STANOVENÍ CUT-OFF SKÓRE ŠKÁLY DŮVĚRY V AI POMOCÍ ROC ANALÝZY

ÚVOD

Pojem umělá inteligence (artificial intelligence, AI) označuje soubor technologií, které umožňují počítačovým systémům vykonávat úkoly běžně spojované s lidskou inteligencí, jako je učení, uvažování, řešení problémů nebo rozhodování (Russell & Norvig, 2010). Vztah člověka k technologiím je přitom významně ovlivňován mírou důvěry, kterou jim přisuzuje (Hoff & Bashir, 2015). Při interakci člověka s umělou inteligencí se důvěra promítá do ochoty uživatele spoléhat se na systém, přenechat mu určitou míru kontroly nebo jednat podle jeho doporučení (Lee & See, 2004). Ideálním stavem je taková míra důvěry, která odpovídá skutečným schopnostem systému a rizikům dané situace. Pokud je AI přesná a potenciální rizika jsou nízká, může být míra důvěry vyšší než v situacích, kdy je systém nejistý nebo mohou mít jeho chyby závažné důsledky, například v medicíně. Obecně se nedoporučuje ani nekritická důvěra v AI, ani její apriorní odmítání (Lee & See, 2004; Mayer et al., 1995; NIST, 2023).

V rámci předmětu Psychometrika 1 byla vytvořena vlastní Škála důvěry v AI, zaměřená na důvěru uživatelů k AI chatbotům, jako je ChatGPT, Gemini apod. Cílem této práce je stanovit cut-off skóre Škály důvěry v AI pomocí ROC analýzy (Receiver Operating Characteristics). Stanovené cut-off skóre umožní identifikovat respondenty, kteří mají tendenci ve většině případů neověřovat informace poskytnuté AI jiným zdrojem. Identifikace těchto osob by mohla být prakticky využitelná například při cílení vzdělávacích programů zaměřených na mediální a digitální gramotnost.

Pro ROC analýzu byla použita data od 430 respondentů, která byla sesbíraná při tvorbě škály. Věk respondentů se pohyboval od 18 do 90 let, průměrný věk byl 28,13 roků a medián 23 let. Ženy tvořily 67,21 % souboru a muži 32,79 %.[1]

PROMĚNNÉ

Hrubé skóre celkové škály důvěry v AI. Rozsah skóre, kterého může respondent ve škále dosáhnout je 13 – 65 bodů. Vyšší hrubé skóre naznačuje vyšší důvěru v AI.

Dichotomická proměnná vyjadřující, zda respondent ve většině případů neověřuje informace poskytnuté AI jiným zdrojem. Tato proměnná byla kódována jako 0 = respondent informace ověřuje a 1 = respondent informace ve většině případů neověřuje. Hodnota této proměnné byla odvozena z odpovědi na validační otázku: *„V kolika procentech případů ověřujete informace, které Vám poskytne AI, z jiného zdroje?“* Respondenti zadávali hodnotu v rozmezí 0–100 % (např. 0 % = nikdy, 100 % = vždy). Respondenti, kteří ověřují informace v méně než 50 % případů, byli klasifikováni jako osoby, které ve většině případů informace z AI neověřují. Mezi frekvencí ověřování informací z AI a hrubým skóre škály byl zjištěn středně silný negativní vztah (Spearmanovo $\rho = -0{,}39$; $p < 0{,}001$).

[1] Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4/zprava.php?id=702

Tento výsledek naznačuje, že respondenti s vyšší důvěrou v AI mají tendenci méně často ověřovat informace, které AI poskytuje.

Prevalence respondentů v našem souboru, kteří ve většině případů neověřují informace poskytnuté AI jiným zdrojem. Ta se rovná 24,65 %.

