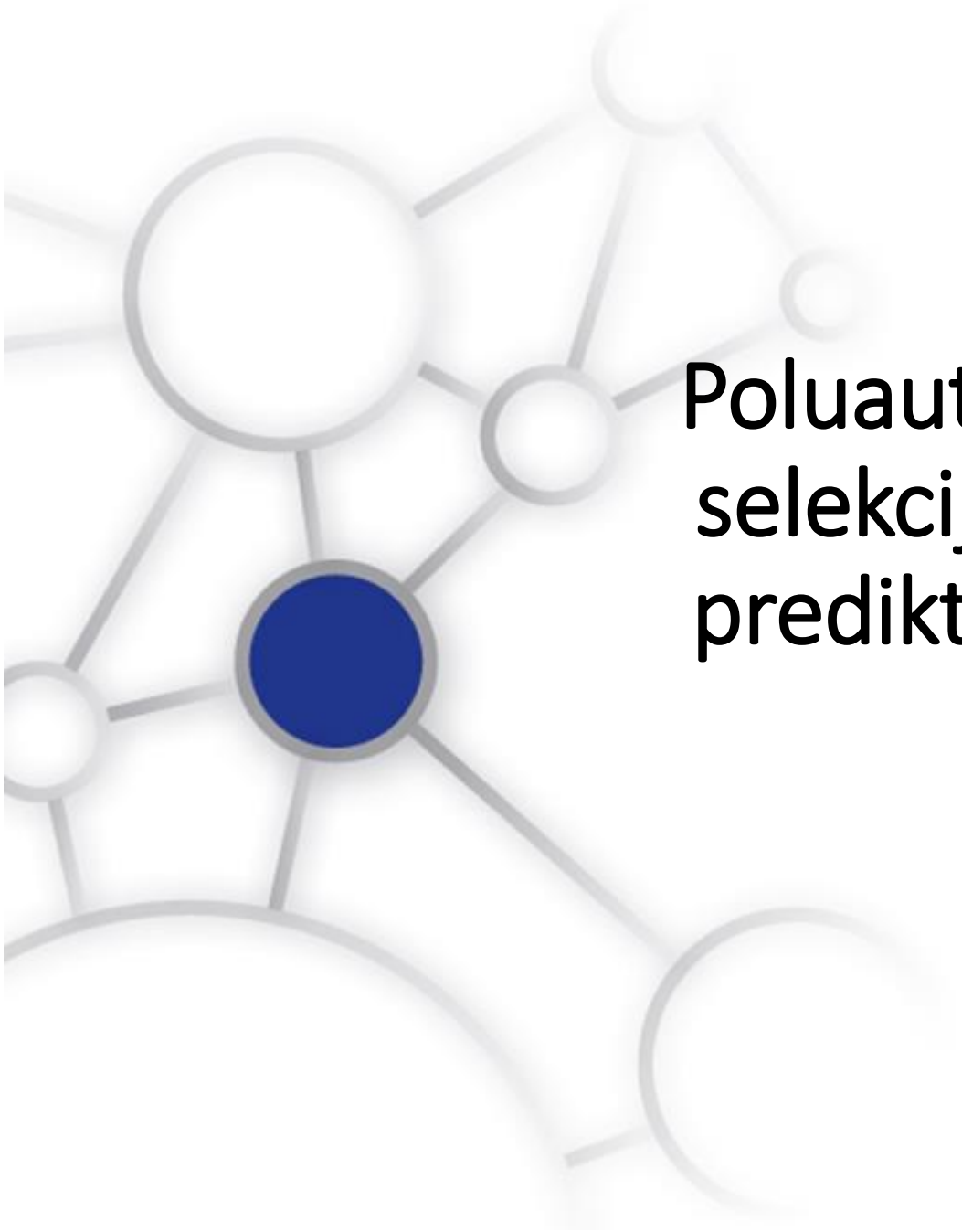


Poluautomatizirana selekcija varijabli u prediktivnoj analizi

Multicom

- **Glavna područja ekspertize:**
  - Data Mining
  - Obračun i naplata (**Billing**)
  - Upravljanje matičnim podacima (**MDM**)
  - Skladišta podataka (**DWH**) i Poslovna Inteligencija (**BI**)
  - **B2B**
  - Upravljanje korisničkim procesima (**CRM**)

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

Cro Cert ISO 9001:2008 certified

Prediktivna analiza

predviđanje budućnosti?

