BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Framework pro modelování a predikci událostí ve fotbale

Maroš Geffert

Abstrakt

V tejto práci sa skúmali súčasné metódy predikcie futbalových udalostí, napríklad počet gólov v zápase, výsledok, alebo či oba tímy dajú gól. Analyzované boli modely neurónovej siete, XGBoost a RandomForest. Boli zhromaždené podrobné historické dáta o zápasoch a hráčoch s cieľom overiť, či detailné príznaky pozitívne ovplyvňujú presnosť predikcií, ako historické dáta ovplyvňujú kvalitu predikcií a či je možné s takýmito modelmi uspieť na stávkovom trhu. Výsledky ukázali, že podrobné štatistiky zvyšujú presnosť predikcií. Experiment nepotvrdil hypotézu o zlepšení presnosti predikcií len s údajmi z posledných sezón, pričom optimálny počet sezón sa líšil pre jednotlivé modely a nie je možné výsledky zovšeobecniť. Modely RandomForest a neurónová sieť dosiahli významné zisky, respektíve 29.04% a 24.61% po analýze 2000 zápasov.

xgeffe00@stud.fit.vutbr.cz, *Fakulta Informačných Technologií, Vysoké Učení Technicke v Brně*

1. 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. Tento výskum využíva modely neurónových sietí [1], XGBoost [2] a RandomForest [3], na zlepšenie presnosti predpovedí tým, že efektívne spracúvajú rozsiahle dátové sady.

Práca sa zameriava na predikciu udalostí ako počet gólov, alebo výsledok zápasu. Vzhľadom na variabilitu v počte gólov, ktorých môže byť kľudne aj 10, a rôznych možných výsledkov zápasu (výhra domácich, remíza, výhra hostí), je predikcia komplexná. Pre uľahčenie sú problémy pretransformované na binárnu klasifikáciu. Pri počte gólov sa predikuje, či ich bude menej alebo viac než stanovená hranica, napríklad 2.5. V prípade výsledkov zápasov sa predpovedá buď výhra domácich, alebo že hostia neprehrajú (alebo naopak).

Hucaljuk a Rakipović [4] testovali modely Random Forest, Bayessovsku sieť a metódu K-Najbližších Susedov, kde niektoré modely dosiahli presnosť nad 60%. Výsledky však boli obmedzené na malý počet zápasov (96), čo môže výsledky skresliť. Zaujímavé bolo, že jednoduchšie príznaky často dosahovali rovnakú alebo vyššiu presnosť ako expertné.

Herbinet [5] vytvoril modely na základe Elo hodnotenia [6], kde SVM s lineárnym jadrom dosiahol najlepšiu presnosť. Použil postupné predikovanie od útočných akcií, cez počet gólov, až po výsledok zápasu.

Hubáček [7] zdôraznil dôležitosť dekorelácie od kurzov stávkových kancelárií pri modelovaní neurónových sietí, čo môže viesť k ziskovosti vďaka identifikácii trhových nekonzistencií.

Tieto štúdie ukazujú potenciál i obmedzenia súčasných modelov pri predikcii futbalových udalostí a naznačujú dôležitosť optimalizácie modelov na základe trhových neefektívností.

Kľúčovým elementom je efektívne využitie rozsiahlych historických dát z rôznych lig, dekorelácia modelu od bookmakera a optimalizácia výberu príznakov.

Môj prístup k predikcii futbalových udalostí spočíva v efektívnom spojení pokročilých analytických metód a hlbokého poznania športových dát. Moje modely neurónovej siete a RandomForest, výrazne prevyšujú základné modely v predikčnej presnosti a dosiahli ziskovosť na stávkovom trhu. Konkrétne zaznamenali zisky 29.04% a 24.61% po analýze 2000 zápasov.

1.1 Hodnotiace metriky v športovej predikcii

Chápať základné metriky ako presnosť a návratnosť investície (ROI) [8] je kľúčové pri skúmaní športových prediktívnych modelov. Presnosť je zavádzajúca, ak distribúcia tried nie je rovnomerná, čo umožňuje

jednoduchým modelom dosiahnuť vysokú presnosť, pričom skutočné ROI môže byť rôzne.

Bookmakeri nastavujú kurzy na základe presnosti predikcií, ale kurz sa mení v reakcii na stávky, čo môže skresliť reálnu pravdepodobnosť výsledkov. Efektívne predikčné modely identifikujú trhové nekonzistencie a využívajú ich na ziskové stávky, nezávisle od celkovej presnosti predikcií.

2. Navrhovaná metóda

2.1 Použité modely

Preto túto prácu sa použili tri typy modelov.

- RandomForest [3]
- XGBoost [2]
- Neurónová sieť [1]

Tieto modely boli vybrané na základe relevantných štúdií, ich experimentálnych výsledkov, popularity a vhodnosti pre riešenie daného problému. Modely ako RandomForest a XGBoost sú schopné efektívne spracovať veľké množstvo príznakov aj pri menšom objeme dát a dobre zvládajú chýbajúce hodnoty, čo je v tejto úlohe kľúčové. Neurónová sieť je zase vynikajúca v zachytávaní komplexných vzťahov a interakcií medzi premennými.

2.2 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 [9]. 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:

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

kde $\hat{p}^i$ je výstup modelu, y_i skutočný výsledok, o_i sú kurzy, c je citlivosť dekorelácie, a N počet vzoriek.

2.3 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:

- Tímové a hráčske štatistiky
- Štatistiky domácej výhody
- Bodové výsledky a forma tímov
- Štatistiky head-to-head zápasov
- Štatistika očakávaných gólov [10]
- Elo rating [6]

Klasifikácia príznakov

Príznaky rozdeľujem do dvoch hlavných kategórií:

- Na úrovni tímu
- Na úrovni hráčov, kde dáta sú reprezentované maticou s hráčmi a ich atribútmi.

Model pre dáta hráčov je trénovaný pomocou konvolučnej neurónovej siete a autoenkodéru, transformujúc maticu na 1D vektory. Po transformácii sa nahrádza dekodér plne prepojenou vrstvou na generovanie predikcií. Tieto predikcie sa kombinujú so štatistikami na úrovni tímu na vytvorenie finálnej predikcie.

2.4 Odhad ziskovosti predikčných modelov

Očakávaný zisk z predikčného modelu sa vypočíta porovnaním predikovanej pravdepodobnosti udalosti p_n s ponúkaným kurzom o_n podľa vzorca:

$$E[Z] = (p_n \cdot (o_n - 1)) - (1 - p_n), \qquad (2)$$

kde p_n je pravdepodobnosť výhry, o_n kurz udalosti, a $E[Z]$ očakávaný zisk z jednotkovej stávky.

Výpočet reflektuje potenciálny zisk a pravdepodobnosť prehry. Pozitívna hodnota $E[Z]$ naznačuje, že stávka je výhodná.

2.5 Experimenty a výsledky

Experimenty boli zamerané na analýzu predikčných modelov z viacerých uhlov:

1. Skúmanie vplyvu historických dát a kurzov na predikcie.
2. Hodnotenie efektivity predikčných modelov v porovnaní s trhom stávkových kancelárií.

3. Výsledky a vyhodnotenie

Experimenty ukázali, že modely RandomForest a neurónové siete efektívne predpovedajú udalosti vo futbale. Použitie kurzov ako prediktívnych príznakov vo všeobecnosti znižuje ziskovosť, ale model XGBoost

v špecifickom prípade ukázal zlepšenie s kurzami. Detailné štatistiky zvyšujú predikčnú efektivitu, umožňujúc modelom dosiahnuť zisky na stávkovom trhu, kde modely RandomForest a neurónová sieť zaznamenali návratnosť investície 29.04% a 24.61%.

Literatúra

[1] James Chen. What is a neural network? online.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[3] Sruthi E R. Understand random forest algorithms with examples. online.

[4] Josip Hucaljuk and Alen Rakipović. Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1623–1627, 2011.

[5] Corentin Herbinet. Predicting football results using machine learning techniques. online, 2018.

[6] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470, 2010. Sports Forecasting.

[7] Ondřej Hubáček, Gustav Šourek, and Filip Železný. Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2):783–796, 2019.

[8] Tim Stobierski. How tp calculate roi to justify a project. online.

[9] Tadgh Hegarty and Karl Whelan. Calculating the bookmaker's margin: Why bets lose more on average than you are warned. 2023.

[10] Fred Garratt-Stanley. What is expected goals (xg)? online.