Maroš Geffert
Ústav počítačové grafiky a multimédií, Fakulta Informačných Technológií
Vysoké Učení Technické v Brně

FRAMEWORK PRE MODELOVANIE A PREDIKCIU UDALOSTÍ VO FUTBALE

ÚVOD

Futbal, často označovaný ako "kráľovský šport", je globálny fenomén ovplyvňujúci kultúru, ekonomiku a spoločnosť. Jeho schopnosť predpovedať výsledky zápasov je kľúčová pre kluby a investorov v snahe o získanie konkurenčnej výhody.

MOTIVÁCIA

Futbal, známy svojou dynamikou a nepredvídateľnosťou, predstavuje výzvu pre prediktívne modely ako sú neurónové siete, XGBoost a RandomForest, ktoré sľubujú zvýšenie presnosti predpovedí športových výsledkov. Táto práca sa zameriava na využitie týchto metód na predpovedanie dôležitých futbalových udalostí, ako počet gólov či výsledok zápasu, s dôrazom na adekvátny výber príznakov. Hlavné ciele práce zahŕňajú:

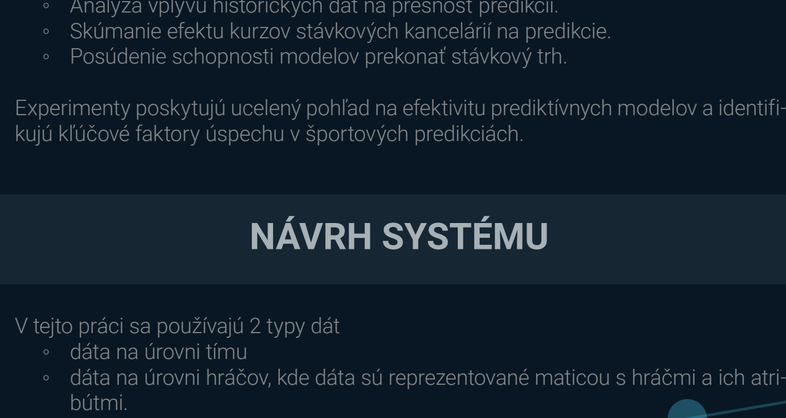
- Hodnotenie efektivity prediktívnych modelov oproti základným modelom.
- Analýza vplyvu historických dát na presnosť predikcií.
- Skúmanie efektu kurzov stávkových kancelárií na predikcie.
- Posúdenie schopnosti modelov prekonať stávkový trh.

Experimenty poskytujú ucelený pohľad na efektivitu prediktívnych modelov a identifikujú kľúčové faktory úspechu v športových predikciách.

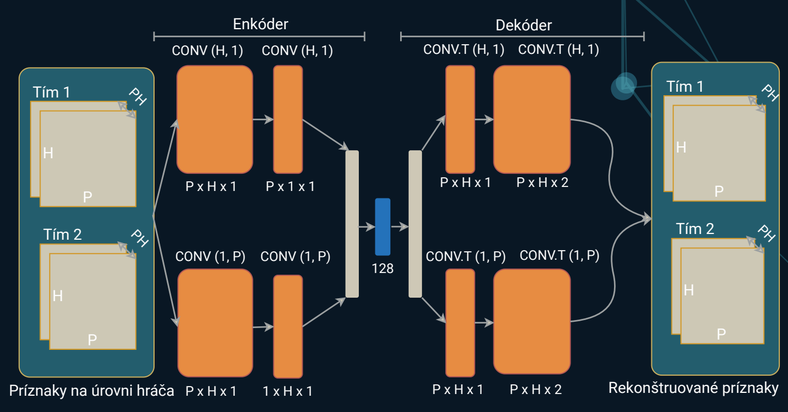
NÁVRH SYSTÉMU

V tejto práci sa používajú 2 typy dát
- dáta na úrovni tímu
- dáta na úrovni hráčov, kde dáta sú reprezentované maticou s hráčmi a ich atribútmi.

Dáta na úrovni hráčov sú transformované do 1D vektoru pomocou konvolučnej neurónovej siete ktorá je trénovaná pomocou autoenkóderu, čo pomáha transformovať dáta tak, aby sa stratilo čo najmenej podstatných informácií. Následne je dekodér odpojený a pripojená je plne prepojená vrstva na ktorej sa to dotrénuje. Výsledkom je pravdepodobnosť udalostí, ktorá sa pripojí do datasetu na úrovni tímov a následne sa trénuje a predikuje udalosť znova, teraz už finálna a už s celým komplexným datasetom a výberom jedného z troch dostupných modelov.



Obr. 1.1: Schéma predikčných modelov systému, ktorá najprv spracováva detailné štatistiky o hráčoch. Výsledkom je pravdepodobnosť konkrétnej predikovanej udalosti. Táto pravdepodobnosť je potom spájaná s datasetom na úrovni tímov, čím vzniká komplexný vstupný dataset. Následne sa tento dataset využíva ako vstup do jedného z troch dostupných modelov, ktorý produkuje konečnú predikciu udalosti.



Obr. 2.1: Konvolučná neurónová sieť trénovaná s využitím autoenkodéra. Cieľom siete je naučiť sa efektívne komprimovať dáta do jednorozmerného vektora, pričom sa snaží zachovať maximum informácií o pôvodných zostavách hráčov jednotlivých tímov. Premenná **H** označuje počet hráčov v matici, zatiaľ čo premenná **P** reprezentuje počet príznakov pre každého hráča. Treťou dimenziou vstupných dát je premenná **PH**, ktorá symbolizuje posty jednotlivých hráčov, pričom informácie o pozícii sú rozkopírované naprieč celým riadkom.

DATASET

Celkový počet vzoriek v datasete je 27 118, čo pokrýva 12 sezón siedmich top futbalových líg. Dáta na úrovni hráčov pochádzajú z populárnej série futbalových hier FIFA, kde hráčom a tímom sú pridelené hodnotenia reflektujúce ich schopnosti ako zakončenie, zrýchlenie, či fyzická kondícia. S viac ako 18 000 hráčmi a dátami z 38 kôl za 12 rokov dosiahlo celkové množstvo hráčskych hodnotení viac ako 8,2 milióna. Na vyhodnotenie modelov sa využilo 25 118 zápasov, zatiaľ čo zvyšných 2 000 slúžilo na testovanie.

DEKORELÁCIA MODELU OD BOOKMAKERA

V stávkovom trhu nie sú stávky takzvane „férové", teda sa vytvára marža $m=p-q$ z pravdepodobnosti udalosti p a ponúkanej pravdepodobnosti q od stávkovej kancelárie, čo znižuje výhry stávkujúcich v dlhodobom horizonte. Ako riešenie sa navrhujú techniky na zníženie korelácie medzi modelom a stávkovými kurzami ako váhovanie vzoriek v prípadoch kde vyhral outsider (udalosť ktorá bola bookmakerom nadhodnotená), alebo modifikácia loss funkcie v neurónových sieťach:

$$L = \sum_{i=1}^{N} \left((p^i - y_i)^2 - c \cdot (p^i - \frac{1}{o_i})^2 \right) \quad (1)$$

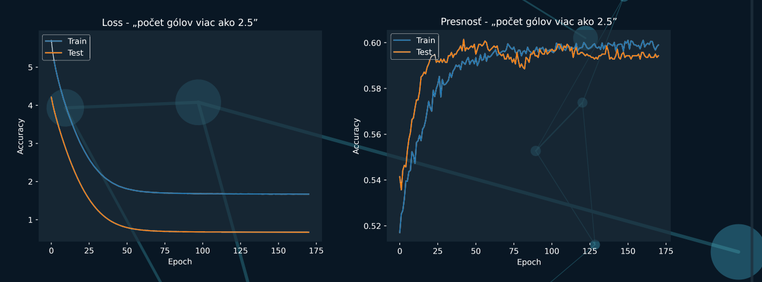
kde p^i je výstup modelu, y_i je skutočný výsledok, o_i sú kurzy, c je konštanta (citlivosť dekorelácie) a N je počet vzoriek.

PRÍZNAKY

Futbal ovplyvňuje množstvo faktorov, ako pripravenosť hráčov, meteorologické podmienky, alebo domáce prostredie. Preto je dôležité identifikovať kľúčové príznaky pre presnú predikciu udalostí. V tejto práci sa analyzovali rôzne štatistiky z viacerých zdrojov na zníženie náhodnosti a lepšie pochopenie dynamiky hry:

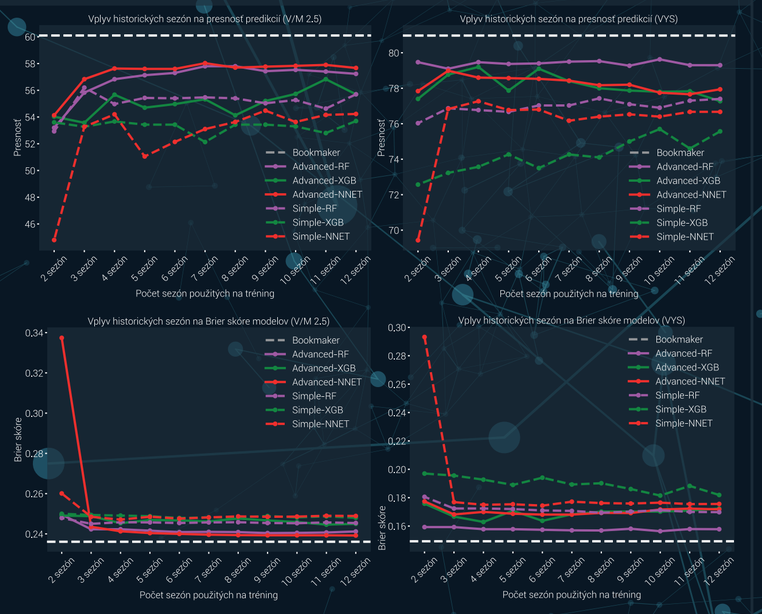
- Štatistiky domácej výhody
- Bodové výsledky a forma tímov
- Štatistiky head-to-head zápasov
- Štatistika očakávaných gólov
- Elo rating

Počet všetkých extrahovanych príznakov činí 661, ale každá udalosť používa iba podmnožinu z týchto všetkých dostupných príznakov.



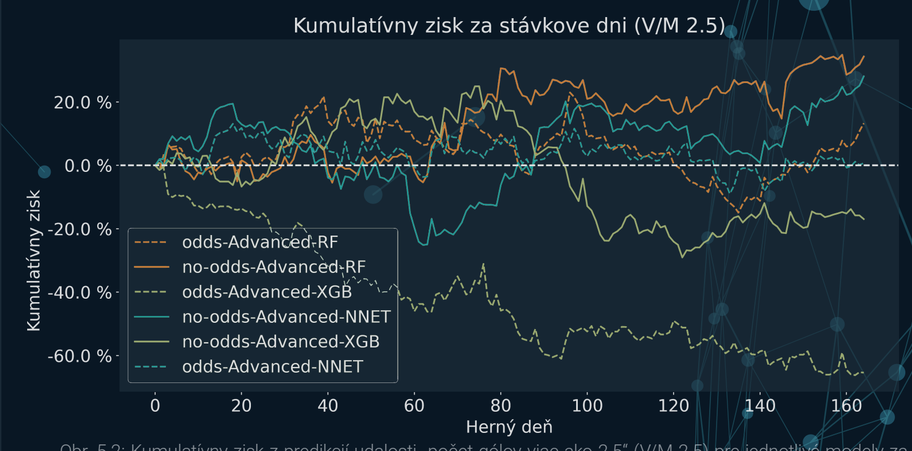
Obr. 3.1: Loss a presnosť neurónovej siete počas trénovania

EXPERIMENT 1

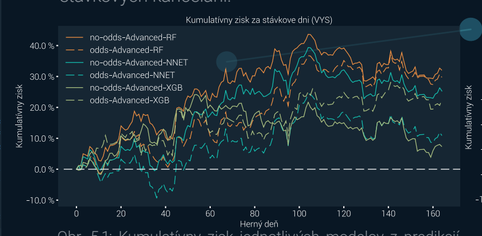


Obr. 4.1: Porovnanie vplyvu historických sezón na presnosť a brier skóre jednotlivých modelov pri predikcii udalostí „výsledok zápasu" (VYS), „počet gólov viac ako 2.5" (V/M 2.5).

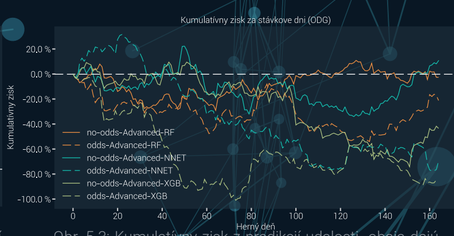
EXPERIMENT 2)



Obr. 5.2: Kumulatívny zisk z predikcií udalosti „počet gólov viac ako 2.5" (V/M 2.5) pre jednotlivé modely za 175 stávkových dní obsahujúcich 2000 zápasov. Pri tejto udalosti boli ziskovejšie modely bez použitia kurzov stávkových kancelárií.

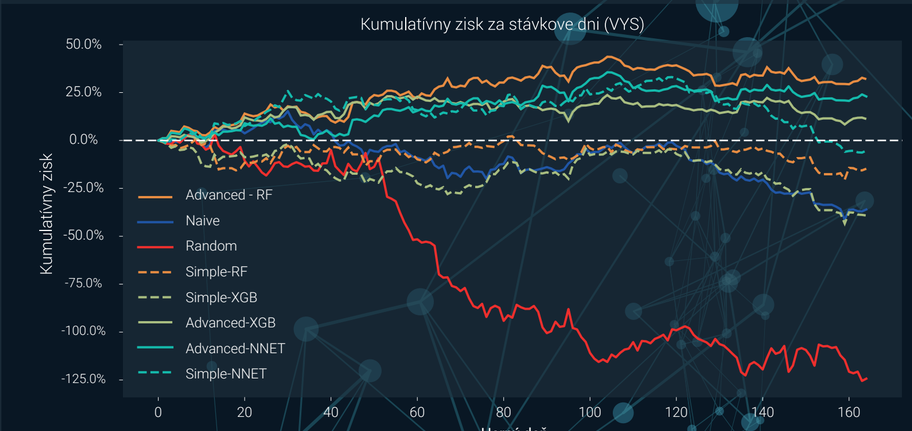


Obr. 5.1: Kumulatívny zisk jednotlivých modelov z predikcií udalosti „výsledok zápasu" (VYS) za 175 stávkových dní. Zaujímavé je, že všetky modely boli ziskové, čo by mohlo byť spôsobené častejším výskytom pravdepodobnejších udalostí s nižšími kurzami v danom období, ak sa modely efektívne naučili predpovedať tieto udalosti, mohlo to viesť k ich ziskovosti.

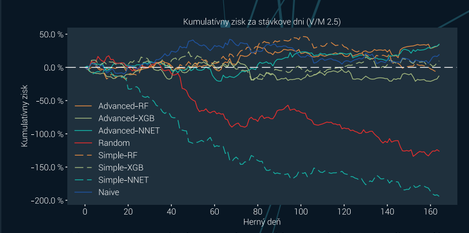


Obr. 5.3: Kumulatívny zisk z predikcií udalosti „obaja dajú gól" (ODG) pre jednotlivé modely počas 175 stávkových dní. Podobne ako pri udalosti V/M 2.5, aj tu boli všetky modely ziskovejšie bez použitia kurzov stávkových kancelárií.

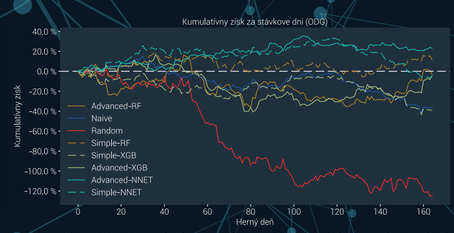
EXPERIMENT 3)



Obr. 6.1: Kumulatívny zisk predikcie udalosti „výsledok zápasu" (VYS) pre jednotlivé modely počas 175 stávkových dní. Väčšina modelov výrazne prekonala základné modely, ktoré by pravdepodobne skončili v bankrote. Najlepšie výsledky ukázali Advanced modely s detailnejšími príznakmi, najmä modely RandomForest (20.84%) a neurónová sieť (16.57%).



Obr. 6.2: Kumulatívny zisk predikcie udalosti „počet gólov viac ako 2.5" (V/M 2.5) pre jednotlivé modely počas 175 stávkových dní. Naivný model, ktorý konzistentne predikoval väčšinovú triedu, sa ukázal byť v tomto časovom horizonte ziskový, čo naznačuje, že modely, ktoré boli ziskové, ale neprekonali jednoduchý model, tak si pravdepodobne neosvojili ani základnú stratégiu predikcie. Najlepšie výsledky predviedli Advanced-NNET (34.9%) a Advanced-RF (34.34%).



Obr. 6.3: Kumulatívny zisk predikcie udalosti „obaja dajú gól" (ODG) pre jednotlivé modely počas 175 stávkových dní. Pri tejto udalosti dominovali modely neurónovej siete a RandomForest. Zaujímavostou je, že model RandomForest dosiahol lepšie výsledky s jednoduchými príznakmi ako s detailnejšími. Najvyšší zisk 20.42% zaznamenal model Advanced-NNET.

	Presnosť	Presnosť (škálovaná)	Brier skóre	ROI	Podobnosť Bookmakera
Výsledok zápasu					
Advanced-NNET	79.27%	50.44%	0.1755	16.57%	0.0258
Simple-NNET	69.37%	49.85%	0.2666	-5.33%	0.0864
Advanced-RF	76.41%	51.15%	0.1732	20.84%	0.0057
Simple-RF	63.11%	49.19%	0.2374	-14.7%	0.0286
Advanced-XGB	76.31%	50.36%	0.1879	11.21%	0.0229
Simple-XGB	64.56%	48.45%	0.2476	-39.18%	0.0442
Naive	70.7%	49.1%	0.293	-36.02%	0.1244
Random	50.8%	47.21%	0.492	-111.66%	0.3231
Počet gólov viac ako 2.5					
Advanced-NNET	59.68%	50.95%	0.2533	34.9%	0.0317
Simple-NNET	43.28%	45.02%	0.2972	-194.08%	0.0509
Advanced-RF	59.1%	50.86%	0.2383	34.34%	0.0046
Simple-RF	56.45%	49.86%	0.2435	-5.43%	0.0087
Advanced-XGB	56.2%	49.69%	0.2439	-12.51%	0.0111
Simple-XGB	55.25%	50.24%	0.2469	9.76%	0.0128
Naive	56.0%	50.48%	0.44	19.19%	0.2259
Random	48.75%	46.85%	0.514	-126.05%	0.2648
Obaja dajú gól					
Advanced-NNET	50.31%	51.56%	0.2501	20.42%	0.0054
Simple-NNET	45.7%	48.52%	0.2803	-22.4%	0.0281
Advanced-RF	55.85%	49.92%	0.2464	-3.26%	0.0039
Simple-RF	55.0%	50.29%	0.2484	11.4%	0.0073
Advanced-XGB	51.77%	49.81%	0.2521	-4.46%	0.014
Simple-XGB	53.35%	49.05%	0.2501	-37.87%	0.0086
Naive	54.4%	49.07%	0.2498	-35.99%	0.2887
Random	47.85%	45.93%	0.2564	-162.93%	0.2648

Tabuľka 6.1:Experimentálne výsledky vykonnosti jednotlivych modelov v porovnaní so základnými modelmi. V tabuľke sú hodnotené metriky ako klasická a škálovaná presnosť, Brier skóre, návratnosť investície (ROI), a podobnosť s predikciami bookmakera.

Obr. 6.4: Kumulatívny zisk z predikcií top troch modelov za 175 stávkových dní obsahujúcich 2000 zápasov. Tieto modely boli natrénované desaťkrát s rôznymi inicializáciami váh, aby sa overilo, či výsledky nezávisia len od vhodnej inicializácie. Výsledné hodnoty boli následne spriemerované.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ | FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Excel@FIT 2024

Maroš Geffert
Ing. Igor Szőke Ph.D.
Ústav počítačové grafiky a multimédií, Fakulta Informačných Technológií
Vysoké Učení Technické v Brně