Univerza v Ljubljani

Fakulteta za elektrotehniko, Biotehniška fakulteta,
Ekonomska fakulteta, Fakulteta za družbene vede,
Fakulteta za matematiko in fiziko, Fakulteta za
računalništvo in informatiko, Medicinska fakulteta

Urh Peček

# Mera pričakovanih zadetkov v nogometu in njena uporaba

Magistrsko delo

Magistrski študijski program druge stopnje Uporabna statistika

Mentor: doc. dr. Nataša Kejžar

Ljubljana, 2022

# Vsebina

# Seznam slik

# Seznam tabel

# Povzetek

V delu je obravnavana mera pričakovanih zadetkov v nogometni igri. Koncept in izračun pričakovanih zadetkov je predstavljen tako teoretično kot praktično. Opisane so spremenljivke, za katere menimo, da v nogometni igri vplivajo na vrednosti pričakovanih zadetkov. Analiziran je vpliv porazdelitve vrednosti pričakovanih zadetkov in njihova uporaba za kvantifikacijo uspešnosti igralca ali ekipe. Na podlagi pričakovanih zadetkov je izpeljana in praktično predstavljena mera pričakovanih točk. Predstavljena je metoda napovedovanja rezultatov nogometnih tekem, ki temelji na Poissonovi porazdelitvi in je posodobljena z upoštevanjem pričakovanih zadetkov. Na praktičnem primeru so primerjane metode z in brez upoštevanja pričakovanih zadetkov.

Obravnavana tema se trenutno zelo hitro razvija. Večina zamisli je objavljena v polstrokovnih člankih na spletu. V pričujočem delu so prevedene v matematični oziroma statistični jezik ter urejeno in celostno predstavljene, dodane pa so tudi nekatere nove, avtorjeve zamisli. Vsem teoretično predstavljenim konceptom so dodani praktični primeri.

Vse analize, simulacije in rezultati so pridobljeni s pomočjo računalniškega statističnega programa R. V sklopu tega so uporabljeni tudi nekateri specifični paketi, kot so *regista*, *StatsBombR*, *ggsoccer*, *SBpitch*, *worldfootballR* in *soccermatics*.

Uporabljeni pristop se je izkazal kot učinkovit. V prihodnosti bi bilo smiselno razširiti analize na preučevanje dejavnikov, ki vplivajo na spremenljivke, iz katerih se izračuna vrednost xG. Lahko bi preučili tudi, kako se vrednosti xG razlikujejo med različnimi ravnmi tekmovanj, ali pa bi za napovedovanje rezultatov nogometnih tekem uporabili modele strojnega učenja, v katere bi vključili

vrednost xG.

**Ključne besede:** pričakovani zadetki, pričakovane točke, Monte Carlo simulacija, logistična regresija, Poissonova porazdelitev, model Dixon-Coles, nogomet

# Abstract

The measure of expected goals in a a football game is discussed in the work. The concept and calculation of expected goals is presented both theoretically and practically. The impact of the distribution of expected goals and their use to quantify player or team performance is described. On the basis of expected goals, the measure of expected points is derived and practically presented. A method for predicting the results of football games based on the Poisson distribution is presented and updated to take expected goals into account. A practical example compares methods with and without taking expected goals into account.

The topic under discussion is currently rapidly developing. Most of the ideas are obtained in semi-professional articles online. In the present work, they are translated into mathematical or statistical language and presented in an orderly and comprehensive manner, and some new ideas of the author are also added. Practical examples are added to all theoretically presented concepts.

All analyses, simulations and results are obtained with the help of the computer statistical program R. As part of this, some specific packages are used, such as *regista*, *StatsBombR*, *ggsoccer*, *SBpitch*, *worldfootballR* and *soccermatics*.

The approach used has proven to be effective. We assessed which variables statistically significant and in what way affect the probability of a goal in a football game at the highest male professional level. In the future, it would be reasonable to extend the analyzes to the study of factors that influence the variables from which the xG value is calculated. We could look at how the xG values vary between different levels of competition. We could also predict the results of football games with some machine learning models and check the impact of the inclusion of the xG value on the quality of the predictions.

**Key words:** expected goals, expected points, Monte Carlo simulation, logistic regression, Poisson distribution, Dixon-Coles model, football

# 1 Uvod

V zadnjih dveh desetletjih podatki in z njimi povezana statistika postajajo vse bolj razširjeni tudi v športu. Po Soccerment [1], je razmah športne industrije spodbudil strokovnjake podatkovne analitike za ustvarjanje novih mer, katerih pravilna uporaba lahko vodi k prednosti tako na športnem igrišču kot v vseh povezanih dejavnostih. Članek navaja tudi, da športne organizacije podatkovno analitiko uporabljajo za izboljšanje iger svojih moštev, preprečevanje poškodb, optimizacijo plač, povečevanje dohodka s strani prodanih vstopnic in športnih artiklov ter mnogo drugega. Velike športne organizacije podatkovno analitiko uporabljajo tudi za določanje razvrstitev pri različnih žrebanjih in sistemih tekmovanj, ki imajo različne finančne in strateške posledice.

Eden prvih analitikov na nogometnem področju je bil Charles Reep, ki je po drugi svetovni vojni s svinčnikom in papirjem začel zbirati in analizirati podatke nogometnih tekem [2]. Že na podlagi tega, da so eno izmed njegovih prvih teorij mnogi analitiki na čelu z Jonathanom Wilsonom zaradi pristranskih zaključkov zavrgli [3], vidimo, da zbiranje in proučevanje podatkov lahko prinese pomembno konkurenčno prednost, vendar zgolj, če je pravilna tudi njihova analiza in interpretacija. Dober športni analitik mora imeti tako poleg samega statističnega in analitičnega tudi odlično notranje poznavanje športne igre, ki jo preučuje. V nasprotnem primeru njihovi zaključki lahko vodijo do napačnih odločitev in tako celo škodijo.

Za začetek športne analitike v širši družbi velja leta 2003 izdana knjiga Moneyball avtorja Michaela Lewisa [4], po kateri je bil posnet tudi film. Zgodba temelji na sestavljanju konkurenčne bejzbolske ekipe kljub majhemu proračunu ekipe Oakland Athletics. Izhodišče je, da je bilo takratno kolektivno poznavanje

bejzbolske igre zastarelo, subjektivno in pogosto napačno. Vodstvo Oaklanda je izkoristilo analitične meritve uspešnosti igralcev in tako zgradila ekipo, ki je kljub majhnemu proračunu uspešno tekmovala z bogatejšimi tekmeci v ligi MLB (Major League Baseball). Kmalu so delu Oaklanda sledile ostale profesionalne športne organizacije in športi, sprva v ZDA kasneje pa tudi drugod, so postajali vse bolj podatkovno usmerjeni.

V članku [1] so opisane različne uporabe podatkovne analitike. Ena izmed uporab podatkovne analitike v nogometni igri je t.i. skavting, tj. odkrivanje novih talentov in igralcev, ki bi klubu prinesli dodano vrednost. Poleg primarnega cilja, dražje prodaje kot nakupa, se nogometni strokovnjaki s pomočjo filtriranja lahko omejijo na manjšo skupino igralcev z želenimi lastnostmi in jih šele kasneje pričnejo podrobneje in v živo proučevati. Med pomembnejšimi vidiki podatkovne analitike je tudi zmanjšanje tveganja za poškodbe in hitrejša rehabilitacija poškodovanih igralcev. Pri tem nogometni analitiki s pomočjo strokovnjakov s področja medicine svetujejo trenerjem glede količine in prilagajanja nogometnih treningov zdravih nogometašev in primernega trenažnega procesa nogometašem, ki se vračajo po poškodbi. Podatkovna analitika se vse več uporablja tudi pri mlajših športnikih in omogoča optimizacijo njihovega razvoja.

Z ogromnim naborom podatkov, ki opisujejo posamezen dogodek nogometne tekme, lahko nogometni analitiki odkrivajo še neznana dejstva o nogometni igri, lastnosti, ki ustrezajo domači ekipi, ali slabosti nasprotne ekipe in tako zgradijo inovativno zmagovalno strategijo. Z vključitvijo različnih primernih statistik in kazalnikov lahko lažje ocenimo vpliv in predstave igralca, podrobneje analiziramo igro in dobimo boljšo predstavo o dogajanju na igrišču. Posamezne mere in spremenljivke lahko združimo v različne kazalnike, ki merijo uspešnost. Ti so velikokrat prikazani v pajčevinastih grafikonih, ki omogočajo celosten pregled nad specifičnimi vidiki delovanja igralcev [5]. Ena izmed takih kazalnikov sta indeks nevarnosti, ki sta ga razvila Maurizio Viscidi in Antonio Gagliardi [6] in kazalnik celotnega prispevka igralca k uspešnosti ekipe [1]. Različne mere se uporabljajo tudi pri računalniških igrah, kjer denimo računalnik določa ali bo šla žoga v gol ali ne [7].

Ena najbolj razširjenih ter preprostih in hkrati naprednih mer v nogometu so pričakovani zadetki. V tem delu bo predstavljen koncept pričakovanih zadetkov in

nekateri vidiki, ki jih je potrebno upoštevati pri njegovi interpretaciji. Opisana bo zamisel izračuna pričakovanih zadetkov in predstavljen bo model logistične regresije, ki jo lahko uporabimo za izračun pričakovanih zadetkov iz širokega nabora spremenljivk. Na podlagi velikega nabora podatkov bo ustvarjen subjektivni model za izračun vrednosti pričakovanih zadetkov. Znotraj tega modela bomo opazovali tudi, katere spremenljivke in kako vplivajo na to vrednost.

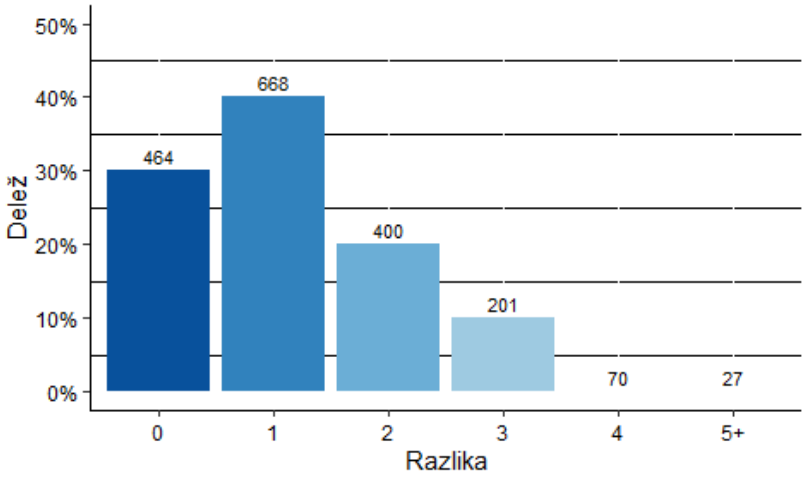
Na primerih bo prikazan vpliv porazdelitve vrednosti pričakovanih zadetkov na število dejanskih zadetkov oziroma njihovo verjetnost, kot tudi na verjetnost zmage določene ekipe. Na podlagi pričakovanih zadetkov so bile razvite mnoge druge mere, med njimi pričakovane točke. Predstavljen bo izračun pričakovanih točk, ki bo praktično prikazan na primeru dejanske tekme. Priložen bo tudi algoritem v statističnem programu R. S pričakovanimi točkami lahko izračunamo tudi t.i. lestvico pravičnosti. Ta bo praktično prikazana na primeru sezone 2017/18 angleške Premier League [8]. Definirana bosta tudi napadalni in obrambni izkoristek in iz njiju skupni faktor izkoristka, ki dajejo informacijo o statistični vzdržnosti ter šibkih in močnih točkah ekipe.

Na koncu bodo predstavljene osnove napovedovanja rezultatov nogometnih tekem. Modeli napovedovanja rezultatov nogometnih tekem, ki se ne poslužujejo konceptov strojnega učenja, temveč napovedujejo s pomočjo Poissonove porazdelitve in izračuna napadalnih in obrambnih moči ekip, bodo nadgrajeni s pričakovanimi zadetki, ki v izračunu lahko zamenjajo dejanske zadetke. Predstavljeni modeli bodo tudi praktično uporabljeni. Prek različnih mer za ocenjevanje kakovosti napovedi bodo primerjane napovedi modelov, ki temeljijo na dejanskem številu zadetkov, in modelov, ki v napovedih upoštevajo pričakovano število zadetkov. S tem bo ovrednoten vpliv pričakovanih zadetkov na kakovost napovedi rezultatov nogometnih tekem oziroma števila zadetkov na tekmah.

# 2 Pričakovani zadetki

## 2.1 Motivacija

Na rezultate v nogometu, morda bolj kot v katerem koli drugem športu, vpliva slučajnost. Pomen slučajnosti poveča tudi dejstvo, da v povprečju na nogometni tekmi igralci dosežejo zgolj nekaj več kot dva zadetka in pol [9] in veliko tekem se konča z neodločenim izidom ali pa je razlika le v enem golu. Slika 2.1 na podlagi podatkov pridobljenih s spletne strani FBref [10] prikazuje delež tekem v odvisnosti od končne razlike v zadetkih za 1830 odigranih tekem v petih največjih evropskih nogometnih ligah (Anglija, Španija, Italija, Nemčija in Francija) v sezoni 2020/21.




Slika 2.1: Delež tekem in njihovo število glede na razliko v zadetkih na koncu tekme za pet največjih evropskih lig, sezona 2020/21.

Zaradi velikega vpliva slučajnosti je uspešnost ekipe ali igralca težko ovre-

dnotiti samo s končnim izidom ali številom doseženih zadetkov. Vedno se lahko vprašamo, ali je končna zmaga produkt sreče in ali igralci svoje priložnosti izkoriščajo v skladu s pričakovanji, oziroma ali bi pričakovali boljšo realizacijo. Tovrstne koncepte je težko oceniti na 'pamet' in z grobimi podatki, zato stremimo k temu, da jih čim bolje kvantificiramo.

## 2.2 Koncept pričakovanih zadetkov

Uspešnost strelov bi pred dobrim desetletjem ocenjevali prek deleža števila strelov, ki so končali v golu, ali deleža strelov, ki so zadeli okvir gola, ter nekaterih subjektivnih mnenj. Tudi to so uporabne mere za ocenjevanje napadalne uspešnosti in izkoristka priložnosti, vendar ne povedo celotne zgodbe. Vsak strel na gol je edinstven in na strel ter priložnost za dosego zadetka vpliva veliko dejavnikov, na podlagi katerih lahko kakovost priložnosti kvantificiramo. V ospredju vseh mer, ki opisujejo strel, je mera pričakovanih zadetkov, ki jo označimo z xG (angl. expected goals).

Pričakovane zadetke je iz hokeja na ledu v nogometni svet leta 2012 vpeljal Sam Green iz podjetja Opta Sports [11]. Mera pričakovanih zadetkov je izračunana na podlagi številnih spremenljivk, ki opisujejo situacijo na igrišču v času strela, s čimer kvantificira kakovost določene priložnosti in predstavlja verjetnost, da se točno določen strel pretvori v zadetek. Te spremenljivke so lahko razdalja od mesta strela do gola, kot med strelcem in golom, položaj ostalih igralcev, del telesa, s katerim je bil izveden strel, tip podaje, ki je vodil do strela, in druge. Ker gre za verjetnost, zavzame vrednost na intervalu [0,1], kjer vrednost 0 predstavlja strel oziroma priložnost, iz katere je zadetek nemogoče doseči, in vrednost 1 strel, kjer žoga vsakič konča v golu. Če smo za strel izmerili vrednost $xG = 0,25$, to pomeni, da povprečni igralec iz takšne priložnosti doseže zadetek na vsake 4 strele.

Mera xG je tako verjetnost oziroma pričakovana vrednost upoštevajoč povprečnega igralca, da strel pristane v golu, ki je ocenjena na podlagi velikega števila strelov iz enakih oziroma podobnih okoliščin. Pričakujemo torej, da bo pri večjem naboru identičnih oziroma podobnih priložnosti delež zadetkov pristal

v ravnovesju z xG. Vidimo, da je pri kvantifikaciji in interpretaciji xG pomembno upoštevati raven tekmovanja, iz katerega črpamo strele za izračun verjetnosti in s tem t.i. 'povprečnega' igralca. Igralci v manj kakovostnih tekmovanjih imajo namreč stopnjo konverzije nižjo kot profesionalni igralci na najvišjih ravneh tekmovanj.

Pričakovane zadetke lahko uporabimo za določanje števila zadetkov, ki bi jih od igralca ali ekipe na posamezni tekmi pričakovali. To lahko razširimo na število zadetkov znotraj izbranega obdobja ali trenerjevega mandata in objektivno izmerimo uspešnost igralca ali ekipe. Kakovostnejše priložnosti kot si igralec ali ekipa ustvari, večja je vrednost xG, in večja kot je razlika med dejanskimi in pričakovanimi zadetki, boljša je konverzija ponujenih priložnosti. Obratno velja v obrambi: manj in slabše priložnosti kot ekipa dopušča, manjša je vsota vrednosti xG nasprotne ekipe in boljša je njena obramba. Obenem velja, da večja kot je razlika med pričakovanimi in dejanskimi zadetki, boljše so obrambe vratarja.

Seveda ima mera xG kot vsaka novost nekatere pomanjkljivosti, vendar pomeni pomemben preskok pri razumevanju nogometne igre in jo je potrebno dobro razumeti in pravilno uporabljati in interpretirati. Po spletni strani `theanalyst.com` [12] je največkrat napaka predpostavka, da bi ekipa z večjo skupno vrednostjo xG tekmo morala dobiti. Če denimo ekipa hitro v tekmi povede, ne gre pričakovati njene ofenzive, temveč pričakujemo, da si bo nasprotnik, ki zaostaja, v preostanku tekme ustvaril večje število priložnosti za zadetek in tako prišel do večje skupne vrednosti xG. Mera pričakovanih zadetkov tako meri samo kakovost priložnosti in ne pričakovanega izida tekme. Velikokrat se predpostavi tudi, da bo igralec ali ekipa, ki z dejanskimi zadetki preseže pripadajočo (kumulativno) vrednost xG, v nadaljevanju igral pod pričakovanji, da bi se z dejanskimi zadetki vrnil na pričakovano raven v skladu z xG. Vendar ni tako: ti dogodki so se že zgodili in ne vplivajo na prihodnost, zato v nadaljevanju pričakujemo število dejanskih zadetkov v skladu z vrednostjo xG.

Z vrednostjo xG in na podlagi nje razvitimi merami torej lahko kvantificiramo vprašanja, ki zanimajo vsakega navijača in predvsem trenerja. Kateri igralec ima težave z zaključkom? Katera ekipa igra bolje, kot prikazuje trenutna lestvica? Kateri igralec bi lahko dobil več igralnih minut? V prihodnosti lahko s poglobljeno analizo koncepta pričakovanih zadetkov pričakujemo implementacijo taktičnih

strategij, ki bi lahko dodobra spremenile nogometno igro.

## 2.3 Primer: primerjava xG

Za osnovno interpretacijo koncepta pričakovanih zadetkov z uporabo podatkov s strani FBref [10] primerjajmo dva nogometaša na podlagi njunih tekem v sezoni 2020/21. To sta Timo Werner iz kluba Chelsea angleške Premier League in Rodrigo De Paul iz Udineseja italijanske Serie A. Oba igralca sta v obravnavani sezoni na gol ustrelila natanko 80-krat in oba, izključujoč enajstmetrovke, dosegla 6 zadetkov. Z uporabo mere xG kvantificiramo kakovost njunih priložnosti in s tem njune strele in število zadetkov postavimo v kontekst.

Tabela 2.1: Primerjava kumulativnih vrednosti xG, Timo Werner in Rodrigo De Paul, sezona 2020/21.

| Timo Werner | Igralec | Rodrigo De Paul |
|---|---|---|
| Napad | Igralno mesto | Sredina |
| 80 | Streli | 80 |
| 6 | Zadetki | 6 |
| 11,9 | Pričakovani zadetki [kumulativni xG] | 4,3 |
| 0,15 | Pričakovani zadetki na strel [xG na strel] | 0,06 |

Na podlagi tabele 2.1 bi, upoštevajoč priložnosti, ki jih je imel Timo Werner, pričakovali, da povprečen igralec doseže 12 zadetkov (xG = 11,9); glede na priložnosti De Paula bi pričakovali, da povprečen igralec doseže le 4 zadetke (xG = 4,3). Profile njunih strelov lahko primerjamo preko pričakovanega števila zadetkov na strel (xG na strel), kjer je pri Wernerju pričakovati zadetek na slabih 7 strelov, od De Paula pa bi zadetek pričakovali na slabih 17 strelov. Upoštevajoč kakovosti priložnosti je realizacija De Paula, ki je nadpovprečna, bistveno boljša od podpovprečne Wernerjeve.

Opazimo lahko povezanost xG z igralnim položajem, ki posredno vpliva na vrednosti spremenljivk, ki določajo xG. De Paul je vezist in njegovi streli so v splošnem iz težjih položajev na igrišču. Na drugi strani je Werner napadalec in celotna igra njegove ekipe sloni na cilju, da mu pripravijo čim lepšo priložnost

za zadetek - on je tisti, ki je zadolžen za zadetke in zadnji cilj podaje soigralcev. Posledično je kakovost njegovih priložnosti večja.

## 2.4 Koncept izračuna xG

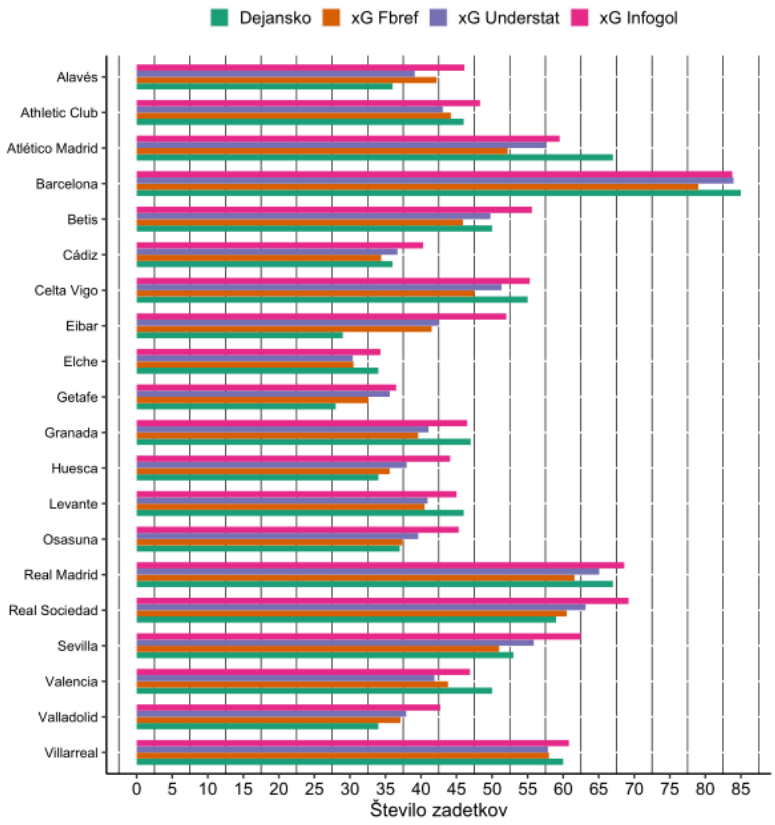
Mera pričakovanih zadetkov je izračunana na podlagi številnih spremenljivk, ki opisujejo situacijo na igrišču v času strela, in predstavlja verjetnost, da se strel pretvori v zadetek. Želimo torej model, kjer bomo ob upoštevanju nekaterih spremenljivk izračunali, kakšna je verjetnost, da se strel pretvori v zadetek, in hkrati opredelili, kako vsaka od spremenljivk vpliva na to verjetnost.

Pričakovane zadetke bi v teoriji izračunali z uporabo frekventističnega pristopa na podlagi velikega števila strelov, ki so kategorizirani glede na spremenljivke, ki ga opisujejo. Vsak strel bi na podlagi spremenljivk razvrstili v določeno skupino in vrednost xG bi predstavljala delež strelov, ki so pristali v golu. Vendar tak pristop ni najboljši. V teoriji niti dva strela na gol na podlagi spremenljivk, ki strel opisujejo, nista identična. Tako bi morali imeti število skupin enako številu strelov in izračun vrednosti xG s tovrstnim pristopom je neizvedljiv.

Edini strel, ki ga lahko ocenimo s frekventističnim pristopom, je enajstmetrovka, kjer so vse spremenljivke, ki strel opisujejo, načeloma identične (npr. položaj vratarja in igralcev za žogo ne nujno). Na podlagi raziskave podjetja InStat, kjer je bilo upoštevanih več 100 tisoč enajstmetrovk, je bila izračunana stopnja realizacije enajstmetrovk 0,7557 [13].

V praksi je vrednost xG največkrat ocenjena na podlagi regresijskih modelov, kjer je najbolj pogost model logistične regresije. Njegova prednost je, da v primerjavi s številnimi modeli strojnega učenja omogoča enostavnejšo interpretacijo spremenljivk, kar je poleg samega izračuna verjetnosti v različnih krogih uporabnikov zelo pomembno. Možnost interpretacije tako odtehta morda nekoliko slabšo napoved izbranega modela. Obstaja več analitičnih podjetij in spletnih mest (Opta, StatsBomb, Understat, Smartodds, Infogol itd.), ki so s podatkovnimi bazami več sto tisoč strelov na gol in povezanih podatkov različnih nogometnih lig za več sezon zgradili svoje algoritme in modele, ki vsakemu strelu kvantificirajo vrednost xG. Vsako podjetje ima svoj algoritem, količino, kakovost

in vrsto podatkov, in na podlagi tega se njihove vrednosti xG razlikujejo. Na sliki 2.2 so predstavljene vsote vrednosti xG celotne sezone 2020/21 španske La Lige, podane s strani FBref [10], Understat [14] in Infogol [15].




Slika 2.2: Primerjava kumulativnih vrednosti xG različnih spletnih strani (FBref, Understat, Infogol), La Liga, sezona 2020/21.

Velja, da so podatki kumulativnih vrednosti xG, torej vsot vrednosti xG posameznih strelov, na ravni posameznih tekem, sezon ali tekmovanj, za večje nogometne lige in tekmovanja prosto dostopni. Vrednosti xG posameznih strelov, kot tudi obsežnejši podatki, ki vključujejo vrednosti vseh spremenljivk, ki strel opisujejo, načeloma niso prosto dostopni.

# 3 Praktični izračun xG

Eden glavnih ciljev naloge je ustvariti model za izračun vrednosti xG. Preden se lotimo njegove gradnje, moramo razmisliti, kakšni podatki nas zanimajo. Potrebujemo veliko zbirko podatkov o strelih ter vseh dejavnikih, ki opisujejo situacijo na igrišču, znotraj katere je bil strel izveden. Nogometni podatki so običajno razdeljeni v dve obliki: podatke o dogodkih in podatke o sledenju [16]. Podatki o dogodkih beležijo vse dogodke, povezane z žogo in tem, kje na igrišču so se zgodili. Tu upoštevamo strele, podaje, preigravanja, prekrške in podobno. Podatki o sledenju beležijo položaj igralcev in žoge skozi celotno tekmo v rednih, izjemno kratkih časovnih intervalih.

Podatki na podlagi katerih bomo ustvarili model za izračun xG so last podjetja StatsBomb [17] in so v zadovoljivi meri za naše potrebe prosto dostopni. Zajemajo podatke o vseh vidnejših dogodkih večjega števila tekem najvišje profesionalne ravni skozi časovno obdobje, ki ne sega več kot 15 let nazaj. Vsak dogodek je predstavljen z okoli 180 spremenljivkami in vseh dogodkov, ki predstavljajo strel, je nekaj več kot 15.500. Za vsak strel je v podatkih podana tudi vrednost xG, izračunana s strani lastnika podatkov. Omejili se bomo na moške nogometne tekme in predpostavljamo, da se nogometna igra v zadnjih 15 letih ni spremenila v tolikšni meri, da podatkov iz leta 2005 v naši analizi ne bi smeli upoštevati. Pri izračunu xG se bomo omejili na strele iz odprte igre, kamor ne štejemo neposrednih strelov iz kotov in prostih strelov ter enajstmetrovk. Tovrstnih strelov je 14.099. Velikost vzorca je precejšnja in bodoče ugotovitve lahko posplošimo na celotno nogometno igro najvišje moške profesionalne ravni, kar opisujejo naši podatki.

Vrednosti xG vsakega strela bomo ocenjevali z modelom logistične regresije.

V nadaljevanju bomo predstavili model logistične regresije in spremenljivke, ki so nam v podatkih na voljo in bi lahko igrale vlogo pri izračunu xG, zato smo jih uvrstili v nabor za možnost vključitve v model. Po izbiri najboljšega modela in prikazu njegovih lastnosti bomo dobljene vrednosti xG primerjali s podanimi vrednostmi xG podjetja StatsBomb, ki veljajo za ene boljših in bi se jim radi približali.

## 3.1 Logistična regresija

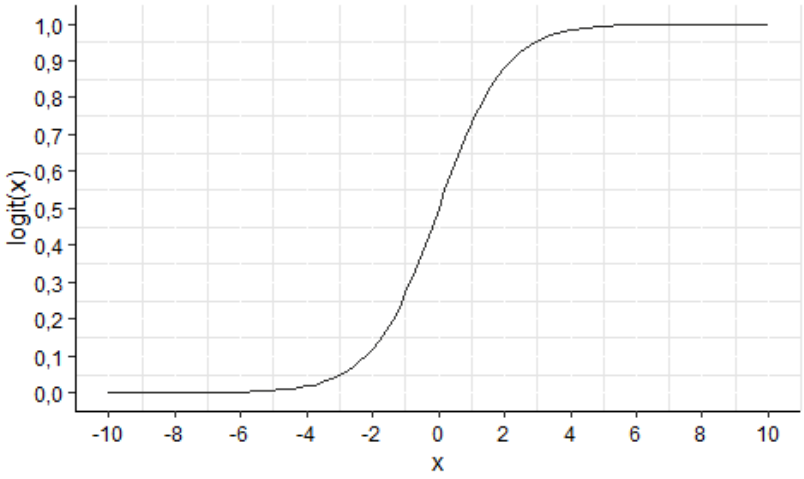
Regresijska analiza je vrsta napovedne analize, ki omogoča preučevanje odnosa med dvema ali več spremenljivkami, ki nas zanimajo. Z njo preverjamo vpliv neodvisnih spremenljivk na odvisno spremenljivko. Neodvisnim spremenljivkam pravimo tudi prediktorji oziroma napovedovalci, medtem ko odvisno spremenljivko lahko poimenujemo tudi odzivna ali izhodna spremenljivka.

V statistiki se model logistične regresije (model logit) uporablja za modeliranje verjetnosti določenega razreda binarne odvisne spremenljivke, torej spremenljivke tipa da ali ne oziroma 1 ali 0, pogojno na vrednosti nabora neodvisnih spremenljivk. Vsak opažen dogodek ima natanko enega izmed dveh možnih izidov z verjetnostjo med 0 in 1, pri čemer je njuna vsota 1.

Model logistične regresije v svoji osnovi za modeliranje verjetnosti pozitivnega razreda binarne spremenljivke uporablja logistično funkcijo, definirano kot:

$$\text{logistic}(x) = \frac{1}{1 + e^{-x}}. \tag{3.1}$$

Tako kot vse oblike regresijske analize, se tudi logistična regresija osredotoča na pogojno porazdelitev odziva glede na vrednosti parametrov, ki napovedujejo verjetnost pozitivnega razreda odvisne spremenljivke ($Y$). Model logistične regresije predpostavlja, da je logit transformacija verjetnosti pozitivnega razreda odzivne spremenljivke definirana kot linearna kombinacija koeficientov ($\boldsymbol{\beta}$), pomnoženih z naborom pojasnjevalnih spremenljivk ($\boldsymbol{X}$).




Slika 3.1: Logistična krivulja.

$$\ell = \ln[\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}] = \beta_0 + \beta_1 \boldsymbol{X_1} + \cdots + \beta_p \boldsymbol{X_p}. \tag{3.2}$$

Logit transformacija verjetnosti pozitivnega razreda odzivne spremenljivka nam tudi omogoča enostavno interpretacijo parametrov modela. Brez uporabe logit transformacije je verjetnost pozitivnega razreda odzivne spremenljivke definirana kot

$$p(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 \boldsymbol{X_1}+\cdots+\beta_p \boldsymbol{X_p}}}{1+e^{\beta_0+\beta_1 \boldsymbol{X_1}+\cdots+\beta_p \boldsymbol{X_p}}}. \tag{3.3}$$

Pri interpretaciji moramo upoštevati, da $\beta$ določa multiplikativno povečanje verjetnosti pozitivnega razreda odvisne spremenljivke $Y$ ob spremembi neodvisne spremenljivke $X$. Če se $X$ poveča za eno enoto mere, se verjetnost pozitivnega razreda $Y$ poveča $e^{\beta}$-kratno oziroma se verjetnost pozitivnega razreda poveča za $(e^{\beta}-1)\%$. Bolj kot same vrednosti parametrov modela ($\boldsymbol{\beta}$) nas tako zanimajo njihove eksponente vrednosti $e^{\boldsymbol{\beta}}$.

Za oceno parametrov modela ($\boldsymbol{\beta}$) uporabimo metodo največjega verjetja, ki išče njihove vrednosti tako, da je razlika med pričakovano verjetnostjo pozitivnega razreda odvisne spremenljivke $Y_i$ tj. $\hat{p}(\boldsymbol{X_i})$ in opaženo vrednostjo $Y_i$ minimalna. Formalno torej iščemo parametre $\boldsymbol{\beta}$, ki maksimizirajo funkcijo verjetja:

$$\ell(\boldsymbol{\beta}) = \prod_{i:y_i=1} p(\boldsymbol{X_i}) \prod_{i:y_i=0} \Big(1 - p(\boldsymbol{X_i})\Big). \tag{3.4}$$

Pri izbiri modela za analizo je potrebno upoštevati primernost modela. Dodajanje neodvisnih spremenljivk regresijskemu modelu bo vedno povečalo pojasnjeno variabilnost odvisne spremenljivke, vendar z dodajanjem prekomernega števila neodvisnih spremenljivk zmanjšujemo splošnost modela, saj bi bile lahko nekatere spremenljivke pomembne samo zaradi naključja. Upoštevamo zgolj spremenljivke, ki statistično značilno vplivajo na odvisno spremenljivko, in izločimo spremenljivke, ki bi lahko dodale šum, kar bi poslabšalo napovedi modela.

Kot rečeno, logistični model modelira verjetnost pozitivnega razreda odvisne spremenljivke, vendar se lahko uporablja tudi za uvrščanje, kjer verjetnostim, večjim od izbrane mejne vrednosti, pripišemo pozitivni razred, in verjetnostim, manjšim od mejne vrednosti, pa pripišemo negativni razred odvisne spremenljivke. Dobljene regresijske ocene uporabljamo za razlago razmerja med odvisnimi in neodvisnimi spremenljivkami, natančneje za določanje jakosti prediktorjev, napovedovanje vpliva sprememb (kako se spreminja odvisna spremenljivka s spremembo ene ali več neodvisnih spremenljivk) ter za napovedovanje trendov in prihodnjih vrednosti.

Pri logistični regresiji predpostavljamo, da so opazovane enote medsebojno neodvisne. Najlažji način, da preverimo to predpostavko, je z grafom ostankov glede na vrstni red opazovanj. Če v grafu ne opazimo vzorca, predpostavka drži, sicer je kršena. Posledica kršene predpostavke so pristranske ocene standardnih napak parametrov in intervali zaupanja so ožji, kot bi morali biti.

Naslednja predpostavka je, da med pojasnjevalnimi spremenljivkami ni resne multikolinearnosti. Ta nastopi, ko sta dve ali več pojasnjevalni spremenljivki med seboj močno povezani. Kršena predpostavka lahko vodi v napačno interpretacijo parametrov modela in otežuje identifikacijo statistično značilnih napovednih spremenljivk. Najpogostejši način za preverjanje te predpostavke in odkrivanje multikolinearnih spremenljivk je uporaba variančno inflacijskega faktorja VIF. Ta meri moč korelacije med napovednimi spremenljivkami v modelu.

Predpostavljamo tudi, da v naboru podatkov ni ekstremno izstopajočih vre-

dnosti, tako imenovanih osamelcev. To predpostavko lahko preverimo z uporabo Cookove razdalje za vsako opazovanje. Če odkrijemo, da izstopajoče enote obstajajo, jih lahko odstranimo, zamenjamo s povprečjem ali mediano, ali pa jih v modelu obdržimo in to upoštevamo v njegovi interpretaciji.

