Alkalmazott Matematikai Lapok **32** (2015), 41–61.

JÉGKORONGCSAPAT ÖSSZEÁLLÍTÁSÁNAK VALÓS IDEJŰ OPTIMALIZÁLÁSA ADATBÁNYÁSZATI ESZKÖZÖK SEGÍTSÉGÉVEL

SÜDY BARBARA

Az adatelemzés egyik legizgalmasabb ága a sportadatokohoz köthető prediktív analitika területe. A masszív adatbázisokból kinyert információ – az adott sportág kellően mély ismerete mellett – számos előrejelzés alapjául szolgálhat.

Az alábbi cikkben egy olyan modellt mutatunk be, amely adatbányászati módszerek segítségével játék közben optimalizálja egy jégkorongcsapat sorösszeállítását. A modell előre jelzi, hogy a következő játékmegszakításig melyik öt mezőnyjátékosnak lesz a legnagyobb esélye gólt szerezni, ennek alapján az edző játék közben módosíthatja a sorösszeállítást. A modell három alegységének bemutatása mellett értékeljük azok teljesítményét, valamint említést teszünk a modell potenciális javíthatóságáról is.

1. Bevezetés

A napjainkban nagy népszerűségnek örvendő adatbányászatot (angolul *Data Mining*) az üzleti, mérnöki és tudományos élet számos területén alkalmazzák már évtizedek óta. Célja a statisztika és a mesterséges intelligencia eszközeivel masszív adatbázisokban rejtőző, eddig nem ismert, hasznos összefüggések feltárása. Az adatbányászatot többek között biztosítótársaságok, bankok, kereskedelmi vállalatok, egészségügyi szervezetek alkalmazzák vásárlási szokások elemzésére, kereskedési és kockázati modellek, befektetési stratégiák létrehozására, direktmarketing stratégia meghatározására, pénzügyi portfolió optimalizálására, betegségek modellezésére, súlyossági esetek kiszűrésére.

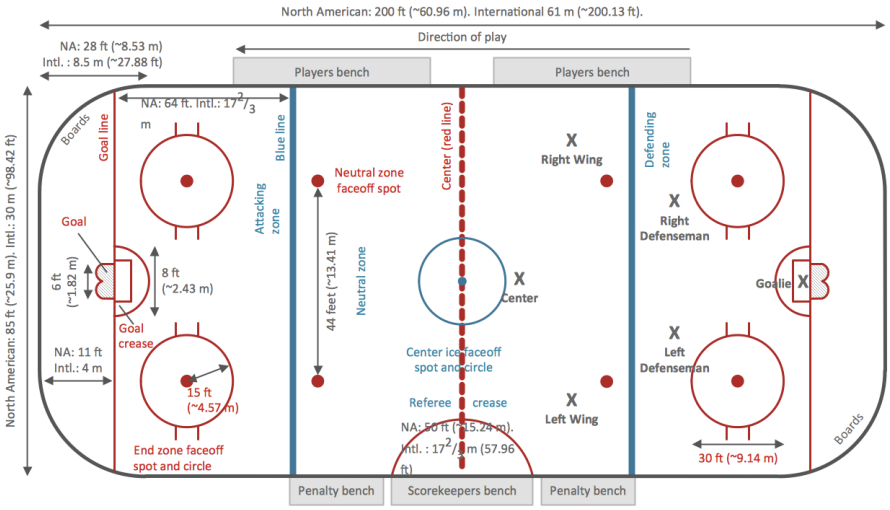
A 2000-es évek óta egyre több sportszervezet ismeri fel az adatbányászatban rejlő lehetőségeket [7]. Az adatok elemzésével feltárt információ új perspektívát nyitott számos területen: játékosmegfigyelők eddig ismeretlen, új metrikák alkalmazásával új tehetségeket fedezhetnek fel; az edzői stáb árnyaltabb képet kaphat a játékosok teljesítményéről, és még sorolhatnánk.

Ebben a cikkben egy, a jégkorongsportban eddig nem használt, új alkalmazást mutatunk be. A jégkorong intenzivitása miatt a játékosok átlagban 45 másodpercenként cserélik egymást. Játékmegszakítás esetén az edző gyakran – bár nem

minden esetben – egy teljesen új sort küld be a korongbedobáshoz. Modellünk célja, hogy a játék során eddig történt, egymást követő események alapján előrejelezzük, hogy melyik öt mezőnyjátékos játékba állítása maximalizálná a gólszerzés valószínűségét. Mivel egymást követő, sorrendfüggő adatsorokkal dolgozunk, a feladat idősorelemzésként fogható fel. A modell alapján az edző minden korongbedobás előtt, valós időben optimalizálhatja a csapatösszeállítást.

