VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS

Sporto prognozių modelių charakteristikų tyrimas

Sport predictions models characteristics research

Magistro baigiamasis darbas

Atliko:	Edgaras Karka	(parašas)
Darbo vadovas:	Linas Petkevičius	(parašas)
Recenzentas:	doc. dr. Kristina Lapin	(parašas)

Vilnius – 2018

TURINYS

Įvadas

Temos aktualumas

Dėl savo dinaminės prigimties, prognozavimo uždaviniai yra vieni iš sunkiausiai išsprendžiamų. Dažniausiai tokių uždavinių rezultatams turi įtakos daugybė skirtingų faktorių, todėl norint gauti tikslius rezultatus reikia įvertinti skirtingų kintamųjų įtaką prognozuojamo įvykio baigčiai.

Futbolas yra vienas populiariausių komandinių žaidimų pasaulyje. Kiekvieną dieną milijonai žmonių žiūri futbolo varžybas, diskutuoja apie rezultatus. Taip pat pastaruoju metu tarp futbolo aistruolių ypač populiaru prognozuoti artėjančias varžybas, jų eigą bei būsimus rezultatus. Tokį elgesį gali motyvuoti noras turėti gerą laiką arba tikintis laimėti pinigų lažybose.

Šuo metu stipriai populiarėjanti grandinių technologija (*angl. block-chain*), iš esmės keičia lažybose naudojamas technologijas (*angl. bettech*)[FT15] [JCH+16]. Decentralizuotos lažybų platformos leis nevaržomai naudotis lažybų paslaugomis. Tačiau šioms platformoms bus reikalingos įvykių prognozavimo paslaugos (*angl. oracles*), gebančios gerai nustatyti įvykių baigčių tikimybes. Tokiose platformose teoriškai įmanoma vykdyti lažybas automatizuotai, o tam reikalingi geri baigčių prognozavimo modeliai ir jų programų sistemos.

2018 metais internetinių lažybų rinka siekė 51 mlrd. JAV dolerių, o 2020 prognozuojama, kad ji sieks 60 mlrd. JAV dolerių. Mokslininkai Volfgangas Breueris, Guidonas Hautenas, Klaudia Kreuz iš Vokietijos Acheno universiteto mano, kad sporto lažybos gali būti alternatyva tokiems investavimo instrumentams, kaip investavimas į akcijas [BHK09]. Didėjant kompiuterių skaičiavimų galingumams, bei augant informacijos pasiekiamumui, tiesioginiu laiku galime stebėti augantį įvairių futbolo prognozių modelių kiekį. Tačiau norint lažybas panaudoti kaip investavimo instrumentą, reikalingos pelningos lažybų strategijos, o pelningos lažybų strategijos pagrindas yra tikslus prognozių modelis.

Didelę įtaką sėkmingam investavimui turi turimo portfelio valdymas, tai ne išimtis ir sporto lažybose. Viena iš pagrindinių problemų lažybose yra rasti pakankamai užtikrintas varžybų baigtis. Su panašia problema susiduria ir investuotojai į akcijas, kai reikia pasirinkti galimos grąžos ir rizikos santykį. Ekonomistai atliko daug tyrimų siekiant atrasti tinkamiausias pinigų valdymo strategijas [Bro99]. Šios pinigų valdymo strategijos gali būti pritaikomos ir sporto lažybose [Tho06].

Egzistuojantys sporto prognozių statistiniai matematiniai modeliai dažniausiai yra išbandyti su įvairiomis sporto rūšimis, neatsižvelgiant į sporto lažybų agentūrų siūlomus varžybų baigčių tikimybių įvertinimus. Atsiradus galimybei realiu laiku gauti bei apdoroti didelius kiekius informacijos, galime įvertinti modelių rezultatus, bei jų tinkamumą kuriant alternatyvius investavimo įrankius.

Pastaruoju metu stipriai populiarėjant informacinėms technologijoms, galime stebėti daugėjant įvairių tyrimų, susijusių su sporto varžybų baigčių prognozavimu. Tokie tyrimai labai aktualūs lažybų agentūroms ir asmenims besidomintiems sporto baigčių prognozavimu. Sporto varžybų baigčių prognozavimo modelio įgyvendinimui ir vertinimui yra svarbu naudojami programiniai įrankiai ir duomenys. Prognozių vartotojams labai svarbu patogiai ir laiku gauti patikimą informa-

ciją. Atsižvelgiant į tai, suprojektuotas ir įgyvendintas programų sistemos prototipas, leidžiantis modelių kūrėjams nesirūpinti dėl sporto duomenų ir vertinimo įrankių.

Temos naujumas

Internetas, kompiuteriai, dirbtinis intelektas, grandinių technologijos (*angl. blockchain*) - tampa mūsų kasdienybė, todėl neišvengiamai šios technologijos skverbiasi į visas gyvenimo sritis. Ne išimtis yra ir sporto pasaulis bei sporto prognozių rinka.

Šiuo metu lažybos keliasi ir į virtualia erdvę. Kuriamos įvairios sporto išsivaizduojamos lygos (*angl. fantazy sports*) ir elektroninio sporto (*angl. eSports*) turnyrai. Šiose lygose ir turnyruose vyksta virtualios varžybos su galimybe lažintis už įvairių varžybų baigtis. [Sch15].

Šiuo metu vyksta sistemų, veikiančių grandinių technologijų pagrindu (*angl. blockchain*), revoliucija [Vog15]. Tai keičia lažybose naudojamų informacinių technologijų pasaulį *(angl. bet-tech)*.

Išpopuliarėjus internete veikiančioms sporto prognozių brokerių agentūroms, atsirado naujų galimybių norintiems uždirbti iš sporto prognozių. Norint tai pasiekti, reikalinga pelninga lažybų strategija. Pelningos lažybų strategijos pagrindas yra sporto varžybų baigčių prognozavimo modelis, leidžiantis atsirinkti tikslias varžybų baigčių prognozes. Paplitus socialinių tinklų platformoms atsirado galimybė kurti tikslesnius modelius, pasinaudojant šių tinklų kaupiamais duomenimis. [SJL16].

Šiuolaikinės technologijos leidžia atlikti sporto prognozių modelių analizes, taip pat įvertinti labiausiai tinkančius pinigų valdymo mechanizmus. Radus tikslius sporto baigčių prognozavimo modelius ir įgyvendinus automatizuotą sporto varžybų baigčių prognozavimo sistemą, galėtume automatizuoti lažybų procesą. Tai ir yra šio darbo tikslas. Pilnai įgyvendinta sėkminga tokio tipo sistema galėtų būti alternatyva standartiniams investavimo instrumentams.

Darbo tikslas

Atlikti futbolo varžybų prognozių modelių tyrimą, taip sukuriant futbolo varžybų rezultatų automatizuotos prognozavimo sistemos matematinį ir kompiuterinį modelį.

Uždaviniai

1. Atlikti literatūros analizę siekiant išsiaiškinti tinkamiausius sporto varžybų baigčių prognozavimo modelius;
2. Atlikti sporto varžybų baigčių prognozavimo modelių analizę;
3. Įvertinti lažybininkų gebėjimą prognozuoti skirtingų futbolo lygų varžybų baigtis;
4. Suprojektuoti ir įgyvendinti automatizuotą sporto įvykių baigčių prognozavimo informacinę sistemą.

5. Pagal gautus rezultatus įvertinti modelių galimybes juos naudoti automatizuojant lažinimosi procesą.

1 Literatūros apžvalga

1.1 Modelių apžvalga

Egzistuoja įvairių modelių, skirtų prognozuoti sporto varžybų rezultatą ar jų baigtį. Juos galima skirstyti į rūšis pagal naudojamus deterministinius ir statistinius metodus, intensyvumo parametrų tipus ar jų kiekį, apskaičiuotų prognozių tipą, naudojamus duomenis ir tikslumo ar pelningumo rodiklius.

1.2 Puasono pasiskirstymo dėsniu paremti modeliai

Puasono pasiskirstymo dėsnio pagrindu sudaryti modeliai, dažniausiai literatūroje sutinkamas matematinis modelis, naudojamas sporto varžybų rezultatų prognozavimo uždaviniuose. Šį modelį panaudojo ir aprašė Maheris dar 1982 metais [Mah82]. Modelio pagrindą sudaro Puasono pasiskirstymo dėsnis. Modelio intensyvumas priklauso nuo penkių parametrų: namų komandos puolimo stiprio koeficiento α, svečių komandos gynybos stiprio koeficiento β, svečių komandos puolimo stiprio koeficiento γ, namų komandos gynybos stiprio koeficiento δ ir namų komandos pranašumo koeficiento k.

Maheris savo Puasono pasiskirstymo dėsniu paremtą modelį įvertino panaudojęs 1973-1975 sezonų, Anglijos futbolo lygų varžybų rezultatus. Skaičiavimų pagrindą sudarė didžiausio tikimumo metodas *(angl. maximum likelihood)*. Jo pagalba autorius savo darbe įvertino komandų puolimo ir gynybos parametrus. Matheris palygino parametrų įtaką modelių sporto prognozių rezultatams, įtraukdamas į vertinamas modelių variacijas skirtingus parametrus.

Modelis 0 $\quad \beta_i = \alpha, \beta_i = \beta, \gamma_i = \gamma; \sum_i \alpha_i = \sum_i \beta_i$

Modelis 1A $\quad \beta_i = \alpha_i, \beta_i = \beta, \gamma_i = \gamma; \sum_i \alpha_i = \sum_i \beta_i$

Modelis 1B $\quad \alpha_i = \alpha, \beta_i = \beta; \sum_i \alpha_i = \sum_i \beta_i$

Modelis 2A $\quad \delta_i = k\alpha, \gamma_i = k\delta_i; \sum_i \alpha_i = \sum_i \beta_i$

Modelis 3A $\quad \delta_i = \alpha_i; \sum_i \alpha_i = \sum_i \beta_i$

Modelis 3B $\quad \gamma_i = \beta_i; \sum_i \alpha_i = \sum \beta_i$

Modelis 4 $\quad \sum_i \alpha_i = \sum \beta_i; \sum_i \gamma_i = \sum_i \beta_i$

Čia i-itoji komanda. Modelyje 0 naudojami vienodi parametrai. Modelių hierarchijos viršuje esantis modelis 4 naudojamas su skirtingais komandų parametrais.

Įvertinęs modelių variacijas Matheris pastebėjo, jog namų komandos puolimo α ir svečių komandos gynybos β parametrų įvertinimai kritiškai svarbūs modelio prognozių rezultatams, tačiau namų komandos gynybos δ ir svečių komandos puolimo γ parametrai neturi didelės įtakos įvertintoms tikimybėms. Įvertinus visus modelius pastebėta, jog modelis 2A prognozuoja tiksliausiai. 2A modelyje naudojami visi komandų stiprių parametrai.

Matheris, įvertinęs modelį (1. lentelė) pasirinktiems duomenų rinkiniams, pastebėjo mažus, tačiau sisteminius skirtumus. Pastebėta, kad modelis ne pakankamai įvertina mažo rezultatyvumo (0-1, 1-0, 1-1, 2-0, 0-2) ir pervertino aukšto rezultatyvumo (>4 įvarčiai) varžybų rezultatų tikimybes. Gauti rezultatai parodė, jog varžybų įvarčių pasiskirstymas artimas Puasono tikimybių pasiskirstymui. Matherio tyrimo apibendrinimai nedaug skyrėsi nuo Morneys ir Benjamino [Kla62] teiginių. Šių mokslininkų darbai pagrinde skiriasi nuo Matherio tuo, kad Matheris sudarinėjo intensyvumo parametrus kiekvienoms varžyboms atskirai, o Mornio ir Benjamino sudarytas modelis taikomas visoms komandoms vienodai.

Lentelė 1. Matherio pasiūlyto modelio namų ir svečių komandų pelnytų įvarčių skaičiaus prognozės ir realūs rezultatai.

Namų komandos rezultatai					
Įvarčių skaičius	0	1	2	3	>4
Rezultatas	0.217	0.321	0.254	0.130	0.078
Prognozė	0.230	0.318	0.238	0.128	0.086
Svečių komandos rezultatai					
Rezultatas	0.388	0.371	0.177	0.051	0.014
Prognozė	0.406	0.352	0.166	0.056	0.020

Matheris atkreipė dėmesį, jog komandų rezultatai nepriklausomi. Pavyzdžiui baigiantis rungtynėms pralaiminti komanda turi rizikuoti ir daugiau atakuoti taip susilpninant gynybą, o dėlto padidėja tikimybė, jog bus pelnytas įvartis. Remiantis tokia logika Matheris panaudojo dvimatį Puasono pasiskirstymo dėsnį, į modelį įtraukdamas $L_{ij} = X_{ij} - Y_{ij}$. Matherio atliktame moksliniame darbe ištirta dviejų tipų Puasono tikimybių pasiskirstymo pagrindu sudaryti prognozavimo modeliai. Pirmajame daroma prielaida, kad namų ir svečių komandų rezultatai yra nepriklausomi, o jiems prognozuoti naudojamas nepriklausomas Puasono pasiskirstymo dėsnis su parametrais $\alpha_{ij}\beta_{ij}$ ir $\gamma_{ij}\delta_{ij}$. Čia α_i ir β_j atitinka namų komandos puolimo stiprį ir svečių komandos gynybos pajėgumus, o γ ir δ intensyvumo parametrai atitinka namų ir svečių komandų puolimo ir gynybos stiprius.

Didžiausio tikėtinumo metodu *(angl. maximum likelihood estimation)* parodė, jog tik α ir β parametrai reikalingi norint pakankamai tiksliai prognozuoti futbolo varžybų įvarčių kiekį. Šio modelio bandymai, su 24 skirtingų duomenų rinkiniais, parodė, jog 19-ios duomenų rinkinių prognozuojami įvarčių skaičiai nesiskyrė daugiau nei 5% nuo tikrojo komandų pelnytų įvarčių skaičiaus. Iš tokių rezultatų galima teigti, jog Puasono pasiskirstymo dėsnis gerai susitvarko su tokio tipo uždaviniais. Šio modelio nukrypimai nedideli ir sistemingi visiems duomenų rinkiniams. Antroje tyrimo dalyje Matheris panaudojo dvimatį Puosono pasiskirstymo dėsnį. Kaip matome iš 2 lentelės, dvimatis modelio variantas rezultatus prognozavo tiksliau.

Diksonas ir Kolis 1997 metais, panaudoję Anglijos lygos futbolo duomenimis nuo 1992 iki 1995 metų, sukūrė ir išbandė matematinį futbolo varžybų rezultatų prognozavimo modelį [DC97]. Modelis sukurtas tikintis aptikti lažybų rinkos siūlomų varžybų baigčių koeficientų neatitikimus

Lentelė 2. Prognozuoti ir išmatuoti pelnytų įvarčių skirtumų dažniai, naudojant nepriklausomą ir dvimatį Puasono pasiskirstymo dėsnį.

Z	<=-3	-2	-1	0	1	2	3	4	>=5
Rezultatas	8	26	72	129	105	69	31	16	6
Prognozuota $\varrho = 0$	14.4	30.3	69.8	113.0	104.9	68.7	35.8	15.8	9.3
Prognozuota $\varrho = 0.2$	9.9	25.3	68.0	126.6	111.7	67.7	32.6	13.4	7.1

tarp 1995 ir 1996 metų lažybų pasiūlos. Modelio pagrindas didžiausio tikėtinumo metodas, bei Puasono pasiskirstymo dėsnis. Autoriai savo darbe susidūrė su duomenų struktūros ir komandų rodyklių dinamiškumo problemomis. Diksonas ir Kolis, tikėdamiesi tikslesnių tikimybių įvertinimo, praplėtė Maherio siūlytą modelį. Autoriai sukurtame prognozavimo modelyje įtraukė įvairiais modifikacijas, leidžiančias prognozuoti futbolo varžybas, kai susitinka komandos iš skirtingų lygų. Rezultate, jiems pavyko sukurti pelningą lažybų strategiją, naudojant modifikuotą Matherio pasiūlytą sporto varžybų rezultatų prognozavimo modelį.

Panašios tematikos ankstesni darbai orientavosi ne į konkrečių varžybų rezultato prognozavimą o į rezultatų pasiskirstymą. Tokio tipo uždavinį ištyrė Mornis dar 1954 metais [TM54]. Jis padarė išvadą, jog nors Puasono pasiskirstymo dėsnis puikiai tinka įvertinant futbolo varžybų rezultatų pasiskirstymą, geresnių rezultatų galima tikėtis naudojant neigiamą binominį pasiskirstymo dėsnį. Hilas 1974 metais palygino futbolo ekspertų prognozes galutinei turnyrinei lentelei [Hil74]. Gauti rezultatai leido padaryti reikšmingų išvadų. Literatūroje galima aptikti šiek tiek skirtingo pobūdžio darbų. Šie darbai tiria specifinius varžybų aspektus. Įvairių darbų autoriai bandė įvertinti, kaip žaidėjo išsiuntimas iš aikštelės įtakoja rezultatą. Iš šių darbų galime daryti išvadą, jog nėra sudėtinga prognozuoti komandų pozicijas turnyrinėje lentelėje ar įvarčių dažnių pasiskirstymą sezono pabaigoje. Visai kas kita įvertinti konkrečių varžybų rezultatų tikimybes.

Su lažybų strategijomis susijusių darbų galime aptikti ekonomikai skirtoje literatūroje. Daug panašių darbų yra atlikta išskirtinai amerikietiškajam futbolui ar žirgų lenktinėms, tačiau ne daugelyje iš jų rasime statistinių metodų panaudojimą. Taip pat mokslininkų darbuose randame diskusijų apie įvairias lažybų ar turimų pinigų valdymo strategijas.

Individualios komandos pasirodymas varžybų metu gali būti įtakotas daugybės išorinių veiksnių. Pavyzdžiui naujo žaidėjo prisijungimas prie komandos ar trenerio atleidimas ir pan.. Nors šią informaciją galime pasiekti viešai, tačiau ją sudėtinga panaudoti prognozavimo modeliuose. Šią informaciją, dėl duomenų subjektyvumo, sunku formalizuoti ir strukturizuoti. Dėl to Diksono ir Koli aprašytame prognozavimo modelyje panaudota tik istorinė (3 metų) varžybų rezultatų statistinė informacija. Nors autoriai ir nepanaudojo kitų rūšių informacijos, apžvelgtame darbe rasime šiai temai skirtą skyrelį [DC97].

Diksono ir Koli tyrime buvo naudojami duomenys iš keturių sezonų varžybų. Šių varžybų baigčių santykis yra 46:27:27 atitinkamai, namų komandos pergalė, lygiosiomis ir svečių pergale. Šis dažnis parodo atsitiktinių varžybų baigties tikimybes. Pavyzdžiui, tikimybė, kad atsitiktinai parinktos varžybos baigsis namų komandos pergalę lygi 46%. Dėl naudotos duomenų bazės didžio,

gautos tikimybės gan tiksliai atitinka atsitiktinai parinktų varžybų baigčių dažnį. Diksono ir Koli pagrindinis tikslas buvo sukurti modelį gebanti prognozuoti varžybų baigtis įvertinant įvairias komandų charakteristikas.