Pomembna predpostavka je tudi linearna povezanost med vsako od pojasnjevalnih spremenljivk in vrednostjo logit transformacije odzivne spremenljivke. Predpostavko lahko preverimo grafično ali pa z uporabo Box-Tidwellovega testa. Čeprav to najpogosteje ni težava, moramo zagotoviti tudi, da število pojasnjevalnih spremenljivk ne presega števila opazovanj ($p < n$). Z zadostnim številom opazovanih enot, to je za vsako napovedno spremenljivko vsaj 10 dogodkov (pozitivnih izidov) vsake izmed vrednosti odzivne spremenljivke, zagotovimo, da iz ocenjenega modela naredimo veljavne zaključke.

V nasprotju z linearno regresijo logistična regresija nima predpostavke o normalni porazdeljenosti ostankov ter njihovi konstanti varianci oziroma homoskedastičnosti.

## 3.2 Ocena kakovosti regresijskega modela

Fazi učenja regresijskega modela, kjer na podlagi enot, za katere poznamo pripadnost razredu, zgradimo model oz. pravilo, na podlagi katerega bomo novim enotam predpisali verjetnosti pozitivnega razreda. Sledi faza preverjanja, kjer želimo oceniti kakovost zgrajenega modela. Verjetnosti oziroma razrede, ki jih napove model, želimo primerjati z dejanskimi. Tovrstno oceno kakovosti lahko naredimo na več načinov [18].

Prvi način je, da model ocenimo na učni množici in ga preverimo na neodvisni testni množici podatkov. Ta vsebuje enake spremenljivke kot učna množica in v splošnem tudi tu poznamo dejanske razrede odvisne spremenljivke. Kot testna množica lahko služi po slučajnosti razbita učna množica podatkov.

Druga in najpogosteje uporabljena strategija preverjanja kakovosti napovedi modela je prečno preverjanje s $k$ pregibi ($k$-fold cross validation). Bistvo postopka je, da enote $\{1, ..., n\}$ razbijemo na $k$, običajno $k = 5$ ali $k = 10$, (približno) enako

velikih podmnožic (pregibov). Postopek ocenjevanja je sestavljen iz zaporedja $k$ iteracij, kjer na vsakem koraku za učno množico uporabimo $k-1$ podmnožic in klasifikator preizkusimo na preostali podmnožici. To ponovimo za vse množice $1, ..., k$ in model tako zaporedoma preizkusimo na vseh enotah.

Kakovost napovedi modela lahko ocenimo tudi z metodo samovzorčenja (angl. bootstrap). Pri samovzorčenju iz originalne učne množice podatkov velikosti $n$ z vzorčenjem s ponavljanjem pridobimo novo učno množico podatkov, prav tako velikosti $n$, ki jo uporabimo kot novo učno množico, na kateri zgradimo model in njegovo kakovost preverimo na originalni učni množici, ki tako postane testna množica podatkov. Ta postopek mnogokrat ponovimo in tako dobimo oceno kakovosti napovedi modela, ki jo lahko popravimo za različne navidezne natančnosti prileganja modela. Tako prečno preverjanje kot samovzorčenje ne potrebujeta neodvisne testne množice temveč zgolj učno množico podatkov, kar nam omogoča večji nabor podatkov za končno učenje izbranega modela.

Pri zveznih klasifikatorjih, kjer v našem primeru napovedujemo verjetnost zadetka, oceno kakovosti modela lahko predstavimo s krivuljo ROC (angl. receiver operating characteristic) in ploščino pod njo (AUC - area under the ROC curve). Z njima predstavimo relativni kompromis med dejansko pozitivnimi enotami (zadetki) in lažno pozitivnimi enotami. Krivulja ROC je definirana v enotskem koordinantnem sistemu, ki ga določata občutljivost na osi $y$ in (1-specifičnost) na osi $x$. V prostoru ROC je neka točka boljša od druge, če leži bližje točki $(0, 1)$. Klasifikatorji, katerih dosežki ležijo pod diagonalo oziroma je AUC $< 0{,}5$, delujejo slabše od slučajnega klasifikatorja; in klasifikatorji katerih dosežki ležijo nad diagonalo oziroma je AUC $> 0{,}5$, delujejo bolje od slučajnega klasifikatorja. Formalno je mera AUC enaka verjetnosti, da bo klasifikator slučajno izbrano pozitivno enoto rangiral višje od slučajno izbrane negativne enote.

Pri oceni kakovosti modela lahko uporabimo tudi kalibracijo oziroma umerjanje modela, kjer primerjamo napovedane in opazovane vrednosti. Opazovane vrednosti v primeru logistične regresije pomenijo opažene deleže na vzorcu. Verjetnosti lahko prikažemo na enotskem koordinatnem sistemu, kjer vsaki napovedani verjetnosti na osi $x$ pripada opazovana vrednost na osi $y$: bolj kot so vrednosti na premici $y = x$, boljše so modelske napovedi. Ker v našem primeru ne poznamo opazovanih vrednosti, jih moramo oceniti. V teoriji bi vsak strel v podatkih mno-

gokrat ponovili in delež zadetkov bi predstavljal 'pravilno' oceno verjetnosti. Ker tega ni mogoče storiti, si pomagamo z dejstvom, da če strelu pripišemo verjetnost $a\%$, potem mora biti delež zadetkov strelov s to pripisano verjetnostjo čim bližje ravno $a\%$. Ker so napovedane verjetnosti zvezne, jih moramo za to vrsto izračuna diskretizirati oziroma razporediti v intervale, na podlagi katerih lahko za vsak interval napovedanih verjetnosti izračunamo delež zadetkov [19]. Primerjavo verjetnosti lahko naredimo prek prečnega preverjanja ali samovzorčenja, kjer napovedi modela in izračun opazovanih vrednosti večkrat ponovimo.

Poleg opisanih metod za oceno kakovosti modela, ki temeljijo na izračunanih verjetnostih, pri logistični regresiji - predvsem za primerjavo dveh ali več modelov med seboj - upoštevamo različne teste in mere, kot so test razmerja verjetij, Akaikejev informacijski kriterij, razlika med variabilnostjo ostankov polnega modela v primerjavi z ničelnim modelom in različni koeficienti psevdo- $R^2$, kot sta Nagelkerkejev in McFaddenov.

## 3.3 Pomembne pojasnjevalne spremenljivke modela xG

Iz nabora prosto dostopnih in razpoložljivih podatkov podjetja StatsBomb bomo iz razpoložljivih 180 spremenljivk, ki opisujejo vsak strel, predstavili tiste, za katere menimo, da bi lahko značilno vplivale na verjetnost zadetka in tako na izračun/oceno xG ter bodo sestavljale množico spremenljivk, ki imajo možnost vključitve v model.

V uvodu smo dejali, da bomo regresijski model omejili na strele iz odprte igre. Ostale tri vrste neposrednih strelov, tj. proste strele, strele iz kota in enajstmetrovke, bi za primeren izračun xG zahtevale svoje modele, ki jih v nalogo ne bomo vključili. Vseeno jih krajše predstavimo, da dobimo vpogled v spremenljivke, ki lahko vplivajo na verjetnost zadetka.

### 3.3.1 Enajstmetrovke, streli iz kota in prosti streli

Kot in enajstmetrovka sta izvedena pri (bolj ali manj) enakih pogojih, torej z (pri kotu vsaj približno) enako postavitvijo igralcev na igrišču in z enakega položaja

(pri kotu se menja zgolj stran igrišča). Na podlagi tega bi enajstmetrovki in strelu iz kota pripisali fiksno vrednost xG, ki izhaja iz deleža konverzije. Na podlagi raziskave podjetja InStat bi enajstmetrovkam pripisali vrednost xG = 0,7557. Za strel iz kota pa velja, da je to večkrat posledica slabe podaje kot namenskega strela. Zaradi izjemne zahtevnosti bi strelom iz kota morali pripisati vrednost xG blizu 0, kar pa morda ne bi bilo v skladu z realno stopnjo konverzije, ki je verjetno zavajajoče visoka.

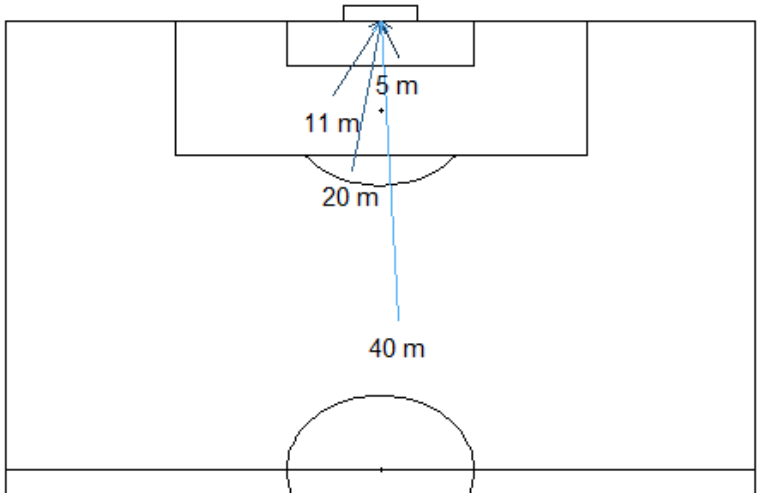
Pri prostem strelu bi na verjetnost zadetka moral vplivati zgolj položaj na igrišču, torej oddaljenost od gola in kot med žogo in golom. Za udarce s prostega strela velja, da so vsi izvedeni izven kazenskega prostora, torej pravokotnika okoli gola, ki je od središča gola in od obeh vratnic oddaljen 16,5 metrov. Večina neposrednih udarcev s prostega strela je izvedena v bližini kazenskega prostora, saj s položajev bolj proti stranski črti igrišča prihaja bolj do podaj kot strelov. V naših podatkih je med vsemi streli delež strelov neposredno s prostega strela enak 7% in od teh jih je v golu končalo 6,4%.

### 3.3.2 Oddaljenost od gola

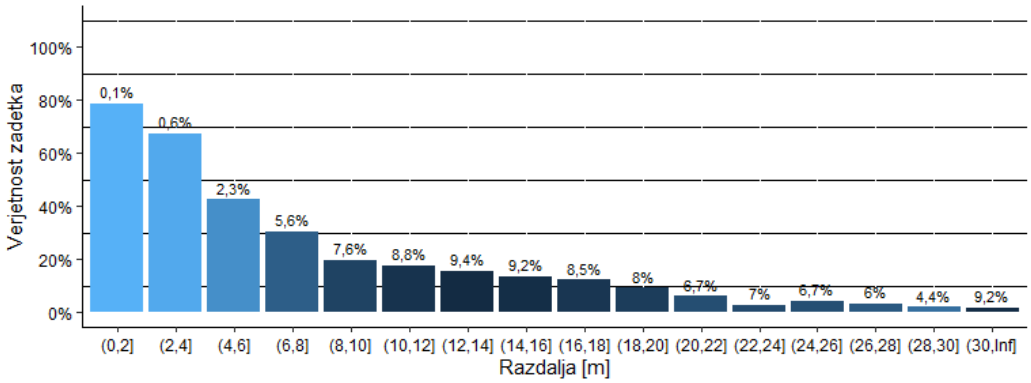
Lastnost strela, za katero domnevamo, da močno vpliva na verjetnost zadetka, je oddaljenost od gola. Večina vrednosti, ki v naših podatkih predstavlja razdaljo, je izražena relativno glede na velikost igrišča, kjer so upoštevane dimenzije 120 m × 80 m. Za boljšo predstavo razdalje si na sliki 3.2 lahko ogledamo, kako na nogometnem igrišču izgleda razdalja od točke strela do sredine nogometnega gola.

Velja, da se blizu gola ne zgodi veliko strelov, saj je do strela s takega položaja težko priti. Na sliki 3.3 vidimo, da do večine strelov pride na razdalji 8 metrov in več in ti so približno enakomerno porazdeljeni. Največji delež strelov, ki se pretvorijo v gol, je pričakovano najbližje golu. Z oddaljenostjo od sredine nogometnega gola se verjetnost zadetka zmanjšuje.

Izkaže se, da polovica zadetkov pade na razdalji od 7,8 metrov do 16 metrov od gola, torej v kazenskem prostoru. V tem območju se sicer ne zgodi niti polovica strelov na gol. Le majhen del zadetkov pade na razdalji bližje od 8 metrov ali



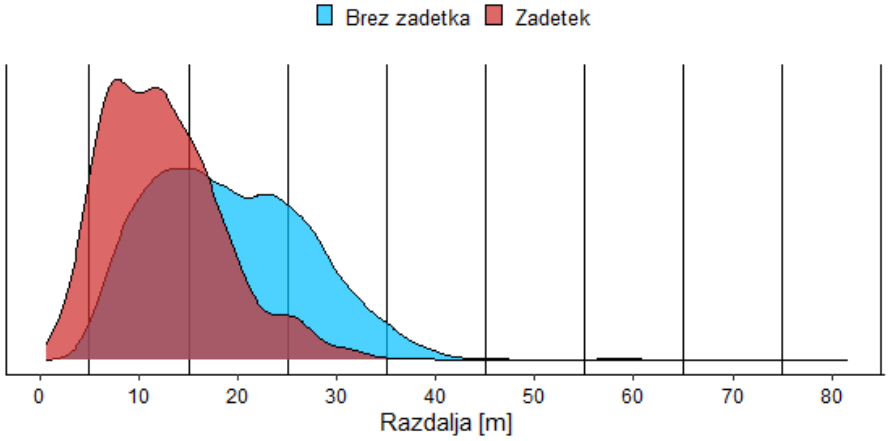

Slika 3.2: Prikaz razdalje od mesta strela do sredine gola na nogometnem igrišču.




Slika 3.3: Verjetnost zadetka (višina stolpca) in delež strelov glede na razdaljo od mesta strela do sredine gola.

več kot 25 metrov od gola, kar pa je tudi posledica majhnega števila strelov na tem območju.

V podatkih imamo tudi spremenljivko o razdalji od mesta strela do vratarja, ki je običajno nekoliko manjša kot do sredine gola. Spremenljivki sta močno korelirani in podajata informacijo približno enake kakovosti. V regresijskem modelu bomo upoštevali kvečjemu eno od njiju.



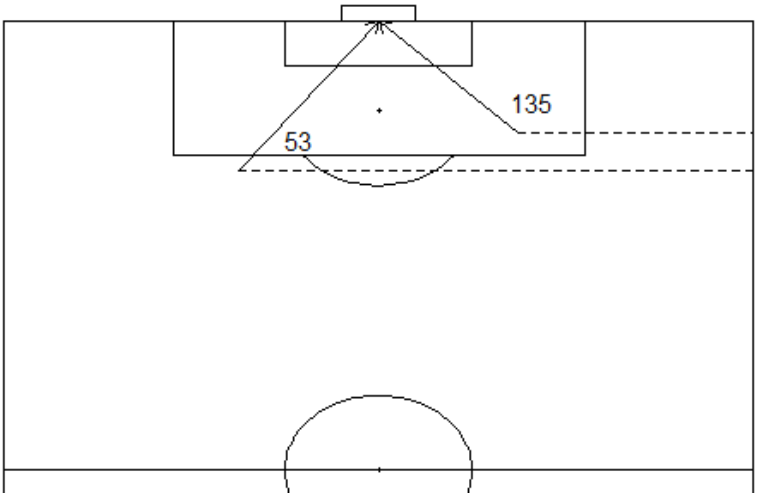

Slika 3.4: Porazdelitev strelov glede na razdaljo od mesta strela do sredine gola.

### 3.3.3 Kot strela

V podatkih imamo tri spremenljivke, ki opisujejo kot strela na gol. To so kot med točko strela in sredino gola, kot med točko strela in vratarjem ter njuna razlika. Kot med točko strela in sredino gola je notranji kot med črto, ki povezuje desni stranski rob igrišča s točko strela, in črto, ki povezuje mesto strela s sredino gola. Izračunan kota je predstavljen na sliki 3.5. Podobno je izračunan tudi kot med točko strela in vratarjem. Tovrstni izračun kota, kjer upoštevamo stran igrišča, je primeren ob upoštevanju preferenčne noge igralca, česar v podatkih nimamo, ali pa noge igralca, s katero je bil strel izveden, kar je informacija, ki jo imamo.

Če ne upoštevamo preferenčne noge igralca ali noge, s katero je bil strel izveden, je za oceno verjetnosti zadetka vseeno, če levo in desno stran igrišča obravnavamo enakovredno. Tako upoštevamo *absolutno vrednost kota*, kar je notranji kot med črto, ki povezuje sredino širine igrišča s točko strela (90 stopinj) in črto, ki povezuje mesto strela s sredino gola. Tovrstni kot je prikazan na sliki 3.6. Če je kot manjši od 90 stopinj, torej na levi strani igrišča, ohranimo vrednost kota, sicer pa ga odštejemo 180 stopinjam.

V teoriji velja, da bližje kot je absolutni kot 0 stopinjam, večja je verjetnost za dosego zadetka, saj vratar lahko pokrije manjši del gola. Vendar po sliki 3.7 vidimo, da bistvenih razlik v deležu strelov, ki se pretvorijo v zadetek, glede na absolutno vrednost kota ni. Obenem velja, da več strelov pride iz večjih



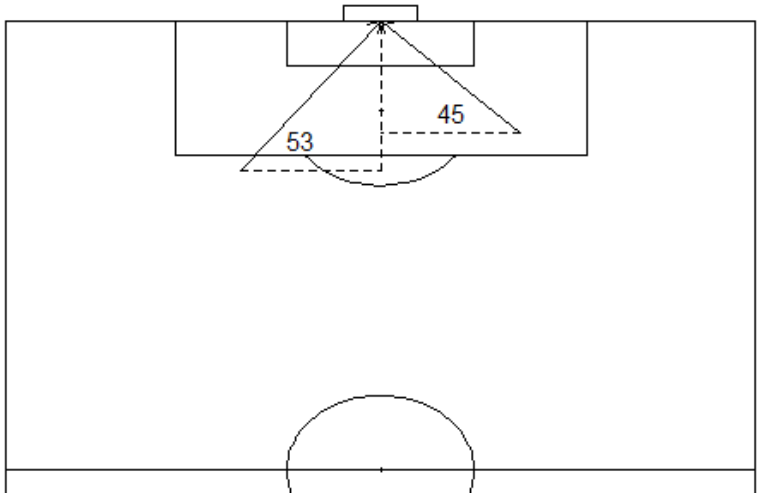

Slika 3.5: Polni kot med točko strela in sredino gola.

absolutnih kotov.

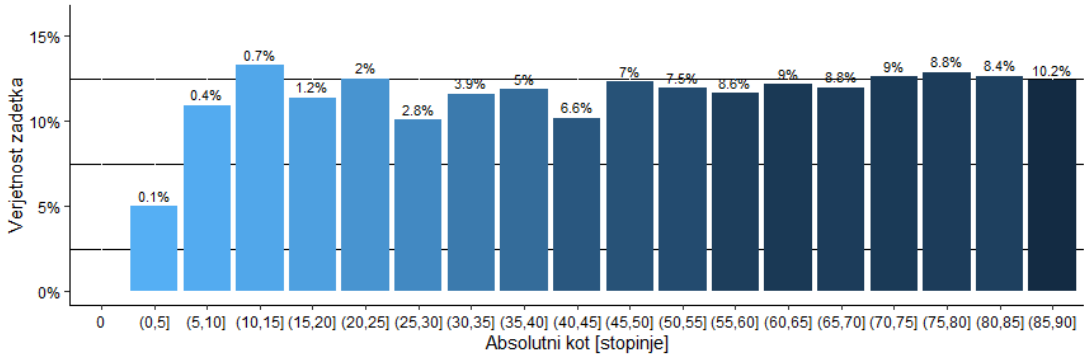
Za kot med točko strela na gol in sredino gola bi lahko rekli, da je pomemben posredno oziroma v kombinaciji z oddaljenostjo strela. Sklepali bi, da je najpomembnejši kot med točko strela in obema vratnicama. Ta določa prostor ki se odpira igralcu in je morda tudi lažje razumljiv. Poimenujmo ga *kot odprtja*. Prikazan je na sliki 3.8 in je notranji kot med črtama, ki povezujeta točko strela in vratnici. Kot odprtja 180 stopinj lahko dosežemo zgolj, če smo popolnoma na sredini širine igrišča in na golovi črti. Podobno kot odprtja 0 stopinj dosežemo le v primeru, da smo na golovi črti in na eni oziroma drugi strani vratnice. Bližje kot smo golu in bolj na sredini igrišča smo, večji je naš kot odprtja.

Verjetnost zadetka razumljivo raste s kotom odprtja in večina strelov se zgodi med 15 in 40 stopinjami. To so večinoma streli s polrazdalje pod kotom do sredine gola. Tovrstni tip strelov lahko vidimo na sliki 3.8. Glede na sliko 3.9 pri manj kot 15 stopinjah in več kot 40 stopinjah kota odprtja ni veliko strelov, in sicer zaradi majhne nevarnosti za gol (manj kot 15 stopinj) ali velike pozornosti nasprotnih igralcev (več 40 stopinj). Pomembno vlogo, predvsem pri strelih s strani, igra tudi pozicija vratarja, ki dodatno zapre kot, ki je na voljo strelcu.

Čeprav je kot odprtja lažje razumljiv in morda pomembnejši od absolutnega kota med točko strela in sredino gola, se pri modelu logistične regresije pojavi težava multikolinearnosti. Kot odprtja je izračunan na podlagi položaja strelca

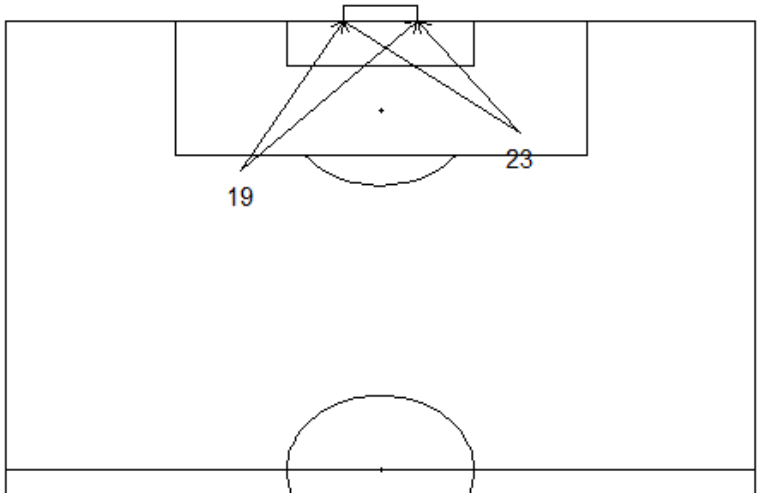



Slika 3.6: Absolutni kot med točko strela in sredino gola.

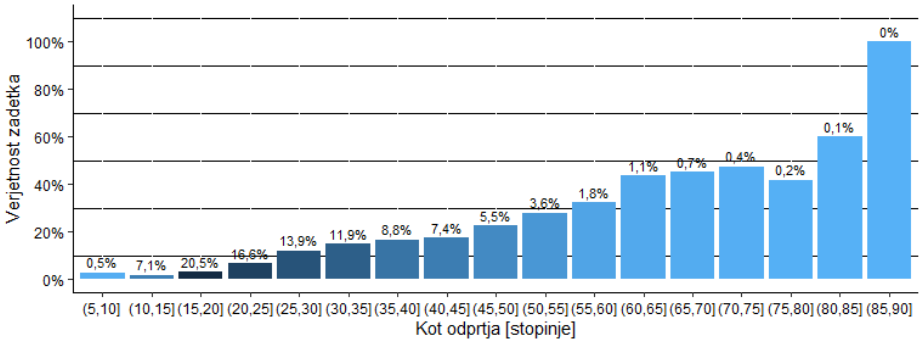



Slika 3.7: Verjetnost zadetka (višina stolpca) in delež strelov glede na kot med točko strela in sredino gola.

in je tako koreliran z oddaljenostjo strela do sredine gola, ki je ena od pomembnejših spremenljivk in bo v našem modelu gotovo nastopala. Na drugi strani absolutna vrednost kota in razdalja do sredine gola nista povezana, zato lahko skupaj nastopata v regresijskem modelu. Med spremenljivko kot odprtja in kombinacijo spremenljivk razdalja od gola in absolutni kot tako najverjetneje velja izbrati slednjo.




Slika 3.8: Kot odprtja.

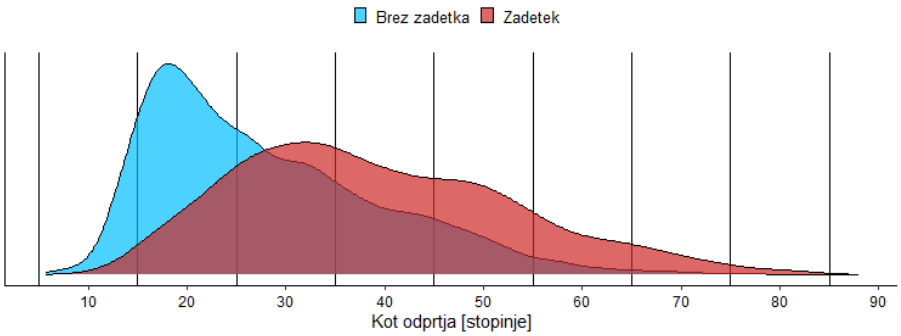



Slika 3.9: Verjetnost (višina stolpca) zadetka in delež strelov glede na kot odprtja.

### 3.3.4 Tip strela

Dejavnik, ki (posredno) določa težavnost strela in tako verjetnost zadetka, je tip strela na gol. Ločimo običajen strel (ne spada v nobeno drugo kategorijo), volej (žoga pred strelom ni padla na tla), polvolej (žoga se pred strelom odbije od tal), udarec s peto, lob (spodkopana žoga preko vratarja), nizek strel z glavo (žoga je ob strelu z glavo v višini pasu ali nižje) in škarjice (v času strela je igralec obrnjen s hrbtom proti golu in žogo pošlje prek lastne glave).

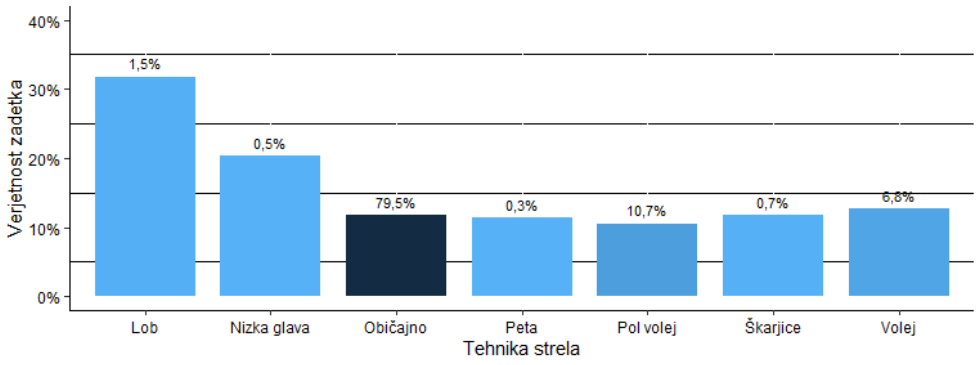
Velja sicer, da je tip strela odvisen od situacije, znotraj katere je strel izveden. Racionalen igralec bi se vedno odločil za najlažji oziroma najustreznejši strel, zato verjetnost zadetka narekujejo druge spremenljivke, ki jih v danem tipu strela




Slika 3.10: Porazdelitev strelov glede na kot odprtja, pogojno na indikator zadetka.

upoštevamo posredno.

Na sliki 3.11 vidimo, da je največ strelov običajnega tipa, dobrih 80%. Z manj kot 10% sledita polvolej in volej, ostali tipi strelov pa so redki. Visoka stopnja konverzije je dosežena z lobom in tudi z nizkim strelom z glavo. To je predvsem posledica majhnega deleža tovrstnih strelov in izvedbe v okoliščinah, ki prinašajo veliko verjetnost zadetka. Majhen delež zadetkov običajnega tipa je posledica večjega števila tovrstnih strelov. Koncept verjetnosti zadetka ostalih strelov je podobno povezan z deležem strelov in njihovo težavnostjo.

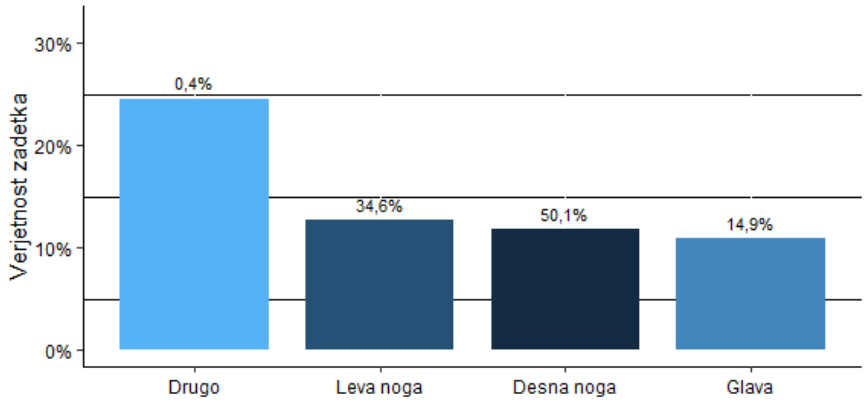



Slika 3.11: Verjetnost zadetka in delež strelov glede na tehniko strela.

### 3.3.5 Del telesa

Poglejmo, ali se delež strelov pretvorjenih v zadetek vidno razlikuje tudi glede dela telesa, s katerim je bil strel izveden. Strel je lahko izveden z nogo, kjer ločimo levo in desno, z glavo ali z drugim delom telesa (ki ni roka), kar pa je redko.

Po sliki 3.12 vidimo, da je najvišja stopnja realizacije dosežena z delom telesa, ki ni glava ali noga. To je posledica izredno majhnega števila tovrstnih strelov, ki so običajno posledica odbite žoge in srečnih situacij, ki se zgodijo zelo blizu gola in se v 25% pretvorijo v zadetek. Zanimivo je tudi dejstvo, da je realizacija z levo nogo nekoliko višja kot z desno. Z deležem levičarjev oziroma desničarjev v populaciji približno sovpada delež strelov z eno ali drugo nogo. Če ne upoštevamo strani strela glede na širino igrišča, bomo zaradi neupoštevanja preferenčne noge igralca v nadaljnji analizi ta dva strela enačili, saj bi nam to sicer prineslo lažno informacijo (npr. naj desničarji raje poskušajo streljati z levo nogo).



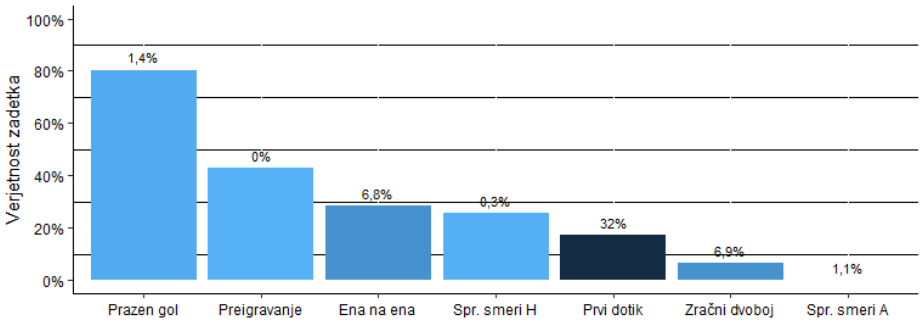

Slika 3.12: Verjetnost (višina stolpca) zadetka in delež strelov glede na del telesa.

### 3.3.6 Binarne spremenljivke

V podatkih imamo nekaj binarnih spremenljivk. Te spremenljivke so: prazen gol v času strela, sprememba smeri žoge zaradi nasprotnega (A) igralca, sprememba smeri žoge zaradi ekipnega (H) igralca, 'ena na ena' z vratarjem (med strelcem in vratarjem ni nasprotnih igralcev), strel iz prvega dotika, strel po preigravanju in strel iz dobljenega zračnega dvoboja. Kot že pri tipu strela, so te spremenljivke

lahko posredno določene z vrednostmi nekaterih ostalih spremenljivk in z njimi močno korelirane.

Slika 3.13 prikazuje, da največjo verjetnost zadetka prinaša strel na gol, v katerem ni vratarja. Razmeroma majhna stopnja konverzije je pri soočenju igralca z vratarjem brez vmesnih obrambnih igralcev, slabih 30%, kar pove, da to v resnici ni tako velika priložnost, kot se nam morda zdi. Skoraj ničelna je uspešnost strelov po spremembi žoge zaradi nasprotnega igralca. Večina teh strelov je uspešno blokiranih.




Slika 3.13: Verjetnost (višina stolpca) zadetka in delež strelov glede na binarno spremenljivko.

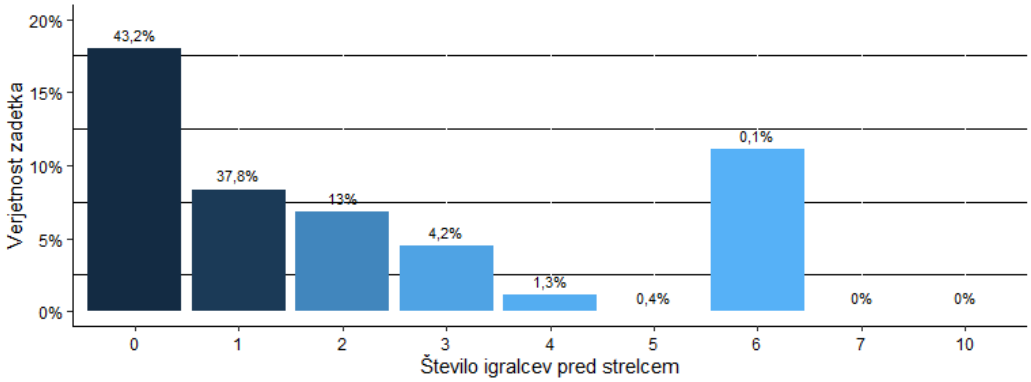
### 3.3.7 Postavitev obrambnih igralcev

Pomembna informacija pri določanju verjetnosti zadetka je postavitev obrambnih igralcev v času strela. V podatkih imamo več močno koreliranih spremenljivk, ki določajo postavitev obrambnih igralcev. To so: število obrambnih igralcev za žogo, število obrambnih igralcev med strelcem in golom (v trikotniku med vratnicama in strelcem), najmanjša kvadratna površina, ki pokrije vse branilce, gostota obrambnih igralcev za žogo med strelcem in golom (v trikotniku med vratnicama in strelcem), ki se izračuna kot agregirana inverzna razdalja za vsakega branilca za žogo, ter razdalja med strelcem in najbližjim branilcem in strelcem in drugim najbližjim branilcem.

Menimo, da največ informacije nosijo spremenljivke o številu branilcev v trikotniku med strelcem in obema vratnicama, njihova gostota in morda razdalja do

najbližjega branilca. Prvi dve spremenljivki sta močno korelirani, nekoliko lažja za predstavo pa je informacija o številu branilcev.

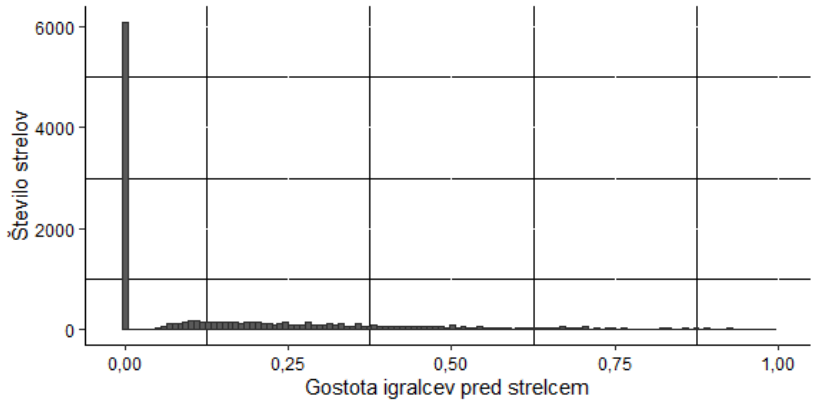
Število branilcev med strelcem in obema vratnicama v teoriji variira med 0 in 10, kolikor je nogometašev (brez vratarja) nasprotne ekipe na igrišču, vendar so višje vrednosti redkejše. Pri številu šestih igralcev, kjer je glede na sliko 3.14 verjetnost zadetka visoka, je potrebno upoštevati, da je delež tovrstnih strelov skoraj ničelen. Do večine strelov, kot tudi deleža zadetkov, pride pri 0, 1 ali 2 nasprotnih igralcev med strelcem in vratnicama.




Slika 3.14: Verjetnost zadetka (višina stolpca) in delež strelov glede na število obrambnih igralcev v trikotniku med vratnicama in strelcem.

