Dobývání z dat (data mining)

Praxe v programu RapidMiner

Metodologie CRISP-DM (CRoss-Industry Standard Process for Data Mining)

1. porozumění problematice (Business Understanding)
2. porozumění datům (Data Understanding)
3. příprava dat (Data Preparation)
4. modelování (Modeling)
5. vyhodnocení výsledků (Evaluation)
6. využití výsledků - implementace vytvořeného modelu (Deployment)

Pozn.:
Fáze č. 4 – Dolování dat (data mining)

Využití data miningu v marketingové kampani

1. fáze

POROZUMĚNÍ PROBLEMATICE

Úloha

- přímý marketing – prodej finančního produktu (termínovaný vklad)
- jsou k dispozici data o předchozí marketingové kampani
- analýza dat pomocí metod DM a zvýšení efektivity prodeje v nové kampani

Přímý marketing

výhody

- přesné zacílení,
- vysoká efektivita,
- okamžité a jednoznačné výsledky,
- možnost testování nejlepších řešení,
- prognózování výsledků kampaně

nevýhody

- vysoké náklady
  - poštovné,
  - výroba zásilek,
  - cena telefonních hovorů,
  - náklady spojené s využitím call centra

Základní pojmy a příklad kampaně

- **response rate -** 1 až 10%
- **converse rate -** menší než response rate
- **return of investment (ROI) = výnosy / investice * 100 -** návratnost investice (chceme maximalizovat)

Příklad:
cena oslovení potencionálního klienta = 100Kč
výnos z klienta = 1 000Kč

Oslovím **10 000** klientů
výdaje = **1 000 000** Kč,
při converse rate = **5%** mám
příjmy = 0.05 * 10 000 * 1 000 = **500 000**
výsledek kampaně = ROI = 50%

Je potřeba kampaň lépe připravit!!

=>

2. fáze

POROZUMĚNÍ DATŮM

Seznámení s daty

- k dispozici reálná data z portugalské banky (získáno z UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)
- zhruba 45000 záznamů z marketingové kampaně

1 záznam obsahuje:

- 16 vstupních atributů – informace o klientovi - viz další slide
- 1 výstupní (klasifikační) atribut – údaj o tom, zda si klient koupil finanční produkt

- v ukázce pracujeme s 10%ním vzorkem dat

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no

Atributy – obecné informace o klientovi

1 - **age** (numeric)

2 - **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self- employed", "retired", "technician","services")

3 - **marital** : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - **education** (categorical: "unknown","secondary","primary","tertiary")

5 - **default**: has credit in default? (binary: "yes","no")

6 - **balance**: average yearly balance, in euros (numeric)

7 - **housing**: has housing loan? (binary: "yes","no")

8 - **loan**: has personal loan? (binary: "yes","no")

age	job	marital	education	default	balance	housing	loan
30	unemployed	married	primary	no	1787	no	no
33	services	married	secondary	no	4789	yes	yes
35	management	single	tertiary	no	1350	yes	no

Atributy – informace o kontaktu v aktuální kampani + ostatní atributy

9 - **contact**: contact communication type (categorical: "unknown","telephone","cellular")

10 - **day**: last contact day of the month (numeric)

11 - **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - **duration**: last contact duration, in seconds (numeric)

Ostatní atributy:

13 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - **previous**: number of contacts performed before this campaign and for this client (numeric)

16 - **poutcome**: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

contact	day	month	duration	campaign	pdays	previous	poutcome
cellular	19	oct	79	1	-1	0	unknown
cellular	11	may	220	1	339	4	failure
cellular	16	apr	185	1	330	1	failure

3. fáze

PŘÍPRAVA DAT

Software a metody pro analýzu dat

software: 

(volně ke stažení)

Metody:

