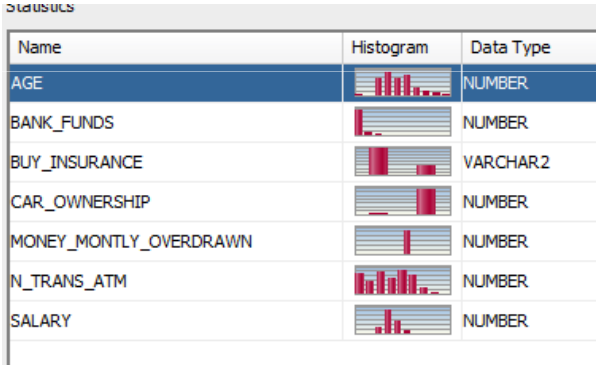
Proces izgradnje modela u ODM-u

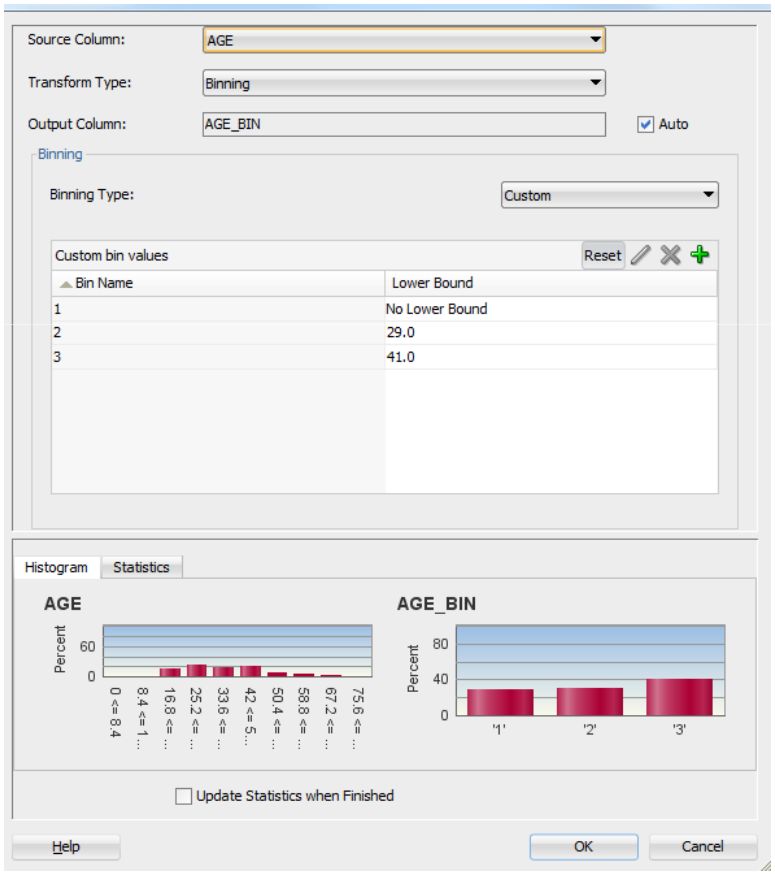
- U završnom procesu izgradnje modela se koristi Oracle Data Miner
- Koraci:
  - Kategorizacija varijabli pomoću Transformation čvora
  - Izračun WOE, IV-a, te WOE grafa
  - Zamjena vrijednosti u varijablama sa WOE vrijednostima iz WOEApplyTGTrain
  - Izgradnja regresijskog modela