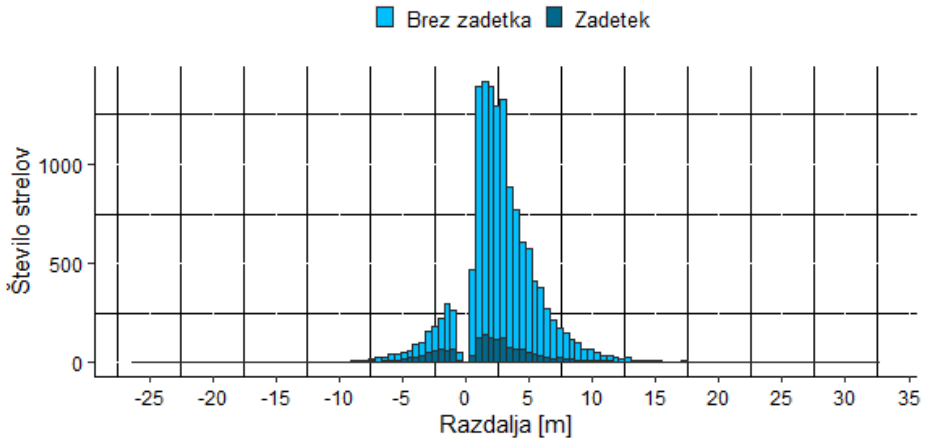
Gostota obrambnih igralcev med strelcem in vratnicama poleg števila igralcev upošteva tudi razdaljo do gola in zavzame vrednosti med 0 in 1. Kot smo opazili na sliki 3.14 in prikazuje tudi slika 3.15, je večina strelov izvedena z ničelno gostoto obrambnih igralcev. Do večjih gostot pride predvsem ob večji oddaljenosti strelca do gola, kjer pa je verjetnost zadetka manjša. Spremenljivka ima tako nizko variabilnost in nosi malo informacije o verjetnosti zadetka.

Na sliki 3.16 si lahko ogledamo še razdaljo od strelca do prvega obrambnega igralca. Ta je podobna razdalji do drugega obrambnega igralca, vendar nosi večjo informacijo. Zavzame tako pozitivne kot negativne vrednosti. Negativne vrednosti ustrezajo obrambnim igralcem, ki so v času strela za žogo, pozitivne pa obrambnim igralcem, ki so med strelcem in golom. Bližje ničli kot smo, bližje je nasprotni branilec strelcu in težje je doseči zadetek. Pri negativnih vrednostih je delež zadetkov večji in raste s povečevanjem (negativne) razdalje. Pri razdalji




Slika 3.15: Porazdelitev gostote obrambnih igralcev v trikotniku med vratnicama in strelcem.

0 je viden globok padec, ki je posledica skoraj nemogočega strela ob tesnem pokrivanju nasprotnega igralca.




Slika 3.16: Porazdelitev razdalje do prvega obrambnega igralca, pogojno na indikator zadetka.

### 3.3.8 Podaja za strel

Nazadnje si oglejmo, katere vrste podaj vodijo do strela in kako so od njih odvisni zadetki. V naših podatkih do 78% strelov iz odprte igre pride iz predhodne ključne podaje in tako je 22% strelov iz odprte igre brez podaje. Od vseh strelov iz podaje

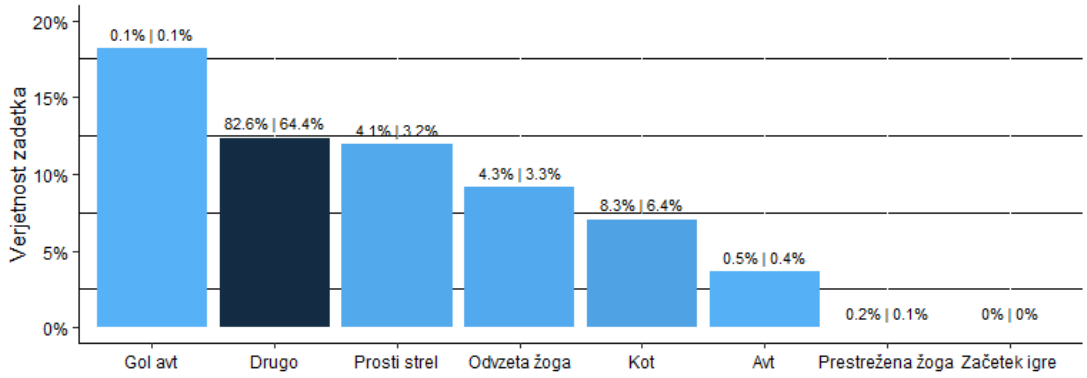
je delež zadetkov enak 11,7%, medtem ko je delež zadetkov, ki so posledica strela brez ključne podaje, enak 13,1%. Med spremenljivkami, ki opisujejo podajo, bomo upoštevali tiste, za katere menimo, da bi lahko pomembneje vplivale na verjetnost zadetka iz strela, ki podaji sledi:

- ali je bila podaja nazaj (znotraj 5-metrskega prostora nazaj znotraj kazenskega prostora),
- ali je bila podaja predložek (podaja s strani na nasprotno stran kazenskega prostora),
- ali je bila podaja v prostor (za zadnjo linijo obrambe),
- višina podaje (po tleh, polvisoko - pod rameni, visoko - nad rameni),
- ali je šlo za prenos igre (žoga prepotuje vsaj 50% širine igrišča),
- dolžina podaje,
- vrsta podaje (nič od naštetega, odvzeta žoga, kot, prosti strel, avt, prestrežena žoga, začetek igre, gol avt) in
- tehnika podaje iz kota (zavrtinčeno proti golu, zavrtinčeno od gola, naravnost).

Pri interpretaciji podaj velja opozoriti, da opazujemo verjetnost zadetka kot posledico strela po eni izmed vrst podaje. Ta verjetnost torej ne določa nujno verjetnosti zadetka po sami podaji. Če bi želeli svetovati glede podaj, bi morali vključiti informacijo o deležu tovrstnih podaj, ki vodijo v strel, katerega posledico opazujemo (pogojna verjetnost), s čimer pa se v tem delu ne ukvarjamo.

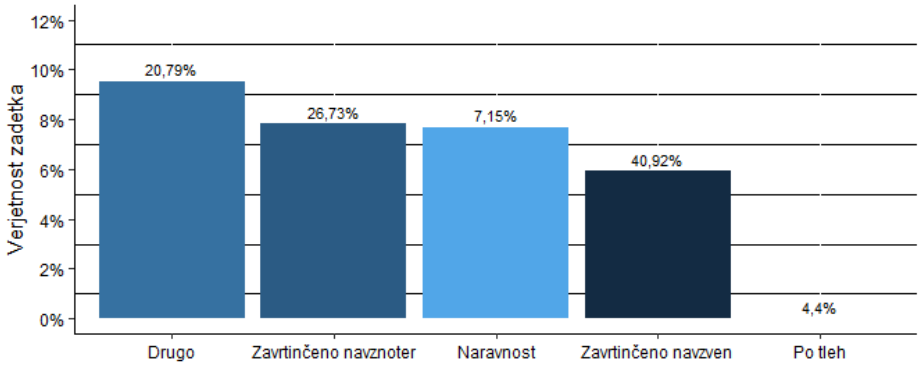
Na sliki 3.17, kjer so dodani deleži ustreznih strelov zgolj glede na strele iz podaje in deleži ustreznih strelov glede na vse strele, torej iz podaj in brez predhodne podaje, vidimo, da je največja verjetnost zadetka po strelu, ki je posledica podaje iz gol avta, torej prostega udarca vratarja pred svojim golom. Tovrstnih strelov je izjemno malo, v naših podatkih zgolj 0,1%. Če pride do strela po tovrstni podaji, je velika verjetnost, da je šlo za napako obrambe nasprotne ekipe in zato za veliko verjetnost, da pride do lepe priložnosti za zadetek. To tako ni spodbuda za dolge podaje iz gol avta napadalcu. Podobno velja tudi pri podajah s prostega strela. Večina strelov kot posledica podaje s prostega strela se zgodi iz primerne lokacije prostega strela, torej napadalne polovice blizu kazenskega prostora. Pri podajah iz kota je (skoraj) vedno cilj podaja za strel,

ki pa ne uspe vedno in tako pri interpretaciji ponovno upoštevamo precej manjši odstotek verjetnosti zadetka. Pri strelu kot posledica podaje iz avta velja podobno kot pri prostih strelih.




Slika 3.17: Verjetnost zadetka (višina stolpca) in delež strelov glede na podaje in skupno, glede na vrsto podaje.

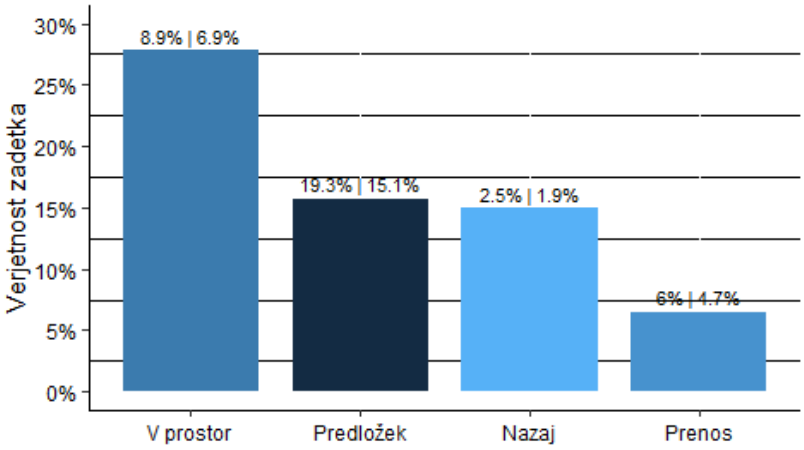
Podaje iz kota, kjer do zadetka pride v dobrih 6% strelov, lahko nekoliko razčlenimo. Kot je prikazano na sliki 3.18, je največji delež podaj iz kota, po katerih pride do strela, zavrtinčenih navzven, stran od gola, vendar te ne prinesejo največje verjetnosti zadetka. Največji delež zadetkov pride iz strelov, ki so posledica podaj iz kota, ki ne spadajo v nobeno drugo skupino. Verjetnost zadetka je v tem primeru slabih 10%.




Slika 3.18: Verjetnost zadetka (višina stolpca) in delež strelov, glede na vrsto podaje iz kota.

Podajo lahko obravnavamo tudi z vidika binarnih spremenljivk, med katerimi

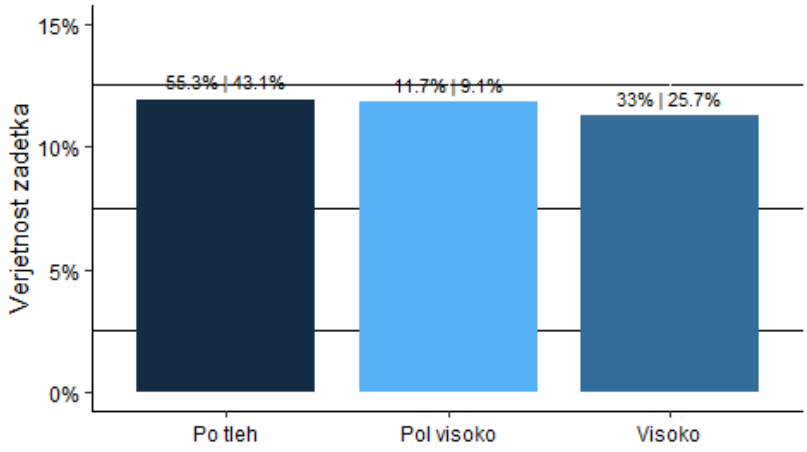
so nekatere močno korelirane z že predstavljenimi. Tu upoštevamo podaje v prostor (za zadnjo linijo obrambe), predložek (s strani na drugo stran kazenskega prostora), podaje nazaj (znotraj kazenskega prostora nazaj) in prenos igre (žoga prepotuje vsaj 50% igrišča). Glede na sliko 3.19 je v primeru strela na gol največja verjetnost zadetka po podaji v prostor, kjer se strelec velikokrat znajde sam pred vratarjem. Sledita predložek s strani in strel kot posledica podaje nazaj, ki je skoraj vedno vedno blizu nasprotnega gola.




Slika 3.19: Verjetnost zadetka (višina stolpca) in delež strelov glede na podaje in skupno, glede na binarno spremenljivko.

Višino podaje upoštevamo preko ordinalne spremenljivke s tremi kategorijami: podaja po tleh, polvisoka podaja (nižje od ramen) in visoka podaja (višje od ramen). Slika 3.20 prikazuje, da je največji delež strelov posledica podaje po tleh in najmanjši delež strelov posledica polvisoke podaje. Med verjetnostjo zadetka po uspešno izvedenem strelu med upoštevanimi višinami podaje ni večjih razlik.

Podaje, ki vodijo do strela, lahko razdelimo glede na njihovo dolžino. Na sliki 3.21 hitro opazimo, da podaja med 85 in 90 metri, če do strela pride, v 50% vodi v zadetek. V veliki večini so to opisane podaje vratarja iz gol avta. Med ostalimi dolžinami podaj ni večjih razlik v verjetnosti zadetka po strelu.



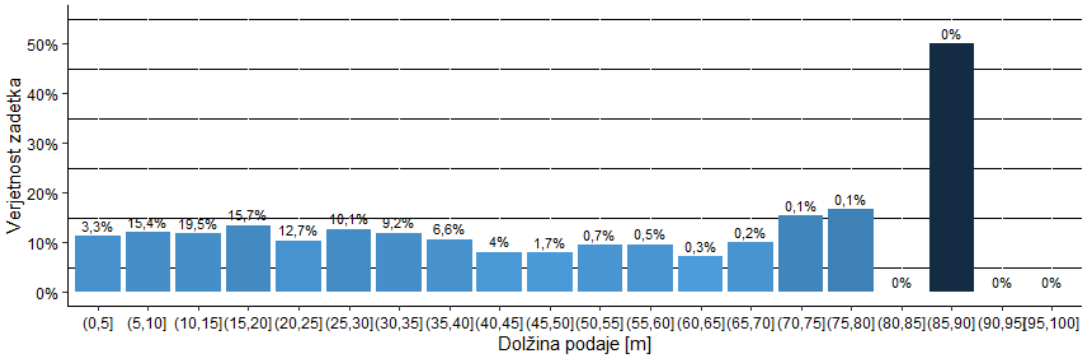

Slika 3.20: Verjetnost zadetka (višina stolpca) in delež strelov glede na podaje in skupno, glede na višino podaje.

## 3.4 Predstavitev modela za izračun xG

Na podlagi opisanega smo zgradili model logistične regresije za oceno verjetnosti zadetka oziroma mere xG. Končni model smo izbrali na podlagi mer ocenjevanja kakovosti regresijskega modela in subjektivne presoje, predvsem glede interpretacije in smiselnosti.

Pred in med modeliranjem smo preverili vse predpostavke modela logistične regresije, ki so zadovoljivo izpolnjene. Za opazovane strele lahko rečemo, da so neodvisni, čeprav lahko več strelov pripada istemu igralcu ali pa sta dva strela posledica iste akcije, vendar bomo to možnost zanemarili. V postopku modeliranja smo z izbiro spremenljivk poskrbeli, da ni prišlo do multikolinearnosti. Ugotovili smo, da je nekaj (manj kot 10) vrednosti nekoliko izstopajočih, t.i. osamelcev, vendar jih iz podatkov nismo izključili, saj niso posledica napačne meritve. V nogometu namreč pride do enkratnih in neobičajnih dogodkov, ki jih je smiselno upoštevati. Predpostavko linearnega razmerja med številskimi neodvisnimi spremenljivkami in logit transformacijo odvisne spremenljivke smo preverili grafično in s tem ni bilo težav. Število zadetkov pa tudi presega 10-kratnik števila napovednih spremenljivk.

Krajše bomo predstavili postopek gradnje logističnega modela, nekatera spo-




Slika 3.21: Verjetnost zadetka (višina stolpca) in delež strelov glede na dolžino podaje.

znanja pri tem in katere spremenljivke in zakaj smo v model vključili. Predstavili in interpretirali bomo vpliv spremenljivk. Grafično in z nekaj merami bomo prikazali in kvantificirali kakovost modela. Na koncu bomo z uporabo krivulje ROC in mere AUC primerjali dobljene vrednosti z vrednostmi StatsBomb.

### 3.4.1 Izbira spremenljivk

Pri opisu pojasnjevalnih spremenljivk nismo omenili lokacije strela, podane s koordinatama $x$ in $y$. Ta posredno določa oddaljenost od gola in kot strela. Ker sta para spremenljivk močno korelirana, v modelu lahko nastopa zgolj en par. Izkazalo se je, da lokacija strela pojasni manjši delež variabilnosti zadetka kot kombinacija oddaljenosti od gola in kota strela na gol. Slednji spremenljivki sta tudi veliko lažje interpretabilni. Prvi spremenljivki, ki smo ju vključili v model, sta tako oddaljenost strela do gola (Razdalja) in kot strela na gol (Kot).

Pri opisu spremenljivk smo predstavili tri različne kote strela na gol. Opazili smo, da je kot odprtja močno koreliran z oddaljenostjo od gola, zato ga v model ne moremo vključiti. Vprašanje je, ali upoštevati absolutni kot, ki zavzame vrednosti [0,90] stopinj in ne loči med levo in desno stranjo igrišča, ali upoštevati polni kot, ki zavzame vrednosti [0,180] stopinj in tako loči med levo in desno stranjo igrišča. Polni kot bi bilo smiselno upoštevati zgolj v interakciji z nogo strela (leva, desna) in izkazalo se je, da spremenljivki absolutni kot in del telesa za strel (leva noga, desna noga, glava, ostalo) kot samostojni spremenljivki pojasnita značilno večji

delež variabilnosti verjetnosti zadetka kot interakcija med polnim kotom in delom telesa za strel. V modelu smo tako upoštevali absolutni kot, ki ne loči med levo in desno stranjo igrišča, in na podlagi njegove definicije smo del telesa za strel upoštevali kot binarno spremenljivko, ki loči zgolj, ali je do strela prišlo z nogo ali ne (Noga), torej ne loči med levo in desno nogo ter enači glavo in ostale dele telesa, ki niso noga. Med spremenljivkama nismo upoštevali interakcije.

Pri izbiri spremenljivke postavitve obrambnih igralcev smo razmišljali, ali v model vključiti število obrambnih igralcev med strelcem in golom (Število branilcev) ali spremenljivko gostote obrambnih igralcev (Gostota branilcev). Pearsonov korelacijski koeficient med spremenljivkama znaša 0,53, obenem pa vrednosti variančno inflacijskega faktorja spremenljivk v modelu nista kritični. Ker sta obe spremenljivki v modelu statistično značilni in nista premočno korelirani, smo v model vključili obe. Število obrambnih igralcev smo zaradi možnega različnega preskoka (tj. nemonotonosti) vpliva med številom igralcev upoštevali kot kategorično spremenljivko, in sicer z 8 razredi (0, ..., 7 in več).

Nekaj težav smo imeli pri kategoričnih spremenljivkah podaj. Te zavzamejo vrednosti zgolj pri strelih, ki so neposredna posledica podaje in imajo tako v originalnih podatkih precej manjkajočih vrednosti. Manjkajoče vrednosti smo uvodoma nadomestili z razredom 'Brez podaje', kar pa se ni izkazalo najbolje, saj ob vključitvi dveh ali več tovrstnih spremenljivk v modelu pride do popolne multikolinearnosti. V ta namen smo v negativni razred (binarne spremenljivke) ali v razred 'Ostalo', vključili razred brez podaje in odpravili težavo multikolinearnosti. Definirali smo tudi novo binarno spremenljivko Podaja, ki je označevala, ali je do podaje sploh prišlo. Na koncu se je izkazalo, da sta edini značilni spremenljivki, ki opisujeta podajo, njena višina (Višina podaje) in podaja v prostor (Podaja v prostor). Ker nam model dopušča vključitev ene spremenljivke z razredom 'Brez podaje', smo ta razred pustili v spremenljivki višine podaje, ki tako zavzame vrednosti 'Brez podaje', 'Po tleh', 'Pol visoko' in 'Visoko', medtem ko je podaja v prostor binarna spremenljivka in zavzame vrednosti Da in Ne.

Pri opisu strela so se kot značilne binarne spremenljivke v modelu izkazale: strel po dobljenem zračnem dvoboju (Zračni dvoboj), strel iz prvega dotika (Prvi dotik), sprememba smeri žoge zaradi soigralca (Sprememba smeri), strel na prazen gol (Prazen gol) in strel pri soočenju z vratarjem ena na ena (Ena na ena).

Precej neodločni smo bili pri vključitvi spremenljivke tehnike strela (Tehnika strela), kjer ločimo med običajnim strelom, lobom, nizko glavo, strelom s peto, polvolejem, volejem in škarjicami. Spremenljivka je kot celota v osnovnem logističnem modelu značilna in večina mer za oceno kakovosti modela je kazala v izboljšanje modela (test razmerja verjetij, AUC, Nagelkerkejev psevdo-$R^2$, McFaddenov psevdo-$R^2$), vendar pa je primerjava verjetnosti na podlagi samovzorčenja kazala v nekolikšno poslabšanje kakovosti modela. Naposled smo se tudi na podlagi subjektivnega mnenja le odločili za vključitev spremenljivke v model, čeprav da grafična primerjava verjetnosti nekoliko slabšo sliko.

Spremenljivke oddaljenost prvega obrambnega igralca, vrsta podaje, tehnika podaje, dolžine podaje in binarne spremenljivke (sprememba smeri zaradi nasprotnega igralca, predložek, podaja nazaj in prenos igre) ob upoštevanju vseh ostalih spremenljivk v modelu ne vplivajo značilno na verjetnost zadetka in v model niso vključene. Poskusili smo tudi ustvariti nekatere nove spremenljivke in združevati spremenljivke v skupine, vendar to ni prineslo izboljšanja modela.

Pri vključitvi spremenljivk v model smo uporabili ročno metodo poskusov in napak ter sprotno dodajanje spremenljivk v model, kjer smo začeli z ničelnim modelom in v model postopoma dodajali spremenljivke, ki so se nam zdele najbolj pomembne pri pojasnjevanju verjetnosti zadetka. Spremenljivke smo tako izbrali glede na njihovo značilnost v modelu, na podlagi izboljšanja kakovosti modela, glede na izbrane mere in nekaj subjektivne presoje. Pri izbiri spremenljivk bi si lahko pomagali tudi z metodami kot so Lasso ali Ridge in Elastic net, ki s pomočjo regularizacije izberejo spremenljivke, ki bi jih bilo v model smiselno vključiti. Izkazalo se je, da v našem primeru te metode ne delujejo dobro, oziroma tovrstna izbira spremenljivk ne prinese izboljšanja modela, saj so metode izbire spremenljivk preveč odvisne od parametrov, ki določajo krčenje koeficientov proti 0 in je zaradi težave multikolinearnosti še vedno potrebna subjektivna izbira spremenljivk.

### 3.4.2 Vpliv spremenljivk

V tabeli 3.1 so za uporabljeni model logistične regresije, za posamezne spremenljivke oziroma njihove razrede, poleg vrednosti $p$ navedene vrednosti koeficientov

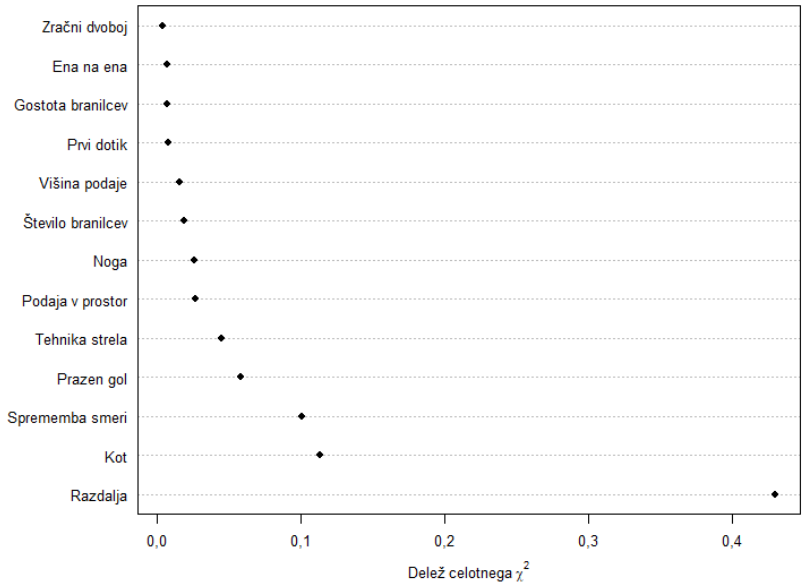
$b_i$. Dodani sta tudi spodnja in zgornja meja 95% intervalov zaupanja, izračunani kot (Koeficient $\pm$ $1,96 \cdot$ Standardna napaka). Tako koeficienti kot spodnje in zgornje meje njihovih 95% intervalov zaupanja so za lažjo interpretacijo predstavljeni na eksponentni lestvici in jih lahko imenujemo tudi razmerje obetov (OR, angl. Odds Ratio). Pri interpretaciji velja upoštevati, da vsako povečanje ali zmanjšanje verjetnosti zadetka zaradi njegove splošno majhne verjetnosti pomeni večjo spremembo, kot se nam morda zdi.

Tabela 3.1: Spremenljivke in koeficienti modela xG

| Spremenljivka | OR | Spodnja meja IZ OR | Zgornja meja IZ OR | vrednost $p$ |
|---|---|---|---|---|
| Konstanta | 0,322 | 0,229 | 0,454 | $<$ 0,001 |
| Razdalja | 0,855 | 0,845 | 0,865 | $<$ 0,001 |
| Kot | 1,021 | 1,018 | 1,024 | $<$ 0,001 |
| Noga | 2,333 | 1,803 | 3,018 | $<$ 0,001 |
| Število Branilcev - 1 vs. 0 | 0,733 | 0,613 | 0,875 | $<$ 0,001 |
| Število Branilcev - 2 vs. 0 | 0,600 | 0,457 | 0,788 | $<$ 0,001 |
| Število Branilcev - 3 vs. 0 | 0,402 | 0,254 | 0,637 | $<$ 0,001 |
| Število Branilcev - 4 vs. 0 | 0,092 | 0,021 | 0,397 | 0,001 |
| Število Branilcev - 5 vs. 0 | 0,000 | 0,000 | 0,000 | 0,945 |
| Število Branilcev - 6 vs. 0 | 1,393 | 0,157 | 12,334 | 0,766 |
| Število Branilcev - 7 in več vs. 0 | 0,000 | 0,000 | 0,000 | 0,987 |
| Gostota Branilcev | 0,642 | 0,499 | 0,827 | $<$ 0,001 |
| Tehnika strela - Peta vs. Običajno | 0,289 | 0,106 | 0,783 | 0,015 |
| Tehnika strela - Nizka glava vs. Običajno | 1,433 | 0,731 | 2,806 | 0,294 |
| Tehnika strela - Polvolej vs. Običajno | 0,582 | 0,466 | 0,727 | $<$ 0,001 |
| Tehnika strela - Lob vs. Običajno | 2,488 | 1,766 | 3,505 | $<$ 0,001 |
| Tehnika strela - Škarjice vs. Običajno | 0,430 | 0,226 | 0,817 | 0,01 |
| Tehnika strela - Volej vs. Običajno | 0,644 | 0,497 | 0,834 | $<$ 0,001 |
| Zračni dvoboj | 0,656 | 0,472 | 0,911 | 0,012 |
| Prvi Dotik | 1,298 | 1,123 | 1,501 | $<$ 0,001 |
| Sprememba smeri | 10,982 | 7,560 | 15,954 | $<$ 0,001 |
| Prazen gol | 6,779 | 4,587 | 10,018 | $<$ 0,001 |
| Ena na ena | 1,401 | 1,148 | 1,709 | $<$ 0,001 |
| Višina podaje - Po tleh vs. Brez podaje | 0,803 | 0,681 | 0,947 | 0,009 |
| Višina podaje - Visoko vs. Brez podaje | 0,628 | 0,513 | 0,769 | $<$ 0,001 |
| Višina podaje - Polvisoko vs. Brez podaje | 0,644 | 0,512 | 0,810 | $<$ 0,001 |
| Podaja v prostor | 1,816 | 1,516 | 2,176 | $<$ 0,001 |

Vse številske spremenljivke so seveda močno statistično značilne, medtem ko to ne velja nujno za vse razrede kategoričnih spremenljivk. Pri interpretaciji spremenljivk in njihovi pomembnosti je pri pojasnjevanju verjetnosti zadetka poleg vrednosti koeficientov smiselno upoštevati tudi vrednosti $\chi^2$, prikazane na sliki 3.22. Gre za deleže celotnega $\chi^2$ upoštevanih spremenljivk, katerih značilnost

preverjamo z Waldovim testom. Velja, da večji kot je delež oziroma vrednost $\chi^2$, bolj pomembna je spremenljivka v modelu.




Slika 3.22: Delež $\chi^2$ spremenljivk xG modela v Waldovem testu.

Po pričakovanju daleč največji delež variabilnosti zadetka oziroma njegove verjetnosti pojasni razdalja od mesta strela do gola. Z vsakim metrom, ko se oddaljimo od gola, se obeti zadetka v povprečju zmanjšajo 0,145-kratno oziroma za 14,5% prejšnje vrednosti. Zaradi močne značilnosti spremenljivke je interval zaupanja ozek in enak [0,135, 0,155].

Ostale spremenljivke so glede na delež $\chi^2$ precej podobne. Izpostavimo lahko kot strela na gol in binarno spremenljivko sprememba smeri strela zaradi soigralca. Z vsako stopinjo povečanja kota, ki se odraža v strelu bolj iz sredine igrišča, se obeti zadetka v povprečju pričakovano povečajo, in sicer 1,021-kratno. To je morda manj kot bi pričakovali, a s strani je velikokrat celo lažje zadeti kot s strela popolnoma na sredini širine igrišča. Pri spremembi smeri žoge velja, da ta v povprečju poveča obete zadetka kar 11-kratno; 95% interval zaupanja je precej širok: [7,5-kratno, 16-kratno].

Za število branilcev med strelcem in golom velja, da večje kot je število bra-

nilcev, manjša je verjetnost zadetka. Faktor zmanjšanja obetov zadetka od nič do treh branilcev je relativno konstanten in velja, da se ob povečanju števila branilcev za 1 obeti zadetka v povprečju zmanjšajo 0,15-kratno. Preskok iz treh v štiri branilce je večji, in sicer se obeti zadetka v povprečju zmanjšajo 0,3-kratno. Med številom branilcev 5, 6, 7 ali več in 0 ni značilnih razlik v verjetnosti zadetka; razlog je najbrž v majhnem vzorcu strelov pri več kot 4 branilcih. Tudi s povečanjem gostote branilcev med strelcem in golom se verjetnost zadetka zmanjša, kar pa je težje interpretirati.

Podobno interpretiramo ostale spremenljivke, kjer velja, da je verjetnost zadetka po strelu z nogo pričakovano večja od verjetnosti po strelu z glavo. Pri tehniki strela velja, da strel z lobom v primerjavi z običajnim strelom poveča verjetnost zadetka, kar pa, kot je bilo že razloženo, velja upoštevati z rezervo. Na drugi strani strel s peto, polvolejem, volejem ali škarjicami v primerjavi z običajnim strelom pričakovano zmanjša verjetnost zadetka, med strelom z nizko glavo in običajnim strelom pa ni značilnih razlik. Strel po dobljenem zračnem dvoboju nekoliko zmanjša verjetnost zadetka, medtem ko strel na prvi dotik, situacija ena na ena z vratarjem, podaja v prostor in še posebej strel na prazen gol povečajo verjetnost zadetka.

Zanimivo velja, da je ob upoštevanju ostalih spremenljivk v modelu verjetnost zadetka po predhodni ključni podaji manjša, kot če do podaje ne pride. To je verjetno spet posledica tega, da se nogometaši za strele odločajo le, ko imajo nekoliko več prostora in tako večjo priložnost za dosego zadetka, sicer pa podajo ostalim. Hrkrati velja, da višja kot je podaja, manjša je verjetnost zadetka, čeprav razlike v verjetnosti, še posebej med podajami po zraku, niso velike.

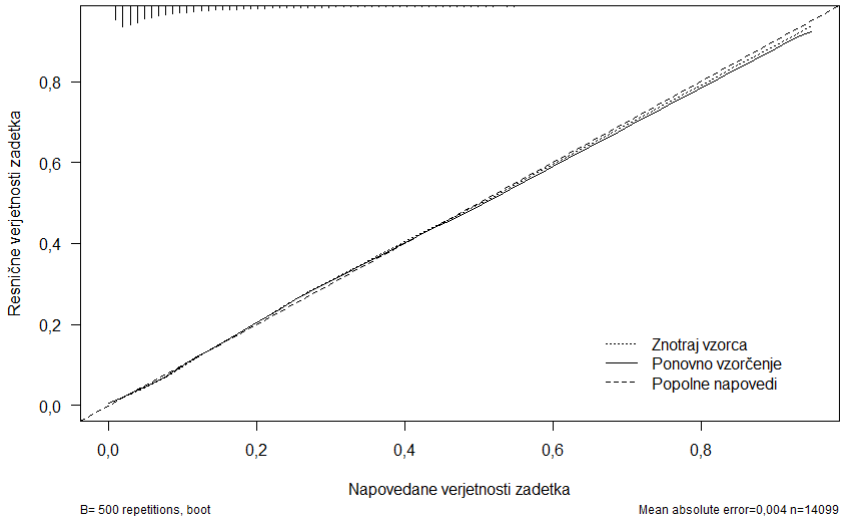
### 3.4.3 Kakovost modela

Končni model smo izbrali na podlagi različnih mer kakovosti. Zgolj za informacijo, čeprav nam te vrednosti same po sebi ne povedo veliko, ampak nam bolj pomagajo pri primerjavi modelov, je v končnem modelu McFaddenov psevdo-$R^2$ enak 0,221, Nagelkerkejev psevdo-$R^2$ enak 0,288 in Akaikejev informacijski kriterij enak 8119.

Bolj se velja osredotočiti na primerjavo napovedanih in opazovanih vrednosti.

Te smo pridobili prek samovzorčenja, kjer smo vzorčenje s ponavljanjem ponovili $B = 500$-krat.

Na grafu 3.23 vertikalne črtice nad osjo $x$ ponazarjajo porazdelitev napovedanih verjetnosti. Dolge presekane črtice prikazujejo popolno napoved, kjer so napovedane verjetnosti enake opazovanim vrednostim. Kratke presekane črtice prikazujejo kalibracijo oziroma napoved znotraj vzorca, kjer je testna množica enaka učni. Polna črta pa prikazuje kalibracijo oziroma napoved s postopkom ponovnega vzorčenja in prilagodi napovedi za preprileganje, s čimer dobimo predstavo, kako bi napoved delovala na neodvisni testni množici oziroma kako dobro se model posplošuje. Povprečna absolutna napaka (angl. mean absolute error) je povprečna absolutna razlika med napovedanimi in opazovanimi vrednostmi, ki je v idealnem primeru enaka 0.



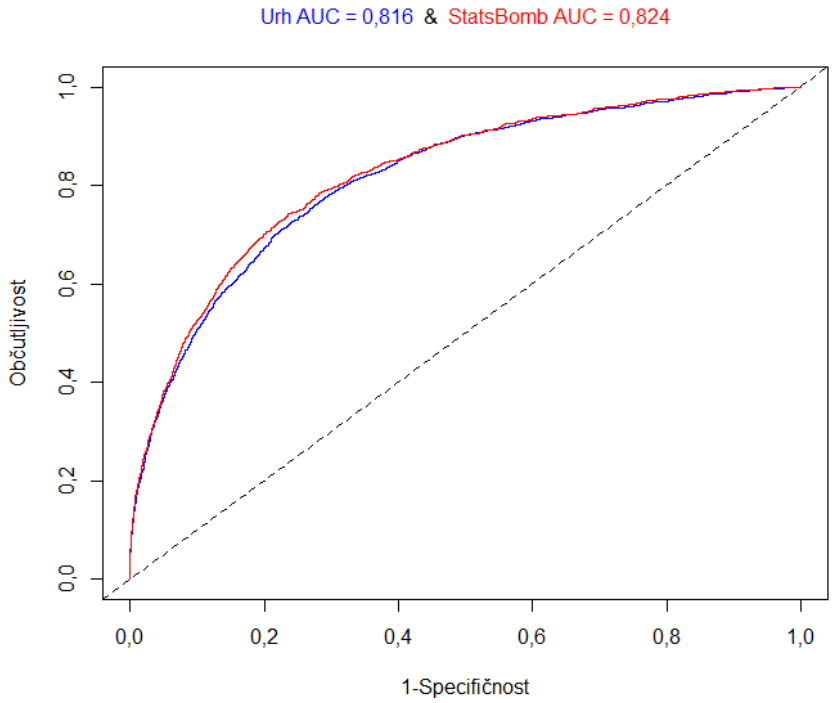

Slika 3.23: Primerjava s samovzorčenjem (B = 500) pridobljenih opaženih in napovedanih vrednosti modela xG.