METODA

Pro nalezení optimálního cut-off skóre byly nejprve identifikovány všechny hodnoty hrubého skóre, kterých respondenti ve standardizačním souboru dosahovali. Pro každou z těchto hodnot byly následně vypočítány ukazatele, které umožňují posoudit vhodnost daného cut-off skóre:

- **TP** (true positive) – počet respondentů, které dané cut-off skóre označí jako osoby, které informace z AI neověřují, a kteří je ve skutečnosti opravdu neověřují.
- **FP** (false positive) – počet respondentů, které cut-off skóre označí jako osoby, které informace z AI neověřují, avšak ve skutečnosti je ověřují.
- **TN** (true negative) – počet respondentů, které cut-off skóre označí jako osoby, které informace z AI ověřují, a kteří je skutečně ověřují.
- **FN** (false negative) – počet respondentů, které cut-off skóre klasifikuje jako osoby, které informace z AI ověřují, přestože je ve skutečnosti neověřují.
- **Senzitivita** – podíl správně identifikovaných respondentů, kteří informace z AI neověřují. Vypočítá se pomocí vzorce TP/(TP+FN). Udává pravděpodobnost, s jakou cut-off skóre správně identifikuje respondenta, který informace z AI ve většině případů neověřuje.
- **Specificita** – podíl správně identifikovaných respondentů, kteří informace z AI ověřují. Vypočítá se pomocí vzorce TN/(TN+FP). Udává pravděpodobnost, s jakou cut-off skóre správně identifikuje respondenta, který informace z AI ověřuje.

VÝSLEDKY

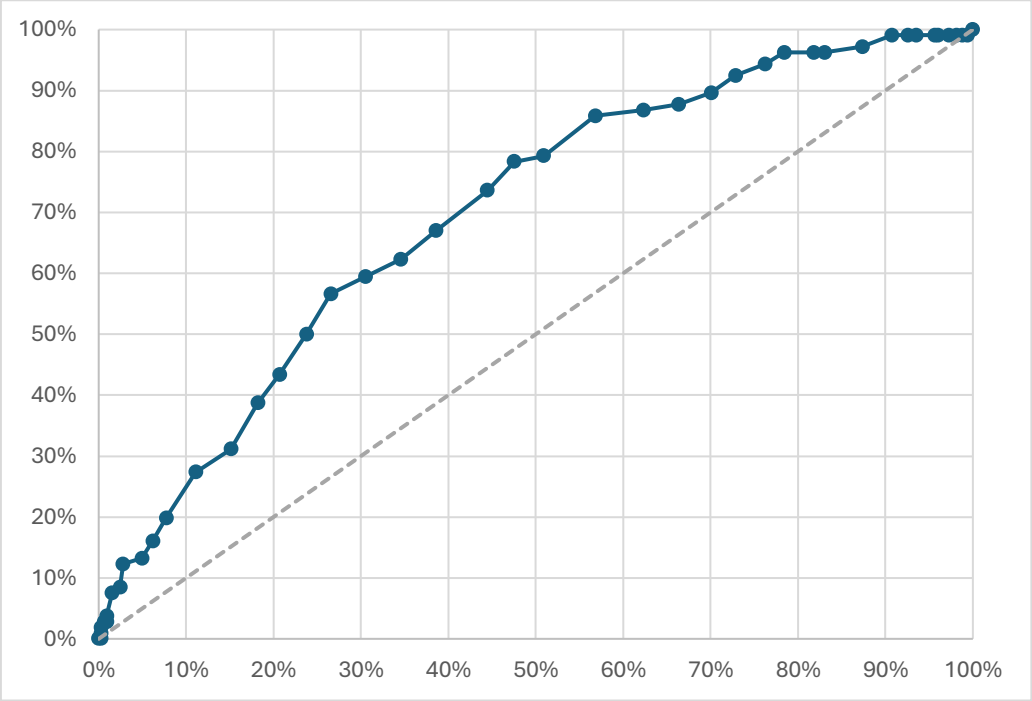
Na základě hodnot senzitivity a specificity byl pro každé možné cut-off skóre vypočten **Youdenův index (J)**. Čím vyšší je jeho hodnota, tím lépe dané cut-off skóre rozlišuje mezi skupinami. Nejvyšší hodnota Youdenova indexu byla v našem souboru dosažena při **cut-off skóre HS = 34**, které se tak jeví jako nejvhodnější hranice podle tohoto kritéria.

Youdenův index přikládá stejnou váhu senzitivitě a specificitě a nezohledňuje skutečné zastoupení jednotlivých skupin v souboru. V našem souboru je však zastoupení skupin nerovnoměrné, protože respondentů, kteří informace z AI ověřují, je výrazně více než těch, kteří je ve většině případů neověřují. Z tohoto důvodu byla pro každé cut-off skóre vypočtena také **statistika I** která zohledňuje skutečné zastoupení obou skupin v souboru. Hodnota statistiky I vyjadřuje pravděpodobnost správné klasifikace respondenta při daném cut-off skóre.

