Predikcia budúcich nákladov na pacienta

Marián Kravec
školiteľ: MSc. František Dráček

Obsah

Motivácia a ciele práce

- Slovenské zdravotníctvo má nízku mieru využitia dostupných dát.
- Predikcia budúcich nákladov pacienta môže zlepšiť plánovanie a prerozdelenie zdrojov.
- Ciele:
  - Navrhnúť a implementovať spôsob transformácie záznamov pacienta do číselných vektorov (embedding).
  - Navrhnúť a implementovať systém na predikciu budúcich nákladov pacienta na základe jeho historických záznamov.
- Výsledky práce:
  - Embedované atribúty spĺňali vlastnosť blízkosti podobných záznamov.
  - Systém na predikciu budúcej nákladovej skupiny pacienta mal presnosť 39.9% (84.4% v prípade povolenia chyby o jednu kategóriu).

Obsah

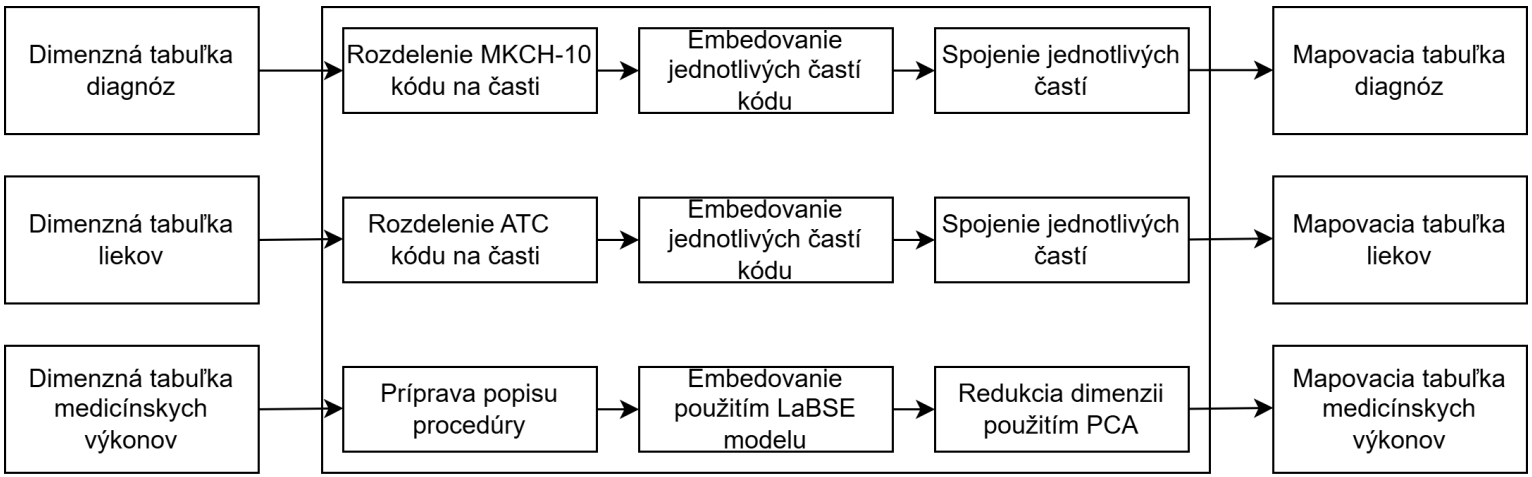
Dáta

- Umelo generované dáta podľa štruktúry NCZI: 173 355 pacientov, 133 miliónov záznamov.
- Typ dát: poistné dáta.
- Dve skupiny záznamov (datasety):
  - Výkony ambulantnej zdravotnej starostlivosti
  - Predpísané lieky
- Jeden záznam: informácia o jednom výkone/lieku konkrétneho pacienta.
- Použité atribúty:
  - Dátum záznamu
  - Vek pacienta
  - Medicínsky výkon
  - Liek
  - Diagnóza
  - Cena výkonu/lieku

Obsah

Vytváranie embeddingov záznamov

- Každý záznam reprezentovaný vektorom (embedding) z 4 častí:
  - Časová pečiatka (vek pri udalosti)
  - Diagnóza (MKCH-10, hierarchické embeddingy)
  - Liek (ATC, hierarchické embeddingy)
  - Výkon (LaBSE embedding popisu + PCA)
- Cieľ: podobné záznamy majú podobné vektory (Euclidovská vzdialenosť).
- Overenie: klastrovanie, výpočty podobností.