Vidimo, da je porazdelitev verjetnosti zadetka asimetrična v desno, kjer je večina verjetnosti manjših od 0,4. Če upoštevamo zgolj povprečno absolutno napako, se zdi, da model dobro napoveduje verjetnosti. Povprečna absolutna napaka med napovedanimi in opazovanimi vrednostmi je zgolj 0,004, kar pri ciljni vrednosti 0,13 le 2,7% napaka. Do občutnejšega odstopanja med napovedmi znotraj vzorca in napovedmi na podlagi ponovnega vzorčenja prihaja zgolj pri verjetno-

stih, višjih od 0,4, kjer se model zaradi majhnega vzorca strelov preveč prilagodi učni množici. V splošnem verjetnosti modela sledijo opazovanim vrednostim, do odstopanja pa prihaja predvsem na obeh koncih porazdelitve verjetnosti zadetka, kjer je vzorec učne množice majhen.

S krivuljo ROC in ploščino pod njo (AUC) lahko primerjamo kakovost napovedanih verjetnosti našega modela in dane verjetnosti s strani StatsBomb. Verjetnosti našega modela smo pridobili z metodo prečnega preverjanja v dveh sklopih, kjer smo učno množico podatkov sprva razdelili na 10 približno enako velikih pregibov, nato pa v vsaki iteraciji parametre modela ocenili z 10-kratnim prečnim preverjanjem na devetih pregibih in verjetnosti napovedali na testni množici, ki jo je predstavljal preostali pregib. Na ta način smo dobili napovedane verjetnosti za vsak strel v podatkih in jih primerjali z danimi verjetnostmi StatsBomb. Cilj modeliranja sicer ni bil, da bi se čim bolj približali verjetnostim StatsBomb, temveč ustvariti model, ki da najboljše vrednosti mer splošne kakovosti modela.

Na grafu 3.24 je krivulja ROC StatsBomb nekoliko višja od naše, kar posledično velja tudi za vrednosti AUC. To pomeni, da so verjetnosti StatsBomb nekoliko boljše ocenjene od naših. Naša vrednost AUC je enaka 0,816, kar je več od 0,5, kar bi dal slučajni klasifikator, medtem ko je za StatsBomb AUC enaka 0,824. Na podlagi modeliranja in opazovanja vrednosti AUC različnih modelov lahko rečemo, da sta si vrednosti precej blizu in smo se tako, vsaj glede kakovosti ocen, močno približali ocenam s strani StatsBomb. Če primerjamo pare ocen, je povprečna absolutna razlika verjetnosti enaka 0,042 oziroma 4,2 odstotne točke, kar je glede na izhodiščno vrednost ocene StatsBomb ($\frac{|\text{Urh xG - StatsBomb xG}|}{\text{StatsBomb xG}}$) v povprečju enako 50% (zaradi močno desno asimetrične porazdelitve je mediana enaka 33%). To ni znak, da je naš model slab, saj ni nujno, da so verjetnosti StatsBomb najboljše in smo videli, da je kakovost napovedanih verjetnosti precej blizu. Lahko dodamo še, da je povprečna ocena xG našega modela enaka 0,120, povprečna ocena StatsBomb pa 0,115.




Slika 3.24: Primerjava z dvojnim 10-kratnim prečnim preverjanjem pridobljene krivulje ROC in AUC vrednosti našega modela in vrednosti StatsBomb.

# 4 Pomen porazdelitve pričakovanih zadetkov

Sedaj, ko smo spoznali koncept pričakovanih zadetkov in jih znamo interpretirati, se lahko vprašamo, kako ovrednotiti odstopanje vrednosti xG od dejanskih zadetkov posameznega igralca ali ekipe, kakšen vpliv imajo absolutne vrednosti xG na verjetnost dejanskega števila zadetkov in kako to vpliva na verjetnost zmage ekipe.

## 4.1 Porazdelitev xG in uspešnost igralca

Zanima nas, kako močno v pozitivni oziroma negativni smeri dejanski zadetki igralca odstopajo od njegovih pričakovanih zadetkov na podlagi vrednosti xG.

Vsak strel si lahko predstavljamo kot (v praksi ne popolnoma) neodvisno Bernoullijevo spremenljivko z verjetnostjo dogodka $p = \mathrm{xG}$. Število zadetkov igralca, ki je izvedel $n$ strelov z vrednostmi $\mathrm{xG}_1, \mathrm{xG}_2, ..., \mathrm{xG}_n$, se tako porazdeljuje po posplošeni binomski porazdelitvi, ki jo imenujemo Poisson-Binomska porazdelitev: $Y \sim Poiss - Bin(\mathrm{xG}_1, \mathrm{xG}_2, ..., \mathrm{xG}_n)$.

Njena pričakovana vrednost je enaka $E[Y] = \sum_{i=1}^{n} \mathrm{xG}_i$ in varianca je enaka $Var(Y) = \sum_{i=1}^{n} (1 - \mathrm{xG}_i) \cdot \mathrm{xG}_i$. Verjetnost, da je igralec dosegel $0 \leq k \leq n$ zadetkov, lahko izračunamo kot:

$$P(Y = k) = \sum_{A \in F_{k,n}} \Big( \prod_{i \in A} \mathrm{xG}_i \prod_{j \in F_{k,n} \setminus A} (1 - \mathrm{xG}_j) \Big), \tag{4.1}$$

kjer $F_{k,n}$ označuje množico vseh podmnožic $\{1, ..., n\}$ velikosti $k$.

V [20] je predstavljena ideja, da število zadetkov sledi poisson-binomski porazdelitvi z verjetnostmi $\mathrm{xG}_i$ in da tako ob znanih vrednostih $\mathrm{xG}_i$ lahko za vsako število zadetkov izračunamo njihovo verjetnost. S tem izmerimo odstopanje dejanskega števila zadetkov od pričakovane vrednosti relativno in ne samo z razliko med številom dejanskih zadetkov in kumulativno vrednostjo xG.

Za primer vzemimo dva navidezna igralca z enako kumulativno vrednostjo xG. Recimo, da je igralec A na gol ustrelil 120-krat in je vsak njegov strel imel vrednost $\mathrm{xG} = 0,1$, medtem ko je igralec B na gol ustrelil 20-krat in je vsak njegov strel je imel vrednost $\mathrm{xG} = 0,6$. Skupna vrednost xG obeh igralcev je enaka 12 in od obeh pričakujemo 12 zadetkov.

Recimo še, da je igralec A dosegel 17 zadetkov, igralec B pa 15 zadetkov. Zgolj na podlagi razlik med dejanskim in pričakovanim številom zadetkov bi sklepali, da je igralec A uspešnejši. To je z vidika športnega dosežka res, saj so edino zadetki tisto, kar šteje, vendar pa kvantificirajmo kakovost njune realizacije in tako 'pravičneje' prikažimo njuno uspešnost.

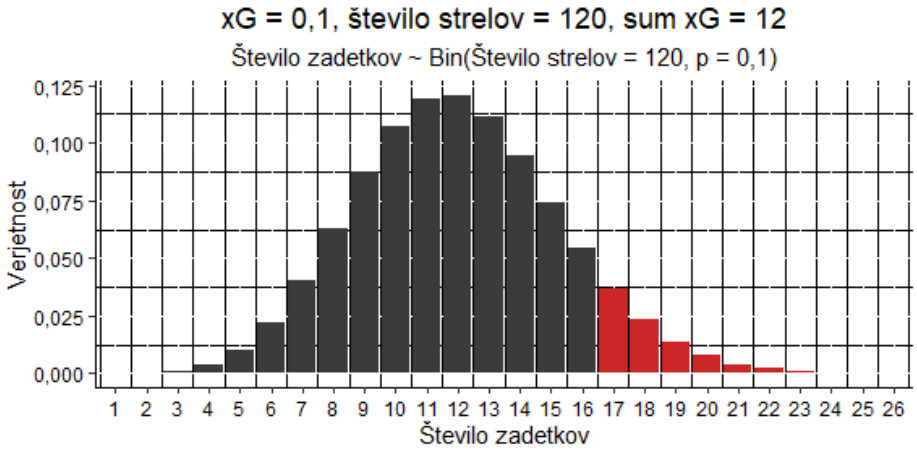
Za igralca A, katerega porazdelitev zadetkov je prikazana na sliki 4.1, na podlagi posamičnih vrednosti xG izračunamo verjetnost, da doseže dejanskih 17 zadetkov ali več kot:

$$P(Y \geq 17 | \mathrm{xG}_i = 0,1; \sum_{i=1}^{n} \mathrm{xG}_i = 12) = 0,053.$$

Podobno za igralca B, katerega porazdelitev zadetkov je prikazana na sliki 4.2, na podlagi posamičnih vrednosti xG izračunamo verjetnost, da doseže dejanskih 15 zadetkov ali več kot:

$$P(Y \geq 15 | \mathrm{xG}_i = 0,6; \sum_{i=1}^{n} \mathrm{xG}_i = 12) = 0,051.$$

Na podlagi dejanskih zadetkov in razlike s kumulativno vrednostjo xG igralca A označimo kot uspešnejšega, saj je dosegel dva zadetka več od igralca B. Če pa



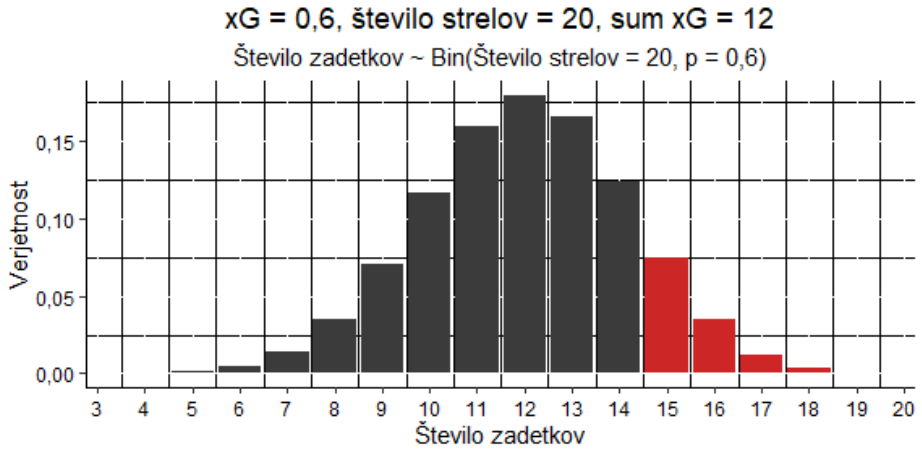

Slika 4.1: Porazdelitev pričakovanega števila zadetkov in dejansko število zadetkov igralca A.

upoštevamo še kakovost njunih priložnosti, ki se odraža v posameznih vrednostih xG vidimo, da je igralec B kljub manjšemu številu dejanskih zadetkov nekoliko bolj presegel pričakovanja kot igralec A.

Spoznali smo, da posamezne vrednosti xG ponujajo veliko več informacije kot vsota zadetkov in kumulativna vrednost xG. Pričakovano število zadetkov v primerjavi z dejanskimi zadetki z upoštevanjem posameznih vrednosti xG postavimo v kontekst in bolje ocenimo stopnjo konverzije posameznega igralca ali ekipe ter njegovo realno uspešnost pretvarjanja priložnosti v zadetek.

## 4.2 Porazdelitev xG in verjetnost zmage

Podobno kot porazdelitev xG vpliva na verjetnost števila zadetkov in njihovo odstopanje od pričakovane vrednosti, velja tudi na ravni tekme. Verjetnost zmage, poraza oziroma remija ekipe je odvisna od kakovosti in količine priložnosti, torej posameznih xG, ne samo vsote njihovih vrednosti. V [21] so bile primerjane verjetnosti števila zadetkov iz velikih in malih priložnosti, z enako kumulativno vrednostjo xG, da bi preizkusili 'moč' velikih priložnosti. Na ta način, si na teoretičnem primeru tekme poglejmo, kako zgolj primerjava kumulativne vrednosti xG ne zadošča za napovedovanje pričakovanega končnega zmagovalca tekme.




Slika 4.2: Porazdelitev pričakovanega števila zadetkov in dejansko število zadetkov igralca B.

Za primer vzemimo tekmo med ekipo A in ekipo B. Naj bo kumulativna vrednost xG obeh ekip enaka 2,1. Zgolj na podlagi tega podatka bi pričakovali, da bo na dolgi rok največja verjetnost neodločenega rezultata, verjetnost zmage ene ali druge ekipa pa enaka. Predpostavimo še, da je ekipa A na gol ustrelila 3-krat in je vsak strel imel vrednost $\mathrm{xG} = 0,7$. Na drugi strani je ekipa B na gol ustrelila 21-krat in je vsak strel imel vrednost $\mathrm{xG} = 0,1$.

V poglavju 5.2 opisana Monte Carlo simulacija z 20.000 ponovitvami, katere rezultati so podani v tabeli 4.1, kaže večjo verjetnost zmage ekipe A, ki je na gol ustrelila zgolj 3-krat vendar bi lahko vsak strel označili kot zelo veliko priložnost. Ekipa B, ki je na gol ustrelila 21-krat, v idealnem scenariju doseže kar 21 zadetkov, medtem ko ekipa A lahko doseže največ 3 zadetke, pa je vseeno verjetnost zmage večja na strani ekipe A. Število mogočih rezultatov, ki prinašajo zmago ekipi z veliko majhnimi priložnostmi, je sicer večje, vendar pa je vsota verjetnosti rezultatov, ki prinašajo zmago ekipi z manjšim številom velikih priložnosti, večja.

Zaključimo lahko, da je kumulativna vrednost xG pomemben in informativen dejavnik pri ocenjevanju 'pravičnega' zmagovalca tekme, vendar ne pove celotne zgodbe, kjer je pomembna tudi porazdelitev vrednosti xG posameznih strelov. Ob predpostavki enake kumulativne vrednosti xG je manjše število večjih priložnosti v teoriji verjetnosti za končno zmago boljše kot veliko število manjših priložnosti. To znanje lahko pomaga trenerjevim taktičnim zamislim in interpretaciji tekme,

Tabela 4.1: Verjetnost zmage v odvisnosti od porazdelitve xG

| | Ekipa A | Ekipa B |
|---|---|---|
| Porazdelitev xG | $0{,}7 \cdot 3$ | $0{,}1 \cdot 21$ |
| Povprečni zadetki | 2,1 | 2,1 |
| Standardni odklon zadetkov | 0,791 | 1,372 |
| Verjetnost zmage [%] | 39,96 | 35,22 |
| Verjetnost remija [%] | 24,82 | 24,82 |
| Verjetnost poraza [%] | 35,22 | 39,96 |

kjer je ena ekipa večino časa tekme v podrejenem položaju in cilja predvsem na protinapade, ki ne prinesejo nujno manjše verjetnosti zmage.

## 4.3 Napadalni in obrambni izkoristek

Kumulativne vrednosti xG ekipe, tako dosežene v napadu kot prejete obrambi, v kombinaciji z dejanskimi doseženimi in prejetimi zadetki predstavljajo pomembno informacijo o izkoristku napadalnega in obrambnega potenciala ekipe oziroma njenih napadalcev in vratarja, razlika prejetih zadetkov oziroma strelov pa da informacijo o povprečni obrambni sposobnosti igralcev oziroma vratarja.

Absolutna vrednost xG v napadu prikazuje kakovost priprave priložnosti in večja kot je, boljše priložnosti si ekipa ustvari. Razlika med dejanskimi in pričakovanimi doseženimi zadetki pa predstavlja kakovost realizacije in primerja napadalno sposobnost igralcev v primerjavi s povprečnim igralcem na tej ravni tekmovanja. Gledano iz obrambne perspektive, absolutne vrednosti xG nasprotne ekipe predstavljajo kakovost dovoljenih priložnosti in velja, da manjša kot je vrednost xG nasprotne ekipe, boljša je obramba ekipe. Dejanska razlika med prejetimi xG in dejanskimi zadetki je bolj informacija o sposobnosti vratarja kot obrambnih igralcev.

Lahko bi rekli, da je glavna informacija, ki prikazuje napadalni izkoristek ekipe, razlika med dejanskimi in pričakovanimi zadetki: večja kot je, večji je napadalni izkoristek ekipe. Glavna informacija obrambnega izkoristka ekipe oziroma vratarja je razlika med pričakovanimi in dejanskimi prejetimi zadetki: večja

kot je, večji je obrambni izkoristek ekipe. Večji kot je napadalni ali obrambni izkoristek ekipe, manj je prostora za izboljšave znotraj trenutnih predstav ekipe.

Tekom sezone izkoristek ekipe predstavlja informacijo o statistični vzdržnosti. Višji kot je izkoristek, manj je dejansko število doseženih oziroma prejetih zadetkov vzdržno in v nadaljevanju sezone je ob identičnih predstavah ekipe pričakovati padec števila doseženih zadetkov oziroma porast števila prejetih zadetkov. Ob koncu sezone izkoristek lahko uporabimo za analizo izkoriščanja priložnosti in obrambnih predstav ter tako najdemo morebitne šibke ali močne člene ekipe.

V nadaljevanju na podlagi [22] predstavili, kako lahko kvantificiramo stopnjo statistične vzdržnosti in s tem prostora za izboljšave ter uvedli mero izkoriščanja napadalnega oziroma obrambnega potenciala ekipe in definirali skupni faktor izkoristka. Analiza bo temeljila na podatkovnem okvirju sezone 2017/18 Premier League, pridobljenem iz [23], sicer pa v lasti podjetja Stats Perform.

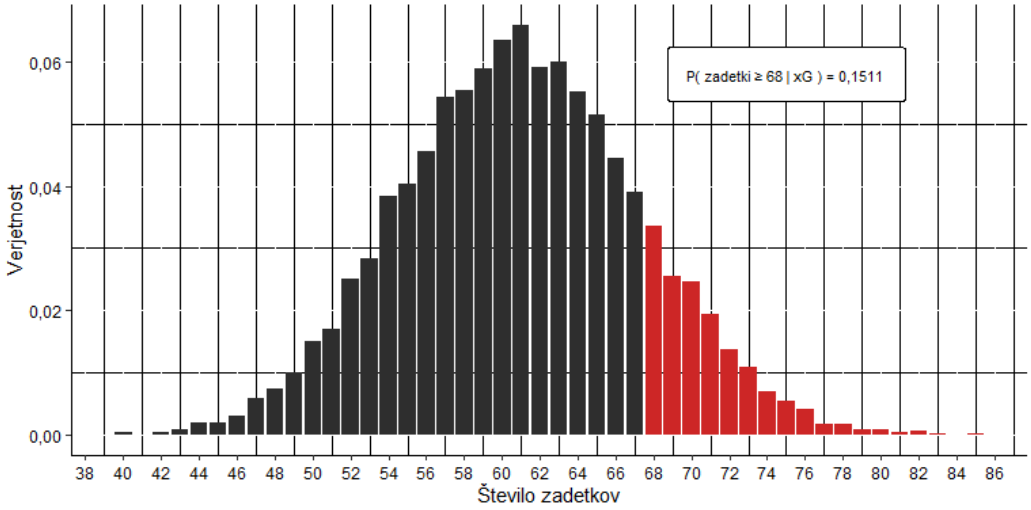
### 4.3.1 Napadalni izkoristek

Osnovna zamisel napadalnega izkoristka ekipe je razlika med dejanskimi in pričakovanimi zadetki. Večja kot je razlika, bolje ekipa izkorišča svoje priložnosti in manj je prostora za povečanje števila doseženih zadetkov. Poleg absolutne razlike je seveda pomembna tudi njena relativna vrednost. Razlika 15-ih zadetkov ne nosi enake informacije pri 80 ali 40 doseženih zadetkih.

Zaradi boljše informacije o napadalnem izkoristku dejansko število zadetkov predstavimo skupaj z njihovo verjetnostno porazdelitvijo na podlagi vrednosti xG. Da dobimo verjetnosti za število zadetkov, ki bi jih ekipa lahko dosegla, ponovno uporabimo v poglavju 5.2 opisano simulacijo Monte Carlo. Na podlagi vrednosti xG vseh strelov znotraj obdobja, ki ga preučujemo, simuliramo vse strele, ter beležimo število zadetkov. Simulacijo posameznega strela si lahko predstavljamo kot realizacijo Bernoullijeve slučajne spremenljivke $\pi = \mathrm{xG}$, strele čez celotno obdobje pa kot realizacijo poisson-binomsko porazdeljene slučajne spremenljivke $Y \sim Poiss - Bin(\mathrm{xG}_1, \mathrm{xG}_2, ..., \mathrm{xG}_n)$, kjer je $n$ število vseh strelov v obdobju. Simulacijo strelov znotraj obdobja ponovimo velikokrat in izračunamo delež posameznega števila zadetkov, ki ga interpretiramo kot verjetnost.

Napadalni izkoristek definiramo kot verjetnost, da pogojno na posamezne vrednosti xG ekipa doseže večje ali enako število zadetkov od dejanskega števila. Bližje ničli kot je ta verjetnost, bolj izkoriščen je napadalni potencial in manj možnosti je za večje število doseženih zadetkov. Obratno, večja kot je verjetnost, več prostora je za izboljšavo.

Za primer vzemimo ekipo Manchester Uniteda. V obravnavanem tekmovanju so igralci Manchestra na nasprotni gol ustrelili 514-krat in dosegli 68 zadetkov. Na podlagi vrednosti xG vsakega strela s pomočjo simulacije dobimo na sliki 4.3 prikazano porazdelitev števila pričakovanih zadetkov.




Slika 4.3: Napadalni izkoristek, Manchester United, Premier League, sezona 2017/18.

Verjetnost, da dosežejo večje ali enako število zadetkov od dejanskih 68, je enaka 0,1511. Njihov napadalni izkoristek je torej visok, čeprav nekaj možnosti za izboljšave še ostaja.

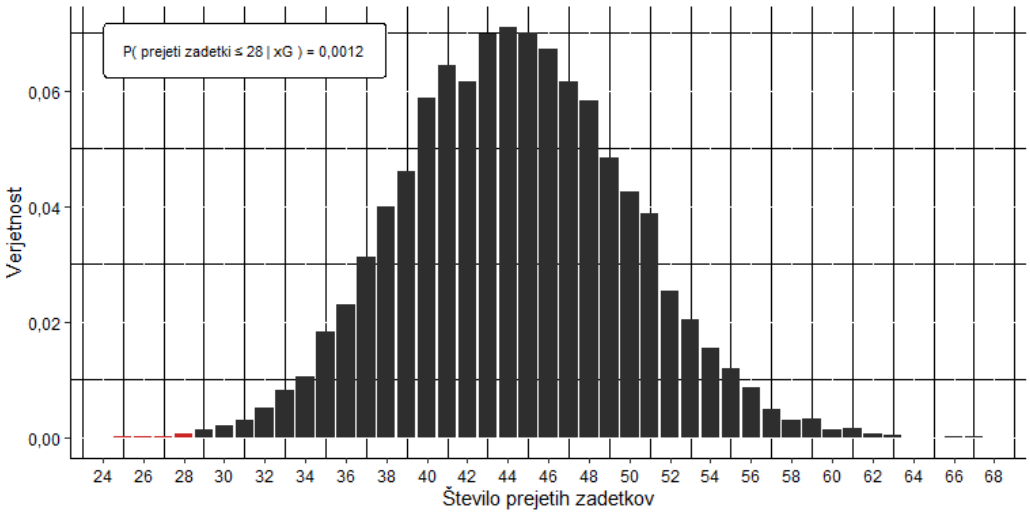
### 4.3.2 Obrambni izkoristek

Podobno napadalnemu izkoristku predstavimo tudi koncept obrambnega izkoristka ekipe. Ponovno nadgradimo razliko med pričakovanimi in dejanskimi prejetimi zadetki, kjer velja, da večja kot je razlika, boljša je ekipa v obrambi oziroma

boljši je njen vratar in manj je prostora za zmanjšanje števila prejetih zadetkov. Tudi obrambni izkoristek ekipe predstavimo na podlagi dejanskega števila prejetih zadetkov in verjetnostne porazdelitve prejetih zadetkov, ki jih ekipa prejme na svoj gol, izračunane na podlagi posameznih vrednosti xG.

Obrambni izkoristek definiramo kot verjetnost, da pogojno na posamezne vrednosti xG ekipa prejme manjše ali enako število zadetkov od dejanskega števila. Bližje ničli kot je ta verjetnost, bolj izkoriščen je obrambni potencial in manj možnosti je za manjše število prejetih zadetkov. Obratno, večja kot je verjetnost, več prostora je v izboljšavi obrambne igre ekipe.

Tudi tu za primer vzemimo ekipo Manchester Uniteda. Na svoj gol so prejeli 434 strelov in 28 jih je končalo v golu. Na podlagi vrednosti xG vsakega strela s pomočjo simulacije dobimo na sliki 4.4 prikazano porazdelitev števila prejetih zadetkov.




Slika 4.4: Obrambni izkoristek, Manchester United, Premier League, sezona 2017/18.

Verjetnost, da Manchester United prejme manjše ali enako število zadetkov od dejanskih 28, je minimalna, saj znaša zgolj 0,0012. Obrambni izkoristek je torej skoraj optimalen in tako obrambna sposobnost ekipe izkoriščena.

### 4.3.3 Faktor izkoristka

Z verjetnostjo na intervalu [0,1] predstavljena napadalni in obrambni izkoristek ekipe lahko prikažemo še s kazalnikom oziroma faktorjem izkoristka, definiranim v [22]. Ta zavzame vrednosti na intervalu [-1,1] in nam da morda bolj intuitivno predstavo izkoristka, kjer pozitivna vrednost pomeni boljši, negativna vrednost pa slabši izkoristek, kot je pričakovan glede na vrednosti xG ekipe. Bližje kot je faktor izkoristka 1, manj je prostora za izboljšave in bolj je potencial izkoriščen; bližje kot je faktor izkoristka -1, več je prostora za izboljšave in manj je izkoriščen potencial ekipe.

Napadalni faktor izkoristka definiramo kot normalizirano (znotraj intervala [-1,1]) verjetnost, da ekipa doseže manjše ali enako število zadetkov kot dejansko:

$$\text{Napadalni faktor} = \Big(P(\text{zadetki} \leq \text{dejanski zadetki}) - 0,5\Big) \cdot 2. \tag{4.2}$$

Obrambni faktor izkoristka definiramo kot normalizirano verjetnost, da ekipa prejme večje ali enako število zadetkov kot dejansko:

$$\text{Obrambni faktor} = \Big(P(\text{prejeti zadetki} \geq \text{dejanski prejeti zadetki}) - 0,5\Big) \cdot 2. \tag{4.3}$$

Nazadnje definiramo še skupni faktor izkoristka, ki je aritmetična sredina obrambnega in napadalnega faktorja izkoristka in predstavlja izkoristek ekipe kot celote:

$$\text{Skupni faktor} = \Big(\frac{\text{napadalni faktor} + \text{obrambni faktor}}{2} - 0,5\Big) \cdot 2. \tag{4.4}$$

### 4.3.4 Primer

Za praktično interpretacijo vseh treh faktorjev izkoristka si pomagajmo s podatkovnim okvirjem tekem sezone 2017/18 angleške Premier League. S pomočjo v

poglavju 5.2 opisanih Monte Carlo simulacij smo izračunali pričakovano število skupnih doseženih in prejetih zadetkov in njihove porazdelitve za vsako izmed ekip. V kombinaciji z dejansko doseženimi zadetki smo izračunali vse tri faktorje izkoristka.

Največji oziroma najmanjši (negativen) faktor izkoristka bi načeloma pričakovali pri ekipah z največjo razliko med pričakovanimi in dejanskimi (prejetimi) zadetki. Vendar pa je pri interpretaciji potrebno upoštevati, da po absolutni vrednosti največji faktor izkoristka ne pomeni nujno največje razlike med zadetki - upoštevati je potrebno tako njihovo dejansko vrednost kot tudi vrednost xG posameznih strelov.

Na podlagi tabele 4.2 je največja pozitivna razlika med dejanskimi in pričakovanimi zadetki pri ekipah Manchester Citya in West Ham Uniteda. To se pozna tudi pri faktorju napadalnega izkoristka, ki je prikazan na sliki 4.5. Svoje priložnosti je najbolje izkoriščal West Ham, ki je ob pričakovanih 37,6 zadetkih dosegel 48 dejanskih zadetkov. Kljub najvišji razliki 13 zadetkov (106 proti 93,2) Manchester Citya, je ta na drugem mestu glede na faktor napadalnega izkoristka. Najslabšo napadalno realizacijo ima Crystal Palace, ki ima tudi največjo negativno razliko med pričakovanimi in doseženimi zadetki. Ob pričakovanih 57,9 zadetkih je dosegel le 45 zadetkov.

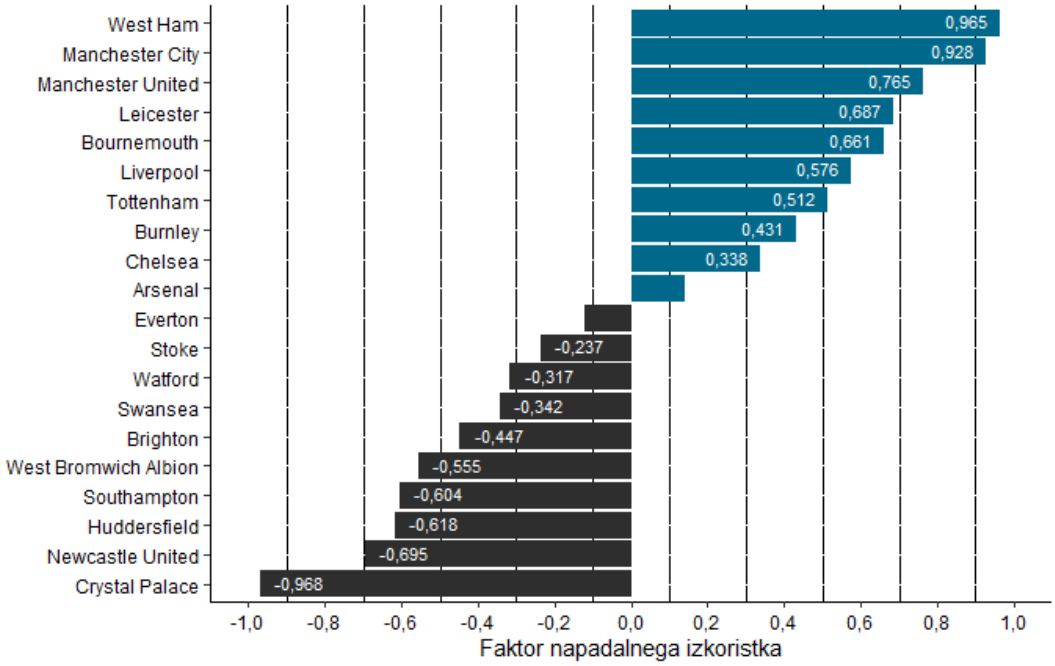
Razlika med pričakovanimi in dejanskimi prejetimi zadetki je prikazana v tabeli 4.3. Največja razlika je pri ekipah Manchester Uniteda in Burnleya, kjer bi na vrednost dejanskih 28 oziroma 39 zadetkov pričakovali 16,2 oziroma 13,6 prejetih zadetkov več. Po sliki 4.6 sta ti dve ekipi skoraj popolnoma izkoristili svoj obrambni potencial in razlika med njima je minimalna, česar pa razlika med pričakovanimi in dejanskimi zadetki ne prikazuje najbolje. Pri ekipah s slabo obrambno predstavo je precejšnja gneča, kjer je 5 ekip prejelo veliko preveč zadetkov glede na vrednosti xG. Razlika v faktorju obrambnega izkoristka pri njih je majhna, saj je razmerje med dejanskimi in pričakovanimi prejetimi zadetki podobno. Na izhodiščne vrednosti od 56 do 68 dejanskih zadetkov so te ekipe prejele od 8 do 9 zadetkov preveč.

Za konec si oglejmo še skupni faktor izkoristka, ki izkoristek ekipe predstavi celostno. Upoštevajoč tako obrambni kot napadalni izkoristek je daleč najboljša

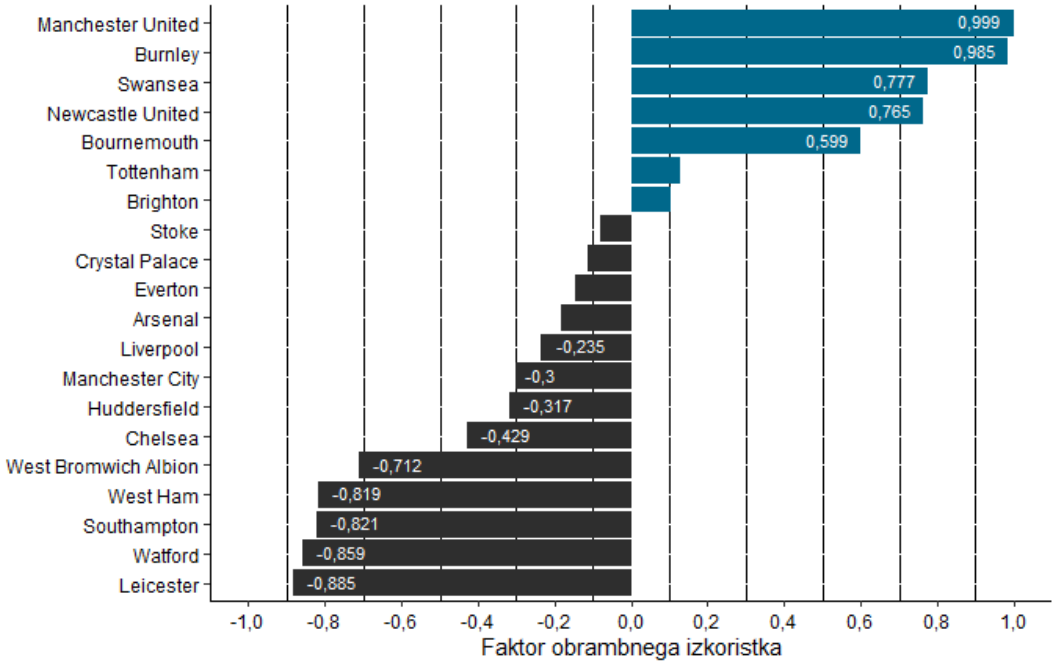
Tabela 4.2: Dejanski in pričakovani zadetki, Premier League, sezona 2017/18

| Ekipa | Dejanski zadetki | Pričakovani zadetki | Razlika v zadetkih |
|---|---|---|---|
| Manchester City | 106 | 93,150 | 12,850 |
| West Ham | 48 | 37,595 | 10,405 |
| Manchester United | 68 | 61,023 | 6,977 |
| Liverpool | 84 | 78,578 | 5,422 |
| Leicester | 56 | 50,789 | 5,211 |
| Bournemouth | 45 | 40,329 | 4,671 |
| Tottenham | 74 | 69,886 | 4,114 |
| Chelsea | 62 | 59,740 | 2,260 |
| Burnley | 36 | 33,852 | 2,148 |
| Arsenal | 74 | 73,330 | 0,670 |
| Everton | 44 | 45,225 | -1,225 |
| Stoke | 35 | 37,183 | -2,183 |
| Swansea | 28 | 30,694 | -2,694 |
| Watford | 44 | 46,869 | -2,869 |
| Brighton | 34 | 37,454 | -3,454 |
| West Bromwich Albion | 31 | 35,455 | -4,455 |
| Huddersfield | 28 | 32,484 | -4,484 |
| Southampton | 37 | 42,062 | -5,062 |
| Newcastle United | 39 | 45,126 | -6,126 |
| Crystal Palace | 45 | 57,911 | -12,911 |

ekipa Manchester Uniteda, ki je pri vrhu v napadalnem faktorju izkoristka in na samem vrhu obrambnega faktorja izkoristka (slika 4.7). Visok skupni faktor izkoristka ima tudi Burnley, predvsem zaradi visokega obrambnega izkoristka, medtem ko so pri napadalnem izkoristku zgolj nekoliko nadpovprečni. Skupno najslabši ekipi sta Southampton in West Bromwich Albion, ki sta v spodnjem delu tako ustvarjene lestvice pri napadalnem in obrambnem faktorju izkoristka. Zanimiva je ekipa Leicesterja, ki ima najslabši obrambni faktor izkoristka, vendar pa je zaradi 4. najboljšega napadalnega faktorja izkoristka njen skupni faktor izkoristka zgolj nekoliko negativen.



