Manažer modelování kreditních rizik

část I: **pravděpodobnost selhání a ztráta ze selhání**

Vojtěch FILIPEC

whoami

☢ jaderná fyzika

🏢 profesní zkušenosti:

- banky: vývoj a validace modelů pro řízení kreditního rizika, monitoring, automatizace…
- fintech, farma, FMCG: “analytik - konzultant”

🎯 data scientist

obsah dopolední části kurzu

1. komponenty očekávané ztráty: PD, LGD, EAD
2. modelování PD: logistická regrese, klasifikační stromy, náhodné lesy
3. modelování LGD: vážené průměry, lineární regrese, regresní stromy

Co se chcete dovědět?

Proč odhadovat PD a LGD?

část 1: komponenty očekávané ztráty

$$EL = PD \times LGD \times EAD$$

[CZK] [%] [%] [CZK]

EL … očekávaná ztráta, expected loss

PD … pravděpodobnost selhání, probability of default

LGD … loss given default, ztráta z defaultu

EAD … exposure at default, expozice v čase defaultu

Než se pustíme do odhadu

dostupnost a relevance historických dat

horizont predikce

kalibrace: Point-in-Time vs. Through-the-Cycle

očekávaná vs. neočekávaná ztráta: opravné položky a kapitál

dokumenty Basel II a III: standardizovaný nebo IRB přístup

Jak odhadovat PD?

část 2: logistická regrese, klasifikační stromy, náhodné lesy

Strategie odhadu defaultu

klasifikace vs. regrese

diskrétní veličina × modelování spojité pravděpodobnosti (PD)

trénovací vzorek vs. horizont odhadu PD:

- co nejnovější data, avšak
- dostatečně stará na to, abychom věděli, zda default nasta
- nové proměnné ve starých datech?!

O modelech (statistické minimum)