Časová pečiatka

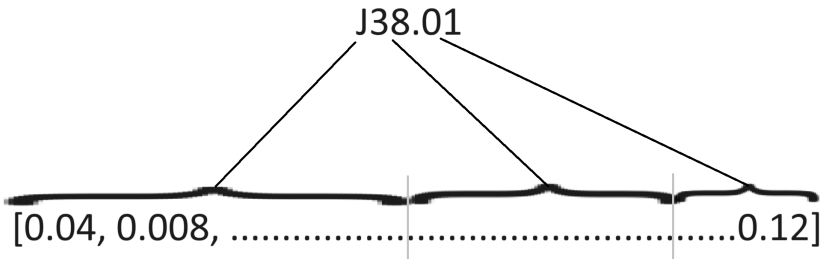
- Odhad veku pacienta v dňoch v čase záznamu.
- Určená podľa veku pacienta v rokov v čase prvého záznamu a časovým rozdielom (v dňoch) medzi dátumami prvého a daného záznamu.
- Chyba: v priemere štvrťrok, nanajvýš polrok.

Diagnózy

- Embedujeme MKCH-10 kódy.

[znak][dvojčíslie].[dvojčíslie]

- Hierarchický kód.
- Každá časť kódovaná samostatne a následne spojené dohromady.
- Príklad:
  - J - Choroby dýchacej sústavy
  - J3 - Iné choroby dýchacích ciest
  - J38 - Choroba hlasiviek a hrtana
  - J38.0 - Obrna hlasiviek a hrtana
  - J38.01 - Jednostranná čiastočná obrna hlasiviek a hrtana



Lieky

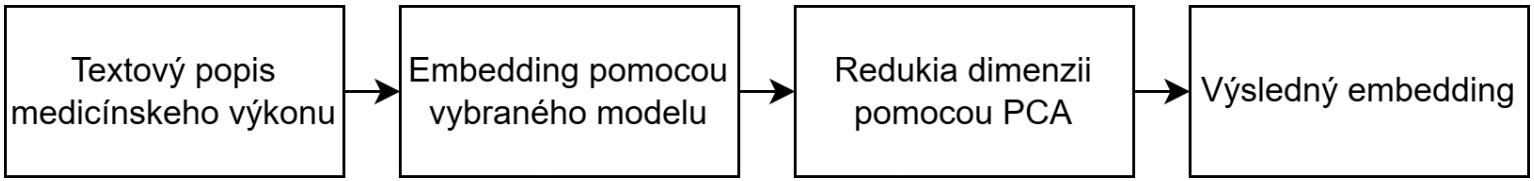
- Embedujeme ATC kód.

  [znak][dvojčísle][dvojica znakov][dvojčíslie]

- Hierarchický kód.
- Každá časť kódovaná samostatne a následne spojené dohromady.
- Príklad:
  - N - Centrálna nervová sústava
  - N02 - Analgetiká
  - N02B - Iné analgetiká a antipyretiká
  - N02BA - Kyselina salicylová a deriváty
  - N02BA01 - Acylpyrín

Výkony

- Embedujeme textový popis výkonu.
- Vyskúšané modely:
  - Language-agnostic BERT sentence embedding model (LaBSE)
  - Lemmatizer + Word2vec model
- Redukcia dimenzionality pomocou PCA, počet dimenzii vyberaný tak aby zachovávali 90% variancie.



Validácia embeddingov

- Diagnózy a lieky:
  - Porovnanie podobností (obrátená hodnota vzdialenosti) dvojíc niekoľkých náhodne vybraných prípadov.
  - Klastrovanie (K-means) a následné kontrola distribúcii hlavných kategórii prípadov v klastroch.
- Výkony: klastrovanie a následná vizuálna kontrola obsahu náhodne vybraných klastrov.