Iš gautų rezultatų (3 lentelė) galime teigti, jog Puasono pasiskirstymo dėsnis gerai įvertina futbolo varžybų rezultatų pasiskirstymą. Tokie rezultatai darbo autoriams leido pasitikėti nepriklausomu Puasono pasiskirstymo dėsniu.

Lentelė 3. Dixon ir Coles sukurto modelio varžybų rezultatų tikimybių ir jų rezultatų matrica.

Namų įvarčiai	Išmatuotos rezultatų tikimybės					
	Svečių	0	1	2	3	4
		33.4(0.74)	36.4(0.57)	19.5(0.49)	7.9(0.42)	2.1(0.16)
0	22.1(0.36)	8.2(0.32)	7.4(0.28)	4.5(0.23)	1.4(0.13)	0.4(0.06)
1	33(0.65)	10.3(0.38)	12.7(0.30)	6.4(0.24)	2.7(0.15)	0.6(0.07)
2	24.5(0.51)	8.2(0.31)	9.1(0.25)	4.8(0.22)	1.9(0.14)	0.5(0.09)
3	12.6(0.4)	4.2(0.25)	4.5(0.25)	2.3(0.19)	1.2(0.11)	0.4(0.06)
4	5.3(0.1)	1.6(0.14)	1.8(0.13)	1.1(0.13)	0.06(0.07)	0.1(0.04)

Matheris darė prielaidą, kad per varžybas pelnyti namų ir svečių komandų įvarčiai yra nepriklausomi Puasono kintamieji. Šie parametrai atitinka komandų gynybos ir puolimo kokybę [Mah82]. Tiksliau, varžybose, tarp dviejų komandų i ir j, X_{ij} ir Y_{ij} - namų ir svečių komandų įvarčių skaičius.

Taigi

$$X_{ij} \sim \mathcal{P}(\alpha_i \beta_j \gamma),$$
$$Y_{ij} \sim \mathcal{P}(\gamma_i \delta_j)$$

,kur $X_i j$ ir $Y_i j$ yra nepriklausomi, α_i, $\beta_j > 0$. $\forall i$, α_i aprašo namų komandos puolimo stiprį, β_j svečių komandos gynybos stiprį ir $\gamma > 0$ namuose žaidžiančios komandos pranašumą.

Diksonas ir Koli 1997 metais išsikėlė tikslą patikrinti modelio prognozuojamų futbolo varžybų rezultatų tikimybes su lažybininkų siūlomomis [DC97]. Prognozavimo modelis sudarytas naudojant dvimatį Puasono pasiskirstymo dėsnį. Taip įtraukiant komandų pelnytų įvarčių skaičius, bei parametrus nurodančius ankstesnių varžybų rezultatus. Diksonas ir Koli naudota lažybų strategija gana paprasta: jie lažinosi už varžybų baigtis pagal tikimybių reikšmes. Diksonas ir Koli [DC97] pasiūlė praplėsti Matherio [Mah82] modelį įtraukiant priklausomybės parametrą ρ.

$$P(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x,y) \frac{\lambda^x exp(-\lambda)}{x!} \frac{\mu^y exp(-\mu)}{y!}$$

,kur

$$\lambda = \alpha_i \beta_j \gamma,$$
$$\mu = \alpha_i \beta_j,$$

ir

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1-\lambda\mu\rho; & \text{Jei } x=y=0 \\ 1+\lambda\rho; & \text{Jei } x=0, y=1 \\ 1+\lambda\mu; & \text{Jei } x=1, y=0, \\ 1-\rho; & \text{Jei } x=1, y=1 \\ 1; & \text{kita} \end{cases}$$

Šiame modelyje ρ atitinka:

$$max(-1/\lambda, -1/\mu) \leqslant \rho \leqslant, min(-1/\lambda\mu, 1)$$

Šis parametras pakoreguoja Matherio [Mah82] pasiūlyto modelio tikimybes rezultatams: 0-0, 0-1, 1-0 ir 1-1.

Naudojant Diksono ir Koli aprašytą prognozių modelį reikia įvertinti $\{\alpha_1 ... \alpha_n\}$ namų komandos atakos parametrus, $\{\beta_1 ... \beta_n\}$ svečių komandos gynybos stiprio parametrus, ρ priklausomybės parametrą, ir namų komandos pranašumo parametrą. Darbo autoriai norėdami apsaugoti nuo per didelio parametrų skaičiaus įtraukė suvaržimus:

$$n^{-1} \sum_{n=1}^{n} \alpha_i = 1$$

Diksonas ir Koli [DC97] tirdami prognozavimo modelį naudojo futbolo varžybas iš Anglijos priemier bei 1-3 diviziono lygų. Taigi jų modelyje $n = 92$, o visas parametrų skaičius 185. Darbe autoriai tikimybių vertinimui naudojo didžiausio tikėtinumo funkciją, čia $k = 1...n$ komandų indeksai, o (X_k, Y_k) prognozuojamas rezultatas:

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, .., n) = \prod_{k=1}^{N} \{\tau_{\lambda_k,\mu_k}(x_k, y_k) exp(-\lambda_k) \lambda_k^{x_k} exp(-\mu_k) \mu_k^{y_k}\}$$

,kur

$$\lambda = \alpha_i \beta_j \gamma,$$
$$\mu = \alpha_i \beta_j$$

Tačiau ši modelio modifikacija turi struktūrinių trukumų. Prognozavimo modelyje naudojami statiniai parametrai. Komandų puolimo α ir gynybos β stipriai yra konstantos ir laike nekinta. Realybėje komandų žaidimo kokybė yra dinaminė ir skirtingu metu yra skirtinga, todėl logiška, kad tai turėtu būti įtraukta sudarant prognozavimo modelį.

Komandų žaidimo kokybė yra susijusi su paskutinėmis varžybomis. Taigi Diksonas ir Koli [DC97] darė prielaida, kad naujesnė informacija yra naudingesnė, nei senesnė. Tai autoriai panaudojo praplėsdami modelio modifikaciją panaudodami laiko parametrą t:

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, .., n) = \prod_{k \in A_t} \{\tau_{\lambda_k,\mu_k}(x_k, y_k) exp(-\lambda_k) \lambda_k^{x_k} exp(-\mu_k) \mu_k^{y_k}\}^{\phi(t-t_k)},$$

,kur t_k komandos k žaistų varžybų laiko momentas t , $A_t = \{k : t_k < t\}$. t turi tokia pačią reikšmę kaip ir ankstesnėje didžiausio tikėtinumo funkcijos modifikacijoje. ϕ - laiko parametro įtakos funkcija.

Diksonas ir Kolis [DC97] naudojo keletą skirtingų laiko įtakos funkcijos reikšmių. Viena iš jų:

$$\phi(t) = \begin{cases} 0 & t > t_0 \\ 1 & t \leqslant t_0 \end{cases}$$

Šiuo atveju, jei varžybos žaistos anksčiau nei laiko momentas t_0 , varžybos neįtraukiamos į prognozių skaičiavimus. Savo darbe Diksonas ir Kolis [DC97] varžybų žaidimo laiko įtakai įvertinti naudojo funkciją:

$$\phi(t) = exp(-\xi t)$$

Naudojant tokią laiko parametro funkciją, varžybų įtaka mažėja eksponentiškai. Esant didesnei ξ reikšmei, didesnę įtaką turi naujausi rezultatai. Tai gi statiniame modelyje ši reikšmė būtų lygi 0.

Diksonas ir Kolis [DC97] daugiau dėmesio skyrė prognozuodami varžybų baigtis, o ne varžybose pelnytų įvarčių skaičių. Taip pat jie daug dėmesio skyrė tinkamo xi parametro didžio parinkimui. Autoriai norėdami įvertinti namų komandos pergalės tikimybę naudojo:

$$p_x^H = \sum_{l,m \in B_H} Pr(X_k = l, Y_k = m)$$

,kur $B_H = \{(l, m) : l > m\}$. Rezultatų tikimybės vertinamos naudojant didžiausią tikėtinumo metodą. Panašiai autoriai vertino svečių komandos pergalės p_k^A bei lygiųjų p_k^D tikimybes. Autoriai ištyrė, jog nuo laiko priklausanti didžiausio tikėtumo funkcija geriausia maksimizuojama, kai $\xi = 0.065$.

Martinas Krovderis ir Markas Diksonas savo darbe panaudojo 92 Anglijos lygų komandų varžybas nuo 1992 iki 1997 metų. Futbolo varžybų baigčių tikimybių įvertinimui jie naudojo modifikuotą Diksono ir Kolio [DC97] darbe naudotą prognozavimo modelį. Pagrindinis modifikacijos tikslas supaprastinti prognozių skaičiavimus. Šiuose modeliuose atsisakyta skirstymo į namų bei svečių komandų gynybos ir puolimo parametrus. Darbuose naudojama taškų sistemos, leidžiančios įvertinti parametrų reikšmingumą. Šiam tikslui Koning panaudojo didžiausio tikėtumo metodą, o Knoras Heldas [KR00] naudojo išplėtotąjį Kalmano filtrą kartu su *ad hoc* metodu. Knoras-Heldas [KR00] modifikavo Diksono ir Kolio [DC97] Puasono pasiskirstymo dėsnio pagrindu veikiantį futbolo varžybų baigčių prognozavimo modelį taip, jog eliminuotu rezultatus aukštesnius už 5. Pavyzdžiui esant rezultatui 7-6, autorių naudotame modelyje rezultatas traktuojamas kaip 5-5.

Šie darbų rezultatai parodė, kad Matherio [Mah82] pasiūlytas Puasono pasiskirstymo dėsniu paremto modelio variantas gali būti optimizuojamas, norint sumažinti futbolo varžybų baigčių prognozių skaičiavimus, stipriai neprarandant prognozių tikslumo.

Karlis ir Ntzoufras [KN03] pastebėjo, kad futbolo varžybos baigiasi lygiosiomis dažniau nei tai prognozuoja nepriklausomų parametrų Puasono pasiskirstymo dėsniu paremtas modelis. Todėl savo darbe jie siūlė į prognozavimo modelį įtraukti parametrą λ. Šis parametras padidina lygiųjų prognozės tikimybę varžybose. Be to jie siūlė naudoti dvimatį Puasono pasiskirstymo dėsnį *(angl. bivariate Poisson distribution)* vietoj nepriklausomo, *(angl. independent Poisson distribution)*. Darbe teigiama jog ir nedidele λ reikšmė padidina lygiųjų tikimybę (pvz. $\lambda 3 = 0.05$ ir $\lambda = 0.2$ lygiųjų tikimybę padidina atitinkamai 3.3 % ir 14%).

2015 metais dvimatį Puasono pasiskirstymo dėsnį *Anglijos premier* lygos futbolo varžybų rezultatų analizei ir prognozavimui panaudojo Simas Janas Kopmenas ir Rutgeris Litas [KL15]. Jie savo darbe naudojo laike kintančius pasiskirstymo dėsnio intensyvumo parametrus. Modelis įgyvendintas remiantis būsenų erdvės *(angl. State spaces)* ir *(angl. importance sampling)* metodais.

Apie futbolo komandų gynybos ir puolimo stiprio kitimą laikę jau buvo kalbėta ir Maherio [Mah82] darbe, tačiau daugiau rezultatų, parodonačių kaip prognozavimo modelio kokybę įtakoja laiko parametras savo darbe ištyrę Diksonas ir Kolis [DC97].

1.3 Robastiškumas

Dimitris Carlis ir Ioannis Ntzoufras [KN03] savo darbe bandė įgyvendinti metodą leidžiantį įvertinti prognozavimo modelyje naudotų rezultatų įtaką. Darbo motyvas buvo tai, jog dauguma apžvelgtų modelių į savo struktūrą neįtraukia parametrų robastiškumo t.y. netikslūs ankstesni rezultatai gali neigiamai paveikti ateities prognozių tikimybių tikslumą. Autoriai modelio parametrų svorius skirstė į dvi rūšis: fiksuoto didžio svoriai ir pagal naudojamą prognozavimo modelį pritaikyti svoriai. Pavyzdžiui su fiksuotais svoriais, galime sumažinti tam tikrų rezultatų įtaką (pvz. rezultatai 3-0 ir 4-0 suteikia beveik vienodą informaciją apie komandų puolimo ir gynybos kokybę). Modeliais paremti naudojamų parametrų svorių nustatymai leidžia sumažinti praeities mažų tikimybių rezultatų įtaką modelio vertinamoms varžybų baigčių tikimybėms. Autoriai norėdami tai įgyvendinti pritaikė Vindhamo (1995) metodą. Šis metodas skirtas pagerinti statistinio modelio robastiškumą. Darbe panaudoti duomenys iš UEFA čempionų lygos 2008-2009 metų sezono. Kiekvieną grupę sudarė keturios, pagal *round robin* schemą (kai kiekviena komanda susitinka su kita po du kartus, vieną kartą namuose kitą kartą išvykoje), tarpusavyje besivaržančios komandos.

Lentelė 4. GF - įspirti įvarčiai. GA - praleisti įvarčiai. Simuliacijos rezultatai ir skliaustuose tikrieji rezultatai.

Komandos	Vid.	Vid.	Vid.	Tikimybės (%)			
	Taškai	GF	GA	1	2	2.5-3	3.5-4
AS Roma	11.8 (12)	11.9	6.0	49.2	87.1	10.4	2.5
Chelsia	11.1 (11)	9.0	5.1	36.9	82.7	13.7	3.6
Bordeaux	4.7 (7)	5.0	11.0	1.0	7.4	33.0	59.6
CFR 1907 Cluj	5.6(4)	5.0	8.9	1.8	13.5	42.6	4.39

4 lentelėje pateikta: turnyre surinktu taškų prognozės, pasiektų įvarčių skaičius visose varžybose, tikimybės, kad komanda užims 1-2 vietas ir pateks į kitą turnyro etapą, tikimybės, kad komanda užims trečią vietą (tokiu atveju leidžiama žaisti UEFA taurės varžybose), bei tikimybė jok

komanda užims paskutinę ketvirtą vietą. Tikimybės apskaičiuotos naudojant standartinį nepriklausomą dvimatį Puasono pasiskirstymo dėsnį. Tikrasis surinktas komandų taškų skaičius nurodytas skliaustuose.

Lentelė 5. GF - įspirti įvarčiai. GA - praleisti įvarčiai. Simuliacijos rezultatai po modifikacijos ir skliaustuose tikrieji rezultatai.

Komandos	Vid.	Vid.	Vid.	Tikimybės (%)			
	Taškai	GF	GA	1	2	2.5-3	3.5-4
AS Roma	11.7 (12)	12.2	6.0	40.4	90.1	8.9	1.0
Chelsia	12.4 (11)	11.9	5.1	50.1	92.1	7.2	0.7
Bordeaux	4.9 (7)	5.0	11.0	0.8	6.7	43.4	49.9
CFR 1907 Cluj	4.6(4)	5.1	12.1	0.5	4.6	34.7	60.7

Autoriai kaip pavyzdį leidžiantį suprasti robastiškumą panaudojo Chelsea - Cluj varžybas. Varžybos baigėsi 2-1, nors pagal komandų žaidimo lygį, metinį biudžetą ir pan. tikėtinas rezultatas gali būti ir 5-1. Šie rezultatų pakitimai neturi jokios įtakos komandų reitingams ir galutiniam lygoje surinktų taškų pasiskirstymui. Taigi dėl šios priežasties kas nors gali pamanyti, kad toks rezultatų skirtumas neturi didelės įtakos modelio prognozuojamoms tikimybėms. Tačiau realią įtaka modelio tikimybėms galime matyti iš 5 lentelės. Tokie pakitimai lemia tai, kad atnaujintos prognozės Chelsea komandai prognozuoja pirmą vietą grupėje. Lentelė 5 apibendrina prognozavimo modelių tikimybių pokyčius atlikus šias rezultatų korekcijas. Dimitris Karlis ir Ioanis Nitzoufras savo darbe [KN03] akcentavo tai, kad labai keista, jog toks svarbus aspektas nebuvo anksčiau aptartas panašios tematikos darbuose.

Lentelė 6. GF - įspirti įvarčiai. GA - praleisti įvarčiai. Simuliacijos rezultatai po modifikacijų.

Komandos	Vid.	Vid.	Vid.	Tikimybės (%)			
	Taškai	GF	GA	1	2	2.5-3	3.5-4
AS Roma				-8.8	+3.0	-1.5	-1.5
Chelsia	+1.3	+2.9		+13.2	+9.4	-6.5	-2.9
Bordeaux	4.9 (7)	5.0	11.0	0.8	6.7	+10.4	-9.7
CFR 1907 Cluj						34.7	60.7

Robastiškumas - labai svarbus, tačiau dažnai nepakankamai įvertinamas statistikos aspektas. Didžiausio tikėtinumo metodai gerai žinomi ir yra pakankamai efektyvus. Tačiau šie metodai stipriai pažeidžiami duomenų neatitikimų. Robastiški metodai reikalauja daugiau skaičiavimo, todėl jie praktikoje naudojami retai. Naudodami robastiškus metodus aukojame efektyvumą dėl patikimumo. Taigi Dimitris Karlis ir Ioanis Nitzoufras [KN03] savo darbe ieškojo kompromiso tarp skaičiavimų kiekio ir patikimumo. Savo darbe autoriai naudojo:

- n - sužaistų rungtynių skaičius;
- $X_i, Y_i, i = 1, ..., n$ - namų ir svečių komandų pelnytų įvarčių skaičius;
- θ_i - rungtynėms pritaikomas parametras naudojamas įvertinant jungtine *(angl. joint)* tikimybę;

Autoriai tikimybėms įvertinti naudojo didžiausio tikėtumo metodą:

$$L_w = \sum w_i log f_i(x_i, y_i; \theta_i)$$

,čia w_i svoris priskiriamas i-tajam žaidimui (standartinėje didžiausio tikėtumo lygtyje naudojama $w_i = 1$). w_i leidžia suteikti tam tikrą sužaistų varžybų svorį įtakojantį modelio tikimybes. Autorių nuomone toks būdas tinkamas į prognozavimo modelį įtraukiant tokias charakteristikas kaip: komandos sudėties pasikeitimas, komandų motyvacija, ekspertų nuomonės ir pan.

Dimitris Karlis ir Ioanis Nitzoufras [KN03] save darbe vertindami aptartą reikšmę naudojo schemą:

$$w_i = \begin{cases} i & \text{jei } |x_i - y_i| < m_0 \\ p & \text{kitu atveju} \end{cases}$$

Tokiu būdu daroma prielaida, jog rezultatų skirtumas didesnis nei m_0 turėtų daryti mažesnę įtaka modelio tikimybėms. Dar viena autorių siūlyta svorių schema paremta modeliu svoriais. Pagal šį būdą rezultatai neatitinkantys modelio prognozių turi mažesnę įtaką prognozėms.

