DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

Machine-learning: základy

Jan Hučín

29. března 2019

Osnova

1. Co si pod tím představit
2. Základní typy ML
3. Modelování
4. Vyhodnocení modelu
5. Nasazení modelu

PROFINIT

Účel ML

automatický proces

ze známých vlastností jednotky (člověk, objekt):

- predikce události
  - přestane splácet
  - přežije následující rok
- predikce hodnoty veličiny
  - výše majetku, počet a věk dětí
  - výsledek v testu
- klasifikace
  - je to spam?
  - sestavení tříd podobných případů
- výpočet charakteristiky
  - redukce dimenzionality

PROFINIT

Názvosloví ML

predikovaná/odhadovaná veličina:

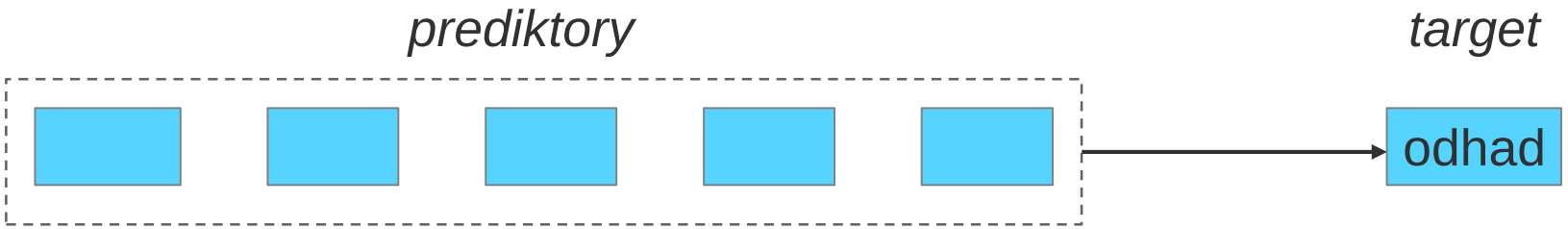
- target
- response
- vysvětlovaná proměnná, cílová proměnná
- závislá proměnná

vstupní veličiny:

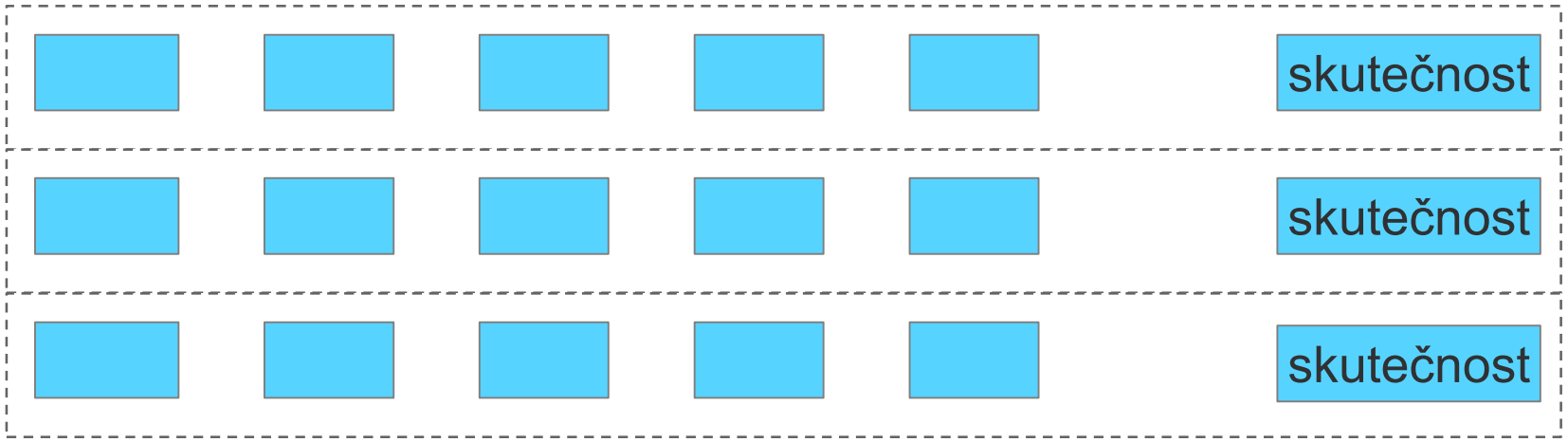
- prediktory, regresory
- příznaky
- vysvětlující proměnné
- nezávislé proměnné

