RUDARENJE PODACIMA U ORACLE R ENTERPRISE-U

Laren Zuliani

Ukratko o Neosu

- Osnovne informacije
  - Osnovano 2002
  - 30+ zaposlenih
  - Specijalizirani za DW/BI sustave, Java/Oracle Custom Development i konzalting
- Dugogodišnje iskustvo
  - Više od 10 godina DW/BI & CD iskustva
  - Certificirani stručnjaci (OCP, Specialists, Experts...)
  - Brojni uspješni projekti i zadovoljni klijenti

Ukratko o Neosu

- Oracle Partner - Gold Level
  - Od 2002
- Specialized Oracle BI Foundation Partner
  - 30-ti u svijetu i prvi u regiji
- Specialized Oracle ADF Partner
  - Jedan od prvih partnera u svijetu

ORACLE Gold Partner
Specialized
Data Warehousing

ORACLE Gold Partner
Specialized
Oracle Business Intelligence Foundation

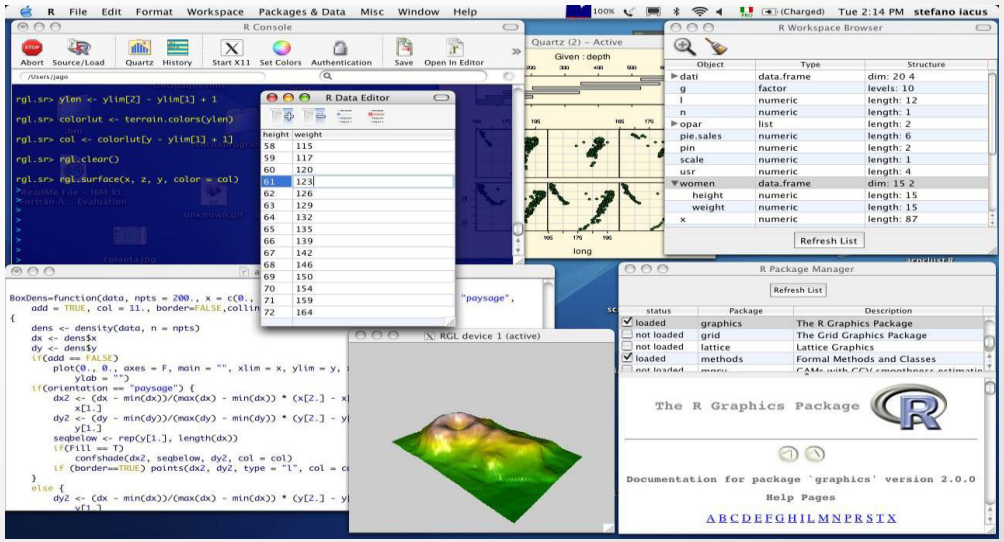
ORACLE Gold Partner
Specialized
Oracle Application Development Framework

Što je R?

- Open source skriptni jezik i okruženje za statističke izračune i grafički prikaz
- pojavio se 1994
- alternativa za SAS, SPSS i druge komercijalne alate i okruženja
- oko 2 milijuna korisnika
- tisuće open source paketa koji povećavaju produktivnost u područjima kao:
  - bioinformatika
  - analiza financijskog tržišta
  - linearno i nelinearno modeliranje

Zašto R?

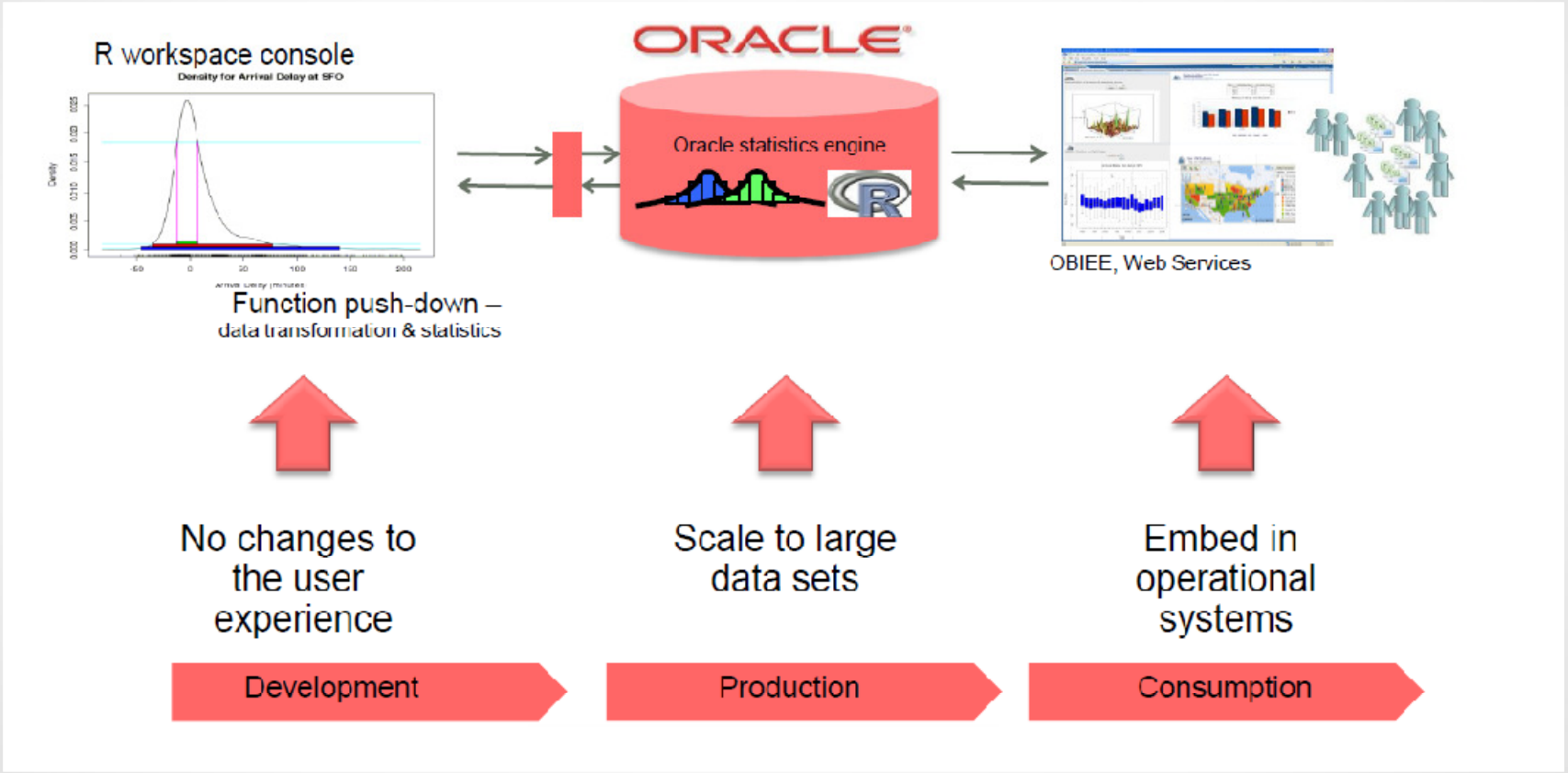
- Moćan
- Nadogradiv
- Lako se instalira i koristi
- Mogućnosti grafičkog prikaza
- *Besplatan je*



R okruženja

Which R interfaces do you use frequently?	
built-in R console (225)	40%
RStudio (135)	24%
Eclipse with StatET (90)	16%
RapidMiner R extension (80)	14.2%
Tinn-R (62)	11%
ESS (Emacs Speaks Statistics) (59)	10.5%
Rattle GUI (53)	9.4%
R Commander (43)	7.7%
Revolution Analytics (31)	5.5%
RKWard (22)	3.9%
JGR (Java Gui for R) (21)	3.7%
RExcel (18)	3.2%
R via a data mining tool plugin (12)	2.1%
Red-R (8)	1.4%
SciViews-R (6)	1.1%
Other (44)	7.8%