1.4 Veibulo pasiskirstymo dėsniu paremti modeliai

Naujausiais darbas rastas atliekant literatūros analizę - Bošnakovo, Kharato ir Makhalo [BKM17] tyrimas. Mokslininkai pasiūlė dvimatį Puasono modelį. Šis prognozavimo modelis naudoja Veibulo *inter-intervall-times-based* skaičiavimo procesą. Darbe modelis palygintas su paprastuoju ir nepriklausomu Puasono pasiskirstymo dėsniais. Jo naudingumas įvertintas su *Kelly-type* lažybų strategija. Ši strategija skirta lažintis už futbolo varžybų baigtis tokias kaip namu, svečių komandų pergalės, lygiosios arba už įvarčių skaičių futbolo varžybose. Vertinimo rezultatai buvo teigiami, t.y mokslininkams pavyko atrinkti pelningas varžybų baigčių prognozes.

1.5 Diskusija ir apžvalga

Sporte komandos nėra vienodos, kiekviena turi stipriąsias ir silpnąsias puses. Faktas yra tai, jog mes turėtume tikėtis, kad daugiau taškų pelnys ir laimės varžybas stipresnė komanda. Bet kaip nustatyti stipresnę komandą? Kokie komandų parametrai yra svarbesni? Kaip vienodi parametrai skiriasi skirtingu laiku? Kokia įtaką turi trenerio pasikeitimas? Tokius ir panašius klausimus sprendžia mokslininkai norėdami sukurti tikslius prognozavimo modelius. Mes galime panaudoti įvairius matematinius įrankius, siekiant padaryti įvairiais išvadas apie rungtynių baigtį. Tam tikslui plačiai naudojamas didžiausio tikėtinumo metodas, tiesiniai modeliai, nepriklausomas ar dvimatis Puasono pasiskirstimo dėsnis ir pan. Sėkmingas prognozių modelis, gali būti naudojamas kaip pelningos lažybų strategijos pagrindas.

Komandos žaidėjų pozicija ir kamuolio kontrolė varžybų metu svarbus veiksnys įtakojantis tikimybę p (komanda pelnys įvartį). Tikimybė p yra maža, tačiau dažnis, kai komanda turi galimybę pelnyti įvartį yra didelis. Jai p yra konstanta, o atakos yra nepriklausomos, tuomet įvarčių

skaičius yra binominis pasiskirstymo dėsnis ir tokiomis aplinkybėmis įvarčių prognozavimo uždaviniui spręsti puikiai tinka Puasono pasiskirstymo dėsniu paremti prognozavimo modeliai.

Literatūroje galime rasti įvairių modeliu skirtų futbolo varžybų rezultatų prognozavimui. Tačiau dėl savo paprastumo ir patikimumo dažniausiai naudojamas Puasono pasiskirstymo dėsnis ir jo variacijos. Atlikus literatūros apžvalgą galime spręsti, jog šis prognozavimo modelis tinka sporto prognozių rezultatų prognozavimo uždaviniams spręsti. Be to Puasono pasiskirstymo dėsniu paremti modeliai lengvai lengvai plečiami įtraukiant įvairius naujus intensyvumo parametrus. Tiksliausi rezultatai gaunami panaudojus dvimatį Puasono pasiskirstymo dėsnį. Tikslumą padidina tokie parametrai kaip: komandų gynybos bei puolimo stiprumas, namų pranašumas, varžybų laikas ir pan. Iš atliktos literatūros apžvalgos galime teikti, kad lygiosios turi didesnę tikimybę įvykti, nei tikimybę įvertina standartinis trijų parametrų, Puasono pasiskirstymo dėsnio pagrindu, sudarytas modelis. Tačiau apžvelgtuose modeliuose pasigendama robatiškumo, leidžiančio tiksliau parinkti parametrų įvertinimus.

Apžvelgtuose darbuose autoriai dažniausiai atkreipia dėmesį į tokias modelių charakteristikas kaip:

- tikslumas - kaip tiksliai modelis geba prognozuoti futbolo varžybų rezultatus ar jų baigtis;
- robastiškumas - kaip keičiasi modelio prognozių tikslumas, atsižvelgiant į duomenų neatitikimą ar nukrypimą;
- prognozėms atlikti ir duomenims apdoroti reikalingi skaičiavimų resursai;
- modelio naudojami intensyvumo parametrai.

Dimitris Karlis ir Ioanis Nitzoufras darbe didžiausias dėmesys skirtas robastiškumui [KN03]. Iš šio darbo rezultatų galime teigti, kad modelio naudojamų duomenų patikrinimas gali būti puikus būdas padidinti statistinio modelio kokybę.

Atlikus ir įvertinus literatūros apžvalgą galime apibendrinti rezultatus ir pasiūlyti šio darbo gaires:

A Į modelį turi būti įtraukta susitinkančių komandų gebėjimai;

B Modelyje turėtu būti įtraukta namų komandos pranašumas;

C Paskiausi rezultatai turėtu daryti didžiausią įtaka prognozėms;

D Modelyje turėtu būti įvertinta komandų puolimo ir gynybos sugebėjimai;

E Modelyje turėtu būti atsižvelgta į komandų tarpusavio rezultatus.

1.6 Susitinkančių komandų gebėjimai

Dažniausi literatūroje sutinkamuose modeliuose atsižvelgiama į komandų puolimo ir gynybos kokybę. Apžvelgtų darbų autoriai dažniausiai vadovaujasi Matherio pasiūlytu modelio variantu,

kai vertinama namų komandos puolimo kokybės rodiklis α, svečių komandos gynybos rodiklis β, namų komandos gynybos rodiklis γ, ir svečių komandos puolimo rodiklis δ.

Atlikus literatūros analizę buvo nuspręstą, jog šiame darbe naudojamo modelio pagrindą sudarys Puosono pasiskirstymo dėsnis su anksčiau aptartais komandų kokybę įvertinančiais parametrais. Jiems apskaičiuoti panaudotas didžiausio tikėtumo metodas. Šis metodas, kaip galima matyti iš literatūros apžvalgos, teikia efektyviais prognozes, bei plačiai naudojamas tokio tipo uždaviniams spręsti. Modelio pasirinkimą lėmė atlikta literatūros apžvalga bei Puasono pasiskirstymo dėsnio tinkamumas sporto varžybų rezultatų prognozavimo uždavinių sprendime.

1.7 Namų komandos pranašumas

Apžvelgtoje literatūroje dažnai sutinkama namų pranašumo parametras, sustiprinantis namuose žaidžiančios komandos tikimybę laimėti varžybas. Paprastai apžvelgtoje literatūroje namuose žaidžiančiai komandai yra priskiriama pranašumo konstanta. Matheris siūlė šią konstantą įvertinti atsižvelgiant į tai, kaip dažniau namų komanda spiria įvarčius.

$$\hat{k}^2 = \frac{\sum_i \sum_{i \neq j} x_{ij}}{\sum_i \sum_{i \neq j} y_{ij}}$$

1.8 Laiko parametras

Dar 1982 metais Maheris kalbėjo, jog norint tiksliai prognozuoti futbolo rungtynes, reikia įvertinti, jog komandų žaidimo kokybė kinta ir ji nėra statinė. Savo darbe tai plačiausiai aprašė Diksonas ir Kolis.

Jie savo darbe naudojo laiko įtakos funkciją ϕ:

$$\phi(t) = exp(-\xi t)$$

pagal šią funkciją ankstesnių rezultatų įtaka mažėja eksponentiškai pagal parametrą $\xi > 0$. Kuo didesnė ξ reikšmė, tuo didesnę įtaką turi naujausi rezultatai. Tai gi statiniame modelyje ši reikšmė būtų lygi 0.

1.9 Duomenų robastiškumas

Logiškai mąstant rezultatai gali būti laikomi patikimais jai jie nenukrypsta nuo standartinių duomenų. Jei ši prielaida teisinga, tuomet tai leistų pagerinti modelių kokybę. Statistikoje tai galima pasiekti pasinaudojant dispersija. Dispersija *(angl. variance)* – statistinė imties charakteristika, atspindinti labiausiai tikėtiną eilinio matavimo vertės nukrypimą nuo aritmetinio vidurkio.

1.10 Išvados

Atliktoje literatūros apžvalgoje išnagrinėti modeliai, labiausiai tinkantys futbolo varžybų baigčių ir rezultatų prognozavimo uždavinių sprendimams. Taigi prognozavimo modelių charakteristikų tyrimas bus atliktas šių modelių kontekste.

Literatūros apžvalgoje sutikti modeliai dažniausiai remiasi Puasono pasiskirstymo dėsniu ir jo modifikacijomis. Iš apžvelgtų modelio rezultatų matyti, kad sudarant statistinį modelį reikėtų atsižvelgti ne vien tik namų ir svečių komandų gynybos bei puolimo stiprius, bet taip pat ir į namų komandos pranašumą bei laiko parametrą, be to rezultatams (0-0, 0-1, 1-0 ir 1-1) suteikti didesnes tikimybes. Taigi šie modeliai sudaro automatizuotos sporto varžybų baigčių prognozavimo sistemos pagrindą.

2 Prognozuojamumo indeksas

Pirmiausia prieš pradedant matematinių prognozavimo modelių tyrimą, reikėjo nuspręsti kokiame kontekste jis bus atliktas. Tyrimą vykdyti buvo galimybė visų futbolo varžybų, lygos ar komandos kontekste, todėl buvo sugalvotas prognozuojamumo indeksas. Šio indekso tikslas aptikti, atvejus kai lažybininkai naudoja netinkamus metodus nustatant varžybų baigčių tikimybes. Todėl atradus kokie algoritmai šioms baigtims tinka labiau, bei jais pasinaudoję galėtume pagerinti šių baigčių prognozuojamumo indeksą bei suteikti galimybę kuriamos sistemos vartotojui pasirinkti pelningas lažybų baigtis.

Kadangi šiam koeficientui apskaičiuoti ir norint objektyviai spręsti apie komandų prognozuojamumą, varžybų skaičiaus sužaistų vienos komandos yra per mažai. Todėl nuspręsta prognozuojamumo indeksą vertinti visų futbolo lygos varžybų kontekste.

Prognozuojamumo indeksui įvertinti sukurtas metodas ir metodu paremtas programinis įrankis. Sukurtas įrankis geba įvertinti kaip dažnai pasitvirtina lažybų agentūrų prognozės. Tam tikslui varžybų lažybų koeficientai paversti į tikimybes naudojant formulę (o varžybų baigties koeficientas):

$$P = (1/o)100$$

,kur o - baigties koeficientas.

$$\bar{P_x} = \frac{1}{n} \sum_{m=1}^{n} \frac{100}{o_{x,m}}$$

,kur $\bar{P_x}$ - prognozuojamumas, o_x - baigties koeficientas, x - baigtis, n - viso varžybų.

$$S_{pr} = L_{pr} - \frac{1}{n} \sum_{i=1}^{n} L_{pr_i}$$

,kur S_{pr} - prognozavimo stabilumas, L_{pr} - lygos prognozavimo charakteristika, L_{pr} - rinkinio prognozuojamumas.

Šio darbo rašymo metu sistema naudojo vieno lažybų organizatoriaus siūlomus varžybų baigčių koeficientus. Nors tikslesnius rezultatus gautume naudojant vidurkį iš daugelio lažybų organizatorių informacijos, tačiau tai stipriai apsunkina prognozuojamumo indekso vertinimo procesą. Šio metodo metu vertinama lažybų organizatorių siūlytų tikimybių pasiteisino procentas. Vertinant prognozuojamumo indeksą, tikimybės skaidomos į intervalus pagal jų reikšmes. Dėl skirtingų lygose sužaistų varžybų skaičiaus, prognozuojamumo indeksas vertinamas naudojant 4 lygių intervalų sistemą (7 lentelė). Tai leido įvertinti lygų su mažai sužaistų varžybų prognozuojamumą. Lygose su didesniais varžybų kiekiais, naudojami mažesni procentų intervalai, kadangi turimų duomenų pakanka kokybiškai įvertinti lygos prognozuojamumą.

Skaičiuojant prognozuojamumo indeksą, pirmiausia varžybos priskiriamos į vieną iš intervalų pagal lažybų organizatorių apskaičiuotas baigties tikimybės. Tikimybės įvertinamos naudojant lažybų agentūrų baigčių koeficientus. Tuomet apskaičiuojama intervale esančių baigčių pasiteisinimo procentas, t.y. koks varžybų procentas baigėsi tam tikra baigtimi. Pavyzdžiui į pirmo lygio

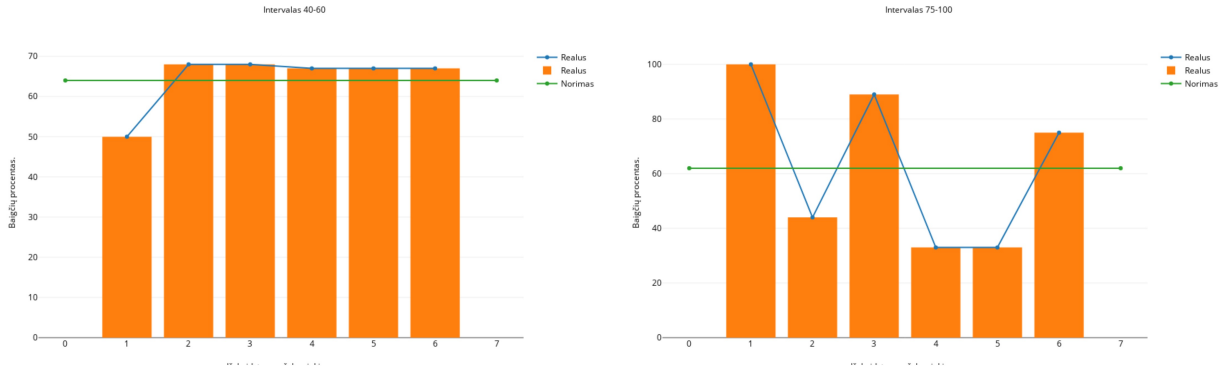
Lentelė 7. Prognozuojamumo indekso, 4 lygių tikimybių intervalai procentais.

Lygis	Intervalai									
1	0-50					50-100				
2	0-33			33-66			66-100			
3	0-20		20-40		40-60		60-80		80-100	
4	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100

pirmą intervalą patenka visos varžybų baigtys su lažybų organizatorių apskaičiuotomis tikimybes nuo 0 iki 50 %. Tarkime šiame rėžyje turime 100 varžybų baigtis. Jei iš 100 varžybų baigčių 60 varžybų baigėsi norima baigtimi, tuomet gautas procentas parodo lažybų agentūrų nesugebėjimą ar sugebėjimą tinkamai įvertinti apžvelgiamos lygos varžybų baigčių tikimybes. Skaičiuojant prognozuojamumo indeksą sudedama visu intervalų nuokrypiai.

2.1 Prognozuojamumo pastovumo indeksas

Šiam tikslui visos į intervalą patenkančios varžybos suskirstomos į lygias dalis pagal laiką. Tuomet vertinamas baigčių dažnis kiekviename intervale bei lyginama su bendrų prognozuojamumo indeksu. Įvertinus intervalus, tikrinama ar jie pakankamai pastovūs laiko atžvilgiu, t.y. ,kad tai yra dėsningumas o ne atsitiktinumas. Norint įsitikinti jog lažybų agentūrų nesugebėjimas tinkamai įvertinti baigčių tikimybių yra pastovus, sukurtas įrankis pateikiantis norimų intervalų prognozuojamumo indekso kitimo laike grafiką. Vartotojas pasinaudojęs sukurtu įrankiu, gali įvertinti ar šis indeksas yra pakankamai stabilus.



1 pav. Stabilaus ir nestabilaus prognozuojamumo indeksų grafiko pavyzdžiai.

Kaip matome iš grafiko (1 pav.), stabilus prognozuojamumo indeksas pasižymi pastovių baigčių procentu nepriklausomai nuo laiko. Nepastovumas pasireiškia baigčių dažnių svyravimais. Intervalo prognozuojamumo pastovumo indekso skaitinė reikšmė išreikšta per intervalo dalių (oranžiniai stulpeliai grafike) vidutinį nuokrypį nuo intervalo baigčių procento (žalia linija). Vertinant stabilumo indeksą sudedama stulpelių nuokrypiai nuo žalios linijos. Žemesnis skaičius reiškia, jog intervalas yra stabilus ir patikimas.

2.2 Rezultatai

Atliekant prognozuojamumo indekso tyrimą buvo ieškoma intervalų kuriuose varžybų baigčių tikimybės nėra pervertinamos t.y. faktinis baigčių procentas yra didesnis nei juos vertina lažybų agentūros. Tyrimo metu surinkta futbolo varžybos iš 566 lygos naudojant varžybų duomenis nuo 2013 metų. Taip pat parašytas bei panaudotas skriptas skirtas apskaičiuoti visų lygų prognozuojamumo indeksus leidžiantis automatizuoti šį procesą. Gauti duomenys išsaugoti duomenų bazėje. Skaičiavimai pasiekiami pasinaudojus sukurta naršyklės programa. Programoje pateikiamos visos duomenų bazėje esančios lygos, jų prognozuojamumo ir prognozuojamumo stabilumo indeksai. Naršyklės aplikacijoje įgyvendinta rūšiavimo funkcija, leidžianti vartotojui greitai ir patogiai surasti lygas su prastais prognozuojamumo indeksais.