IV i WOE te integracija sa ODM-om

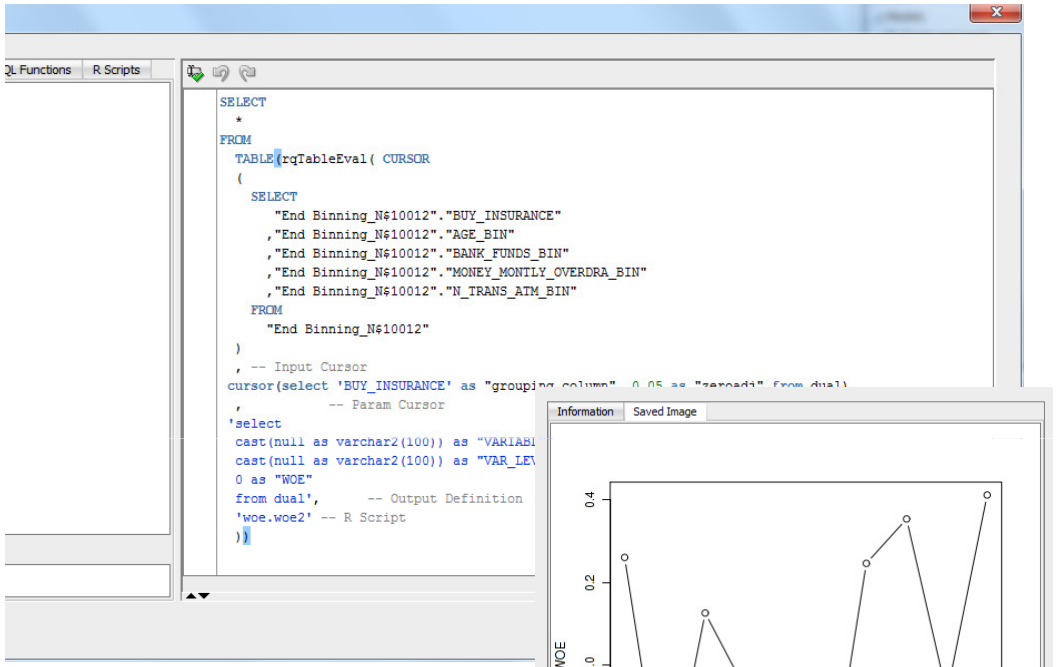
- Primjer na varijabli AGE iz INSUR_CUST_LTV podataka
- Transformacija i kategorizacija numeričke varijable
- Zamjena za IGEN čvor u SAS-u





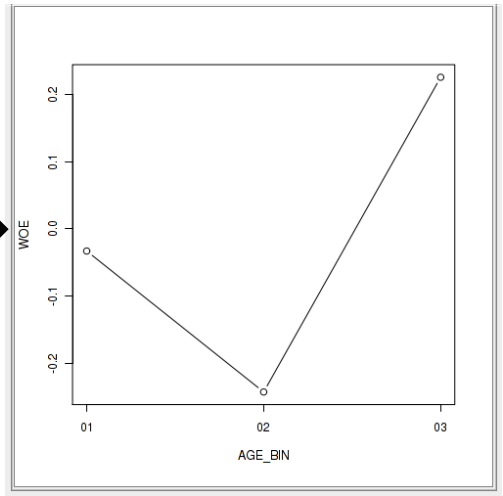
IV i WOE te integracija sa ODM-om

- Izračun WOE vrijednosti i WOE grafa pomoću SQL čvora i skripte spremljene u bazi



