Budapesti Műszaki és Gazdaságtudományi Egyetem

TDK Dolgozat

Magyar bajnoki labdarúgó-mérkőzések eredményének előrejelzése valószínűségi és gépi tanulási modellekkel

Pintér József és Ragács Attila

Témavezető:

Molontay Roland

Tudományos segédmunkatárs, BME, Sztochasztika Tanszék

2020

Tartalomjegyzék

1. Bevezetés

Labdarúgó-mérkőzések végeredményeinek becslése már a 20. század közepe óta nagy népszerűségnek örvendő kutatási terület. A téma első igazán komoly áttörése Maher angol matematikushoz kötődik, aki 1982-ben Poisson-eloszlással modellezte a csapatok támadó és védekező készségeit, majd ezek segítségével becsülte meg az egyes csapatok által szerzett gólok mennyiségét. A kezdeti modell hibáinak kiküszöbölésére számtalan kísérlet született, ezek közül kiemelkednek Dixon és Coles modelljei. A dolgozatban azt is megvizsgáljuk, hogy hogyan teljesítenek a fenti modellek az OTP Bank Liga mérkőzésein, továbbá saját elgondolásaink alapján igyekszünk statisztikai és gépi tanulási modelleket fejleszteni.

Napjainkra a növekvő adatmennyiségnek köszönhetően számos sportegyesület alkalmaz erre építő szakembereket, akik elemzéseikkel a klubokat érintő fontos döntések meghozatalában nyújtanak segítséget. Ilyenek az edzésmunka optimalizálása, új játékosok leigazolása, csapatsportoknál az ideális taktika kiválasztása [1]. Nincs ez másképp a labdarúgásban sem, továbbá a világ sportjait tekintve igen jelentős bevételt hozó sportfogadási terület. Mindemellett mi magunk is rajongunk a sportért, ezért is esett rá a választásunk. A mi célkitűzésünk a szakértők, fogadóirodák oddsait, valószínűségeit is figyelembevéve olyan modellek építése, amelyek magas pontossággal tudják előrejelezni a mérkőzések végkimenetelét. Továbbá, ezen modellek felhasználásával megvizsgáltuk lehetséges-e nyereséges fogadási stratégia kidolgozása.

A dolgozat 2. fejezetében olvasható egy rövid áttekintés a korábbi munkákról a területen, ezután a 3. fejezetben a felhasznált adatról adtunk leírást. A 4-6. fejezetekben a megvizsgált és kidolgozott modellek működése szerepel, a 4. fejezetben a gépi tanulási alapú és az erre épülő fogadási stratégia alapötlete, az 5-6. fejezetekben a statisztikai modellek. A 7. fejezetben találhatók az eredmények, a 8. fejezet pedig tartalmazza a munka rövid összefoglalását és kitekintést a jövőbeni munkát illetően.

2. Irodalomáttekintés

Az elmúlt években jelentősen növekedett az érdeklődés a labdarúgó mérkőzésekhez köthető numerikus előrejelző modellek iránt. Megközelítették számos módon a feladatot, kezdve statisztikai módszerekkel [2], ezt követően a gépi tanulási algoritmusok kerültek előtérbe, napjainkban pedig javarészt neurális hálókat alkalmaznak [3].

A gépi tanulási módszerek közül az egyik legeredményesebb irány a Bayes-hálók használata. A Bayes-háló a változók oksági struktúráját leíró irányított körmentes gráf (DAG), amiben a gráf minden csúcsához tartozik egy valószínűségi változó [4]. A gráf (X,Y) irányított élének jelentése, hogy X közvetlenül függ Y-tól (Y szülője X). Ennek segítségével a változók együttes eloszlása a következő képlettel írható fel:

$$P(X_1, X_2, \ldots, X_n) = \prod_{s=1}^{n} P(X_s | X_{\Gamma(s)})$$

ahol $\Gamma(s)$ az s szüleinek halmaza

Egy ilyen munka az angol Premier League 16 szezonjának végeredményeit és a szakértők szubjektív véleményét vette alapul a modell építésekor [5]. A szerzők így egy a csapat erősségét kifejező értéket tudtak rendelni minden csapathoz, a rendszert pedig egy teljes tesztszezonon tesztelték. Ez a modell összevetve az akkori fogadóirodai eredményekkel igen jól teljesített.

Egy másik szintén Bayes-hálón alapuló megoldás ezt a módszert más algoritmusokkal hasonlítja össze a Tottenham Hotspur mérkőzésein az 1995-97es szezonok során. A szakértők gondosan válogatott tulajdonságokat alkalmaztak a modellezésre (kulcsjátékosok a pályán voltak-e, meccs helyszíne, a csapat egyes részeinek erőssége), így sikerült is minden más akkor elérhető rendszert túlszárnyalniuk. Azonban a szerzők bevallása szerint is ez a módszer az attribútumok érzékenysége és időhöz kötöttsége miatt nem vihető át eredményesen későbbi időszakokra [6].

Graham és Scott fogadóirodák rátáit vették alapul, illetve egy probit modellt építettek az angol liga csapatai erősségének felmérésére [7]. Az oddsokat szintén egy külön modell becsülte, majd a két rendszer kombinációját követően vizsgálták, lehet-e nyereséges fogadási tervet kidolgozni. Végül a vizsgált fogadóiroda számára feltételezhetően elérhető egyéb információk miatt arra jutottak, hogy nem lehetséges profitálni a rendelkezésre álló adatokkal és módszerekkel.

Világi és Sterbenz a magyar sport informatikai és analitikai helyzetét térképezték fel, illetve felmérték az adatfelhasználási lehetőségeket. Négy sportág (kosárlabda, kézilabda, jégkorong, röplabda) első osztályú egyesületeinél azt találták, hogy a kérdőívükre válaszoló csapatok 29%-a végez adatkezelési munkát [1].

3. Adatgyűjtés, adattisztítás, felderítő elemzés

Kutatásunkban OTP Bank Liga eredményeit vettük górcső alál. Az OTB Bank Liga az első osztályú magyar bajnokság, 1901 óta rendezik meg. Hagyományos nevén az NB1 napjainkban 12 csapatos, ebből az első 4 helyezett indulhat az európai kupákban. Jelenlegi szervezője a Magyar Labdarúgószövetség, nemzetközi szinten az Európai Labdarúgó-szövetség (UEFA) képviseli [8].

Nemzetközi szinten az OTP Bank Liga elemzése egy keveset kutatott feladat, így a legnagyobb adatbázisokat felsorakoztató oldalakon sem volt elérhető több szezonra vonatkozó részletes statisztika a bajnokságról. Ennek orvoslására úgy döntöttünk, hogy magunk gyűjtjük a megfelelő formátumban az adatot. A forrás az Eredmények.com weboldal volt, ahol elérhető az összes mérkőzés részletes statisztikája visszamenőleg a 2017-2018-as szezonig [9]. Az annál régebbi évadokban csupán az eredmények által rendelkezésre, ez pedig problémát okozott a gépi tanulási modell illesztésénél, hiszen a gaz-

dag adathalmaz az eredményes működés alapja, enélkül nem tudná a főbb összefüggéseket megtanulni a modell. Így csak a legutóbbi 3 szezon adatait használtuk fel (2017-18, 2018-19, 2019-20-as szezonok).

Az adatot webscraping technikával szereztük be Python segítségével. A BeautifulSoup csomagot használtuk az oldalak szövegtartalmának kiolvasására és a Selenium csomagot az oldalak közötti továbblépésre[10]. A kapott táblában szerepelnek minden mérkőzésről a legfontosabb információk, mint a meccset játszó csapatok, az eredmény, a meccs időpontja. Emellett a statisztikák, ilyenek a kapuralövések száma, labdabirtoklás, szögletek, szabadrúgások száma. Továbbá a táblázat tartalmazta a fogadóirodák által ajánlott oddsokat a hazai győzelem, döntetlen, vendég győzelem végkimenetelekre.

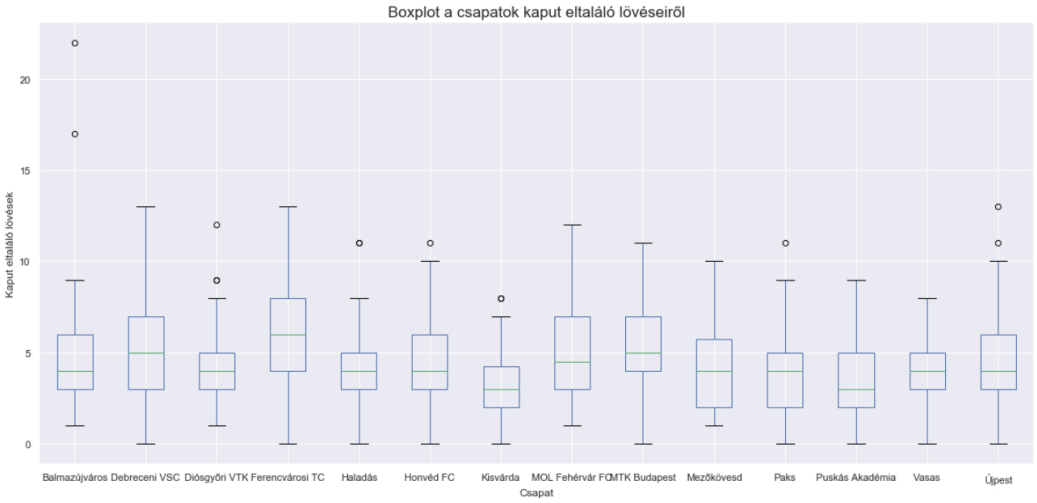
	MeccsID	Dátum	Csapat	Gólok	KapuraLövés	KaputEltaláLólövések	KaputNEMTL	Labdabirtoklás	Odds	Szabadrúgás	Szöglet	Védés	Bedobás	Hazaie
0	1	2017	Paks	2	9	4	5	54	2.75	23	9	0	30	1
1	1	2017	Újpest	2	10	5	5	46	2.38	22	1	3	21	0
2	2	2017	Debreceni VSC	1	18	9	9	48	1.80	16	8	1	36	1
3	2	2017	Mezőkövesd	2	9	6	3	52	4.20	20	3	2	22	0
4	3	2017	Honvéd FC	2	7	5	2	54	1.80	24	7	3	27	1

1. ábra. A begyűjtött adatok egy táblázatban, egy sorban egy csapat egy mérkőzésen nyújtott teljesítménye

Az önálló adatgyűjtés bár időigényes, de egy nagy előnye, hogy az elejétől fogva saját elképzeléseink alapján formálhattuk az adatot. A kezdeti tanítások során kiugró értékeket tapasztaltunk helyenként, így figyelemesebben tanulmányozva az adatot észrevettük, hogy két adatsor esetében szükség van adattisztításra. Az egyik sorban hiányos adatok voltak, elhagyása nem jelentett problémát a további vizsgálódások során. Míg a másik esetben egy konkrét statisztika értéke volt hibás mindkét csapatra nézve, ezek manuálisan könnyen javíthatók voltak.