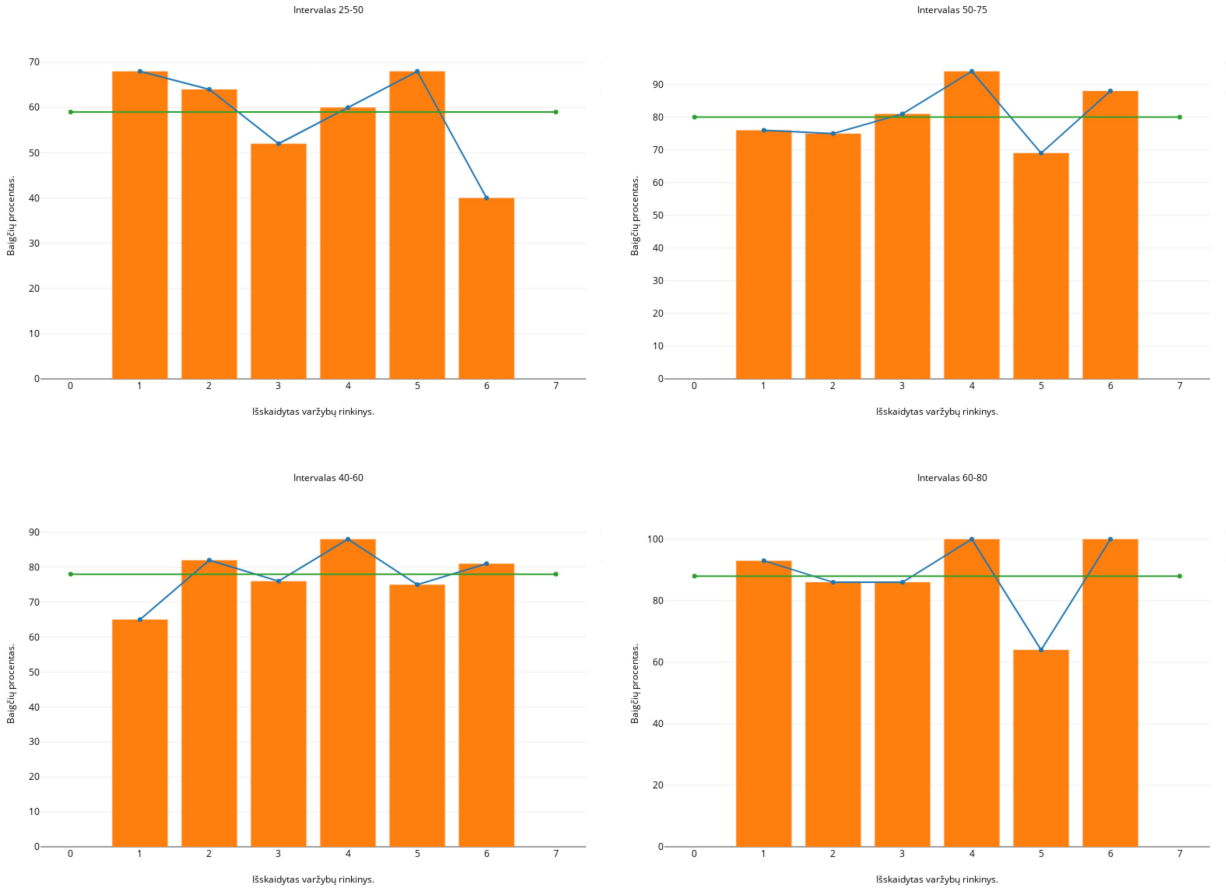
Taigi tolesniam matematinių modelių tyrimui buvo pasirinktos sunkiausiai lažybų agentūrų prognozuojamos lygos. Vertinant prognozuojamumo indeksą tikrinami tik tie intervalai kurių bendras varžybų skaičius nebuvo žemesnis nei 60. Naudojant intervalus su mažesniu varžybų skaičiumi pateikti stabilumo rodikliai nebus tikslūs.

Atlikus lygų prognozuojamumo analizę (8 lentelė), nuspręsta darbe atliktame modelių tyrime naudoti *Taça de Portugal* lygą. Šioje lygoje nuo 2013 metų sužaista 807 rungtynės. Atlikus analizę pastebėta (8 lentelė), jog lažybų agentūromis sunkiai sekasi nustatyti šio lygos varžybų tikimybes t.y. lyga sunkiai prognozuojama. Šios lygos namų komandų pergalių baigtims yra suteikiami per dideli baigčių koeficientai.

Lentelė 8. *Taça de Portugal* lygos tikimybių intervalai, bei jų prognozuojamumo ir prognozuojamumo stabilumo indeksai.

Intervalas, %	Realus, %	Pokytis, %	Stabilumo indeksas	Važybų skaičius
25 - 50	59	+9	-10	150
50 - 75	80	+5	-6	97
40 - 60	78	+18	+5	99
60 - 80	88	+8	+4	83

Pasinaudojus sukurta analizės sistema patikrintas lygos prognozuojamumo stabilumas (2 pav.). Naudodami pateiktus grafikus galime įvertinti ar lažybų agentūrai nesisekė įvertinti varžybų baigčių tikimybių atsitiktinai ar tai yra dėsningumas. Iš gautu rezultatų galime daryti išvadą, kad nesugebėjimas prognozuoti baigčių teisingai yra dėsningumas. Todėl tampa labai aktualu sukurti prognozavimo modelius gebančius tiksliau įvertinti baigčių tikimybes. Šių modelių pagalba galėtume tiksliau nustatyti šios lygos varžybų įvykiu baigtis. Taigi toliau mes pasimaudomai sukurta analizės sistema ištirsime kaip pasirinkti prognozavimo modeliai geba nustatyti *Taça de Portugal* lygos varžybų baigtis.



2 pav. *Taça de Portugal* lygos prognozuojamumas pagal tikimybių intervalus.

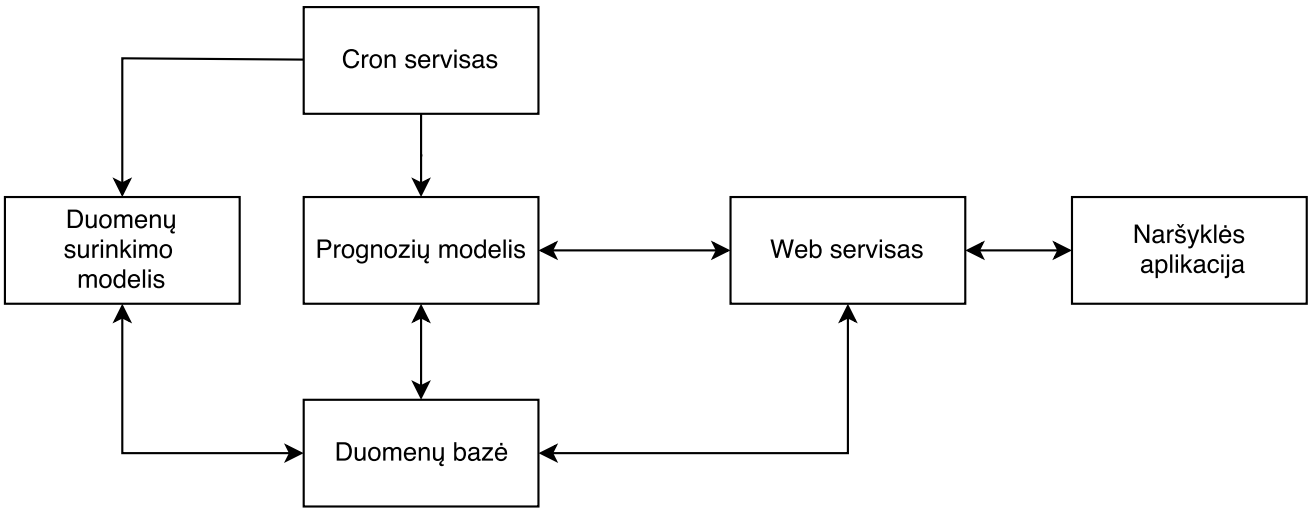
3 Programų sistema

3.1 Programos architektūra

Sistemos įgyvendinimui (naršyklės ir serverio programoms) pasirinkta *JavaScript* programavimo kalba. Tai leido naudoti vieną programavimo kalbą ir taip padidinant programos kūrimo greiti bei kokybę, kadangi įgyvendinant programų sistemą nereikia mokėti skirtingų programavimo kalbų. Naršyklės aplikacija sukurta naudojant *React*, *TypeScript*, *Webpack* technologijas. *TypeScript* leido praplėsti standartinias *JavaScript* programavimo kalbos galimybes. Tai leido naudoti objektų tipus taip palengvinant kodo analizę ir sumažinant klaidų tikimybę. Serveryje veikiančiam *HTTP* serveriui sukurti naudotas *Express* programavimo karkasas, palengvinantis serverio ir naršyklės programų bendravimą *HTTP* protokolu. Dėl gan didelio sistemoje naudojamų duomenų struktūros kintamumo buvo pasirinkta naudoti *MongoDB NoSQL* duomenų bazę.

Sukurtą futbolo varžybų baigčių prognozavimo ir prognozavimo modelių analizės automatizuotą sistemą sudaro:

- Naršyklėje veikianti programa, skirta vartotojo grafinei sąsajai su sistema;
- *HTTP web* servisas skirtas naršyklėje veikiančios programos bendravimui su serveryje veikiančia programų sistema;
- Prognozių modelis - skirtas varžybų baigčių tikimybių skaičiavimams;
- *Cron* servisas - skirtas automatizuoti įvairius sistemos procesus kaip: varžybų informacijos surinkimas ir atnaujinimas, futbolo varžybų baigčių tikimybių įvertinimas, prognozavimo modelių įvertinimų perskaičiavimas;
- Duomenų surinkimo modelis - skirtas futbolo varžybų informacijos automatizuotam surinkimui iš tokia informaciją disponuojančių šaltinių.



3 pav. Sukurtos programų sistemos koncepcinis modelis.

Kaip matome iš 3 pav. naršyklės aplikacija bendrauja su http servisu. Http servisas bendrauja su duomenų baze ir prognozavimo modeliu atsakingu už varžybų baigčių tikimybių įvertinimu. *Cron* servisas turi sąsajas su prognozavimo bei duomenų valdymo modeliais.

Matematiniams modeliams įgyvendinti pasirinkta sistemos architektūra kai sistema su prognozavimo modeliais bendrauja per http protokolą. Tai leido naudoti prognozavimo modelius parašytus skirtingomis programavimo kalbomis.

3.2 *HTTP* serveris

Sukurtas *web* servisas skirtas aptarnauti naršyklės programos *HTTP* užklausas. Servisas atitinka šiuos reikalavimus:

- Komunikacijai naudoja *HTTP* protokolą;
- Bendravimui pakanka 4 pagrindinių operacijų: *GET, POST, PUT, DELETE*;
- Lengvas *web* serviso įgyvendinimas;
- Plačiai palaikomas įvairių programavimo karkasų;
- *Web* servisas pasiekimas naudojant bet kurią interneto naršyklę.

Atsižvelgiant į kuriamos sistemos tipą ir reikalavimus pasirinkta *RESTful* architektūros *web* servisas. Su šio tipo web servisu bendraujama kaip su bet kuriuo kitu web resursu - vadovaujamasi *REST (angl. representationalState transfer)* architektūros principais [Mas11]:

- Konkretų resursą identifikuoja *URI*, kuriuo kreipiamasi į *web* servisą;
- *HTTP GET* užklausa naudojama pasiekti resurso turiniui. Turinys grąžinamas *HTTP* atsakyme;
- *HTTP PUT* arba *POST* užklausa naudojama resursui keisti arba naujam resursui sukurti. Užklausos kūne *(angl. body)* nurodomas naujasis resurso turinys;
- *HTTP DELETE* užklausa naudojama resursui trinti.

Kuriamos sistemos naršyklės programai aptarnauti įgyvendinti šie galiniai serviso taškai (*angl. endpoints*) :

- *GET* / - *html* failas su *Javascript React* karkaso programą;
- *GET /model/data/:modelId/:date* - prognozių modelio skaičiavimų rezultatai pasirinktai dienai;
- *GET /model/:id/estimation/* - gaunamas modelio įvertinimas *IDrawData* formatu;
- *GET /model/:id/summary/* - gaunamas modelio įvertinimas *ISummary* formatu;

- *GET /match/:id* - gražina duomenis apie varžybas pagal varžybų *id*;
- *GET /match/:mid/prediction/:pid* - gražina futbolo varžybų duomenis su apskaičiuotomis tikimybėmis *IMatchWithPredictions* formatu;
- *GET /matches/:date* - gražina varžybų sąrašą pagal norimą datą;
- *GET /team/:id* - gražina duomenis apie komandą pagal komandos *id*;
- *GET /league/:id/:session/matches* - gražina varžybų sąrašą pagal lygos *id* bei sezoną (metai);
- *GET /location/:id/:session/matches* - gražina varžybų sąrašą pagal šalies *id* ir sezoną (metai);
- *POST /update/data/:date* - atnaujinta futbolo varžybų rezultatus pagal datą;
- *POST /update/match/:id* - atnaujinta futbolo varžybų rezultatus pagal varžybų *id*;
- *POST /update/league/:id* - atnaujinta futbolo varžybų rezultatus pagal lygos *id*;
- *POST /update/loaction/:id* - atnaujinta futbolo varžybų rezultatus pagal šalies *id*;
- *POST /update/team/:id* - atnaujinta futbolo varžybų rezultatus pagal komandos *id*;

Kadangi sistemos įgyvendinimui pasirinkta *JavaScript* programavimo kalbą, pranešimų turinys koduojamas *JSON* formatu. Šios kalbos sintaksė aprašanti objektų struktūras yra tokia pat kaip *JSON*. Serverio aplikacijoms kurti pasirinkta naudoti *NodeJS*. Naudojant *JavaScript* programavimo kalbą netik rašant naršyklės programą , tačiau ir serverio programas galime per panaudoti programinį kodą tarp šių skirtingų lygių programų. Taip pat kuriant sistemą užtenką mokėti vieną programavimo kalbą [Rau12].

3.3 Duomenų bazė

Visa informacija apie varžybas sistemoje laikoma *Mongo NoSQL* duomenų bazėje. Pasirenkant ir projektuojant duomenų bazę atsižvelgta, kad duomenų bazė bus nuolat pildoma naujais duomenimis. Taip pat reikia atkreipti dėmesį į tai, jog nors šiame darbe bus saugoma tik futbolo varžybų duomenys, duomenų bazės struktūra turi būti lengvai papildoma ar modifikuoja pagal kitas sporto šakas. Pasirinkta duomenų bazė pasižymi šiomis savybėmis:

- Populiariausių programavimo kalbų palaikymas;
- Galimybė valdyti skirtingas duomenų struktūras;
- Pasirinktai duomenų bazei galime rasti patogių įrankių leidžiančių testuoti sistemos modelius su fiktyviais (*angl. mock*) duomenimis;
- Galimybė daryti atsarginiais kopijas;
- Galimybė naudoti išskirstytus serverius;
- *JSON* duomenų struktūra;

3.3.1 Sistemos duomenų struktūra

Kaip matome iš duomenų struktūros schemos (4 pav.) vienas iš pagrindinių sistemoje naudojamų objektų - varžybas aprašanti duomenų struktūra. Šį objektą aprašo *IMatch* sąsaja. Kiekvienos varžybos gali turėti n ryšių su lygos (*ILeague*) objektais. *ILeague* objekte saugoma informaciją apie futbolo lygas (programos kūrimo metu buvo surinkta ir duomenų bazėje išsaugota 566 lygos objektai). Taip pat sistemoje svarbi, *ITeam* sąsaja aprašanti komandas. Visa informacija apie prognozes saugoma prognozės (*IPrediction*) objekte. *IPrediction* sąsaja apibrėžia ryšius tarp varžybų ir prognozių modelio *IPredictionModel* objektų. Pastarasis skirtas saugoti informacijai susijusiai su sistemoje naudojamais prognozavimo modeliais.

Viena iš problematiškesnių vietų - lažybų agentūrų varžybų baigčių koeficientų saugojimas. Kadangi vienoms varžybos galime priskirti labai daug skirtingų baigčių (namų komandos pergalė, varžybos pasibaigs lygiosiomis, svečių komandos pergalė, tikslus varžybų rezultatas, įvarčių kiekis ir pan), be to duomenų bazėje tenka saugoti skirtingų lažybų agentūrų siūlomus koeficientus visoms šioms baigtims, be to šie koeficientai kinta laike. Dėl šių priežasčių kuriant sistemą buvo susidurta su problema kai dėl didelio koeficientų kiekio sistema veikdavo lėtai. Norint paspartinti sistemos veikimą nuspręsta naudoti tik vienos lažybų agentūros baigčių koeficientus.

3.4 Grafinė sąsaja

Sukurtos sistemos grafinės sąsajos pagrindinis tikslas pateikti vartotojui varžybų baigčių prognozes. Kuriant sistemą įgyvendinti atskiri puslapiai skirti platesnei analizei pagal lygą, komandą ir varžybas. Naršyklės programai įgyvendinti naudota tokios technologijos kaip *React, Redux, TypeScript, Webpack*. Naršyklės kaip ir serverio programa, parašyta *JavaScript* programavimo kalba. Tai leido kuriant sistemą naudoti serverio funkcijas naršyklės programoje, taip perkeliant skaičiavimus į kliento kompiuterį.

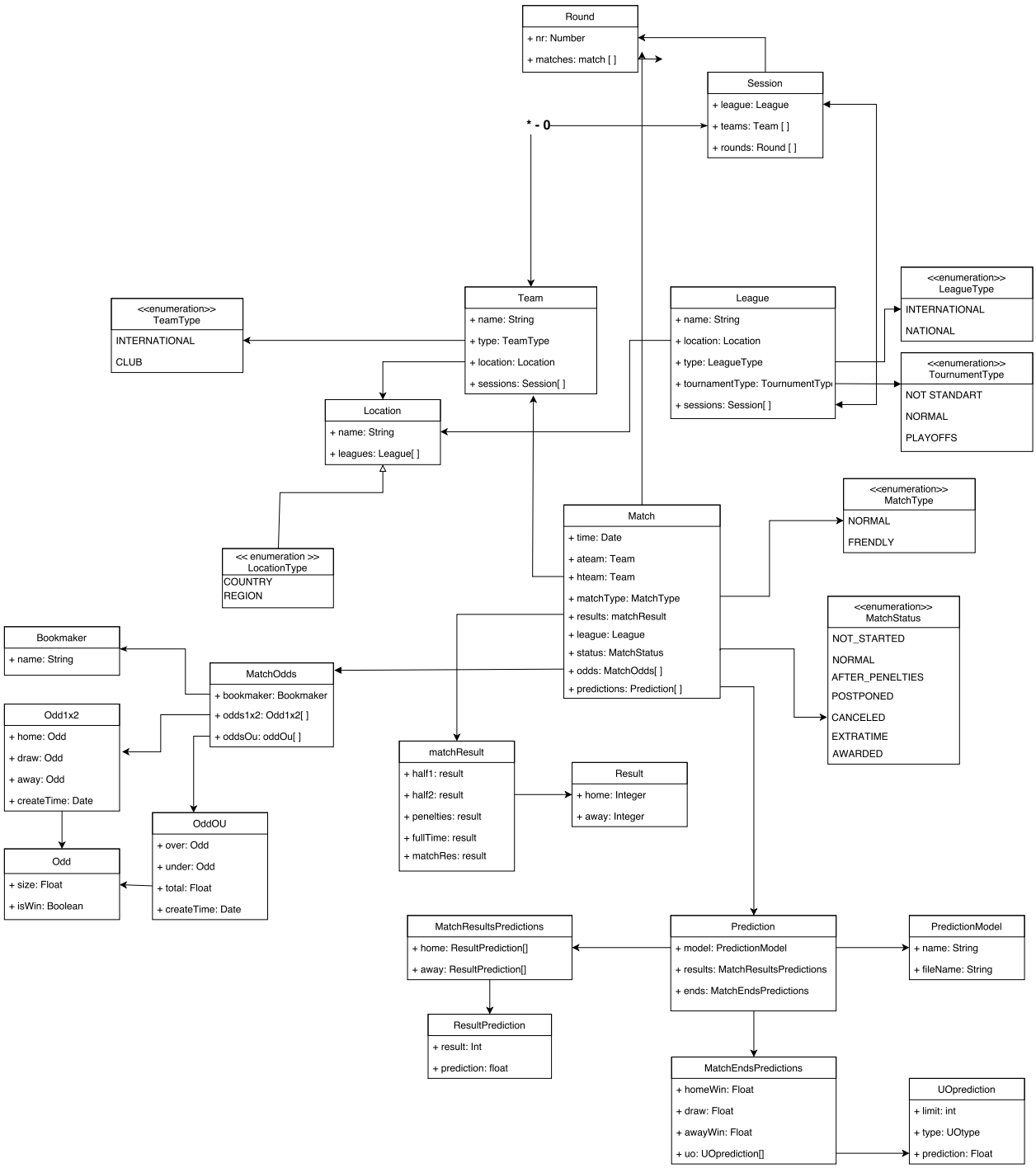
3.4.1 Dienos prognozių puslapis

Šiame puslapyje vartotojui pateikimas dienos varžybų sąrašas (5 pav.) su lažybų agentūros ir sistemoje naudojamų modelių vertinimais bei jų prognozių informacija. Varžybos rūšiuojamos pagal lygos įvertinitą investicijos grąžą *(ROI)*. Naudodamasis šiuo puslapiu vartotojas turi galimybę atsirinkti dienos varžybų prognozes pagal skirtingų modelių įvertinimus.