PROFINIT

Typy ML



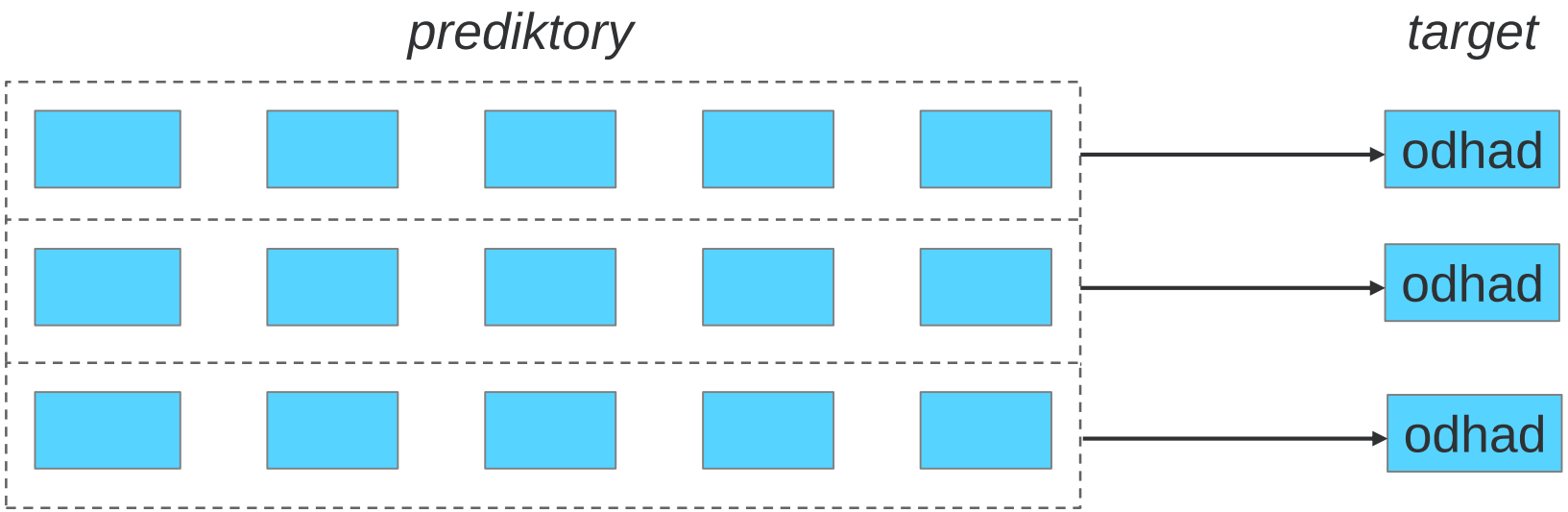
Známé případy (**trénovací množina**)



- **supervised learning** – učení s učitelem
- odhad vychází ze souvislostí pozorovaných na známých případech