Oracle R Enterprise arhitektura



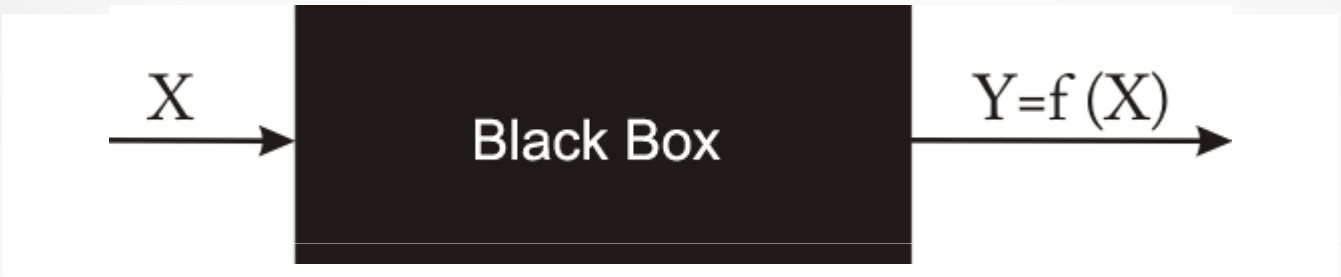
ORE – unapređenje *open source* R-a

- Analiza i manipulacija podacima u Oracle bazi podataka kroz R, transparentno
- Izvođenje R skripti kroz bazu u paraleli
- Korištenje *in-database* algoritama prediktivne analize, neprimjetno kroz R
- *Scoring* R modela u Oracle bazi
- R skripte dinamički se integriraju u SQL
- Integrira R u IT programski stog (*software stack*)

OREdm paket

Algorithm	Main R function	Mining Type/Function
Minimum Description Length	ore.odmAI	Attribute Importance for Classification or Regression
Decision Tree	ore.odmDT	Classification
Generalized Linear Models	ore.odmGLM	Classification Regression
KMeans	ore.odmKMeans	Clustering
Naïve Bayes	ore.odmNB	Classification
Support Vector Machine	ore.odmSVM	Classification Regression Anomaly Detection

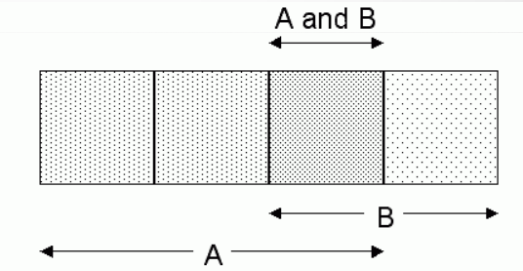
Princip „crne kutije”



- X: ulazni skup parametara - *atributa*
- Y: izlaz – uzima se iz postojećeg povijesnog skupa podataka
- f(X) = ? – određivanje „zakonitosti” među podacima

Naive Bayes algoritam

- Temeljen na uvjetnim vjerojatnostima
- Koristi Bayesov teorem - formula koja izračunava vjerojatnosti brojanjem učestalosti vrijednosti i kombinacija vrijednosti u povijesnim podacima
- Bayesov teorem:
  - *Prob(B given A) = Prob(A and B)/Prob(A)*
- Prednosti:
  - Brza, skalabilna izrada modela i *scoring*
  - Proces izgradnje je paraleliziran
  - Za binarnu i višeklasnu klsifikaciju
- Nedostaci:
  - Algoritam predpostavlja neovisnost među prediktorima



P(A) = 3/4

P(B) = 2/4

P(A and B) = P(AB) = 1/4

P(A|B) = P(AB) / P(B) = (1/4) / (2/4) = 1/2

P(B|A) = P(AB) / P(A) = (1/4) / (3/4) = 1/3

Naive Bayes – ORE primjer

RStudio

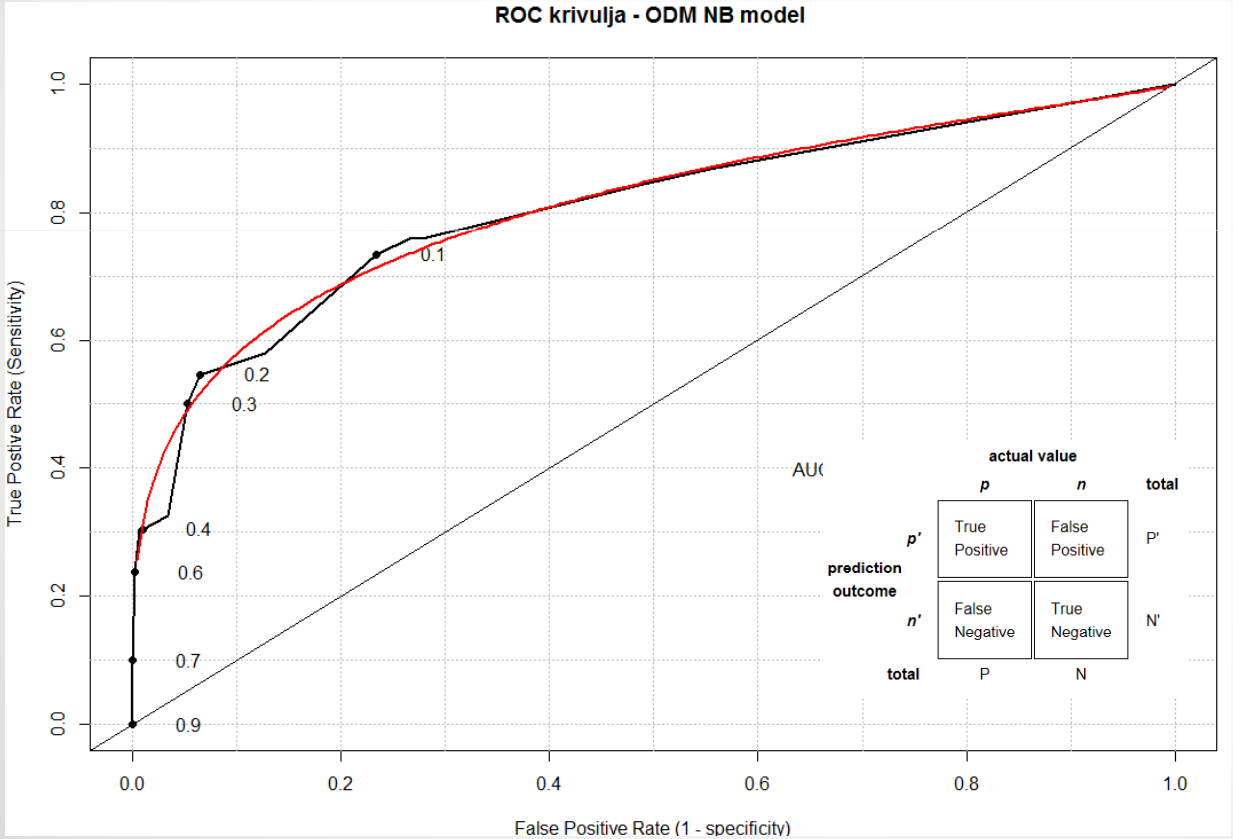
File Edit Code View Plots Session Project Build Tools Help

start_sess.R | Untitled1* | odm_svm.R | R demos | Untitled2* | titanic_nb.R* | Untitled3*

Source on Save

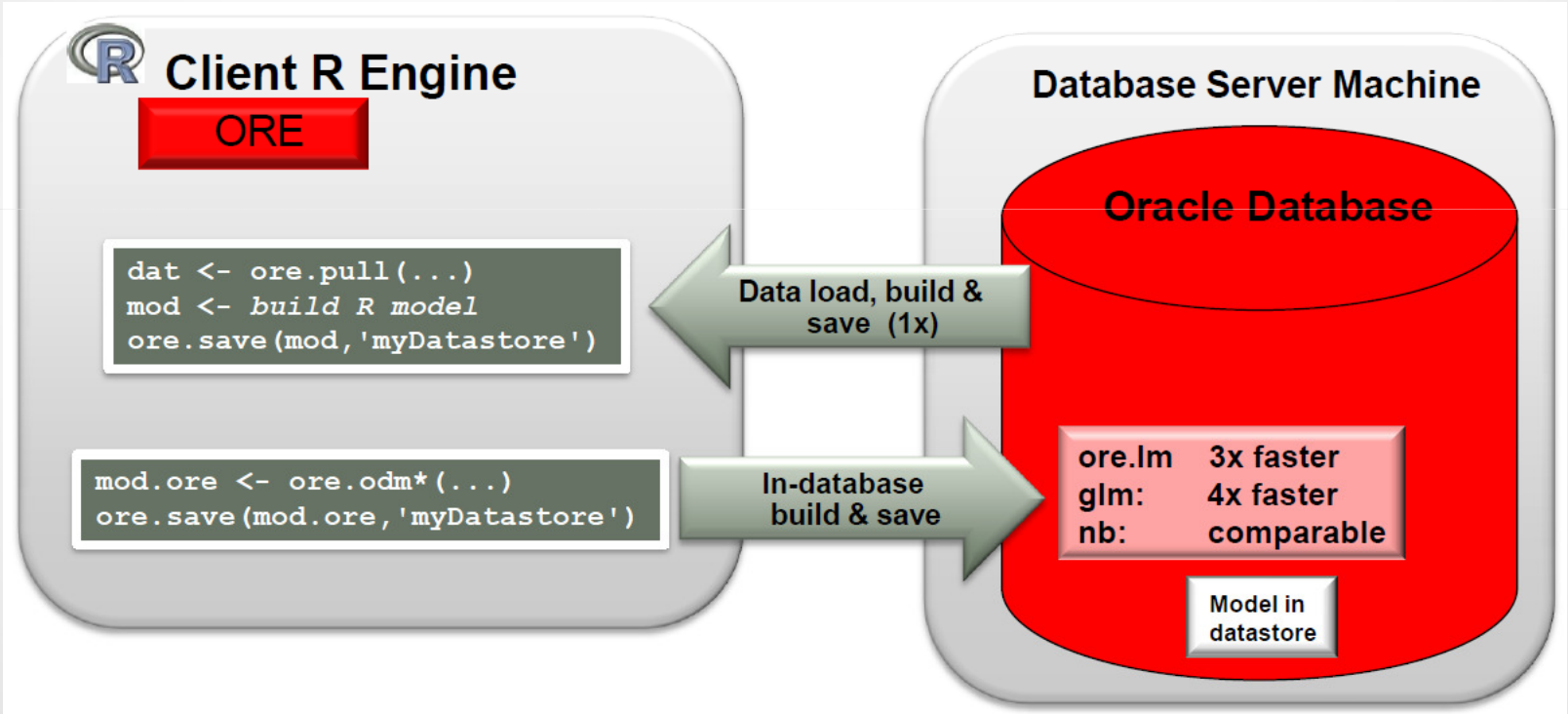
```
1  library(ORE)
2  ore.connect("rquser","orcl","localhost","rquser",all=TRUE)
3
4  data(titanic3,package="PASWR")
5  t3 <- ore.push(titanic3)
6  t3$survived <- ifelse(t3$survived == 1, "Yes", "No")
7
8  n.rows <- nrow(t3)
9  set.seed(seed=6218945)
10 random.sample <- sample(1:n.rows, ceiling(n.rows/2))
11 t3.train <- t3[random.sample,]
12 t3.test <- t3[setdiff(1:n.rows,random.sample),]
13
14 priors <- data.frame(TARGET_VALUE = c("Yes", "No"), PRIOR_PROBABILITY = c(0.1, 0.9))
15
16 nb <- ore.odmNB(survived ~ pclass+sex+age+fare+embarked,t3.train, class.priors=priors)
17
18
19 nb.res <- predict (nb, t3.test,"survived")
20 head(nb.res,10)
21 with(nb.res, table(survived,PREDICTION, dnn = c("Actual","Predicted")))
22
23 library(verification)
24 res <- ore.pull(nb.res)
25 perf.auc <- roc.area(ifelse(res$survived == "Yes", 1, 0), res$"'Yes'")
26 auc.roc <- signif(perf.auc$A, digits=3)
27 auc.roc.p <- signif(perf.auc$p.value, digits=3)
28
29 roc.plot(ifelse(res$survived == "Yes", 1, 0), res$"'Yes'", binormal=T,
30          plot="both",
31          xlab="False Positive Rate (1 - specificity)",
32          ylab="True Postive Rate (Sensitivity)",
33          main= "ROC krivulja - ODM NB model ")
34 text(0.7, 0.4, labels= paste("AUC ROC:", signif(perf.auc$A, digits=3)))
35 text(0.7, 0.3, labels= paste("p-value:", signif(perf.auc$p.value, digits=3)))
36
37 summary(nb)
38 ore.disconnect()
```

ROC (Receiver operating characteristic)

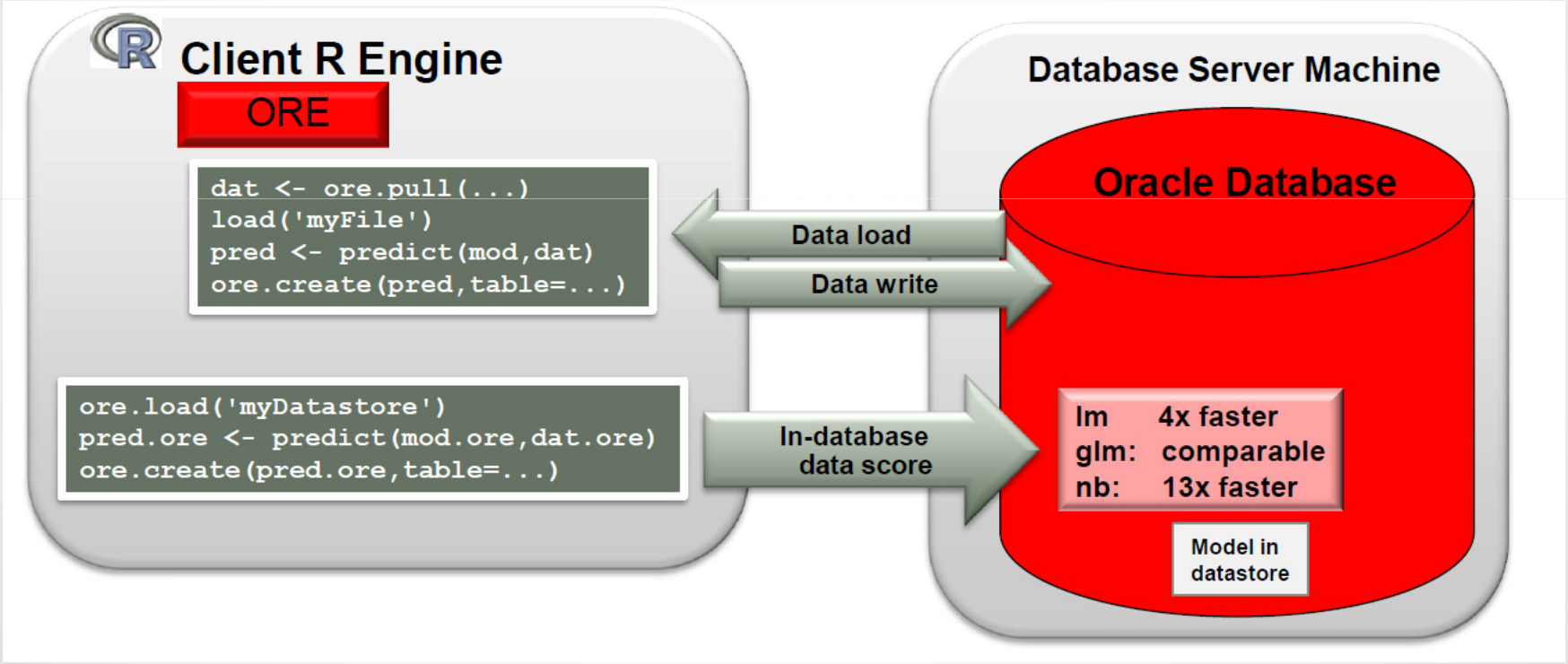


- WW2 – detekcija protivničkih objekata
- Grafički prikaz binarnog klasifikacijskog sustava
- *Confusion matrix*
  TPR=TP/P=TP/(TP+FN)
  FPR=FP/N=FP/(FP+TN)
- Medicina, radiologija, biometrija, *data mining*

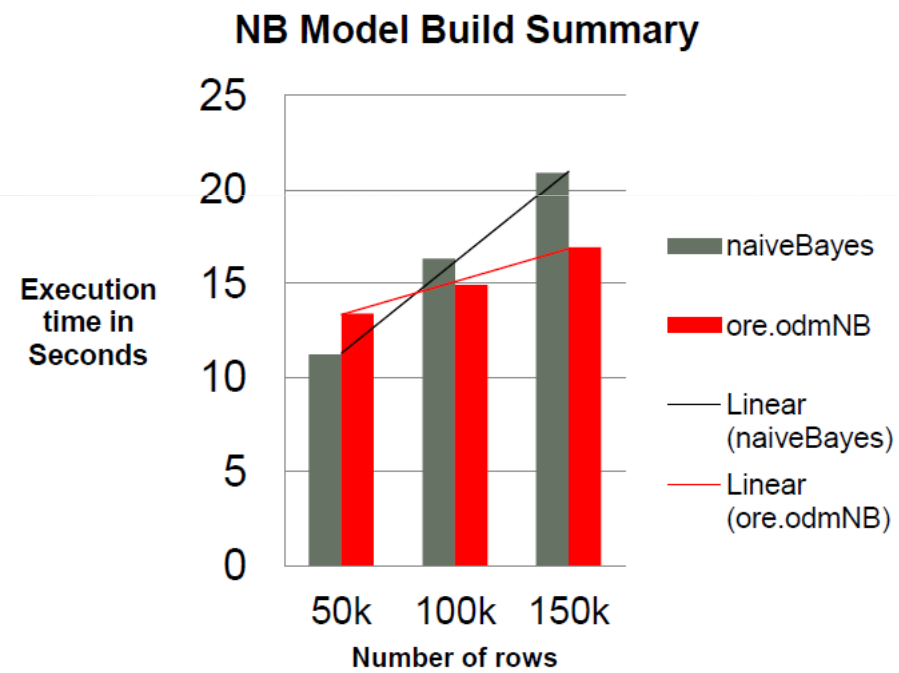
OREdm poboljšanje performansi – izgradnja modela

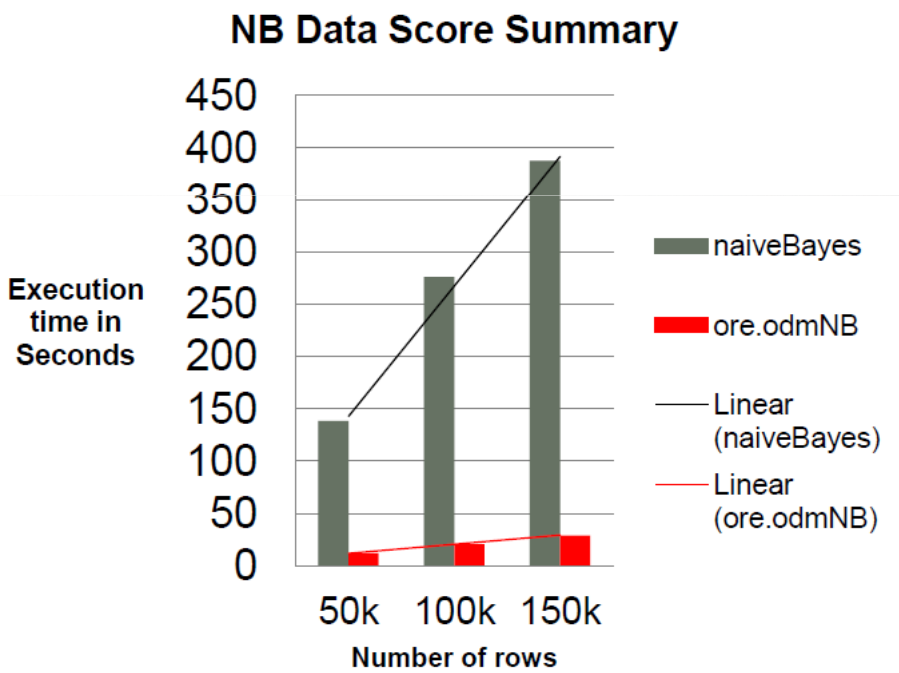


OREdm poboljšanje performansi – *data scoring*



OREdm poboljšanje performansi – grafički prikaz





Data Mining u bankarstvu

- Procjena rizika
- Prodaja dodatnih proizvoda postojećim klijentima
- Zadržavanje postojećih klijenata (CHURN)
- Segmentacija
- Životna vrijednost klijenta (CLV)
- Odaziv
- Aktivacija
- Racionaliziranje poslovanja

Case study 1: HFC Bank otkriva zlouporabe kreditnih kartica

- kreditne kartice – 3mil. britanskih građana
- 9 mil. transakcija mjesečno, 2500 zlouporaba
- izazov detekcije takvih transakcija
- menadžment -> angažirao 5 analitičara
- 60 varijabli za identificiranje sumnjivih transakcija
- suprotno predviđenom negativnom trendu za 2003 (20% povećanje zlouporabe kartica):
  - smanjen broj štetnih transakcija za 55%
  - mjesečna ušteda 220.000 USD
  - prosječan gubitak u kartičnoj industriji je 0,27% ukupnog prihoda, a gubitak HFC Bank je 0,10%.

Case Study 2: First National Bank povećava efikasnost marketinga

- jedna od najvećih afričkih banaka, 3.2 mil. Klijenata
- loše usmjerene marketinške kampanje (targetiranje)
- analiza podataka i personalizacija ponuda
- odaziv na kampanju 9%
- otkriće: najprofitabilniji klijenti (5%) nemaju značajan broj ključnih proizvoda banke
- prve marketinške kampanje vratile ulaganje uz profit 3000%

Zaključak

- Mnoge dragocjene informacije ostaju zauvijek „zatočene” unutar TB podataka -> data mining
- Generiranje sve većih količina podataka -> potrebama za *efikasnim* metodama i alatima
- Oracle R Enterprise ->
  - Koristi benefite Oracle DB Management sustava
  - Zadovoljava R korisnike
  - Oracle Support
- Gartner-ovo istraživanje -> analiza podataka - potreba i potencijal budućnost

Izvori

- http://www.oracle.com/technetwork/database/options/advanced-analytics/r-enterprise/index.html
- http://www.r-project.org/
- Data mining in banking, doc. dr. Mirjana Pejić Bach
- http://en.wikipedia.org/wiki/Receiver_operating_characteristic

Hvala na pažnji !

Q & A

ORACLE Gold Partner
Specialized
Data Warehousing

ORACLE Gold Partner
Specialized
Oracle Business Intelligence Foundation

ORACLE Gold Partner
Specialized
Oracle Application Development Framework