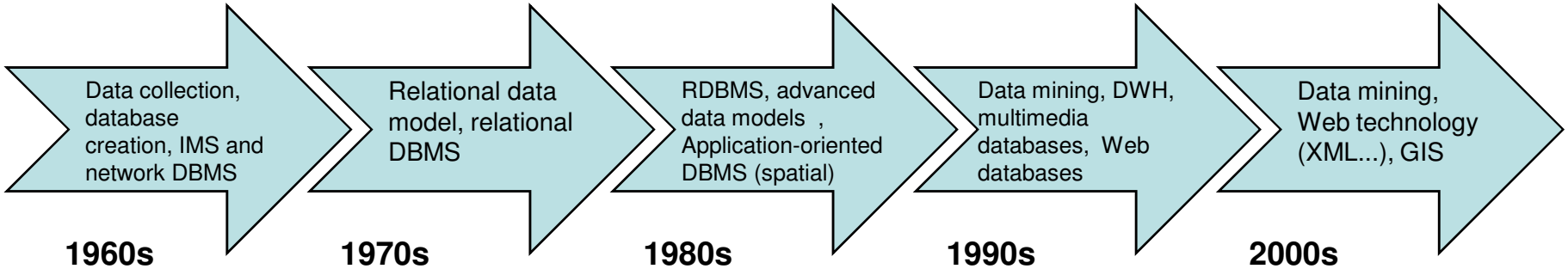
# U rudnik s SQL Developerom
# primjena Oracle data mining alata


K.Bokulić, M.Šipek

# Evolution of Database Technology





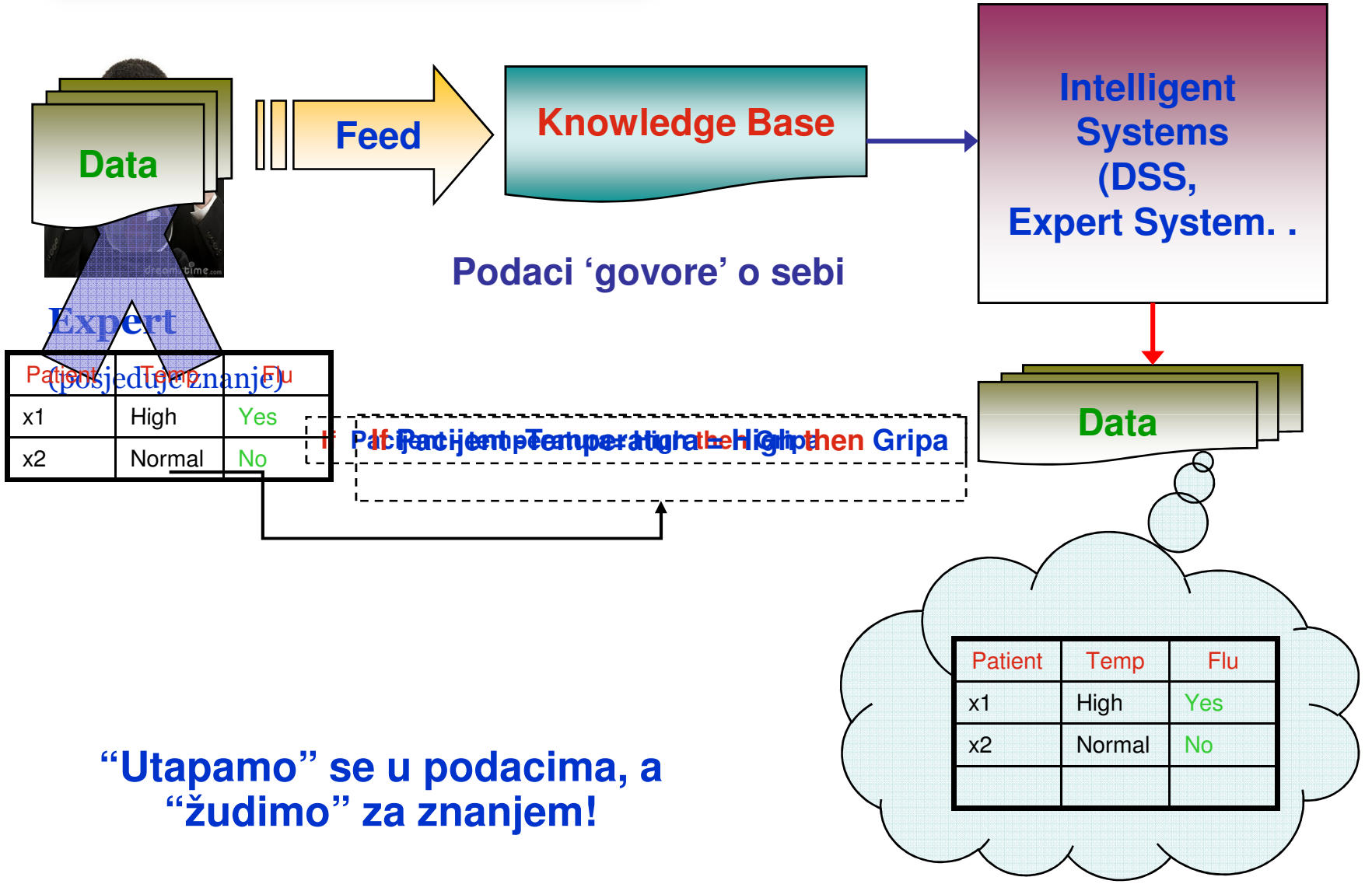
- Ekstrakcija detaljnih informacija
- Daje informaciju -

**Tko je kupio** proizvod A u zadnja 3 mjeseca

- Otkrivanje znanja i uzoraka
- Daje uvid i predikciju -

**Tko će kupiti** proizvod A u iduća 3 mjeseca **i zašto**?

# The New Story: Data Mining and KDD




"Utapamo" se u podacima, a
"žudimo" za znanjem!

# Svjetla budućnost:

*"A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."*

*Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years."*

*Gartner*

# Što je Data mining

- **Rudarenje podataka** (*Data mining*) je postupak *sortiranja, organiziranja ili grupiranja* velikog broja podataka i izvlačenje relevantnih informacija.

- Možemo ga definirati kao proces pronalaženja korisnog znanja kojim se ***otkrivaju odnosi***, ***logičnost***, ***pravilnost*** te općenito bilo kakve ***strukture u podacima***.

- Alternativni nazivi:
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data / pattern analysis, data archeology, data dredging, information harvesting, business intelligence, dr.

# Primjena DM-a

- **Direktni marketing** –najveća vjerojatnost odaziva
- **Izradu profila kupaca** – prilagođena ponuda
- **Segmentacija** – segmente kojima se mogu posebno prilagoditi usluge
- **Istraživanje povezanosti prodaje različitih proizvoda**– analiza kupovne košarice
- **Stimulacija kupovine drugih artikala istog poduzeća** - crossell, upsell
- **Zadržavanje klijenata, životna vrijednost i aktivacija klijenta**
- **Racionalizacija poslovanja**
- **Razni modeli rizika-** odlaska klijenata, pranja novca, kreditni rizik

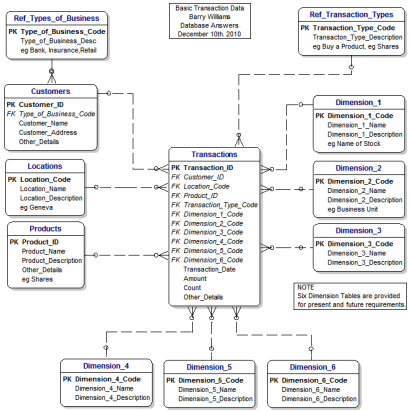
# Tko koristi Data Mining

- Financijske institucije
- Telekomi
- Osiguravajuća društva
- Maloprodaja
- Servisi
- Zdravstvo
- Autoindustija
- Proizvodnja
- Kemijska industrija