PROFINIT

Typy ML

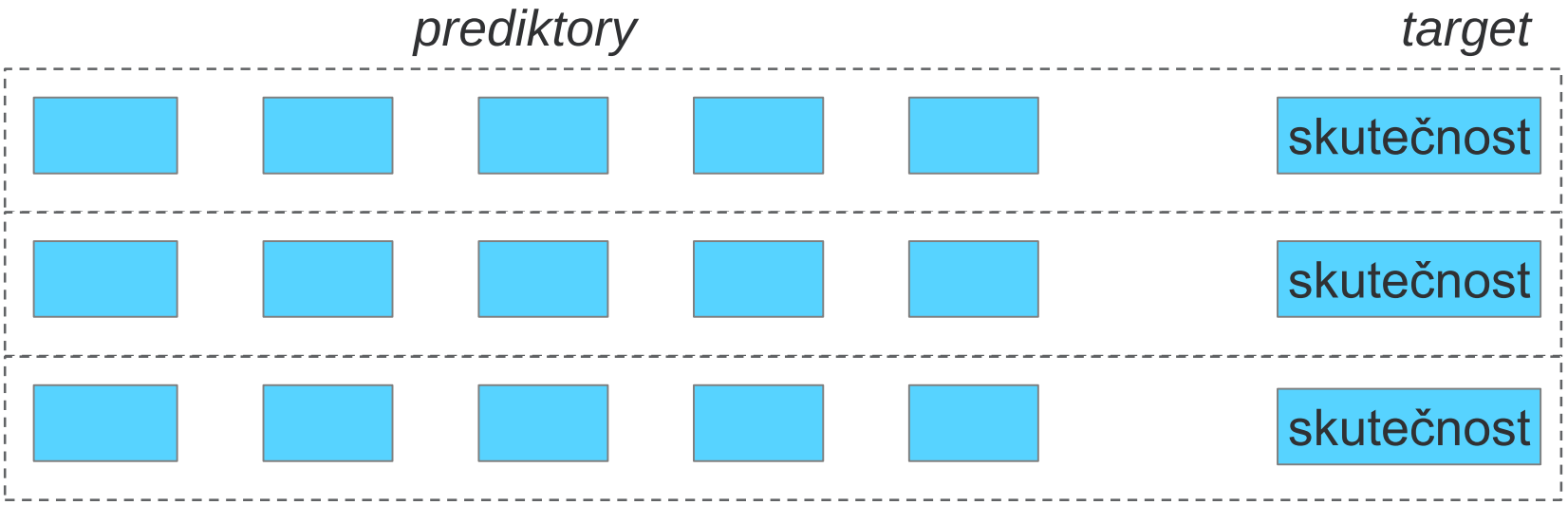


- **unsupervised learning** – učení bez učitele
- odhad vychází ze vzájemných vztahů jednotek

PROFINIT

Supervised learning: modelování

Modelování



hledáme, jak ze vstupní informace odvodit vlastnost:

- matematický vzorec (např. regrese, lineární model)
- posloupnost rozhodnutí (např. rozhodovací strom)
- iterativní algoritmus (např. neuronová síť, gradient boosting)
- podobnost (např. nearest neighbours)

PROFINIT

Modelování – příklady

- odhad celkovým průměrem
- odhad skupinovým průměrem (muži/ženy; mladí/staří)
- odhad funkčním vztahem (hmotnost jako funkce výšky)
- odhad kombinací efektů:
  - celkový průměr + efekt pohlaví + efekt bydliště + funkce věku + ...
- → lineární model (regrese)

PROFINIT

Jak dobrý je můj model?

odhad číselné hodnoty

› MSE (mean squared error) = průměr čtverců odchylek

predikce události

› confusion matrix

	predikce ano	predikce ne
skutečnost ano	true positive	false negative
skutečnost ne	false positive	true negative

PROFINIT

Jak dobrý je můj model?

Problém:

Metriky počítám na datech, která jsem použil k sestavení modelu.