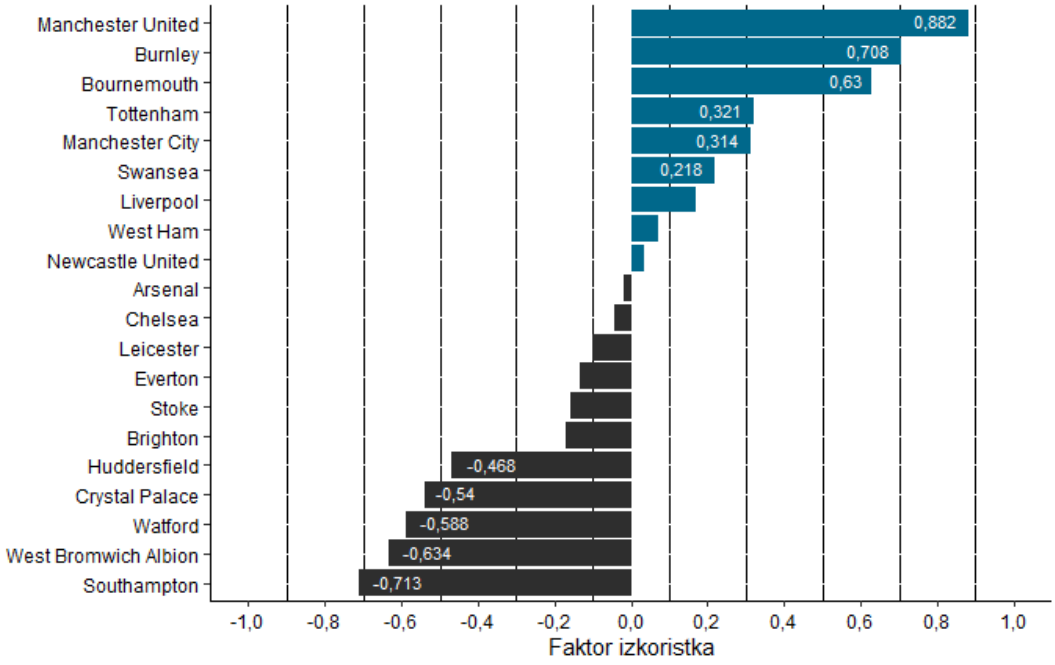

Slika 4.5: Faktor napadalnega izkoristka, Premier League, sezona 2017/18.




Slika 4.6: Faktor obrambnega izkoristka, Premier League, sezona 2017/18.

Tabela 4.3: Dejanski in pričakovani prejeti zadetki, Premier League, sezona 2017/18.

| Ekipa | Dejanski zadetki | Pričakovani zadetki | Razlika v zadetkih |
|---|---|---|---|
| Manchester United | 28 | 44,226 | 16,226 |
| Burnley | 39 | 52,576 | 13,576 |
| Swansea | 56 | 63,236 | 7,236 |
| Newcastle United | 47 | 53,475 | 6,475 |
| Bournemouth | 61 | 66,000 | 5,000 |
| Brighton | 54 | 54,467 | 0,467 |
| Tottenham | 36 | 36,354 | 0,354 |
| Crystal Palace | 55 | 53,878 | -1,122 |
| Stoke | 68 | 66,798 | -1,202 |
| Everton | 58 | 56,436 | -1,564 |
| Arsenal | 51 | 49,284 | -1,716 |
| Liverpool | 38 | 36,192 | -1,808 |
| Manchester City | 27 | 24,910 | -2,090 |
| Huddersfield | 58 | 55,050 | -2,950 |
| Chelsea | 38 | 34,754 | -3,246 |
| West Bromwich Albion | 56 | 49,342 | -6,658 |
| Southampton | 56 | 47,829 | -8,171 |
| West Ham | 68 | 59,087 | -8,913 |
| Watford | 64 | 54,770 | -9,230 |
| Leicester | 60 | 50,074 | -9,926 |




Slika 4.7: Faktor izkoristka, Premier League, sezona 2017/18.

# 5 Pričakovane točke

## 5.1 Koncept pričakovanih točk

Pričakovane točke, ki jih označimo z xP (angl. expected points), poskušajo verjetnost posameznega izida in s tem dodelitve točk na tekmi postaviti v kontekst glede na vrednosti xG posameznih strelov. Predstavljajo pričakovano število točk ekip na tekmi in povedo, kako bi se tekma odvila na dolgi rok oziroma, kako bi se točke razdelile med ekipi na podlagi tekme z identičnimi streli na gol, vendar brez prisotnosti slučaja [24]. Z xP lahko ocenimo, kako verjeten je bil dejanski rezultat in koliko je bil posledica slučaja.

V realnem svetu nogometna ekipa na določeni tekmi običajno (obstajajo sicer tudi drugačni sistemi tekmovanj) za zmago dobi 3 točke, za remi 1 točko in v primeru poraza dobi 0 točk. S pričakovanimi točkami poskušamo pravičneje razdeliti dejanske točke med obe ekipi s pomočjo mere pričakovanih zadetkov xG, s katero kvantificiramo njuno igro.

Največje število pričakovanih točk, ki jih posamezni ekipi na tekmi lahko pripišemo, je 3, kar pa je izjemno redko. Tri pričakovane točke bi pripisali ekipi, ki bi edina streljala na gol, torej bi nasprotna ekipa tekmo zaključila brez strela na gol. Nasprotni ekipi bi v tovrstni situaciji pripisali 0 pričakovanih točk. Prevladujoče ekipe običajno prejmejo med 2,1 in 2,8 pričakovanih točk, medtem ko nasprotne ekipe prejmejo med 0,1 in 0,5 pričakovanih točk [24]. Zaradi možnosti realnega remija, kjer obe ekipi dobita 1 točko, ni nujno, da je seštevek pričakovanih točk obeh ekip enak 3.

Pričakovane točke lahko uporabimo za dodelitev števila točk na eni tekmi,

ali pa za daljše obdobje ali tekmovanje. S seštevkom pričakovanih točk ekip s posameznih tekem lahko na koncu sezone, namesto realne lestvice ekip, kjer so mesta ekip določena z dejanskimi točkami, pogledamo lestvico ekip, določeno s pričakovanimi točkami in dobimo t.i. 'lestvico pravičnosti'.

## 5.2 Izračun pričakovanih točk

Osnovni izračun pričakovanih točk posamezne tekme temelji na večkratni simulaciji tekme z uporabo metode Monte Carlo. Izračun lahko posplošimo v štiri korake [24].

1. Izračunamo vrednost xG vsakega strela na tekmi.
2. Večkrat simuliramo tekmo na podlagi vrednosti xG vsakega strela pogojno na ekipo in zabeležimo končni simuliran rezultat.
3. Na podlagi simuliranih rezultatov izračunamo verjetnost zmage, remija ali poraza vsake ekipe kot deleža ustreznih rezultatov.
4. Izračunane verjetnosti uporabimo v enačbi pričakovanih vrednosti za izračun xP [25].

$$\begin{aligned}\text{xP} &= P(\text{zmaga}) \cdot (\text{št. točk za zmago}) + P(\text{remi}) \cdot (\text{št. točk za remi}) + P(\text{poraz}) \cdot (\text{št. točk za poraz}) \\ &= P(\text{zmaga}) \cdot 3 + P(\text{remi}) \cdot 1 + P(\text{poraz}) \cdot 0 \\ &= P(\text{zmaga}) \cdot 3 + P(\text{remi})\end{aligned} \tag{5.1}$$

Enkratno simulacijo tekme izvedemo tako, da na podlagi vrednosti xG posameznih strelov simuliramo vsakega izmed strelov in tako dobimo simulirano število zadetkov vsake od ekip. Enkratna simulacija strela z vrednostjo xG je realizacija Bernoullijeve slučajne spremenljivke $\pi = \text{xG}$, lahko pa si jo predstavljamo tudi kot realizacijo slučajne spremenljivke $X \sim U(0, 1)$, kjer do zadetka

pride, če je $x < \text{xG}$, sicer pa se zadetek ne zgodi. Ko so bili simulirani vsi streli znotraj tekme, kar si lahko predstavljamo kot realizacijo Poisson-binomsko porazdeljene slučajne spremenljivke $Y \sim Poiss - Bin(\text{xG}_1, \text{xG}_2, ..., \text{xG}_n)$, uporabimo seštevek pozitivnih realizacij spremenljivke oziroma zadetkov pogojno na ekipo, da izračunamo končni izid simulirane tekme in tako število točk, ki jih vsaka ekipa prejme (3, 1 ali 0).

Simulacijo tekme večkrat ponovimo in vrednost xP posamezne ekipe izračunamo prek deleža oziroma verjetnosti zmag in porazov ekipe v kombinaciji z dejanskim številom pripadajočih točk. Večje kot je število simulacij, bolj se vrednost xP približa 'pravi' vrednosti. Kot dodatno in lažje predstavljivo informacijo lahko poleg izračunane vrednosti xP prikažemo tudi verjetnost vsakega rezultata tekme, torej delež simulacij, ki se končajo z določenim rezultatom.

Monte Carlo simulacija je za večje število tekem in vrednosti xG časovno zahtevna. Algoritem Monte Carlo simulacije za izračun pričakovanega števila točk ene izmed tekem na podlagi vrednosti xG je v statističnem programu R dodan v Prilogi A. V nadaljevanju bo predstavljena dodatna metoda, ki temelji na predpostavki verjetnostne porazdelitve končnega rezultata tekme (4.1). Lahko jo uporabimo za izračun verjetnosti posameznega rezultata tekme in s tem tudi pričakovanih točk na analitičen način brez simulacije, ki pa ni tako intuitiven.

Glavna pomanjkljivost tako definirane mere pričakovanih točk je časovna komponenta rezultata, ki lahko napihne vsoto vrednosti xG in ne prikazuje realne razporeditve moči med ekipama. Potek tekme je lahko močno odvisen od slučaja, ki v xG ni upoštevan. Velja, da bo ekipa, ki je morda po slučaju, ki se ni odražal v vrednostih xG, v zaostanku in lovi zadetek, imela veliko večjo težnjo po napadu kot ekipa, ki vodi. Zato bo imela ekipa v zaostanku zaradi osredotočenosti nasprotne ekipe na obrambo večje število priložnosti (z majhnim xG) od ekipe v vodstvu, ki bo imela manjše število priložnosti (z velikim xG).

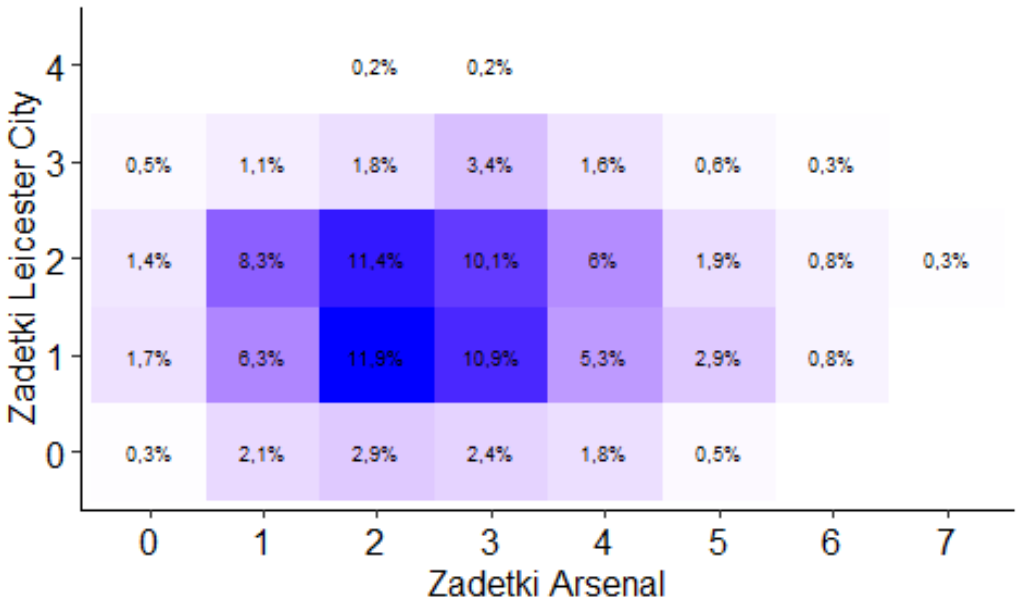
## 5.3 Primer: izračun pričakovanih točk

Opisano Monte Carlo simulacijo izračuna pričakovanih točk si oglejmo na primeru tekme Arsenala in Leicester Citya. Tekma je bila odigrana 11.8.2017 na domačem

stadionu Arsenala in se je končala z rezultatom 4:3 v korist Arsenala. Uporabili bomo znane podatke tekem Premier League sezone 2017/18 [23] z že določenimi vrednostmi xG za vsak strel na tekmi.

Vsota vrednosti xG za Arsenal je bila 2,54, za Leicester City pa 1,46. Ti dve vsoti seveda lahko zavzameta vrednosti na intervalu $[0, \infty)$. V primerjavi z dejanskim številom zadetkov (4:3) vidimo, da so igralci svoje strele zaključevali nad pričakovanji oziroma nadpovprečno.

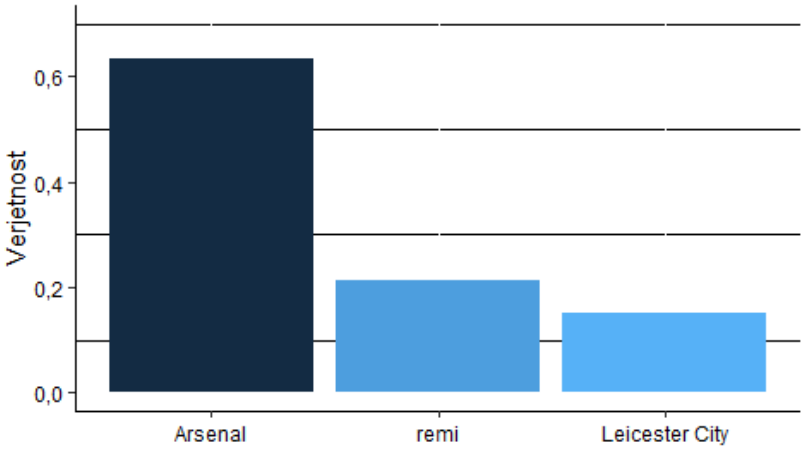
Monte Carlo simulacijo tekme smo izvedli 10.000-krat in izračunane verjetnosti rezultatov, ki so se zgodili v več kot 0,1%, so prikazane na sliki 5.1. Rezultati oziroma njihove simulirane verjetnosti so bolj naklonjene ekipi Arsenala in kot nakazujeta skupni vrednosti xG, so izračunane verjetnosti dejanskega rezultata ali višjih nizke.




Slika 5.1: Rezulati simulacije Monte Carlo (B = 10.000), končni izidi in njihove verjetnosti, Arsenal in Leicester City

Glede na dejanski rezultat (4:3), ki kaže precej, izenačeno tekmo nam relativno visoka verjetnost zmage domačega Arsenala, prikazana na sliki 5.2, pokaže, da je bila zmaga Arsenala zaslužena in razlika morda celo nekoliko nizka.

Da verjetnost zmag in remija bolje kvantificiramo, izračunajmo še pričakovane




Slika 5.2: Rezulati simulacije Monte Carlo (B = 10.000), verjetnost končnega zmagovalca, Arsenal in Leicester City

točke obeh ekip.

$$\text{xP(Arsenal)} = P(\text{zmaga}) \cdot 3 + P(\text{remi}) \cdot 1 = 0,6366 \cdot 3 + 0,2091 \cdot 1 = 2,1189$$

$$\text{xP(Leicester City)} = P(\text{zmaga}) \cdot 3 + P(\text{remi}) \cdot 1 = 0,1543 \cdot 3 + 0,2091 \cdot 1 = 0,6720$$

## 5.4 Primer: lestvica pravičnosti, Premier League, sezona 2017/18

Na podlagi pričakovanih točk bomo izračunali lestvico pravičnosti sezone 2017/18 angleške Premier League. Ponovno bomo uporabili podatkovni okvir 380 odigranih tekem sezone 2017/18 [23], kjer imamo za vsako tekmo poleg končnega rezultata podano vrednost xG vsakega strela na tekmi. V obravnavanem tekmovanju nastopa 20 ekip. Vse ekipe se med seboj pomerijo dvakrat, enkrat doma in enkrat v gosteh.

Lestvica ekip po vseh odigranih tekmah je sestavljena glede na padajoč vrstni red seštevka pridobljenih točk. V realnem svetu ekipa za zmago dobi 3 točke, za remi 1 točko in v primeru poraza ostane brez točk, pričakovane točke

pa izračunamo oziroma simuliramo na podlagi vrednosti xG, kot je predstavljeno v poglavju 5.2. Tako izračunamo dejansko in pravično lestvico ekip po koncu sezone. Po opravljeni simulaciji, kjer smo Monte Carlo simulacijo vsake tekme ponovili 10.000-krat in sešteli pričakovane in dejanske točke, za začetek primerjajmo razvrstitvi ekip na podlagi dejanskega in pričakovanega števila točk. V Premier League (2017/18) je najpomembnejših prvih 7 mest in zadnja 3 mesta. V naslednji sezoni prva 4 mesta namreč vodijo v Ligo prvakov, 5. mesto in potencialno 6. in 7. mesto vodijo v Ligo Evropa, medtem ko zadnje 3 ekipe izpadejo v nižjo ligo.

Zmagovalec tekmovanja je nesporen, to je Manchester City, ki je najboljši tako po dejanski kot pravični razvrstitvi. Tudi pri ekipah Liverpoola in Tottenhama lestvica na podlagi pričakovanega števila točk bistveno ne spremeni vtisa sezone. Chelsea bi moral glede na lestvico pravičnosti igrati v Ligi prvakov, vendar je zasedel dejansko 5. mesto in se uvrstil zgolj v ligo Evropa. Simulacija kaže, da so imeli srečo pri Manchester Unitedu, ki je končal na 2. mestu, kar je 4 mesta višje, kot jim pripisuje lestvica pravičnosti. Pri zadnjih treh ekipah velja, da je West Bromwich Albion nezasluženo izpadel v nižji rang tekmovanja, saj bi jih glede na igro pričakovali bistveno višje, medtem ko izpad Stoke Citya in Swanseaja sovpada z lestvico pravičnosti.

Razlike v pričakovani in dejanski razvrstitvi ekip (pričakovana razvrstitev - dejanska razvrstitev) lahko pokažemo še bolj pregledno na sliki 5.3. Večja kot je vrednost prikazane razlike, bolj je ekipa lahko zadovoljna z dejansko končno razvrstitvijo in več 'sreče' je ekipa imela. Na drugi strani pa manjša kot je razlika (negativna) bolj je ekipa lahko nezadovoljna s končnim izkupičkom dejanskega števila točk.

Najbolj so s končno razvrstitvijo na podlagi dejanskih predstav lahko zadovoljni pri Burnleyu, sledita Bournemouth in Manchester United. Na drugi strani sta imela najmanj sreče Southampton in West Bromwich Albion, ki bi morala biti uvrščena kar 8 oz. 7 mest boljše. Ekipe Tottenhama, Stoke Citya, Manchester Citya in Brightona so razvrščene v skladu z dejanskimi predstavami.

Bolj podroben izhodni rezultat simulacij so pričakovane točke, predstavljene v tabeli 5.2, ki jih primerjamo z realnimi točkami. Te določajo v 5.1 predstavljene

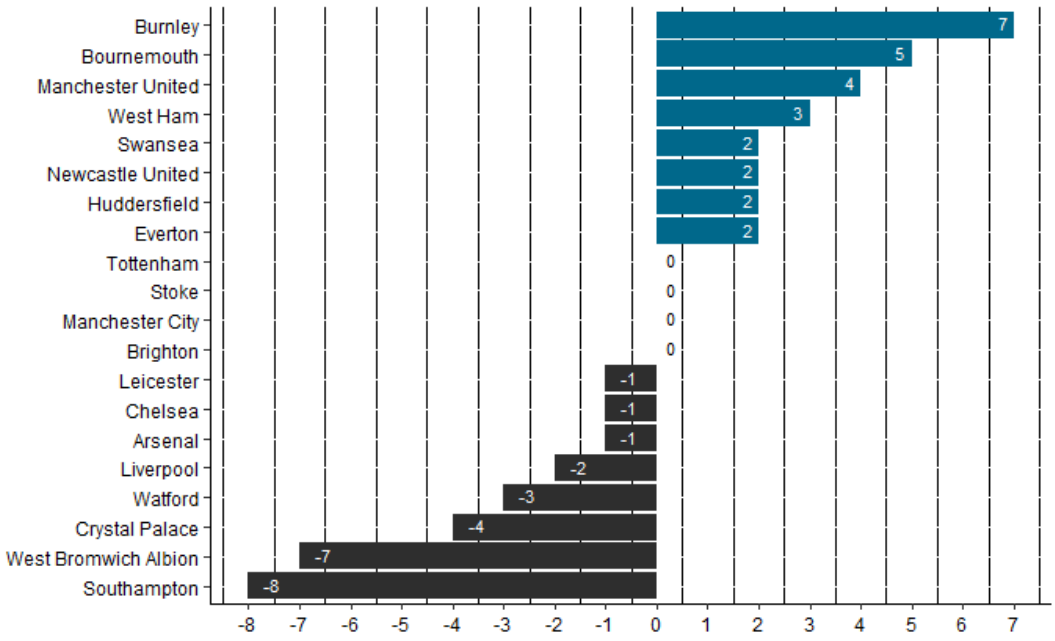
Tabela 5.1: Dejanska in pričakovana razvrstitev, Premier League 2017/18

| Ekipa | Dejanska razvrstitev | Pravična razvrstitev | Razlika v razvrstitvi |
|---|---|---|---|
| Manchester City | 1 | 1 | 0 |
| Liverpool | 4 | 2 | -2 |
| Tottenham | 3 | 3 | 0 |
| Chelsea | 5 | 4 | -1 |
| Arsenal | 6 | 5 | -1 |
| Manchester United | 2 | 6 | 4 |
| Crystal Palace | 11 | 7 | -4 |
| Leicester | 9 | 8 | -1 |
| Southampton | 17 | 9 | -8 |
| Everton | 8 | 10 | 2 |
| Watford | 14 | 11 | -3 |
| Newcastle United | 10 | 12 | 2 |
| West Bromwich Albion | 20 | 13 | -7 |
| Burnley | 7 | 14 | 7 |
| Brighton | 15 | 15 | 0 |
| West Ham | 13 | 16 | 3 |
| Bournemouth | 12 | 17 | 5 |
| Huddersfield | 16 | 18 | 2 |
| Stoke | 19 | 19 | 0 |
| Swansea | 18 | 20 | 2 |

razvrstitve ekip in tako z njimi precej sovpadajo. Razlika je zgolj v tem, da med dvema mestoma na lestvici ostaja prostor za različno število točk, kar nekoliko natančneje ovrednoti predstavo ekip, vendar pri končni oceni uspešnosti sezone ne spremeni ničesar.

Interpretacija slike 5.4 je podobna kot pri končnih razvrstitvah. Večja kot je vrednost, ki predstavlja razliko med dejanskimi in pričakovanimi točkami, bolj je ekipa lahko zadovoljna z dobljenimi točkami oziroma manj bi si jih 'pravično' zaslužili. Na drugi strani velja, da bolj negativna kot je vrednost, več točk bi si ekipa glede na svoje igre zaslužila.

Najbolj zadovoljni glede števila osvojenih točk so lahko pri Manchester Unitedu, ki je bil na podlagi razlike končne razvrstitve šele na 3. mestu. Manchester City, ki je sezono končal na prvem mestu tako na dejanski lestvici kot lestvici pravičnosti, morda ni tako dominiral, saj bi si glede na pričakovane točke zaslužil

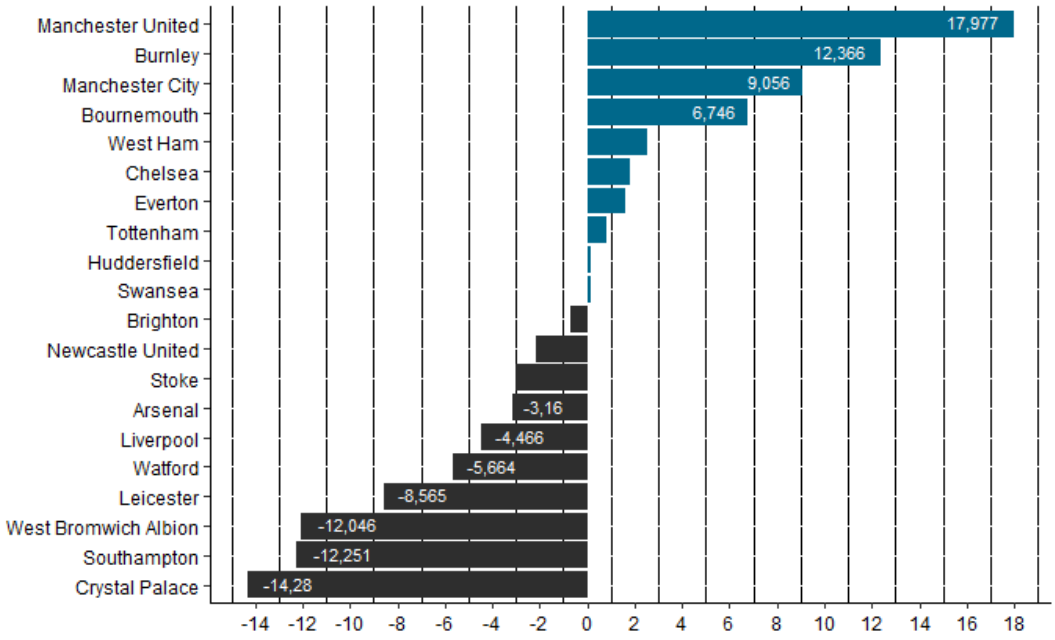



Slika 5.3: Razlika med pravično in dejansko razvrstitvijo, Premier League, sezona 2017/18

9 točk manj. Zanimiv je morda še primer Chelseaja, kjer je število pričakovanih točk manjše od dejanskega, obenem pa mu lestvica pravičnosti narekuje višjo uvrstitev od dejanske.

Tabela 5.2: Dejanske in pričakovane točke, Premier League 2017/18

| Ekipa | Dejanske točke | Pričakovane točke | Razlika v točkah |
|---|---|---|---|
| Manchester City | 100 | 90,944 | 9,056 |
| Liverpool | 75 | 79,466 | -4,466 |
| Tottenham | 77 | 76,156 | 0,844 |
| Chelsea | 70 | 68,215 | 1,785 |
| Arsenal | 63 | 66,160 | -3,160 |
| Manchester United | 81 | 63,023 | 17,977 |
| Crystal Palace | 44 | 58,280 | -14,280 |
| Leicester | 47 | 55,565 | -8,565 |
| Southampton | 36 | 48,251 | -12,251 |
| Everton | 49 | 47,394 | 1,606 |
| Watford | 41 | 46,664 | -5,664 |
| Newcastle United | 44 | 46,163 | -2,163 |
| West Bromwich Albion | 31 | 43,046 | -12,046 |
| Burnley | 54 | 41,634 | 12,366 |
| Brighton | 40 | 40,700 | -0,700 |
| West Ham | 42 | 39,444 | 2,556 |
| Bournemouth | 44 | 37,254 | 6,746 |
| Huddersfield | 37 | 36,849 | 0,151 |
| Stoke | 33 | 35,994 | -2,994 |
| Swansea | 33 | 32,873 | 0,127 |




Slika 5.4: Razlika med dejanskimi in pričakovanimi točkami, Premier League, sezona 2017/18

# 6 Napovedovanje rezultatov nogometnih tekem

Vse od pojava prvih stavnic se športni navdušenci trudijo čim bolje napovedati rezultate. Industrija športnih stav je ena izmed panog, na katero je pomembno vplival razvoj podatkovne analitike.

Napovedovanje izidov nogometnih tekem je mogoče povzeti v dva povezana trenda: izkoriščanje razpoložljivih podatkov ter razvoj novih statistik in računskih metod. Ena izmed najbolj obetavnih statistik, ki bi v svetu z naraščajočim vplivom podatkovne analitike lahko omogočila boljše in naprednejše napovedovanje nogometnih izidov, je obravnavana statistika pričakovanih zadetkov xG. Nove metode so uporabne predvsem za analizo ekip na podlagi njihovih predstav in ne zgolj končnih rezultatov tekem ali bolj osnovnih statistik.

Mero pričakovanih zadetkov bomo poskusili vključiti v uveljavljene modele za napovedovanje nogometnih izidov in jih tako izboljšati. Osnovni model za napovedovanje rezultatov nogometnih tekem je razvil M.J. Maher, kasneje pa sta njegov model nadgradila Mark J. Dixon in Stuart G. Coles (model Dixon-Coles) [26]. Osnovni model predpostavlja, da je število zadetkov, ki jih ekipa na nogometni tekmi doseže, mogoče modelirati z uporabo Poissonove porazdelitve. Ob predpostavki, da so zadetki ene in druge ekipe neodvisni (kar ne drži popolnoma), izid nogometne tekme lahko modeliramo z uporabo dveh neodvisnih Poissonovo porazdeljenih spremenljivk. Dixon in Coles sta osnovni model razširila s prilagoditvijo Poissonove porazdelitve, tako da sta 'odpravila' pomanjkljivosti, kot sta premajhna verjetnost nizkih rezultatov in časovna odvisnost tekem.

Osnovni model in predvsem njegova nadgradnja, model Dixon-Coles, sta

znana kot razmeroma dobra modela za napovedovanje izidov nogometnih tekem. Seveda to nista edina modela, obstajajo tudi drugi, razviti znotraj različnih podjetij, ki pa so strogo tajni, saj jih podjetja uporabljajo za ustvarjanje dobička.

Na začetku bomo predstavili najenostavnejši model za izračun pričakovanega števila zadetkov in ga nadgradili z mero xG. Potem bomo predstavili model Dixon-Coles, ki pričakovano število zadetkov oceni na nekoliko drugačen način, ter tudi tega nadgradili z mero xG. Na koncu bomo na večkrat uporabljenem podatkovnem okviru nogometnih tekem Premier League 2017/18 primerjali kakovost napovedi predstavljenih modelov.

## 6.1 Poissonova porazdelitev

Poissonova porazdelitev je diskretna porazdelitev, ki lahko zavzame vsa nenegativna cela števila $(0, 1, 2....)$, katerih verjetnosti so odvisne od parametra $\lambda > 0$. S Poissonovo porazdelitvijo modeliramo število neodvisnih dogodkov, ki se zgodijo v določenem intervalu. Predpostavka modeliranja s Poissonovo porazdelitvijo je, (a) da so dogodki neodvisni, torej pojav prejšnjega dogodka ne vpliva na verjetnost, da se bo zgodil naslednji dogodek, in (b) da je povprečni čas med dogodki konstanten.

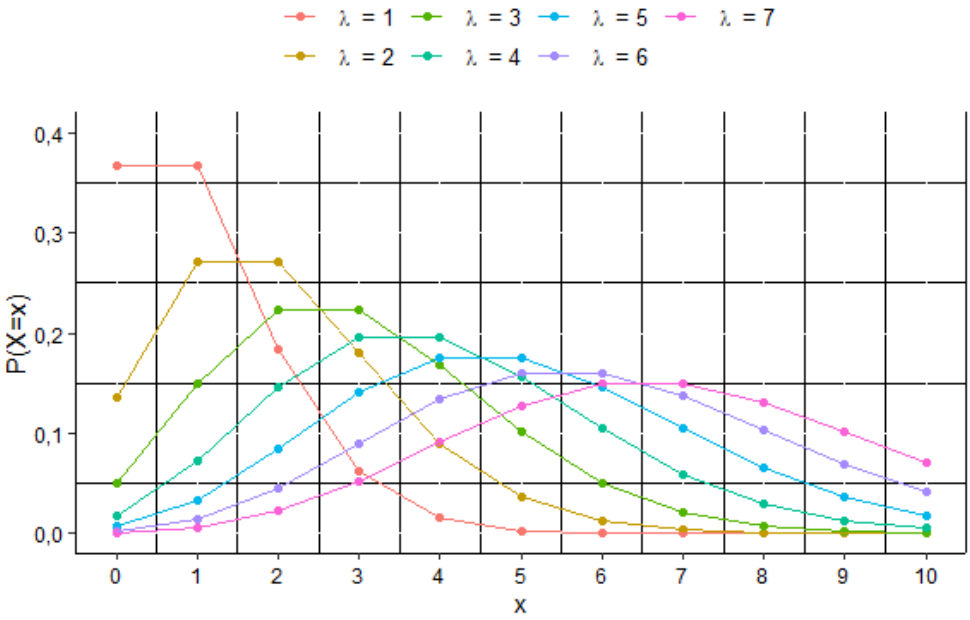
Za slučajno spremenljivko $X \sim \text{Pois}(\lambda)$ velja:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{6.1}$$

kjer je $\lambda$ povprečno število dogodkov v določenem intervalu tj. $E[X] = \lambda$ in $x$ število dogodkov katerega verjetnost iščemo.

Na sliki 6.1 so prikazane verjetnosti, da slučajna spremenljivka $X \sim \text{Pois}(\lambda)$ zavzame vrednosti $\{0, 1, 2, ..., 10\}$ za $\lambda \in \{0, 1, 2, ..., 7\}$.

Predpostavimo, da zgornji predpostavki ustrezata nogometni igri, kjer so dogodki, ki jih opazujemo zadetki. Poissonovo porazdelitev bomo uporabili za izračun verjetnosti določenega števila zadetkov, ki jih posamezna ekipa doseže,




Slika 6.1: Verjetnosti Poissonovo porazdeljene slučajne spremenljivke

in s tem končnega rezultata. V ta namen moramo poiskati parameter $\lambda$, ki predstavlja povprečno oziroma pričakovano število zadetkov, ki jih ekipa doseže.

## 6.2 Osnovni Poissonov model

Najenostavnejši izračun parametra $\lambda$, na podlagi katerega preko Poissonove porazdelitve (6.1) ocenimo verjetnost posameznega števila zadetkov, ki jih ekipa na tekmi doseže, je [27]:

$$\lambda = \frac{\text{število zadetkov}}{\text{število odigranih tekem}}. \tag{6.2}$$

Pri tem določimo relevanten časovni interval upoštevanih tekem. Ta mora biti čim bližje datumu tekme, hkrati pa moramo zagotoviti dovolj širok nabor tekem.

Tovrsten izračun parametra $\lambda$ ni najboljši, saj ne upošteva niti najosnovnejših relevantnih dejavnikov. Ne upošteva namreč, ali ekipa igra doma ali v gosteh.

Prednost domačega igrišča se v splošnem odraža v večjem številu doseženih zadetkov ekip doma v primerjavi z gostovanjem. Ne upošteva niti lastnosti nasprotne ekipe, na primer števila zadetkov, ki jih nasprotna ekipa v povprečju prejme. Upoštevali bi lahko tudi relevantnost v izračun vključenih tekem, ki se s časovno oddaljenostjo zmanjšuje, in poljubno drugih dejavnikov, ki pa jih je v izračun težje vpeljati.

Da ocenimo ustreznejši parameter $\lambda$, kjer bomo upoštevali tudi nasprotno ekipo, izračunamo napadalno in obrambno moč ekip, katerih rezultat tekme želimo napovedati.

Napadalna moč ekipe je povprečno število doseženih zadetkov ekipe v primerjavi z ligaškim povprečjem. Večje kot je število zadetkov ekipe, večja je njena napadalna moč. Izračunamo jo kot [28]:

$$\text{Napadalna moč} = \frac{\text{povprečno število zadetkov ekipe}}{\text{ligaško povprečno število zadetkov}}. \tag{6.3}$$

Podobno je obrambna moč ekipe povprečno število prejetih zadetkov ekipe v primerjavi z ligaškim povprečjem. Manjše kot je število prejetih zadetkov, manjša je njena obrambna moč, kar je v primeru obrambe boljše. Izračunamo jo kot [28]:

$$\text{Obrambna moč} = \frac{\text{povprečno število prejetih zadetkov ekipe}}{\text{ligaško povprečno število prejetih zadetkov}}. \tag{6.4}$$

Povprečno število danih oziroma prejetih zadetkov ekipe preprosto izračunamo kot razmerje med številom vseh danih oziroma prejetih zadetkov ter številom upoštevanih tekem. Podobno je pri ligaškem povprečju, kjer seštejemo zadetke vseh ekip in jih delimo s številom vseh tekem.

Pričakovano število zadetkov ekipe, katere tekma je predmet zanimanja, potem lahko izračunamo kot [28]:

$$\begin{aligned} E[\text{ zadetki ekipa }] = \lambda = \text{napadalna moč ekipe}\times \\ \times \text{ obrambna moč nasprotne ekipe}\times \\ \times \text{ ligaško povprečno število zadetkov.} \end{aligned} \tag{6.5}$$

Če z $\alpha_i$ označimo napadalno moč ekipe, z $\beta_j$ obrambno moč nasprotne ekipe in z $\mu$ ligaško povprečno število zadetkov, zgornjo enačbo lahko zapišemo kot:

$$E[\text{ zadetki ekipa i }] = \lambda_i = \alpha_i \cdot \beta_j \cdot \mu. \tag{6.6}$$

Za nadaljnjo izboljšavo $\lambda$ v modelu upoštevamo tudi prednost domačega igrišča. V tem primeru napadalno in obrambno moč izračunamo pogojno na indikator domačega igrišča. Tako izračunamo napadalno moč doma, kjer upoštevamo le število zadetkov na domačih tekmah, ter napadalno moč v gosteh, kjer upoštevamo zgolj število zadetkov na gostujočih tekmah. Podobno izračunamo tudi obrambno moč doma in obrambno moč v gosteh.

