

# Manažer modelování kreditních rizik

část I: pravděpodobnost selhání a ztráta ze selhání





# 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



# whoami

☢ jaderná fyzika

📄 [linkedin.com/in/vojtech-filipec/](https://www.linkedin.com/in/vojtech-filipec/)

🔍 [github.com/vojtech-filipec](https://github.com/vojtech-filipec)

🎯 data scientist

profesní zkušenosti:

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

# **Proč odhadovat PD a LGD?**

část 1: komponenty očekávané  
ztráty



$$\mathbf{EL = PD \times LGD \times EAD}$$

[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 očekávané ztráty

dostupnost a relevance historických dat

EL krytá opravnými položkami

horizont predikce: typicky 12 měsíců

neočekávaná ztráta krytá kapitálem

kalibrace: PiT vs. TtC: hodnota na následujících 12 měsíců (výhoda: aktuální, nevýhoda: procyklická) vs. hodnota na průměrných 12 měsíců (nižší než konkurence za krize, vyšší než konkurence za konjunktury)

dokumenty Basel II a III: standardizovaný nebo IRB přístup

v této části kurzu se věnujeme jen komponentám PD a LGD

# **Jak odhadovat PD?**

část 2: logistická regrese,  
klasifikační stromy, náhodné  
lesy



# Strategie odhadu defaultu

klasifikace vs. regrese

diskrétní veličina: je praktické modelovat pravděpodobnost defaultu (spojitá veličina, tzv. PD)

trénovací vzorek pro odhad PD:

- co nejnovější data, avšak
- musíme znát, zda default nastal = kratší horizont predikce je lepší





# 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



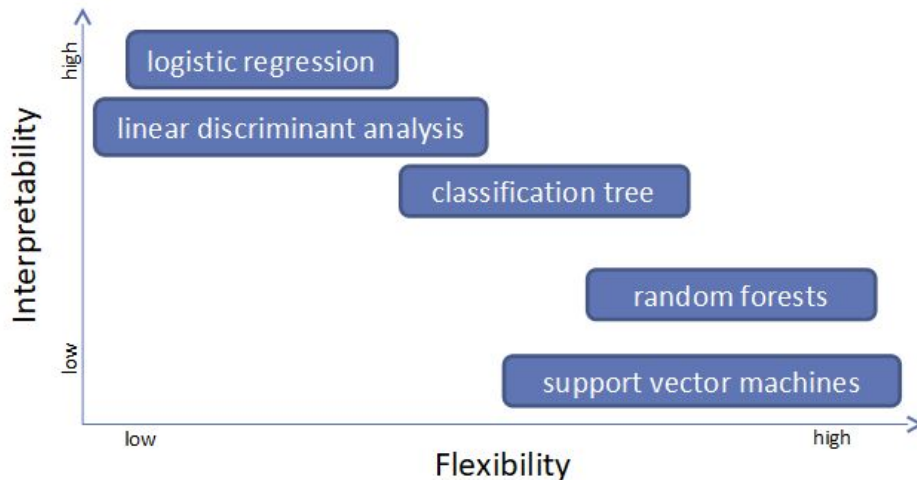
# O modelech (statistické minimum)