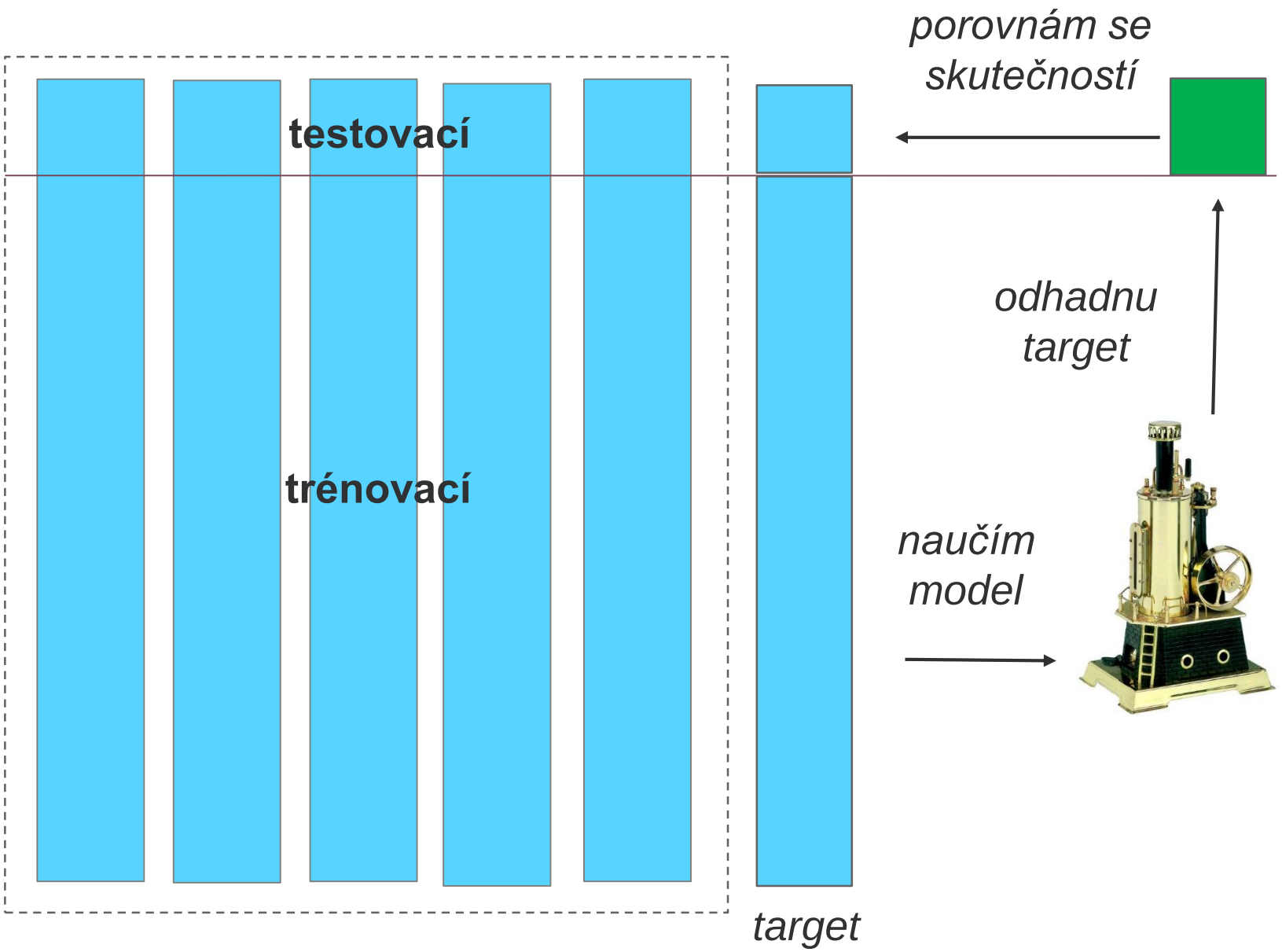
- model si mohu hodně ohnout podle dat
- chybu tím zmenším
- odhad na nových datech ale může být zkreslený
- **overfitting** (přeučení modelu)