Kaip matome 5 pav. varžybų komponentas (1) skirtas pateikti su varžybomis susijusia informaciją: varžybų laiką, lygą, komandų pavadinimus, rungtynių statusą (neprasidėjusios, pasibaigusios, atšauktos ir pan.), rezultatą. Vartotojas paspaudes lygos nuorodą nukreipiamas į lygos analizei skirtą puslapį.

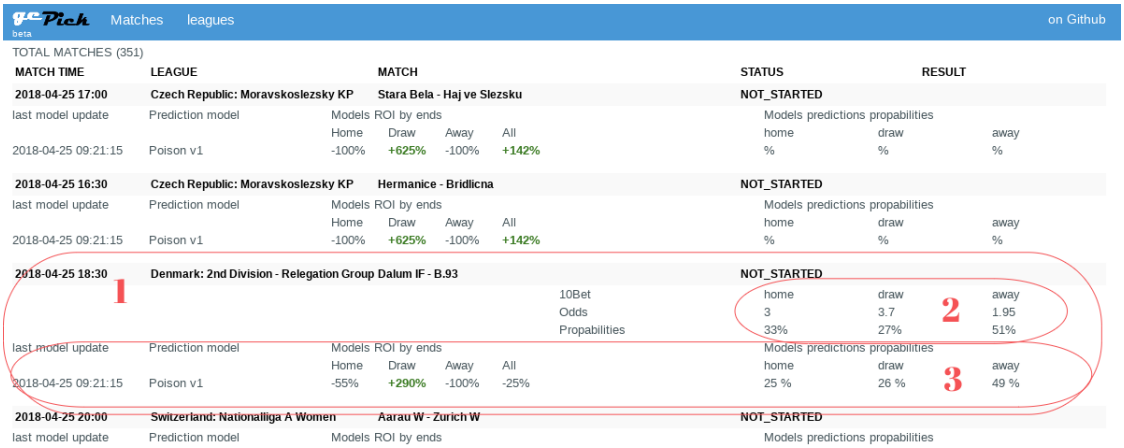
Lažybų agentūros informacinis komponentas (2) pateikia informaciją apie lažybų agentūrų siūlomus koeficientus varžybų baigtims. Taip pat šiame komponente pateikiama baigčių išraiškos procentais. Čia procentai atitinka lažybų organizatorių įvertintas baigčių tikimybes.

Prognozavimo modelių komponente (3) randame informaciją apie sistemoje naudojamus prognozavimo modelius. Informacija pateikta eilutėmis, kiekviena eilutė atitinka skirtingą prog-



4 pav. Duomenų modelių struktūra.

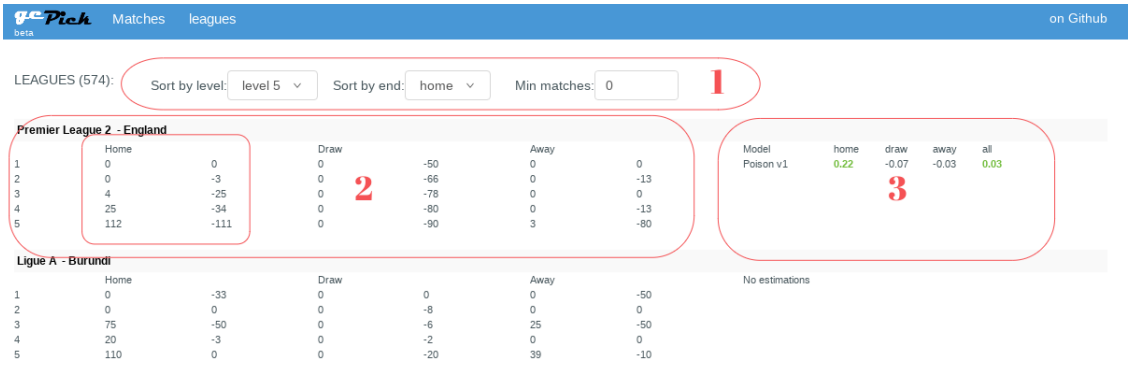
nozavimo modelį. Paskutinio atnaujinamo data nurodo kada paskutinį kartą įvertintas modelio pelningumas. Prognozavimo modelio įvertinimas pateikiamas kiekvienai baigčiai atskirai pagal investicijos grąžą. Modelių prognozių komponente pateikiamos apskaičiuotos konkrečių baigčių tikimybės. Vartotojas palyginęs tikimybes su lažybų agentūros informacija gali nuspręsti apie baigties prognozės vertingumą.



5 pav. Dienos prognozių puslapis.

3.4.2 Lygų puslapis

Šis puslapis skirtas vartotojui pateikti duomenų bazėje esančių lygų sąrašą. Paspaudus lygos nuorodą atidaromas lygos puslapis.



6 pav. Futbolo lygų puslapis.

Puslapio viršuje (6. pav) turime valdymo komponentą. Šio komponento paskirtis leisti vartotojui nustatyti parametrus naudojamus rūšiuojant arba filtruojant futbolo lygų sąrašą (1). *Sort by level dropdown* elementas skirtas vartotojui pasirinkti prognozuojamumo intervalų lygį. *Sort by level end* elementas skirtas nurodyti naršyklės aplikacijai varžybų baigtį (*namų pergalė, lygiosios ar svečių pergalė*).

Lygos informacijos blokas padalintas į dvi dalis: prognozuojamumo indekso informacija (2) ir prognozavimo modelių statistika (3).

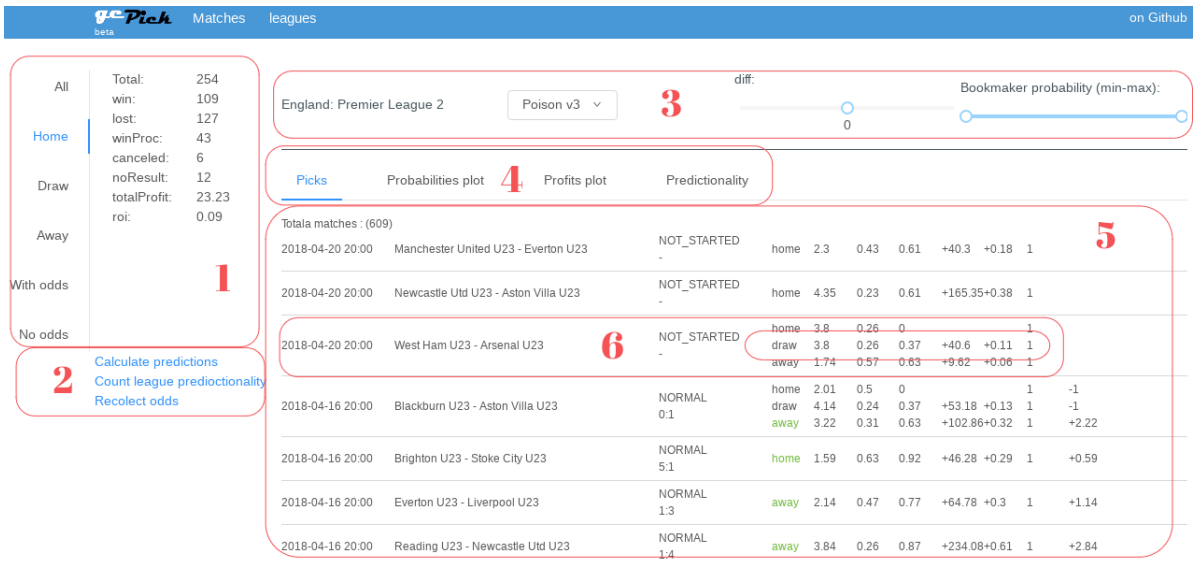
Prognozuojamumo indekso informacija pateikta lentelėje, kurios eilutės atitinka skirtingus intervalų lygius o stulpeliai skirtingas varžybų baigtis. Kiekviena varžybų baigtis susideda iš dviejų stulpelių: pirmasis rodo prognozuojamumo indeksą (kuo šis indeksas didesnis tuo lygą yra sunkiau prognozuojama (darbo atlikimo metu buvo ieškoma lygų su didžiausiais būtent šiais indeksais)), antrasis lygos stulpelis rodo neigiamą reikšmę nurodančią jog lyga buvo pervertinama lažybų agentūrų (t.y. varžybų baigčių tikimybės paprastai pervertinamos ir todėl varžybų baigtims suteikiamas ne pakankamas koeficientas).

Matematinių modelių įvertinimo blokas (3) skirtas pateikti informaciją apie prognozių rezul-

tatus. Rezultatai pateikiami nurodant kokią investicijos grąžą *(ROI)* pasiekia modelis įvertinus atskirų lygų duomenis.

3.4.3 Lygos puslapis

Šis puslapis (7 pav.) skirtas futbolo lygos istorinės informacijos peržiūrai bei analizei. Sukurta naršyklės programa geba pateikti skirtingų prognozavimo modelių rezultatus, jų tikimybių pasiteisinimo bei pelningumo kitimo grafikus, įvertinti ir pateikti prognozuojamumo indeksą ir jo stabilumo grafiką.



7 pav. Lygos puslapis, varžybų sąrašas.

Komponentas (1) skirtas pateikti pasirinkto prognozavimo modelio statistikai. Šis komponentas susideda iš dviejų dalių: kairėje - statistikos meniu, dešinėje pateikiama informacija pagal pasirinktą menių punktą. Vartotojas pasinaudojęs kairėje pusėje esančių meniu, turi galimybę pasirinkti jį dominančia statistiką. Bloke pateikiama tokia statistinė informacija: sužaistų varžybų skaičius, teisingai prognozuotų varžybų skaičius, neteisingai prognozuotų varžybų skaičius, teisingai prognozuotų varžybų procentas, atšauktų varžybų skaičius, varžybų su neatnaujintu rezultatu (nesibaigusios/neprasidėjusios varžybos) skaičius, visas pelnas (pelnas vertinamas darant prielaidą jog buvo lažintasi rizikuojant 1 vienetu ((angl. unit)) ir *ROI* (investicijos grąža procentais)).

Naršyklės programos komandų komponentas (2) skirtas nurodyti programai pradėti užduotis tokias kaip: prognozių skaičiavimas, lygos prognozuojamumo indekso nustatymas ar lygos varžybų koeficientų surinkimas. Šios komandos pradedamos vykdyti vartotojui paspaudus jų nuorodą.

Vartotojas naudodamasis parametrų valdymo komponentu (3), turi galimybę pasirinkti tokius parametrus kaip: prognozavimo modelis, skirtumo dydis tarp prognozavimo modelio ir lažybų agentūros tikimybių, lažybų agentūrų apskaičiuotų varžybų baigčių tikimybių intervalas. Pakeitus šiuos parametrus naršyklės programa perskaičiuoja statistiką, atnaujina varžybų sąrašą pagal pasirinktus parametrus, bei pateikia tikimybių pasiteisinimo ir pelno kitimo laike grafikus.

Pagrindinio turinio meniu komponentas skirtas navigacijai tarp varžybų sąrašo, prognozavimo modelių pasiteisinimo grafikų ir prognozuojamumo indekso elementų. Vartotojui pasirinkus

norimą meniu elementą, naršyklės programa pasirinkimą pateikia pagrindinio turinio komponente (5).

Varžybų sąrašo komponentas skirtas pateikti informacijai apie lygos varžybas. Šis sąrašas sudaromas pagal parametrų valdymo komponento (3) nustatymus. Naršyklė varžybų sąrašą pateikia išrūšiuotą pagal žaidimo datą. Varžybos ir jų informacija pateikiama eilutėmis (6). Eilutėje vartotojui pateikiama: varžybų laikas, susitinkančių komandų pavadinimai, varžybų būsena (neprasidėjusios, sužaistos, atšauktos ir pan.), rungtynių rezultatas bei pasirinkto prognozavimo modelio komponentas. Prognozavimo modelio komponento eilutėje pateikiama: varžybų baigtis, lažybų agentūrų suteiktas koeficientas, lažybų agentūros ir pasirinkto modelio tikimybės ir jų skirtumas.

Taigi naudodamiesi lygos puslapio programą turime galimybę analizuoti futbolo lygą įvairiais pjūviais. Reikėtu išskirti dvi kryptis t.y. prognozių modelių bei lygos prognozuojamumo analizė. Vartotojas analizuodamas prognozavimo modelius turi galimybę atvaizduoti modelio vertinimą, tikimybių dažnį bei pelną atsispindinčius grafikus. Turint šią informaciją vartotojas gali nuspręsti ar verta pasitikėti matematinio modelio veikimu pasirinktoje lygoje.

3.5 Užduotys

Sukurtos programos sistemos architektūroje nuspręsta užduotis skaidyti į atskirus, vienas nuo kito nepriklausomus modulius. Toks sprendimas leidžia atskirti sistemos funkciniais dalis taip išlaikant sistemos struktūros aiškumą.

3.5.1 Duomenų surinkimas ir atnaujinimas

Norint, kad sukurta sistema teiktu šiuo metu aktualias sporto prognozes, reikalingi sprendimai leidžiantys prižiūrėti ir vykdyti procesus susijusius su sistemos duomenų atnaujinimu realiu laiku. Galime išskirti tris pagrindiniais užduotis vykdomas sukurtoje sistemoje palaikant programos tinkamą funkcionalumą ir naudingumą:

- Sporto varžybų informacijos surinkimas;
- Sporto varžybų baigčių prognozių skaičiavimas;
- Rezultatų periodinis atnaujinimas.

Aptartos užduotys vykdomos pasinaudojus *collectDayMatches.js* ir *setupDayMatches.js* skriptais. Šie skriptai gali būti vykdomi atskirai nuo sistemos, vien tik šioms užduotims atlikti. Toks sprendimas leidžia taupyti serverio resursus. Taip pat šie skriptai lengvai vykdomi naudojant tokius įrankius kaip *linux cronjob*. Šiam darbui skirtoje sistemos konfigūracijoje šie failai yra kviečiami naudojant *node-cron js* modulį.

3.5.2 Dienos varžybų pasiūlos surinkimas

Norint užtikrinti duomenų vientisumą naudojamas *cron.js* skriptas. Ši programa kas valandą vykdo *collectDayMatches.js* skriptą. Programa parsisiunčia ir analizuoja puslapio html kodą. Html kode

randama ir iškoduojama (*angl. parse*) dienos varžybų informaciją, pagal gautą informaciją atnaujinami duomenų bazės dokumentai. Taip pat šio skripto vykdymo metu patikrinama ar duomenų bazėje saugoma visa reikalinga informaciją apie varžybų istorinius susitikimus. Programai pastebėjus jog varžybos yra žaidžiamos lygoje kuri dar nėra užregistruota programos duomenų bazėje, iškviečiama funkcija atsakinga už naujos lygos sukūrimą ir išsaugojimą. Taip pat kaip ir aptikus duomenų bazėje neegzistuojančią lygą, dienos varžybų surinkimo metu aptikus naujas komandas, jos išsaugomos duomenų bazėje. Be to šis skriptas tikrina ar surinkti dienos varžybų komandų istoriniai duomenys, jai ne, tuomet pradeda šių duomenų surinkimą. Esamoje sistemos konfigūracijoje į duomenų baze įtraukiama visos komandų varžybos nuo 2013 metų sezono.

3.5.3 Sporto varžybų baigčių prognozių skaičiavimas

Kai turima visa reikalinga informacija apie dienos varžybas, tuomet vykdomas *setupDayMatches.js* skriptas. Šio kodo tikslas apskaičiuoti dienos varžybų prognozes ir jas išsaugoti duomenų bazėje. Skriptas konfigūruojamas nurodant kokius prognozavimo modulius naudoti skaičiuojant varžybų prognozes. Suskaičiuotos prognozės saugomos duomenų bazėje pagal *IMatch* duomenų struktūrą. Ši informacija aplikacijoje naudojama pateikiant varžybų prognozes naršyklės programoje bei vertinant prognozių modelius. Be to skripto vykdymo metu patikrinama ar lygos jau įvertintos, jai lyga neturi įvertimo pagal tam tikrą prognozavimo modelį tuomet skriptas iškviečia ir įvykdo funkciją atsakinga už šia užduotį. Funkcija įvertinimą išsaugo duomenų bazėje pagal *PredictionModelEstimate* schemą. Šie duomenys veliau pateikiami interneto naršyklės programoje.