Cross-sell / up-sell

Otkrivanje prijevara

Churn

Poznatiji alati:

- Mathematica
- Mathlab
- Oracle Advanced Analytics
- Orange
- R
- RapidMiner
- SAP
- SAS

Prediktivna analiza pomoću R-a

R je Open Source jezik i okolina za statističke proračune i grafiku
Stvoren 1994 kao alternativa SAS-u i SPSS-u
Preko 2 milijuna R korisnika u svijetu
Tisuće open source paketa na CRAN mreži
CRAN – Comprehensive R Archive Network

CRAN

Mirrors
What's new?
Task Views
Search

About R

R Homepage
The R Journal

Software

R Sources
R Binaries
Packages
Other

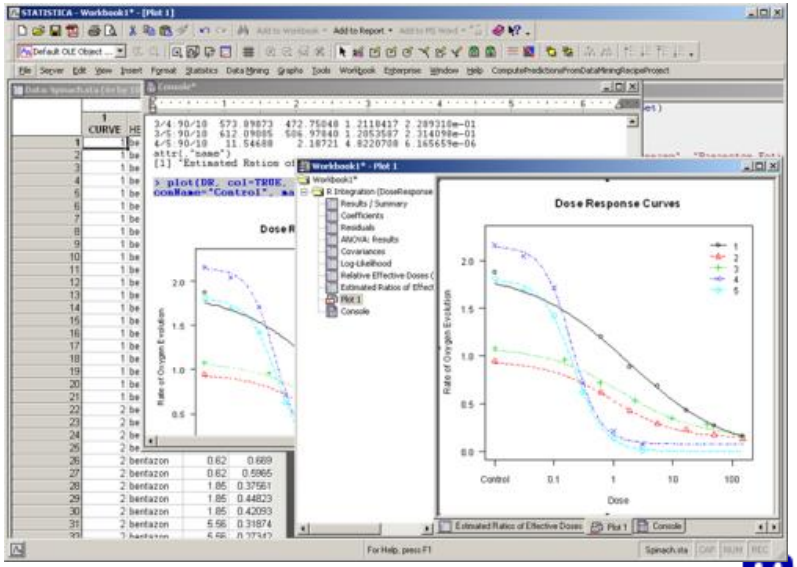
Documentation

Manuals
FAQs
Contributed

CRAN Task Views

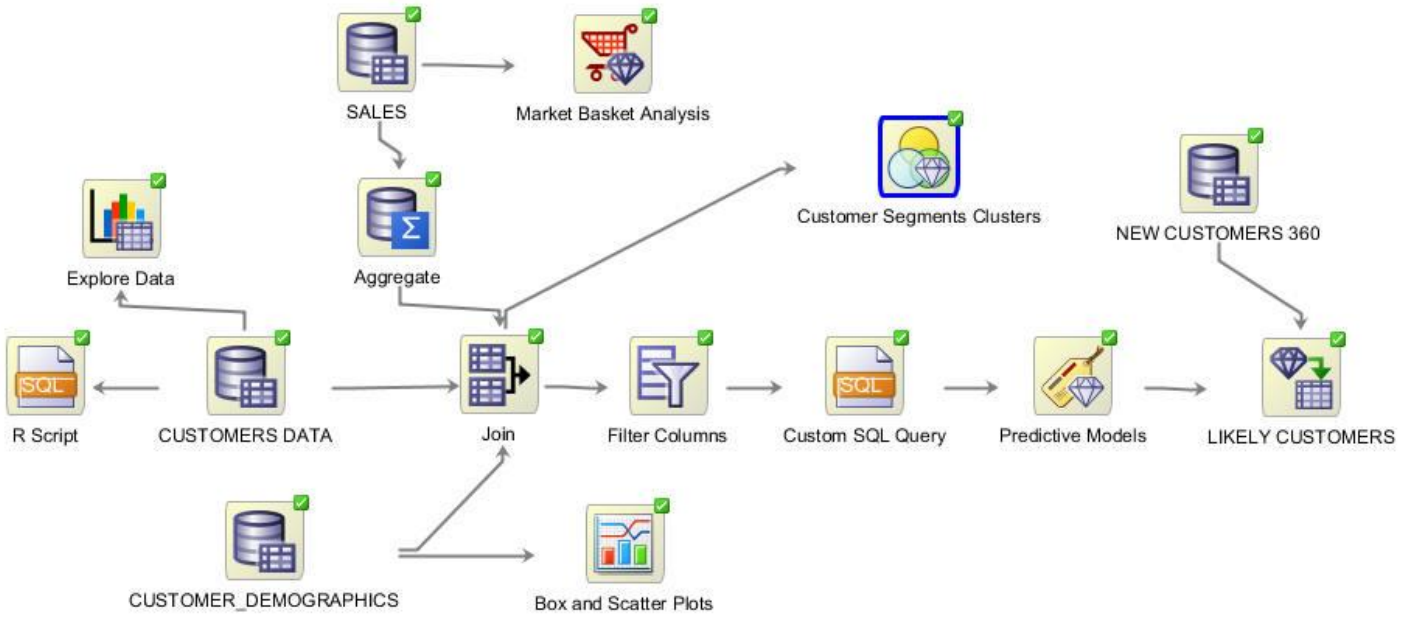
Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods

- Statistika
- Grafika
- Data mining
- R Enterprise



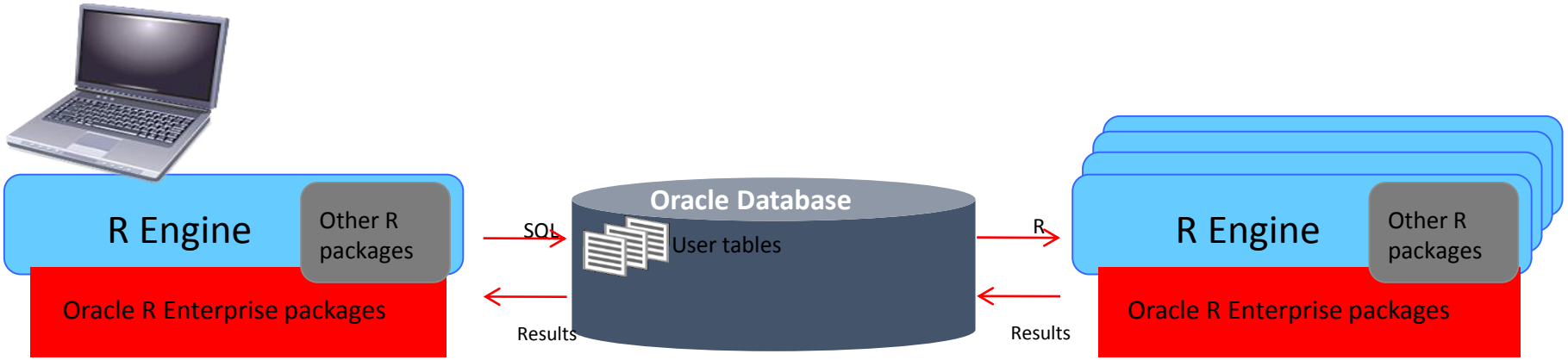
Prediktivna analiza pomoću ODM-a

- Skalabilna unutar baze prediktivna analiza
- Jednostavno za korištenje
- Brzi i pouzdani rezultati



Oracle R Enterprise (SAVRŠENI SPOJ?)