Řešení = cross-validace

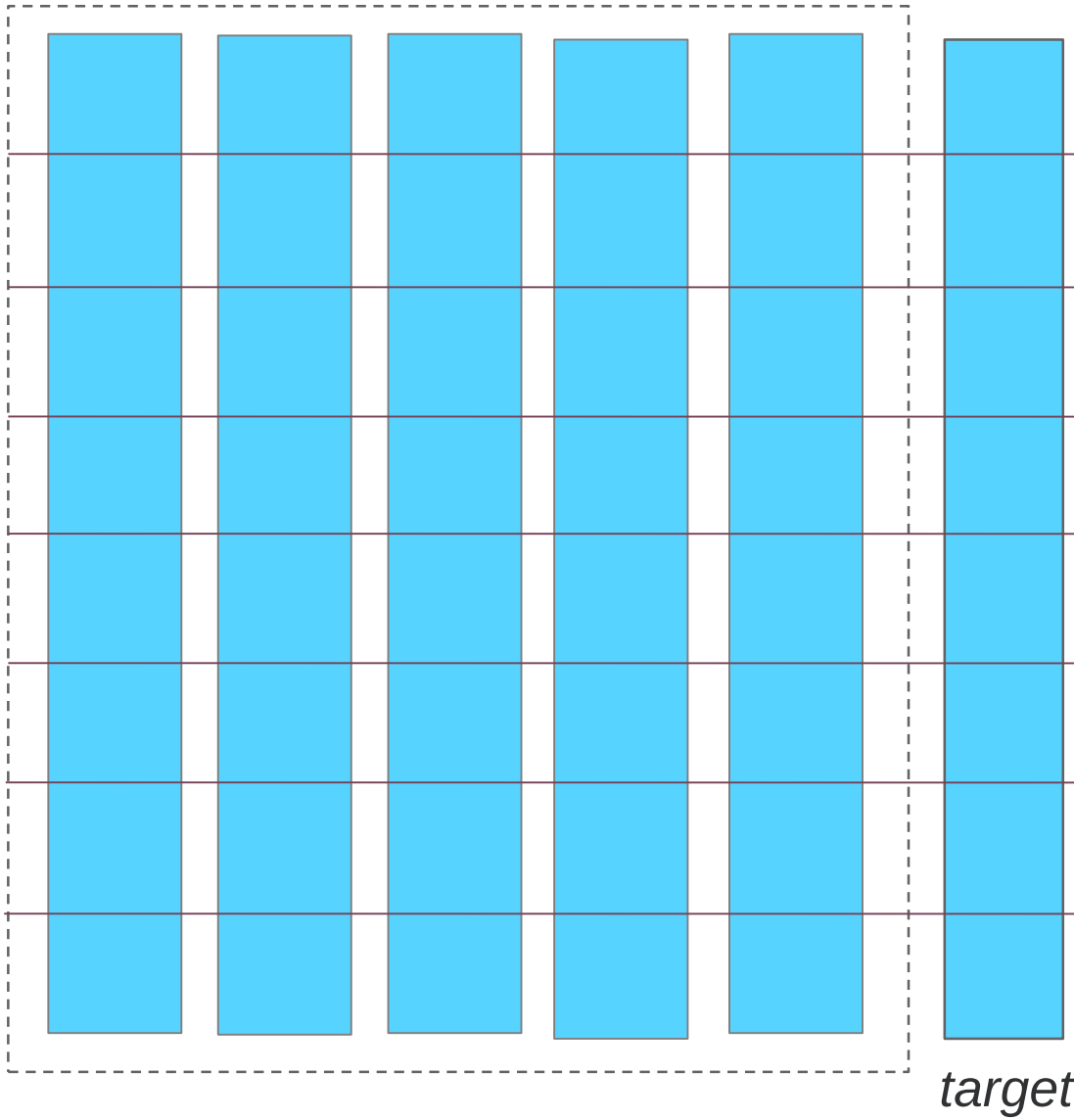
- z dat oddělím malou část – **testovací množina**
- zbytek použitý jako trénovací → sestavení a naučení modelu
- odhady pro testovací množinu pomocí naučeného modelu
- porovnání odhadů a skutečnosti na testovací množině
- opakujeme pro jiné dělení dat

PROFINIT

Cross-validace



Cross-validace



- dělení opakujeme
- každý řádek je (aspoň) jednou v testovací množině a mnohokrát v trénovací

PROFINIT

Nasazení modelu

- Model sestavím a naučím na všech datech.
- Odhad na nových datech pomocí modelu:
  - bodový odhad
  - intervalový odhad (vezme se v úvahu chyba z cross-validace)

PROFINIT

Díky za pozornost

PROFINIT

Profinit, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU