Acta Mathematica Nitriensia
Vol. 6, No. 1, p. 26 – 36
ISSN 2453-6083

Futbalové bitky o Britániu

Football Battles of Britain

Ľubomír Rybanský[a*] – Marek Varga[b]

[a,b]*Department of Mathematics, Faculty of Natural Sciences, Constatntine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra*

Received: March 25, 2020; received in revised form: March 30, 2020; accepted: April 1, 2020

Abstract

In the article we connect football with mathematics, or more precisely football with statistics. We chose the most famous football match in the world - the North West derby. These matches between Liverpool FC and Manchester United FC are known as Battle of England or Battle of Britain. By using their results from 1906 (only in the First division and later in Premier League), we want to predict the results of the next matches, or at least the winner. Finally we compared our predictions with the predictions of bookmakers.

Keywords: derby, forecast, model, ordinal regression.

Classification: K80, K90

Úvod

Predpovedanie výsledkov športových udalostí ako sú zápasy alebo turnaje priťahuje pozornosť nielen športových priaznivcov, ale i vedeckej komunity už dlhší čas. Veľkej popularite sa tešia najmä futbalové a tenisové zápasy. V súčasnosti sú navyše k dispozícii pomerne rozsiahle databázy výsledkov, ktoré je možné využiť aj na tvorbu nových metód a modelov predpovedania výsledkov. Presnosť predpovedí nie je iba akademickou otázkou, ale aj pomerne lukratívnym ekonomickým odvetvím, pretože presnejšie predpovede znamenajú väčšie zisky. Predpovedanie v športe prostredníctvom matematického modelovania je komplexná úloha, pri riešení ktorej je potrebné vyriešiť tri čiastkové úlohy: *proces hodnotenia* - stanoviť kvalitu tímov, ktorá sa zväčša stanovuje na základe matematických modelov, *proces predpovedania* - nájsť alebo vytvoriť štatistický model, ktorým budeme predpovedať a *proces testovania* - vhodnými štatistickými testami porovnať predpovede s reálnymi dátami [8]. V snahe o porozumenie procesu predpovedania sa skúmali a porovnávali predpovedné schopnosti rôznych zdrojov predpovedí, ktoré sa dajú vo všeobecnosti rozdeliť do štyroch kategórií:

1. Ľudský úsudok – predpoveď sa stanoví na základe predpovedí respondentov s rôznym stupňom znalostí, napríklad športových expertov.

2. Rebríčky – na odhad výsledkov budúcich stretnutí alebo turnajov sa používajú oficiálne rebríčky ako napríklad Rebríček FIFA vo futbale, alebo Rebríček ATP v tenise.

*Corresponding author; email: lrybansky@ukf.sk
DOI: 10.17846/AMN. 2020.6.1.26-36

3. Matematické modely – predpovede športových udalostí sa vytvárajú pomocou existujúcich alebo nových matematických a štatistických prístupov.

4. Stávkové kurzy – k predpovedi výsledku športovej udalosti sa použijú stávkovými spoločnosťami resp. bookmakermi odhadnuté kurzy.

Často používaným a všeobecne akceptovaným matematickým prístupom k predpovedaniu v športe je ELO systém, ktorý bol pôvodne navrhnutý pre šach a v súčasnosti je používaný aj pre hodnotenie športových tímov ale i hráčov [7]. Hvattum a Arntzen v [5] použitím logistickej regresie rozšírili ELO systém tak, aby bol schopný predikovať pravdepodobnosti výsledkov športových stretnutí, kde sú tri možné výsledky (domáci/remíza/hosť) na základe hodnoty ELO. Tento systém tvorby odhadov dosahoval lepšie výsledky ako Goddardov prístup založený na ordinálnej logistickej regresii [3], ale horšie výsledky ako dosahovali bookmakeri. Je potrebné zdôrazniť, že doteraz navrhnuté modely používajú k predpovedaniam aj výsledky iných stretnutí než iba výsledky dvojice tímov, ktoré sa aktuálne snažia predpovedať.

V našom článku sa zaoberáme vytvorením modelu, ktorý by predpovedal pravdepodobnosti výsledkov stretnutí v „The North West derby“ medzi FC Liverpool a Manchestrom United a to iba výlučne na základe výsledkov ich predchádzajúcich súbojov pri zohľadnení ich aktuálnej formy. Nezameriavame sa však na odhad výsledku stretnutia, ale chceme modelovať pravdepodobnosti možných výsledkov. Náš prístup je založený na ordinálnom regresnom modeli, čo by sme mohli nazvať štandardným prístupom.

V prvej časti článku sa zameriame na históriu derby a v ďalšej zdôvodníme opodstatnenosť premenných z ktorých budeme model zostavovať, zostavíme model a nakoniec porovnáme jeho predikčnú schopnosť na tých derby stretnutiach, ktoré neprispievali k vytvoreniu modelu a to porovnaním s predpoveďami bookmakerov. Nebudeme však porovnávať percentuálnu úspešnosť predpovedí na základe výsledkov, ale porovnáme Brierovo skóre, ktoré je pre tento účel vhodnejším nástrojom.

História „The North West Derby“

Pojmom „North West derby“ sa vo futbalovom svete rozumejú derby[*] – zápasy medzi historicky najúspešnejšími anglickými klubmi ležiacimi na severozápade tejto krajiny, a to Liverpool Football Club (ďalej LFC) a Manchestrom United Football Club (ďalej MU). Hoci tieto futbalové kluby vznikli v rokoch 1892 resp. 1902, súperenie týchto miest vzniklo už v dobách priemyselnej revolúcie (prelom 18. – 19. st.) – Liverpool bol prístav, a teda mal geografickú

[*]Ako derby zápasy sa označujú v športe najmä dlhoročné súperenia rivalov z jedného mesta, napr. „Old Firms derby“ (najstarší Celtic Glasgow – Glasgow Rangers), „derby dellacapitale“ (AS Roma – SS Lazio), „derby della Madonina“ (AC Milan – Internazionale Milano), „večiti derby“ (Partizan Beograd – Crvena zvezda Beograd), „intercontinental derby“ (Galatasaray Instanbul – Fenerbahce Istanbul), či najslávnejšie mimoeurópske „Superclasico“ (Boca Juniors Buenos Aires – River Plate Buenos Aires). Tieto už nezriedka vyše storočné súperenia boli dané už v okamihu vzniku spomentých klubov – niektoré reprezentujú chudobnú časť mesta, niektoré bohatšie štvrte, niekedy sú v hre náboženské rozdiely. Neskôr sa ako derby začali označovať aj zápasy medzi nezmieriteľnými rivalmi, ktoré vznikali medzi najúspešnejšími klubmi v daných krajinách. Z tých najslávnejších sem môžeme zaradiť „derby d´Italia“ (Juventus Torino – Internazionale Milan), „der Klassiker“ (Bayern Munchen – Borussia Dortmund), „elclasico“ (FC Barcelona – Real Madrid), „leClassique“ (Paris Saint-Germain – Olympique de Marseiile), „nieuwe Klassieker“ (Ajax Amsterdam – PSV Eindhoven), či napokon nami vybraný zápas LFC – MU (podľa údajov z roku 2011 išlo o najsledovanejší zápas sveta vysielaný do 211 krajín [12]
Vo všeobecnosti sú tieto zápasy známe tým, že málokedy závisí ich výsledok od momentálnej formy, či postavenia v tabuľke – oveľa väčší vplyv má v samotnom stretnutí sústredenie hráčov na zdolanie (nebojíme sa použiť slovo) nenávideného protivníka.

výhodu pri obchodovaní so svetom. Koncom 19. st. však vybudovali lodný kanál až priamo do Manchestru (vzdialeného menej ako 60 km), čo nepriaznivo ovplyvnilo zamestnanosť a rozvoj Liverpoolu.

Prvoligové súperenie v najvyššej anglickej futbalovej súťaži sa začalo 25.12.1906 na štadióne Bank Street bezgólovou remízou a len nedávno (19.1.2020) sa dočkalo svojho 170 pokračovania, v ktorom boli úspešnejší hráči LFC na svojom domovskom stánku Anfield v pomere 2:0. V týchto súbojoch sú zatiaľ mierne úspešnejší hráči z Manchestru, ktorí zvíťazili v 66 stretnutiach, kým ich súper iba v 56 stretnutiach. Remízou, ale sotva zmierlivo sa rozišli v 48 prípadoch. United sú úspešnejší aj počte strelených gólov (229:210). Na tomto mieste musíme poznamenať, že do nášho modelu sme použili len výsledky z prvoligových zápasov (prvá liga sa v Anglicku nazývala First Division, od 1992 Premier League), hoci LFC a MU sa stretli samozrejme i v anglických pohároch a dokonca aj v európskej pohárovej súťaži.

Ak vyberieme zápasy s vysokým skóre, v LFC zrejme radi spomínajú na výhry 7 : 4 či 5 : 0, fanúšikom MU sa zrejme páčili zápasy s výhrami 6 : 1 či 4 : 0. Zaujímavejšie je spomenúť dva zápasy, ktoré skončili stavom 3 : 3. Prvý sa odohral v sezóne 1987/88, keď sa majstrom stal LFC, ale oslavy mu remízou naštrbil tradičný rival, hoci MUFC už prehrával 1 : 3. Zrkadlová situácia nastala v roku 1998, keď sa majstrom stal MU – vo vzájomnom zápase viedol už 3 : 0 po 25 minútach, ale LFC žiadnu potupu nedovolil, a napokon remizoval (hoci v tabuľke skončil až na 8. mieste).

Zdroj:
https://en.wikipedia.org/wiki/Liverpool_F.C

Zdroj
https://en.wikipedia.org/wiki/Manchester_United_F.C.

Rivalita medzi klubmi sa samozrejme prenáša do vzťahov medzi hráčmi. Od roku 1964 sa neuskutočnil žiadny transfer medzi týmito klubmi (aj keď sú známe výnimočné prechody cez tretí klub), hráči určite nemajú záujem byť označený v choráloch fanúšikov za „zradcov“ resp. „judášov“. Vzhľadom na kvalitu klubov treba priznať, že ak sa niekto už stal legendou LFC či MU, automaticky šlo aj o legendu svetového futbalu (spomeňme aspoň mená Best, Keegan, Dalglish, Cantona, Giggs, Gerrard...). Všetci hráči samozrejme priznávajú, že North West derby

je oveľa dôležitejší zápas ako ich lokálne derby, t.j. „Manchester derby“ (MU – Manchester City) a „Merseyside derby“ či „friendly derby“ (nenechajme sa však pomýliť názvom, v zápasoch LFC – Everton Liverpool je historicky najvyšší počet červených kariet v Premier League). Zo všetkých príbehov snáď stačí spomenúť posledný štart legendy LFC Stevena Gerrarda („captain fantastic“), ktorý vystriedal do druhého polčasu premotivovaný a do červenej karty strávil na ihrisku 38 sekúnd...

Samostatnou kapitolou tejto rivality sú fanúšikovia oboch tímov. Tento text však naozaj nie je vhodným miestom na to, aby sme mohli popísať chorály určené nenávideným súperom, a tak sa radšej vrátime k najväčším historickým úspechom oboch klubov. LFC vyhral 18 krát anglickú ligu, 6 krát európsku ligu majstrov, celkovo má vo vitríne 63 trofejí. MU bol 20 krát majstrom ligy, 3 krát európskym majstrom, celkovo vlastní 66 víťazných pohárov.

V článku sme sa pokúsili odhadnúť výsledky týchto zápasov, niekedy nazývaných aj „Bitka o Britániu“, pomocou istých vstupných údajov. Keďže ľahko nájdeme v histórii nečakané prekvapenia, zaváhania favoritov či víťazstvá jasných outsiderov, namodelovať rezultáty North West derby rozhodne nebude štandardnou úlohou...

Pravdepodobnosti a model ordinálnej regresie

Stávkové kancelárie štandardne neinformujú o šanciach tímov uvedením pravdepodobnosti výsledku stretnutia, ale formou kurzov. Napríklad v poslednom stretnutí LFC a MU, ktoré sa 19. januára 2020 na Anfield Road skončilo víťazstvom domácich 2:0, boli stávkovou kanceláriou Bet365 vypísané kurzy 1,45-4,68-7,49. Je zrejmé, že nižší kurz má udalosť, ktorá je pravdepodobnejšia a opačne vyšší kurz má udalosť, ktorá je menej pravdepodobná. Z kurzu *k* vieme veľmi jednoducho vypočítať tzv. *implikovanú pravdepodobnosť* udalosti *p* pomocou vzťahu $p = 1/k$. Pre uvedené kurzy by to znamenalo, že implikovaná pravdepodobnosť víťazstva LFC bola $\frac{1}{1{,}45} = 0{,}69$, implikovaná pravdepodobnosť remízy bola $\frac{1}{4{,}68} = 0{,}21$ a implikovaná pravdepodobnosť víťazstva MU bola $\frac{1}{7{,}49} = 0{,}13$. Avšak súčet uvedených implikovaných pravdepodobností $0{,}69 + 0{,}21 + 0{,}13 = 1{,}04$ presahuje hodnotu 1. Hodnota, o ktorú súčet implikovaných pravdepodobností presahuje hodnotu 1 presahuje, predstavuje zisk stávkovej kancelárie z celkovej prestávkovanej sumy (nazýva sa *vigorish* alebo skrátene *vig*), čo sú v tomto prípade 4%. Opačný proces, teda stanovanie kurzu z odhadnutej implikovanej pravdepodobnosti je zrejmý, avšak pri udalostiach, ktorých pravdepodobnosť je veľmi vysoká sa kurz stanovuje inak. Detaily možno nájsť napríklad v článku [4].

Na modelovanie pravdepodobnosti výsledku futbalového stretnutia, ktoré sa môže skončiť tromi rôznymi výsledkami, ako funkcie premenných vzťahujúcich sa k tomuto stretnutiu, je možné použiť ordinálnu logistickú regresiu. Model ordinálnej regresie, ktorý uvedieme, je spracovaný podľa [1] a [6].

Ak modelujeme ordinálnu závisle premennú *y* s *r* usporiadanými kategóriami, tak táto môže vzniknúť zo spojitej náhodnej premennej y^* a z prahov $\theta_0, \dots, \theta_r$ tak, že

$$y = j \; ak \; \theta_{j-1} \leq y^* < \theta_j; j = 1{,}2, \dots, r.$$

Hovoríme, že premenná *y* je kategorizovaná verzia premennej y^*. Napríklad, ordinálna výstupná premenná „Výsledok futbalového stretnutia“ je reprezentovaná ako ohraničená verzia pôvodnej premennej vyjadrujúcej rozdiel v počte strelených gólov súperiacich tímov.

Model pre ordinálnu premennú *y* môžeme vyjadriť kumulatívnymi pravdepodobnosťami τ_j:

$$\tau_j \equiv P(y \leq j) = P(y^* < \theta_j), \qquad j = 1, \dots, r.$$

Cieľom je nájsť vzťah medzi kumulatívnymi pravdepodobnosťami τ_j a vysvetľujúcimi premennými $x_1, x_2, \dots, x_k$.

Predpokladáme, že $y^* = -\boldsymbol{x}'\beta + \varepsilon$, s $E(\varepsilon) = 0$, z čoho vyplýva, že $E(y^*) = -\boldsymbol{x}'\beta$ a teda

$$\tau_j = P(\varepsilon \leq \theta_j + \boldsymbol{x}'\beta).$$

Presný tvar modelu je určený rozdelením ε.

V prípade, že ε má štandardizované logistické rozdelenie:

$$P(\epsilon \leq x) = \frac{1}{1 + e^{-x}}$$

hovoríme o kumulatívnom logistickom modeli resp. o modeli proporcionálnych šancí.

Potom

$$\tau_j = P(\varepsilon \leq \theta_j + \boldsymbol{x}'\beta) = \frac{1}{1 + e^{-(\theta_j + \boldsymbol{x}'\beta)}}, \tag{1}$$

a odtiaľto

$$ln\frac{\tau_j}{1 - \tau_j} = \theta_j + \boldsymbol{x}'\beta, \quad j = 1, \dots, r-1, \tag{2}$$

kde θ_j sú absolútne členy, ktoré závisia iba od *j* ak koeficient β nezávisí od *j*, tak všetkých $r - 1$ modelov v (2) má vzhľadom k vysvetľujúcej premennej *x* rovnakú smernicu. Po odhade parametrov modelu získame odhady pravých strán v (2) z ktorých vypočítame pravdepodobnosti $P(y = j)$ hodnôt náhodnej premennej *y*

$$P(y = j) = P(y \leq j) - P(y \leq j - 1) = \hat{\tau}_j - \hat{\tau}_{j-1}.$$

Hodnotenie kvality modelu logistickej regresie sa dá posúdiť rôznymi metódami. Na testovanie zhody dát s modelom predikovanými hodnotami sa používa *Hosmer-Lemeshowov test* (ak sú vysvetľujúce premenné spojité), prípadne *Pearsonov chí-kvadrát test*. *Testom pomerom vierohodnosti* (LR test) testujeme, či je vhodné do modelu zahrnúť ďalšie parametre. V situácii keď sa rozhodujeme o výbere najvhodnejšieho z viacerých vzájomne si konkurujúcich modelov, je výhodné použiť informačné kritériá, prípadne štatistiku *deviance*, ktorá je definovaná ako záporne vzatý dvojnásobok rozdielu logaritmu vierohodnosti odhadovaného modelu a logaritmu vierohodnosti saturovaného modelu. Najčastejšie využívanými sú *Akaikeho informačné kritérium* (AIC) a *bayesovské informačné kritérium* (SIC), pre ktoré platí, čím menšia hodnota, tým lepší model.

Na stanovenie úspešnosti predpovedí v *N* položkách, z ktorých každá predpoveď je vyjadrená *R*-ticou pravdepodobností, ktorej každý člen $f_{ti}, t = 1, \dots, N;\ i = 1, \dots, R$ vyjadruje odhadovanú pravdepodobnosť nastania každej z *R* možných udalostí sa *Brierovo skóre B*, ktoré je definované vzťahom

$$B = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2,$$

kde o_{ti} je binárna premenná nadobúdajúca hodnotu 1, ak sa v *t* - tej položke realizovala *i* - ta udalosť a hodnotu 0, ak sa realizovala iná ako *i* - ta z *R* -tice možných udalostí. Čím nižšia je hodnota Brierovho skóre, tým presnejšie sú predpovede. Minimálna a súčasne najlepšia hodnota Brierovho skóre je 0. Maximálna hodnota pre binárne položky je 1 a pre položky s tromi možnými výsledkami 2. Bez ohľadu na počet možných výsledkov, je možné Brierovo skóre upraviť tak, aby bola maximálna hodnota rovná 1.

Výstavba modelu a dáta

Pri zostavovaní vhodného modelu sme vychádzali zo 160 stretnutí medzi LFC a MU, pričom desať stretnutí z posledných 26 zápasov sme náhodne vybrali pre porovnanie úspešnosti predikčnej schopnosti modelu s predpoveďami bookmakerov. Aby bolo možné porovnanie realizovať, museli sme mať k dispozícii stávkové kurzy pre tieto stretnutia. Najstaršie nám dostupné záznamy z [13]siahajú do sezóny 2007/2008 a to k zápasu hranému 16.12.2007. Spolu to je práve 26 stretnutí, z ktorých sme náhodne vybrali už spomínaných desať. Dodatočné informácie o hracom kole ligového ročníka v ktorom sa tímy stretli a počte získaných bodov do dňa stretnutia v danom ligovom ročníku sme získali z [11].

Našim cieľom je zostaviť model, ktorý by odhadoval pravdepodobnosť výsledku NWD a to iba z historických výsledkov ich vzájomných súperení a aktuálnej formy tímov.

Modelovanou ordinálnou závisle premennou *NWD* je výsledok stretnutia z pohľadu LFC, ktorá nadobúda tri hodnoty: *Liv* – víťazstvo Liverpoolu, *Remíza* – remíza, *MU* – víťazstvo Manchestru Utd.; $Liv > Remíza > MU$).

Ktoré ukazovatele by mohli byť užitočné pri odhade pravdepodobností výsledkov stretnutia?

Vo futbalových stretnutiach hrá nesporne dôležitú úlohu výhoda domáceho prostredia a nie inak tomu je v North West Derby, pretože LFC má bilanciu v domácom prostredí 38-22-24 a MU dokonca až 42-26-18. Premenná *DV* vyjadruje, ktorý z tímov bol domácim a nadobúda dve hodnoty: *Liv* (domácim tímom je LFC) a *MU* (domácim tímom je MU) Pre zohľadnenie efektu rivality sme uvažovali výsledok posledného vzájomného ligového stretnutia (premenná *PZ,* ktorá je ordinálna a nadobúda tri hodnoty: *Liv* – víťazstvo LFC, *Remíza* – remíza, *MU* – víťazstvo MU; $Liv > Remíza > MU$). Z historických údajov z pohľadu LFC vyplýva, že po víťazstve v NWD bol LFC v nasledujúcom NWD úspešnejší (25-13-17), po remíze bol mierne úspešnejší MU (15-13-20) a po prehre LFC bol úspešnejší MU (16-21-29). Rozdiel vo výkonnosti tímov v príslušnej sezóne sme vyjadrili premennou *Rozdiel*. Pri jej definovaní sme museli zohľadniť v ktorom kole (hracom dni) súťažného ročníka sa tímy stretli, pretože rozdiel napríklad 12 bodov po dvadsiatom kole a po desiatom kole je kvalitatívne rozdielny. Druhý prípad signalizuje väčší rozdiel vo výkonnosti tímov, než prvý prípad. Ponúka sa tak riešenie, definovať rozdiel vo výkonnosti tímov ako podiel počtu získaných bodov do vzájomného stretnutia a počtu odohratých stretnutí. Tento prístup by však nebol korektný, pretože v roku 1981 sa zmenil bodovací systém z pôvodného: 2 body za výhru, 1 bod za remízu a 0 bodov za prehru, na trojbodový systém: 3 body za výhru, 1 bod za remízu a 0 bodov za prehru. Preto sme namiesto bodového rozdielu uvažovali *rozdiel* vyjadrený v počte zápasov, ktorý by pri odstupe 12 bodov pri dvojbodovom systéme predstavoval šesť stretnutí, ale pri trojbodovom systéme iba štyri stretnutia. Premennú *Rozdiel* sme zaviedli ako podiel *rozdielu* vyjadreného v zápasoch a počtu odohratých stretnutí do vzájomného stretnutia. Kladné hodnoty premennej *Rozdiel* zodpovedajú prípadom keď mal v tabuľke navrch LFC a záporné zasa prípadom keď bol úspešnejší MU.

Pri výstavbe modelu sme uvažovali aj o interakciách premenných, ktoré je možné teoreticky zdôvodniť nasledovne:

- *DV×PZ* - efekt domáceho prostredia sa môže vzhľadom na výsledok posledného vzájomného stretnutia líšiť, pretože je zrejme rozdiel v tom, či bolo posledné víťazstvo dosiahnuté doma alebo vonku,
- *DV×Rozdiel* - ten istý výkonnostný rozdiel môže mať iný vplyv na šance tímov v domácom a v súperovom domácom prostredí,
- *PZ×Rozdiel* - pri tejto interakcii by sa uplatňoval iný efekt rozdielu vo výkonnosti vzhľadom na výsledok posledného vzájomného stretnutia.

Hodnoty kritérií (deviance a AIC) pre viacero uvažovaných modelov sú uvedené v tabuľke 1. Ukazuje sa, že najnižšie hodnoty a teda za najlepší spomedzi uvažovaných modelov môžeme považovať model v ktorom vystupujú všetky tri uvažované premenné (*DV, Rozdiel, PZ*) spolu s interakciou domáceho prostredia a výsledkom posledného vzájomného stretnutia.

Tab. 1: *Výber modelu pre výsledok NWD. Najlepší model je hrubo zvýraznený. (Δ - deviance, AIC - hodnota Akaikeho informačného kritéria)*

Model	**Δ**	**AIC**
DV	344,96	350,96
PZ	354,43	359,43
Rozdiel	347,03	353,03
DV+PZ	329,60	339,60
DV+Rozdiel	336,48	345,48
PZ+Rozdiel	340,62	350,62
DV+PZ+Rozdiel	328,21	340,21
DV+PZ+Rozdiel+DV×PZ	**321,81**	**337,81**
DV+PZ+Rozdiel+DV×Rozdiel	326,96	340,96
DV+PZ+Rozdiel+PV×Rozdiel	327,55	343,55
DV+PZ+Rozdiel+DV×PZ+DV×Rozdiel	321,12	339,12
DV+PZ+Rozdiel+DV×PZ+PZ×Rozdiel	321,29	341,29
DV+PZ+Rozdiel+PZ×Rozdiel+DV×Rozdiel	326,02	344,02
DV+PZ+Rozdiel+DV×PZ+PZ×Rozdiel+DV×Rozdiel	320,45	342,45

Model pre odhad pravdepodobností výsledkov NWD

Pre model, ktorý sme na základe AIC a hodnoty deviance vybrali ako najlepší, sme odhadli parametre, ktorých hodnoty, štatistickú významnosť a 95% intervaly spoľahlivosti uvádzame v tabuľke 2.

Tab. 2: Výsledok NWD, ordinálna regresia.

		$\hat{\beta}$	SE	$OR = e^{\hat{\beta}}$	t	p	-95%	+95%
Abs. člen 1 (MU)		-0,05	0,21	0,96	-0,21	0,834		
Abs. člen 2 (Remíza)		1,28	0,24	3,59	5,31	<0,001		
DV	1							
MU	0	0,00	0,00	1,00				
Liv	1	1,10	0,32	3,02	3,46	0,001	1,63	5,71
PZ	2							
MU	0	0,00	0,00	1,00				
Remíza	1	0,05	0,37	1,05	0,12	0,901	0,51	2,17
Liv	1	-0,23	0,36	0,79	-0,66	0,511	0,39	1,59
Rozdiel	1	0,83	0,65	2,28	1,27	0,203	0,66	8,47
DV x PZ	2							
Liv, MU	0	0,00	0,00	1,00				
Liv, Remíza	1	1,21	0,53	3,34	2,26	0,024	1,19	9,68
Liv, Liv	1	0,66	0,56	1,94	1,18	0,239	0,64	5,88

($\hat{\beta}$ – odhad parametra β, SE – štandardná chyba odhadu parametra β, OR – pomer šancí, t – testová štatistika Waldovho testu štatistickej významnosti parametra, *p* – p hodnota pre Waldov test, -95% - dolná hranica 95% intervalu spoľahlivosti pre OR, +95% - horná hranica 95% intervalu spoľahlivosti pre OR).

Z analýzy 160 vzájomných stretnutí NWD ordinálnou regresiou pre závisle premennú *NWD* s úrovňami: MU ($j = 1$), Remíza ($j = 2$), Liv ($j = 3$) vyplýva, že kumulatívne pravdepodobnosti $\tau_j, j = 1{,}2{,}3$, ktoré sú modelované premennými *PZ, DV* a *Rozdiel* sú

$$ln\frac{\hat{\tau}_j}{1-\hat{\tau}_j} = \hat{\theta}_j + 1{,}10x_1 + \cdots + 0{,}66x_{6,}, \qquad j = 1{,}2{,}3,$$

kde $\hat{\theta}_1 = -0{,}05$ (víťazstvo MU), $\hat{\theta}_2 = 1{,}28$ (remíza). Premenná x_1 je indikátorová premenná pre domáci tím, x_2, x_3 pre výsledok posledného vzájomného zápasu, x_4 pre rozdiel vo výkonnosti a x_5, x_6 pre interakciu domáceho prostredia a výsledku posledného vzájomného stretnutia.

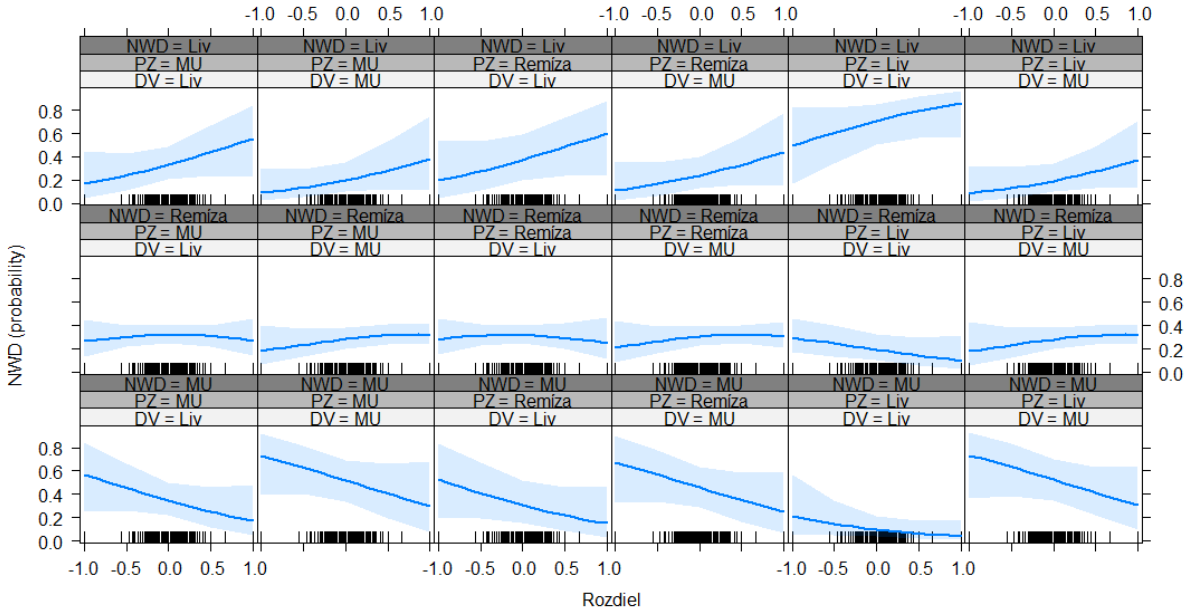
Výsledok NWD menší ako *j*, pre $j = 1{,}2$, znamená buď prehru Liverpoolu ($j = 1$), alebo remízu, prehru LFC, $j \leq 2$.

Ukazuje sa, že výkonnostný rozdiel jednej "jednotky" premennej *rozdiel* má pre LFC za následok 2,28 násobné (o 128%) zvýšenie šance dosiahnuť výsledok *j* alebo lepší.

Keďže model obsahuje interakciu domáceho prostredia a výsledku posledného vzájomného stretnutia, tak efekt týchto premenných musíme vyhodnotiť súčasne. Referenčnou kategóriou pre domáce prostredie je *MU* a pre výsledok posledného vzájomného stretnutia je víťazstvo *MU* (najnižšia kategória z pohľadu Liverpoolu). Potom napríklad šanca, že LFC v domácom prostredí po víťazstve v predchádzajúcom vzájomnom súboji dosiahne výsledok *j* alebo lepší,

je $3{,}02 \times 0{,}79 \times 1{,}94 = 4{,}64$ krát väčšia v porovnaní s tým, ak predchádzajúci vzájomný zápas vyhrali hráči MU a nasledovný zápas by hrali v domácom prostredí.

Z praktického hľadiska je však názornejšie uviesť priamo pravdepodobnosti hodnôt modelovanej závisle premennej v závislosti od hodnôt prediktorov. Tieto uvádzame v grafickej podobe na obrázku 1. Môžeme si všimnúť, že pravdepodobnosť nerozhodného výsledku sa nezávisle od toho, ktorý z tímov hrá v domácom prostredí a akým výsledkom sa skončil posledný vzájomný zápas, pohybuje približne na úrovni 0,3 a výraznejší vplyv na ňu nemá ani rozdiel vo výkonnosti tímov. Pravdepodobnosť víťazstva LFC sa bez ohľadu na hodnoty zvyšných dvoch prediktorov so zvyšujúcim sa výkonnostným rozdielom medzi tímami zvyšuje (horná šestica grafov na obrázku) a naopak pravdepodobnosť prehry sa znižuje (dolná šestica grafov na obrázku). Neprekvapuje, že pravdepodobnosť výhry LFC podľa zostaveného modelu je najväčšia v prípade, že sa hrá na Anfield Roade a predchádzajúci zápas dopadol lepšie pre "The Reds".



Obr. 1: *Modelom predikované pravdepodobnosti výsledku NWD v závislosti od hodnôt prediktorov.*

Hosmer – Lemeshowov test $(\chi^2(17) = 10{,}45; p = 0{,}883)$ indikuje, že medzi pozorovanými a modelom odhadovanými početnosťami nie je rozdiel, čiže model vykazuje s dátami dobrú zhodu. Nagelkerkeov koeficient pseudo-R^2 nadobúda hodnotu 0,18.

Porovnanie predikcií modelu a bookmakerov

Predikčnú schopnosť modelu sme skúmali na desiatich náhodne vybratých stretnutiach z obdobia rokov 2008-2018 ku ktorým sme vedeli dohľadať aj predikcie bookmakerov. V databázach kurzov pre anglickú Premier League na internetovej stránke [13] sa nachádzajú kurzy viacerých stávkových kancelárií: Bet365, BlueSquare, Bet&Win, Gamebookers, Interwetten, Ladbrokes, Pinnacle, SportingOdds, Sportingbet, Stan James, Stanleybet, VC Bet, William Hill. Nie je prekvapujúce, že odhady bookmakerov z rôznych stávkových kancelárií nie sú totožné, ale rozdiely nebývajú veľké. Z tohto dôvodu sme náhodne vybrali dve z týchto kancelárií (Bet365, Pinnacle) pre ktoré sme zo zverejnených kurzov po zohľadnení *vig* vypočítali nimi odhadované pravdepodobnosti výsledkov. Tieto sú spolu s odhadmi výsledkov získanými našim modelom uvedené v tabuľke 3.

Tab.3: *Pravdepodobnosti výsledkov v desať náhodne vybraných stretnutiach odhadnutých nami vytvoreným modelom a stanovených bookmakermi (Bet*365*, Pinnacle).(*1 - *pravdepodobnosť víťazstva domáceho tímu,* X *– pravdepodobnosť nerozhodného výsledku,* 2 *– pravdepodobnosť víťazstva hosťujúceho tímu).*

			Model				Bet365				Pinnacle		
Dátum	**Zápas**	**Skóre**	**1**	**X**	**2**		**1**	**X**	**2**		**1**	**X**	**2**
23.3.2008	MU-LIV	3–0	**0,54**	0,28	0,18		**0,48**	0,28	0,24		**0,48**	0,30	0,22
13.9.2008	LIV-MU	2–1	**0,40**	0,32	0,28		**0,33**	0,31	0,36		**0,34**	0,32	0,34
25.10.2009	LIV-MU	2–0	**0,63**	0,24	0,13		**0,39**	0,30	0,32		**0,38**	0,30	0,32
6.3.2011	LIV-MU	3–1	**0,23**	0,30	0,47		**0,31**	0,28	0,41		**0,34**	0,30	0,37
11.2.2012	MU-LIV	2–1	**0,45**	0,31	0,25		**0,51**	0,28	0,21		**0,53**	0,26	0,21
23.9.2012	LIV-MU	1–2	0,16	0,26	**0,58**		0,31	0,29	**0,40**		0,36	0,29	**0,35**
17.1.2016	LIV-MU	0–1	0,28	0,32	**0,40**		0,39	0,29	**0,31**		0,39	0,29	**0,32**
17.10.2016	LIV-MU	0–0	0,34	**0,32**	0,34		0,46	**0,28**	0,27		0,45	**0,28**	0,28
14.10.2017	LIV-MU	0–0	0,27	**0,31**	0,42		0,38	**0,28**	0,35		0,38	**0,28**	0,34
16.12.2018	LIV-MU	3–1	**0,38**	0,32	0,30		**0,62**	0,22	0,16		**0,64**	0,22	0,14

Vidíme, že napríklad pre stretnutie hrané 25.10.2009 model odhadol pravdepodobnosť víťazstva domáceho tímu ako pomerne vysokú (0,63) v porovnaní s bookmakermi, kým 23.9.2012 odhadol šance domáceho tímu výrazne nižšie ako bookmakeri. Podobnosť medzi modelom a bookmakermi je zrejmá pri odhade pravdepodobnosti nerozhodného výsledku, ktorá sa bez ohľadu na to, ktorý tím je domáci, pohybuje v rozpätí 0,24-0,32 pre náš model a 0,22-32 pre bookmakerov a vykazuje tak pomerne nízku variabilitu. Z hľadiska porovnania predikcií je však rozhodujúcim ukazovateľom Brierovo skóre, ktoré dosahuje hodnotu 0,26 pre odhady získané modelom a hodnotu 0,29 pre odhady oboch stávkových kancelárií. Model v porovnaní s dvojicou stávkových kancelárií dosahuje v desiatich náhodne vybratých stretnutiach mierne lepšiu predikčnú schopnosť.

Záver

V článku sme sa zaoberali modelovaním pravdepodobností výsledku „Bitky o Britániu“, ktorá sa už viac než jedno storočie odohráva na futbalových trávnikoch. Ide o jedno z najslávnejších futbalových derby na svete. Nami zvolený prístup k modelovaniu sa od štandardných prístupov líši v tom, že zostavený model využíva výlučne údaje pochádzajúce z duelov týchto tímov. Napriek tomu, že vytvorený model bol mierne presnejší ako odhady bookmakerov, tak je potrebné vziať do úvahy skutočnosť, že porovnávacia vzorka bola veľmi malá. Model neberie do úvahy iné ukazovatele, ako napríklad hráčska maródka alebo vyťaženosť tímov v rámci sezóny, ktoré pri odhadoch využívajú bookmakeri. Tieto a iné aktuálne informácie nezahrnuté do modelu mohli byť dôvodom prečo sa naše odhady líšili od odhadov bookmakerov, čo však nechápeme ako nedostatok, ale dôsledok inakosti tvorby modelov. Autorom článku nie je z literatúry, ani iných zdrojov známe, že v článku prezentovaný postup by už realizovali iní autori.

Literatúra

1. Agresti, A. 2002. *Categorical data analysis 2nd edition*. New York: John Wiley and Sons.
2. Forrest, D., Goddard, J., Simmons, R. 2005. *Odds-setters as forecasters. The case of English football*. International Journal of Forecasting 21 (3): 551-564.

3. Goddard, J. 2005. *Regression models for forecasting goals and matchresults in association football*. International Journal of Forecasting 21 (2): 331-340.
4. Graham, I., Stott, H. 2008. *Predicting bookmaker odds and efficiency for uk football*. Applied Economics, 40:99–109.
5. Hvattum, L. M., Arntzen, H. 2010. *Using ELO ratings for match result prediction in association football*. International Journal of Forecasting 26 (3): 460-470.
6. de Jong, P., Heller, G. Z. 2008. *Generalized linear models for insurance data*. CambridgeUniversity Press.
7. Lasek, J., Szlávik, Z., Bhulai, S. 2013. *The predictive power of ranking systems in association football*. International Journal of Applied Pattern Recognition, 1 (1), 27-46.
8. Wunderlich, F., Memmert, D. 2018. *The Betting Odds Rating System: Using soccer forecasts to forecast soccer*. PLoS ONE 13(6):e0198668. https://doi.org/10.1371/journal.pone.0198668
9. https://www.eloratings.net/
10. https://en.wikipedia.org/wiki/Liverpool_F.C.%E2%80%93Manchester_United_F.C._rivalry
11. https://www.enfa.co.uk/
12. https://www.dailymail.co.uk/sport/football/article-2049908/211-countries-tuned-watch-Liverpool-vs-Manchester-United.html
13. https://www.football-data.co.uk/englandm.php