Integrira open source programski jezik R
unutar Oracle baze podataka



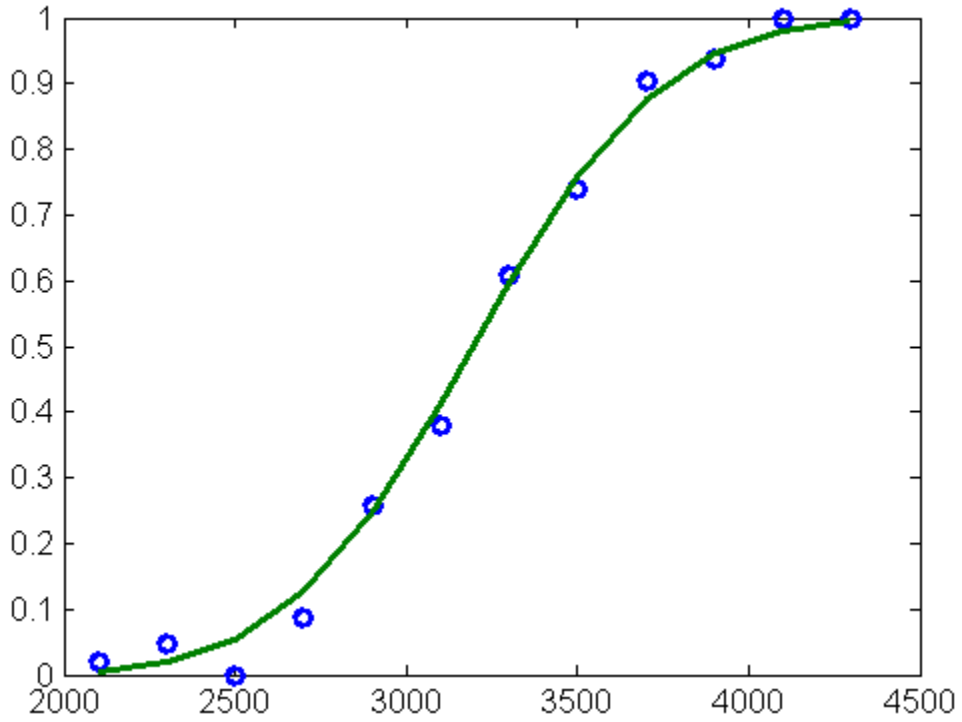
R na desktop računalu

Sustav izračuna unutar baze

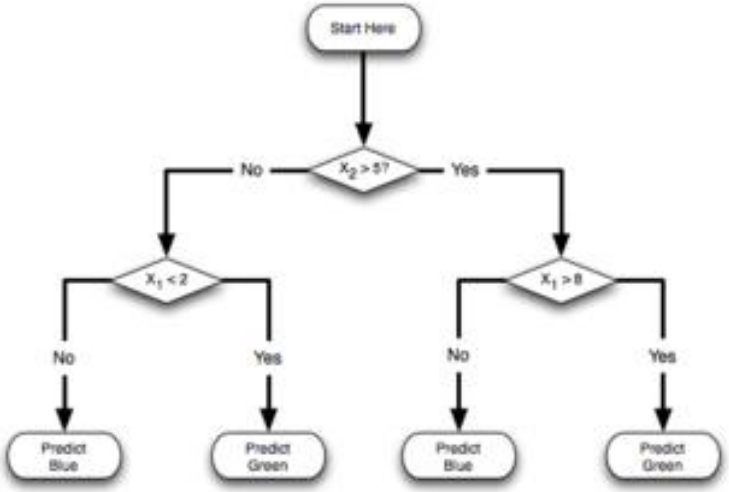
R sustav prilagođen od strane Oracle-a

Koristljivi algoritmi

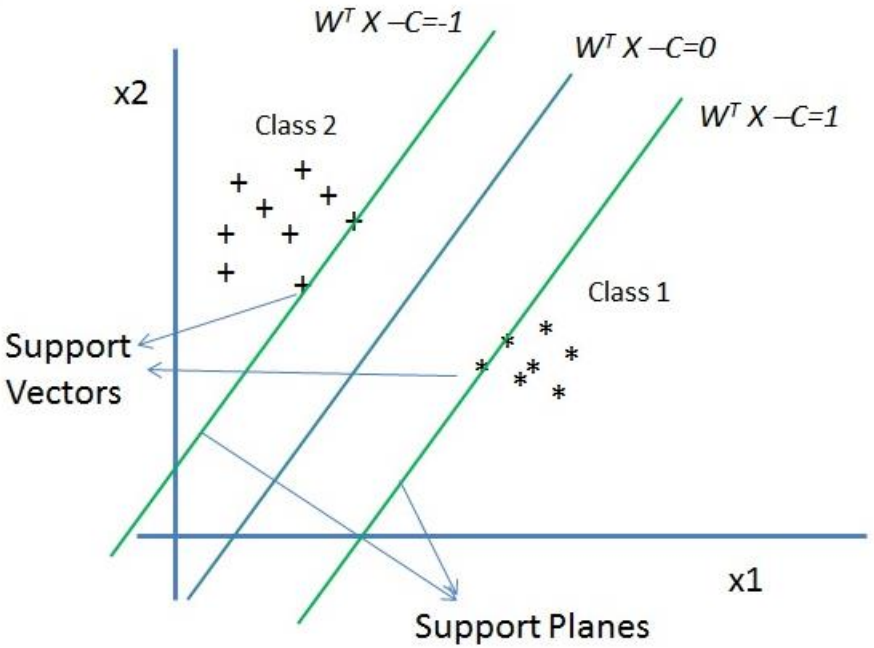