spojité, ordinální, nominální proměnné

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

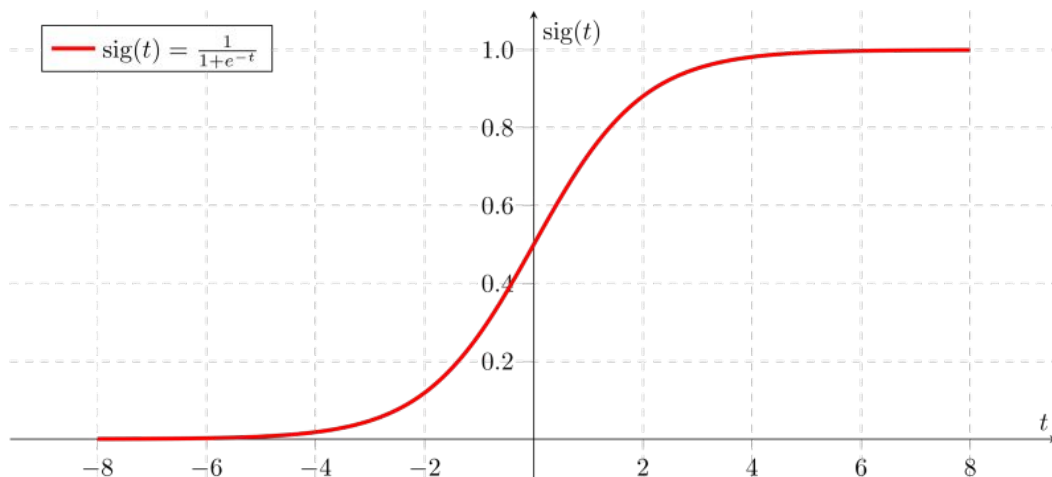
historická vs. budoucí data: stabilita populace, population stability index

komplexita vs. interpretabilita





# Logistická regrese



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

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

**předpoklady:**

- $x \propto \log\text{odds}(y(x))$  ... transformace
- extrémní hodnoty ... transformace
- korelace prediktorů ... výběr

**výhoda:** interpretabilní

**nevýhody:** transformace: nutné, leč obtížná validace



# Logistická regrese: výstup

model: **trénovací** data:

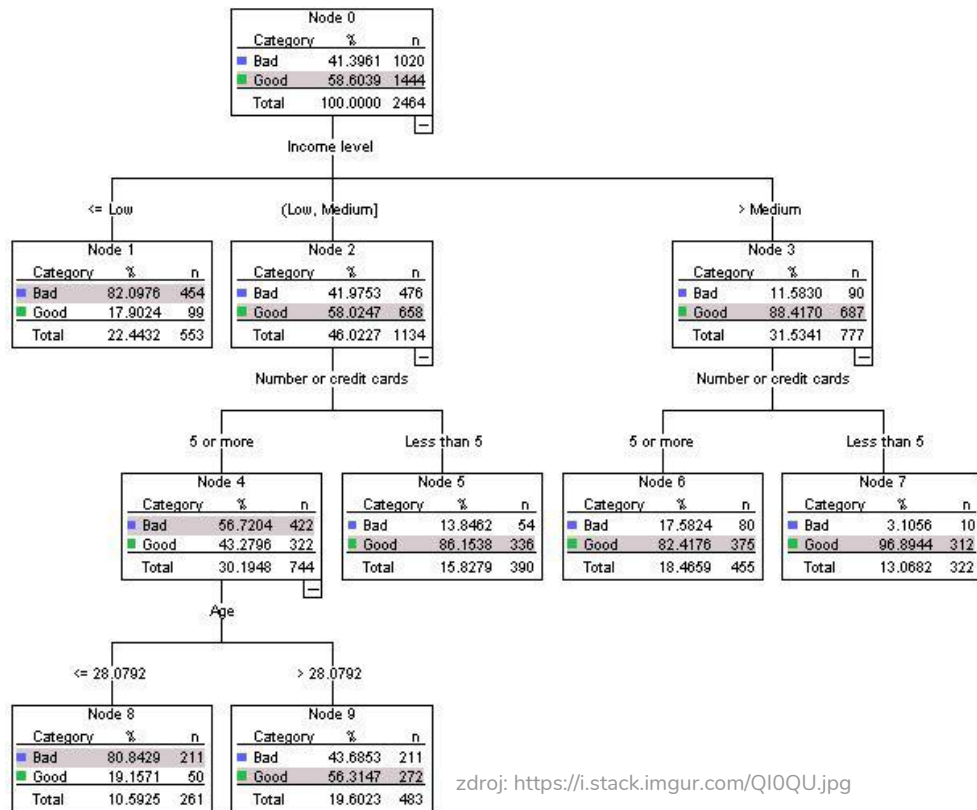
Parameter	DF	Estimate
Intercept	1	<b>-1.99</b>
Credit TOV to debt (WOE)	1	<b>-1.03</b>
Credit registry scores (WOE)	1	<b>-0.55</b>
Family status (WOE)	1	<b>-0.6</b>
Behavioral scores (WOE)	1	<b>-0.49</b>
Affordability (WOE)	1	<b>-0.47</b>
Time with bank (WOE)	1	<b>-0.63</b>