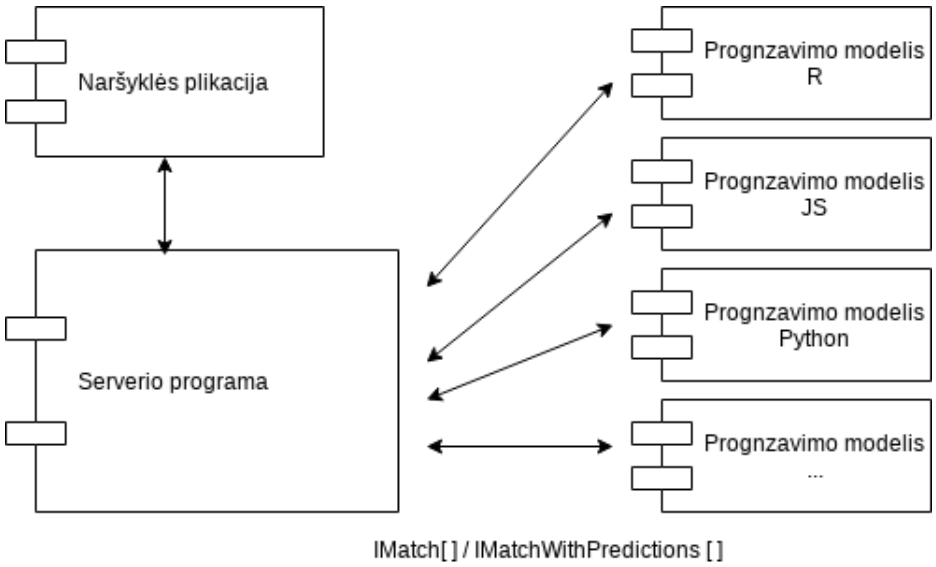
3.6 Lažybų organizatorių koeficientų surinkimas

Programoje varžybų koeficientai surenkami iš interneto puslapių *html* kodo. Šiai užduočiai sukurta *fastAddOdds* funkcija. Funkcija kaip argumentą priima varžybų masyvą, tuomet surenka masyve esančių varžybų koeficientus. Paprastai ši funkcija naudojama surenkant dienos varžybų duomenis ar prognozių skaičiavimo metu.

3.7 Sistema ir prognozavimo modeliai

Sukurtoje automatizuotoje sporto baigčių prognozavimo sistemoje vienas iš pagrindinių sistemos elementų - prognozavimo modeliai. Projektuojant sistemą nuspręsta neprisirišti prie vienos programavimo kalbos įgyvendinant prognozavimo algoritmus. Šis sprendimas leidžia programuotojams kurti ir testuoti prognozavimo modelius sukurtoje sistemoje įvairiomis programavimo kalbomis.

Kaip matome 8 pav. pavaizduotoje schemoje, prognozavimo modeliai veikia kaip atskiri servisai. Sistema su šias komponentais bendrauja *HTTP* protokolo pagalba. Į servisą siunčiami du masyvai su varžybų sąrašais. Viename siunčiamas prognozuojamų varžybų sąrašas. Antrame masyvas su istoriniais varžybų duomenimis, šie duomenys naudojami vertinant varžybų baigčių tikimybes. Iš serviso tikimasi gauti varžybų prognozių masyvą. Atliekant tyrimą pasirinkti matematiniai prognozavimo modeliai įgyvendinti *JavaScript* programavimo kalba. Sukurto *NodeJS*



8 pav. Sistemos ir prognozavimo modelių servisų koncepcinė diagrama.

serviso pagalba vyko bendravimas tarp prognozavimo sistemos ir prognozavimo serviso. Prie sistemos prijungti servisai matomi lygos puslapyje esančiame *dropdown* elemente.

4 Prognozavimo modeliai

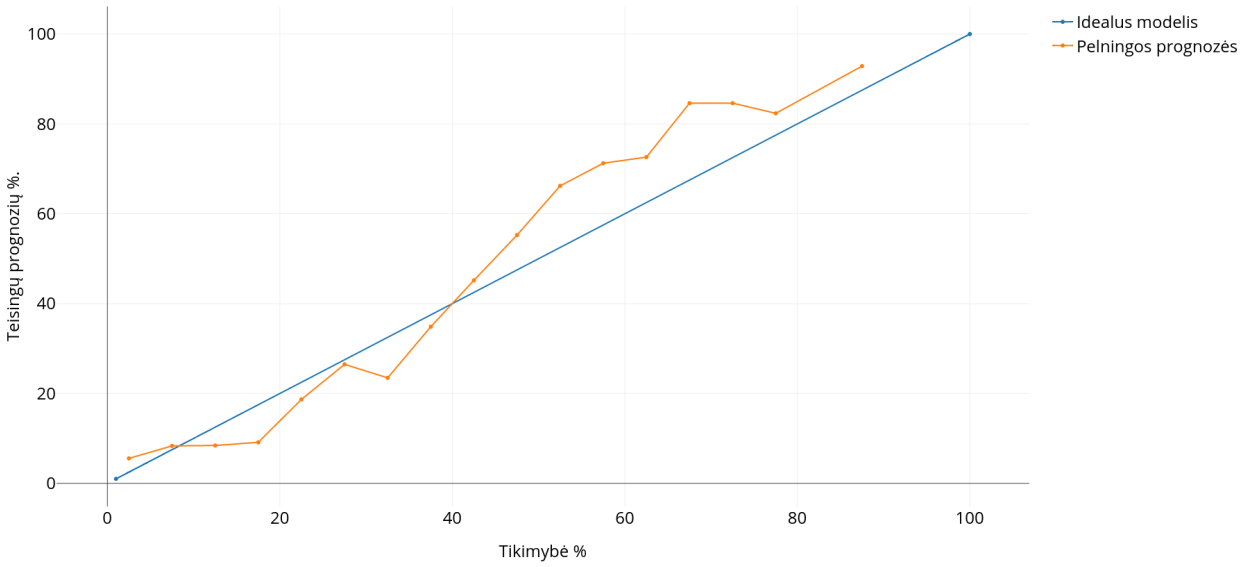
Atlikus literatūros analizę matome, kad egzistuoja įvairių modelių, skirtų prognozuoti futbolo varžybose pelnytų įvarčių skaičių ar jų baigtį. Juos galima skirstyti į rūšis pagal naudojamus matematinius ir statistinius metodus, intensyvumo parametrų tipus ir jų kiekį, taip pat pagal apskaičiuotų prognozių tipą, naudojamus duomenis, tikslumo ar pelningumo rodiklius. Tačiau dažniausiai sutinkami Puasono pasiskirstymo dėsnio pagrindu veikiantys sporto varžybų rezultatų prognozavimo modeliai. Šie modeliai yra lengvai įgyvendinami, paprasti ir aiškūs. Dėl šių priežasčių modelių charakteristikų tyrimas įgyvendintas šių modelių kontekste. Modeliams vertinti buvo sukurti metodai ir įrankiai, kai modeliai vertinami pagal jų tikslumą ir pelningumą.

4.1 Prognozavimo modelių vertinimas

Sporto varžybų baigčių prognozavimo modelių vertinimui sukurtas ir naudotas įrankis leidžiantis prognozavimo modelius vertinti pagal jų tikslumą ir investicijos grąžą.

4.1.1 Modelių vertinimas pagal prognozių tikslumą

Modelių vertinimui buvo sukurtas metodas, kai modeliai vertinami pagal jų prognozių tikslumą. Geriausiu atveju modelio rezultatų tikslumo įvertinimo grafikas (9. pav) - tolydi tiesė. Iš grafiko galima spręsti apie modelio nukrypimus nuo norimų rezultatų. $x \in [0, 100]\%$ ašys atitinka varžybų baigčių tikimybių reikšmių intervalus, $y \in [0, 100]\%$ - realus prognozių pasiteisinimo dažnis.



9 pav. Pelningo prognozių modelio grafiko pavyzdys.

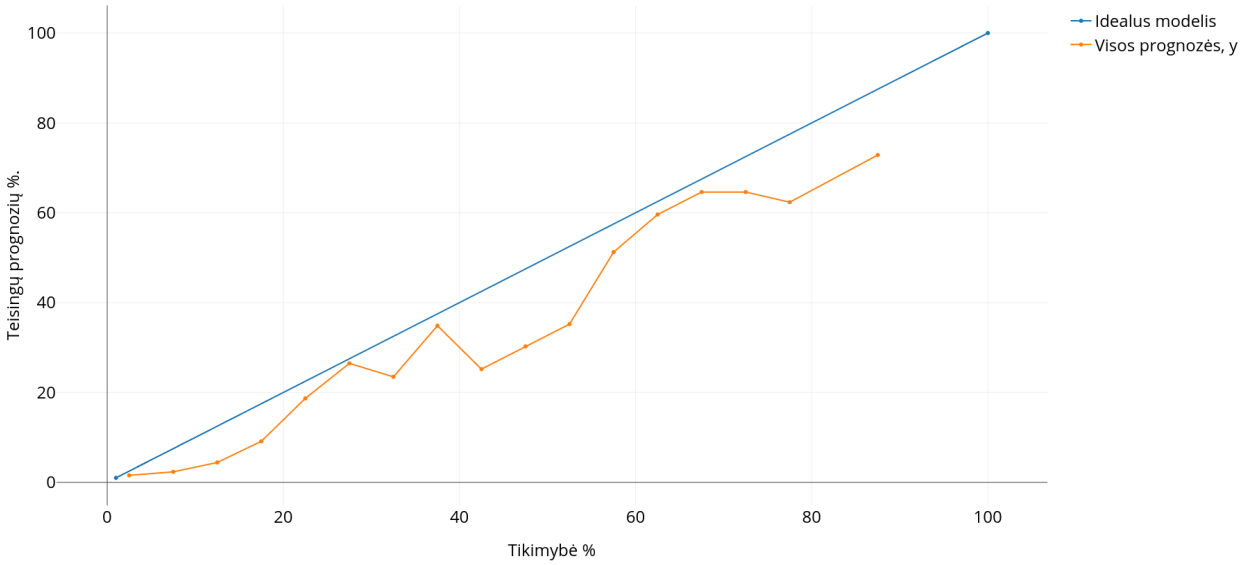
Vertinimo modelis kaip parametrus priima varžybų (su įvertintomis prognozėmis) masyvą ir prognozuojamą baigtį (*home, draw, away)*). Kaip rezultatą gražiną duomenis atitinkančius *IResponse* duomenų struktūrą. Mažas nuokrypis nuo norimo (žalia kreivė) grafiko rodo, jog modelis tiksliai prognozuoja varžybų baigtis ir atvirkščiai, didelis nuokrypis - netikslaus prognozav-

imo modelio požymis. Vertinimo modelio gražinami parametrai: nuokrypių suma (1), vidutinis nuokrypis (2), didžiausias nuokrypis (nuokrypiai skaičiuojamas nuo etaloninio modelio grafiko ($x = y$)).

$$D_{L_1} = \sum_{i=1}^{N} |I(p_i) - \hat{R}(p_i)| \tag{1}$$

$$D_{avg.} = \frac{1}{N} \sum_{i=1}^{N} (I(p_i) - \hat{R}(p_i)) \tag{2}$$

$$D_{max.} = \max_{i=1,\dots,N} (I(p_i) - \hat{R}(p_i)) \tag{3}$$



10 pav. Nuostolingo prognozių modelio grafiko pavyzdys.

Sukurtame programiniame įrankyje skirtame pateikti prognozių modelių tikslumo grafikus yra galimybe nurodyti mažiausią sužaistų varžybų skaičių procentų intervale, kad rezultatai būtų laikomi patikimais. Šiame darbe naudotas 10 varžybų minimumas. Prognozavimo modelio vertinimo grafikas laikomas patikimu jai varžybų duomenų pakanka įvertinti 10 tikimybių intervalų (t.y. 0-10%, 10-20%, 20-30% ... 90-100%).

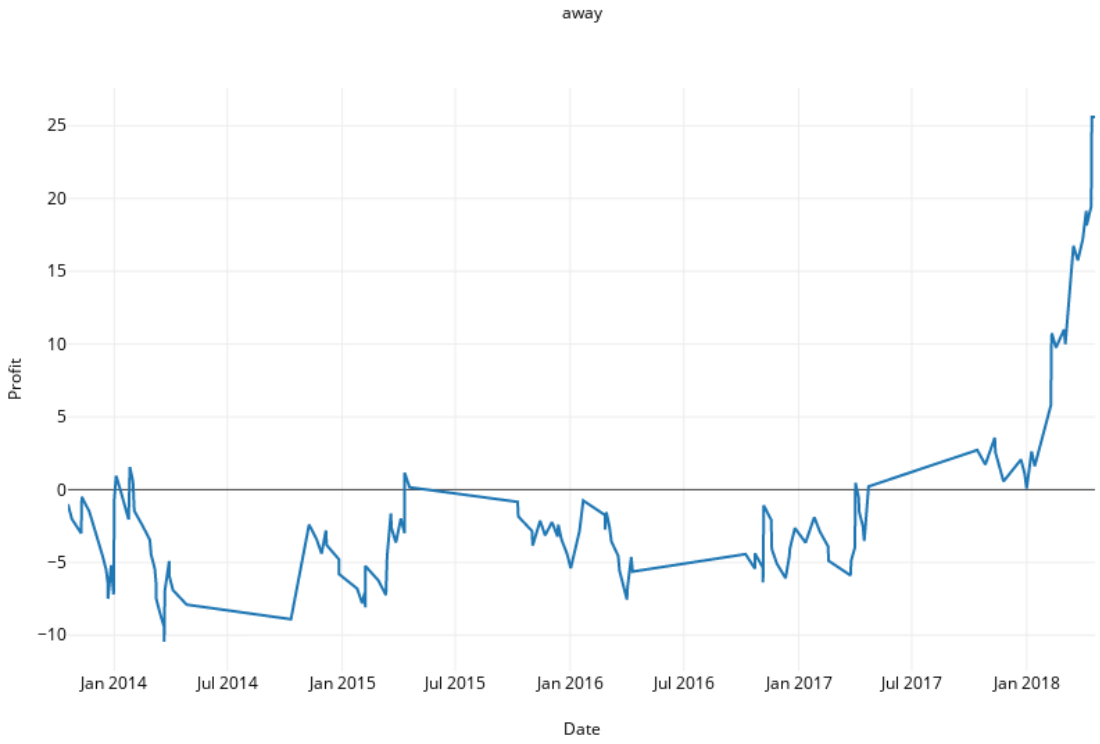
4.1.2 Modelių vertinimas pagal prognozių pelningumą

Sukurtoje automatizuotoje sporto varžybų baigčių prognozavimo sistemoje naudojamas metodas kai prognozavimo algoritmai yra vertinami pagal jų pelningumą Pr_m. Vartotojas modelio pelningumą gali įvertinti pasirenkant skirtingas charakteristikas: prognozių tikimybių intervalas, lažybų organizatorių ir prognozavimo modelio įvertintų baigčių tikimybių skirtumas, prognozavimo modelis, prognozuojamos baigties tipas.

$$\text{Pr}_m = \sum_{i=1}^{n} \left(\mathbb{1}\left\{ \textit{i-oji teisinga prognozė} \left(o_i - s \right) \right\} - \mathbb{1}\left\{ \textit{i-oji neteisinga prognozė} - s \right\} \right)$$

$$\mathbb{1}(x) = \begin{cases} 1, & \text{jei teisinga} \\ 0, & \text{jei neteisinga} \end{cases}$$

Kur s - lažybų suma, n - prognozių skaičius.



11 pav. Modelio pelningumo kitimo grafikas.

Grafikas (11 pav.) sudaromas atsižvelgiant į pasirinkto modelio pelno kitimą laike. Iš grafiko vartotojas gali įžvelgti ar pasirinktas prognozavimo modelis pelno stabilumo rodiklius.

4.2 Modelių vertinimo rezultatai

Atlikus literatūros analizę ir įvertinus įvairius prognozavimo modelius buvo nuspręsta tyrimui pasirinkti Poisono pasiskirstymo dėsnio pagrindu paremtus modelius ir jų variacijas. Skyrelyje skirtame lygų prognozuojamumo indeksui įvertinti nustatyta, jog lažybų organizatoriams sunkiausiai pavyksta nustatyti *Taça de Portugal* lygos varžybų tikimybes. Todėl nuspręsta plačiau ištirti Puasono pasiskirstymo dėsnio pagrindu sudarytus modelius šios lygos kontekste, tikintis tiksliau prognozuoti šios lygos baigtis, nei tai daro lažybų organizatoriai.

Šiam tikslui *JavaScript* programavimo kalba sukurtas varžybų baigčių prognozavimo servisas. Servise įgyvendintos Maherio [Mah82] darbe aprašyto algoritmo variacijos. Šis servisas

sujungtas su sukurta prognozavimo ir analizės sistema. Gauti rezultatai įvertinti naudojantis sukurta naršyklės programa.

Prognozavimo modelio pagrindą sudaro komandų puolimo ir gynybos stipriai: α_i - namų komandos puolimo stiprio rodiklis, β_j - svečių komandos gynybos stiprio rodiklis, γ_i - namų komandos gynybos stiprio rodiklis, δ_j - svečių komandos puolimo stiprio rodiklis.

$$PR(X_i = x, Y_j = y) = Poison(\alpha_i \beta_j) Poison(\gamma_i \delta_j) \tag{4}$$

$$S_x = \sum_i \sum_j x_{ij} \tag{5}$$

$$S_y = \sum_i \sum_j y_{ij} \tag{6}$$

$$\alpha_i = \sum_i x_j / \sqrt{S_x} \tag{7}$$

$$\beta_j = \sum_i x_{ij} / \sqrt{S_x} \tag{8}$$

$$\gamma_j = \sum_j x_{ij} / \sqrt{S_y} \tag{9}$$

$$\delta_j = \sum_j x_{ij} / \sqrt{S_y} \tag{10}$$

Šiame prognozavimo modelyje naudojami komandų stipriai sudaryti ne pagal visos lygos varžybas, o pagal konkrečių komandų istorinius duomenis. Pirmiausia modelis ištirtas naudojant komandų istorinius duomenis nepriklausomai nuo ar komandą žaidė namuose ar išvykoje. Atliktame tyrime įvert kokią įtaką turi naudojamų istorinių duomenų skaičius prognozavimo rezultatams. Šio modelio varžybų baigčių prognozavimui naudojamas Puasono tikimybių pasiskirstymo dėsnis.

$$PR(X_i = x, Y_j = y) = Poison(\alpha_i \beta_j) Poison(\gamma_i \delta_j) \tag{11}$$

Susitinkant komandoms i ir j, šių komandų stipriai yra apskaičiuojami pagal formules:

$$\alpha_i^{(t)} = \sum_{m=1}^{n} X_i^{t-m} \tag{12}$$

$$\beta_j^{(t)} = \sum_{m=1}^{n} Y_j^{t-m} \tag{13}$$

$$\gamma_i^{(t)} = \sum_{m=1}^{n} Y_i^{t-m} \tag{14}$$

$$\delta_j^{(t)} = \sum_{m=1}^{n} X_j^{t-m} \tag{15}$$

n - skaičius nurodo kiek istorinių duomenų reikia įtraukti i formulę, HG - namų komandos įvarčių skaičius istoriniuose duomenyse, AG - svečių komandos įvarčių skaičius. Komandos istorinių duomenų sąrašas sudaromas neatsižvelgiant kur komanda žaidė, t.y. namuose ar išvykoje.

Šis modelis įvertintas su $n = 1, 3, 5$

Vertinimo lentelėse naudojami trumpiniai atitinkantys: v.v- viso varžybų, v.sk - vidutinis nuokrypis, m.sk - didžiausiais nuokrypis, *ROI* - investicijos grąžos rodiklis, T - prognozių tipas, A - visos prognozės, V - potencialiai pelningos.