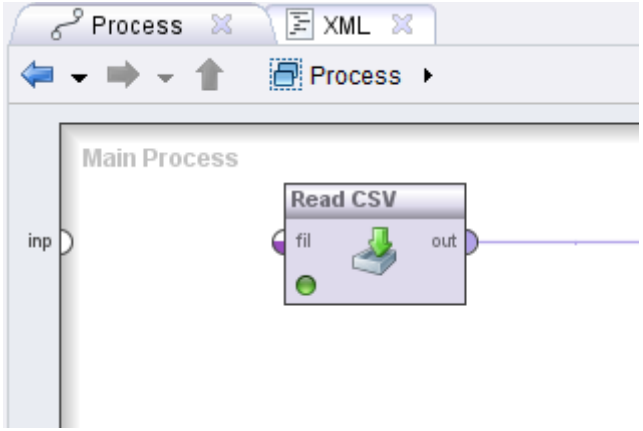
1. rozhodovací strom
2. pravidla

Import dat

- data – uložena v souboru bank.csv

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	yes

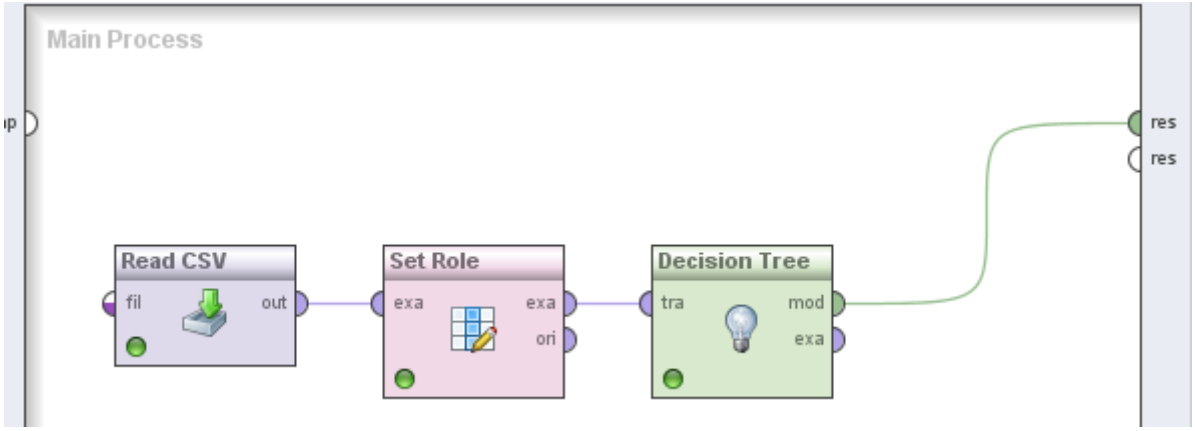
- operátor readCSV -> import dat

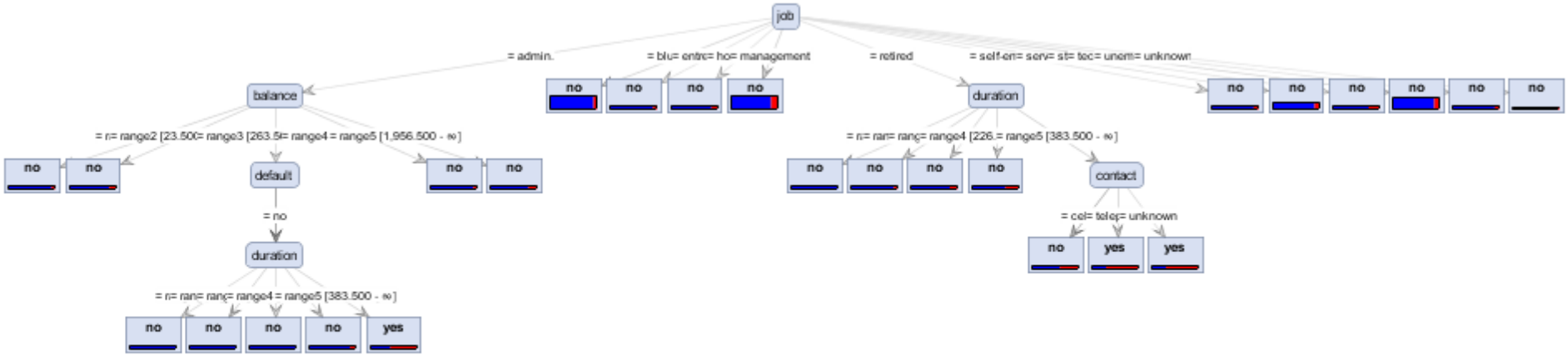


4. fáze

MODELOVÁNÍ

Modelování pomocí rozhodovacích stromů

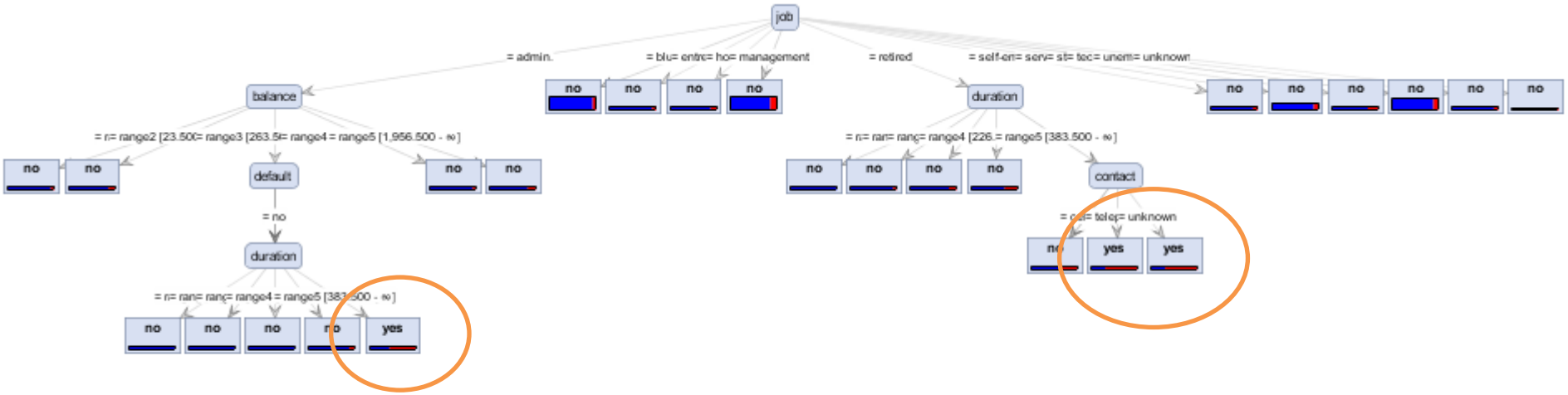


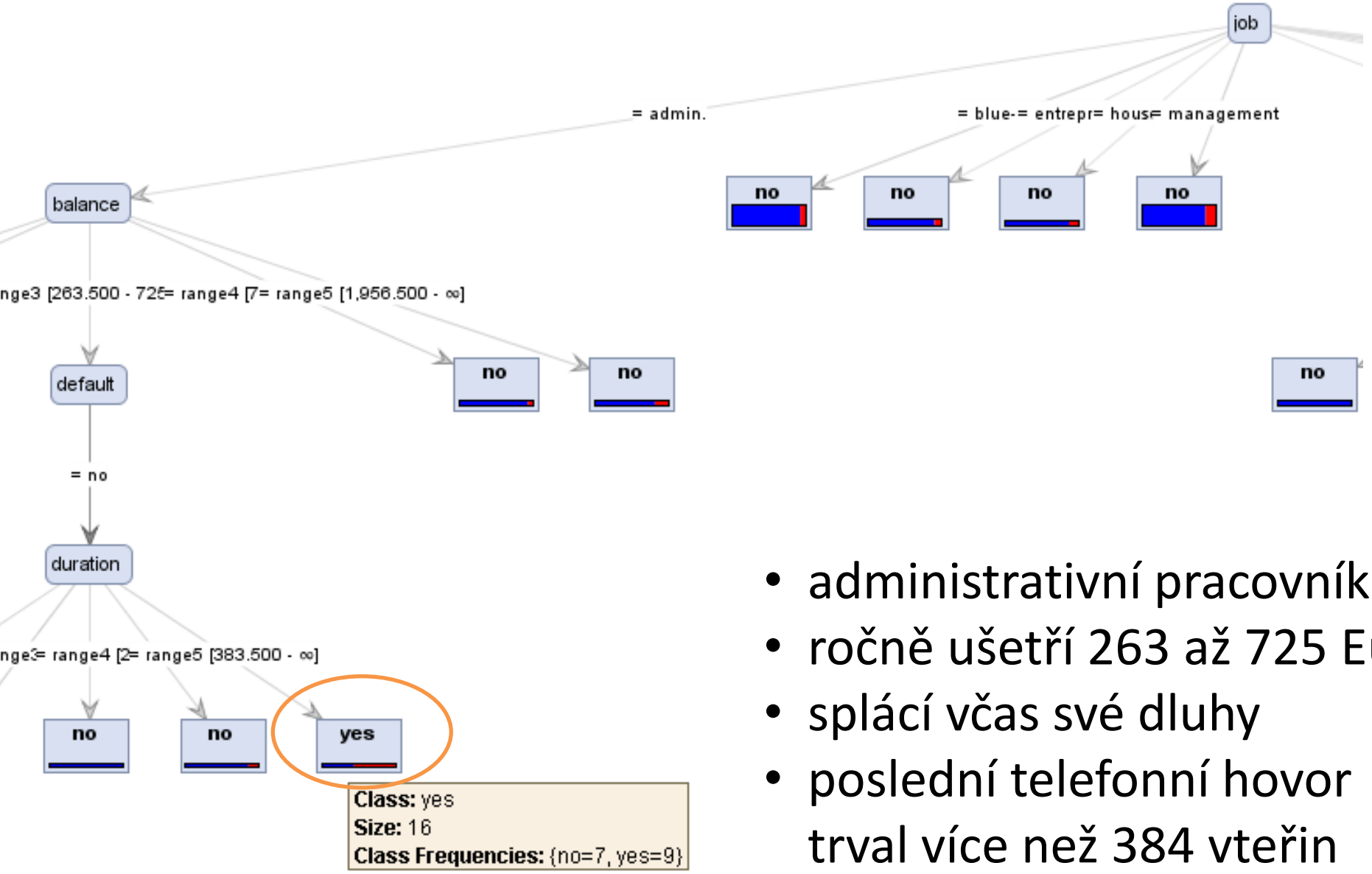


5. fáze

VYHODNOCENÍ VÝSLEDKŮ

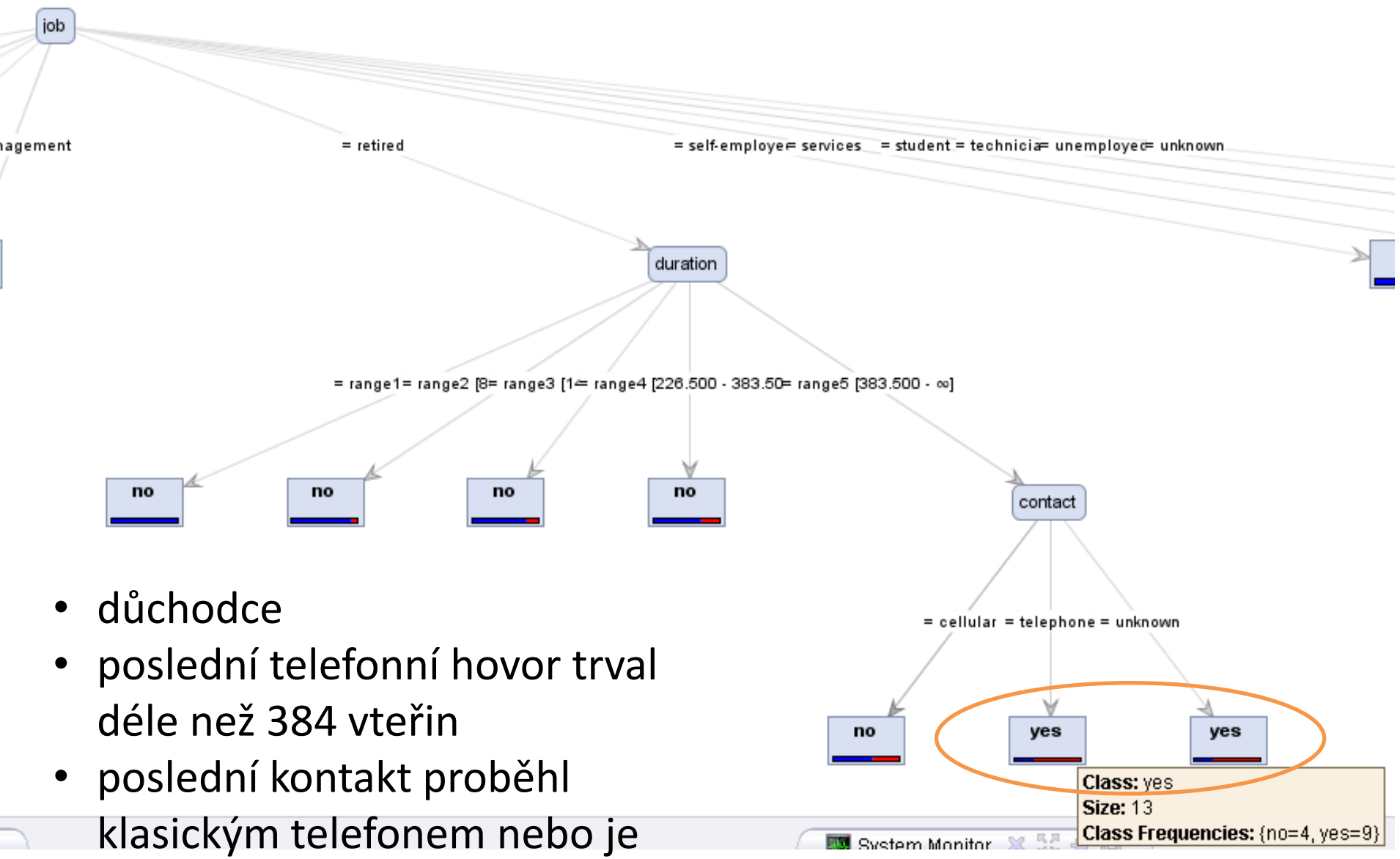
Rozhodovací strom





- administrativní pracovník
- ročně ušetří 263 až 725 Eu
- splácí včas své dluhy
- poslední telefonní hovor trval více než 384 vteřin

Strom – levá část



- důchodce