A felderítő elemzés során fontos, hogy mivel nagy adathalmazzal dolgozunk, vizualizáljuk azt. Így jobb képet kaphatunk a háttérben meghúzódó összefüggésekről vagy igazolhatjuk esetleges előzetes sejtéseinket. A 2. ábrán egy boxplot látható a 2017-18-as és 2018-19-es szezonokban szereplő csapatok kapuralövési adatairól. Egy boxploton az egy csapathoz tartozó ábrázolást

a következőképpen lehet értelmezni: a belső téglalapot az 1. és 3. kvartilisek (azaz a mediánnál nagyobb és kisebb elemek mediánjai) határolják, az ebben húzott vonal az adat mediánját jelöli. A téglalapokból kinyúló rész az adat értelmezési tartománya, azonban ez a rész legfeljebb olyan hosszú, mint a téglalap hosszának másfélszerese. Az ezen kívül eső adatpontokat lyukas pöttyök jelzik (ezek az úgy nevezett outlier adatpontok). Megfigyelhető, hogy az outlier pontokat leszámítva a legmagasabb értékekkel valóban azok a csapatok rendelkeznek, amelyeket mi magunk is jobb képességűnek gondolunk korábbi eredményeik alapján (Ferencvárosi TC, MOL Fehérvár FC, Újpest).



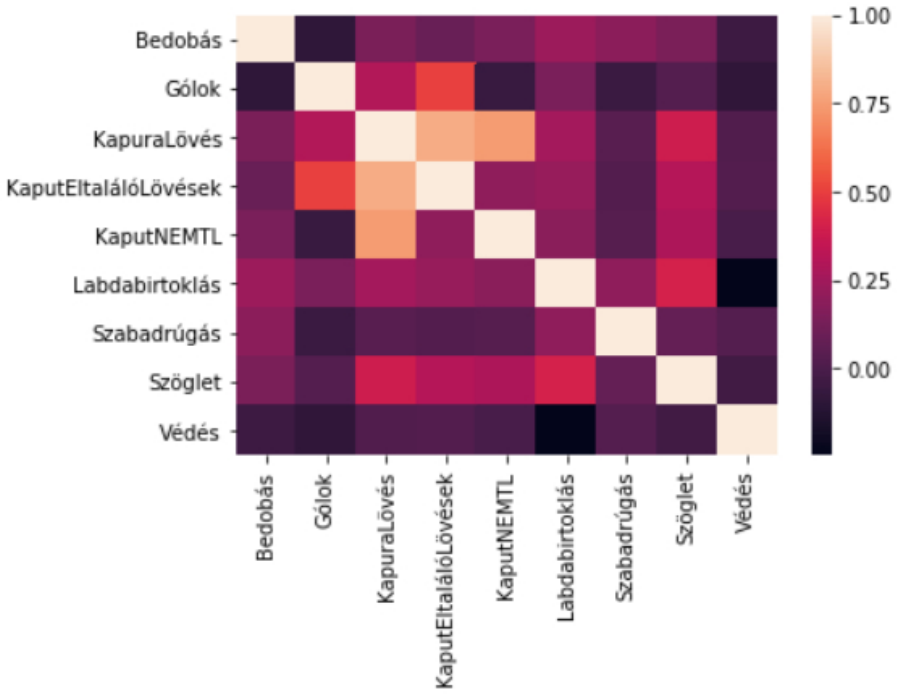
2. ábra. A kaput találó lövések boxplotja

A 3. ábrán egy pont egy csapat szezononkénti átlagos teljesítményét jelöli, az x koordináta a kapuralövések számát illetően, míg az y koordináta a labdabirtoklását. Igazolódni látszik az az előzetes feltevés, hogy amely csapatok többet birtokolják a labdát, több alkalommal jutnak el kapuralövésig. Ez alól kivételt képez 2 adatpont (a két legalsó), ez mindkét esetben a Debreceni VSC csapatához tartozik, minden bizonnyal ez a csapat olyan taktikát alkalmaz, ami nem a labdabirtoklásra épül mégis veszélyes támadójátékot

eredményez.

3. ábra. A kapuralövések és labdabirtoklás átlagos értékeinek összefüggése a tanítóadatban

A következő ábrázolás a változóink közötti összefüggések megtalálásnak céljából készült. A későbbiekben a csapatok által szerzett gólok számának jóslását szeretnénk véghezvinni, így arra vagyunk kiváncsiak, mely statisztikák vannak erre a legnagyobb hatással. A 4. ábrán látjuk, hogy egyes statisztikák mennyire függnek egymástól, minél világosabb egy mező (a korreláció közelebb van 1-hez), annál nagyobb az összefüggés az ahhoz az oszlophoz és sorhoz tartozó változó között. Ebből az olvasható le, hogy a kapuralövések száma, azon belül a kaput eltaláló lövések számítanak a legjobban, amikor a gólok számát szeretnénk vizsgálni.



4. ábra. A változók közti összefüggések

4. Gépi tanulási modell

A jelentősen megnövekedett adatmennyiségnek köszönhetően egyre hatásosabbak a gépi tanulási algoritmusokat alkalmazó modellek. Ennek is betudható, hogy rengeteg korábbi munka elérhető a világ számos bajnokságáról, elsősorban az angol Premier League küzdelmeiről[11] [5]. Egy ilyen munka szolgált a mi kutatásunk alapjául is, amelyet komolyabb átalakítások után tudtunk használni a saját adatunkon is [12]. Ebben a fejezetben ennek a modellnek a működését mutatjuk be, illetve kitérünk az eredmények alkalmazására is a sportfogadás területén.

4.1. Az induló modell

A modellek felépítése során Python környezetben dolgoztunk, a felhasznált algoritmusok pedig a sci-kit learn csomag részét képezik [13]. Első lépésben fontos tényező, hogy a használni kívánt algoritmusok tanításához numerikus és beszédes adatokra van szükségünk, ezért az 1. ábra táblázatából elhagyhatjuk a MeccsID, Dátum, Csapat oszlopokat, hiszen ezek nem árulnak el sokat egy csapat teljesítményéről (a későbbiekben az eredményeink értelmezéséhez ezeket visszahozzuk). Továbbá a 4. ábrán megfigyelhető az is, hogy a Gólok változó oszlopában a legsötétebb mező a Bedobás sorában van, azaz ezen két változó között minimális az összefüggés. Így elhagyjuk a táblázatunkból a Bedobás oszlopot is, egyszerűsítve ezzel a modellt. Illetve a KapuraLövés oszlopot is elhagyhatjuk, hiszen ez a KaputEltalálóLövések és KaputNEMTL oszlopok összege, így nem hordoz extra információt.

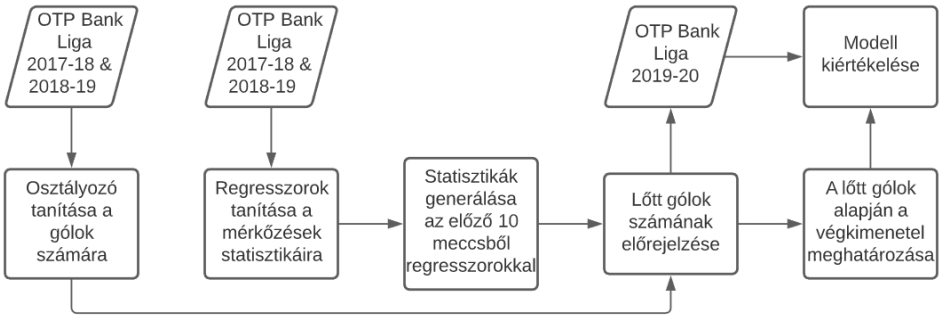
	Gólok	KaputEltalálóLövések	KaputNEMTL	Labdabirtoklás	Odds	Szabadrúgás	Szöglet	Védés	Hazaie
0	2.0	4.0	5.0	54.0	2.75	23.0	9.0	0.0	1
1	2.0	5.0	5.0	46.0	2.38	22.0	1.0	3.0	0
2	1.0	9.0	9.0	48.0	1.80	16.0	8.0	1.0	1
3	2.0	6.0	3.0	52.0	4.20	20.0	3.0	2.0	0
4	2.0	5.0	2.0	54.0	1.80	24.0	7.0	3.0	1

5. ábra. A tanításhoz felhasználandó numerikus adatok

Tanításra használtuk a 2017-2018-as és 2018-2019-es évadok adatait, a modellünk előrejelző képességét pedig a 2019-2020as szezonon teszteltük. Az első tanítási lépés során egy osztályozó algoritmust tanítottunk, amely magyarázóváltozói 5. ábra oszlopai kivéve a gólok száma (ezt átmenetileg elhagyjuk és úgy veszünk egy táblát), ezek alapján szeretnénk az ismeretlen célváltozót becsülni. Ez a célváltozó, pedig a Gólok változó. Ez a feladat kezelhető lenne regressziós feladatként is, azaz generálhatnánk folytonos értékeket, amelyeket utána kerekítünk, de a tesztelés során arra jutottunk, hogy az osztályozás jobb eredményre vezet. Osztályozáskor a jóslatunk egy

diszkrét halmazból kerül ki, ennek elemei jelen esetben egész számok 0-tól 5-ig. Felügyelt gépi tanulásról van szó, azaz a magyarázóváltozókhoz tartozó célváltozót ilyenkor még tanulmányozhatja az algoritmus és megtanulhatja a legfontosabb összefüggéseket (például a magasabb KaputEltalálóLövések érték magasabb gólszámhoz vezet, stb.). A választott algoritmus az AdaBoostClassifier volt, ennek egy sajátossága, hogy az eredeti tanítás után újabb másolatait tanítja az adaton az osztályozónak, nagyobb hangsúlyt fektetve az első tanítás során problémás esetekre [14].