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/Q10QU.jpg>

**metoda odhadu:** rozdělit pozorování dle  $x_1, \dots, x_j$  do homogenních skupin

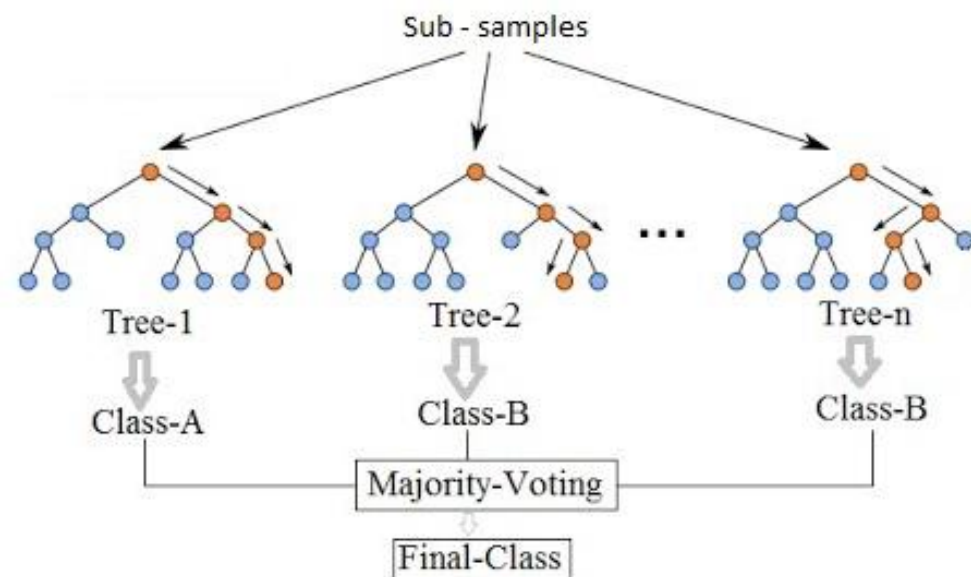
**předpoklady:** dostatek dat (“greedy” algoritmus)

**výhoda:** neparametrické, interpretabilní

**nevýhody:** overfitting

(jak trénovat, lineární separabilita)

# Náhodný les



**metoda odhadu:** mnoho stromů -> průměrovat výsledek

**předpoklady:** žádné!

**výhody:** robustní, neparametrické, implicitní validace

**nevýhoda:** interpretace

(bagging, Variable Importance Factor)

# 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

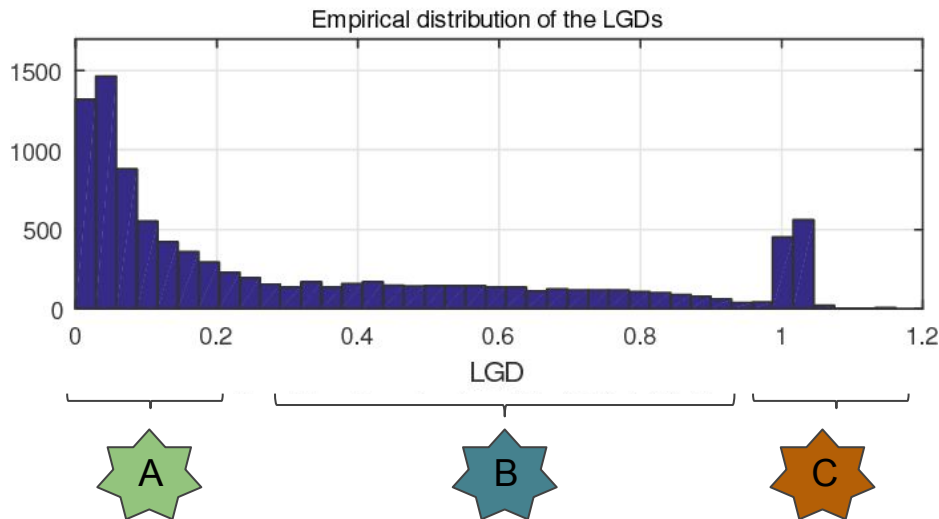
trénovací vzorek pro odhad LGD:

- co nejnovější data, avšak
- musíme znát skutečnou ztrátu





# Expertní segmentace



Umím rozdělit pohledávky do skupin?