Vyšší hodnota tedy znamená vhodnější klasifikační hranici. Na základě tohoto kritéria se jako nejvhodnější cut-off skóre jeví **HS = 48**. Vzhledem k tomu, že v souboru převažují respondenti, kteří informace z AI ověřují, dává toto kritérium větší váhu specificitě, což vede k posunu optimální hranice směrem k vyšším hodnotám skóre. Srovnání všech možných cut-off skóre je uvedeno v tabulce 1.

V grafu 1 je zobrazena ROC křivka. Čím více se křivka odchyluje od diagonály náhodné klasifikace a přibližuje se k levému hornímu rohu grafu, tím lepší je schopnost testu rozlišovat mezi jednotlivými skupinami. Plocha pod ROC křivkou se označuje jako AUC (area under the curve) a představuje celkovou diskriminační schopnost škály. Hodnota AUC vyjadřuje pravděpodobnost, že náhodně vybraný respondent, který informace z AI ve většině případů neověřuje, dosáhne vyššího skóre škály než respondent, který informace z AI ověřuje. Hodnota AUC pro naši škálu činí 0,70, což naznačuje přijatelnou diskriminační schopnost škály.

Graf 1
ROC křivka



ZÁVĚR

Na základě statistiky I bylo jako optimální cut-off skóre zvoleno HS = 48. Pokud respondent ve škále dosáhne 48 nebo více bodů, lze předpokládat, že ve většině případů neověřuje informace poskytnuté AI jiným zdrojem.

Tabulka 1
Cut-off skóre

Cut-off	Senzitivita	Specificita	J	I
13	100 %	0 %	0,00	24,65 %
14	99 %	1 %	0,00	24,88 %
15	99 %	1 %	0,00	25,35 %
16	99 %	2 %	0,01	25,81 %
17	99 %	3 %	0,02	26,51 %
18	99 %	4 %	0,03	27,44 %
19	99 %	4 %	0,03	27,67 %
20	99 %	6 %	0,06	29,30 %
21	99 %	7 %	0,06	30,00 %
22	99 %	9 %	0,08	31,40 %
23	97 %	13 %	0,10	33,49 %
24	96 %	17 %	0,13	36,51 %
25	96 %	18 %	0,14	37,44 %
26	96 %	22 %	0,18	40,00 %
27	94 %	24 %	0,18	41,16 %
28	92 %	27 %	0,20	43,26 %
29	90 %	30 %	0,20	44,65 %
30	88 %	34 %	0,21	46,98 %
31	87 %	38 %	0,24	49,77 %
32	86 %	43 %	0,29	53,72 %
33	79 %	49 %	0,28	56,51 %
34	**78 %**	**52 %**	**0,31**	**58,84 %**
35	74 %	56 %	0,29	60,00 %
36	67 %	61 %	0,28	62,79 %
37	62 %	65 %	0,28	64,65 %
38	59 %	69 %	0,29	66,98 %
39	57 %	73 %	0,30	69,30 %
40	50 %	76 %	0,26	69,77 %
41	43 %	79 %	0,23	70,47 %
42	39 %	82 %	0,20	71,16 %
43	31 %	85 %	0,16	71,63 %
44	27 %	89 %	0,16	73,72 %
45	20 %	92 %	0,12	74,42 %
46	16 %	94 %	0,10	74,65 %
47	13 %	95 %	0,08	74,88 %
48	**12 %**	**97 %**	**0,09**	**76,28 %**
49	8 %	98 %	0,06	75,58 %
50	8 %	98 %	0,06	76,05 %
51	4 %	99 %	0,03	75,58 %
53	3 %	99 %	0,02	75,35 %
57	3 %	99 %	0,02	75,58 %
58	2 %	100 %	0,02	75,58 %
60	1 %	100 %	0,01	75,35 %
62	0 %	100 %	0,00	75,12 %
63	0 %	100 %	0,00	75,35 %

LITERATURA

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: a modern approach* (3rd ed). Pearson.