- poslední telefonní hovor trval déle než 384 vteřin
- poslední kontakt proběhl klasickým telefonem nebo je neznámo jak

Strom – pravá část

6. fáze

VYUŽITÍ VÝSLEDKŮ - IMPLEMENTACE VYTVOŘENÉHO MODELU

předpokládaný R(

Úspěch		Neúspěch
9	I. skupinka (adm. prac.)	7
9	II. skupinka (důchodci)	4
7	III. skupinka (důchodci)	3
-----		----
25	celkem	14

Vše za předpokladu, že by se při nové kampani zákazníci chovali stejně jako v předchozí!!

converse rate = 25/(25 + 14) = **64%**

příjmy = 25 * 1 000 = 25 000 Kč
výdaje = 39 * 100 = 3 900 Kč
(na vzorku 4500 klientů)

ROI = 25 000 / 3 900 * 100 = **641%**

Na plném vzorku (45000 klientů): výnos kampaně 211 000 Kč

Pravidla

RuleModel

```
if duration = range1 [-∞ - 185.500] then no  (2205 / 63)
if contact = unknown then no  (602 / 58)
if previous = range1 [-∞ - 0.500] and housing = yes then no  (443 / 93)
if previous = range1 [-∞ - 0.500] and balance = range1 [-∞ - 444.500] then no  (224 / 53)
if poutcome = failure and balance = range1 [-∞ - 444.500] then no  (77 / 16)
if previous = range1 [-∞ - 0.500] and contact = cellular then no  (222 / 81)
if poutcome = failure and loan = yes then no  (18 / 1)
if housing = yes and poutcome = failure then no  (60 / 21)
if poutcome = other and housing = yes then no  (50 / 18)
if poutcome = success and housing = no then yes  (16 / 52)
if housing = no and poutcome = failure then no  (32 / 17)
if housing = no and loan = yes then no  (5 / 0)
if housing = yes and loan = yes then yes  (0 / 4)
if poutcome = other and campaign = range1 [-∞ - 2.500] then no  (19 / 11)
if contact = cellular and campaign = range2 [2.500 - ∞] then yes  (1 / 4)
if contact = cellular and education = tertiary then yes  (3 / 5)
if education = unknown then yes  (0 / 2)
if contact = telephone and education = secondary then no  (8 / 6)
if previous = range1 [-∞ - 0.500] and campaign = range2 [2.500 - ∞] then yes  (2 / 4)
if campaign = range2 [2.500 - ∞] then no  (2 / 0)
if education = tertiary and balance = range2 [444.500 - ∞] then yes  (2 / 3)
if previous = range2 [0.500 - ∞] and balance = range2 [444.500 - ∞] then yes  (3 / 4)
```