Validácia embeddingov

Diagnózy

Code A	Code B	Similarity
G47.30	G40.09	2.77
G47.30	H40.09	0.53
G47.30	H18.80	0.46
G40.09	H40.09	0.54
G40.09	H18.80	0.45
H40.09	H18.80	0.84

Lieky

Code A	Code B	Similarity
C01EB15	C01CA04	1.24
C01EB15	C10AA07	0.54
C01EB15	J01CA04	0.45
C01CA04	C10AA07	0.64
C01CA04	J01CA04	0.48
C10AA07	J01CA04	0.38

Validácia embeddingov

Diagnózy

Cluster ID	Cluster size	Frequency of first level values
0	654	C: 654,
1	2092	M: 2092,
2	766	Y: 766,
3	543	S: 543,
4	786	Z: 786,
5	1100	T: 1100,
6	1067	X: 1067,
7	620	H: 496, U: 124,
8	752	S: 752,
9	1770	M: 1770,
10	739	Q: 739,
11	584	D: 584,
12	507	F: 507,
13	958	W: 958,
14	556	E: 556,
15	491	A: 491,
16	553	O: 553,
17	627	K: 627,
18	872	V: 872,
19	521	G: 521,
20	463	L: 463,
21	420	R: 420,
22	465	B: 465,
23	717	J: 319, P: 398,
24	568	I: 568,
25	535	N: 535,

Lieky

Cluster ID	Cluster size	Frequency of first level values
0	8645	C: 8645,
1	19132	N: 19132,
2	9322	L: 8657, S: 665,
3	11734	A: 11734,
4	7159	B: 7159,
5	9526	V: 9526,
6	4681	R: 4681,
7	14732	C: 14732,
8	7237	V: 7237,
9	4031	M: 3987, P: 44,
10	3599	G: 3599,
11	2367	N: 2367,
12	9032	D: 1266, G: 897, H: 1136, J: 5733,
13	6093	N: 6075, P: 18,

Validácia embeddingov

Medicínske výkony

- Kluster 215:

```
Výber vhodných príjemcov pre kadaverózny transplantát z listiny cakatelov
Transplantácia pečene (UH+90301)
Transplantácia obličky (UH+90101)
Transplantácia obličky
Voľná transplantácia šliach
Odobratie chrupkového alebo kostného materiálu na voľnú transplantáciu
Odber pečene na transplantáciu
Transplantácia pečene
Transplantácia pankreasu
Odobratie orgánov alebo časti orgánov na transplantáciu: Pankreas
Odobratie orgánov alebo časti orgánov na transplantáciu: Oblička
Odobratie orgánov alebo časti orgánov na transplantáciu: Srdce
Odobratie orgánov alebo časti orgánov na transplantáciu: Kostná dreň
Odobratie orgánov alebo časti orgánov na transplantáciu: Rohovka
Odobratie orgánov alebo časti orgánov na transplantáciu: Týmus
Odobratie kostného alebo chrupkového materiálu na transplantáciu
Odber kostnej drene na účely transplantácie
Indikácia darcu na odber orgánov na transplantáciu
Celotelové ožarovanie pre transplantáciu kostnej dreni
Transplantácia obličiek
Transplantácia srdca
Transplantácia pečene
Transplantácia pankreasu
Transplantácia pľúc
Transplantácia rohovky
Transplantácia skléry
Transplantacia sklery - naklady suvisiace s odberom sklery
Voľný šľachový transplantát
voľný šľachový transplantát
Transplantácia kostnej drene
```

Validácia embeddingov

Medicínske výkony

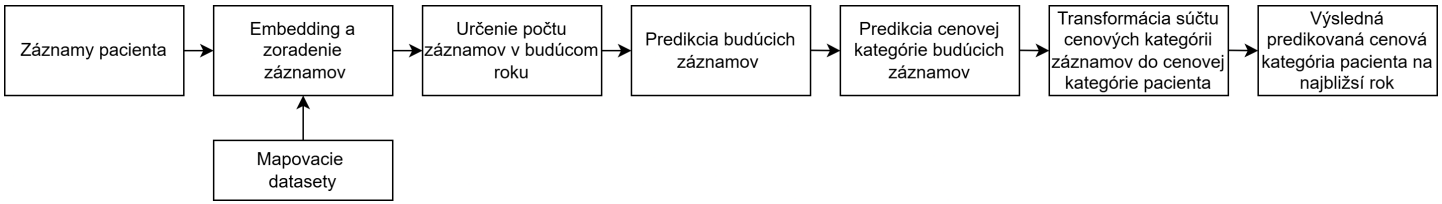
- Kluster 45:

```
Rozbor a plánovanie (komplexná analýza).
Zhodnotenie výsledkov komplexného hemokoagulacného vyšetrenia a klinická interpretácia porúch
Vyhodnotenie KOS a záverecná správa.
Zhodnotenie výsledkov a záver
Resekcia močovodu a reanostomáza
Resekcia a rekonštrukcia žlčových ciest pri nádoroch
Vyhodnotenie KOS a záverečná správa
Vyhodnotenie sociálnej starostlivosti a záverečná správa
vWF antigén - vyšetremie farmakokinetiky a monitorovanie liečby
vWF Ricof - vyšetremie farmakokinetiky a monitorovanie liečby
Faktor VII - vyšetrenie farmakokinetiky a monitorovanie liečby
Faktor VIII - vyšetrenie farmakokinetiky a monitorovanie liečby
Faktor IX - vyšetrenie farmakokinetiky a monitorovanie liečby
Počítačové zhodnotenie polysomnografického záznamu a zhodnotenie lekárom
Papilosfinkterotómia a odstránenie konkrementov zo žlčových ciest
Papilosfinkterotómia a odstránenie konkrementov
Papilosfinkterektómia a odstránenie konkrementov zo žlčových ciest alebo pankreatického vývodu
SVLZ - Spoločné vyšetrovacie a liečebné zložky
```

Obsah

Proces predikcie

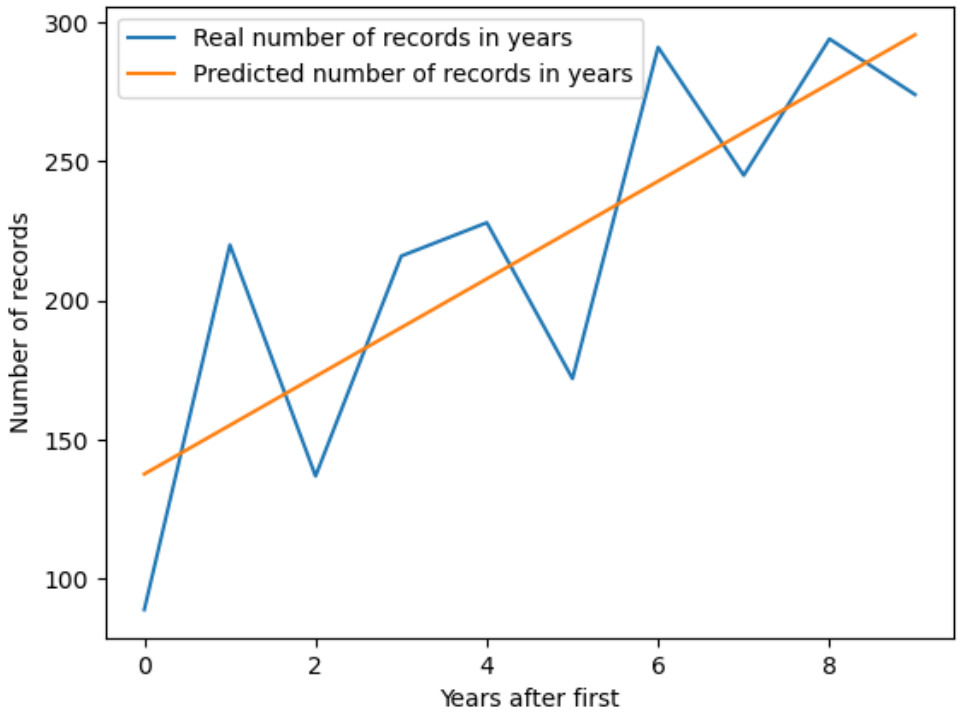
- Kroky pri predikovaní budúcej cenovej kategórie pacienta:
  1. Načítanie dát pacienta, embedding a normalizácia.
  2. Výpočet počtu záznamov na budúci rok.
  3. Predikcia budúcich záznamov.
  4. Predikcia cenových kategórii budúcich záznamov.
  5. Výpočet cenovej kategórie pacienta.



- Hyperparametre modelov nastavované lokálne na podskupine pacientov.
- Finálne modely trénované na serveri na plnohodnotnom datasete.

Výpočet počtu budúcich záznamov

- Testované metódy:
  - Polynomialna regresia počtu záznamov z predchádzajúcich rokov.
  - Generovanie nových záznamov kým generovaná časová pečiatka nepresiahne rok od posledného skutočného záznamu.
- Najlepší výsledok: lineárna regresia.

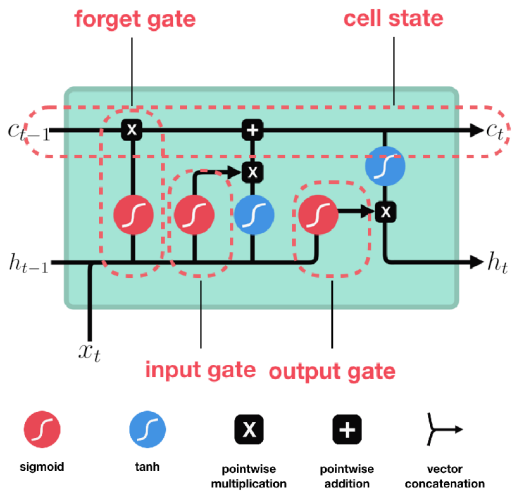


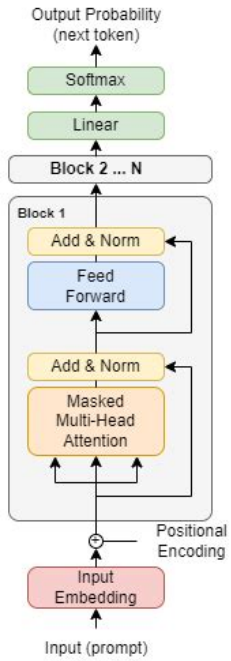
Predikcia budúcich záznamov

- Testované modely:
  - Long short-term memory.
  - Decoder-only transformer.
- Vstup: posledných *n* (veľkosť okna/kontextu) záznamov pacienta.
- Výstup: predikcia nasledujúceho záznamu.
- Chybová funkcia: Subpart weighted MSE

$$SubpartWeightedMSE = \frac{1}{p}\sum_{j=1}^{p}(\frac{1}{l_j}\sum_{i=1}^{l_j}(Y_{s_j+i} - \hat{Y}_{s_j+i})^2).$$

- Optimalizované hyperparametre:
  - Hĺbka modelu (počet vrstiev).
  - Šírka modelu (LSTM).
  - Počet hláv (Transformer).
  - Dropout rate.





Predikcia budúcich záznamov - trénovanie a validácia

LSTM

Number of layers	Width of layer	Train loss	Validation loss
3	196	0.4921	0.6389
	392	0.4472	0.6517
	784	0.4114	0.6535
6	196	0.5156	0.6326
	392	0.4786	0.6452
	784	0.4285	0.6493
12	196	0.5983	0.6278
	392	0.5822	0.6292
	784	0.5780	0.6268

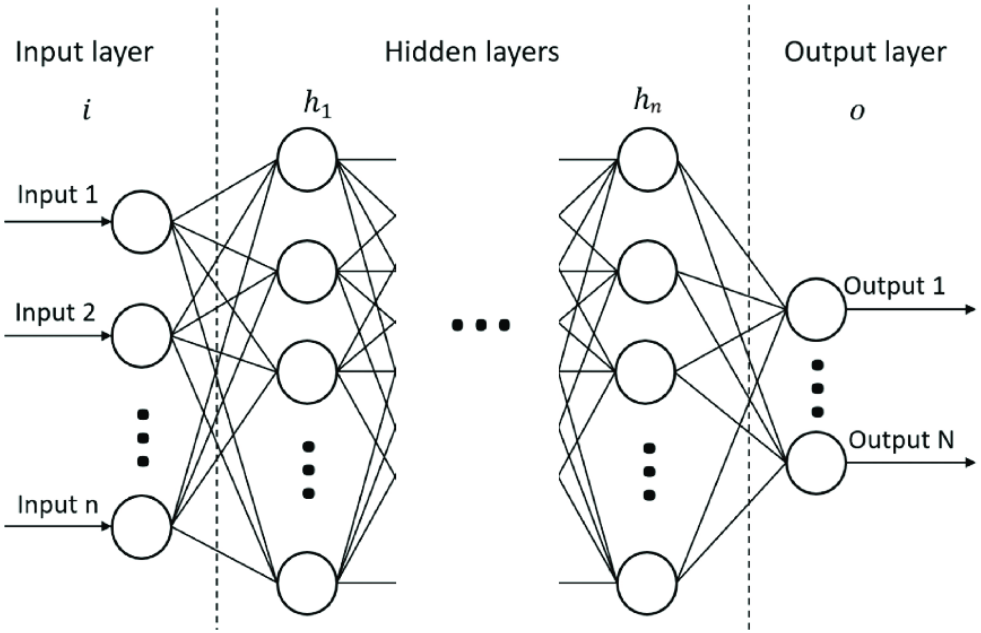
Transformer

Number of layers	Number of heads	Train loss	Validation loss
3	7	0.5057	0.6276
	14	0.5035	0.6266
	49	0.5103	0.6297
6	7	0.4737	0.6416
	14	0.4714	0.6398
	49	0.4716	0.6439
12	7	0.4624	0.6537
	14	0.4301	0.6704
	49	0.4119	0.6729

- Najlepší model pri lokálnom trénovaní: Decoder-only Transformer, 3 vrstvy, 14 hláv, 20% dropout rate.
- Chyba pri trénovaní na serveri:
  - Trénovacia: 0.4005
  - Validačná: 0.5200

Predikcia cenovej kategórie záznamu

- Použitý model: Multilayer perceptron.
- Vstup: jeden záznam pacienta.
- Výstup: jedna z 9 cenových kategórii.
- Vyskúšané jednoduchšie modely: Gradient boosting, Ridge regression



Category	Interval
1	[0,1)
2	[1,5)
3	[5,10)
4	[10,20)
5	[20,50)
6	[50,100)
7	[100,200)
8	[200,500)
9	[500,∞)

Predikcia cenovej kategórie záznamu - trénovanie a validácia

- Optimalizované hyperparametre:
  - Hĺbka modelu.
  - Veľkosti jednotlivých vrstiev.
  - Aktivačné funkcie medzi vrstvami.
  - Chybová funkcia (MSE, Cross Entropy, NLL).
- Miera kvality modelu: presnosť.

Depth	Layer sizes	Activation functions	Mean square error		Cross entropy		Negative log likelihood	
			Test accuracy	Validation accuracy	Test accuracy	Validation accuracy	Test accuracy	Validation accuracy
0	[]	[]	63.2%	63.0%	63.5%	63.3%	62.9%	62.6%
1	[98]	[GELU]	74.5%	74.3%	72.8%	72.6%	73.2%	73.0%
		[Tanh]	71.3%	71.0%	70.9%	70.6%	71.1%	70.9%
	[392]	[GELU]	76.6%	76.3%	74.2%	74.0%	75.5%	75.2%
		[Sigmoid]	70.8%	70.6%	69.9%	69.6%	70.2%	70.0%
	[196]	[GELU]	75.9%	75.6%	74.1%	73.8%	74.1%	73.8%
		[LeakyReLU]	76.0%	75.8%	75.5%	75.3%	74.1%	74.0%
3	[98, 48, 24]	[GELU, GELU, GELU]	74.4%	74.2%	72.0%	71.8%	72.2%	72.0%
		[Sigmoid, ReLU, Tanh]	69.7%	69.5%	69.7%	69.4%	70.3%	70.1%
	[392, 196, 98]	[GELU, GELU, GELU]	77.7%	77.5%	75.2%	75.0%	75.5%	75.3%
		[SiLU, ReLU, GELU]	77.4%	77.2%	74.8%	74.5%	75.2%	75.0%
	[196, 196, 196]	[GELU, GELU, GELU]	77.2%	77.0%	74.7%	74.3%	75.0%	74.7%
		[ReLU, Sigmoid, SiLU]	76.7%	76.5%	73.9%	73.7%	75.0%	74.8%

Predikcia cenovej kategórie záznamu - trénovanie a validácia

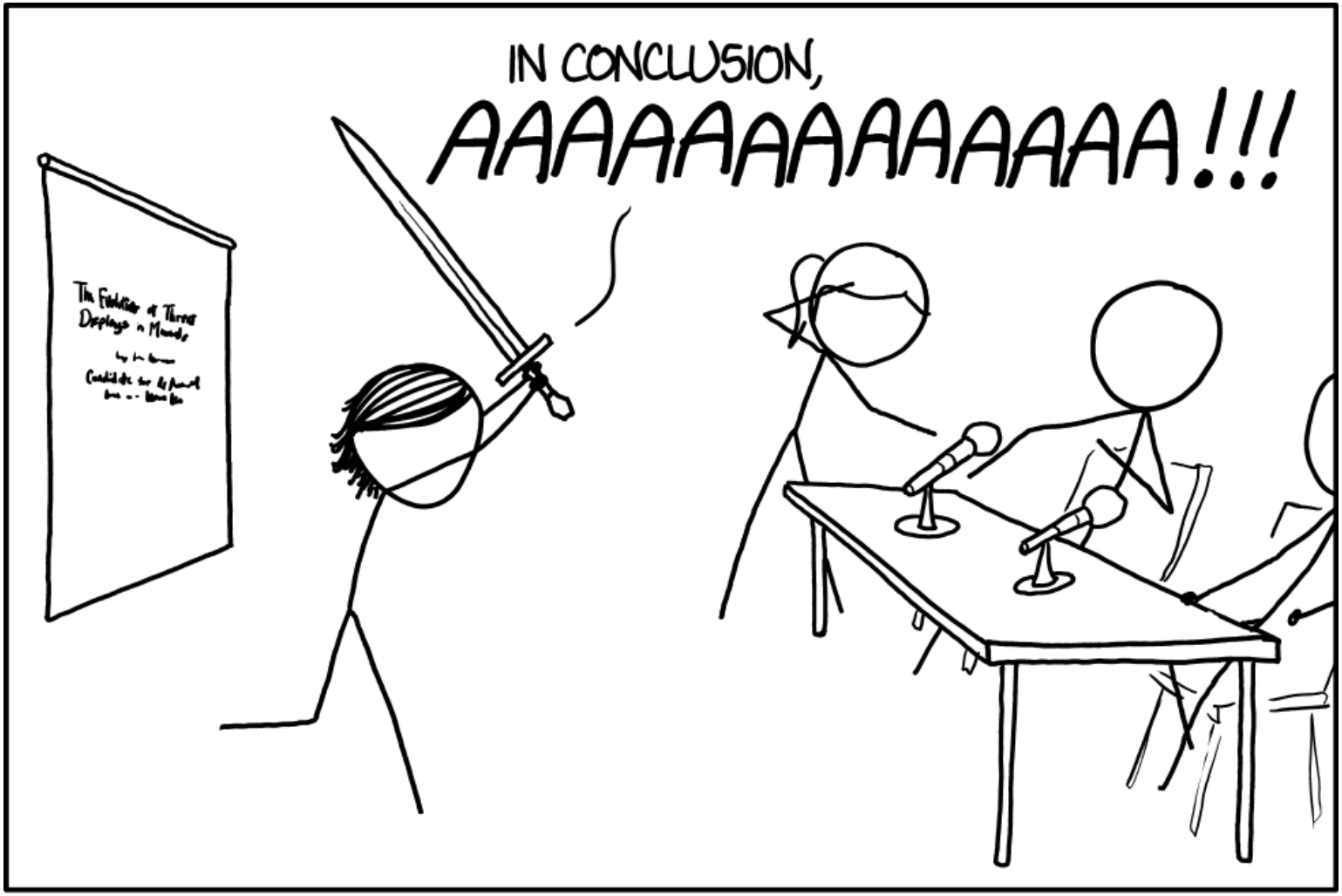
- Najlepší model pri lokálnom trénovaní:
  - 8 vrstiev.
  - Šírky vrstiev: [588, 294, 147, 49, 98, 36, 18, 9].
  - Aktivačné funkcie medzi vrstvami: [SiLU, GELU, Sigmoid, GELU, SiLU, GELU, LeakyReLU, Softmax].
  - Chybová funkcia: MSE.
- Presnosť pri lokálnom trénovaní:
  - Trénovacia: 78.3%
  - Validačná: 78.0%
- Presnosť pri trénovaní na serveri:
  - Trénovacia: 80.2%
  - Validačná: 80.1%
- Validačná presnosť najlepšieho Gradient boosting: 71.4%
- Validačná presnosť najlepšieho Ridge regression: 66.9%

Predikcia cenovej kategórie pacienta - validácia

- AAAAAAAAAAAAAAA

Ďakujem za pozornosť

Obsah



zdroj: XKCD (`https://xkcd.com/1403/`)