Modelio rezultatai vertinami pagal du skirtingus parametrų nustatymus. Pirmuoju atveju vertinamos visos namų komandos pergalės prognozės. Antruoju atveju vertinamos tik tos namų komandos pergalės prognozės kurių tikimybė šiai baigčiai buvo aukštesnė nei lažybininkų tikimybė. Taip pat antruoju atveju vertinamos tik tos baigtys kurioms lažybų agentūros suteikė 20-80% tikimybes. Šis intervalas pasirinktas atlikus *Aça de Portugal* prognozuojamumo vertinimą (plačiau skyrelyje "Prognozuojamumo indeksas").

4.3 Puasono v1 modelis

Ši modelio variacija kaip istorinius duomenis naudoja paskutiniais susitinkančių komandų varžybas. Modelis vertindamas nekreipia dėmesio ar paskutinės varžybos vyko namuose ar išvykoje.

Lentelė 9. Puasono v1 modelio rezultatai prognozuojant *Taça de Portugal* lygos varžybų namų komandos pergalę.

Lažybininkų tikimybių intervalas	min. tikimybių skirtumas	vidutinis nuokrypis	nuokrypių suma	didžiausias nuokrypis	viso prognozių	visas pelnas	ROI
0-100 %	-	11,5	46	21	433	56.97	13%
20-80 %	0.05	11,2	56	21	96	46,42	48%

Kaip matome iš rezultatų lentelės (10 lentelė), Puasono v1 modelis viso apskaičiavo 433 varžybų namų komandos pergalės tikimybes. Vidutinis tikimybių nuokrypis nuo modelio prognozuojamos tikimybės 11,5 %. Taip pat iš šių prognozių buvo atrinktos tik prognozės (antra 10 lentelės eilutė) su tikimybe aukštesne nei lažybininkų tai pačiai baigčiai suteiktos tikimybes. Be to lažybininkų suteiktų tikimybių vertės turėjo svyruoti tarp 20-80%. Šiems reikalavimams atitiko 96 prognozės. Šios modelio apskaičiuotos tikimybės vidutiniškai skyrėsi 11,2% nuo realių tikimybių. Iš gautu rezultatų reikėtu atkreipti dėmesį į pelną apskaičiuotą darant prielaidą, jog buvo lažintasi pagal modelio baigčių prognozes. Kadangi lažybininkams sunkiai sekėsi prognozuoti varžybų baigtis, lažintis už visų varžybų namu komandų pergales, pasiektas 56,97 vienetų arba 13% nuo statumos sumos pelnas. Tačiau pasirinkus tik prognozes su tikimybėmis aukštesnėmis už lažybininkų tikimybes pelnas atitiktu 46,42 vienetus arba 48% nuo statomos sumos.

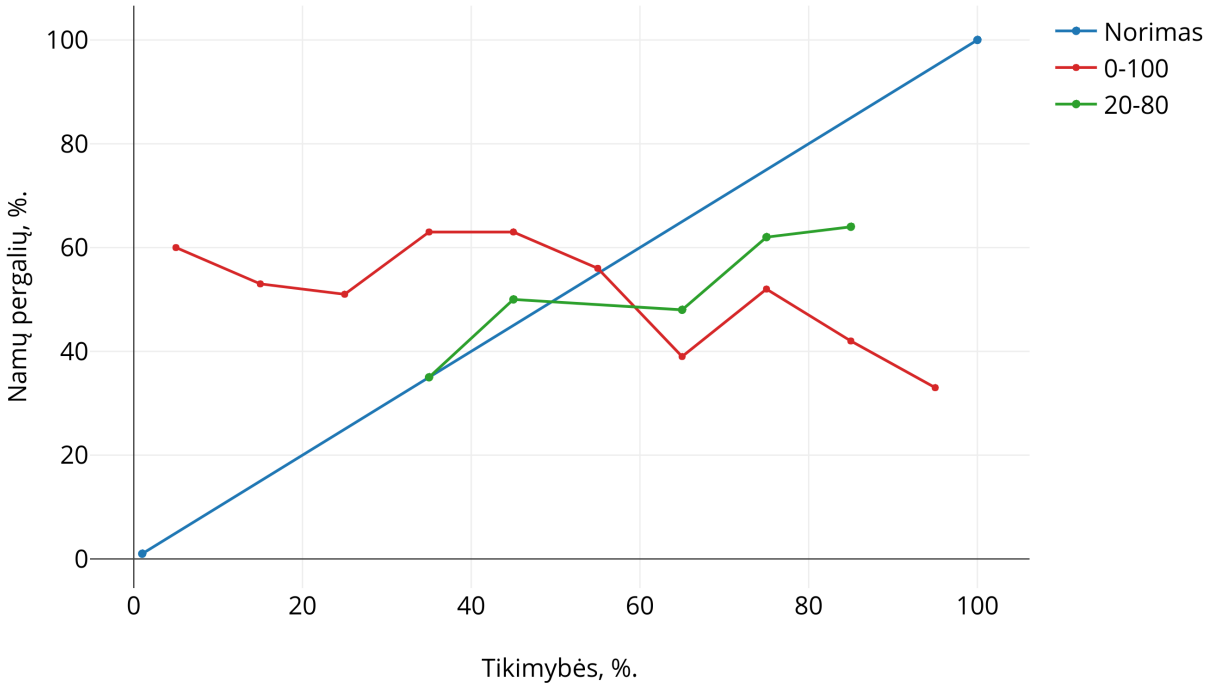
11 lentelėje matome kaip pasiskirstė Puasono v1 modelio tikimybės pagal tikimybių intervalus. Daugumos varžybų namų komandų pergalėms apskaičiuota tikimybė svyravo tarp 60-80% (99 varžybos), mažiausiai tarp 0-20 %. Tiksliausiai prognozuota baigtys su tikimybėmis nuo 40 - 60 %. Šio intervalo varžybos namų komandos pergalė baigėsi 61% kartų. Blogiausias tikslumas užfik-

Lentelė 10. Puasono v1 modelio tikimybių intervalų vertinimo rezultatai.

intervalas		0-20%	20-40%	40-60%	60-80%	80-100%
0-100	namų pergalė	55% +45	59% +29	61% +11	43% -27	40% -50
	viso prognozių	20	36	61	99	69
20-80	namų pergalė	-	30% 0	55% +5	53% -17	56% -34
	viso prognozių	-	20	20	18	18

suotas vertinant baigtis aukštomis tikimybėmis, namų komandos pergalę šventė 40%, nors modelio apskaičiuotos tikimybės baigtis įvertino 80-100%. Stebint kaip modelis suteikia žemas tikimybes namų komandos pergalei, matome, jog tikimybės yra mažesnės 30-45, nors mažų tikimybių varžybų prognozuota mažiausiai.

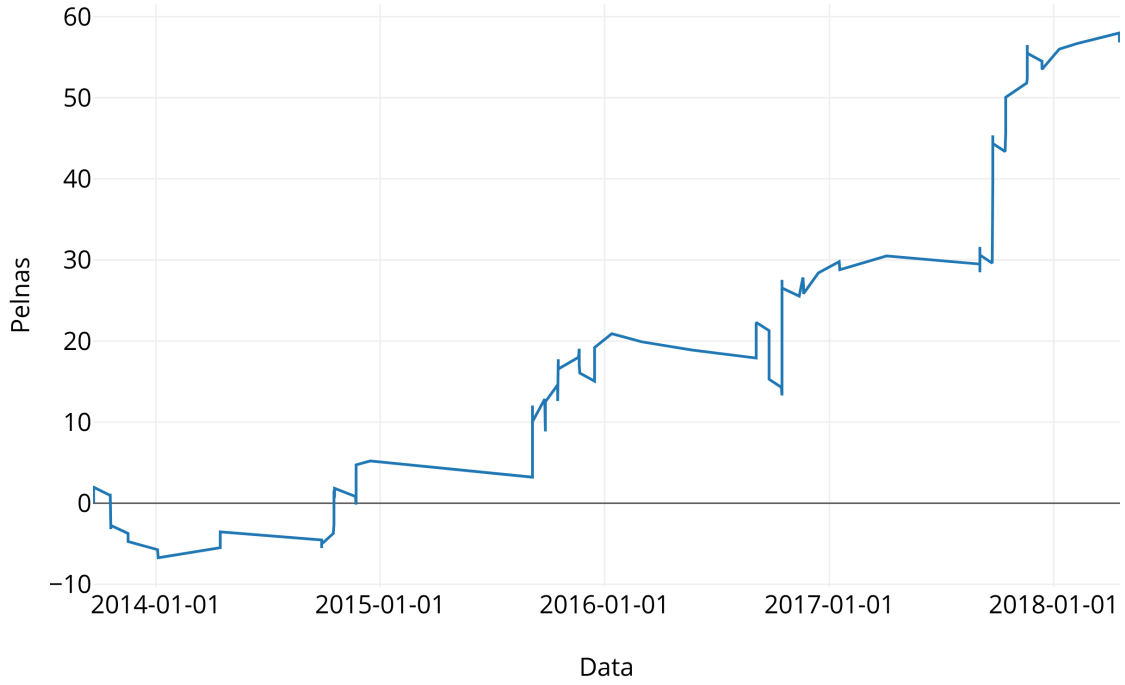
Vertinant modelį pagal varžybas atrinktas pagal parametrus (lažybininkų tikimybių intervalą (20 -80%) ir tik varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais), matome jog varžybų skaičius sumažėjo daugiau nei 4 kartus nuo 433 iki 96. Modelis labai tiksliai (paklaida +5%) apskaičiavo varžybų tikimybes intervale 20-60%. Prastesni rezultatai matomi prognozuojant namų komandos pergales su tikimybėmis įvertintomis intervale nuo 60% iki 100 %. Šiame intervale esančios baigtys vidutiniškai buvo pervertinamos 25%.



12 pav. Puasono V1 modelio prognozių tikimybių pasiteisinimo grafikas.

Grafike (12 pav.) matome Puasono v1 modelio varžyboms apskaičiuotų tikimybių pasiteisinimo procentą. Raudona kreivė atitinka visų modelio prognozių rezultatus, o žalia pagal parametrus

(lažybininkų tikimybių intervalą (20 -80%) ir varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais) atrinktų varžybų rezultatus. Iš grafiko galime pastebėti jog modeliui prastai sekasi nustatyto žemų ir aukštų tikimybių baigtis.



13 pav. Puasono V1 modelio pelningumo laike grafikas.

Iš pelno kitimo laike grafiko (13 pav.) galime pastebėti ,kad grafiko kreivė dėsningai kyla į viršų. Tai rodo ,jog naudodamiesi modelio prognozėmis galime tikėtis pastovaus pelno ilguoju laikotarpiu. Matomas grafikas sudarytas atrenkant prognozes kurios atitinka parametrus (lažybininkų tikimybių intervalą (20 -80%).

Gauti rezultatai nuteikia optimistiškai ieškoti prognozavimo modelių gebančių dar tiksliau prognozuoti šios futbolo lygos namų komandos pergales.

4.4 Puasono v2 modelis

Ši prognozių modelio variacija kaip istorinius duomenis naudoja paskutinių trijų varžybų rezultatus. Modelis vertindamas nekreipia dėmesio ar paskutinės varžybos vyko namuose ar išvykoje.

Kaip matome iš rezultatų lentelės (10 lentelė), Puasono v2 prognozavimo modelis apskaičiavo 333 varžybų namų komandos pergalės tikimybes. Vidutinis tikimybių nuokrypis nuo modelio prognozuojamos tikimybės 20,67 %. Taip pat iš visų prognozių atrinktos tik prognozės (antra 10 lentelės eilutė) su tikimybe aukštesne nei lažybininkų suteiktos tikimybes tai pačiai baigčiai. Be to atrinktos lažybininkų suteiktų tikimybių vertės svyruoja intervale tarp 20-80%. Šiems reikalavimams atitiko 69 prognozės. Šios modelio apskaičiuotos tikimybės vidutiniškai skyrėsi 10,5% nuo realių tikimybių. Iš gautu rezultatų reikėtu atkreipti dėmesį į pelną apskaičiuotą darant prielaidą

Lentelė 11. Puasono v2 modelio rezultatai prognozuojant Portugalijos Taça de Portugal lygos varžybų namų komandos pergalę.

Lažybininkų tikimybių intervalas	min. tikimybių skirtumas	vidutinis nuokrypis	nuokrypių suma	didžiausias nuokrypis	viso prognozių	visas pelnas	ROI
0-100 %	-	20.67	124	42	333	84.83	25%
20-80 %	0.05	10.5	42	15	69	44.55	65%

,jog lažintasi pagal modelio baigčių prognozes. Kadangi lažybininkams sunkiai sekėsi prognozuoti varžybų baigtis, lažintis už visų varžybų namu komandų pergales, pasiektas 84,83 vienetų arba 25% pelnas nuo statumos sumos. Tačiau pasirinkus prognozes su tikimybėmis aukštesnėmis už lažybininkų tikimybes pelnas atitiktu 44,55 vienetus arba 65% nuo statomos sumos.

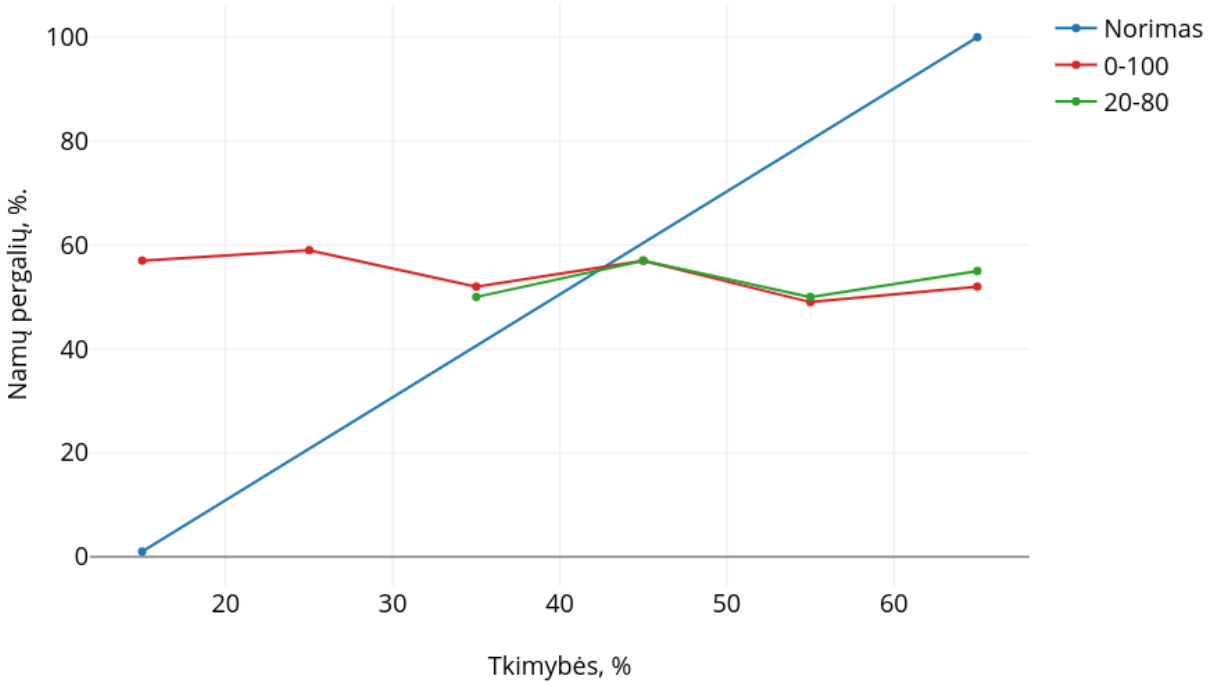
Lentelė 12. Puasono v2 modelio tikimybių intervalų vertinimo rezultatai.

intervalas		0-20%	20-40%	40-60%	60-80%	80-100%
0-100	namų pergalė	54% +44	55% +25	54% +4	47% -23	-
	viso prognozių	28	148	123	34	-
20-80	namų pergalė	-	47% +17	54% +4	40% -17	-
	viso prognozių	-	19	35	15	-

13 lentelėje matome kaip pasiskirstė Puasono v2 modelio tikimybės pagal tikimybių intervalus. Daugumos varžybų namų komandų pergalėms apskaičiuota tikimybė svyravo tarp 40-60% (123 varžybos), mažiausiai tarp 0-20 % (28 varžybos). Tiksliausiai prognozuota baigtys su tikimybėmis nuo 40 - 60 %. Šio intervalo varžybos namų komandos pergalė baigėsi 54% kartų. 40-60% intervale aptariamas prognozavimo modelis, nuo norimo prognozavimo tikslumo nukrypo tik 4%. Blogiausias tikslumas užfiksuotas vertinant baigtis su žemomis tikimybėmis, namų komandos pergalę šventė 54%, nors modelio apskaičiuotos tikimybės baigtis įvertino 0-20%, tai rodo jog prognozių modelis suteikia mažesniais tikimybes, nei tikroji jų reikšmė. Stebint kaip modelis suteikia aukštas tikimybes namų komandos pergalei, matome, jog tikimybės yra mažesnės vidutiniškai 23%, nors aukštų tikimybių varžybų prognozuota mažiausiai.

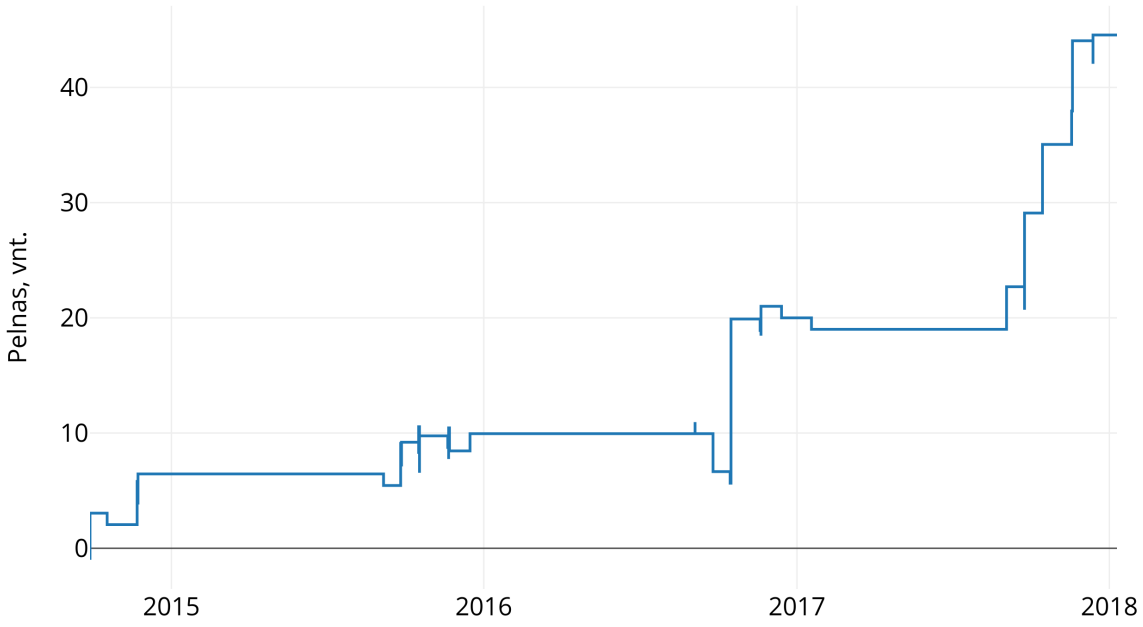
Vertinant modelį atrinkus varžybas pagal parametrus (lažybininkų tikimybių intervalą (20 -80%) ir tik varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais), matome jog varžybų skaičius sumažėjo beveik 5 kartus nuo 333 iki 69. Aptariamas prognozavimo modelis labai tiksliai (paklaida +4%) apskaičiavo varžybų tikimybes intervale 40-60%. Prastesni rezultatai matomi prognozuojant namų komandos pergales su tikimybėmis įvertintomis intervaluose 20-40% ir 60-80% . Šiame intervale esančių vidutinis nuokrypis siekia 17%.