Varijabla sa 10 kategorija

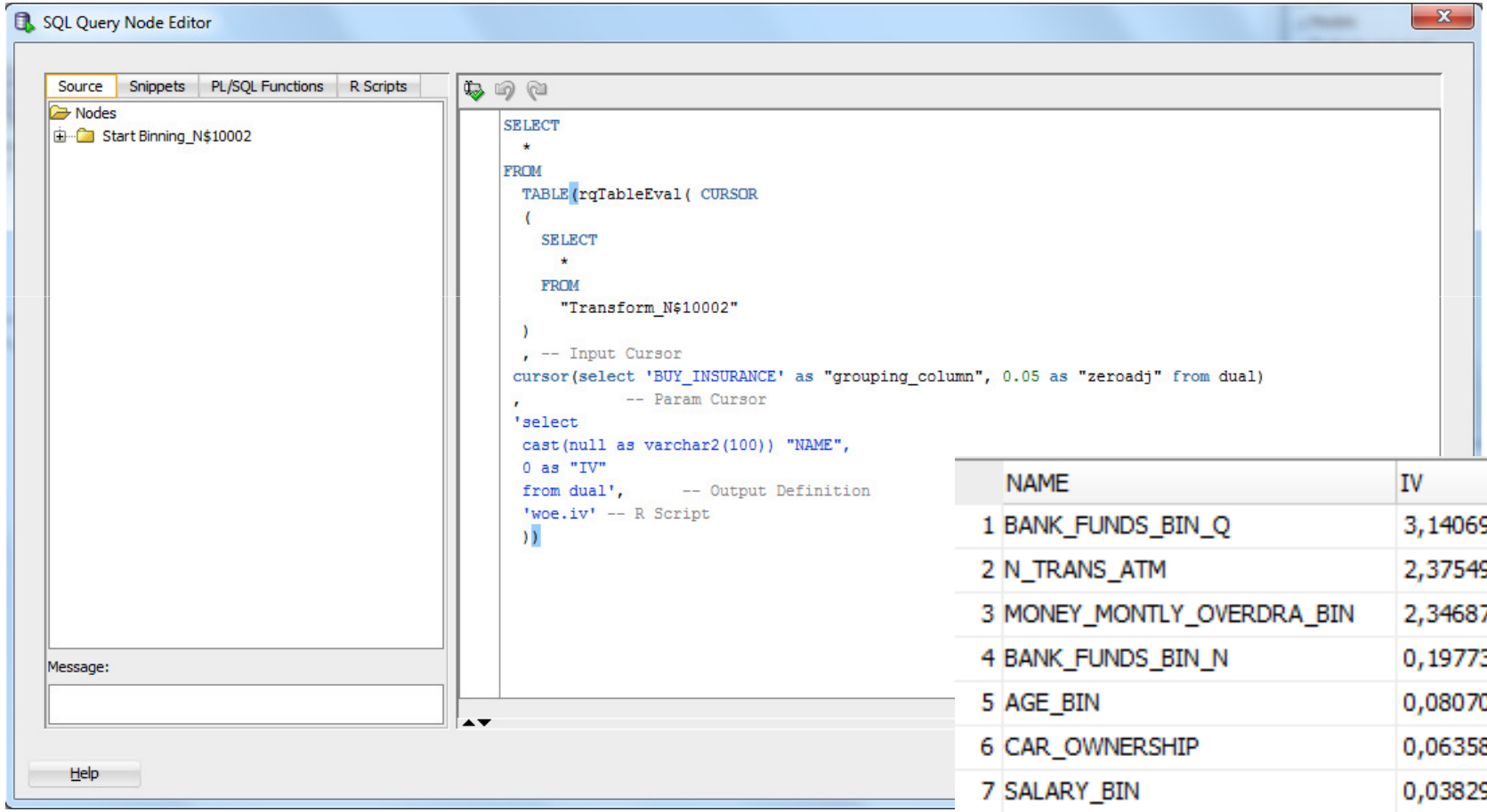
Ideja: Kategorizirati varijable tako da kategorizacija ima poslovni smisao



Varijabla sa 3 kategorije

IV i WOE te integracija sa ODM-om

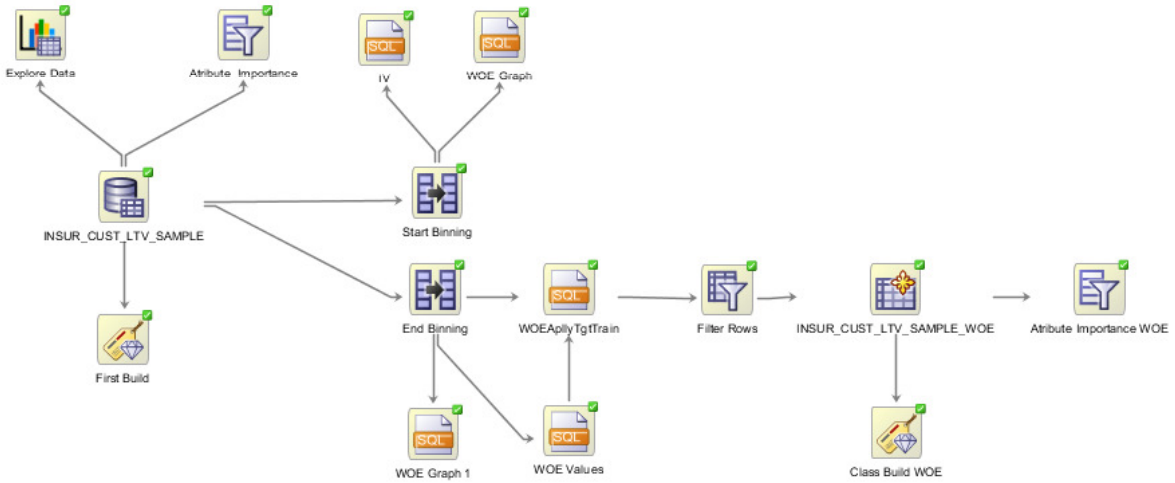
- Primjer na varijabli AGE iz INSUR_CUST_LTV podataka
- Transformacija i kategorizacija numeričke varijable
- Izračun IV-a pomoću SQL čvora koristeći R skriptu u bazi