Podobno kot zgoraj, z $\alpha_{Hi}$ in $\alpha_{Ai}$ označimo domačo in gostujočo napadalno moč ekipe in z $\beta_{Hj}$ in $\beta_{Aj}$ označimo domačo in gostujočo obrambno moč nasprotne ekipe. Z $\mu_H$ in $\mu_A$ označimo še povprečno število zadetkov domačih in gostujočih ekip.

Pričakovano število zadetkov domače ekipe potem izračunamo kot

$$E[\text{ zadetki domača ekipa i proti ekipi j }] = \lambda_i = \alpha_{Hi} \cdot \beta_{Aj} \cdot \mu_H \tag{6.7}$$

in pričakovano število zadetkov gostujoče ekipe izračunamo kot

$$E[\text{ zadetki gostujoča ekipa j proti ekipi i }] = \lambda_j = \alpha_{Aj} \cdot \beta_{Hi} \cdot \mu_A. \tag{6.8}$$

Drugi pristop upoštevanja prednosti domačega igrišča, ki se ga v osnovnem modelu sicer ne bomo poslužili, je prek korekcijskega faktorja domačega igrišča

$\gamma$. Ta je enostaven multiplikativni faktor pri verjetnosti doseganja zadetkov na domačem igrišču in ga izračunamo kot: $\gamma = \frac{\mu_H}{\mu_A}$. V tem primeru napadalne in obrambne moči ekip ne ocenjujemo ločeno glede na domače igrišče. Ponovno z $\alpha_i$ označimo napadalno moč ekipe, z $\beta_j$ označimo obrambno moč nasprotne ekipe in z $\mu$ označimo povprečno število zadetkov vseh ekip.

Pričakovano število zadetkov domače oziroma gostujoče ekipe izračunamo kot:

$$E[\text{ zadetki domača ekipa i proti ekipi j }] = \alpha_i \cdot \beta_j \cdot \mu \cdot \gamma \tag{6.9}$$

$$E[\text{ zadetki gostujoča ekipa j proti ekipi i }] = \alpha_j \cdot \beta_i \cdot \mu. \tag{6.10}$$

Za izračun verjetnosti posameznega rezultata upoštevamo izračunana parametra $\lambda$ domače in gostujoče ekipe ter verjetnost posameznega rezultata izračunamo iz dveh neodvisnih Poissonovih porazdelitev kot produkt verjetnosti ustreznih vrednosti zadetkov domače in gostujoče ekipe.

V skladu z zgornjim, indeks $i$ označuje domačo ekipo in $j$ gostujočo ekipo. Z $X_{ij}$ označimo število zadetkov domače ekipe in z $Y_{ij}$ število zadetkov gostujoče ekipe. Ustrezno napadalno in obrambno moč ekip zapišemo kot vektorja $\boldsymbol{\alpha} = (\alpha_{Hi}, \alpha_{Aj})$ in $\boldsymbol{\beta} = (\beta_{Hi}, \beta_{Aj})$.

Verjetnost posameznega rezultata potem izračunamo kot [26]:

$$P(X_{ij} = x, Y_{ij} = y|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{Pois}(x|\alpha_{Hi}, \beta_{Aj}) \cdot \text{Pois}(y|\alpha_{Aj}, \beta_{Hi}). \tag{6.11}$$

Če upoštevamo korekcijski faktor prednosti domačega igrišča $\gamma$, $\boldsymbol{\alpha} = (\alpha_i, \alpha_j)$ predstavlja napadalni moči ekip ter $\boldsymbol{\beta} = (\beta_i, \beta_j)$ predstavlja obrambni moči ekip, verjetnost posameznega rezultata izračunamo kot [26]:

$$P(X_{ij} = x, Y_{ij} = y|\alpha, \beta, \gamma) = \text{Pois}(x|\alpha_i, \beta_j, \gamma) \cdot \text{Pois}(y|\alpha_j, \beta_i). \tag{6.12}$$

### 6.2.1 Primer napovedovanja izida tekme

Na podlagi vseh 380 odigranih tekem španske La Lige v sezoni 2020/21 napovejmo rezultat tekme med domačim Real Madridom in gostujočo Valencio. Skupno število vseh ekip v ligi je 20 in vsaka ekipa z vsako igra enkrat doma in enkrat v gosteh, skupno torej ena ekipa odigra 38 tekem, od katerih je 19 domačih in 19 gostujočih. Domačo in gostujočo napadalno in obrambno moč ekip bomo izračunali ločeno in ne bomo uporabili korekcijskega faktorja $\gamma$. Podatke smo pridobili s spletišča FBref [10].

Izračunana napadalna in obrambna moč ekipe na podlagi celotne sezone je prikazana v tabeli 6.1.

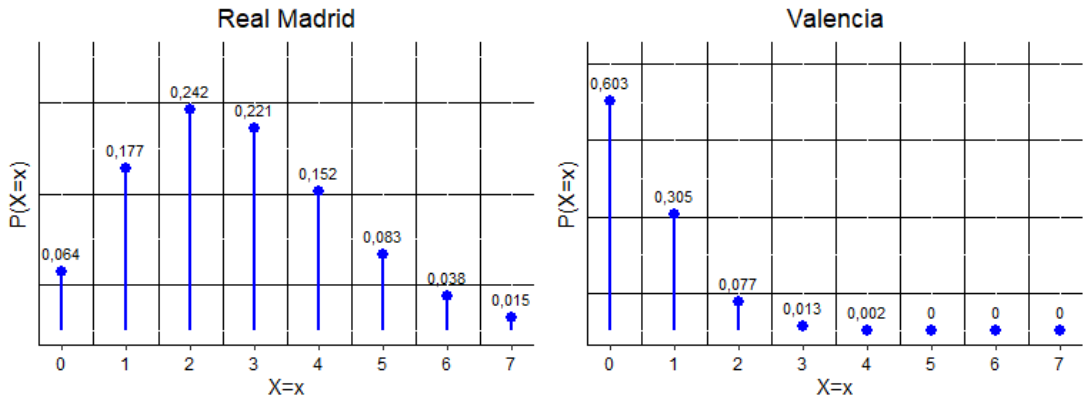
Tabela 6.1: Osnovna Poissonova napoved, obrambna in napadalna moč, Real Madrid in Valencia, La Liga, sezona 2020/21.

| | Doseženi zadetki | Prejeti zadetki |
|---|---|---|
| Domače ekipe | 1,368 | 1,139 |
| Gostujoče ekipe | 1,139 | 1,368 |
| Real Madrid doma | 1,737 | 0,684 |
| Real Madrid moč doma | 1,269 | 0,600 |
| Valencia v gosteh | 0,842 | 1,579 |
| Valencia moč v gosteh | 0,739 | 1,154 |

Pričakovano število zadetkov Real Madrida je
$\lambda_H = 1,269 \cdot 1,579 \cdot 1,368 = 2,741$. Podobno je pričakovano število zadetkov Valencie $\lambda_A = 0,739 \cdot 0,600 \cdot 1,139 = 0,505$. Verjetnosti posameznega števila zadetkov, dobljene prek Poissonove porazdelitve, so prikazane na sliki 6.2.

Verjetnost posameznega rezultata izračunamo kot produkt verjetnosti števila zadetkov ene in druge ekipe, ki ustreza opazovanem izidu. Sedem najbolj verjetnih rezultatov je predstavljenih na sliki 6.3.

Za izračun verjetnosti zmage domače oziroma gostujoče ekipe ali remija je potrebno zgolj sešteti verjetnosti ustreznih rezultatov. Te verjetnosti so: zmaga Real Madrid 83,2%, remi 11,5% in zmaga Valencia 4,6%.




Slika 6.2: Osnovni Poissonov model, verjetnost zadetkov, Real Madrid in Valencia, La Liga, sezona 2020/21

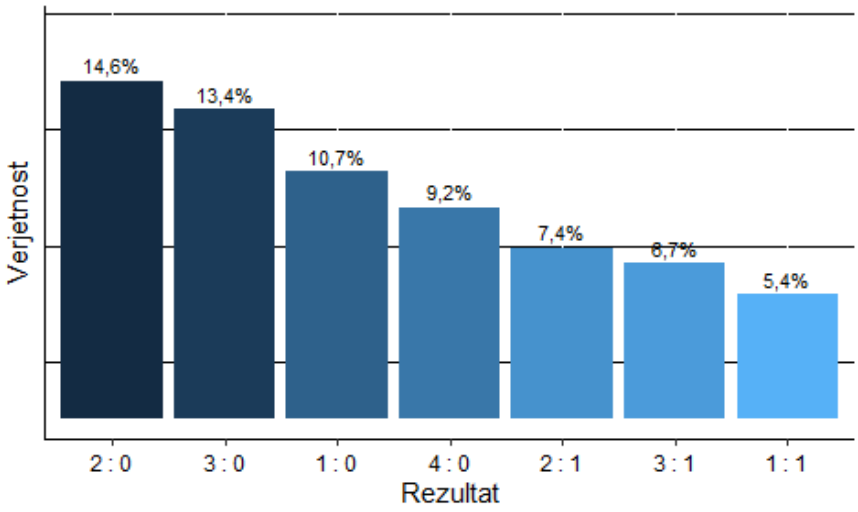
### 6.2.2 Osnovni Poissonov model upoštevajoč kumulativne xG

Predstavljeni osnovni model za izračun napadalne in obrambne moči in s tem parametra $\lambda$ upošteva dejanske rezultate nogometnih tekem. Videli smo, da ti zanemarjajo precejšen del informacije o poteku tekme in tako kakovosti ekipe.

Da bi tudi pri napovedovanju rezultatov zmanjšali vpliv slučaja, lahko pri izračunu $\lambda$ namesto dejanskih zadetkov upoštevamo pričakovane zadetke, torej mero xG. Tako bi morda bolje ocenili napadalno in obrambno moč ekipe, saj ne pričakujemo, da bi se morebiten slučaj zadetkov ponovil na napovedovani tekmi. V osnovnem modelu z xG upoštevamo kumulativno vrednost xG posamezne tekme, torej vsoto vrednosti xG vseh strelov na tekmi. Naj dodamo, da smo v razdelkih 4.1 in 4.2 videli, da tudi to ni najbolj optimalno.

Izračun obrambne in napadalne moči ekip je enak opisanemu v razdelku 6.2 z razliko, da namesto dejanskih zadetkov upoštevamo vsoto posameznih xG:

$$\text{Napadalna moč xG} = \frac{\text{povprečna kumulativna vrednost xG ekipe}}{\text{ligaška povprečna kumulativna vrednost xG}}, \tag{6.13}$$




Slika 6.3: Osnovni Poissonov model, verjetnost končnega izida, Real Madrid in Valencia, La Liga, sezona 2020/21

$$\text{Obrambna moč xG} = \frac{\text{povprečna kumulativna vrednost prejetih xG ekipe}}{\text{ligaška povprečna kumulativna vrednost prejetih xG}}. \tag{6.14}$$

Ponovno lahko upoštevamo tudi prednost domačega igrišča in napadalno in obrambno moč ekipe izračunamo pogojno na indikator domačega igrišča. Z $\alpha_{Hi}^{\text{xG}}$ in $\beta_{Hi}^{\text{xG}}$ označimo domačo napadalno oziroma obrambno moč in z $\alpha_{Aj}^{\text{xG}}$ in $\beta_{Aj}^{\text{xG}}$ označimo gostujočo napadalno oziroma obrambno moč. Z $\mu_H^{\text{xG}}$ označimo še povprečno kumulativno vrednost xG domačih ekip in z $\mu_A^{\text{xG}}$ povprečno kumulativno vrednost xG gostujočih ekip.

Parametra $\lambda$ domače in gostujoče ekipe potem izračunamo kot:

$$E_{\text{xG}}[\text{ zadetki domača ekipa i proti ekipi j }] = \lambda_i^{\text{xG}} = \alpha_{Hi}^{\text{xG}} \cdot \beta_{Aj}^{\text{xG}} \cdot \mu_H^{\text{xG}}, \tag{6.15}$$

$$E_{\text{xG}}[\text{ zadetki gostujoča ekipa j proti ekipi i }] = \lambda_j^{\text{xG}} = \alpha_{Aj}^{\text{xG}} \cdot \beta_{Hi}^{\text{xG}} \cdot \mu_A^{\text{xG}}. \tag{6.16}$$

## 6.3 Model Dixon-Coles

Leta 1995 sta Mark Dixon in Stuart Coles posodobila osnovni Poissonov model napovedovanja rezultatov nogometnih tekem [26]. Model še vedno sledi zamisli, da zadetki pri nogometu sledijo Poissonovi porazdelitvi, kjer pričakovano število zadetkov izračunamo z uporabo parametrov napadalne moči $\boldsymbol{\alpha} = (\alpha_i, \alpha_j)$ in obrambne moči $\boldsymbol{\beta} = (\beta_i, \beta_j)$, ki še vedno temeljita na dejanskem številu zadetkov, ter korekcijskega faktorja prednosti domačega igrišča $\gamma$:

$$\begin{aligned} P(X_{ij} = x, Y_{ij} = y | \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) &= \text{Pois}(x | \alpha_i, \beta_j, \gamma) \cdot \text{Pois}(y | \alpha_j, \beta_i) \\ &= \frac{e^{-\lambda_i} \lambda_i^x}{x!} \cdot \frac{e^{-\lambda_j} \lambda_j^y}{y!}, \end{aligned} \tag{6.17}$$

kjer $\lambda_i = \alpha_i \cdot \beta_j \cdot \gamma$ označuje pričakovano število zadetkov domače ekipe in $\lambda_j = \alpha_j \cdot \beta_i$ označuje pričakovano število zadetkov gostujoče ekipe. V primerjavi z osnovnim modelom v enačbah za izračun $\lambda_i$ in $\lambda_j$ ni parametra $\mu$, ki označuje povprečno število zadetkov vseh ekip. V modelu Dixon-Coles namreč napadalne in obrambne moči ekip ocenjujemo skozi funkcijo največjega verjetja in ne ročno. V tem primeru je parameter $\mu$ že vključen v $\alpha$ oziroma $\beta$. Funkcijo verjetja zapišemo kot [29]:

$$\begin{aligned} L(\alpha_i, \beta_i, \gamma;\ i = 1, ..., n) &= \prod_{k=1}^{N} P(X_{ik,jk} = x_k, Y_{ik,jk} = y_k) \\ &= \prod_{k=1}^{N} \frac{e^{-\lambda_{ik}} {\lambda_{ik}}^{x_k}}{x_k!} \frac{e^{-\lambda_{jk}} {\lambda_{jk}}^{y_k}}{y_k!} \\ &\propto \prod_{k=1}^{N} e^{-\lambda_{ik}} {\lambda_{ik}}^{x_k} e^{-\lambda_{jk}} {\lambda_{jk}}^{y_k}, \end{aligned} \tag{6.18}$$

kjer $N$ predstavlja število vseh odigranih tekem in $k$ točno določeno tekmo. Funkcijo verjetja $L$ maksimiziramo tako, da poiščemo vse parcialne odvode parametrov, ki jih ocenjujemo, in jih enačimo z 0. Za matematično udobje lahko namesto monotono naraščajoče funkcije verjetja $L$ maksimiziramo njen logaritem $ln(L) = LL$ in tako produkte prevedemo v vsote:

$$LL(\alpha_i, \beta_i, \gamma;\ i = 1, ..., n) \propto \sum_{k=1}^{N} \ln(-\lambda_{ik} + x_k ln(\lambda_{ik}) - \lambda_{jk} + y_k \ln(\lambda_{jk})). \quad (6.19)$$

Pri iskanju parametrov, ki maksimizirajo funkcijo verjetja, določimo pogoja, da sta povprečna napadalna in obrambna moč ekipe v tekmovanju enaka 1, torej $\frac{1}{n}\sum_{i=1}^{n}\alpha_i = 1$ in $\frac{1}{n}\sum_{i=1}^{n}\beta_i = 1$.

Z numerično optimizacijo funkcije verjetja za upoštevan nabor tekem dobimo ocene napadalne moči $\alpha_i$ in obrambne moči $\beta_i$ vsake od upoštevanih ekip. Na podlagi teh vrednosti lahko po osnovnem Poissonovem modelu, ki je opisan v razdelku 6.2, z enačbama (6.9) in (6.10) izračunamo pričakovano število zadetkov vsake izmed ekip za poljubno izbrano tekmo.

Osnovnemu Poissonovem modelu sta Dixon in Coles dodala dve novosti. Prva je parameter $\rho$, ki rahlo spremeni verjetnost nizkega števila zadetkov, drugi pa je dodatek časovne komponente, s katero pri ocenjevanju parametrov nedavnim tekmam dodelimo večjo težo kot časovno bolj oddaljenim. Njuni ideji bomo kratko predstavili, vendar se v podrobnosti ne bomo spuščali. Ideji bomo uporabili za vključitev vrednosti xG posameznih strelov v izračun parametra $\lambda$ oziroma napadalnih in obrambnih moči ekip.

Na podlagi zgodovinskih rezultatov je bilo ugotovljeno, da uporaba osnovne Poissonove porazdelitve pri napovedovanju rezultatov nogometnih tekem za nizke izide ni najboljša. Osnovni Poissonov model namreč precenjuje verjetnost izenačenega izida brez zadetkov (0:0) ter podcenjuje remi z enim zadetkom na vsaki strani (1:1). Nekoliko podcenjuje tudi minimalno zmago domače ali gostujoče ekipe (1:0 in 0:1). Da bi popravila verjetnost nizkih izidov, sta Dixon in Coles v model uvedla parameter $\rho$, ki minimalno prilagodi verjetnost posameznega števila zadetkov in ga ocenimo s funkcijo največjega verjetja [26, 29].

Druga in v našem primeru zanimivejša novost modela Dixon-Coles je dodatek uteži tekem v izračunu $\alpha$ in $\beta$. Ker parametre $\alpha_i, \beta_i$ in $\gamma$ (in $\rho$) ocenjujemo z maksimizacijo funkcije največjega verjetja, lahko novejšim tekmam dodelimo večjo težo kot preteklim. To storimo tako, da vrednosti vsake tekme pomnožimo z monotono padajočo funkcijo časa v odvisnosti od parametra, ki predstavlja število

dni, ki je minilo od odigrane tekme do tekme, katere rezultat napovedujemo [26, 29].

Za implementacijo modela Dixon-Coles in izračun parametrov $\alpha, \beta, \gamma, \rho$ z upoštevanjem uteži imamo na voljo R-ov paket *regista* za statistiko in modeliranje v nogometu. Njegovo implementacijo je omogočil Ben Torvaney [23], ki ima tudi pregleden spletni dnevnik z njegovo uporabo [30].

### 6.3.1 Vključitev xG v model Dixon-Coles

Pri izračunu parametrov napadalnih in obrambnih moči ni nujno, da tekme utežujemo glede na časovno komponento. Uporabo uteži lahko razširimo na različne faktorje, kjer pa je nekatere v funkcijo verjetja in njeno optimizacijo težje vključiti kot druge. V našem primeru bomo zaradi boljše ocene parametrov $\alpha$ in $\beta$ ter s tem $\lambda$ namesto časovne komponente upoštevali vrednosti xG posameznih strelov. Možnost implementacije xG v model Dixon-Coles in izračun parametrov je v paketu *regista* že predvidena, zato bo pojasnjeno zgolj njeno matematično ozadje, ki drugje kot v sami R-ovi funkciji ni zapisano.

Za pravilno implementacijo informacije xG posameznih strelov moramo vrednosti xG pretvoriti v celoštevilske zadetke in kot uteži upoštevati njihove verjetnosti. Za izračun verjetnosti posameznega rezultata na podlagi posameznih xG lahko uporabimo Monte Carlo simulacijo iz razdelka 5.2 ali hitrejšo, a ekvivalentno Poisson-binomsko porazdelitev (4.1), s katero analitično ocenimo verjetnosti rezultatov.

Dobljene simulirane rezultate nato v funkciji verjetnosti upoštevamo kot dejanske, njihove verjetnosti pa uporabimo kot uteži. Parametri modela ostajajo isti $(\alpha_i, \beta_i, \gamma)$ in podobno kot prej jih ocenimo z maksimizacijo funkcije verjetja. Znotraj vsake tekme, ki jo določa parameter $k$, moramo uvesti dodatno vsoto, ki prikazuje rezultate, pomnožene z njihovimi verjetnostmi. Pri rezultatih se lahko omejimo do števila zadetkov 7, višje število je zelo redko.

Osnovno enačbo (6.17), ki opisuje dejanski rezultat tekme, v naslednjem koraku upoštevamo na ravni verjetnosti različnih možnih rezultatov, izračunanih z

uporabo vrednosti xG:

$$\begin{aligned}&\sum_{(x_r,y_r)\in\{0,1,...,7\}^2} p_r \cdot P(X_{ij}=x_r, Y_{ij}=y_r|\boldsymbol{\alpha},\boldsymbol{\beta},\gamma) = \\ &= \sum_{(x_r,y_r)\in\{0,1,...,7\}^2} p_r \cdot \frac{e^{-\lambda_i}\lambda_i^{x_r}}{x_r!} \cdot \frac{e^{-\lambda_j}\lambda_j^{y_r}}{y_r!},\end{aligned} \tag{6.20}$$

kjer je $p_r$ izračunana verjetnost rezultata $(x_r, y_r)$; $(x_r, y_r) \in \{0, 1, ..., 7\}^2$.

Funkcijo verjetja potem zapišemo kot:

$$L(\alpha_i, \beta_i, \gamma;\ i = 1, ..., n) \propto \prod_{k=1}^{N} \sum_{(x_r,y_r)\in\{0,1,...,7\}^2} p_{rk} \cdot e^{-\lambda_{ik}} {\lambda_{ik}}^{x_{rk}} e^{-\lambda_{jk}} {\lambda_{jk}}^{y_{rk}} \tag{6.21}$$

in logaritem funkcije verjetja zapišemo kot:

$$LL(\alpha_i, \beta_i, \gamma;\ i = 1, ..., n) \propto \sum_{k=1}^{N} \log\Big( \sum_{(x_r,y_r)\in\{0,1,...,7\}^2} p_{rk} \cdot e^{-\lambda_{ik}} {\lambda_{ik}}^{x_{rk}} e^{-\lambda_{jk}} {\lambda_{jk}}^{y_{rk}} \Big). \tag{6.22}$$

Ponovno z numerično optimizacijo funkcije verjetja za upoštevani nabor tekem dobimo ocene $\gamma$, napadalne moči $\alpha_i$ in obrambne moči $\beta_i$ vsake od ekip ter na podlagi teh vrednosti po principu, enakemu osnovnem Poissonovem modelu, z enačbama (6.9) in (6.10) izračunamo pričakovano število zadetkov vsake izmed ekip za poljubno izbrano tekmo.

### 6.3.2 Primer

Za primer implementacije in primerjave rezultatov osnovnega in z xG uteženega modela Dixon Coles bomo uporabili podatkovni okvir 380 tekem Premier League 2017/18 [23]. Na podlagi vseh odigranih tekem bomo skušali napovedati izid tekme med domačim Liverpoolom in gostujočim Southamptonom.

Vrednosti parametrov smo v obeh primerih ocenili s pomočjo paketa *regista*. V osnovnem modelu pri maksimizaciji funkcije verjetja zgolj izpustimo dodajanje uteži tekem. V drugem primeru pa na podlagi vrednosti xG upoštevamo s Poisson-binomsko porazdelitvijo analitično izračunane rezultate in kot uteži implementiramo njihove verjetnosti. Tako z oceno modela na naših podatkih za vsako ekipo dobimo parametra napadalne in obrambne moči $\alpha$ in $\beta$ in parameter prednosti domačega igrišča $\gamma$. Ocenjen je tudi parameter $\rho$, s katerim lahko popravimo verjetnosti nizkih izidov, ki pa ga v naših napovedih, kjer je glavni namen vključitev pričakovanih zadetkov in ugotavljanje razlik z osnovnim modelom, ne upoštevamo.

Model brez vključitve ostalih uteži (časovna komponenta, dolžina tekem, kakovost nasprotnika) je primeren za naš podatkovni okvir. Upoštevamo zgolj tekme iz ene sezone, vse v istem tekmovanju in vse trajajo enako časa.

V tabeli 6.2 za vse ekipe v tekmovanju primerjamo osnovni model Dixon-Coles z uteženim z vrednostmi xG. Vidimo, da so napadalne in obrambne moči ekip v splošnem podobne, vendar ne povsem enake. To kaže, da razlika v oceni parametrov z uporabo dejanskega števila zadetkov in z vrednostmi xG obstaja tudi na tako velikem podatkovnem okvirju, kot je celotna sezona, kjer bi pričakovali zelo podobne vrednosti xG in dejanske zadetke. Razlika v ocenjenem parametru prednosti domačega igrišča $\gamma$ je majhna, v osnovnem modelu je $\gamma = 1,342$, v uteženem modelu pa je $\gamma = 1,348$.

Kljub majhnim razlikam v ocenah parametrov se te lahko pomembno poznajo pri parametrih $\lambda$ in tako prispevajo h končni napovedi. Poglejmo, kako je z ocenami $\lambda$ pri tekmi med Liverpoolom in Southamptonom.

Najprej upoštevajmo osnovni model Dixon-Coles. Napadalna moč Liverpoola je enaka 1,641 in njegova obrambna moč je 0,899. Napadalna moč Southamptona je 0,738 in njegova obrambna moč je 1,245. Parameter prednosti domačega igrišča je 1,342.

$$E[\text{ zadetki Liverpool (osnovno) }] = 1,641 \times 1,245 \times 1,342 = 2,742$$

$$E[\text{ zadetki Southampton (osnovno) }] = 0,738 \times 0,899 = 0,663$$

Tabela 6.2: Primerjava napadalnih in obrambnih moči osnovnega in uteženega modela Dixon-Coles, Premier League, sezona 2017/18.

| Ekipa | Napad osnovni | Napad uteženi | Obramba osnovni | Obramba uteženi |
|---|---|---|---|---|
| Arsenal | 1,460 | 1,460 | 1,177 | 1,126 |
| Bournemouth | 0,894 | 0,813 | 1,364 | 1,459 |
| Brighton | 0,681 | 0,747 | 1,191 | 1,199 |
| Burnley $^{+}$ | 0,690 | 0,673 | 0,854 | 1,153 |
| Chelsea | 1,207 | 1,169 | 0,859 | 0,780 |
| Crystal Palace * | 0,887 | 1,156 | 1,240 | 1,211 |
| Everton | 0,877 | 0,904 | 1,296 | 1,253 |
| Huddersfield | 0,560 | 0,647 | 1,275 | 1,207 |
| Leicester City $^{+}$ | 1,129 | 1,010 | 1,369 | 1,118 |
| Liverpool | 1,641 | 1,544 | 0,899 | 0,830 |
| Manchester City * | 2,048 | 1,809 | 0,640 | 0,578 |
| Manchester United $^{+}$ | 1,299 | 1,207 | 0,638 | 0,996 |
| Newcastle United | 0,739 | 0,899 | 1,029 | 1,187 |
| Southampton | 0,738 | 0,833 | 1,245 | 1,056 |
| Stoke | 0,705 | 0,750 | 1,501 | 1,472 |
| Swansea | 0,547 | 0,617 | 1,226 | 1,385 |
| Tottenham | 1,431 | 1,372 | 0,816 | 0,826 |
| Watford $^{+}$ | 0,874 | 0,936 | 1,433 | 1,218 |
| West Bromwich Albion | 0,616 | 0,702 | 1,220 | 1,084 |
| West Ham * $^{+}$ | 0,976 | 0,753 | 1,531 | 1,302 |

Z * so označene ekipe, kjer je absolutna razlika med napadalnima močema večja od 0,2 in z $^{+}$ so označene ekipe, kjer je absolutna razlika med obrambnima močema večja od 0,2.

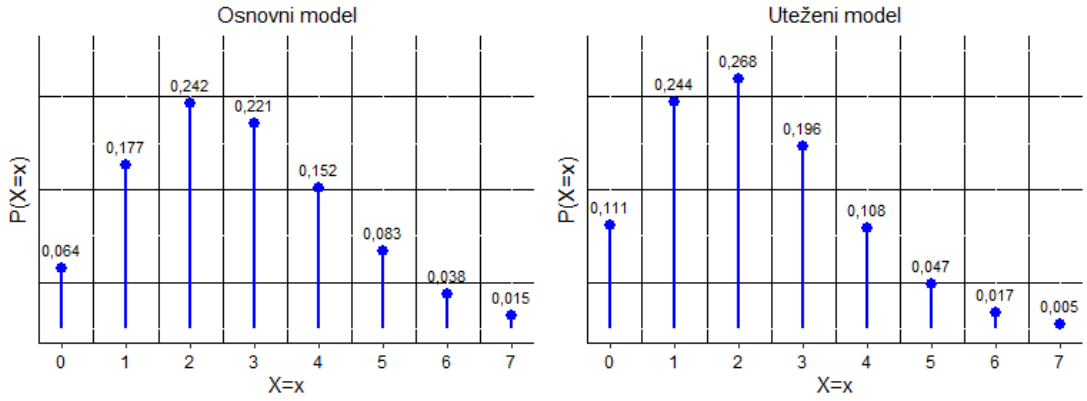
V uteženem modelu Dixon-Coles je napadalna moč Liverpoola enaka 1,544 in njegova obrambna moč je 0,830. Napadalna moč Southamptona je 0,833 in njegova obrambna moč je 1,056. Parameter prednosti domačega igrišča je 1,348.

$$E[\text{ zadetki Liverpool (uteženo) }] = 1,544 \times 1,056 \times 1,348 = 2,198$$

$$E[\text{ zadetki Southampton (uteženo) }] = 0,833 \times 0,830 = 0,691$$

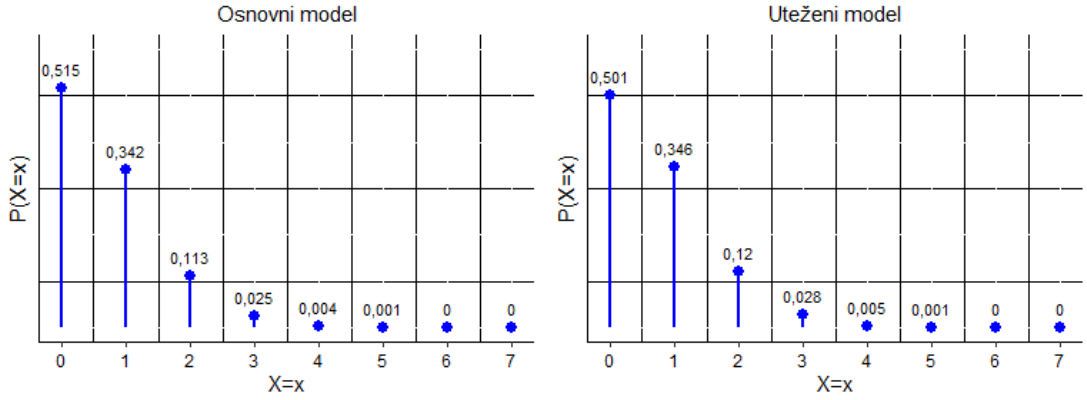
Na sliki 6.4 vidimo, da je pričakovano število zadetkov Liverpoola v uteženem modelu manjše kot v osnovnem modelu. To je posledica zmanjšane napadalne

moči Liverpoola in povečane obrambne moči Southamptona v uteženem modelu.



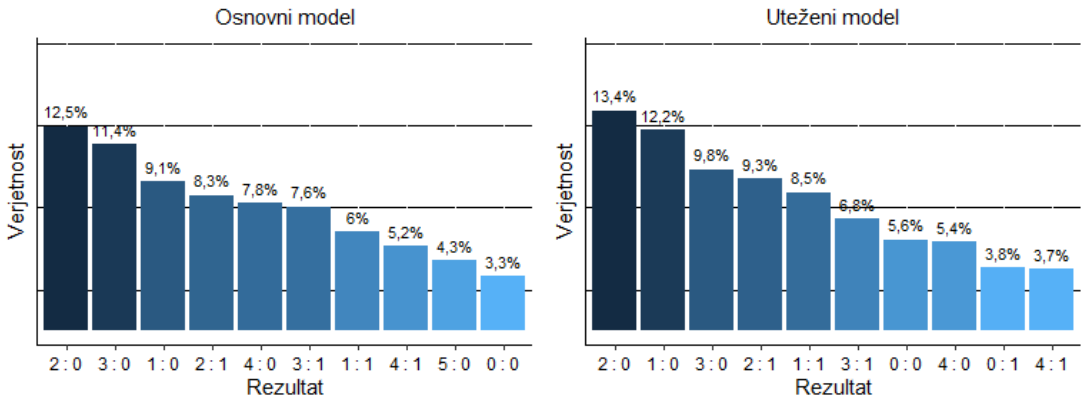

Slika 6.4: Model Dixon-Coles, verjetnost števila zadetkov Liverpoola, Liverpool in Southampton, Premier League, sezona 2017/18.

Obraten učinek kot pri Liverpoolu imajo uteži, izračunane na podlagi xG, pri Southamptonu, kjer po sliki 6.5 pričakujemo nekoliko večje število zadetkov. V tem primeru je to posledica povečane napadalne moči Southamptona, ki za odtenek preseže povečanje obrambne moči Liverpoola.

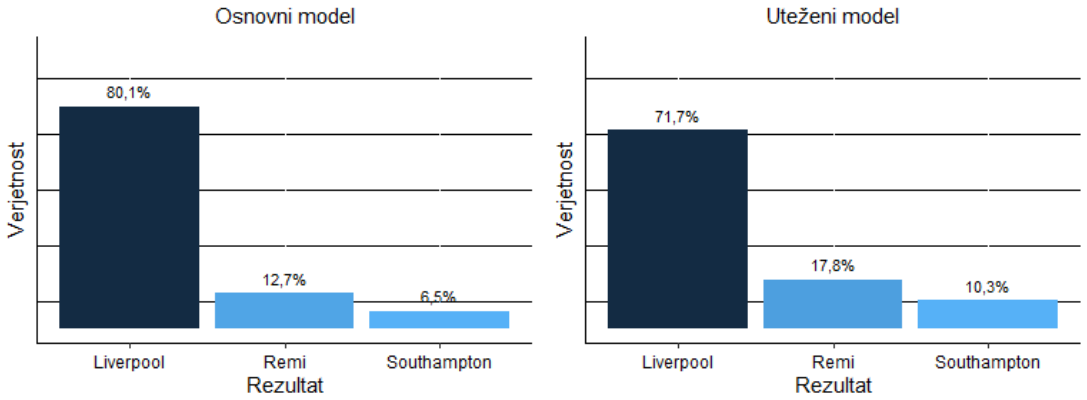



Slika 6.5: Model Dixon-Coles, verjetnost števila zadetkov Southamptona, Liverpool in Southampton, Premier League, sezona 2017/18.

Zgoraj opisano vidimo tudi na slikah 6.6 in 6.7, kjer se verjetnosti različnih izidov razlikujejo. Boljši izkupiček v uteženem modelu v primerjavi z osnovnim lahko pričakuje ekipa Southamptona. To je še posebej razvidno v verjetnosti zmage posamezne ekipe oziroma njunega remija.




Slika 6.6: Model Dixon-Coles, verjetnost končnega rezultata, Liverpool in Southampton, Premier League, sezona 2017/18.




Slika 6.7: Model Dixon-Coles, verjetnost končnega zmagovalca, Liverpool in Southampton, Premier League, sezona 2017/18.

## 6.4 Primerjava napovednih modelov, primer: Premier League 2017/18

Predstavljene modele za napovedovanje rezultatov nogometnih tekem bomo uporabili na podatkovnem okviru Premier League 2017/18 [23]. Kot rečeno, v tekmovanju nastopa 20 ekip in tekmovanje lahko razdelimo na 38 krogov, kjer v vsakem krogu vsaka ekipa igra natanko enkrat. V vsakem od 38 krogov je tako odigranih natanko 10 tekem.

Sezono bomo razdelili na 2 dela, od 1. do 19. kroga in od 20. do 38. kroga,

pri čemer bomo rezultate napovedovali za vse tekme od 20. do 38. kroga in parametre modelov izračunali na podlagi vseh že odigranih tekem. Prva napoved rezultatov bo tako temeljila na vseh odigranih tekmah do vključno 19. kroga, napovedovali pa bomo rezultate v 20. krogu. Podobno bo napoved rezultatov v 21. kolu temeljila na podlagi vseh rezultatov do vključno 20. kroga. Tako bomo napovedovali vse do zadnjega, 38. kroga, katerega rezultate bomo napovedali na podlagi prvih 37. krogov.

Rezultate tekem bomo napovedovali na podlagi modelov, ki temeljijo na, zamislih, predstavljenih v razdelkih 6.2 in 6.3. Tem modelom bomo za primerjavo oziroma referenco dodali tudi najbolj enostaven model in jih primerjali še z napovedmi, izračunanimi prek kvot stavniških hiš.