Következő lépésben továbbra is a fenti táblázatot alapul véve az osztályozó algoritmushoz hasonlóan fogunk eljárni. Ezúttal azonban regressziós algoritmusokat használunk azon attribútumok jóslására, amelyek ismeretlenek egy mérkőzés előtt, ezek: KaputEltalálóLövések, KaputNEMTL, Labdabirtoklás, Szabadrúgás, Szöglet, Védés. Regresszió során a jóslataink egy folytonos halmazból kerülnek ki. 6 algoritmusunk lesz, célváltozóink felváltva az előbb felsorolt változók. Egy másik fontos kitétel, hogy a becslésekkor egy adott csapat legutóbbi 10 mérkőzését vesszük figyelembe (ha feljutó együttesekről van szó, akkor azoknak a csapatoknak a teljesítményét használjuk fel a szezon elején, amelyek az előző szezonban kiestek), és ezek alapján prediktálunk egy értéket a soron következő mérkőzésre. Ennek oka, hogy a régebbi mérkőzések nem adnak pontos képet, az aktuális forma befolyása sokkal nagyobb az elkövetkező mérkőzésen.



6. ábra. A modell tanítás és a predikció folyamatábrája

Amikor elérkezünk a tényleges előrejelzéshez 2019-2020-as szezonon, birtokunkban vannak a betanított algoritmusok, ezek segítségével haladunk soronként, ahol egy sor egy csapat teljesítménye egy mérkőzésen. Először is a regressziós algoritmusokat használjuk, hogy az előző 10 mérkőzés alapján a statisztikákat legeneráljuk. Itt az első 10 fordulóban még az előző szezon mérkőzéseit használjuk, a 11. fordulótól viszont a saját generált adatainkat. A labdarúgásban, mint minden sportban számolni kell a véletlen hatásával is, ezért figyelembe vesszük a változók szórását is, amit a tanítóadatból számoltunk ki. Ezt szorozzuk egy 0 várható értékű, 1 szórású normális eloszlásból véletlenül generált értékkel és hozzáadjuk a becslésünkhöz. Már csak az van hátra, hogy a gólok számát megkapjuk, erre segítségül hívtuk a betanított osztályozó algoritmusunkat, ami az adott mérkőzésre generált statisztikák alapján becsült egy 0 és 5 közötti értéket.

	MeccsID	Dátum	Csapat	Gólok	KaputEltalálóLövések	KaputNEMTL	Labdabirtoklás	Odds	Szabadrúgás	Szöglet	Védés	Hazaie
394	396	2019	Újpest	1	9.031566	7.187097	54.630569	2.00	8.748229	3.173124	4.034325	1
395	396	2019	Puskás Akadémia	0	6.798379	2.523391	56.795874	3.20	15.421862	5.053874	6.595611	0
396	397	2019	Mezőkövesd	1	2.374547	2.154822	49.828718	2.05	25.905769	4.198682	3.058529	1
397	397	2019	Zalaegerszeg	0	1.292169	1.067988	45.488485	3.20	20.188135	7.231682	4.165542	0
398	398	2019	Kisvárda	2	2.809980	3.498167	65.837480	2.10	16.365595	5.066092	5.695083	1
399	398	2019	Paks	1	11.464231	3.216084	36.470135	3.25	20.504032	4.092276	2.439010	0

7. ábra. A 2019-2020-as szezon első 3 mérkőzése generált adatokkal

Ez volt az alapséma, amelyből el tudtunk indulni, kezdeti állapotában azonban ez a modell pontosságban jelentősen elmaradt a kívánt szinttől. A predikciónk végül mindig egy végkimenetel volt (hazai győzelem, döntetlen, vendég győzelem), ami a jósolt gólszámokból adódik. Mindössze a mérkőzések mintegy 30%-ában sikerült helyes jóslatot adnia a végkimenetelre. Viszonyításképpen egy olyan kiértékelés is 43%-ban helyes, amiben mindig hazai győzelmet prediktálunk automatikusan. A valódi feladatot a legjobb gépi tanuló algoritmusok megválasztása és ezek paramétereinek beállítása jelenti, illetve a változókon olyan adatmanipuláció végrehajtása, ami tovább növelheti a modell eredményességét.

4.2. A modellen végzett fejlesztések

Tekintettel arra, hogy másfajta adatok állnak rendelkezésünkre, mint az eredeti modellt ([12]) felállítóknak, érdemes volt a meghatározó függvények részleteit megváltoztatni. Jelentős javulást lehetett először is azzal elérni, hogy a tesztszezonunkban a különböző statisztikák becslésére használt algoritmusokat lecseréltük jobb pontosságot adókra. Az alábbiakban olvasható ezek listája, együtt a megfelelő hiperparaméterekkel, ahol ezek beállítása tovább javította a rendszer eredményességét:

- KaputEltalálóLövések: KNeighborsRegressor(n_neighbors=4)
- KaputNEMTL: KNeighborsRegressor()
- Labdabirtoklás: Ridge(alpha=0.5)
- Szabadrúgás: LassoLars()
- Szöglet: KNeighborsRegressor()
- Védés: GradientBoostingRegressor()

Továbbá a gólok számát eredetileg egy regressziós algoritmussal prediktálták, ehelyett választottuk ki a fentebb említett AdaBoost osztályozót. Ez rögtön egy ésszerű változtatásnak tűnhet, az eredeti implementációban talán a kerekítést használó megoldás jobban teljesített. Azonban egy ilyen kevés értéket felvevő változó esetén, mint a gólok száma nem meglepő, hogy az osztályozás jobban működik. Emellett mivel már diszkrét értékeink vannak, lekérhető ezek valószínűsége, amit további elemzésre használtunk (erről bővebben a következő alfejezetben).

A mérkőzések statisztikáin kívül az Eredmények.com oldalon elérhetők a találkozókra kínált fogadóirodai oddsok. A kínált ráták a bet365-től származnak, ami egy nemzetközileg széles körben használt online fogadási portál.

A következőkben ezeket az oddsokat használjuk mi is, a modellünk teljesítményét ezekhez az adatokhoz hasonlítjuk. Ezek a mutatók olyan modellekből származnak, amelyeket szakértők hoztak létre kifejezetten a fogadóiroda érdekeit szem előtt tartva.

A modellünkben a végkimenetelre vonatkozó oddsok is attribútumként szerepelnek. A tanulási folyamat során csak a hazai és vendég győzelmekre vonatkozó számokat használtuk. Így persze megint jelentősen javult a modell pontossága, nem meglepő módon, hiszen egy igen előrehaladott modell eredményét kapta meg tulajdonképpen. Magyarázható azonban ezen adatok felhasználása azáltal, hogy a konkrét alkalmazásunknál, azaz a tényleges meccsfogadásnál is elérhető, előre ismert mutató a hazai és vendég győzelemekre kínált odds.

4.3. Predikciók felhasználása sportfogadásra

Az előző alfejezetben említett osztályozó algoritmus egy sajátossága, hogy egy függvényhívással megkapható a valószínűsége minden meccs és minden csapat esetén annak, hogy mennyi gólt lő pontosan. Ennek a felhasználását egy példán keresztül szemléltetjük, mégpedig a tesztszezon második mérkőzésén, a Mezőkövesd-Zalaegerszeg (1-0, 2019. 08. 03. 19:30) meccsen.

- Mezőkövesd gólvalószínűségek (0-tól 5-ig): 0.25763, 0.22339, 0.27709, 0.10885, 0.10812, 0.02493
- Zalaegerszeg gólvalószínűségek (szintén 0-tól 5-ig): 0.34636, 0.15013, 0.17368, 0.14402, 0.13045, 0.05536

Ezen két lista megfelelő elemeit összeszorozva megkapható a két csapat mérkőzésén a lehetséges végeredmények valószínűsége. Azzal az egyszerűsítő feltétellel dolgoztunk, hogy egy csapat által lőtt gólok száma nem függ a másik csapat teljesítményétől. Ennek oka, hogy egy adatsorban csupán a csapatok saját mutatói szerepelnek, továbbá a statisztika generáló algoritmusok tanításánál is csak egy csapat saját mérkőzéseit vettük figyelembe.

Ezért feltehető, hogy ezeket a szorzásokat el lehet végezni és tényleges valószínűségeket kapunk eredményül. Tehát a listák legelső elemeit összeszorozva megkapjuk a 0-0 eseményt, a Mezőkövesd lista második elemét a Zalaegerszeg lista első elemével szorozva pedig az 1-0 valószínűsége kapható meg. Az így kapott mátrix:

MEZ \ ZTE	0	1	2	3	4	5
0	0.08923	0.03868	0.04475	0.0371	0.03361	0.01426
1	0.07737	0.03354	0.0388	0.03217	0.02914	0.01237
2	0.09597	0.0416	0.04813	0.03991	0.03615	0.01534
3	0.0377	0.01634	0.01891	0.01568	0.0142	0.00603
4	0.03745	0.01623	0.01878	0.01557	0.0141	0.00598
5	0.00863	0.00374	0.00433	0.00359	0.00325	0.00138

Ebben a mátrixban a főátló elemeinek összege adja a döntetlen valószínűségét, a főátló alatti elemek összege a hazai győzelem, a főátló feletti elemek összege pedig a vendég győzelem valószínűségét. Ezeknek a számoknak a reciprokát véve kapható meg az odds, ami ebben a konkrét esetben a három eseményre (H/D/V): 2.03, 4.39, 3.56. Hasonlítsuk össze ezt az adatbázisunkban szereplő valódi oddsokkal: 2.05, 3.50, 3.20. Kiemelkedő különbség a döntetlen esetben, ezeknek az eseményeknek a modellünk rendszerint túl kicsi valószínűséget ad, ennek orvoslása egy lehetséges jövőbeni fejlesztése a modellnek, mi ennél a modellnél csak a hazai és vendég eseményekkel foglalkoztunk. Ebben az esetben nem olyan kiugró a különbség, így érdekes lehetett vizsgálni, hogy ezek az eltérések mutatnak-e valamilyen mintát és erre építhetünk egy fogadási tervet.

A fent leírt módon megkaptuk az oddsokat a szezon minden mérkőzésén. Fontos megjegyezni, hogy a fogadóirodák által kínált oddsok nem tényleges a valószínűségekre vezethetők vissza, azaz a fenti példával élve : $1/2.05+1/3.5+1/3.2 = 1.086$, míg a saját modellünkre ugyanezt megnézve:

$1/2.03+1/4.39+1/3.56 = 1.001$ (ami mindössze kerekítési hiba, valójában pontosan 1 az összeg) . Tehát a fogadóirodák a valódinál nagyobb valószínűséget párosítanak az eseményekhez, így csökkentve a kínált oddsokat. A fenti példában 9%-os eltérés a fogadóirodák várható profitja, azaz ha a kínált esemény egyszerű érmefeldobás lenne, ahelyett, hogy mindkét eseményre 2-es oddsot kínálna (feltesszük, hogy szabályos érme, a fej és írás is 1/2 valószínűségű), 1.835-as odds lesz a két eseményre ($1/1.835 * 2 = 1.0899$). [2]. A fogadási stratégia kidolgozásakor tehát szem előtt kell ezt is tartani, ha valóban ki akarjuk játszani a fogadóirodát.

Lehetséges még így is nyereséges tervet felállítani az immáron rendelkezésünkre álló adatok segítségével? A következő módon használtuk az adatokat: vegyük minden mérkőzésen a modellünk szerint legvalószínűbb eseményt (ez mindig hazai vagy vendég győzelem lesz), nézzük az ehhez tartozó oddsot. Ha ez a szám alacsonyabb, mint a bet365 oddsa arra az eseményre, akkor arra fogadjunk egy összeggel. Így tehát egyszerre mindig egy mérkőzésre fogadunk rögzített összeggel. Eldöntendő kérdés volt annak a különbségküszöbnek a meghatározása, hogy pontosan mennyivel is legyen legalább kisebb a modell odds, mint a fogadóirodai, ezt az Eredmények fejezetben tárgyaljuk. A további pontos számok szintén ebben a fejezetben találhatók.

5. Valószínűségi modellek

Ebben a fejezetben valószínűségi modelleket fogunk tárgyalni. Ezek a tárgyalt modellek tulajdonképpen semmi mást nem használnak, csak a múlt a mérkőzéseinek eredményeit, azaz a csapatok által az adott meccsen lőtt gólok számát. A legelső eredményes modellt a már említett matematikus, Maher fejlesztette ki 1982-ben [15]. A modelljében azt feltételezte, hogy a csapatok által lőtt gólok száma, mennyisége Poisson-eloszlást követ. Jelöljük X-el egy adott csapat által lőtt gólok számát, ekkor annak a valószínűsége, hogy k gólt lő a csapat:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \lambda > 0$$

Ahol λ nem más, mint a csapat által lőtt gólok várható értéke. A λ értékének meghatározása különböző modellek megalkotására ad lehetőséget. Nézzük meg először Maher modelljét.

5.1. Basic Poisson Modell (1982, J. Maher)

Matematikailag a következőképpen írhatjuk le a modellünket.

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

$$\text{ahol } \lambda = \alpha_i\beta_j\gamma \text{ és } \mu = \alpha_j\beta_i$$

Ahol a változók:

- i és j jelentik az otthoni illetve az idegenbeli csapatot,
- α és β jelöli a megfelelő indexekkel az adott csapat támadó és védekező erejét,
- γ pedig az otthoni csapat hazai pálya által jelentett fölényét jelöli.

A γ paraméter szükségességét az intuíciónk mellett egy egyszerű megfigyeléssel is igazolhatjuk, általánosságban gyakrabban nyer a hazai csapat, mint az idegenbeli. A többi paraméter is magától adódó, hiszen egy mérkőzésen mindkét csapatnak van egyfajta támadó ereje és védekező ereje is, ahol ezek páronként ütköznek egymással. A nagy hazai támadóerő és kicsi idegenbeli védekező erő valószínűleg sok hazai gólt eredményez egy adott meccsen. A paraméterek meghatározásához a jól ismert maximum likelihood módszert alkalmazzuk.
Nézzük meg, hogy mik az előnyei illetve hátrányai ennek a modellnek.
Előnyök:

- Egyszerű és könnyen értelmezhető modell.
- Bevezet hasznos, elengedhetetlen fogalmakat, mint a csapat támadó ereje, védekező ereje, illetve az otthoni pálya bónuszértéke, ezáltal jó alapul szolgálva a későbbi modelleknek.
- Nem a legpontosabb, de azért egész jó eredményeket ad.
- A gépi tanulásos modellekkel ellentétben, csak a múlt meccsek eredményeit használja fel, és nem használ más statisztikát.

Hátrányok:

- A modell nem törődik azzal, hogy a meccsek nem egy időben játszódnak és egy csapat szezon eleji, illetve szezon végi formája jelentősen eltérhet egymástól.
- Bizonyos eredmények felül vannak reprezentálva a modellben.
- Nem használ fel más statisztikát a múlt meccsek eredményein kívül.

Ezzel megismerkedtünk a Basic Poisson Modellel, tekintsük meg ennek egy továbbfejlesztését, a Dixon–Coles-modellt.

5.2. Dixon–Coles-modell, (1997, M. Dixon és S. Coles)

1997-ben megjelent egy irat Mark Dixon és Stuart Coles nevezetű matematikusoktól, mely rávilágított a BP Modell néhány hiányosságára, hibájára, és meg is oldotta azokat [2]. Ezek a hibák nem mások, mint a fent már említett hátrányok közül az első kettő. A szerzők azt állították, hogy a kevés gólos döntetlennel zárult meccsek, mint például: 0-0, 1-1 eredendően alul vannak reprezentálva a modellben. Ezeket természetesen alá is támasztják a fent említett iratukban. Most viszont nézzük meg az általuk felépített modellt:

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

$$\text{ahol } \lambda = \alpha_i \beta_j \gamma \text{ és } \mu = \alpha_j \beta_i$$

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

Láthatóan a különbség a két modellben egy τ függvény, ami egy ρ paramétertől függ. A paraméterek megtalálásához az előző modellhez hasonlóan maximum likelihood módszert alkalmazunk. Tulajdonképpen, ha a meccseket indexeljük, mint $k = 1, \ldots, N$ és nézzük a k-hoz tartozó eredményt, mely (x_k, y_k), akkor a következő a likelihood-függvény, melyet maximalizálni szeretnénk:

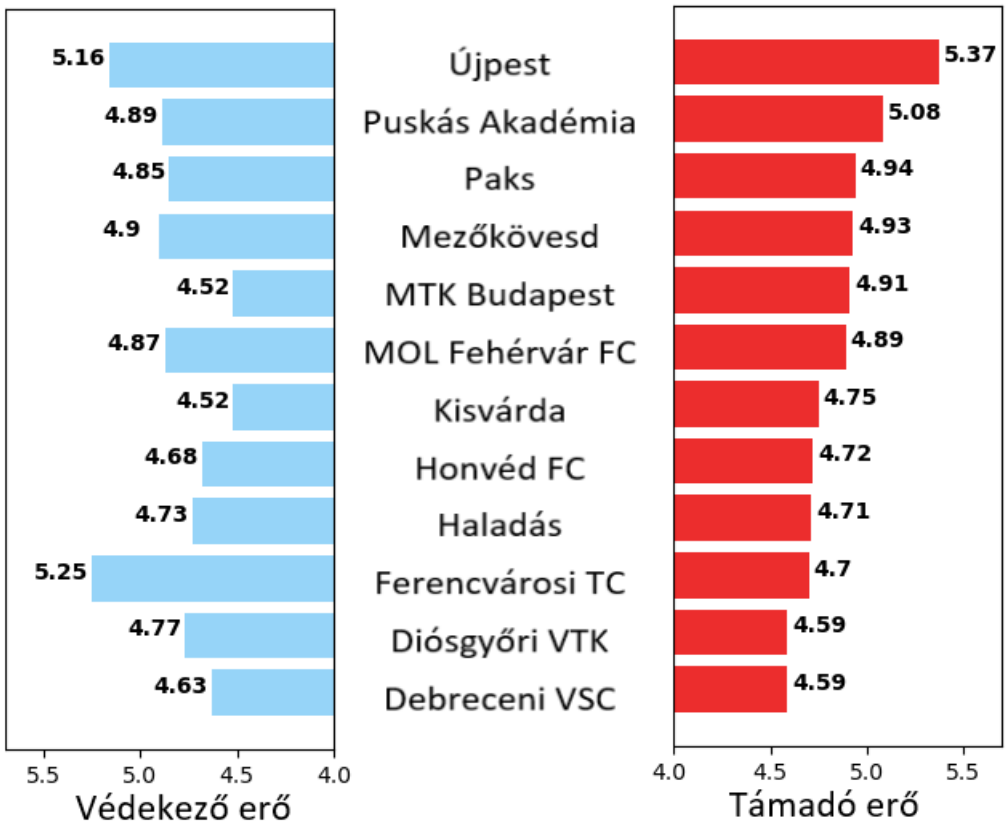
$$L(\alpha_i, \beta_i, \rho, \gamma) = \prod_{k=1}^{N} \tau_{\lambda_k,\mu_k}(x_k, y_k)\ \frac{e^{-\lambda}\lambda^{x_k}}{x_k!}\frac{e^{-\mu}\mu^{y_k}}{y_k!}$$

$$\text{ahol } \ \lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma \text{ és } \mu_k = \alpha_{j(k)}\beta_{i(k)}$$

Itt természetesen $i(k)$ és $j(k)$ a k. meccs csapatait jelenti. Ezzel a modellel is kapunk egy becslést a csapatok támadó és védekező erejére. A 8. ábrán a csapatok értékeit láthatjuk támadó erő szerint csökkenő sorrendbe rendezve, ezzel is közelebb hozva magunkhoz a modellek működését.

Látható például, hogy a Ferencvárosi TC kimagasló támadó és védekező erővel rendelkezik a modell szerint, így számíthatunk rá, hogy ők előkelő helyen fognak végezni a tabellán.

Térjünk most rá arra a Dixon–Coles-modellre, ami megoldja azt a problémát, hogy az eddigi modellek nem kezelik az időbeli eltolódást. Dixon és Coles bevezet egy súlyozást a meccsekre a modellükben, így jobban számítanak bele a paraméterek becslésébe az újabb meccsek. A modell a következő



8. ábra. A csapatok támadó és védekező ereje a Dixon–Coles-modell szerint.

likelihood függvényt eredményezi:

$$L(\xi) = \prod_{k \in A_t} \{\tau_{\lambda_k,\mu_k}(x_k, y_k) \frac{e^{-\lambda}\lambda^{x_k}}{x_k!} \frac{e^{-\mu}\mu^{y_k}}{y_k!}\}^{\phi(t-t_k)},$$

$$\text{ahol } \xi = (\alpha_i, \beta_i, \rho, \gamma)$$

Ebben az egyenletben t_k jelenti azt az időt, amikor k darab meccset játszottak és $A_t = \{k : t_k < t\}$ pedig a t idő előtt lejátszott meccsek halmaza. Ennek a modellnek a helyességét most nem tárgyaljuk, a mellékelt iratban, melyről már többször is szó volt, részletesen szerepel.

A modellek kiértékelésével a 7. Eredmények fejezetben foglalkozunk.

6. Saját modellek

6.1. Első saját modell

Basic Poisson modellnél számos probléma felmerül. Ezek közül, aminek a megoldását mi megcélozzuk az az, hogy függetlennek kezeli az adott meccsen az egyik csapat által lőtt gólok számát, a másik csapat által lőtt gólok számától. Feltételezhetjük, hogy egy bekapott gól teljesen képes átformálni a játék képét, hiszen ekkor a vesztésre álló csapatnak jobban kell támadnia, ami által viszont kinyílik a védelme, így az ellenfél viszont hatásosabban tud ellentámadásokat indítani. Így egyáltalán nem független a kapott gólok száma, hiszen egy kiegyenlített mérkőzésnél, ha egyik csapat gólt kap, akkor utána kénytelen lesz jobban támadni, így nagyobb eséllyel fog gólt lőni, illetve gólt kapni. Ehhez bevezetünk egyfajta elemenkénti szorzót a szokásos eddig is használt predikciós mátrixunkra. Legyenek az új elemek:

$$\hat{m}_{i,j} := m_{i,j} * c^{\delta}$$

Ahol, c egy alkalmas konstans, δ pedig az átlótól vett távolsága az elemnek, tehát közvetlen az átló felett 1, eggyel felette 2 és így tovább. Látható például, hogy így legnagyobb szorzót jelen esetünkben a 3-0, illetve 0-3 eredmények kapják. Vegyük észre, hogy így az új $\hat{M}$ predikciós mátrixunk már nem valószínűségi eloszlás lesz az eredményeken. De ez nekünk nem probléma, a lényeg, hogy egyfajta sorrendet tudunk adni, hogy melyik eredményt tartjuk a legvalószínűbbnek.

6.2. Második saját modell

A modelleknél nagyon fontos szerepet játszik, hogy hogyan választjuk ki a megfelelő predikciót a kapott mátrixból. Választhatjuk például a legnagyobb valószínűségű elemhez tartozó eredményt, így kapva egy konkrét végeredményt a mérkőzésre. De akár azt is megtehetjük, hogy nem konkrét eredményt prediktálunk, hanem azt, hogy hazai győzelemmel, vagy idegenbeli győzelemmel, vagy döntetlennel végződik a mérkőzés. Ennek a módszerére akár igen komplex algoritmust is kidolgozhatunk. Láthatjuk, hogy a Dixon–Coles-modell nagyon kevés esetben prediktál döntetlent. Ennek kiküszöbölésére az ötletünk a Metropolis-algoritmus[1] gondolatán alapszik, melyen alapuló globális optimum kereső algoritmusok (pl.: Szimulált hűtés) egyik kulcs lépése, hogy úgy kerül ki a lokális optimumok csapdájából, hogy minimalizálási feladat esetén néha nem a kisebb érték felé mozog, hanem inkább szembe megy vele, és ellép a nagyobb érték irányába. Vezessünk be egyfajta küszöbindexet ami megadja azt, hogy amennyiben a győzelem valószínűsége egy bizonyos küszöbnél kisebb, akkor bár hiába az a legvalószínűbb, mégis inkább döntetlent fogunk prediktálni. Tehát vezessünk be μ_H és μ_A konstansokat, és ennek a alapján az eredmény prediktálása úgy működik, hogy:

1. Ha $P(H)$ a legnagyobb és $P(H) \geq \mu_H$, akkor H.
2. Ha $P(I)$ a legnagyobb és $P(I) \geq \mu_I$, akkor I.
3. Egyébként D.

[1] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller algoritmusa, mely bekerült a 21. század top 10 algoritmusa közé is. (Láthatjuk, hogy magyar vonatkozása is van Teller Edéék révén.) [16]

7. Eredmények

7.1. Gépi tanulási modell pontossága

Még mielőtt bármilyen modellt építenénk nézzük meg milyen pontossággal dolgoznak a szakértők. Ezt úgy tesszük, hogy a meglévő oddsok alapján minden mérkőzésen jósoljuk a legvalószínűbb eseményt, azaz a legkisebb oddsszal rendelkezőt. Ekkor a következő kiértékelést kapjuk:

- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.522
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.534
- Az idegenbeli győzelmek közül mennyit találtunk el: 0.514
- A döntetlenek közül mennyit találtunk el: 0.0
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.814
- Összességében mennyit találtunk el: 0.530

Látható ebből is, hogy milyen nehéz prediktálni a mérkőzések eredményét, a szakértők is a meccsek alig több, mint felét találták el. Továbbá ennek a modellnek szintén gyengéje, hogy a döntetlenek valószínűségét alulbecsli, hiszen soha nem jósol döntetlent, mivel a kiértékelésünk során a legalacsonyabb odds mindig a hazai vagy vendég győzelemé.

A 3.1. alfejezetben tárgyalt modell mutatói, amin tehát még semmilyen fejlesztés nem történt:

- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.371
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.385

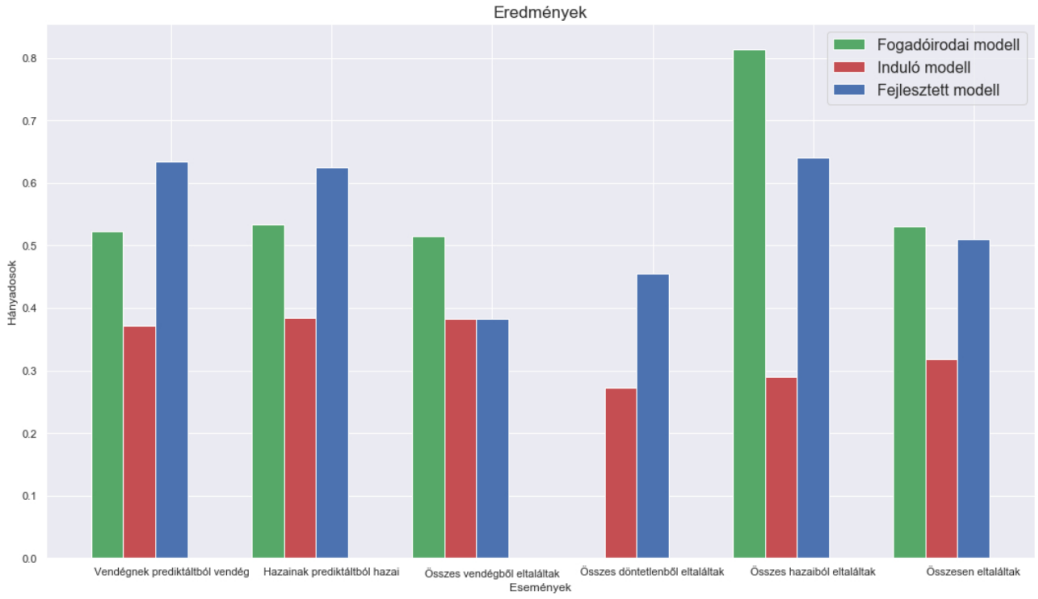
- Az idegenbeli győzelmek közül mennyit találtunk el: 0.382
- A döntetlenek közül mennyit találtunk el: 0.273
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.290
- Összességében mennyit találtunk el: 0.318

Ezek az eredmények még jelentősen elmaradnak a szakértő modell számaitól, a hazai győzelmek terén kifejezetten nagy az elmaradás.

Alább olvasható annak a rendszernek a kiértékelése, amelyben már megtörtént a lehető legjobb eredményt produkáló gépi tanulási algoritmusok kiválasztása, erről részletesebben a 3.2. alfejezetben:

- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.634
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.625
- Az idegenbeli győzelmek közül mennyit találtunk el: 0.382
- A döntetlenek közül mennyit találtunk el: 0.455
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.64
- Összességében mennyit találtunk el: 0.51

Ezek a számok már sokkal jobban megközelítik a szakértői modellt, sőt azon mérkőzéseken ahol hazai vagy vendég győzelmet prediktál, jobb arányban is találja el a végeredményt. Ez annak is betudható, hogy ez a modell már jósol döntetlent is, tehát összesen kevesebb hazai és vendég jóslat lesz. Azonban így a döntetlenek is jobban vizsgálhatók. A 9. ábrán szemléltetjük a gépi tanulási modellek mutatóit és összevetjük a szakértő modell eredményeivel, ezek a fentebb felsorolt értékek.



9. ábra. A gépi tanulási modellek eredményei összehasonlítva a fogadóirodával

Bizonyos esetekben megvizsgálva a modell oddsait igen nagy eltérésekre lettünk figyelmesek a fogadóiroda által kínált számokhoz képest. Néhány ilyen példa:

- Paks-Honvéd (3-1, 2019.08.11.) bet365 oddsok: 2.75/3.1/2.45, ezzel szemben a saját oddsaink: 2.08/5.96/2.85. Jobban ráérzett tehát a saját modellünk az erőviszonyokra.

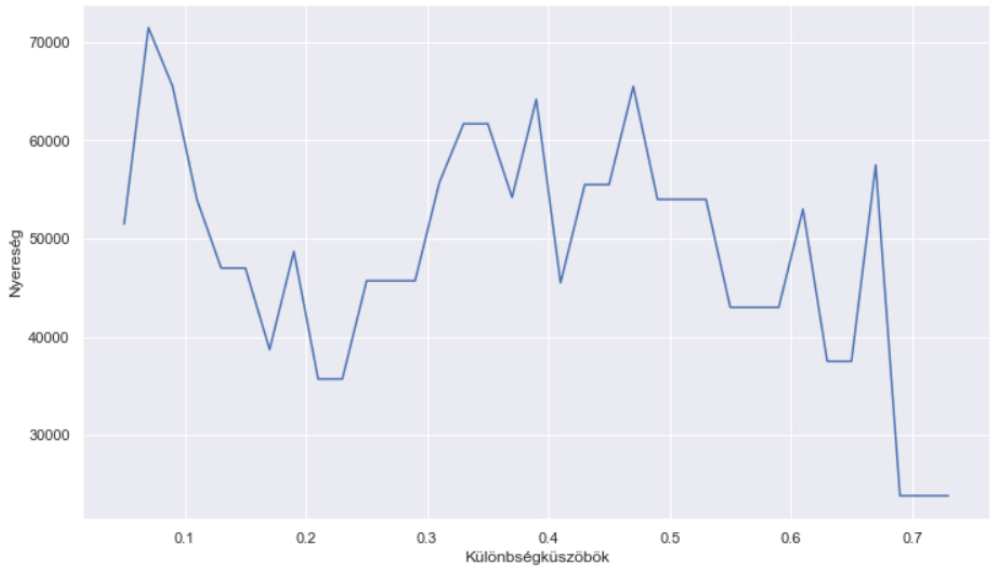
- Mezőkövesd-Diósgyőr (0-1, 2019.12.07.) bet365 oddsok: 1.65/3.8/4.33, saját oddsok: 2.99/6.04/2. Egy újabb példa, ahol még az előzőnél is lényegesen nagyobb különbség észlelhető. Míg az oddsok közelsége az 1. példában bizonytalanságra enged következtetni, ebben az esetben biztosnak tűnik a fogadóirodai modell a hazai győzelemben, épp ellenkezőleg vendég győzelmet jelez a saját modellünk (ráadásul pontosan el is találta a 0-1 végeredményt). Ebben az esetben feltehetőleg túl nagy hangsúlyt fektettek a hazai pálya előnyére.

- Kaposvár-Paks (0-3, 2020.02.05.) bet365 oddsok: 3.3/3.4/2.1, saját oddsok: 6.51/7.61/1.4. A modellek egyetértenek ezúttal, a sajátunk azonban biztosabban jósolja meg a helyes végkimenetelt. A 2. és 3. példánál megjegyezzük, hogy már a 10. forduló után járt a bajnokság, így teljes egészében saját generált statisztikák alapján jelez előre a modell.

7.2. Fogadási terv a gépi tanulási modell alapján

A 3.3. alfejezetben leírt fogadási terv esetében fontos volt meghatározni minimálisan mekkora különbséget engedhetünk meg az oddsok között, azaz legalább mennyivel kell kisebb legyen a saját modellünk által kínált odds a legvalószínűbb eseményre, mint a fogadóiroda oddsa. Ekkor ugyanis azt állíthatjuk, hogy a modellünk biztosabb abban a végkimenetelben, nagyobb valószínűség adódik rá. A folyamat során fontos, hogy csak már megtörtént események alapján optimalizáljuk a küszöböt, ezért azzal a feltevéssel élünk, hogy a tesztszezon első fele már lejátszódott. Ezen adatok alapján beállítunk 2 küszöböt különböző célokat szem előtt tartva, majd ezekkel tesztelünk a szezon második felén (két darab 99 mérkőzésből álló adathalmazzal rendelkezünk). A 10. ábrán az első megközelítés vizsgálata látható, amikor csak azt szeretnénk, hogy maximális legyen a profitunk, legyen az bármennyi fogadásból. Szerepelnek a lehetséges küszöbértékek 0.05-től 0.75-ig és az látható, hogy a legjobb nyereséghez már elég, ha 0.07 különbség van a fogadóirodai odds javára. Ekkor 71500 Ft tiszta nyereséggel zárjuk a szezonnak ezen felét úgy, hogy 10000 Ft-tal fogadunk minden mérkőzésre (persze ez az összeg tetszőlegesen változtatható). Ehhez 47 darab fogadásra volt szükségünk. Innen már kiszámítható, hogy $47 * 10000 = 470000$ forint kockáratételével 71500 forintot nyertünk, azaz 15.21%-os az elméleti hozam, amivel már túltettünk a fogadóiroda 9%-os margóján. Tehát a tesztelés során 0.07 lesz az egyik használt különbségküszöb. Megfigyelhető azonban az is, hogy ha tovább növeljük a különbségküszöböt a végső egyenlegünk nem csökken nagy összeggel.

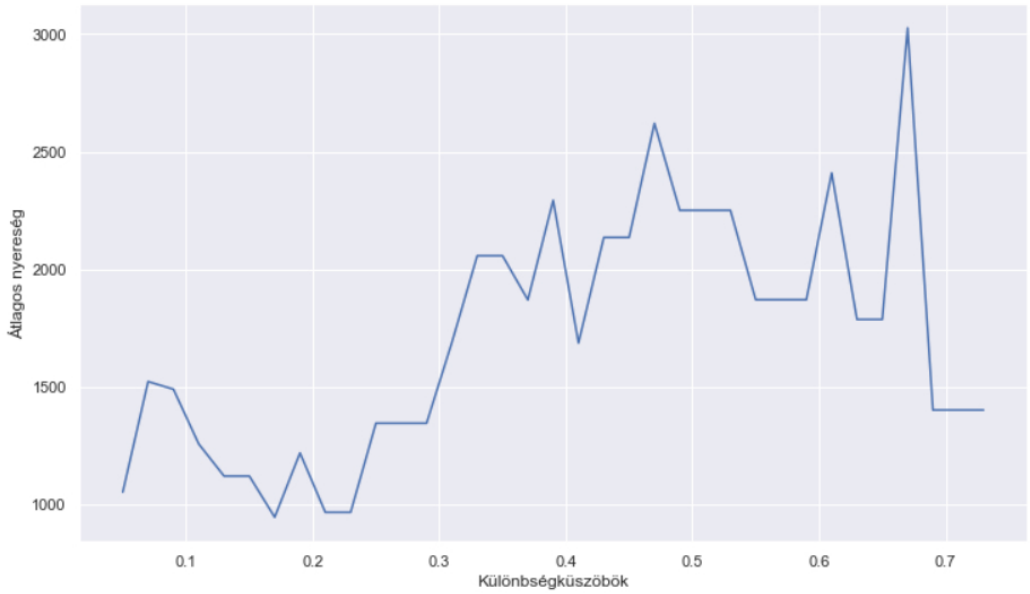
Ekkor a fogadások száma kevesebb lesz, kisebb kockázatot vállalunk.



10. ábra. A végső egyenleg különböző különbségküszöbökre (fogadási összeg = 10000 Ft)

A 11. ábrán látható a fogadásonkénti átlagos hozam a korábban említett különbségküszöbökhöz, továbbra is 10000 Ft-os egymérkőzéses fogadásokkal dolgozunk.

Innen már azt láthatjuk, hogy érdemesebb 0.67-nek választani a különbségküszöböt (tehát legalább ennyivel kisebb a modellünk oddsa, mint a fogadóirodáé), ekkor 3026 Ft-ot nyerünk átlagosan egy fogadással és összesen 19 mérkőzésen voltunk érdekeltek. Ez azt is jelenti, hogy valóban jobban járunk, hiszen az előbbi 15.21%-os elméleti hozamunkat majdnem megduplázva 30.26%-osra növeltük, a végső nyereség most 57500 forint. Ezt a szigorítást lényegében annak érdekében tesszük, hogy jobban tudjon általánosítani a modell. Átigazolások, edzőváltások, tulajdonosváltások jelentősen átformálhatják az erőviszonyokat egy új szezon előtt, ezért törekszünk arra, hogy minél kevesebb fogadással érjük el a maximális nyereséget. Azonban amikor ilyen kevés mérkőzésre fogadunk (a szezon felén dolgozunk és így is a 99-ből



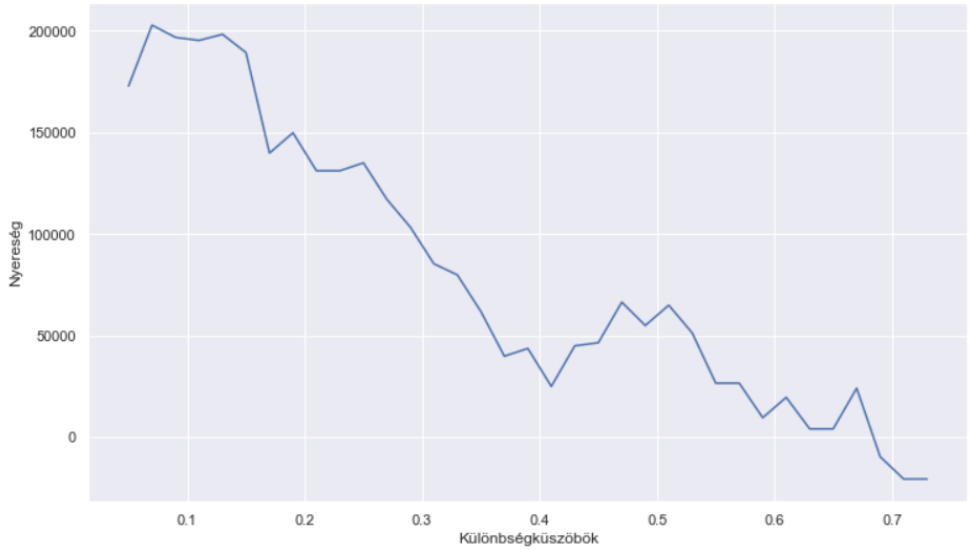
11. ábra. Fogadásonkénti átlagos hozam különböző különbségküszöbökre (fogadási összeg = 10000 Ft)

mindössze 19 meccsen fogadtunk), fennáll a veszélye annak, hogy néhány rossz eredmény, meglepetés sokat ront az eredményünkön a gyakorlatban.

Most teszteljük a két választott különbségküszöböt a 2019-20-as szezon második felén. Ha a küszöb 0.07, ebben a félszezonban is 47 fogadást teszünk, a végső nyereség 131300 forint, ami 27.94%-os hozamot jelent, amivel a gyakorlatban is sikerült túlteljesíteni a fogadóirodák 9%-os biztonsági margóját. Ellenben a 0.67-es küszöbnél az előző bekezdésben említett problémába ütközünk, hiába tűnik ez a biztosabb megoldásnak. Ekkor 17 fogadást tettünk és 33500 forintos mínuszban zártunk, azaz a hozam -17.9%. Ebből arra következtethetünk, hogy érdemesebb kisebb rendelkezésre álló adat (fél szezon) esetén a maximális profitot célként kitűzni optimalizáláskor.

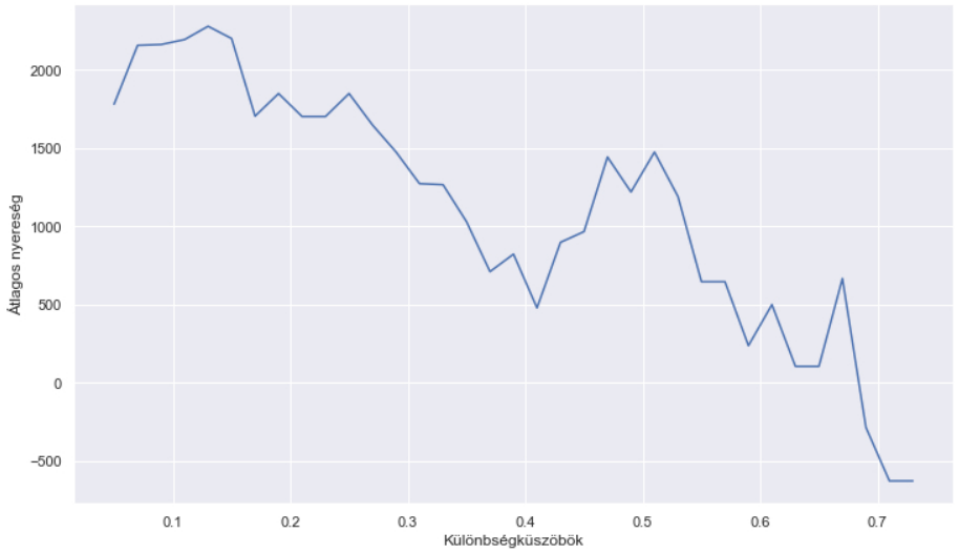
Abban az esetben, ha a modellt szeretnénk a közeljövőben élesben is használni, eltekinthetünk a 2019-20-as szezon felezésétől és felhasználhatjuk az egészet az iménti elméleti különbségküszöbök beállítására. Ekkor a ma-

ximális profit eléréséhez az optimális küszöb szintén 0.07, így 94 fogadásra kerül sor, a bevételünk 202800 forint. Így a hozam 21.57%-os, ez a vizsgálat a 12. ábrán látható.



12. ábra. Fogadásonkénti átlagos hozam különböző különbségküszöbökre (az egész tesztszezont vizsgálva, fogadási összeg = 10000 Ft)

A 10. ábrához hasonlóan észrevehető, hogy a küszöb kis mértékű növelésével nincs jelentős csökkenés a bevételben. Vizsgáltuk az átlagos hozamot is, azt kaptuk, hogy 0.13-nál optimális a küszöb, a fogadások száma 87 (13. ábra). Ekkor az átlagos nyereség fogadásonként 2280 forint, így ezzel a különbségküszöbbel a 21.57%-os hozamunkat 22.8%-osra növeltük. Levonható ebből, hogy több adat rendelkezésre állásakor kisebb szórás lesz a stratégia eredményességét illetően, így ha ténylegesen a gyakorlatban szeretnénk használni a modellt, ez a megoldás a kézenfekvőbb.



13. ábra. Fogadásonkénti átlagos hozam különböző különbségküszöbökre (az egész tesztszezont vizsgálva, fogadási összeg = 10000 Ft)

7.3. Valószínűségi modellek

Kezdjük a valószínűségi modellek kiértékelését a legegyszerűbb modellel, a Basic Poisson Modellel. Megtekintjük, hogy milyen predikciós mátrixot ad a modellünk a Ferencvárosi TC - Újpest mérkőzésre. Jelöljük X-el a Ferencvárosi TC által lőtt gólok számát és Y-al az Újpest által lőtt gólok számát.

```
In:
simulate_match(poisson_model, "Ferencvárosi TC", "Újpest")
Out:
```

	$Y = 0$	$Y = 1$	$Y = 2$	$Y = 3$
$X = 0$	0.08412	0.05863	0.02043	0.00474
$X = 1$	0.14960	0.10428	0.03634	0.00844
$X = 2$	0.13303	0.09270	0.03231	0.00750
$X = 3$	0.07886	0.05497	0.01915	0.00445

Láthatjuk, hogy legnagyobb értékek a Ferencvárosi TC hazai győzelmét preferálják, hiszen azok az $(X, Y) = (1, 0)$ és $(X, Y) = (2, 0)$ cellákban találhatóak. Ez a modell a 2018/19-as adatokon tanult, és a 2019/20-as adatokra prediktált. Így ez a predikció a 2019.10.19. Ferencvárosi TC - Újpest mérkőzés eredményére ad becslést, ami 1-0-ás győzelemmel záródott. Láthatóan a modell sikeresen eltalálta a mérkőzés eredményét.
Tekintsünk meg, hogy összességében hogyan teljesít a modell az adatokon.

- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.644
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.639
- A döntetlennek prediktáltáltak közül valóban mennyi volt döntetlen: 0.271
- Az idegenbeli győzelmek közül mennyit találtunk el: 0.426
- A döntetlenek közül mennyit találtunk el: 0.568
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.453
- Összességében mennyit találtunk el: 0.469

Vegyük észre, hogy a döntetlennek prediktáltak közül, nagyon kevés az, ami igazából is döntetlen. Itt jön elő Dixon és Coles modelljének az alapötlete, hogy felül vannak reprezentálva bizonyos döntetlenek. Ezt próbálja kiküszöbölni az első fajta Dixon–Coles-modell.
Az első fajta Dixon–Coles-modell megpróbál javítani azon, hogy felül vannak reprezentálva bizonyos (kevés gólos) döntetlenek. Ehhez az egyik fajta kiértékelési módszerünk az az, hogy nem a legmagasabb értéket próbáljuk kivenni, hanem összeszummázzuk a lehetséges értékeket a mátrixban az alapján, hogy az otthoni győzelmet, döntetlent, vagy idegenbeli győzelmet adna. Vegyük

észre, hogy ez már egy 4×4-es mátrixnál is nagyon alacsony döntetlen esélyt adhat, hiszen jóval kevesebb cella jellemzi a döntetlent, mint valamelyik fél győzelmét.

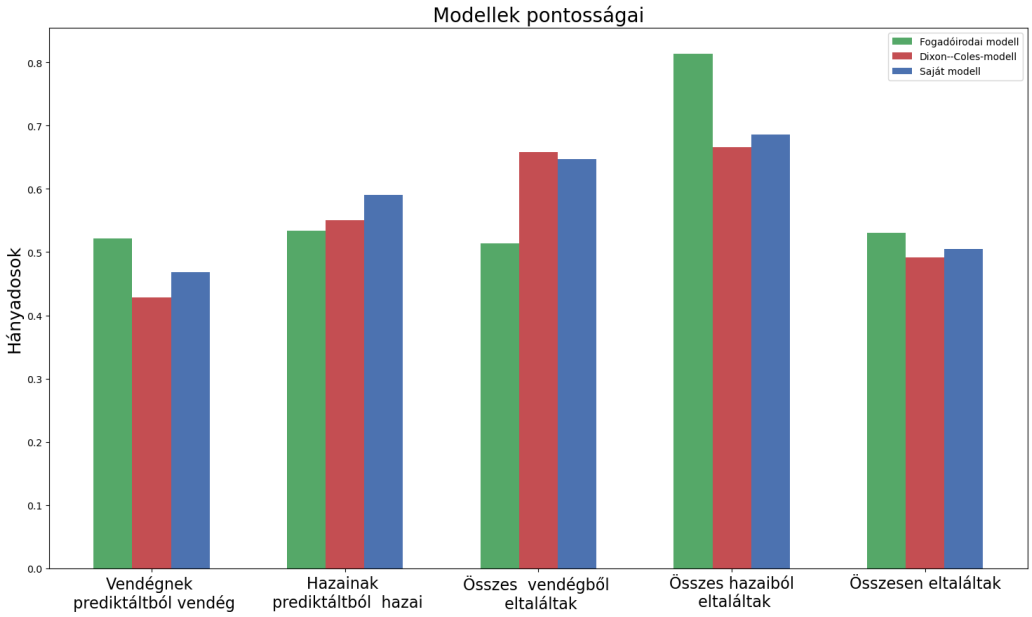
- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.428
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.550
- Az idegenbeli győzelmek közül mennyit találtunk el: 0.658
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.666
- Összességében mennyit találtunk el: 0.492

A modell pontosságán minimális növekedést értünk el, csrébe viszont nem prediktálunk döntetleneket.

Teküntsük meg, hogy milyen eredményeket ad az első saját modellünk. Itt egy új fontos paramétert vezetünk be, amire megpróbálunk optimalizálni. Ez annak a c konstansnak az értéke, ami nagyobb predikciós értéket ad a nagy gólkülönbséges meccseknek c^k-s szorzó alapján, ahol k a gólkülönbséget jelképezi. Azt tapasztaljuk, hogy a $c = 1.1$ megfelelő választás. Ezzel a következőképpen javul a predikciónk:.

- Az idegenbeli győzelemnek prediktáltak közül valóban mennyi volt idegenben győzelem: 0.468
- Az otthoni győzelemnek prediktáltak közül valóban mennyi volt otthoni győzelem: 0.59
- A döntetlennek prediktáltáltak közül valóban mennyi volt döntetlen: 0.25

- Az idegenbeli győzelmek közül mennyit találtunk el: 0.647
- A döntetlenek közül mennyit találtunk el: 0.022
- Az otthoni győzelmek közül valóban mennyit találtunk el: 0.686
- Összességében mennyit találtunk el: 0.505

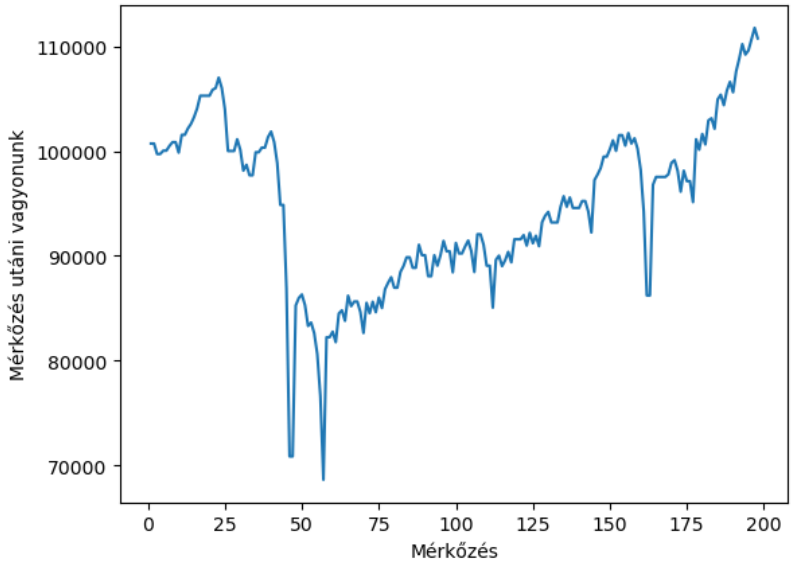


14. ábra. A valószínűségi modellek eredményei összehasonlítva a fogadóirodával

Megjegyezzük, hogy a második fajta Dixon–Coles-modell és a második saját modellünk, ugyan a magyar bajnokságon rosszabbul szerepeltek a többi valószínűségi modellhez képest, de az angol első osztályú labdarúgó-bajnokságban javítottak a predikciókon. Azonban a dolgozat témája a magyar bajnokság labdarúgó-mérkőzéseinek eredményének előrejelzése, melyeken az említett két modell rosszabbul teljesített, mint a társai, így ezeket most nem tárgyaljuk.

7.4. Fogadási terv valószínűségi modell alapján

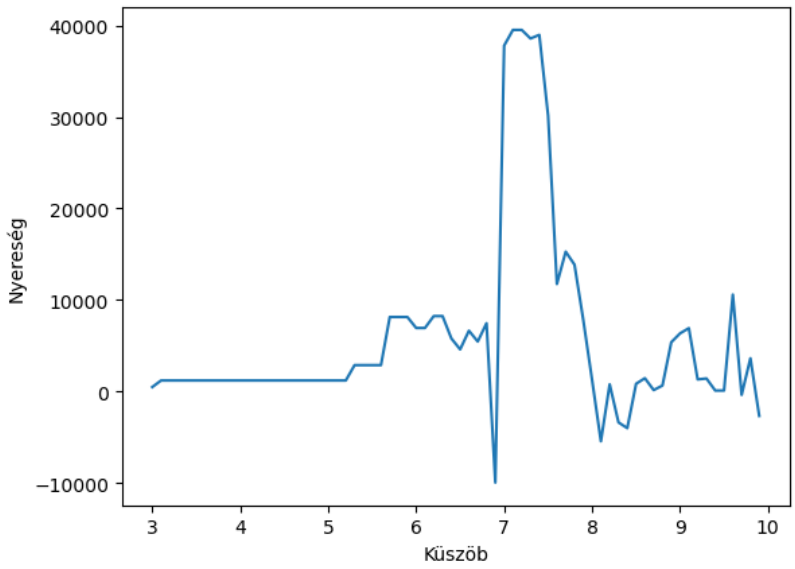
Egy lehetséges fogadási terv, ha összehasonlítjuk a predikciónkat az oddsokkal és akkor fogadunk az adott mérkőzésre, ha a modellünk is a fogadóiroda által előnyben részesített csapat győzelmét jósolja meg. Ekkor fogadjunk a modellünk által preferált csapatra. Ez a fogadási terv egyfajta biztonságos fogadási terv, hiszen akkor fogadunk csak, ha két modell szerint is ugyanaz lesz a győztes. A tételeket tegyük meg a jól ismert duplázós módszer alapján, azaz vereség esetén duplázzuk a tétünket. Legyen a fogadásra félretett vagyonunk 100000 Ft és a kezdő tétünk 1000 Ft. A szezont ezzel a fogadási stratégiával 10760 forint haszonnal zárjuk.



15. ábra. Vagyon alakulása a valószínűségi modellel az első fogadási terv alapján

Egy másik lehetséges terv, ha a gépi tanulásos módszerhez hasonlóan megvizsgáljuk, hogy mennyire biztos a modellünk a predikcióban. A modellünknek a predikciója az adott mérkőzésre a predikciós mátrix legnagyobb

eleméhez tartozó eredményt, így annak a reciprokát véve egy mérőszámot kaphatunk a modell biztonságára. Tegyük meg tétjeinket az előbbiekhez hasonlóan, de vegyük figyelembe azt is, hogy csak abban az esetben fogadjunk, ha egy bizonyos küszöbnél nagyobb a modellünk biztonsága. Ekkor a 16. ábrán látható a nyereségünk a küszöb függvényében.



16. ábra. Nyereség a küszöb függvényében

Láthatóan hatalmas nyereségre tehetünk szert, ha ezt a küszöböt megfelelően választjuk meg. Vajon meg tudunk-e adni olyan formulát a küszöbre, melynek segítségével nagyobb nyereséget érhetünk el? Ehhez több szezonon keresztül kellene vizsgálnunk a küszöbök és nyereségek alakulását. Egyelőre nem rendelkezünk explicit formulával a küszöbre. De a modellünk már így is nyereséges, mint azt az első fogadási tervben láthattuk, azonban a küszöb megfelelő megválasztásával, akár négyszeresére is növelhetnénk a nyereségünket. A küszöb meghatározása egyelőre nyitott feladat számunkra, de a modellben rejtőző potenciál már így is látható. Azonban elmondhatjuk, hogy a gépi tanulásos modell hasznosabbnak bizonyul fogadás szempontjából.

8. Zárógondolatok

8.1. Összefoglalás

Először tájékozódtunk a labdarúgó mérkőzések végeredményének megbecslésének történelméről. Ezután megpróbáltuk különböző gépi tanulási és statisztikai módszerek segítségével megbecsülni az OTP Bank Liga mérkőzéseinek eredményét, illetve ezek segítségével eredményes fogadási tervet kidolgozni. A gépi tanulási algoritmusokat vegyítő modell pontossága megközelítette a szakértők által épített fogadóirodák által alkalmazott rendszer mutatóit. Továbbá ennek a modellnek a segítségével nyereséges fogadási tervet sikerült kidolgozni. A vizsgált fogadóiroda 9%-os margóját is sikeresen túlteljesítettük, a legeredményesebb stratégia 27.94%-os hozamot produkált a gyakorlatban. Kipróbáltunk egyszerűbb statisztikai modelleket, Maher-féle Basic Poisson modellt, és a kétféle Dixon–Coles-modellt. Ezeket megvizsgáltuk, mi is felfedeztük a Basic Poisson modelljének egyik hibáját, és megpróbáltuk egy másik fajta úton kiküszöbölni.
Végső összevetésben a legjobb gépi tanulási modell és statisztikai modell összesített pontossága (sikeresen előrejelzett végkimenetelek leosztva az összes vizsgált mérkőzés számával) között minimális különbség, mindössze 0.05 adódott a gépi tanulási rendszer javára. Így elmondható, hogy mindkét módszerrel jól megközelítettük a szakemberek által fejlesztett rendszert.

8.2. Lehetséges folytatások

Rengeteg nyitott kérdés van ezen a területen, melyek irányába el lehetne indulni. Lehetséges folytatás lehetne valamelyik feltett kérdésünk megválaszolása is akár. Ehhez esetenként több adat beszerzése már elegendő lenne. A nyereséges fogadási terv alkalmazásához továbbtaníthatnánk a modellt, úgy hogy képes legyen az OTP Bank Liga 2020-21-es szezonjának küzdelmeiről

előrejelzést adni. A valószínűségi modelleknél láthattuk, hogy mind a Basic Poisson, mind a Dixon–Coles-modell nehezen tud megküzdeni a döntetlenekkel. Bár az ötletünk minimális javulást elért ezen a téren, az áttöréstől sajnos még messze állunk. Hogyan lehetne kiküszöbölni ezt a problémát? A valószínűségi modelles fogadási terveknél felmerülő küszöbproblémára megoldást találva megsokszorozhatnánk a nyereségünket. Fel lehetne térképezni a neurális hálók irányzatát is, melyhez már szintén számos tanulmány érhető el.

Hivatkozások

[1] Világi Kristóf és Sterbenz Tamás. A hazai sportinformatikai és sportanalitikai helyzet feltérképezése, adatfelhasználási lehetőségek kidolgozása. *Magyar Sporttudományi Szemle*, 19(5):32–35, 2018.

[2] 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.

[3] Md Ashiqur Rahman. A deep learning framework for football match prediction. *SN Applied Sciences*, 2(2):165, 2020.

[4] Irad Ben-Gal. Bayesian networks. *Encyclopedia of statistics in quality and reliability*, 1, 2008.

[5] Anthony C Constantinou, Norman E Fenton, and Martin Neil. pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.

[6] Anito Joseph, Norman E Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[7] I Graham and H Stott. Predicting bookmaker odds and efficiency for uk football. *Applied Economics*, 40(1):99–109, 2008.

[8] Roskó Zoltán Magyar Labdarúgó Szövetség. A magyar labdarúgás stratégiája, 2015.

[9] eredmenyek.com, howpublished = `https://www.eredmenyek.com/`, note = Accessed: 2020-07-15.

[10] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[11] Giovanni Angelini and Luca De Angelis. Parx model for football match predictions. *Journal of Forecasting*, 36(7):795–807, 2017.

[12] Tony Hoang and Thomas Duffy. Predicting english premier league match results with machine learning, 2019.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[15] M. J. Maher. Modelling association football scores. 1982.

[16] Wikipedia contributors. Metropolis–hastings algorithm — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-October-2020].