	NAME	IV
1	BANK_FUNDS_BIN_Q	3,14069519
2	N_TRANS_ATM	2,375496...
3	MONEY_MONTLY_OVERDRA_BIN	2,346874...
4	BANK_FUNDS_BIN_N	0,197735...
5	AGE_BIN	0,08070162
6	CAR_OWNERSHIP	0,063582...
7	SALARY_BIN	0,038294...

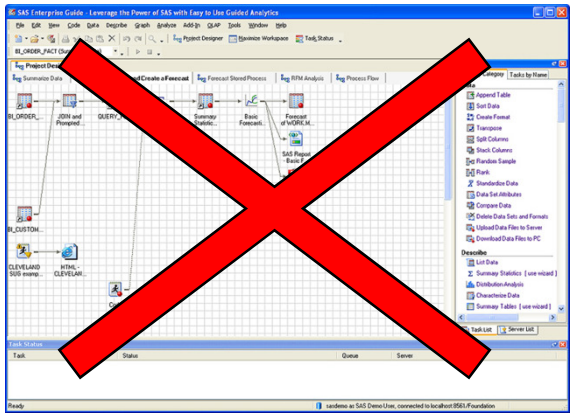
Primjena modela

- Modeli se primjenjuju na produkcijskom sustavu
- Odvojeno od razvojnog sustava
- Deployano pomoću ODM code generation opcije

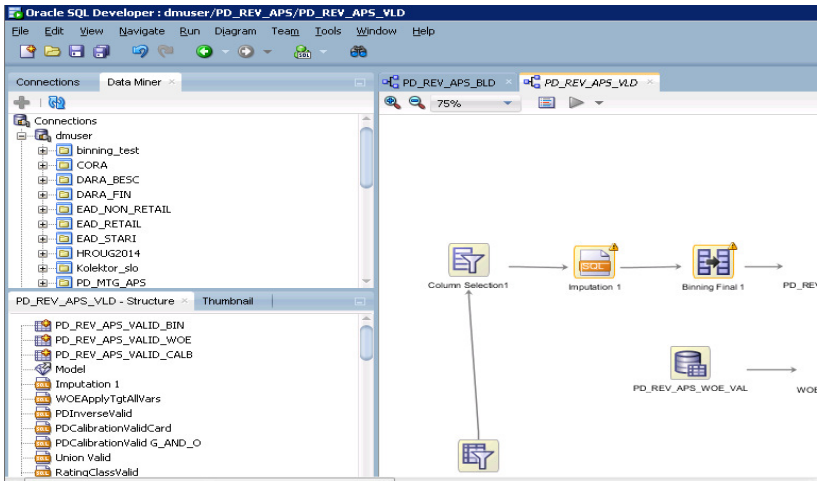


Rezultati

- 12 modela napravljenih na Oracle advanced analytics platformi (OAA)
- Korišteno preko 7000 varijabli u procesu razvoja modela
- 32 worflowa u Oracle data mining (ODM)
- Preko 4000 linija koda u R-u i SQL-u
- mSelect framework razvijen u ORE-u
- Information Value (IV) izračun u ORE i ODM
- Isti modeli s istim rezultatima



MIGRACIJA



Hvala!