### 6.4.1 Ocenjevanje kakovosti napovedi

Za oceno kakovosti napovedi modelov in njihovo primerjavo določimo različne mere, ki bodo ocenile kakovost napovedi in predvsem omogočile njihovo primerjavo.

Končni izid nogometne tekme med ekipama A in B lahko zavzame tri vrednosti: zmaga ekipe A, zmaga ekipe B ali remi. V tem primeru napovedujemo izid opisne spremenljivke s tremi vrednostmi, kar predstavlja problem uvrščanja. Poleg končnega zmagovalca nas zanima tudi verjetnost posameznega rezultata, ki daje informacijo o ravni gotovosti v napoved. Problem tako prevedemo na problem verjetnostnega uvrščanja.

Osnovni in najpogosteje uporabljeni meri za ocenjevanje kakovosti tovrstnih napovedi sta mera log-loss in Brierjev dosežek (angl. Brier score). Obe meri ocenjujeta kakovost napovedi zgolj na podlagi napovedi verjetnosti zmagovalca (ali remija) tekme in ne upoštevata dejanskih rezultatov. Za oceno napovedi samih rezultatov bomo vključili mero, ki pričakovano vrednost zadetkov obeh ekip ($\lambda$) primerja z dejanskimi zadetki. Za še natančnejšo oceno napovedi samih rezultatov bomo zaradi zmanjšanja vpliva slučaja na dejanski rezultat vključili tudi mero, ki napovedane verjetnosti rezultatov tekem primerja s simuliranimi verjetnostmi na podlagi dejanskih vrednosti xG posameznih strelov.

#### 6.4.1.1 Mera log-loss

V osnovi zmagovalca tekme lahko napovemo zgolj na podlagi vrednosti (zmaga, remi, poraz), ki ji pripišemo največjo verjetnost. Vendar pa je pravilno napoved zmagovalca, ki ji pripišemo 90% verjetnost, potrebno vrednotiti drugače kot napoved, ki ji pripišemo 40% verjetnost. Za mero, kako blizu 1 je napovedana verjetnost dejanskega zmagovalca, bomo uporabili mero log-loss [31]:

$$\text{log-loss}_A = log(p_A)\mathbf{1}_A, \tag{6.23}$$

kjer je $p_A$ napovedana verjetnost zmage ekipe A in $\mathbf{1}_A$ indikator dejanske zmage ekipe A. Če torej dejanski rezultat ni zmaga ekipe A, je $\text{log-loss}_A = 0$, v primeru dejanske zmage ekipe A pa je $\text{log-loss}_A = log(p_A)$. Mera log-loss torej upošteva zgolj verjetnost, ki jo pripišemo dejanskemu rezultatu. Definirana je na intervalu $(-\infty, 0]$ in velja, da bližje kot je njena vrednost 0, boljša je napoved.

Mero log-loss za posamezno tekmo definiramo kot:

$$\text{log-loss} = \sum_{i \in \{H,D,A\}} log(p_i)\mathbf{1}_i, \tag{6.24}$$

kjer H označuje zmago domače ekipe, D remi in A zmago gostujoče ekipe.

Za večje število tekem kakovost modela lahko prikažemo s povprečno vrednostjo ali mediano vrednosti log-loss. Povprečje je primerno v primeru simetrične porazdelitve vrednosti, medtem ko je v primeru asimetrične porazdelitve bolj informativna mediana. Povprečje $\overline{\text{log-loss}}$ oziroma mediano Me(log-loss) izračunamo kot:

$$\overline{\text{log-loss}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{i \in \{H,D,A\}} log(p_i)\mathbf{1}_i\Big), \tag{6.25}$$

$$\text{Me(log-loss)} = \text{Me}\Big(\sum_{i \in \{H,D,A\}} log(p_i)\mathbf{1}_i\Big), \tag{6.26}$$

kjer $N$ predstavlja število vseh upoštevanih tekem.

#### 6.4.1.2 Brierjev dosežek

V primerjavi z mero log-loss posplošeni Brierjev dosežek (v osnovi je samo za binarne spremenljivke) poleg verjetnosti končnega zmagovalca enakovredno upošteva tudi verjetnosti ostalih dveh izidov. Brierjev dosežek za napoved posamezne tekme izračunamo kot vsoto kvadratov razlik med napovedanimi verjetnostmi zmagovalca in dejanskim zmagovalcem [32]:

$$\text{Brier-Score} = \sum_{i \in \{H,D,A\}} (f_i - o_i)^2, \tag{6.27}$$

kjer $f_i$ predstavlja napovedovano verjetnost zmagovalca in $o_i$ predstavlja indikator dejanskega zmagovalca (1 - Da, 0 - Ne). Vrednosti $\{H, D, A\}$ predstavljajo možne izide (domača zmaga, remi, gostujoča zmaga).

Najboljša vrednost Brierjevega dosežka za nogometno tekmo je 0 (popolnoma točna napoved) in najslabša vrednost 2 (popolnoma napačna napoved). Manjša kot je vrednost Brierjevega dosežka, boljša je naša napoved.

Za večje število tekem je v primeru simetrične porazdelitve bolj informativna povprečna vrednost Brierjevega dosežka $\overline{\text{Brier-Score}}$ in v primeru asimetrične porazdelitve mediana vrednosti, tj. Me(Brier-Score):

$$\overline{\text{Brier-Score}} = \frac{1}{N} \sum_{i=1}^{N} \Big( \sum_{i \in \{H,D,A\}} (f_i - o_i)^2 \Big), \tag{6.28}$$

$$\text{Me(Brier-Score)} = \text{Me}\Big( \sum_{i \in \{H,D,A\}} (f_i - o_i)^2 \Big), \tag{6.29}$$

kjer $N$ predstavlja število vseh upoštevanih tekem.

#### 6.4.1.3 Mera pričakovanih zadetkov

Naša primarna napoved ni napoved zmagovalca tekme, temveč napoved končnega rezultata oziroma števila zadetkov ene in druge ekipe ter njihovih verjetnosti. V ta namen definiramo mero, ki primerja napovedano število zadetkov in dejanske zadetke ter je tako razlika med pričakovanim in dejanskim številom zadetkov.

Vrednost pričakovanega števila zadetkov vključuje vsa možna števila zadetkov ekipe kot njihove verjetnosti in je tako uteženo povprečje števila napovedanih zadetkov, tj. $\mathrm{E_G} = \sum_{x=0}^{7} x \cdot P(G = x)$. V našem primeru je $\mathrm{E_G}$ v bistvu vrednost $\lambda$, izračunana na podlagi napadalnih in obrambnih moči ekip ter prednosti domačega igrišča, ki jo uporabimo v Poissonovi porazdelitvi napovedovanja števila zadetkov.

Za posamezno tekmo, mero, ki jo označimo kot $\mathrm{E_G}$-diff, izračunamo kot:

$$\mathrm{E_G\text{-}diff} = |\mathrm{E_G^H} - \mathrm{G_H}| + |\mathrm{E_G^A} - \mathrm{G_A}|, \tag{6.30}$$

kjer $\mathrm{E_G^H}$ označuje pričakovano število zadetkov domače in $\mathrm{E_G^A}$ gostujoče ekipe in $\mathrm{G_H}$ oziroma $\mathrm{G_A}$ dejansko število njunih zadetkov.

Da izračunamo mero $\mathrm{E_G}$-diff za večje število tekem, lahko, podobno kot zgoraj, izračunamo povprečno vrednost $\overline{\mathrm{E_G\text{-}diff}}$ ali mediano vrednosti $\mathrm{Me}(\mathrm{E_G}\text{-diff})$:

$$\overline{\mathrm{E_G\text{-}diff}} = \frac{1}{N} \sum_{i=1}^{N} \Big( |\mathrm{E_G^H} - \mathrm{G_H}| + |\mathrm{E_G^A} - \mathrm{G_A}| \Big), \tag{6.31}$$

$$\mathrm{Me}(\mathrm{E_G}\text{-diff}) = \mathrm{Me}\Big( |\mathrm{E_G^H} - \mathrm{G_H}| + |\mathrm{E_G^A} - \mathrm{G_A}| \Big), \tag{6.32}$$

kjer $N$ predstavlja število vseh upoštevanih tekem.

#### 6.4.1.4 Primerjava verjetnosti števila zadetkov

Pri zgoraj predstavljenih merah napovedane rezultate primerjamo z dejanskimi, pri katerih ima velik vpliv slučaj. To se lahko še posebej pozna pri majhnem

napovednem vzorcu, kjer ima lahko večinski del tekem, katerih rezultat napovedujemo, izid, ki morda ni realen glede na dejansko predstavo ekip na igrišču. Slaba napoved je tako lahko posledica slučaja in ne nujno slabega modela.

Za zmanjšanje vpliva slučaja na dejanski rezultat smo definirali mero, ki primerja verjetnosti napovedanih rezultatov z verjetnostmi 'simuliranih' oziroma 'pravičnih' rezultatov na podlagi vrednosti xG posameznih strelov. Napovedi modelov bomo ocenili na podlagi verjetnosti napovedanih in 'simuliranih' zadetkov, ki jih bomo primerjali za vsako ekipo posebej.

Napovedane verjetnosti zadetkov pogojno na ekipo, bomo kot doslej izračunali z uporabo Poissonove porazdelitve s pričakovano vrednostjo $\lambda$, izračunano na podlagi napadalnih in obrambnih moči ekip ter prednosti domačega igrišča:

$$P_{\text{nap}}(\text{G}_i = x|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^x}{x!}. \tag{6.33}$$

'Simulirane' verjetnosti zadetkov pogojno na ekipo bomo izračunali z uporabo Poisson-binomske porazdelitve na podlagi dejanskih vrednosti xG posameznih strelov: $P_{\text{sim}}(\text{G}_i = x|\text{xG}_1, ..., \text{xG}_n) = \text{Pois-Bin}(x, (\text{xG}_1, ..., \text{xG}_n))$.

V obeh primerih bomo za računsko preprostost izračunali verjetnosti števila zadetkov $x \in \{0, ..., 7\}$, saj je večje število zadetkov zelo redko. Vrednost mere, ki jo označimo z $\text{E}_{\text{xG}}$-diff, za posamezno tekmo izračunamo kot:

$$\begin{aligned}&\text{E}_{\text{xG}}\text{-diff} = \\ &= \frac{1}{2}\Big(\frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_H = x) - P_{\text{sim}}(\text{G} = x)| + \frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_A = x) - P_{\text{sim}}(\text{G}_A = x)|\Big),\end{aligned} \tag{6.34}$$

kjer $\text{G}_H$ označuje zadetke domače in $\text{G}_A$ gostujoče ekipe.

Za večje število tekem mero ponovno povprečimo preko vseh tekem, tj. $\overline{\text{E}_{\text{xG}}\text{-diff}}$, ali pa izračunamo mediano vrednosti $\text{Me}(\text{E}_{\text{xG}}\text{-diff})$:

$$\overline{\text{E}_{\text{xG}}\text{-diff}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\Big(\frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_H = x) - P_{\text{sim}}(\text{G} = x)| + \right. \\ \left. + \frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_A = x) - P_{\text{sim}}(\text{G}_A = x)|\Big)\right), \tag{6.35}$$

$$\text{Me}(\text{E}_{\text{xG}}\text{-diff}) = \text{Me}\left(\frac{1}{2}\Big(\frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_H = x) - P_{\text{sim}}(\text{G} = x)| + \right. \\ \left. + \frac{1}{7}\sum_{x=0}^{7}|P_{\text{nap}}(\text{G}_A = x) - P_{\text{sim}}(\text{G}_A = x)|\Big)\right), \tag{6.36}$$

kjer $N$ predstavlja število vseh upoštevanih tekem.

#### 6.4.1.5 Primerjava napovedi s trgom

Za primerjavo modelskih napovedi s tržnimi bomo izračunali odstopanje napovedanih verjetnosti končnega zmagovalca tekme oziroma neodločenega rezultata s strani modelov od napovedanih verjetnosti, izračunanih prek povprečnih kvot vseh večjih stavnih hiš na trgu (od 32 do 43 stavniških hiš), pridobljenih s spletišča Footbal-Data [33].

Upoštevali bomo torej več modelov, ki pripadajo različnim stavniškim hišam. Potrebno se je zavedati, da so kvote stavniških hiš odvisne tudi od stavničarjev, ki kvote in posledično iz njih izračunane verjetnosti lahko spremenijo. Ta vpliv sicer zmanjšamo z upoštevanjem povprečne vrednosti več stavniških hiš.

Verjetnost posameznega rezultata lahko iz stavniške kvote, ki je v evropskem merilu podana v decimalnih številkah (npr. 3,2), izračunamo kot verjetnost = $\frac{1}{\text{kvota}}$ [34] . V našem primeru bo, kot rečeno, uporabljena kvota, ki predstavlja povprečje vseh kvot večjih stavniških hiš na trgu.

Za posamezno tekmo definiramo mero, ki jo označimo z Market-diff in jo izračunamo kot:

$$\text{Market-diff} = \frac{1}{3} \sum_{i \in \{H,D,A\}} |f_i - b_i|, \tag{6.37}$$

kjer $f_i$ predstavlja napovedovano verjetnost rezultata s strani posameznega modela in $b_i$ predstavlja povprečno napovedano verjetnost rezultata na trgu.

Tudi tokrat za večje število tekem vrednosti Market-diff povprečimo po vseh tekmah ($\overline{\text{Market-diff}}$) ali pa v primeru asimetrične porazdelitve izračunamo njihovo mediano (Me(Market-diff)):

$$\overline{\text{Market-diff}} = \frac{1}{N} \sum_{i=1}^{N} \Big( \frac{1}{3} \sum_{i \in \{H,D,A\}} |f_i - b_i| \Big), \tag{6.38}$$

$$\text{Me}(\text{Market-diff}) = \text{Me}\Big( \frac{1}{3} \sum_{i \in \{H,D,A\}} |f_i - b_i| \Big), \tag{6.39}$$

kjer $N$ predstavlja število vseh upoštevanih tekem.

### 6.4.2 Primerjava modelov, rezultati

Z merami definiranimi v razdelku 6.4.1, ki ocenijo kakovost napovedi modela, bomo primerjali štiri modele za napovedovanje rezultatov nogometnih tekem, predstavljene v razdelkih 6.2, 6.2.2, 6.3 in 6.3.2.

Prvi je osnovni Poissonov model, kjer obrambne in napadalne moči ekip in tako parametre $\lambda$ izračunamo ročno, na podlagi dejanskih zadetkov; v nadaljevanju ga imenujemo Osnovni Poisson. Nadgradili smo ga tako, da smo namesto dejanskih zadetkov upoštevali vsoto vrednosti xG posameznih tekem; ta model poimenujmo xG Poisson. Videli smo, da obrambne in napadalne moči ekip lahko ocenimo tudi skozi funkcijo verjetja z modelom Dixon-Coles, na katerem temeljita ostala dva uporabljena modela. Prvi model poimenujmo Osnovni Dixon-Coles, kjer v funkcijo verjetja ne vključimo uteži, v drugem modelu, ki ga poimenujemo xG Dixon-Coles, pa v funkcijo verjetja vključimo uteži v smislu xG vrednosti posameznih strelov. V nobenem izmed modelov Dixon-Coles ne bomo uporabili

zraven izračunanega popravka verjetnosti nizkih izidov, tj. parametra $\rho$, saj je naš cilj oceniti izboljšanje napovedi z vključitvijo xG.

Za primerjavo napovedi modelov z napovedmi trga bomo dodali napovedane verjetnosti zmagovalca tekme ali neodločenega rezultata, izračunane preko povprečnih kvot 32 do 43 večjih stavnih hiš na trgu, pridobljenih s spletišča Football-Data [33]. Za boljšo predstavo kakovosti napovedi modelov bomo pri vseh merah kakovosti napovedi vključili tudi t.i. referenčno napoved, kjer je napoved neodvisna od ekip, katerih rezultat tekme napovedujemo, in tako najosnovnejša, ki ni povsem slučajna. Za vsako mero bo referenčna napoved definirana nekoliko drugače in bo predstavljena sproti.

#### 6.4.2.1 Mera log-loss

Z vidika najosnovnejše mere log-loss, ki ocenjuje zgolj napoved verjetnosti dejanskega zmagovalca tekme, imata najboljše napovedi Poissonov model glede na skupno vrednost xG in model Dixon-Coles glede na vrednost xG posameznih strelov. Modela, ki temeljita na dejanskih zadetkih, nekoliko slabše napovesta zmagovalca tekme, pri čemer je model Dixon-Coles z uporabljeno funkcijo verjetja za odtenek slabši.

Referenčni model zmagovalca tekme v tem primeru napove na podlagi zmag domačih in gostujočih moštev. Zmagovalcu tekme oziroma izenačenemu rezultatu pripiše verjetnost, ki je enaka deležu zmag domačih oziroma gostujočih moštev ali remijev v učnem naboru podatkov.

Na sliki 6.8 vidimo, da je porazdelitev vrednosti mere log-loss pri večini modelov rahlo asimetrična. Tako je ob upoštevanju vseh tekem nekoliko bolj kot povprečje vrednosti informativna njihova mediana.

Na sliki 6.8 so prikazane tudi povprečne vrednosti mere log-loss in njihove mediane. Razlika je tako med povprečnimi vrednostmi kot medianami majhna, nekoliko večja sicer pri medianah. Če nekoliko obrnemo enačbo (6.25) in izračunamo povprečno napovedano verjetnost dejanskega zmagovalca kot $\overline{p_z} = exp\{\overline{\text{log-loss}}\}$, z najboljšim, xG Poissonovim modelom, v povprečju dejanskem zmagovalcu

tekme pripisujemo 37,2% možnosti za zmago, pri najslabšem, osnovnem modelu Dixon-Coles, pa dejanskem zmagovalcu tekme pripisujemo 36,2% možnosti za zmago. Z referenčnim modelom dejanskemu zmagovalcu pripisujemo 34,3% možnosti, trg pa dejanskemu zmagovalcu pripisuje nekoliko višje možnosti in, sicer 39,9%.

Nekoliko slabši povprečni napovedi osnovnega Poissonovega modela in osnovnega modela Dixon-Coles sta posledici močnejše asimetrije porazdelitve vrednosti v levo kot pri preostalih modelih in posledično sta mediani teh dveh modelov v absolutnem smislu najnižji in tako najboljši. Pri xG modelih je povprečje manjše od mediane, kar v našem primeru negativnih vrednosti kaže na asimetrijo v desno. Podobno kot pri povprečju izračunamo, da je polovica verjetnosti pravilnih napovedi zmagovalca pri osnovnem modelu Dixon-Coles oziroma Poissonovem modelu večja oziroma manjša od 38,7% oziroma 38,2%, medtem ko je ta vrednost pri modelu xG Poisson ali modelu xG Dixon-Coles enaka 36,4%. Referenčni model polovico verjetnosti napove kot večje oziroma manjše od 29,1%. Če torej upoštevamo mediano vrednosti, so kakovosti napovedi modelov rahlo boljše in v primerjavi z povprečnimi vrednostmi mer je razlika kakovosti napovedi modelov glede na referenčni model, večja. Najboljša sta osnovna modela, nekoliko kvalitetnejši je sicer model Dixon-Coles.

#### 6.4.2.2 Brierjev dosežek

Brierjev dosežek, ki upošteva vse možne končne izide (domača zmaga, gostujoča zmaga, remi), vodi do nekoliko drugačnih zaključkov kot mera log-loss. Najboljša modela glede na Brierjev dosežek sta oba Poissonova modela, glede na dejanske zadetke in glede na skupno vrednost xG. Sledita oba modela Dixon-Coles, katerih kakovost napovedi je v povprečju prav tako zelo podobna.

Pri Brierjevem dosežku je referenčni model enak opisanemu modelu pri meri log-loss. Zmagovalcu tekme oziroma izenačenemu rezultatu pripiše verjetnost, ki je enaka deležu zmag domačih oziroma gostujočih moštev ali remijev v učnem naboru podatkov.

Tudi pri Brierjevem dosežku, na sliki 6.9, vidimo morda rahlo asimetrično

porazdelitev vrednosti v desno.

Na sliki 6.9 so tudi rezultati povprečnih vrednosti in mediane vrednosti. Po enačbi (6.28), kjer enakovredno upoštevamo vse tri možne izide, in izračunu $\overline{p} = \sqrt{\frac{\overline{\text{Brier-Score}}}{3}}$, Poissonova modela pri končnem izidu posamezne tekme v povprečju zgrešita za 44,4%. Za modela Dixon-Coles velja, da pri končnem izidu v povprečju zgrešita za 44,9%. Z referenčnim modelom pri posameznem izidu zgrešimo za 46,4%, trg pa pri posameznem izidu zgreši za 43,7%.

Razlike med napovedmi modelov so v primerjavi s povprečnimi vrednostmi večje pri medianah. Ponovno sta najboljša osnovna modela, nekoliko boljši sicer model Dixon-Coles, medtem ko med modeloma, ki upoštevata vrednosti xG, ni večjih razlik. Podobno kot pri povprečni vrednosti lahko izračunamo, od katere vrednosti je polovica vrednosti manjša oziroma večja. Tudi interpretacija je podobna. Pri osnovnem modelu Dixon-Coles je porazdelitev vrednosti nekoliko asimetrična v desno in omenjena vrednost je enaka 43,8%; pri osnovnem Poissonovem modelu je 44,1%, porazdelitev je tudi tu rahlo asimetrična v desno; pri xG modelih je ta vrednost enaka 45,1%, kjer pa je porazdelitev rahlo asimetrična v levo. Za primerjavo: pri referenčem modelu v polovici primerov zgrešimo za manj oziroma več kot 50,6%.

#### 6.4.2.3 Mera pričakovanih zadetkov

Mera pričakovanih zadetkov ne upošteva zgolj napovedi zmagovalca temveč upošteva napovedi števila zadetkov in njihovih verjetnosti. Primerja pričakovano število zadetkov, tj. $\lambda$ posameznega modela, in dejansko število zadetkov ekip. Ker za napovedi trga teh podatkov nimamo, mere za povprečno tržno napoved tu ne moremo oceniti.

Ko se v napovedi spustimo na nivo zadetkov in ne zgolj končnega zmagovalca, najboljša postaneta modela, zgrajena na podlagi pričakovanih zadetkov. Najboljši model je xG Poisson model, ki mu sledi model xG Dixon-Coles, kar je v skladu z definicijo mere. Sledita osnovni Poissonov model in osnovni model Dixon-Coles.

Pri meri pričakovanih zadetkov referenčni model za pričakovano število zadetkov domače oziroma gostujoče ekipe upošteva povprečno število zadetkov domačih oziroma gostujočih ekip v učnem naboru podatkov.

Tokrat je (slika 6.10) asimetrična porazdelitev mere pričakovanih zadetkov v desno večja in je tako v primerjavi s povprečjem bolj informativna mediana.

Tudi rezultati povprečnih vrednosti in mediane so prikazani na sliki 6.10. Upoštevajoč enačbo (6.31), xG Poissonov model za posamezno tekmo v povprečju napove 1,74 zadetkov preveč ali premalo, kar je 0,87 zadetka na ekipo, medtem ko osnovni model Dixon-Coles za posamezno tekmo v povprečju napove 1,86 zadetkov preveč oziroma premalo, kar je 0,93 zadetka na ekipo. Za primerjavo, referenčni model za posamezno tekmo v povprečju napove 0,96 zadetka preveč oziroma premalo. Razlike med povprečnimi absolutnimi ocenami napovedi modelov so ponovno zelo majhne.

Upoštevajoč mediano, ki je pri meri pričakovanih zadetkov izrazito bolj informativna, so v primerjavi s povprečjem modeli bolj enakovredni. Najboljši je model xG Dixon-Coles in sledi mu najosnovnejši Poissonov model. Razlike med ocenami napovedi modelov in tako tudi napovedmi števila zadetkov so majhne in se gibljejo okrog 1,6 zadetka na tekmo oziroma 0,8 na ekipo. V primerjavi s povprečnimi vrednostmi so napovedi modelov nekoliko boljše upoštevajoč mediano, kar je posledica asimetrije porazdelitve v desno.

#### 6.4.2.4 Primerjava verjetnosti števila zadetkov

Nadgradnja mere pričakovanih zadetkov je primerjava verjetnosti posameznega števila zadetkov. S tem iz dejanskega števila zadetkov poskušamo izničiti slučajnost in upoštevati 'pravično' število zadetkov in njihove verjetnosti na podlagi posameznih vrednosti xG ter te verjetnosti primerjati s tistimi, ki jih dobimo prek pričakovanega števila zadetkov oziroma $\lambda$ posameznega modela. Ponovno so izpuščene povprečne tržne napovedi.

Rangiranje modelov je enako kot pri meri pričakovanih zadetkov. Najboljša modela sta modela, ki namesto dejanskih zadetkov upoštevata pričakovane za-

detke (xG Poisson in xG Dixon-Coles); sledita jima modela, ki v izračun $\lambda$ vključita dejanske zadetke (Osnovni Poisson in Osnovni Dixon-Coles).

Referenčni model - podobno kot pri meri pričakovanih zadetkov - za pričakovano število zadetkov in s tem $\lambda$ upošteva povprečno število zadetkov domačih oziroma gostujočih ekip v učnem naboru podatkov.

Dobljene rezultate agregirane mere $E_G$-diff interpretiramo kot povprečno oziroma, v primeru mediane, sredinsko razliko med napovedano oziroma pričakovano verjetnostjo števila zadetkov $x \in \{0, ..., 7\}$ in 'pravično' verjetnostjo števila zadetkov na podlagi xG posameznih strelov, kjer ekipi obravnavamo enakovredno.

Tudi tokrat je (slika 6.11) porazdelitev vrednosti mere verjetnosti števila zadetkov nekoliko asimetrična v desno in pričakujemo rahlo manjše vrednosti mediane, ki je tako morda rahlo bolj informativna.

Na isti sliki so rezultati povprečnih vrednosti in mediane. Pri najboljšem, xG Poissonovem modelu, v povprečju pri posameznem številu zadetkov $x \in \{0, ..., 7\}$ napovedana verjetnost od 'pravične' verjetnosti po absolutni vrednosti odstopa za 0,063, pri najslabšem, osnovnem modelu Dixon-Coles, pa pri posameznem številu zadetkov napovedana verjetnost od 'pravične' verjetnosti odstopa za 0,071. Za primerjavo, referenčni model za posamezno tekmo posameznem številu zadetkov v povprečju napove verjetnost, ki od 'pravične' odstopa za 0,071, kar je enako kot osnovni model Dixon-Coles. Če želimo napovedi, ki so vsaj nekoliko boljše od najenostavnejše napovedi, moramo upoštevati pričakovano namesto dejanskega števila zadetkov. Pri tem velja, da vsa števila zadetkov $x \in \{0, ..., 7\}$ oziroma njihove verjetnosti pri izračunu povprečja obravnavamo enakovredno in to nekoliko zmanjša napako napovedi, saj so pri večjem številu zadetkov razlike zaradi samih vrednosti veliko manjše kot pri manjšem številu zadetkov. Če bi upoštevali zgolj zadetke $x \in \{0, ..., 3\}$, bi bile razlike med povprečnimi napakami napovedi modelov večje.

Zaradi rahle asimetričnosti porazdelitev v desno so mediane vrednosti nekoliko manjše od njihovih povprečij. Interpretacija kakovosti napovedi modelov je podobna interpretaciji upoštevajoč povprečne vrednosti in razlike med modeli so podobne.

#### 6.4.2.5 Primerjava napovedi s trgom

Na koncu napovedi zmagovalca tekme, ki jih podajo obravnavani štirje modeli, še neposredno primerjamo s povprečnimi tržnimi napovedmi. Mera Market-diff izračuna povprečno absolutno razliko med napovedano verjetnostjo zmagovalca tekme s strani modela in povprečno tržno napovedano verjetnostjo.

Referenčni model je v tem primeru enak referenčnemu modelu pri merah log-loss in Brierjevem dosežku. Zmagovalcu tekme pripiše verjetnost, ki je enaka deležu zmag domačih oziroma gostujočih moštev ali remijev v učnem naboru podatkov.

Tokrat je porazdelitev vrednosti modelov (slika 6.12) močno asimetrična v desno in bolje je upoštevati mediane. Rezultati povprečnih vrednosti in mediane so na sliki 6.12. Povprečnim tržnim napovedim so najbolj podobne napovedi Poissonovih modelov medtem, ko napovedi modelov Dixon-Coles precej bolj odstopajo. V povprečju se napovedane verjetnosti xG Poissonovega modela od povprečnih tržnih napovedi razlikujejo zgolj za 0,056 medtem, ko se napovedi osnovnega modela Dixon-Coles razlikujejo za 0,101. Verjetnosti referenčnega modela se razlikujejo za 0,129. Tu so razlike med modeli nekoliko večje kot pri prejšnjih merah.

Pričakovano so zaradi asimetrije porazdelitve v desno mediane vrednosti manjše od njihovih povprečij. Ponovno so razlike med modeli podobne kot pri povprečnih vrednostih in interpretacija kakovosti napovedi modelov je podobna interpretaciji upoštevajoč povprečne vrednosti. Razlika je zgolj v referenčnem modelu, ki se bolj približa tržnim napovedim.

Razlike med modelskimi in tržnimi napovedmi so, kot smo lahko videli že pri meri log-loss in Brierjevem dosežku, majhne. V skladu z zgornjimi ugotovitvami to pomeni, da tudi tržne napovedi, ki poleg začetnih kvot izračunanih s strani tržnih modelov upoštevajo še subjektivna mnenja in veliko ostalih informacij, ki jih v model ne moremo enostavno vključiti, relativno slabo napovedujejo dejanske zmagovalce tekem. Torej je napovedovanje zmagovalcev kot tudi števila zadetkov nogometnih tekem zares zahtevno delo, še posebej v angleški Premier League.

### 6.4.3 Ugotovitve

Ugotovili smo, da so razlike med modeli, ki v svojih napovedih upoštevajo dejanske zadetke in pričakovane zadetke xG, majhne. Obenem smo videli, da so tudi razlike med modelskimi in tržnimi in referenčnimi napovedmi majhne. Če upoštevamo, da so glede na uporabljene mere napovedane verjetnosti končnega zmagovalca ali dejanskega števila zadetkov relativno daleč od dejanskih izidov oziroma zadetkov, obenem pa napovedane verjetnosti zadetkov od 'pravičnega' števila zadetkov, dobljenega na podlagi xG posameznih strelov, niso tako različne, vidimo, da je napovedovanje izidov nogometnih tekem zelo težavno, saj je vpliv slučaja velik. Menimo, da so kljub majhnim razlikam med modeli napovedi strokovno pomembne in modelov ni smiselno enačiti.

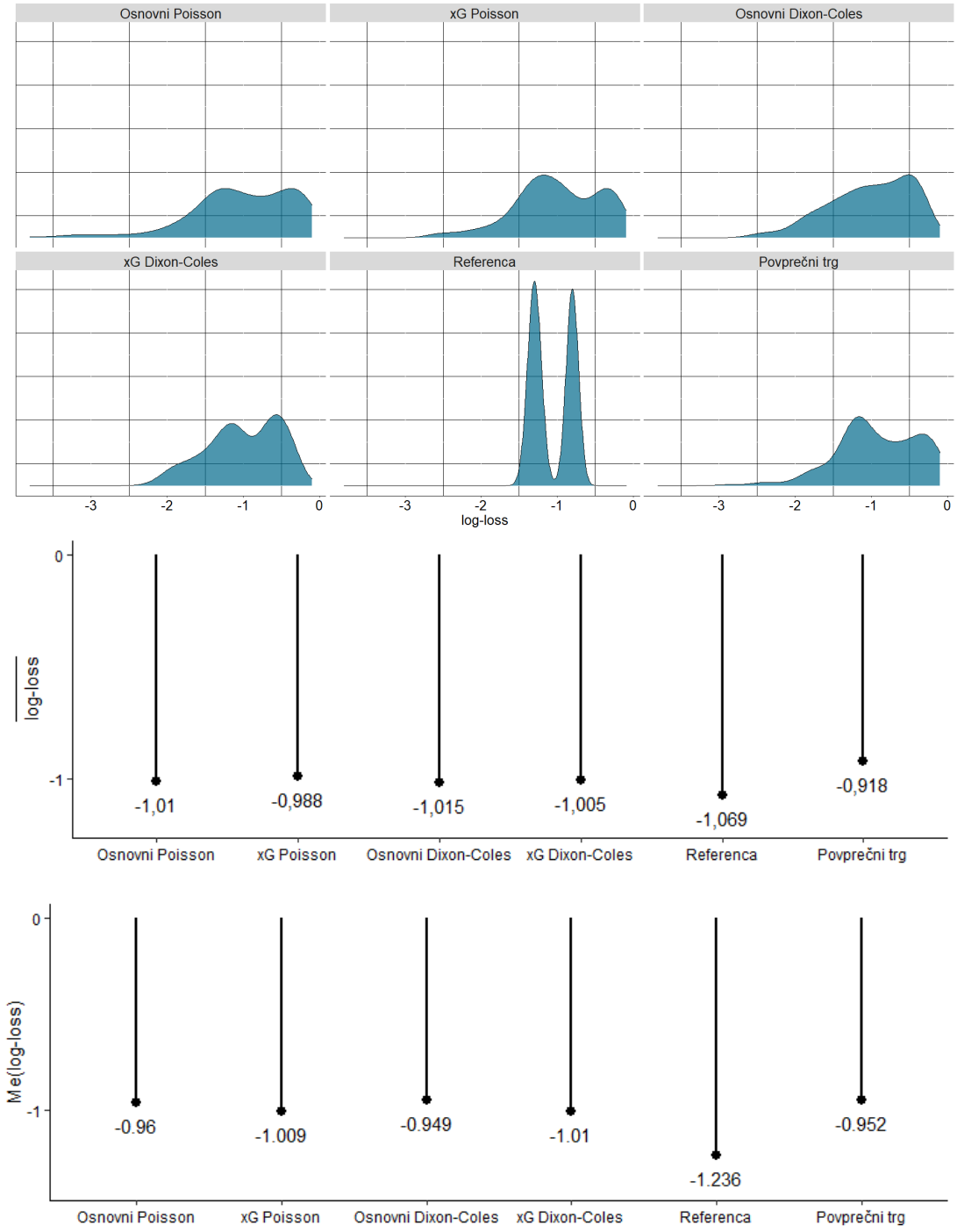
Dodati velja, da smo za napovedovanje rezultatov verjetno izbrali najtežji podatkovni okvir, saj angleška liga slovi kot najbolj izenačena, kjer vsaka ekipa lahko premaga vsako in je dejanske zmagovalce, kaj šele rezultate, izjemno težko napovedati. Če bi vzeli špansko ligo ali drugo tekmovanje, kjer je razkorak med kakovostjo ekip večji, bi bile razlike med napovedanimi verjetnostmi in dejanskimi rezultati verjetno veliko manjše.

V vseh primerih, ne glede na to ali želimo napovedati zgolj zmagovalca tekme ali število zadetkov posamezne ekipe, se za najboljši model izkaže Poissonov model na podlagi skupne vrednosti xG, ki mu sledi model na podlagi vrednosti xG posameznih strelov. Vidimo torej, da upoštevanje vrednosti xG nekoliko izboljša napovedane verjetnosti zmagovalca tekme in še posebej končnega rezultata. Napovedane verjetnosti lahko s povprečnimi tržnimi primerjamo samo pri napovedi zmagovalca in velja, da so tržne vrednosti v povprečju boljše. To je razumljivo, saj napovedane verjetnosti s strani tržnih modelov popravijo še glede na ostale informacije, ki jih v naše modele nismo vključili oziroma jih je vključiti izjemno težko.