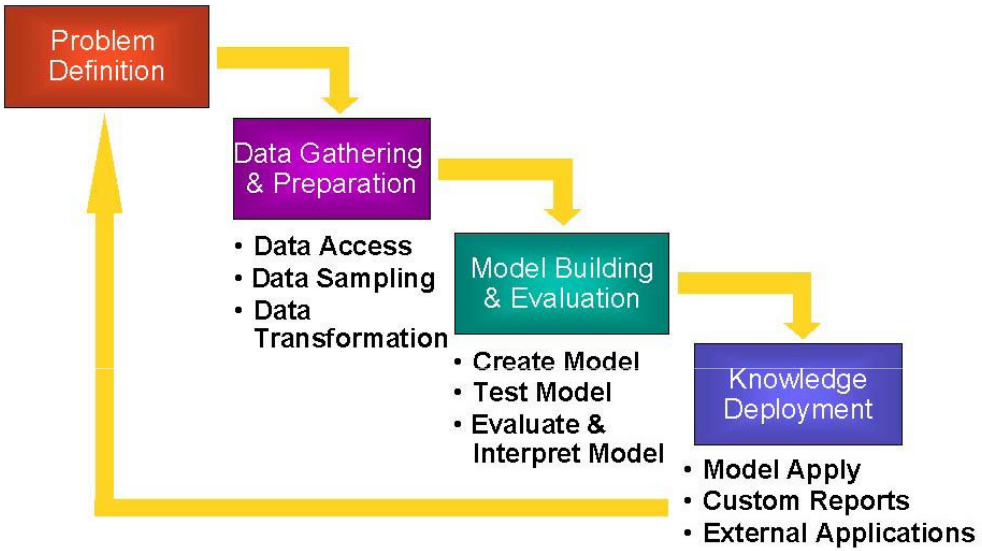
....

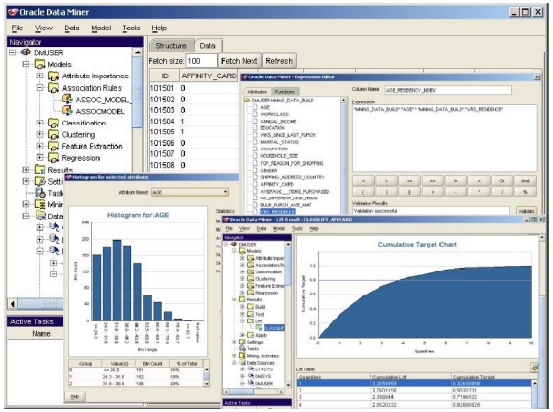

# Proces data mininga

## DWH model




## The Data Mining Process



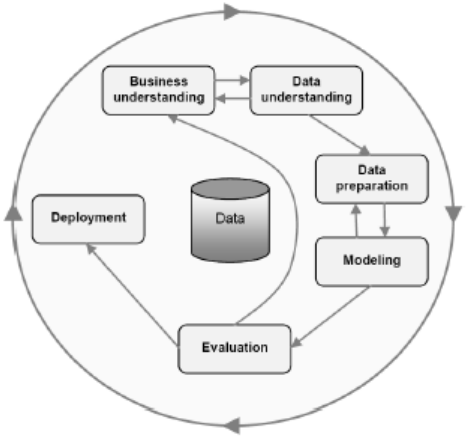




- Odabir metode ovisi o cilju analize.
- Proces rudarenja podataka je **iterativan!**
- Svaka iteracija → precizniji model

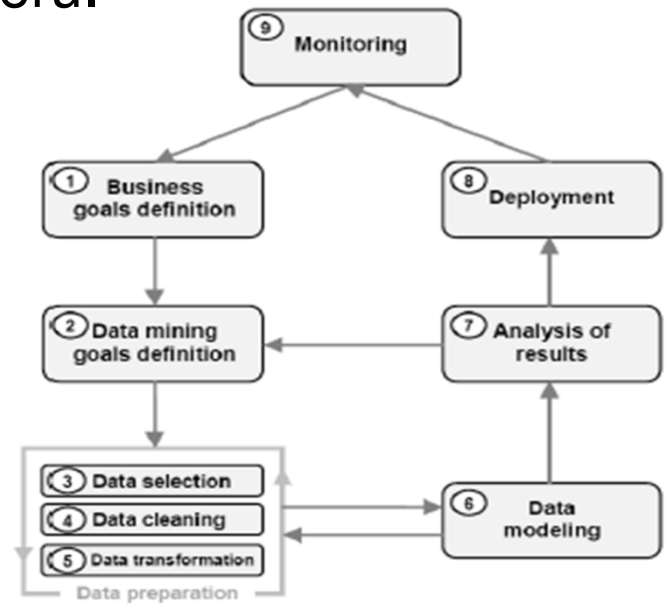
# Metodologija i framework

- Često govorimo o upotrebi rudarenja podataka radi vidljivih “benefita”, međutim, metodologija koja se koristi i koraci unutar frameworka nisu uvijek jasni.

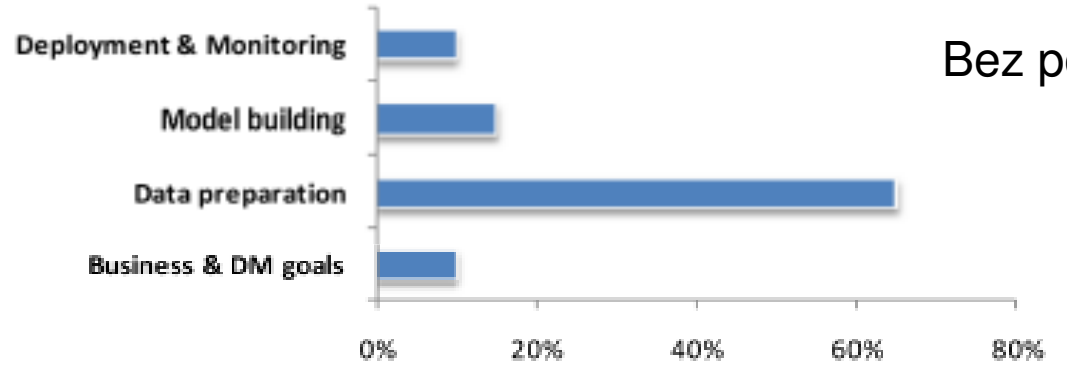
- Za izradu modela korištena je **CRISP-DM** metodologija, preporučena prema Gartneru.



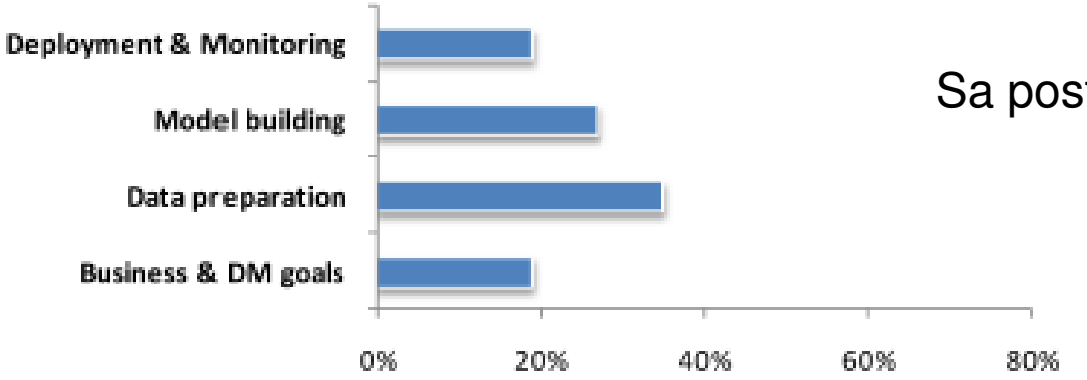

CRISP-DM



# Utrošeno vrijeme po koracima metodologije




Bez postojećeg DWH




Sa postojećim DWH

# Tipovi DM-a

- Dva osnovna tipa rudarenja podataka:
  - ***verifikacija hipoteze*** – cilj je provjeriti da li je neka ideja ili dojam o važnosti odnosa među određenim podacima utemeljen ili ne
  - ***otkrivanje novih znanja*** – među nekim pojavama mogu postojati neki još nepoznati, a statistički važni odnosi koje čovjek ni iskustvom niti svojim intelektualnim sposobnostima ne može spoznati

# Kako DM model uči?
# Tipovi dubinske analize

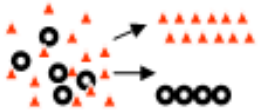
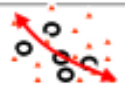
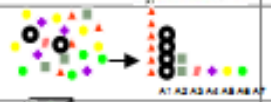
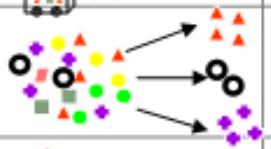
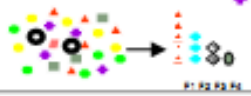
- Za otkrivanje znanja primjereni su postupci učenja iz primjera.
- Nadzirano učenje (eng. ***Supervised/directed data mining***) podrazumijeva pristup „odozgo prema dolje“ i koristi se kada analitičar zna što traži ili što želi predvidjeti. Stoga se ovaj stil često naziva **prediktivnim modeliranjem**.
- S druge je strane nenadzirano učenje (eng. ***Unsupervised / Undirected data mining***) koje se često naziva i **deskriptivnim modeliranjem** te podrazumijeva pristup „odozdo prema gore“. Problem učenja definiran je kao traženje grupa sličnih primjera (engl. *clustering*).