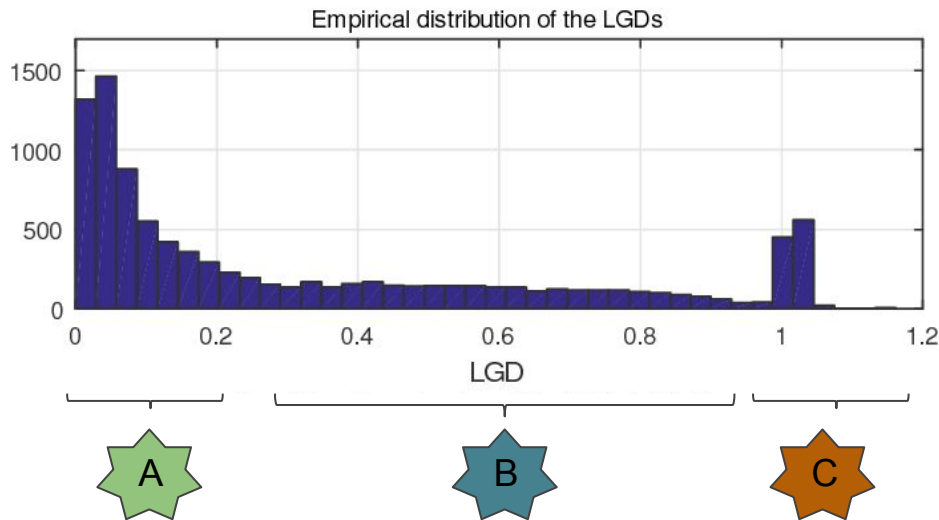
- ano => odhad po skupinách
- ne => odhad dohromady

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,  $\text{LGD} \propto f(x_0, \dots, x_j)$

## předpoklady regrese:

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

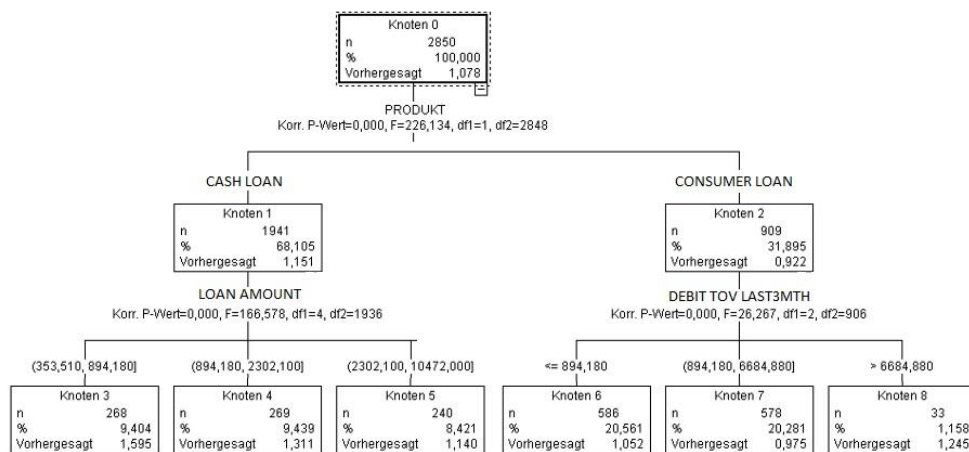
**výhoda:** interpretace

**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  $x_1, \dots, x_j$  do homogenních skupin

**předpoklady:** dostatek dat (“greedy” algoritmus)

**výhoda:** segmentaci vytvoří strom sám!

**nevýhody:**

- těžké zohlednit vliv expertů
- overfitting



## 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>