předpokládaný ROI

Úspěch	Neúspěch
52	16
4	0
4	1
5	3
2	0
4	2
3	2
4	3
-----	----
78	**27**

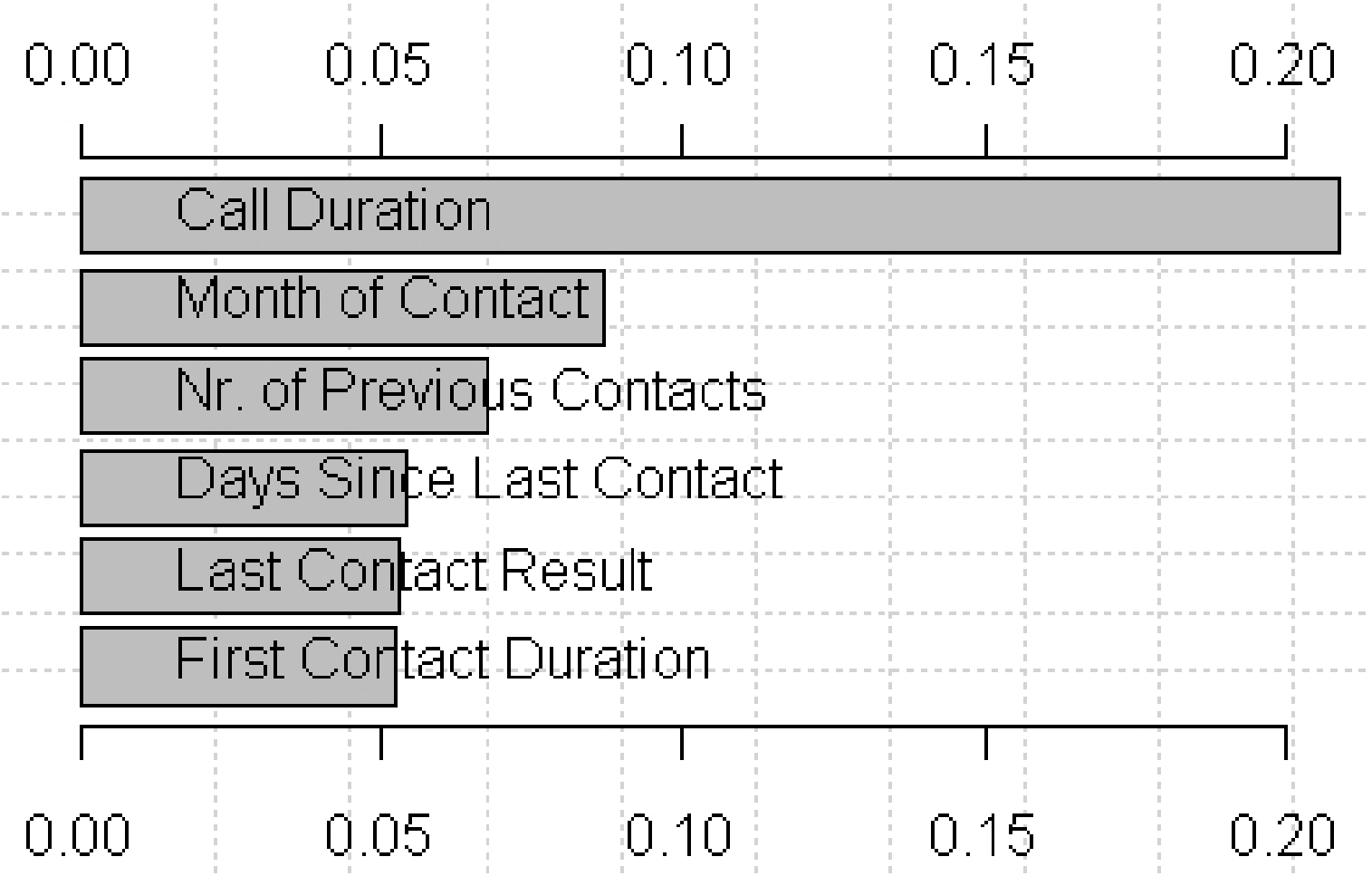
converse rate = 78/(78 + 27) = **74% >> 5%**

příjmy = 78 * 1 000 = 78 000 Kč
výdaje = 105 * 100 = 10 500 Kč
(na vzorku 4500 klientů)

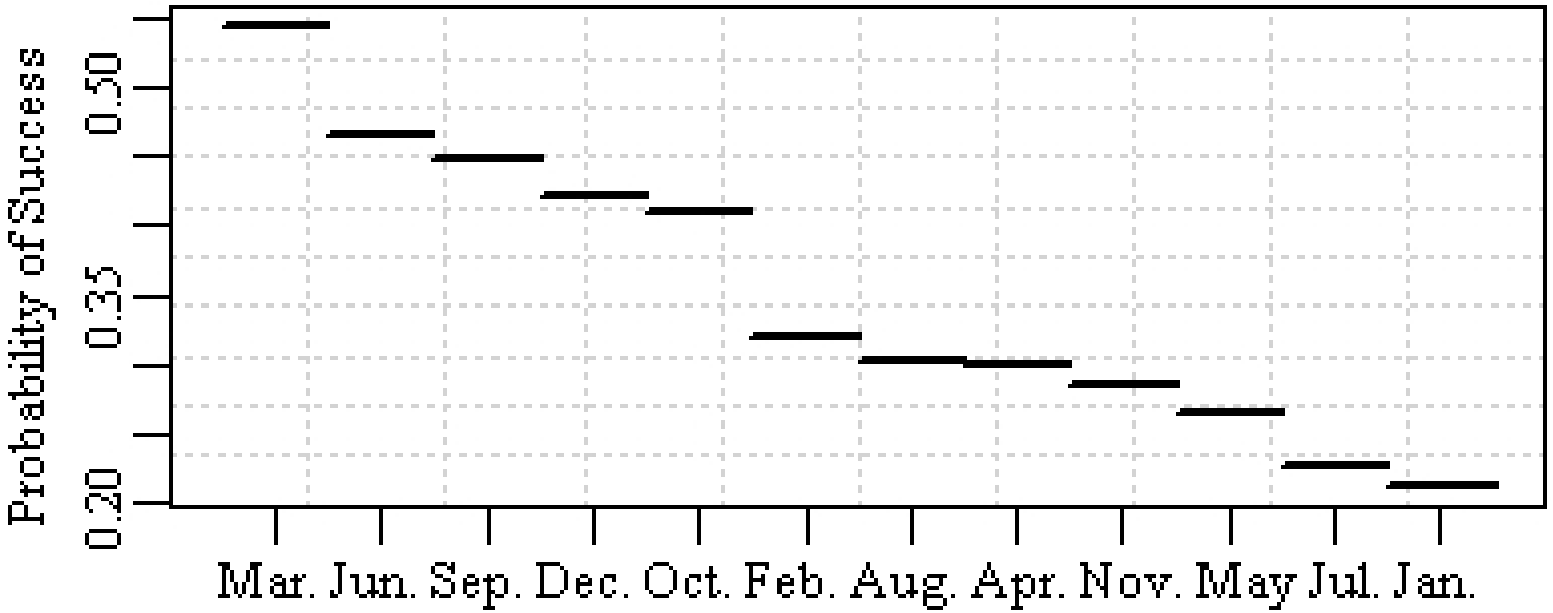
ROI = 78 000 / 10 500 * 100 = **743%**

Na plném vzorku (45000 klientů): výnos kampaně 675 000 Kč

Nejvíce relevantní atributy (dle SVM)



Vliv měsíce na úspěšnost kampaně (dle SVM)



Největší úspěch při oslovování potencionálních klientů je vždy na konci čtvrtletí

=>

je výhodné posunout kampaň právě na tyto měsíce

Děkuji za pozornost