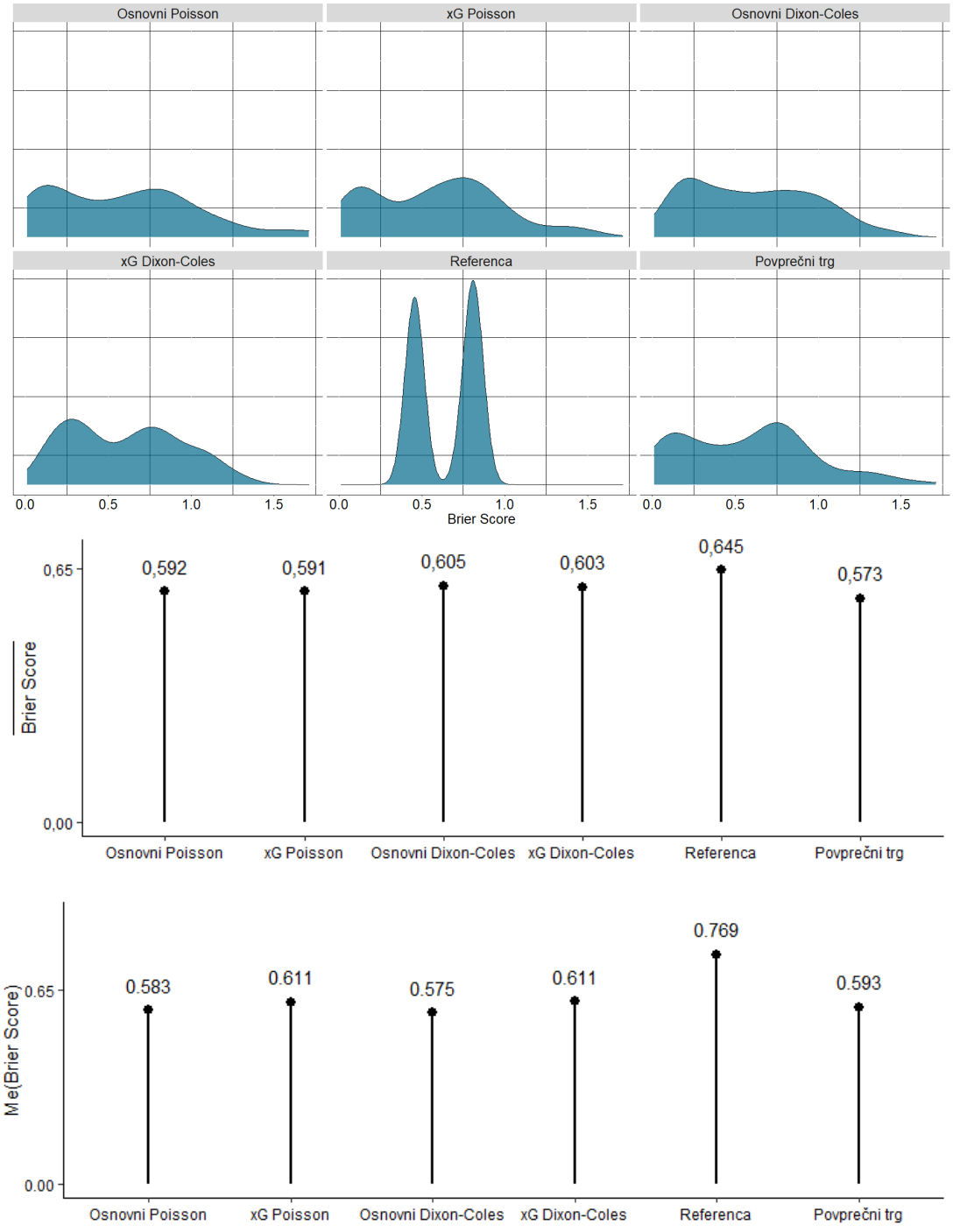
Lahko sklenemo, da je napovedovanje nogometnih izidov zaradi slučajnosti zelo težavna naloga, saj so razlike med dejanskimi in pričakovanimi zadetki majhne, enako pa velja tudi za zmagovalce tekem v primerjavi s poraženci. To se zlasti pozna pri upoštevanju in povprečenju rezultatov in kakovosti napovedi več tekem, tako v učnih kot testnih podatkih, saj dejanski zadetki na dolgi rok

sledijo pričakovanim. Poleg tega nam iz že omenjenih razlogov, kot so navidezna napihnjenost vrednosti xG, konstantno preseganje ali nedoseganje pričakovanih predstav na podlagi xG, obrambni in napadalni izkoristek ter taktične zamisli ekip, več informacije lahko nudi tudi dejanska statistika zadetkov.

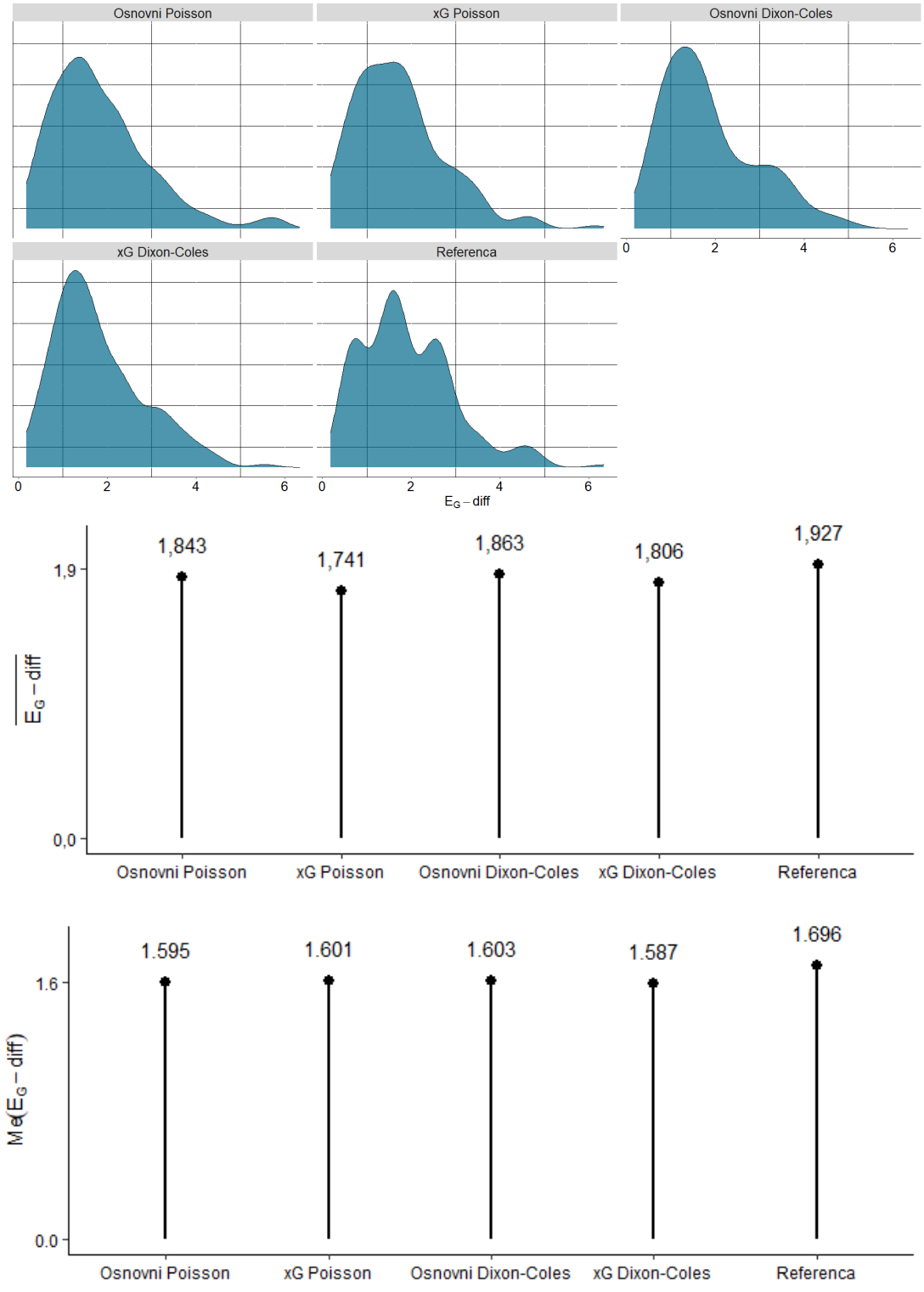
V praksi za najboljše napovedi izidov tekem seveda ni priporočljivo upoštevati zgolj modelov, pač pa je smiselno vključiti tudi svoje poznavanje nogometa in nekatera subjektivna mnenja. Statistika pričakovanih zadetkov nam tako lahko služi kot pomembno dopolnilo pri preučevanju ekip, ne pa kot nadomestek upoštevanja dejanskih zadetkov.



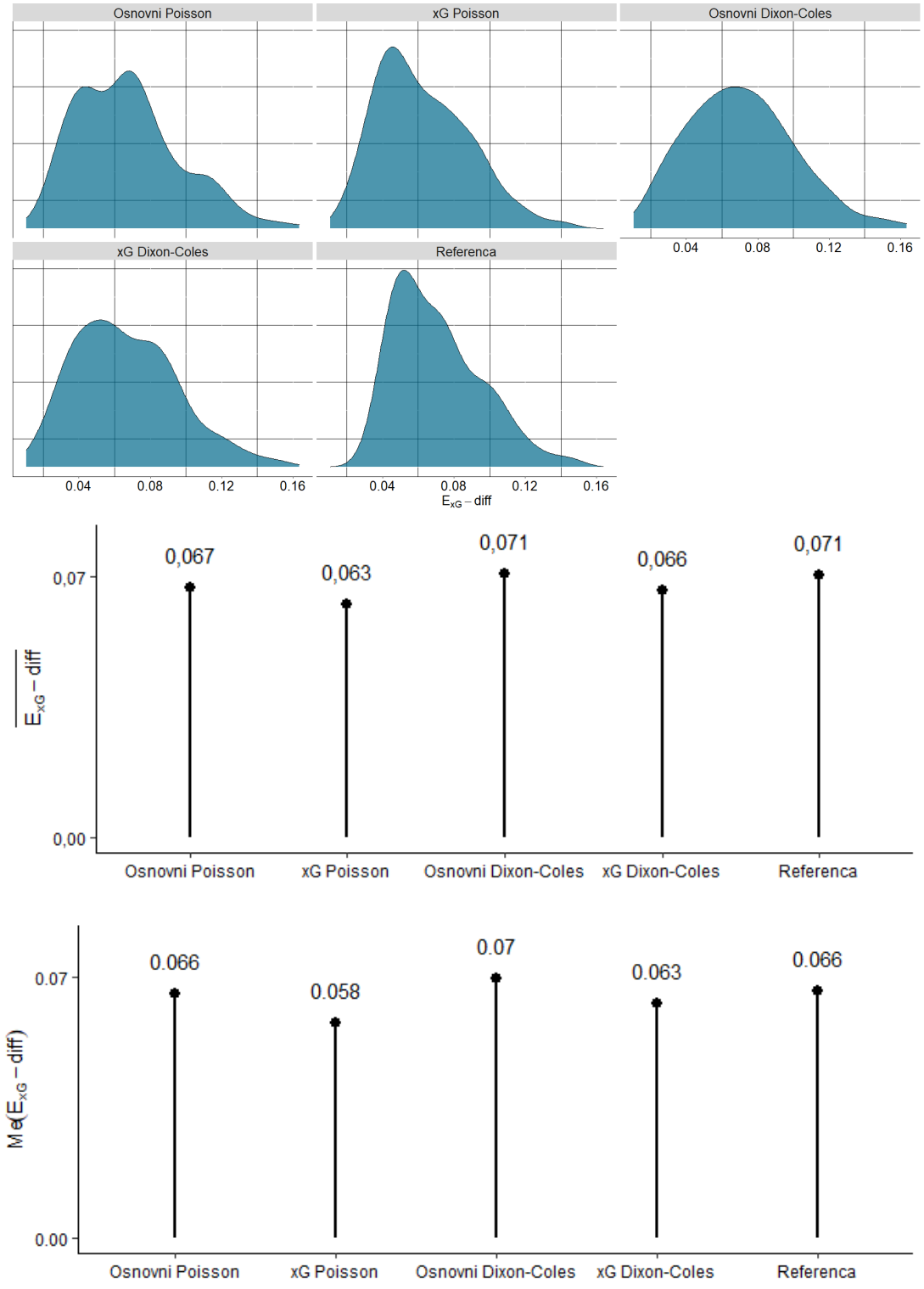

Slika 6.8: Mera log-loss, porazdelitev vrednosti, mediana in povprečje, Premier League, sezona 2017/18, krogi 20-38.



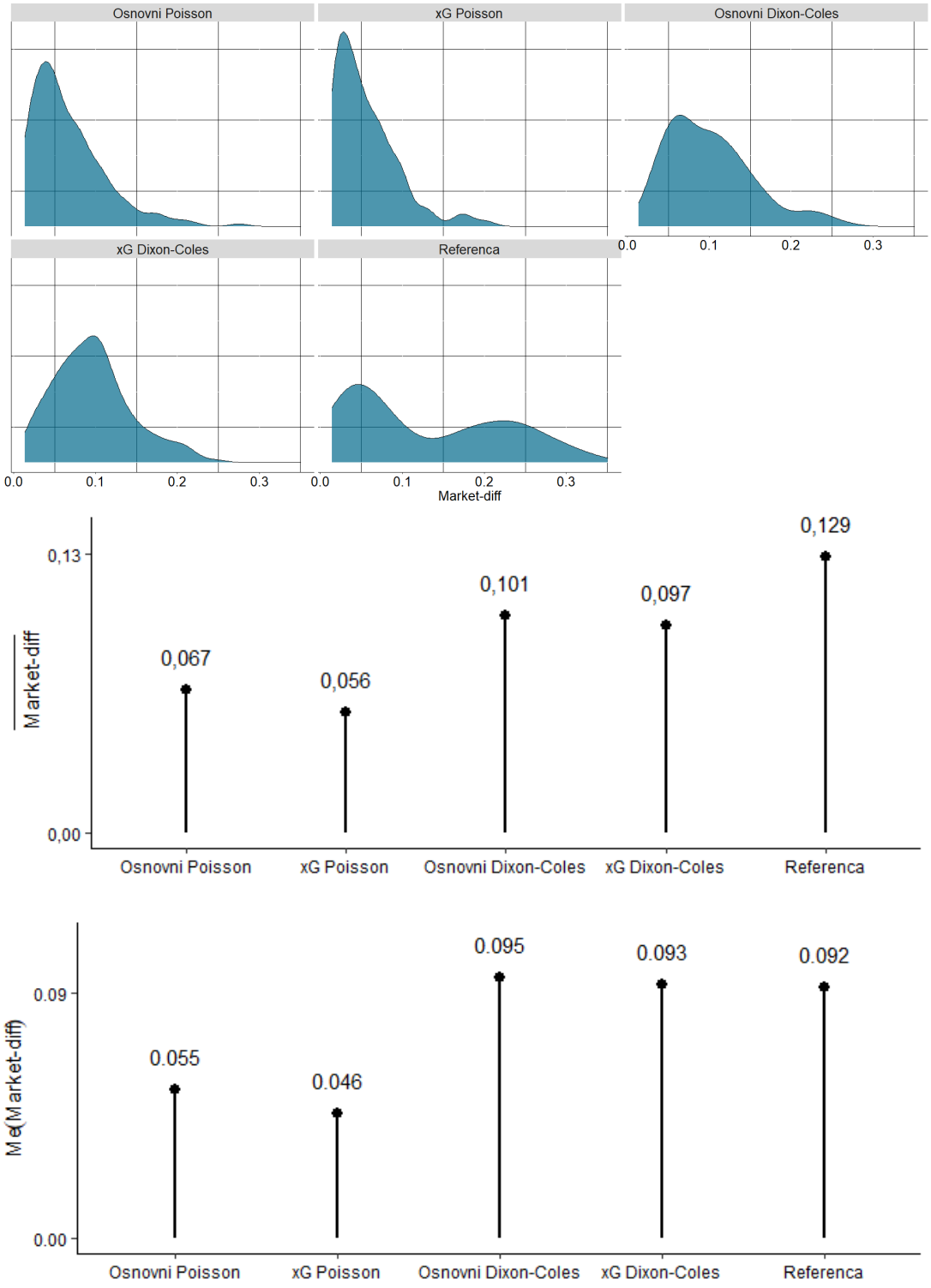

Slika 6.9: Brierjev dosežek, porazdelitev vrednosti, povprečje in mediana, Premier League, sezona 2017/18, krogi 20-38.




Slika 6.10: Mera pričakovanih zadetkov, porazdelitev vrednosti, povprečje in mediana, Premier League, sezona 2017/18, krogi 20-38.




Slika 6.11: Primerjava verjetnosti števila zadetkov, porazdelitev vrednosti, povprečje in mediana, Premier League, sezona 2017/18, krogi 20-38.




Slika 6.12: Primerjava napovedi zmagovalca s trgom, porazdelitev vrednosti, povprečje in mediana, Premier League, sezona 2017/18, krogi 20-38.

# 7 Zaključek

Podatkovna analitika je v zadnjih dveh desetletjih postala pomemben del številnih športnih organizacij. S pomočjo strokovnjakov s specifičnimi znanji podatkovno analitiko vključujejo v procese skavtinga, zmanjšujejo tveganje poškodb in omogočajo hitrejšo rehabilitacijo, optimizirajo trenažni proces in načrtujejo razvoj mladih športnikov, trenerji pa na njej gradijo različne taktične zamisli. Pri večjih športnih organizacijah se podatkovna analitika uporablja tudi za določanje razvrstitev pri različnih žrebanjih in sistemih tekmovanj. Z razvojem športne analitike so strokovnjaki razvili različne statistike in kazalnike, ki omogočajo kvantifikacijo nekaterih opažanj. Bistveno je, da mora imeti dober športni analitik poleg statističnega tudi odlično notranje poznavanje športne igre, ki jo preučuje. Enako ali celo bolj pomembna od zbranih podatkov je njihova pravilna analiza in interpretacija.

V ospredju mer, ki se danes uporabljajo v nogometu in drugih športih z žogo, je mera pričakovanih zadetkov, ki jo označimo z xG (angl. expected goals). Izračuna se jo na podlagi številnih spremenljivk, ki opisujejo situacijo na igrišču v času strela, in predstavlja verjetnost, da se določen strel pretvori v zadetek. Porazdelitev vrednosti xG je močno asimetrična v desno s povprečno vrednostjo okrog 0,12. V praksi je vrednost xG največkrat ocenjena na podlagi regresijskih modelov, kjer je najbolj pogosta logistična regresija. Ta omogoča enostavno interpretacijo spremenljivk modela, kar je za mnoge končne uporabnike mere zelo pomembno.

Na podlagi prosto dostopnih podatkov podjetja StatsBomb, kjer je vsak strel opisan z mnogo spremenljivkami, smo predstavili spremenljivke, za katere smo menili, da vplivajo na verjetnost zadetka, in predstavili njihovo povezanost z

zadetki. S pomočjo teorije logistične regresije in metod ocenjevanja kakovosti modela smo tudi sami ustvarili model za izračun vrednosti xG. Z njim smo predstavili, katere spremenljivke značilno in na kakšen način vplivajo na verjetnost zadetka pri nogometni igri najvišje moške profesionalne ravni. Spremenljivki, ki najbolje razložita verjetnost zadetka, sta oddaljenost strela od gola in kot strela na gol. Izkazalo se je, da verjetnost zadetka ni odvisna od interakcije strani strela in noge strela na gol in je vseeno, če upoštevamo kot, ki ne loči med levo in desno stranjo igrišča. Dlje kot smo od gola, manjša je verjetnost zadetka, in bolj na sredini širine igrišča kot je strel izveden, večja je verjetnost zadetka. Pomembni pojasnjevalni spremenljivki verjetnosti zadetka sta tudi število obrambnih igralcev med strelcem in golom in njihova gostota. Z večanjem vrednosti teh spremenljivk se verjetnost zadetka zmanjšuje. Zanimiva je ugotovitev, da je verjetnost zadetka večja, če pred strelom ne pride do predhodne uspešne ključne podaje, kar seveda ne pomeni, da odsvetujemo podajanje. Podaja po zraku in s tem tudi strel z glavo v primerjavi s strelom z nogo pričakovano zmanjša verjetnost zadetka. Pri tehniki strela je najboljši strel običajen, medtem ko bolj atraktivni streli prinašajo manjšo verjetnost zadetka. Seveda strel na prazen gol, 'ena na ena' z vratarjem in strel na prvi dotik povečajo verjetnost zadetka, kar velja tudi za strel po uspešni podaji v prostor. Nekatere ostale spremenljivke, ob upoštevanju navedenih v modelu niso značilno vplivale na verjetnost zadetka.

Z vrednostmi xG posameznih strelov lahko ovrednotimo, v kolikšni meri dejansko število zadetkov igralca ali ekipe odstopa od pričakovanega števila zadetkov. Tu je poleg kumulativne vrednosti posameznih xG pomembna tudi njihova absolutna vrednost oziroma porazdelitev. Ta nosi več informacije kot zgolj razlika med kumulativnim pričakovanim in dejanskim številom zadetkov. Relativno odstopanje dejanskega števila zadetkov od pričakovane vrednosti lahko ovrednotimo s pomočjo Poisson-binomske porazdelitve. Izkaže se, da je ob predpostavki enake kumulativne vrednosti xG verjetnost določenega števila zadetkov manjša v primeru manjšega števila strelov z večjo vrednostjo xG. Na podoben način, a v obratni smeri, je verjetnost zmage, poraza ali remija na določeni tekmi odvisna od porazdelitve xG. Ob enaki kumulativni vrednosti xG nasprotnih ekip je verjetnost zmage na strani ekipe z manjšim številom strelov vendar večjo vrednostjo xG posameznih strelov večja.

Pričakovane zadetke lahko z dejanskimi zadetki uporabimo tudi za kvantifi-

kacijo statistične vzdržnosti ali prostora za izboljšave ekipe oziroma igralca, tako v napadu kot v obrambi. V ta namen definiramo napadalni in obrambni izkoristek in iz njiju faktor izkoristka. Pri izkoristku ekipe dejansko število zadetkov predstavimo z njihovo verjetnostno porazdelitvijo na podlagi vrednosti xG. Napadalni oziroma obrambni izkoristek je definiran kot verjetnost, da pogojno na vrednosti xG igralec oziroma ekipa doseže večje ali enako število oziroma manjše ali enako število zadetkov od dejanskega števila zadetkov. Višji kot je izkoristek, manj je dejansko število doseženih oziroma prejetih zadetkov vzdržno in v nadaljevanju sezone je ob podobnih predstavah pričakovati padec števila doseženih oziroma porast prejetih zadetkov. Bližje ničli kot je izkoristek, bolj izkoriščen je napadalni oziroma obrambni potencial. Faktor izkoristka je definiran na intervalu [-1,1] in je normalizirana vrednost napadalnega oziroma obrambnega izkoristka. Pozitivna vrednost faktorja izkoristka pomeni boljši, negativna vrednost pa slabši izkoristek, kot je pričakovan glede na vrednosti xG.

Na meri pričakovanih zadetkov je zgrajen tudi koncept pričakovanih točk, ki jih označimo z xP (angl. expected points). Pričakovane točke v kontekst poskušajo postaviti verjetnost posameznega izida (zmaga, poraz, remi) in tako dodelitve točk na tekmi. Pričakovane točke, kot samo ime pove, predstavljajo pričakovano število točk ekip na tekmi in povedo, kako bi se tekma odvila na dolgi rok oziroma, kako bi se točke razdelile med ekipi na podlagi tekme z identičnimi streli na gol, vendar brez prisotnosti slučaja. Osnovni izračun pričakovanih točk temelji na večkratni simulaciji tekme z uporabo metode Monte Carlo, kjer tekmo večkrat simuliramo na podlagi vrednosti xG vsakega strela in na podlagi simuliranih rezultatov in njihovih verjetnosti oziroma deleža izračunamo pričakovane točke z upoštevanjem dejanskega pravila dodeljevanja točk. S seštevkom pričakovanih točk čez daljše obdobje (npr. celo sezono) lahko izračunamo lestvico pravičnosti in jo primerjamo z dejansko lestvico ter kvantificiramo pravičnost uvrstitve ekip.

Kot zadnje smo v nalogi predstavili napovedovanje izidov nogometnih tekem z uporabo Poissonove porazdelitve in vanj vključili vrednosti xG tako na posamični kot kumulativni ravni. Ob predpostavki, da so zadetki med seboj neodvisni, izid nogometne tekme lahko modeliramo prek dveh neodvisnih Poissonovo porazdeljenih slučajnih spremenljivk. V ta namen potrebujemo vrednosti parametrov $\lambda$, ki predstavljata oceni pričakovanega števila zadetkov ekip. V osnovni obliki v oceni

$\lambda$, ki je sestavljena iz produkta napadalnih in obrambnih moči ekip, upoštevamo zgolj število doseženih oziroma prejetih zadetkov in odigranih tekem ekip, katerih tekmo ocenjujemo, in povprečno število zadetkov upoštevanega tekmovanja. Oceno $\lambda$ lahko nadgradimo z upoštevanjem domačih in gostujočih in - če podatki to dopuščajo - poljubnimi drugimi dejstvi. Parameter $\lambda$ lahko izračunamo ročno ali pa z maksimizacijo funkcije verjetja. Model za izračun $\lambda$ lahko nadgradimo s pričakovanimi zadetki, tako da pri ročnem izračunu namesto dejanskega števila zadetkov upoštevamo kumulativno vrednost pričakovanega števila zadetkov na tekmi, pri maksimizaciji funkcije verjetja pa namesto dejanskih zadetkov upoštevamo simulirane zadetke skozi posamične vrednosti xG in njihove verjetnosti upoštevamo kot uteži.

Na podlagi uporabe mer ocenjevanja kakovosti napovedi modelov smo ugotovili, da je napovedovanje izidov nogometnih tekem zelo težavna naloga in da so razlike med modeli, ki v svojih napovedih upoštevajo dejanske zadetke, in modeli, ki upoštevajo pričakovane zadetke, majhne, vendar pričakovani zadetki, še posebej na posamični ravni, vseeno nekoliko izboljšajo napovedi števila zadetkov. Majhne razlike med modeli so tudi posledica dejstva, da dejanski zadetki na dolgi rok sledijo pričakovanim in v nekaterih primerih navidezno napihnjenim vrednostim xG. Statistika pričakovanih zadetkov nam tako služi kot dopolnilo pri preučevanju ekip in ni nujno popoln nadomestek upoštevanja zgolj dejanskih zadetkov.

Uporabljeni pristop se je izkazal kot učinkovit. V prihodnosti bi bilo smiselno razširiti analize na preučevanje dejavnikov, ki vplivajo na spremenljivke, iz katerih se izračuna vrednost xG. Lahko bi preučili tudi, kako se vrednosti xG razlikujejo med različnimi ravnmi tekmovanj. Rezultate nogometnih tekem bi lahko napovedovali tudi z modeli strojnega učenja, v katere bi vključili vrednosti xG in na tak način preverili vpliv vključitve xG na kakovost napovedi.

# Dodatek

# A Programska koda za Monte Carlo simulacijo izračuna pričakovanega števila točk

Edini vhodni podatek so vrednosti xG vsakega strela na tekmi, pogojno na ekipo. V ta namen je podan podatkovni okvir *xG_tekma* z dvema spremenljivkama, *side* in *xg*. Spremenljivka *side* zavzame vrednosti *h* (domača ekipa) in *a* (gostujoča ekipa) in označuje kateri ekipi pripada vrednost xG, ki je podana v spremenljivki *xg*.

```
# Pridobimo vrednosti xG vsakega strela pogojno na ekipo
DomaciStreli <- xG_tekma %>% filter(side == "h") %>% pull(xg)
GostujociStreli <- xG_tekma %>% filter(side == "a") %>% pull(xg)

# Pripravimo podatkovni okvir za shranjevanje rezultatov
# Za vsako simulacijo tekme shranimo stevilo zadetkov vsake ekipe
Rezultat <- data.frame(Domaci = NULL, Gosti = NULL)

# Enkratno simulacijo tekme ponovimo 10.000-krat
N <- 10000
for(i in 1:N){

  # Zacetni rezultat je 0:0
  DomaciZadetki <- 0
  GostujociZadetki <- 0

  # Vsak strel domace ekipe simuliramo na podlagi xG
  for(j in 1:length(DomaciStreli)){
    # Ce je U(0,1) < xG domacim pristejemo zadetek, sicer ne
```

```
20      if(runif(1) < DomaciStreli[j]){DomaciZadetki = DomaciZadetki
            + 1}
21    }
22
23    # Vsak strel gostujoce ekipe simuliramo na podlagi xG
24    for(j in 1:length(GostujociStreli)){
25      # Ce je U(0,1) < xG gostom pristejemo zadetek, sicer ne
26      if(runif(1) < GostujociStreli[j]){GostujociZadetki =
            GostujociZadetki + 1}
27    }
28
29    # Po simulaciji vsakega strela znotraj tekme shranimo koncni
          rezultat ene simulacije
30    Rezultat <- rbind(Rezultat, c(DomaciZadetki, GostujociZadetki))
31  }
32
33  # Iz simuliranih rezultatov izracunamo pricakovane tocke
34  xP <- Rezultat %>%
35
36    # Za vsako simulacijo tekme dolocimo zmagovalca
37    mutate(Zmagovalec = factor(ifelse(DomaciZadetki >
          GostujociZadetki, "ZmagaDomaci",
38                                      ifelse(DomaciZadetki <
                                               GostujociZadetki, "
                                               ZmagaGosti",
39                                             "Remi")),
40                              levels = c("ZmagaDomaci", "Remi", "
                                    ZmagaGosti"))) %>%
41
42    # Izracunamo verjetnost zmagovalca kot delez ustreznih tekem
43    count(Zmagovalec) %>% mutate(Verjetnost = n / N) %>%
44    # Izracunanim verjetnostnim pripisemo ustrezno stevilo tock
45    mutate(TockeDomaci = c(3,1,0), TockeGosti = c(0,1,3)) %>%
46
47    # Izracunamo pricakovane tocke kot sestevek produkta
          verjetnosti in ustreznih tock
48    # xP Domaci = P(Zmaga Domaci) * 3 + P(Remi) * 1 + P(Zmaga Gosti
          ) * 0
49    # xP Gosti = P(Zmaga Gosti) * 3 + P(Remi) * 1 + P(Zmaga Domaci)
            * 0
50    summarize(xP_Domaci = sum(Verjetnost * TockeDomaci),
51              xP_Gosti = sum(Verjetnost * TockeGosti))
```

# B Slovarček uporabljene terminologije

Spodaj so zapisani in prevedeni tuji, v delu manj pogosto uporabljeni statistični izrazi:

| **angleški pojem** | **slovenski pojem** |
|---|---|
| k-fold cross validation | prečno preverjanje s $k$ pregibi |
| bootstrap | samovzorčenje |
| classification | razvrščanje |
| ROC curve | ROC krivulja (krivulja operativne karakteristike) |
| AUC (area under the ROC curve) | območje pod krivuljo ROC |
| calibration | umerjanje |
| odds ratio | razmerje obetov |
| Brier score | Brierjev dosežek |

# Literatura


[1] Soccerment Research, "The Growing Importance of Footbal Analytics". Dosegljivo: https://soccerment.com/the-importance-of-football-analytics/. [Dostopano: 30. 6. 2022].

[2] G. Martinez Arastey, "History of Performance Analysis: The Constroversial pioneer Charles Reep". Dosegljivo: https://www.sportperformanceanalysis.com/article/history-of-performance-analysis-the-controversial-pioneer-charles-reep. [Dostopano: 30. 6. 2022].

[3] J. Sykes, N. Paine, "How One Man's Bad Math Helped Ruin Decades Of English Soccer". Dosegljivo: https://fivethirtyeight.com/features/how-one-mans-bad-math-helped-ruin-decades-of-english-soccer/. [Dostopano: 30. 6. 2022].

[4] M. Lewis, "Moneyball", 2003.

[5] The Data Visualisation Catalogue, "Radar Chart". Dosegljivo: https://datavizcatalogue.com/methods/radar_chart.html. [Dostopano: 30. 6. 2022].

[6] Total Footbal Italia, "L'indice Di Pericolosità: La Prestazione Oggettiva". Dosegljivo: https://totalfootballitalia.com/lindice-di-pericolosita-andare-oltre-al-risultato/l. [Dostopano: 30. 6. 2022].

[7] N. Walsh, "Radar Chart". Dosegljivo: https://realsport101.com/fifa/fifa-22-heat-maps-expected-goals-xg-new-features-more/. [Dostopano: 30. 6. 2022].

[8] Premier League. Dosegljivo: https://www.premierleague.com/. [Dostopano: 30. 6. 2022].

[9] Betting Offers, "What is the Average Number of Goals Scored Per Game in Football?". Dosegljivo: https://www.bettingoffers.org.uk/football/what-is-the-average-number-of-goals-scored-per-game-in-football/#:~:text=In%20the%20five%20seasons%20between,1.51%20in%20the%20second%20half. [Dostopano: 30. 6. 2022].

[10] FBbref. Dosegljivo: https://fbref.com/en/. [Dostopano: 30. 6. 2022].

[11] S. Gregory, "Expected Goals in context". Dosegljivo: https://www.statsperform.com/resource/expected-goals-in-context/. [Dostopano: 30. 6. 2022].

[12] J. Whitmore, "What Are Expected Goals (xG)?". Dosegljivo: https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/. [Dostopano: 30. 6. 2022].

[13] InStat, "Massive research of Penalties by InStat". Dosegljivo: https://instatsport.com/football/article/penalty_research. [Dostopano: 30. 6. 2022]

[14] Understat. Dosegljivo: https://understat.com/. [Dostopano: 30. 6. 2022].

[15] Infogol. Dosegljivo: https://www.infogol.net/en. [Dostopano: 30. 6. 2022].

[16] I. Dragulet, "An Exploration of Expected Goals". Dosegljivo: https://towardsdatascience.com/a-guide-to-expected-goals-63925ee71064. [Dostopano: 30. 6. 2022].

[17] StatsBomb. Dosegljivo: https://statsbomb.com/. [Dostopano: 30. 6. 2022].

[18] F. E. Harrell, "*Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis*". New York: Springer, 2001.

[19] A. Fairchild, K. Pelechrinis, M. Kokkodis, "Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality". *Journal of Sports Analytics*, vol. 4, no. 3., str. 165-174, 2018.

[20] FS, "Expected goals and conversion - a probabilistic approach". Dosegljivo: https://fblsim.medium.com/expected-goals-and-conversion-a-probabilistic-approach-3eea0ff87077. [Dostopano: 30. 6. 2022].

[21] M. Taylor, "Twelve Shots Good, Two Shots Better". Dosegljivo: http://thepowerofgoals.blogspot.com/2014/02/twelve-shots-good-two-shots-better.html. [Dostopano: 30. 6. 2022].

[22] S. Khatri, "Attempting to Quantify Team xG Under/OverPerformance Uniformly: Introducing Overperformance Factor". Dosegljivo: https://shreyaskhatri.medium.com/attempting-to-quantify-team-xg-under\protect\@normalcr\relax-overperformance-uniformly-overperformance-factor-3798e5a3de2a. [Dostopano: 30. 6. 2022].

[23] B. Torvaney (GitHub), "Premier League 2014/15 - 2017/18 xG per shot". Dosegljivo: https://gist.github.com/Torvaney, https://gist.githubusercontent.com/Torvaney/507440f0fd004f6beba5d270b07b3e80/raw/2e58a72f654f73cb77d7f1e72a70d29e00f2fa2d/premier_league_xg.csv. [Dostopano: 30. 6. 2022].

[24] A. Eadson, "What are Expected Points (xP) and how to use them in football betting". Dosegljivo: https://www.bettingodds.com/news/what-are-expected-points-xp-football-betting. [Dostopano: 30. 6. 2022].

[25] Tactics no tantics. "Expected Points Models – Part I". Dosegljivo: https://tacticsnotantics.org/statistical-models-and-analyses/expected-points-models-part-i/. [Dostopano: 30. 6. 2022].

[26] M. J. Dixon, S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market.". *Applied Statistics*, vol. 46, no. 2, str. 265-280, 1997. [Dostopano: 30. 6. 2022].

[27] P. Winchester, "What's the Score?". Dosegljivo: https://philipwinchester.github.io/whats-the-score/. [Dostopano: 30. 6. 2022].

[28] Smarkets. "How to calculate Poisson distribution for football betting". Dosegljivo: https://help.smarkets.com/hc/en-gb/articles/115001457989-How-to-calculate-Poisson-distribution-for-football-betting. [Dostopano: 30. 6. 2022].

[29] P. Winchester, "Dixon Coles Model". Dosegljivo: https://philipwinchester.github.io/dixon-coles-model/. [Dostopano: 30. 6. 2022].

[30] Stats and snakeoil, "Dixon Coles". Dosegljivo: https://www.statsandsnakeoil.com/tags/dixon-coles/. [Dostopano: 30. 6. 2022].

[31] octosport.io, "Football Prediction Performance: How to Calculate Hit-ratio and Log-loss". Dosegljivo: https://medium.com/geekculture/football-prediction-performance-how-to-calculate-hit-ratio-and-\protect\@normalcr\relaxlog-loss-1e5e22310497. [Dostopano: 30. 6. 2022].

[32] BEATTHEBOOKIE2017, "Define variables: Brier score for market odds". Dosegljivo: https://beatthebookie.blog/2017/08/08/define-variables-brier-score-for-market-odds/#:~:text=The%20Brier%20score%20is%20a,feature%20for%20a%20predictive%20model. [Dostopano: 30. 6. 2022].

[33] Football-data. Dosegljivo: https://www.football-data.co.uk/. [Dostopano: 30. 6. 2022].

[34] Online Sports Betting, "Soccer Betting Odds". Dosegljivo: https://www.onlinesportsbetting.net/soccer/odds.html#. [Dostopano: 30. 6. 2022].