spojité, ordinální, nominální proměnné

trénování a validace, overfitting, chybu měříme na **validačním** vzorku

historická vs. budoucí data: stabilita populace, population stability index

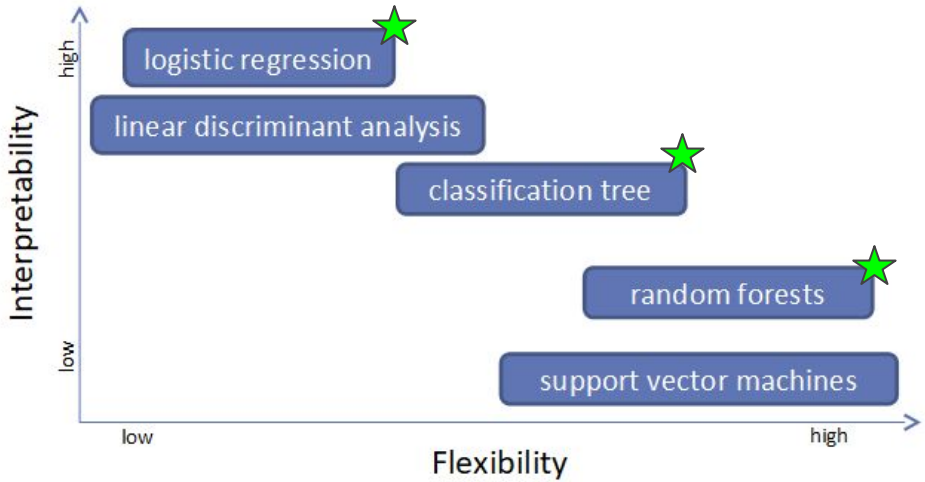
O statistických modelech

spojité, ordinální, nominální proměnné

trénování a validace, overfitting, chybu měříme na **validačním** vzorku

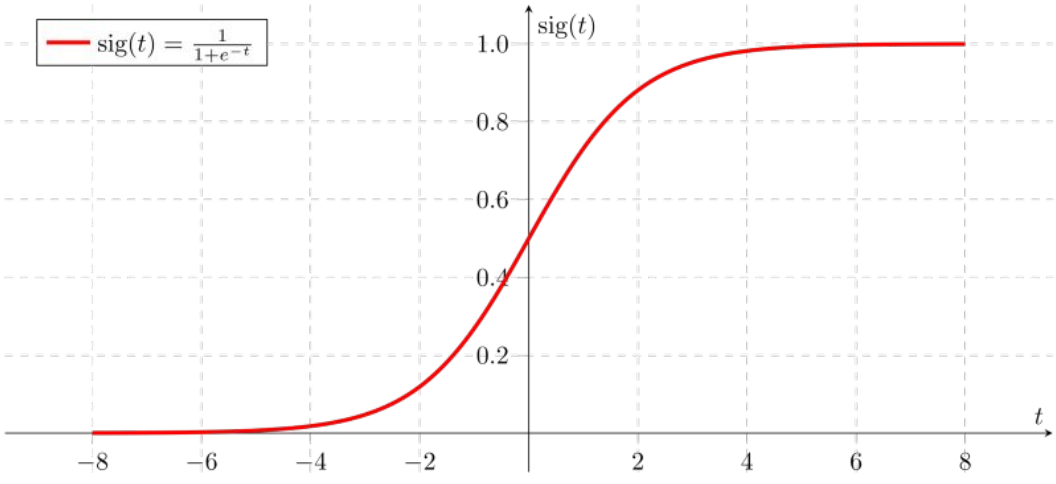
historická vs. budoucí data: stabilita populace, population stability index

komplexita vs. interpretabilita



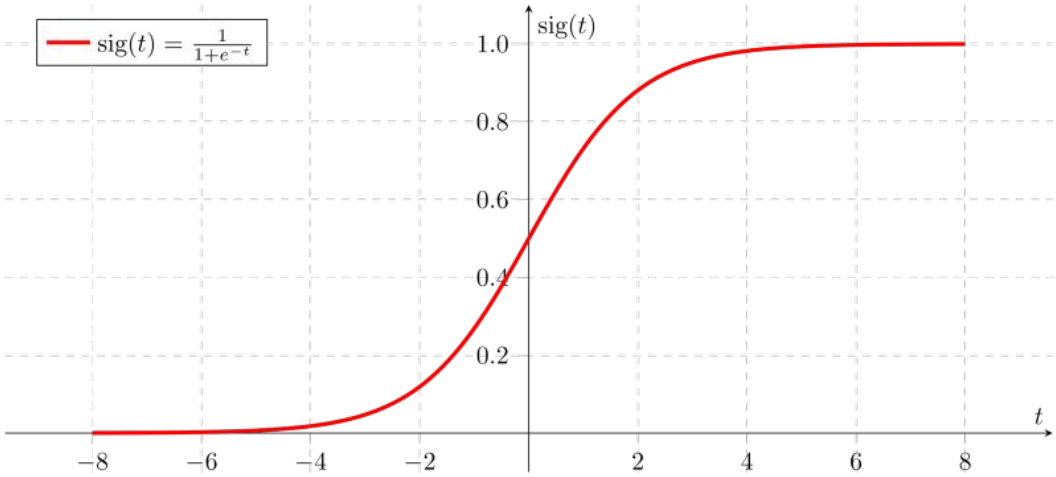
(graphics based on James, Witten, Hastie, Tibshirani: Introduction to Statistical Learning, Springer 2013)

Logistická regrese



zdroj: https://towardsdatascience.com/@iArunava

Logistická regrese



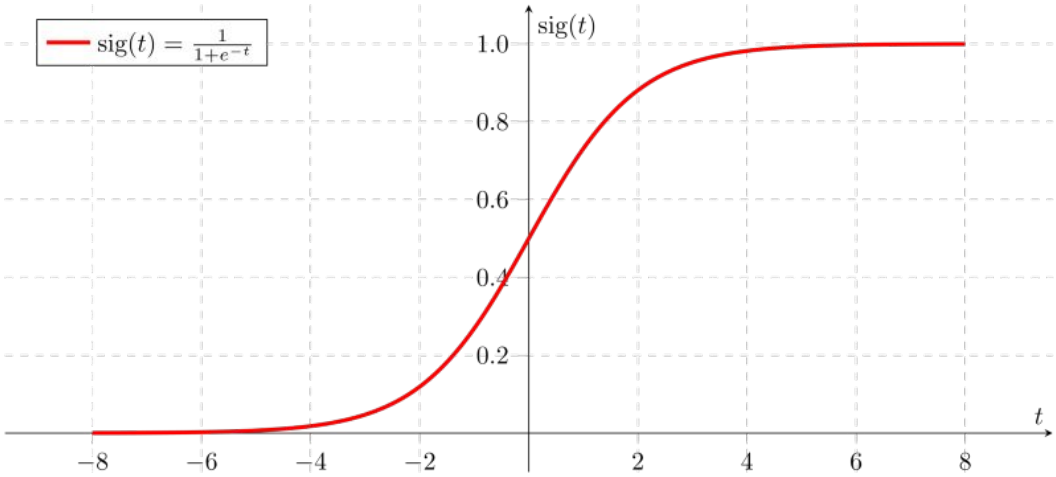
zdroj: https://towardsdatascience.com/@iArunava

metoda odhadu: sestrojit $t = f(x_0, \dots x_j)$

předpoklady:

- $x \propto logodds(y(x))$ … transformace
- extrémní hodnoty … transformace
- korelace prediktorů … výběr

Logistická regrese



zdroj: https://towardsdatascience.com/@iArunava

metoda odhadu: sestrojit $t = f(x_0, \ldots x_j)$

předpoklady:

- $x \propto logodds(y(x))$ ... transformace
- extrémní hodnoty ... transformace
- korelace prediktorů ... výběr

výhoda: interpretabilní

nevýhody: transformace: nutné, leč obtížná validace

Logistická regrese: výstup

model: **trénovací** data:

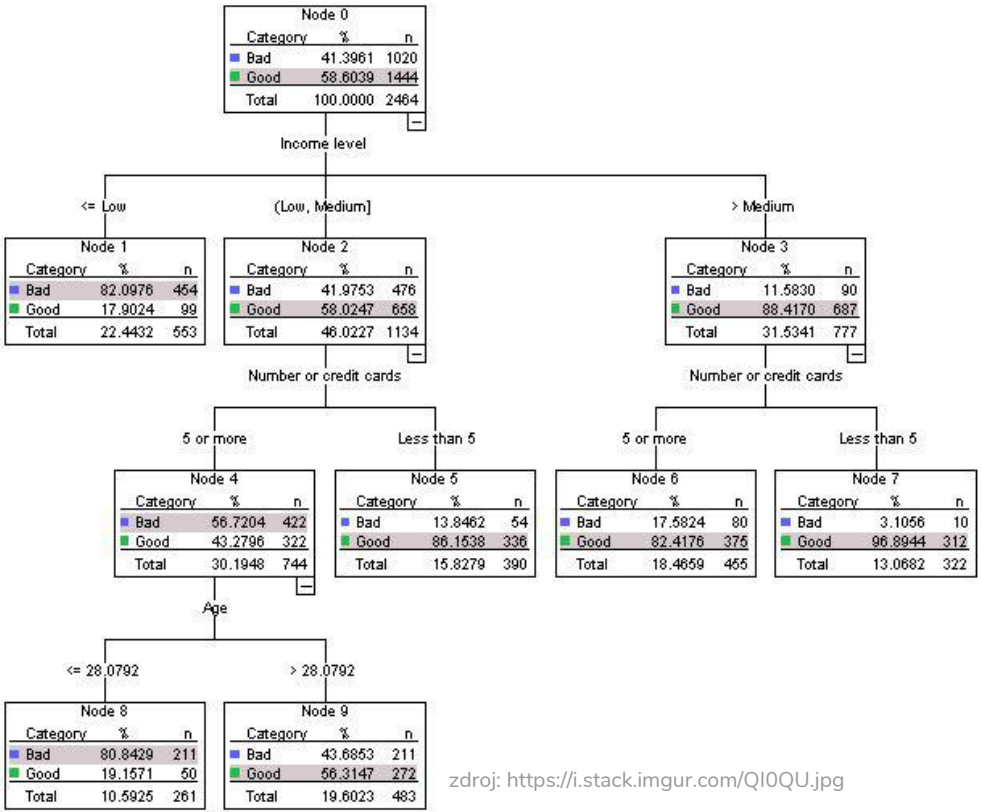
Parameter	DF	Estimate
Intercept	1	**-1.99**
Credit TOV to debt (WOE)	1	**-1.03**
Credit registry scores (WOE)	1	**-0.55**
Family status (WOE)	**1**	**-0.6**
Behavioral scores (WOE)	1	**-0.49**
Affordability (WOE)	1	**-0.47**
Time with bank (WOE)	1	**-0.63**

výkonnost: **testovací** data:

Statistics	Estimate
% Concordant	74.3
% Discordant	25.1
% Tied	0.6
Somers' D	0.49
C-statistics	0.75
Kolm-Smirn	0.28

(znaménka, DF, standardní odhady
transformace,
definice statistik)