- GLM – general linear model



Decision tree



Support vector machine

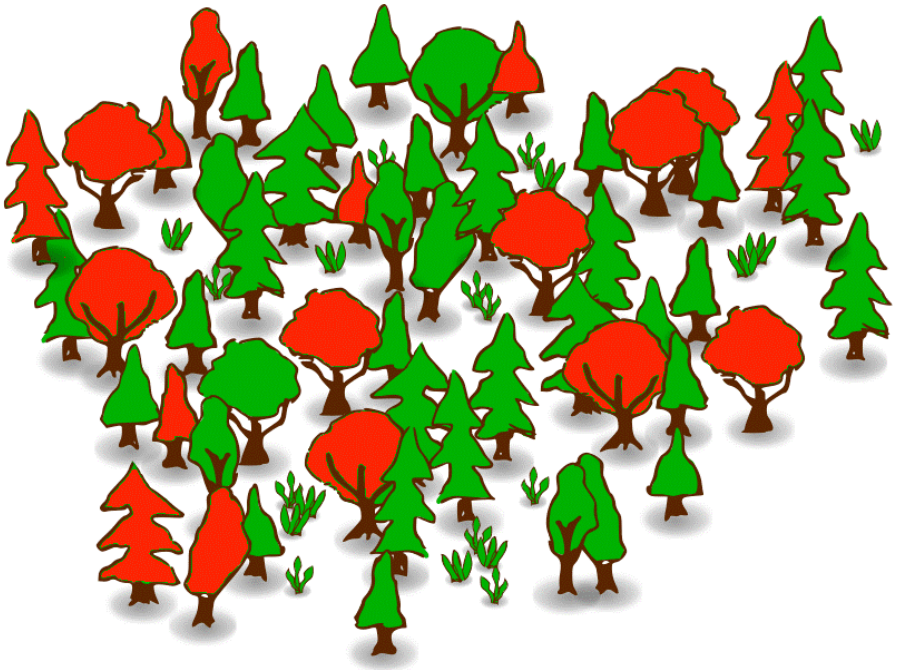


Naive Bayes

$$\log p(C_k|\mathbf{x}) \propto \log\left(p(C_k)\prod_{i=1}^{n} {p_{ki}}^{x_i}\right)$$
$$= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki}$$
$$= b + \mathbf{w}_k^\top \mathbf{x}$$

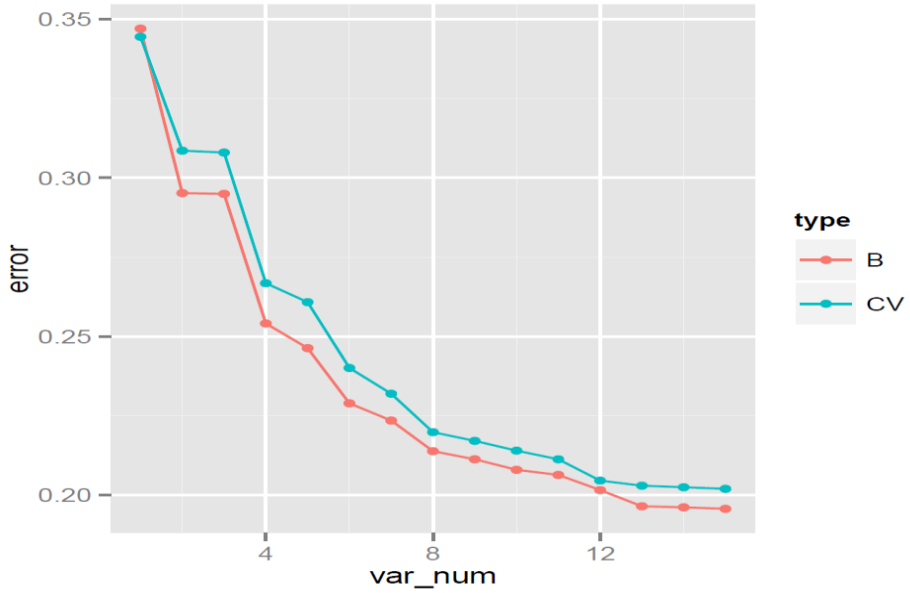
$$b = \log p(C_k)$$
$$w_{ki} = \log p_{ki}$$

Random forest?



m.builder rješenje

- Sampliranje (random, stratified)
- Attribute importnace
- Iterativni odabir varijabli
- izbor algoritma
- Oracle R Enterprise – pouzdano i brzo rješenje



Ulazne vrijednosti

```
R> results <- ore.doEval(FUN.NAME = "m.VarSelect",
                         model = "GLM",
                          build = "TABLE",
                         attribute_importance=TRUE,
                         tgt = "TGT",
                         columns_to_skip=c("USER_ID","SERIAL_NUMBER"),
                         sample_type="ST",
                         c_per=0.3,
                         ore.connect=TRUE )
```