Grafike (14 pav.) pavaizduota Puasono v2 modelio varžyboms apskaičiuotų tikimybių pasiteisinimo procentą. Raudona kreivė atitinka visų modelio prognozių rezultatus, o žalia pagal parametrus (lažybininkų tikimybių intervalą (20 -80%) ir varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais) atrinktų varžybų rezultatus. Iš grafiko galime pastebėti jog aptariamo prognozavimo modelio prognozės pasiteisina apie 50% nepriklausomai



14 pav. Puasono V2 modelio tikimybių pasiteisinimo procentų grafikas.

nuo apskaičiuotų tikimybių,



15 pav. Puasono V2 modelio pelningumo laike grafikas.

Iš pelno kitimo laike grafiko (15 pav.) galime pastebėti ,kad grafiko kreivė dėsningai kyla į viršų. Tai rodo ,jog naudodamiesi modelio prognozėmis galime tikėtis pastovaus pelno ilguoju laikotarpiu. Matomas grafikas sudarytas atrenkant prognozes atitinkančias parametrus: lažybininkų tikimybė patenka į intervalą tarp 20 -80% bei modelio apskaičiuota tikimybė aukštesnė už lažybininkų apskaičiuotą tikimybę.

4.5 Puasono v3 modelis

Lentelė 13. Puasono v3 modelio rezultatai prognozuojant Portugalijos Taça de Portugal lygos varžybų namų komandos pergalę.

Lažybininkų tikimybių intervalas	min. tikimybių skirtumas	vidutinis nuokrypis	nuokrypių suma	didžiausias nuokrypis	viso prognozių	visas pelnas	ROI
0-100 %	-	14.5	58	21	238	53.6	23%
20-80 %	0.05	4	8	5	40	22.79	57%

Kaip matome iš rezultatų lentelės (14 lentelė), Puasono v3 modelis viso apskaičiavo 238 varžybų namų komandos pergalės tikimybes. Vidutinis tikimybių nuokrypis nuo modelio prognozuojamos tikimybės 14,5 %. Taip pat iš šių prognozių atrinktos tik prognozės (antra 14 lentelės eilutė) su tikimybe yra aukštesne nei lažybininkų tai pačiai baigčiai suteiktos tikimybes. Be to lažybininkų suteiktų tikimybių vertės turėjo svyruoti tarp 20-80%. Šiems reikalavimams atitiko 40 prognozių. Modelio apskaičiuotos tikimybės vidutiniškai skyrėsi 14,5% nuo realių tikimybių. Iš gautu rezultatų reikėtu atkreipti dėmesį į pelną apskaičiuotą darant prielaidą ,jog buvo lažintasi pagal modelio baigčių prognozes. Kadangi lažybininkams sunkiai sekėsi prognozuoti varžybų baigtis, lažintis už visų varžybų namu komandų pergales, pasiektas 53,6 vienetų arba 23% nuo statumos sumos pelnas. Tačiau pasirinkus tik prognozes su tikimybėmis aukštesnėmis už lažybininkų tikimybes pelnas atitiktu 22,79 vienetus arba 57% nuo statomos sumos.

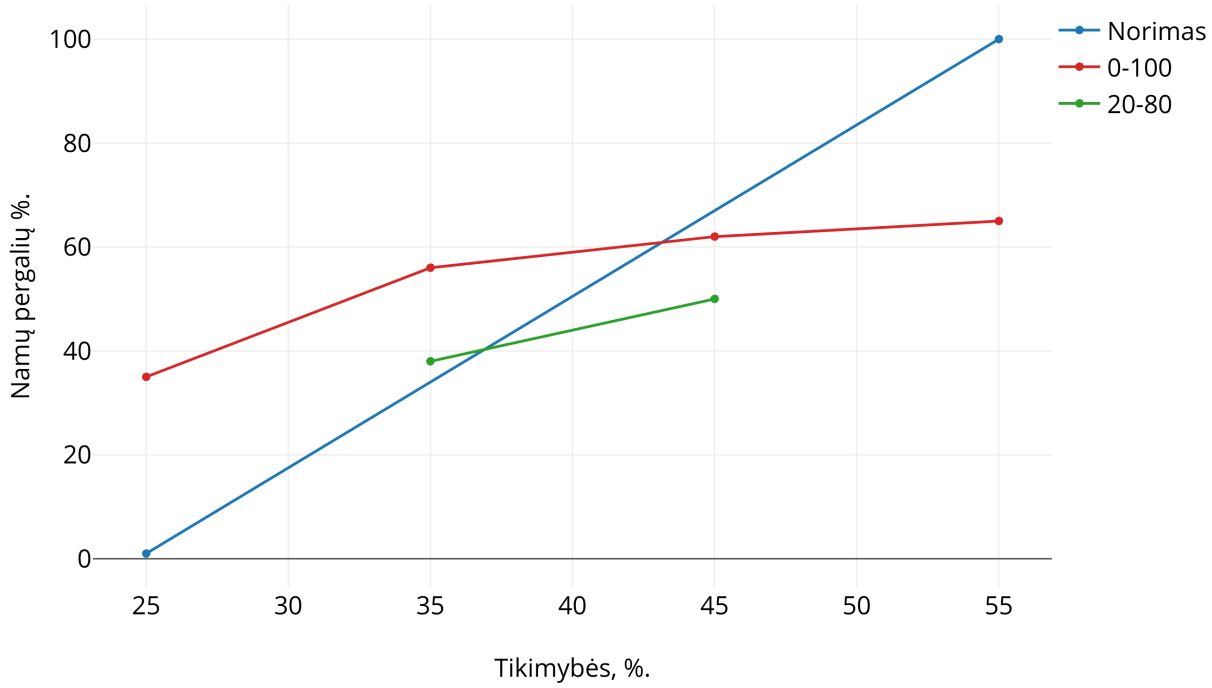
15 lentelėje matome kaip pasiskirstė Puasono v3 modelio tikimybės pagal tikimybių intervalus. Daugumos varžybų namų komandų pergalėms apskaičiuota tikimybė svyravo tarp 20-40% (136 varžybos), mažiausiai tarp 0-20 % (7 varžybos). Tiksliausiai prognozuota baigtys su tikimybėmis nuo 0-20 %. Šio intervalo varžybos namų komandos pergalė baigėsi 29% kartų. 40-60% intervale aptariamas prognozavimo modelis, nuo norimo prognozavimo tikslumo nukrypo 13%. Blogiausias tikslumas užfiksuotas vertinant baigtis su tikimybėmis esančiomis 60-80% intervale, namų

Lentelė 14. Puasono v3 modelio tikimybių intervalų vertinimo rezultatai.

intervalas		0-20%	20-40%	40-60%	60-80%	80-100%
0-100	namų pergalė	29% +9	48% +18	63% +13	33% -23	-
	viso prognozių	7	136	86	9	-
20-80	namų pergalė	-	36% +4	57% +7	33% -47	-
	viso prognozių	-	14	23	3	-

komandos pergalę šventė 33%, nors modelio apskaičiuotos tikimybės baigtis įvertino 60-80%, tai rodo jog prognozių modelis suteikia didesnes tikimybes, nei tikroji jų reikšmė.

Vertinant modelį tik atrinktoms varžyboms (lažybininkų tikimybių intervalą (20 -80%) ir tik varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais), matome jog varžybų skaičius sumažėjo beveik 6 kartus nuo 238 iki 40. Aptariamas prognozavimo modelis labai tiksliai (paklaida +4%) apskaičiavo varžybų tikimybes intervale 20-40%. Prastesni rezultatai matomi prognozuojant namų komandos pergales su tikimybėmis įvertintomis 60-80% intervale. Šiame intervale esančių vidutinis nuokrypis siekia 47%.



16 pav. Puasono V3 modelio tikimybių pasiteisinimo procentų grafikas.

Grafike (16 pav.) matome Puasono v3 modelio varžyboms apskaičiuotų tikimybių pasiteisinimo procentą. Raudona kreivė atitinka visų modelio prognozių rezultatus, o žalia pagal parametrus (lažybininkų tikimybių intervalą (20 -80%) ir varžybos su aukštesnėmis namų komandos pergalės tikimybėmis lyginant su lažybininkais) atrinktų varžybų rezultatus. Iš grafiko galime pastebėti jog modeliui prastai sekasi nustatyto žemų ir aukštų tikimybių baigtis. Taip pat galime pastebėti jog kreivės kyla į viršų. Toks kreivių elgesys - gero prognozių modelio pavyzdys.

4.6 Apibendrinimas

Sukurtas ir panaudotas programinis įrankis skirtas įvertinti lygos prognozuojamumo indeksą ir šio indekso stabilumą. Sukurta programinė priemonė leido atlikti visų futbolo lygų prognozuojamumo vertinimą. Indekso vertinimo metu surinkta ir įvertinta 500 tūkst. varžybų, bei apie 2 mln. lažybų tarpininkų siūlomų baigčių koeficientų. Gauti rezultatai parodė, kad pačios populiariausios

pasaulio lygos gerai išanalizuotos, todėl lažybų agentūros sugeba tikslai nustatyti šių lygų varžybų baigčių tikimybes, paklaidos svyruoja kelių procentų intervale. Tačiau atlikus tyrimą aptikta lažybininkų sunkiai prognozuojamų lygų. Sukurtas prognozuojamumo stabilumo indeksas parodo, kad lygos prognozuojamumo charakteristika yra dėsningumas, o ne atsitiktinumas. Indeksas įvertina lygos prognozuojamumo kitimą laiko atžvilgiu.

Atlikus Puasono pasiskirstymo dėsnio pagrindu sudarytų sporto varžybų baigčių prognozavimo modelių analizę, įvertinta prognozavimo modelių tikslumas bei pelningumas. v.n. - vidutinis nuokrypis, n.s. - nuokrypių suma, d.n. - didžiausias nuokrypis. *ROI* investicijos grąža.

Lentelė 15. Puasono modelio rezultatai prognozuojant namų komandos pergalę įvertintų tikimybių intervaluose.

modelis	prognozių tikimybių intervalas	min. tikimybių skirtumas	v.n.	n.s.	d.n.	viso prognozių	visas pelnas	ROI
P. v1	0-100 %	-	11,5	46	21	433	56.97	13%
P. v1	20-80 %	0.05	11,2	56	21	96	46,42	48%
P. v2	0-100 %	-	20.67	124	42	333	84.83	25%
P. v2	20-80 %	0.05	10.5	42	15	69	44.55	65%
P. v3	0-100 %	-	14.5	58	21	238	53.6	23%
P. v3	20-80 %	0.05	4	8	5	40	22.79	57%

15 lentelėje pateikti visų modelių tikslumo vertinimo rezultatai. Kaip matome mažiausias (4%) vidutinis nuokrypis nuo norimo tikslumo pasiektas naudojant V3 modelio variaciją, didžiausias (20,79%) fiksuotas naudojant V2 prognozavimo modelį. Daugiausia (433), bei mažiausiai (40) baigčių tikimybių įvertinta atitinkamai naudojant V1 ir V3 modelių variacijas. Didžiausias (84,83 vienetai) ir mažiausias (22,79 vienetai) pelnas pasiektas naudojant atitinkamai V2 ir V3 prognozavimo modelių variacijas. Vertinant pelną pagal investicijos grąžą matome, jog naudojant modelio v1 variaciją fiksuotą 13% investicijos grąža. Didžiausia investicijos grąža (65% nuo lažybų sumos) fiksuota naudojant v2 prognozavimo modelį.

Vertinant modelių tikslumą pagal tikimybių intervalus (16. lentelė) matome, jog tiksliausiai prognozavo V1 modelis, 20-40% intervale. Šiame intervale prognozių tikslumas idealus (paklaida 0%). Didžiausias (50%) nuokrypis nuo norimo tikslumo matomas vertinant tikimybes 80-100% intervale, naudojant V1 prognozavimo modelį. Mažiausias (3 prognozės) tikimybių tankis matomas V3 prognozių modelio 60-80% tikimybių intervale. Daugiausia, 148 prognozių, patenka į V2 prognozavimo modelio 20-40% intervalą.

Taip pat vienas iš šio darbo rezultatų - automatizuota futbolo baigčių prognozavimo, bei prognozavimo modelių ir futbolo lygų analizės sistema. Ši sistema jos vartotojui teikia sporto prognozes, o sporto modelių kūrėjams programinį įrankį, leidžiantį įvertinti modelių charakteristikas. Suprojektuota ir įgyvendinta programų sistema jos vartotojams pateikia kiekvienos dienos futbolo varžybų prognozes. Vartotojas, naudodamasis sistema, gali įvertinti įvairių modelių pelningumą ir tikslumą, įvertinti lygų prognozuojamumą ir jų prognozuojamumo pastovumą. Tai sistemos vartotojui suteikia galimybę lažintis už pelningas futbolo varžybų baigtis.

Lentelė 16. Puasono modelio tikimybių intervalų vertinimo rezultatai.

modelis	intervalas		0-20%	20-40%	40-60%	60-80%	80-100%
Puasono V1	0-100	namų pergalė	54% +44	59% +29	61% +11	43% -27	-
		viso prognozių	20	36	61	99	-
	20-80	namų pergalė	-	30% 0	55% +5	53% -17	-
		viso prognozių	-	20	20	18	-
Puasono V2	0-100	namų pergalė	54% +44	55% +25	54% +5	47% -23	-
		viso prognozių	28	148	123	34	-
	20-80	namų pergalė	-	47% +17	54% +4	40% -17	-
		viso prognozių	-	19	35	15	-
Puasono v3	0-100	namų pergalė	29% +9	48% +18	63% +13	33% -23	-
		viso prognozių	7	136	86	9	-
	20-80	namų pergalė	-	36% +4	57% +7	33% -47	-
		viso prognozių	-	14	23	3	-

5 Rezultatai ir išvados

Atlikus išsikeltus darbo uždavinius gauti **rezultatai**:

- Išbandyta ir įvertinta įvairios Puasono pasiskirstymo dėsniu paremtos modelių variacijos;
- Sukurtas įrankis skirtas futbolo varžybų rezultatų prognozavimo modelių charakteristikų vertinimui ir analizei;
- Sukurta automatizuota futbolo varžybų rezultatų prognozavimo sistema;
- Įvertinta futbolo lygų prognozavimo ir prognozavimo stabilumo charakteristikos;
- Atlikta futbolo rezultatų prognozavimo modelių pelningumo ir tikslumo charakteristikų analizės.

Iš gautų rezultatų galime daryti **išvadas**:

- Vienodi modeliai skirtingai veikia skirtingose futbolo lygose;
- Puasono pasiskirstymo dėsniu paremti modeliai gali būti sporto varžybų rezultatų prognozavimo modelių pagrindas;
- Ištirti Puasono pasiskirstymo dėsniu veikiantis modeliai tiksliausiai prognozuoja vidutinių tikimybių intervaluose;
- Futbolo lygos skiriasi savo prognozuojamumu;
- Egzistuoja sunkiai lažybų tarpininkams prognozuojamų futbolo lygų;
- Naudojant sukurtą sistemą galime lengvai įvertinti skirtingų prognozavimo modelių charakteristikas:
  - tikslumą;
  - pelningumą;
  - lygos prognozuojamumą;
  - lygos prognozuojamumo stabilumą.
- Panaudojus prognozuojamumo charakteristiką ir Puasono pagrindu sukurtus prognozavimo modelius galime sukurti pelningą lažinimosi strategiją.

Šaltiniai

[BHK09] Wolfgang Breuer, Guido Hauten, and Claudia Kreuz. Financial instruments with sports betting components: marketing gimmick or a domain for behavioral finance? *Journal of Banking & Finance*, 33(12):2241–2252, 2009.

[BKM17] Georgi Boshnakov, Tarak Kharrat, and Ian G McHale. A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466, 2017.

[Bro99] Sid Browne. Reaching goals by a deadline: digital options and continuous-time active portfolio management. *Advances in Applied Probability*, 31(2):551–577, 1999.

[DC97] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997.

[FT15] George Foroglou and Anna-Lali Tsilidou. Further applications of the blockchain. In *12th Student Conference on Managerial Science and Technology*, 2015.

[Hil74] ID Hill. Association football and statistical inference. *Applied statistics*:203–208, 1974.

[JCH+16] Viktor Jacynycz, Adrian Calvo, Samer Hassan, and Antonio A Sánchez-Ruiz. Betfunding: a distributed bounty-based crowdfunding platform over ethereum. In *Distributed Computing and Artificial Intelligence, 13th International Conference*, pp. 403–411. Springer, 2016.

[KL15] Siem Jan Koopman and Rutger Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):167–186, 2015.

[Kla62] Hugh J Klare. *Anatomy of prison*. Penguin Books, 1962.

[KN03] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[KR00] Leonhard Knorr-Held and Günter Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

[Mah82] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[Mas11] Mark Masse. *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. " O'Reilly Media, Inc.", 2011.

[Rau12] Guillermo Rauch. *Smashing Node. js: JavaScript Everywhere*. John Wiley & Sons, 2012.

[Sch15] Sue Schneider. Esport betting: the intersection of gaming and gambling. *Gaming Law Review and Economics*, 19(6):419–420, 2015.

[SJL16] Robert P Schumaker, A Tomasz Jarmoszko, and Chester S Labedz. Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decision Support Systems*, 88:76–84, 2016.

[Tho06] Edward O Thorp. The kelly criterion in blackjack, sports betting and the stock market. *Handbook of asset and liability management*, 1:385–428, 2006.

[TM54] Robert M Trueblood and Robert J Monteverde. A bibliography on the application of statistical methods to accounting and auditing. *Accounting Review*:251–254, 1954.

[Vog15] Nick Vogel. The great decentralization: how web 3.0 will weaken copyrights. *J. Marshall Rev. Intell. Prop. L.*, 15:136, 2015.