Klasifikační strom



zdroj: https://i.stack.imgur.com/QI0QU.jpg

metoda odhadu: rozdělit pozorování dle x_1, ... x_j do homogenních skupin

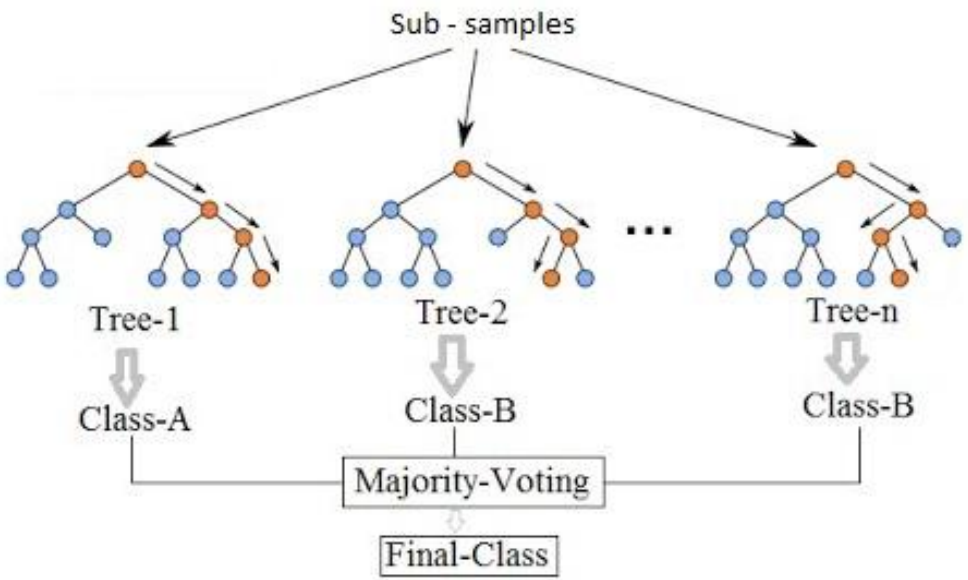
předpoklady: dostatek dat (“greedy” algoritmus)

výhoda: neparametrické, interpretabilní

nevýhody: overfitting

(jak trénovat, lineární separabilita)

Náhodný les



metoda odhadu: mnoho stromů -> průměrovat výsledek

předpoklady: žádné!

výhody: robustní, neparametrické (hyperparametry: dvojí sampling), implicitní validace

nevýhoda: interpretace

(bagging, Variable Importance Factor)

zdroj: https://www.kdnuggets.com/wp-content/uploads/rand-forest-1.jpg

Jak odhadovat LGD?

část 3: vážené průměry, lineární regrese, rozhodovací stromy

Strategie odhadu ztráty ze selhání

spojitá veličina

málo pozorování a proměnných: LGD = loss **given** default = ztráta podmíněná selháním

segmentace

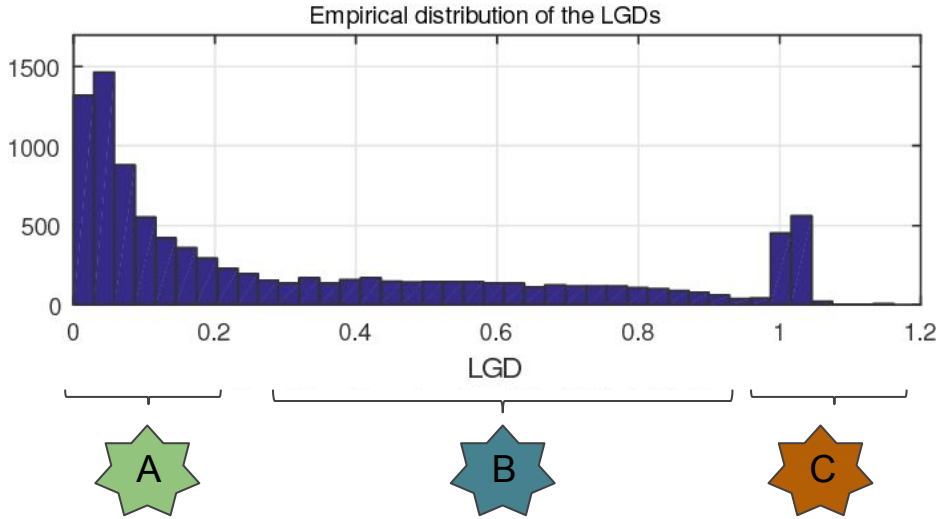
LGD > 100 %?

vazba na ekonomický cyklus: silnější než u PD

trénovací vzorek pro odhad LGD:

- co nejnovější data, avšak
- musíme znát skutečnou ztrátu

Expertní segmentace

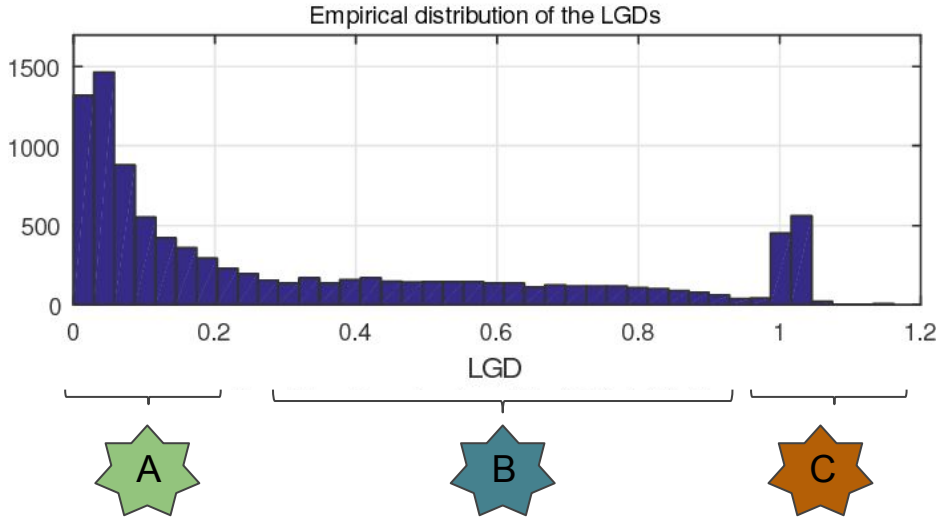


Umím rozdělit pohledávky do skupin?

- ano => odhad po skupinách
- ne => odhad dohromady

zdroj:
https://www.semanticscholar.org/paper/Loss-functions-for-LGD-model-comparison-Hurlin-Leymarie/bb392ddb155cbb4d45635789f5fb9a70ec4d3070

Vážený průměr a regrese



metody odhadu

- se segmentací nebo bez ní
- průměr: vážení historického LGD
- lineární / beta regrese, LGD $\propto f(x_0, \ldots x_j)$

předpoklady regrese:

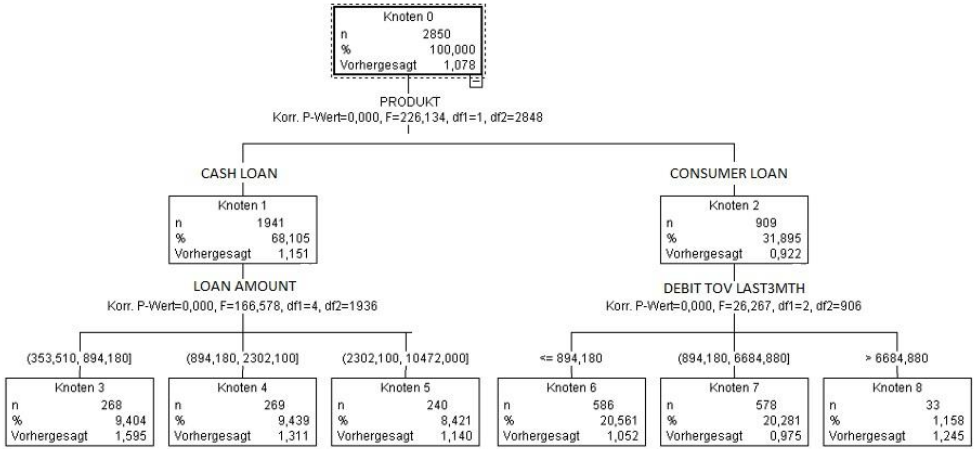
- statistické (lineární: homoskedasticita)
- praktické: dostatek proměnných?

výhoda: interpretace

nevýhody: omezená síla + vliv extr. hodnot

zdroj:
https://www.semanticscholar.org/paper/Loss-functions-for-LGD-model-comparison-Hurlin-Leymarie/bb392ddb155cbb4d45635789f5fb9a70ec4d3070

Regresní strom



metoda odhadu: rozdělit pozorování dle x_1, ... x_j do homogenních skupin

předpoklady: dostatek dat (“greedy” algoritmus)

výhoda: segmentaci vytvoří strom sám!

nevýhody:

- těžké zohlednit vliv expertů
- overfitting

zdroj: simulovaná data, SPSS

Doporučené zdroje

- Vývoj skórkaret prakticky: https://support.sas.com/documentation/cdl/en/emcsgs/66024/PDF/default/emcsgs.pdf
- Další klasifikační a regresní algoritmy: https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0

Děkuji za pozornost

 linkedin.com/in/vojtech-filipec/

🔎 github.com/vojtech-filipec