# Informacije koje utvrđujemo

Rudarenjem je moguće utvrditi sljedeće vrste informacija:

- **Klase**, (unaprijed definirane klase)
- **Klastere** ( bez unaprijed zadanih klasa)
- **Asocijacije** (uvjetovanost događaja)
- **Sekvence** ( događaji koji u određenoj vjerojatnosti slijede jedan za drugim)
- **Prognoze** ( prognozira se budućnost iz postojećih podataka)

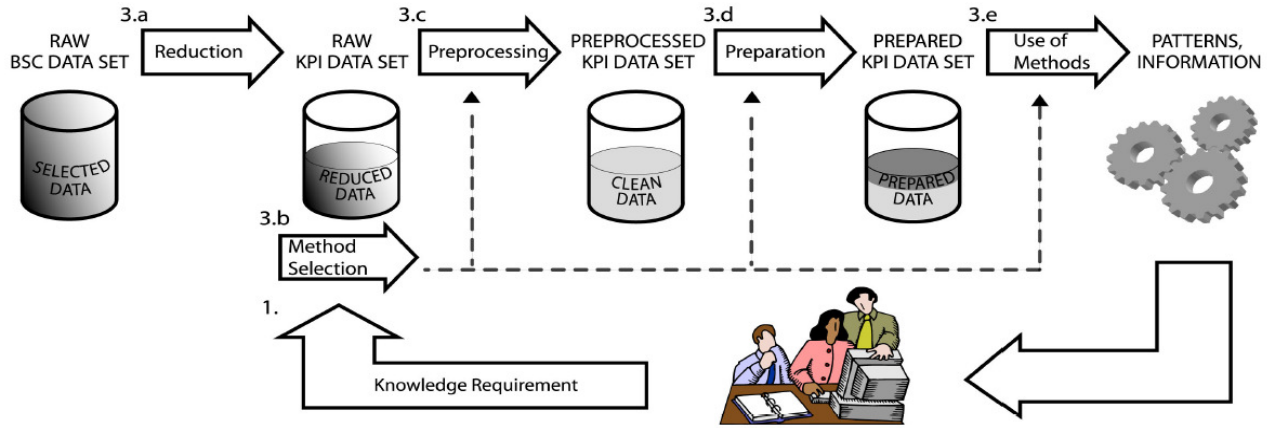
# Modeli i algoritmi Oracle DM-a

| Problem | Algorithm | Applicability |
|---|---|---|
| Classification  | Logistic Regression (GLM)<br>Decision Trees<br>Naïve Bayes<br>Support Vector Machine | Classical statistical technique<br>Popular / Rules / transparency<br>Embedded app<br>Wide / narrow data / text |
| Regression  | Multiple Regression (GLM)<br>Support Vector Machine | Classical statistical technique<br>Wide / narrow data / text |
| Anomaly Detection  | One Class SVM | Lack examples of target field |
| Attribute Importance  | Minimum Description Length (MDL) | Attribute reduction<br>Identify useful data<br>Reduce data noise |
| Association Rules  | Apriori | Market basket analysis<br>Link analysis |
| Clustering  | Hierarchical K-Means<br>Hierarchical O-Cluster | Product grouping<br>Text mining<br>Gene and protein analysis |
| Feature Extraction  | Nonnegative Matrix Factorization | Text analysis<br>Feature reduction |

# Data Preprocessing - Zašto?

- Podaci su stvarnom svjetu su “prljavi”
  - nekompletni: nedostatak vrijednosti atributa, nedostatak određenih atributa od interesa ili postojanje samo agregiranih podataka
  - “bučni”: sadrže grešku ili krajnost
  - nekonzistentni: sadrže diskrepanciju u nazivu ili šifri

- Bez kvalitetnih podataka nema ni rezultata  data mining-a!
  - Kvalitetne odluke temelje se na kvalitetnim podacima



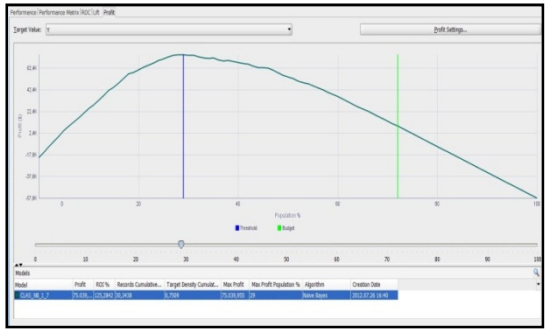
# Evaluacija modela

- **Potvrda modela -** Nakon izgradnje modela, provjeravaju se rezultati i tumači signifikantnost. Stopa točnosti pronađena tijekom testiranja odnosi se samo na podatke na kojima je model izgrađen

- **Confusion matrica -** Za klasifikacijske probleme, pokazuje stvaran odnos predviđenih klasa vrijednosti. Stupci pokazuju stvarne klase, redovi pokazuju predviđene klase, a dijagonala pokazuje sva točna predviđanja

| *Prediction* | *Actual* Class A | Class B | Class C |
|---|---|---|---|
| **Class A** | 45 | 2 | 3 |
| **Class B** | 10 | 38 | 2 |
| **Class C** | 4 | 6 | 40 |

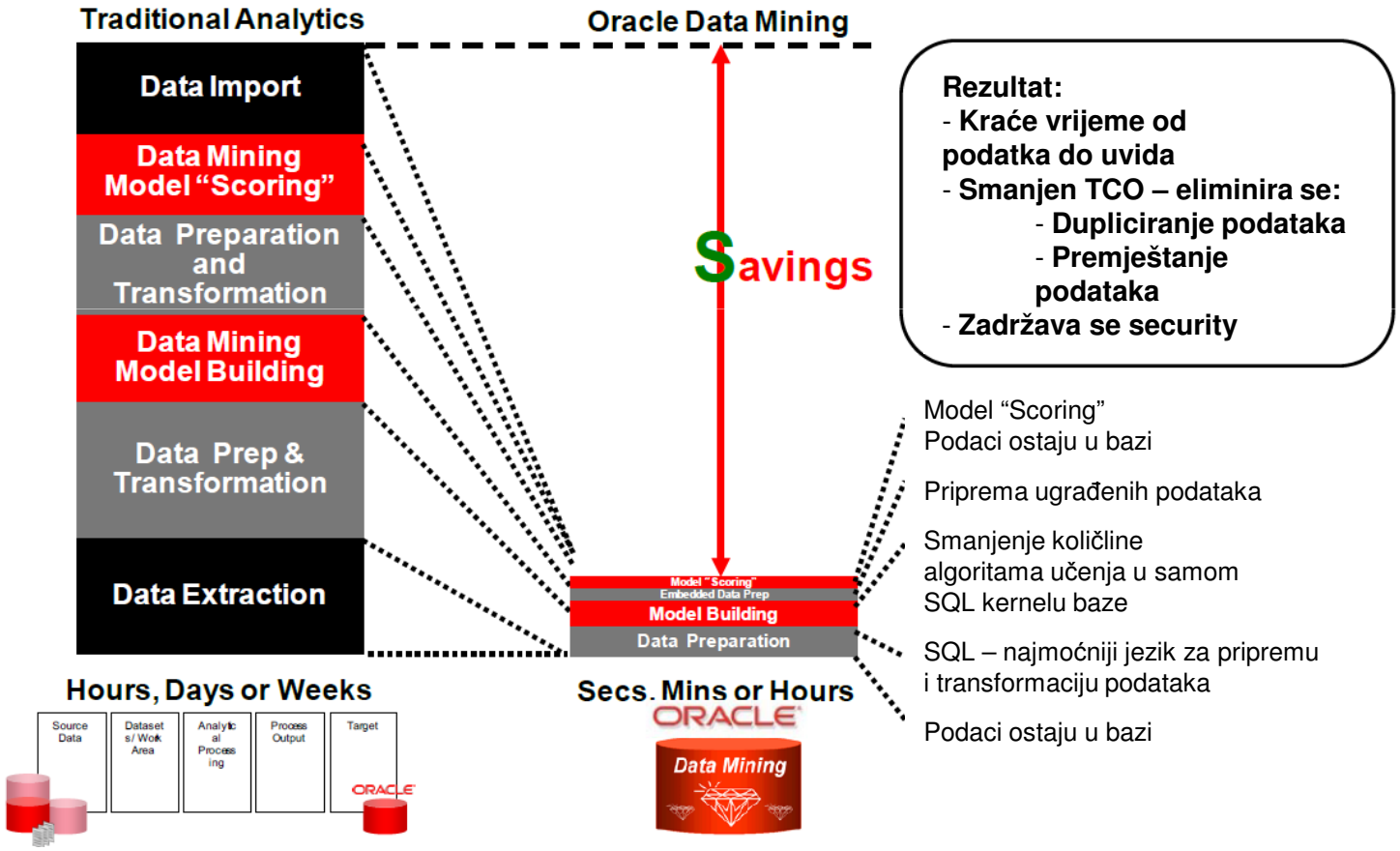
**Confusion matrics**



**ROI**

# Zašto Oracle?

## In-Database Data Mining




**Rezultat:**
- **Kraće vrijeme od podatka do uvida**
- **Smanjen TCO – eliminira se:**
  - **Dupliciranje podataka**
  - **Premještanje podataka**
- **Zadržava se security**

Model "Scoring"
Podaci ostaju u bazi

Priprema ugrađenih podataka

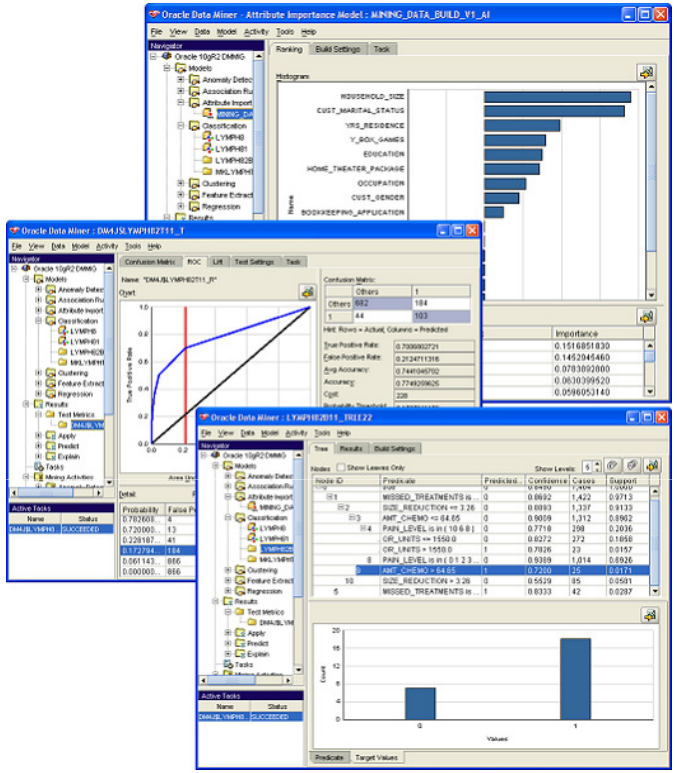
Smanjenje količline algoritama učenja u samom SQL kernelu baze

SQL – najmoćniji jezik za pripremu i transformaciju podataka

Podaci ostaju u bazi

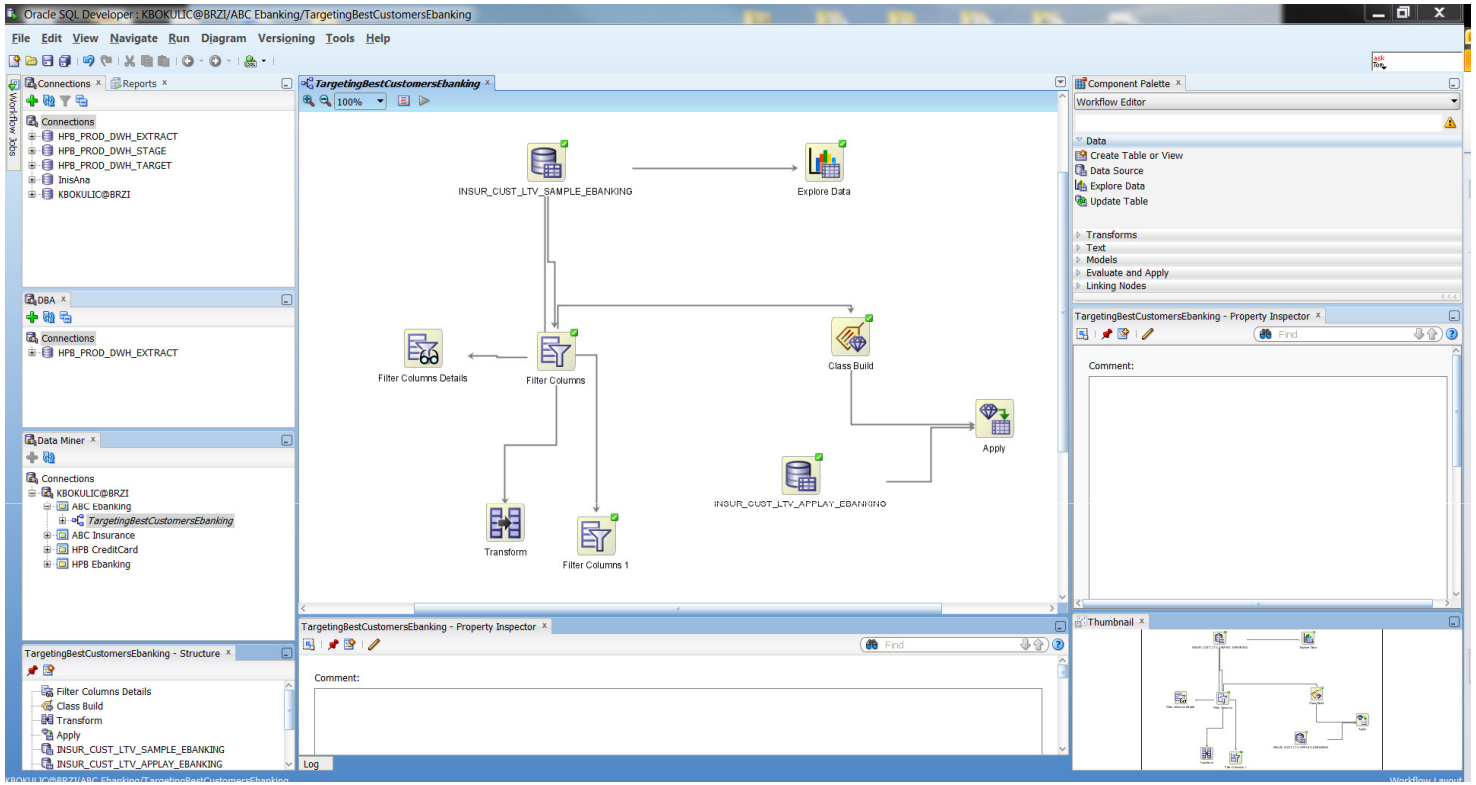
# Oracle Data Miner

- 10 godina sa nama, od sada integrirana opcija SQL Developera
- Obuhvaćeni svi aspekti: dohvat, transformacija, procesiranje, izrada i evaluacija modela
- Mogućnost utjecaja na sve elemente procesa
- Interaktivna vizualizacija i izrada izvještaja, uz set predefiniranih izvještaja i analitika
- Alat za analitičare



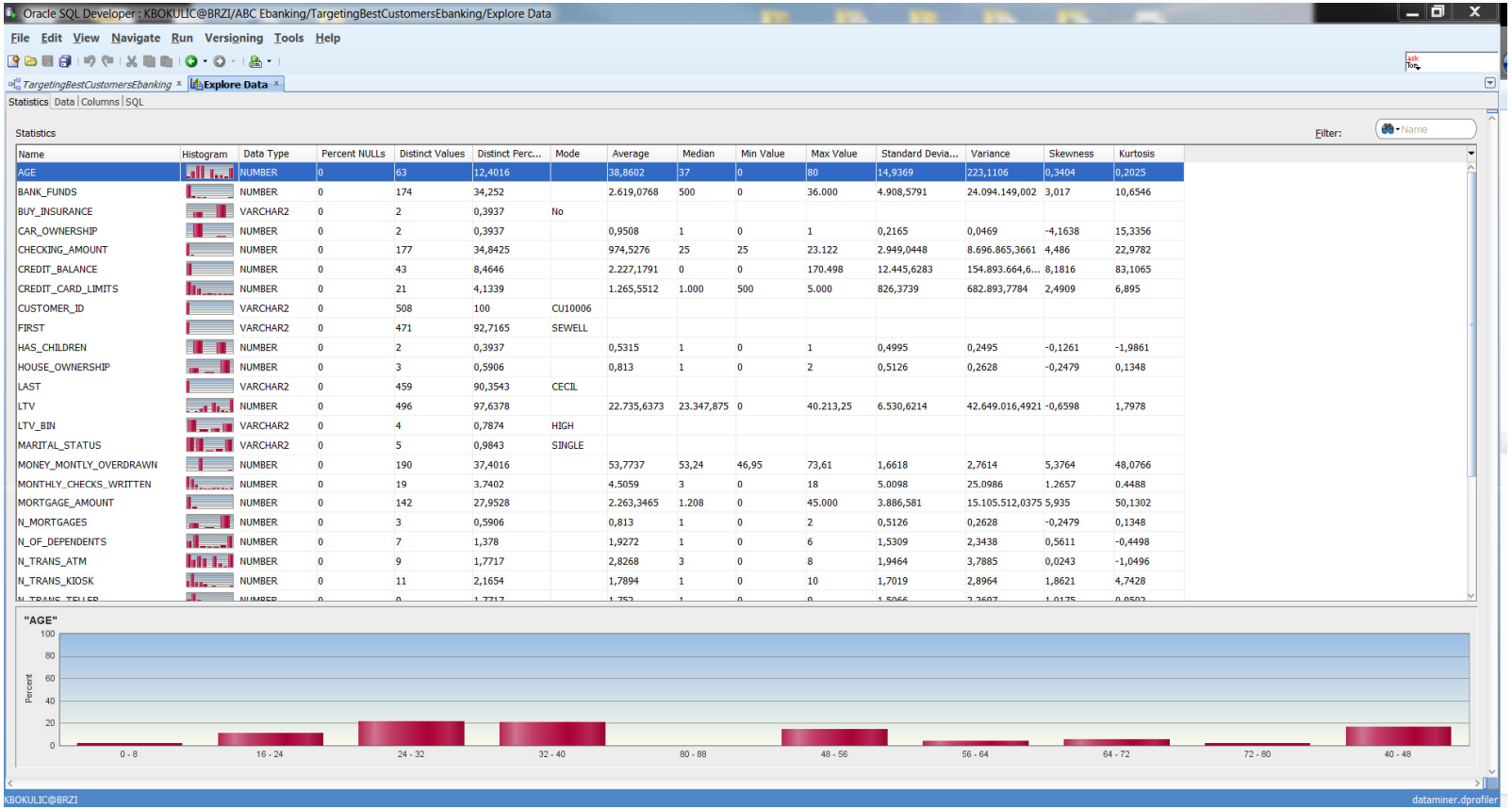
**Oracle Data Mining**
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the "right" data to mine
- ODM "Verbs"
  - PREDICT
  - DETECT
  - CLUSTER
  - CLASSIFY
  - REGRESS
  - PROFILE
  - IDENTIFY FACTORS
  - ASSOCIATE

# Oracle Data Miner



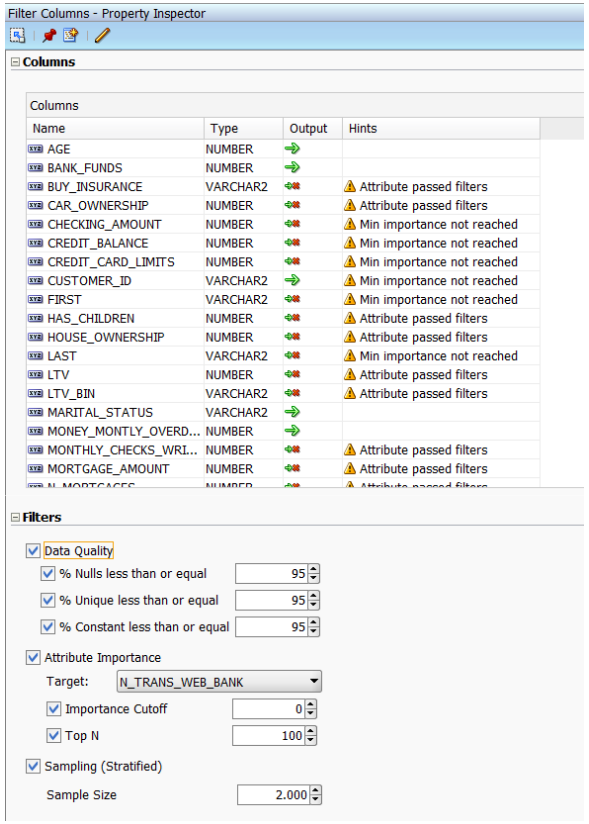

- Repozitorij unutar poznatog sučelja SQL Developera
- Izrada modela kroz work-flow
- Jednostavna rekalkulacija i remodulacija modela

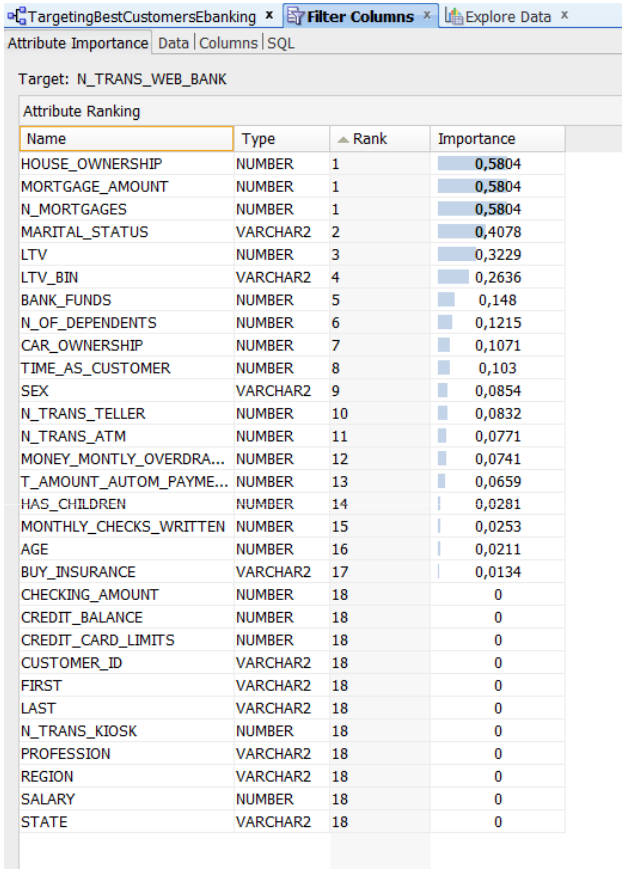
# Oracle Data Miner




- Statistička anliza podataka i uzorka
- Histogram, prosjek, mode, meidan, standardna deviacija, skewness, kurtosis …

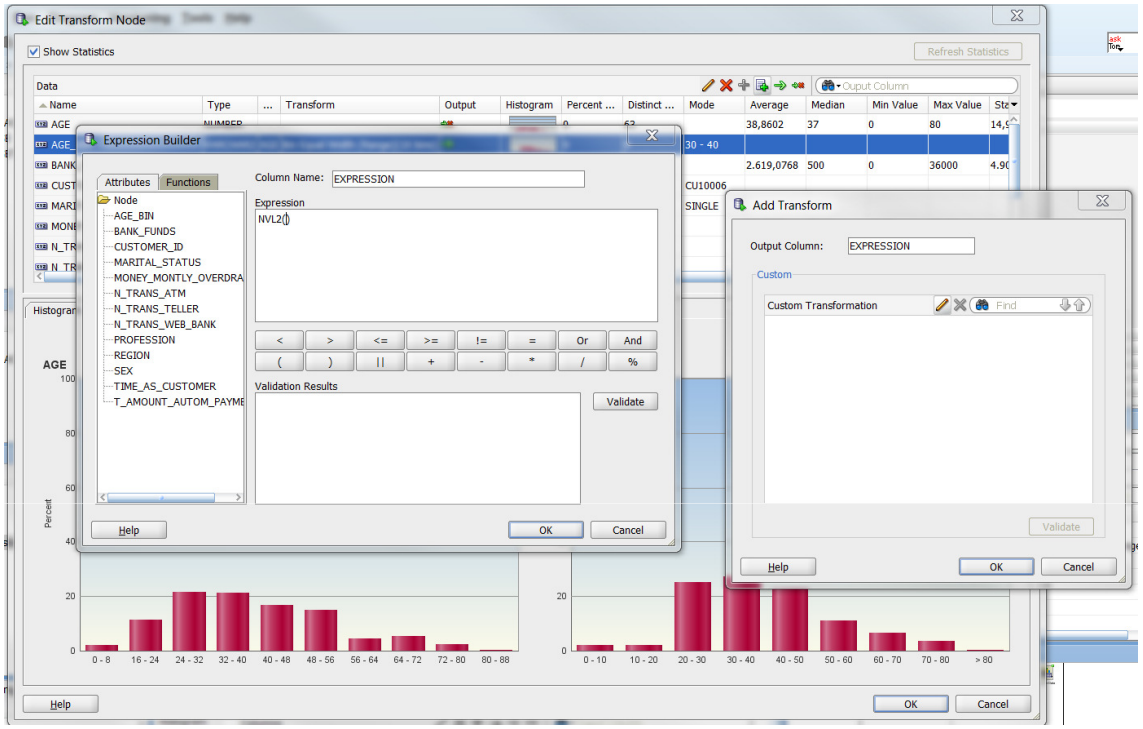
# Oracle Data Miner

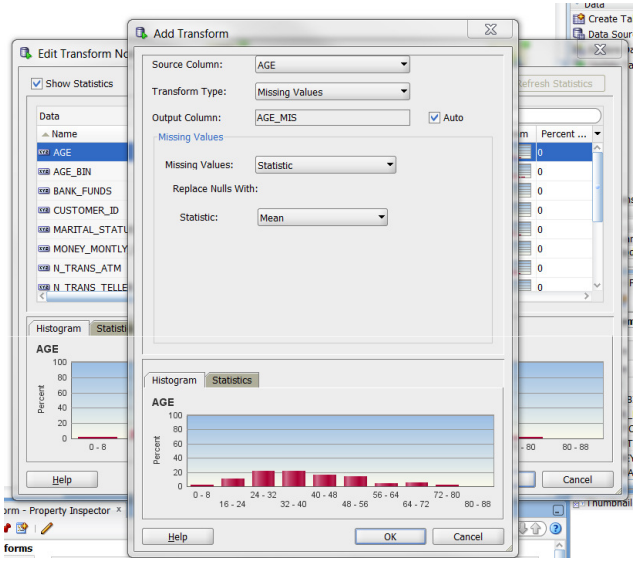



- Analiza kvalitete i prediktivnosti varijabli
- MDL algoritam

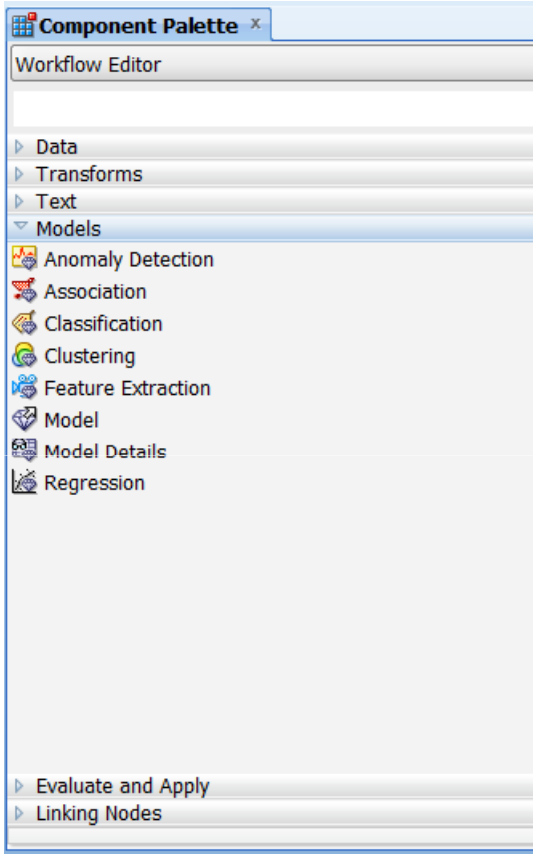
# Oracle Data Miner



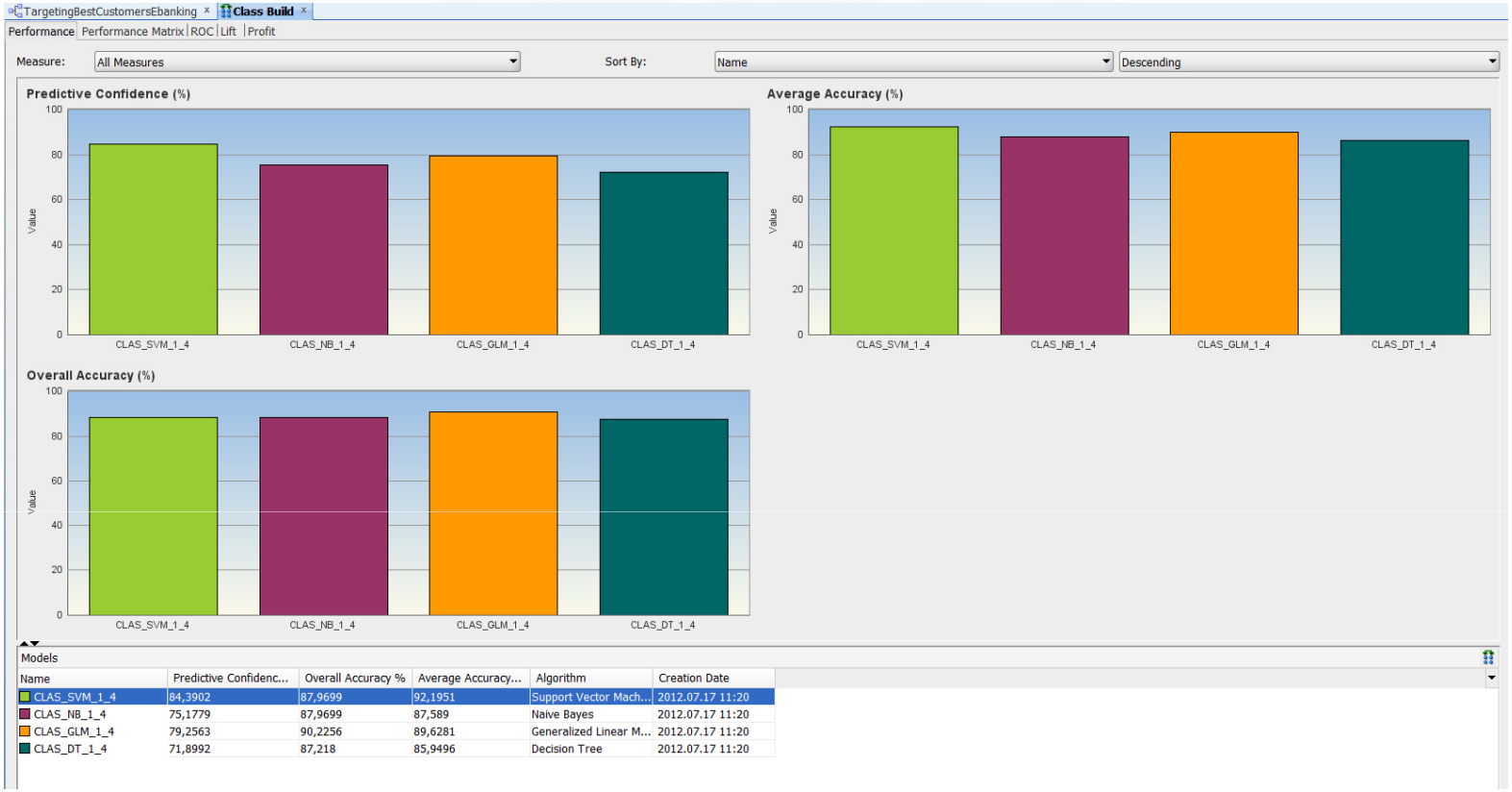

- Transformacije nad podacima, pomoću pl/sql izraza
- Popunjavanje nedostojećih podataka …

# Oracle Data Miner




- Učitavanje podataka
- Tranformacije i analiza podataka
- Izrada DM modela za ne strukturirane podatke
- Podržani svi industriski standardni modeli za DM
- Jednostavna primjena modela nad podacima iz različitih izvora

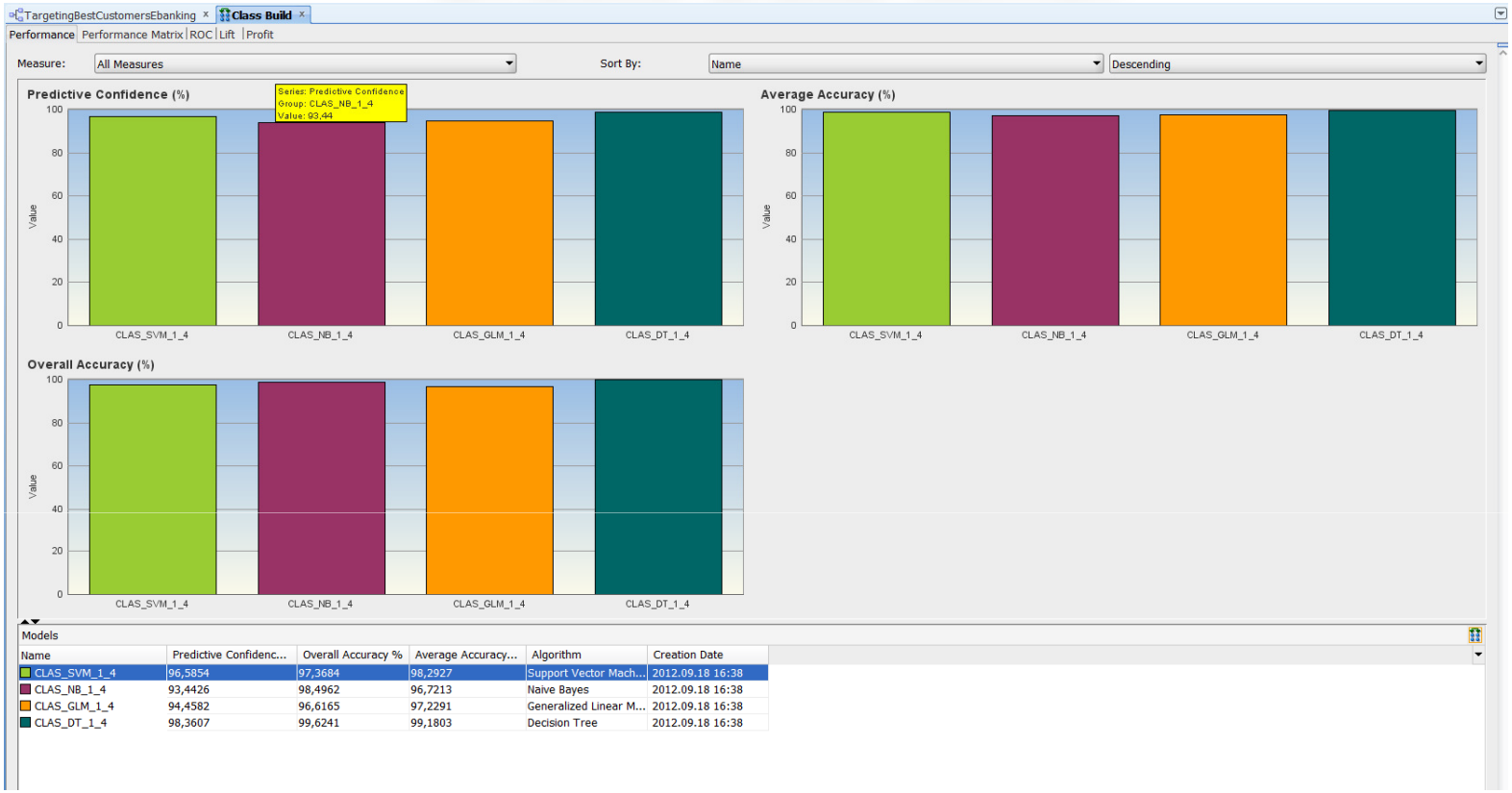
# Oracle Data Miner




Rezultati modela za predikciju kupovine usluge internet bankarstva

- Istovremena izgradnja višestrukih modela različitim algoritmima
- Grafička usporedba rezultata više modela

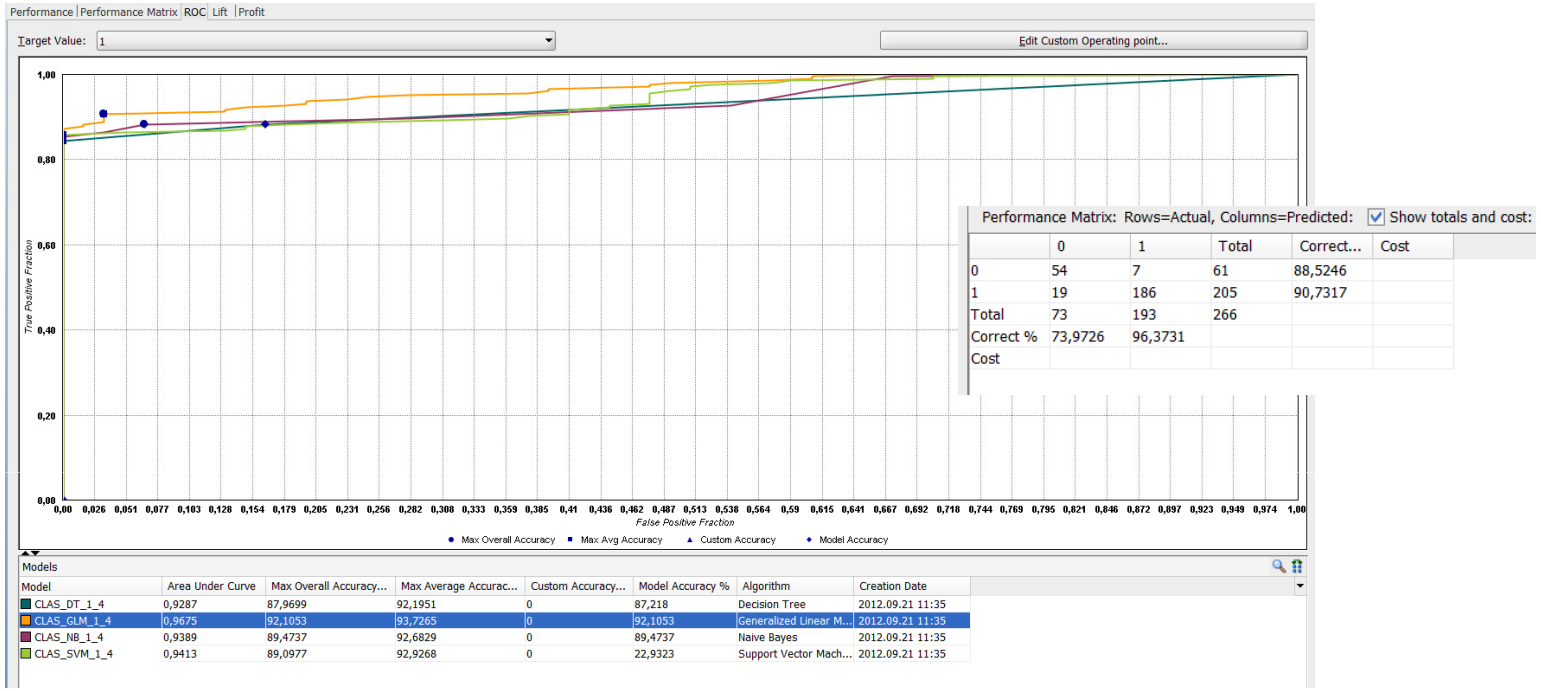
# Oracle Data Miner




Rezultati modela koji koristi dodatnu varijablu **`vlasnik_nekretnine`**

- Modeli su osjetljivi na dobar odabir varijabli s kvalitetnim podacima
- Kvalitetna varijabla s dobrom prediktivnošću - bolja prediktivnost modela

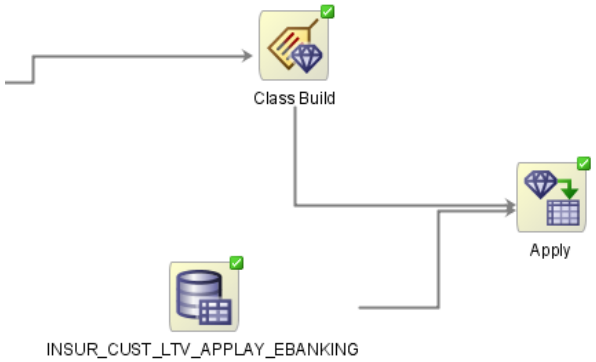
# Oracle Data Miner




ROC krivulja i performance matrica

- ROC krivulja nam prikazuje omjer dobrih i loših predivđanja u testnom uzorku za ciljanu vrijednost

# Oracle Data Miner




Data | Columns | SQL

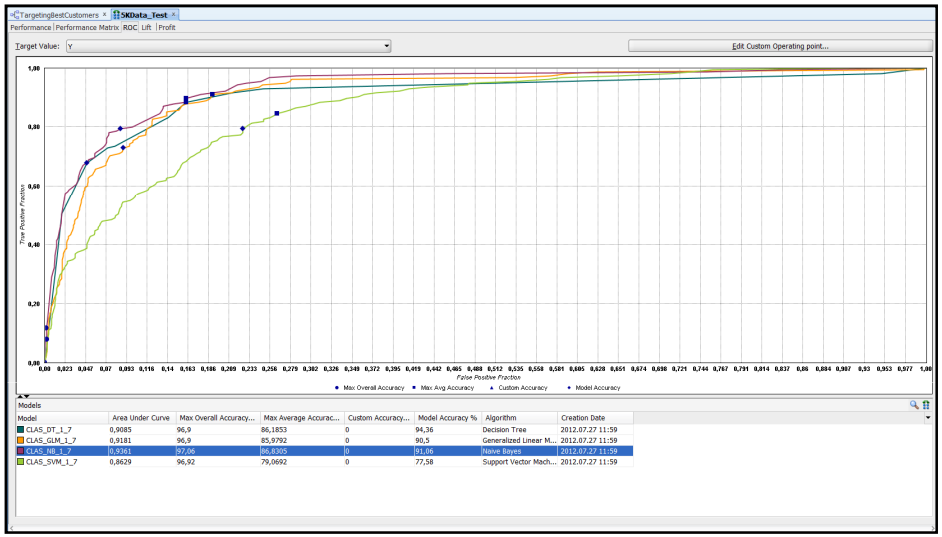
View: Actual Data | Sort... | Filter: Enter Where Clause

| | CLAS_SVM_1_4_PRED | CLAS_SVM_1_4_PROB | CUSTOMER_ID | FIRST | STATE | SEX | AGE |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,8262 | CU11209 | ABRAHAM | CA | M | 33 |
| 2 | 0 | 0,7777 | CU13483 | ADAMS | NY | F | 32 |
| 3 | 0 | 0,8218 | CU9648 | ADLER | FL | M | 34 |
| 4 | 0 | 0,7796 | CU13357 | ALLISON | NV | M | 0 |
| 5 | 1 | 0,8129 | CU10508 | ALONSO | FL | M | 62 |
| 6 | 1 | 0,7943 | CU6758 | ALONZO | NY | M | 62 |
| 7 | 1 | 0,8229 | CU11959 | ALONZO | NY | M | 31 |
| 8 | 1 | 0,8457 | CU3486 | ALSTON | CA | F | 48 |
| 9 | 1 | 0,8231 | CU5473 | ALVARADO | NY | M | 42 |
| 10 | 1 | 0,8361 | CU10308 | ANDERSON | NY | F | 42 |
| 11 | 0 | 0,8253 | CU10540 | ANDERSON | NY | M | 33 |
| 12 | 1 | 0,8224 | CU15165 | ARELLANO | NV | M | 20 |
| 13 | 1 | 0,8346 | CU3654 | ARNOLD | NY | F | 50 |
| 14 | 1 | 0,8233 | CU14806 | ARTHUR | MI | M | 66 |
| 15 | 0 | 0,7821 | CU12993 | ATKINSON | NY | F | 23 |
| 16 | 1 | 0,8254 | CU14750 | AUSTIN | MI | M | 48 |
| 17 | 1 | 0,8195 | CU5056 | BABCOCK | CA | M | 32 |
| 18 | 1 | 0,8515 | CU2410 | BAER | CA | F | 49 |
| 19 | 0 | 0,8185 | CU14975 | BAEZ | WI | M | 35 |
| 20 | 1 | 0,824 | CU10228 | BAGLEY | NY | M | 27 |
| 21 | 0 | 0,8177 | CU14291 | BALL | NY | M | 0 |
| 22 | 1 | 0,8299 | CU5351 | BARBER | MI | F | 28 |
| 23 | 1 | 0,8447 | CU6501 | BARGER | CA | F | 66 |
| 24 | 1 | 0,8234 | CU15116 | BARON | CA | M | 45 |
| 25 | 0 | 0,835 | CU14140 | BARRERA | NY | M | 36 |
| 26 | 1 | 0,8193 | CU10706 | BARRON | NY | M | 26 |
| 27 | 1 | 0,8303 | CU3044 | BARTH | WA | F | 49 |
| 28 | 1 | 0,8576 | CU3229 | BEAULIEU | MN | F | 37 |

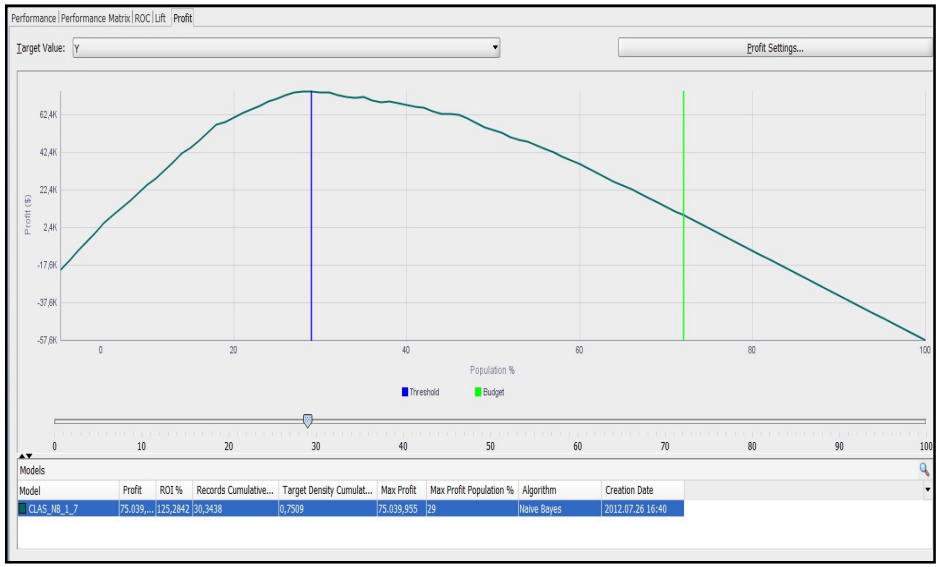
Rezultat modela s vjerojatnostima ishoda

- Podaci nad kojima je primijenjen model smještaju se u tablicu u bazi

# Oracle Data Miner




ROC krivulje




Projekcija profita kampanje

**Primjer:** stvarni rezultati modela povećanje prodaje usluge (up-sell)

# Oracle Data Miner

**Izazovi modeliranja:**

- Kako definirati problem ?

- Koliko varijabli za model ?
- Koje varijable za model ?
- Koji algoritam ?

- Koji uzorak za trening podatke ?
- Koliko je podataka dovoljno?
- Da li je model ‘pretreniran’ ili nije dovoljno ‘treniran’ ?
- Koliko i koji podaci za test modela ?

- Koje razdoblje trening podataka ?
- Koje razdoblje za ishod modela ?

# Pitanja?

Mario.Sipek@multicom-is.hr

Kresimir.Bokulic@multicom-is.hr