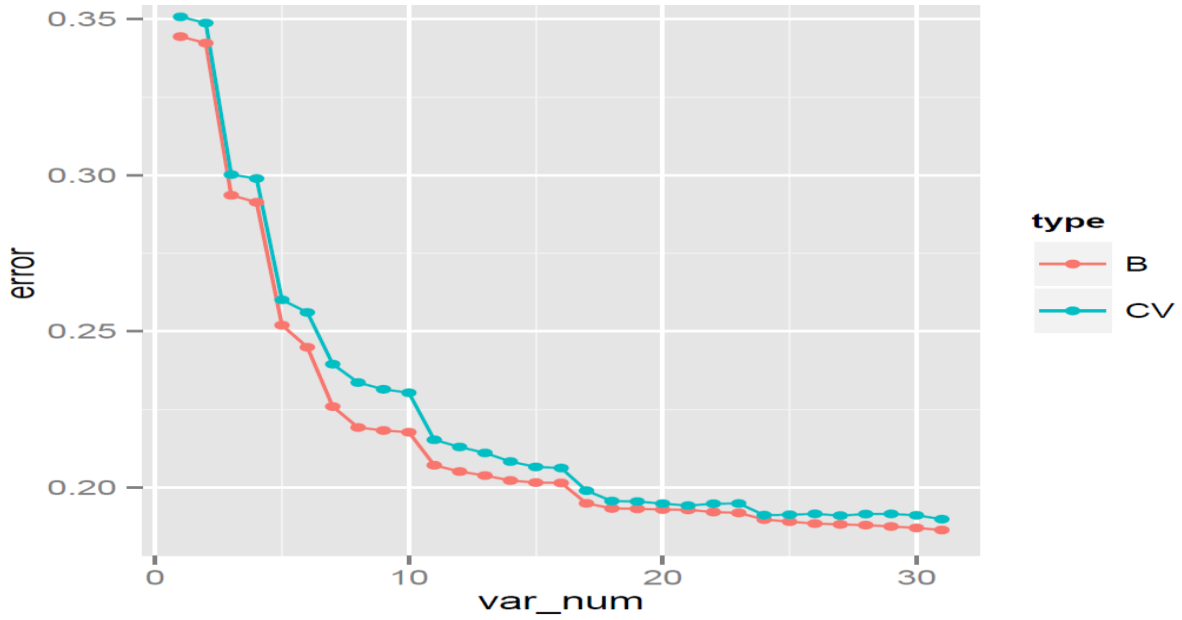
build	Set na kojem se gradi model za testiranje korisnosti varijabli
model	Izbor vrste modelal: GLM (general linear model), DT (decision tree), NB (Naive Bayes) or SVM (support vector machine)
tgt	Ime stupca koji označava ciljnu varijablu
cross_valid	Ime seta za krosvalidaciju
attribute_importance	Izbor dali da se uzimaju u obzir samo varijable koji imaju attribute importance veći od nule (TRUE/FALSE)
columns_to_skip	Koje varijable bi se trebale preskočiti u procesu odabiranja
c_per	Koliko veliki dio BUILD seta treba uzeti za krosvalidaciju, ukoliko set za krosvalidaciju nije dezigniran (0<=c_per<=1).
sample_type	Koja vrsta sempliranja bi se trebala koristiti pri uzimanj krosvalidacijskog seta iz build seta: ST (stratified) or R (random sampling).

Izlazne vrijednosti

sample_size	Velićina BUILD seta
var_number	Broj varijable po redu koja je automatski predviđena da će biti korisna u prediktivnom modelu (ΔAUC >0, attribute importance >0)
var_number_full	Broj varijable po redu koja je testirana (attribute importance >0)
added_variable	Ime varijable pridodane za zadnju iteraciju prediktivnog modela (ΔAUC >0, attribute importance >0)
added_variable_full	Ime varijable pridodane za zadnju iteraciju prediktivnog modela (AUC >0)
R2_build	R na kvadrat modela na BUILD setu
R2_test	R na kvadrat modela na krosvalidacijskom setu
auc_build	Površina ispod ROC krivulje za trenutni prediktivni model na BUILD setu (ΔAUC >0, attribute importance >0)
auc_build_full	Površina ispod ROC krivulje za trenutni prediktivni model na BUILD setu sa restrikcijom za attribute importance >0.
auc_test	Površina ispod ROC krivulje za trenutni prediktivni model na krosvalidacijskom setu
m_start	Vrijeme pokretanja gradnje zadnjeg prediktivnog modela

Rezultati

- Značajna ušteda vremena
- Pojednostavljivanje prediktivnog modela
- Lakša interpretacija modela
- Smanjena mogućnost „overfittinga”



...Hvala!
Pitanja?

Ivan.osman@multicom.hr