1.1. Jégkorong

A jégkorong gyors, dinamikus csapatsport. A játék célja, hogy a játékosok a korongot egy ütő segítségével az ellenfél kapujába juttassák. Egy csapat általában 18 mezőnyjátékosból és 2 kapusból áll. Ebből 1 kapus és 5 mezőnyjátékos van egyszerre a jégen, kivéve, ha szabálytalanság miatt a csapat 2 vagy 5 percre emberhátrányba kerül.



1. ábra. Jégkorong pálya

Egyszerre több játékos is kiállítható, de a kapus mellett legalább három mezőnyjátékos mindig a jégen tartózkodik. Amennyiben több, mint két játékost állítanak ki, akkor a harmadik játékos akkor kezdi letölteni a büntetését, amikor az első játékosé véget ér. Az ellenfél által lőtt gól automatikusan törli az éppen aktuális 2 perces büntetést, ám az 5 perces büntetést nem.

Mivel a jégkorong nagyon gyors és intenzív játék, az 5 játékosból álló úgynevezett sorok átlagban 45 másodpercenként váltják egymást a jégen.

A mezőnyjátékosok két csoportba sorolhatók: támadók és védők. A támadók lehetnek balszélsők, centerek vagy jobbszélsők. Egy sorban 3 támadó játékos sze-

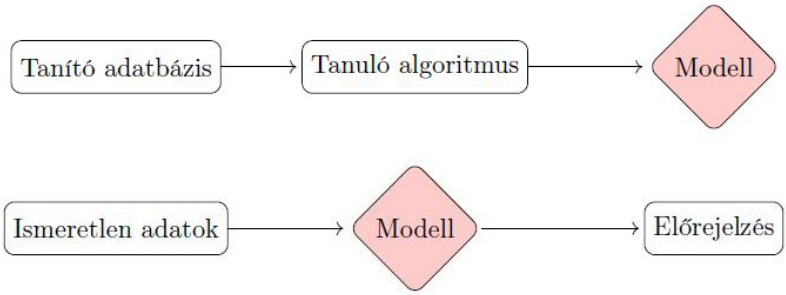
repel. A védők lehetnek bal- vagy jobbhátvédek, és általában párban játszanak. A sorcserék alkalmával általában egy új védőpár és egy új támadósor lép a jégre. Az edzők a sorokat a mérkőzés alatt bármikor cserélhetik, nem szükséges, és nem is szokás mindig megvárni a következő játékmegszakítást.

A jégkorongmérkőzés három, 20 perces harmadból áll. A 20 perc tiszta játékidőt jelent, vagyis a játék megszakítása alkalmával az órát megállítják. Játékmegszakítás után a korongot a bedobással (bulival) hozzák játékba. A két center szemben áll a bedobóhelyen, a játékvezető pedig bedobja közéjük a korongot, amit ők igyekeznek ütőjükkel saját csapattársukhoz juttatni.

1.2. Adatbányászat és gépi tanulás

Az adatbányászat célja a masszív adatbázisokban rejtőző szabályszerűségek, minták feltárása. A minták ismeretlen adatokra való alkalmazásával előrejelzéseket tehetünk, amelyek potenciálisan befolyásolhatják a felhasználó jövőbeni döntéseit. A feladat általában egy minél megbízhatóbb előrejelzéseket adó modell kifejlesztése.

A gyakorlatban a fenti probléma megoldását gyakran egy gépi tanulási (*Machine Learning*) algoritmus szolgáltatja. A gépi tanuló algoritmusnak az úgynevezett tanulási szakaszban példaadatokat (tanító adatbázis) szolgáltatunk, amely ezek alapján szabályszerűségeket határoz meg, azaz „tanul”. A szabályszerűségeket leíró modellt új, eddig ismeretlen adatokra alkalmazva „megjósolhatjuk” a hozzájuk tartozó célfüggvényértéket (2. ábra).



2. ábra. A gépi tanulás folyamatábrája

A gépi tanulási algoritmusok főbb típusai a *felügyelt*, *felügyelet nélküli*, illetve *megerősítéses* tanulás. Modellünkben felügyelt és megerősítéses tanulási algoritmusokat használunk fel.

Felügyelt tanulás esetén a célunk egy függvény közelítése. A tanító adatbázis a megtanulni kívánt függvény bemeneti és kimeneti értékeit tartalmazza. Formá-

lisan: legyenek adottak (x, y) párok, ahol $x \in \mathbb{R}^{d\times m}, y \in \mathbb{R}^{1\times m}$; előállítandó az az $f : \mathbb{R}^{d\times m} \to \mathbb{R}^{1\times m}$ függvény, amely minden (x_j, y_j) párra teljesíti, hogy $y_j = f(x_j), j = 1 \dots m$.

Az x mátrix oszlopait *független* változóknak nevezzük, az y vektort *függő változónak* vagy *célváltozónak*. Minden x_j érték valamely objektum vagy esemény leírása (pl. egy játékos hány másodperce van a jégen). Az y_j értékek az x_j értékekből történő következtetéseket reprezentálják (pl. az adott játékost lecserélik a következő másodpercben). Feltételezzük, hogy a tanulás során az y_j értékek előre meghatározottak. A bemenő adatok, azaz x_j értékei lehetnek számszerűek (pl. eltelt másodpercek száma), kategorizáltak (pl. igen/nem), de lehetnek valamely adat-előfeldolgozás eredményeként kapott értékek is (pl. átlag, maximum, minimum). Ha y-nak csak két lehetséges értéke van (pl. igen/nem), akkor fogalmi tanulásról (concept learning) beszélünk. Ebben az esetben a tanító példákat két diszjunkt részhalmazra lehet bontani: a pozitív és a negatív példák halmazára.

Diszkrét értékkészletű f függvény tanulását osztályozásnak (classification), folytonos értékkészletűét regressziónak (regression) nevezzük. A modellünkben különböző osztályozó algoritmusokat használunk, például döntési fákat [4, 3], logisztikus regressziót [2], k-legközelebbi szomszéd algoritmust [5] stb.

1.3. Mérőszámok osztályozás hatékonyságának jellemzésére

1.1. Definíció. Tekintsünk egy bináris osztályozást. A két lehetséges kimenetet nevezzük pozitívnak (P), illetve negatívnak (N) [1, 861–874. oldal]. Jelölje TP (igaz pozitív, *true positive*) a helyesen pozitívként osztályozott, FP (hamis pozitív) a tévesen pozitívként osztályozott eseteket. Hasonlóan jelölje TN és FN az igaz negatív, illetve hamis negatív előrejelzéseket. A következő mérőszámok alkalmasak az osztályozás teljesítményének értékelésére:

1. Pontosság (*Accuracy*):
$$ACC = \frac{TP + TN}{P + N}$$
2. Osztályozási hiba (*Classification Error*):
$$CE = 1 - ACC$$
3. Igaz pozitív arány (*Érzékenység*):
$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$
4. Hamis pozitív arány (*False positive rate*):
$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

1.2. Definíció. Tekintsünk egy általános osztályozási problémát, legyen az osztályok száma n, $n \geq 2$. Legyen a $C \in \mathbb{R}^{n \times n}$ mátrix olyan, hogy $C_{i,j}$ megegyezik azon esetek számával, amelyek az i-edik osztályba tartoznak, de a j-edik osztályba soroltuk őket. Egy ilyen mátrixot tévesztési mátrixnak (*confusion matrix*) nevezünk. A korrekt predikciók számát a diagonális elemek összege adja:

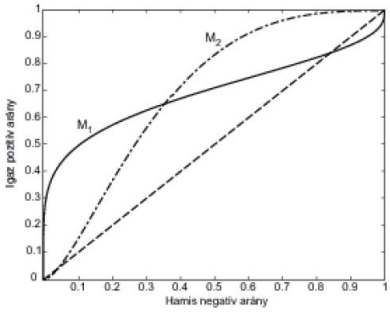
$$|Korrekt\ predikciók| = \sum_{i=1}^{n} C_{i,i}$$

Osztályok	Valódi érték	
Predikció	TP	FP
	FN	TN

1. táblázat. Bináris osztályozás tévesztési mátrixa

Vevő működési karakterisztika görbe

A vevő működési karakterisztika (Receiver Operating Characteristic vagy ROC) görbe grafikus módszer az osztályozó algoritmus hatékonyságának jellemzésére. A görbe (3. ábra) generálásakor az érzékenységet (TPR) az y tengelyen ábrázoljuk, a hamis pozitív arány (FPR) pedig az x tengelyen látható [8]. Egy osztályozási modell akkor a legjobb, ha a TPR maximális, FPR pedig minimális, ebben az esetben a ROC-görbe áthalad az egységnégyzet bal felső csúcsán. Egy véletlenszerűen találgató modell görbéje közelítőleg a főátló mentén fog elhelyezkedni.



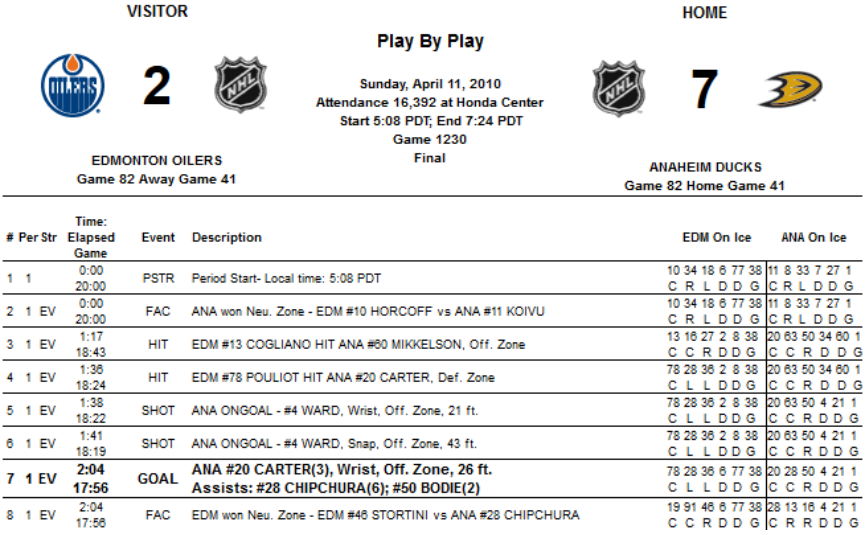
3. ábra. Két különböző osztályozás ROC görbéje

A ROC-görbe alkalmazható több osztályozó algoritmus összehasonlítására is. Ilyenkor a különböző algoritmusok eredményei egy ROC-ábrán jeleníthetők meg, a görbék egymáshoz való viszonyuk és az egységnégyzetbeli elhelyezkedésük alapján hasonlíthatók össze.

A hatékonyság számszerű értékét a görbe alatti terület nagyságával (ROC AUC) adhatjuk meg. Minél közelebb van ez az érték az 1-hez, annál hatékonyabb az algoritmusunk.

2. Az adatok

Az adatokat az NHL (National Hockey League) hivatalos weboldalán (www.nhl.com) található mérkőzések statisztikái szolgáltatják. A modellépítés során az Anaheim Ducks 2009 és 2012 között lejátszott mérkőzéseit használtuk.



VISITOR 2 — HOME 7

Play By Play

Sunday, April 11, 2010
Attendance 16,392 at Honda Center
Start 5:08 PDT; End 7:24 PDT
Game 1230
Final

EDMONTON OILERS
Game 82 Away Game 41

ANAHEIM DUCKS
Game 82 Home Game 41

#	Per	Str	Time: Elapsed Game	Event	Description	EDM On Ice	ANA On Ice
1	1		0:00 20:00	PSTR	Period Start- Local time: 5:08 PDT	10 34 18 6 77 38 C R L D D G	11 8 33 7 27 1 C R L D D G
2	1	EV	0:00 20:00	FAC	ANA won Neu. Zone - EDM #10 HORCOFF vs ANA #11 KOIVU	10 34 18 6 77 38 C R L D D G	11 8 33 7 27 1 C R L D D G
3	1	EV	1:17 18:43	HIT	EDM #13 COGLIANO HIT ANA #60 MIKKELSON, Off. Zone	13 16 27 2 8 38 C C R D D G	20 63 50 34 60 1 C C R D D G
4	1	EV	1:36 18:24	HIT	EDM #78 POULIOT HIT ANA #20 CARTER, Def. Zone	78 28 36 2 8 38 C L L D D G	20 63 50 34 60 1 C C R D D G
5	1	EV	1:38 18:22	SHOT	ANA ONGOAL - #4 WARD, Wrist, Off. Zone, 21 ft.	78 28 36 2 8 38 C L L D D G	20 63 50 4 21 1 C C R D D G
6	1	EV	1:41 18:19	SHOT	ANA ONGOAL - #4 WARD, Snap, Off. Zone, 43 ft.	78 28 36 2 8 38 C L L D D G	20 63 50 4 21 1 C C R D D G
7	**1**	**EV**	**2:04 17:56**	**GOAL**	**ANA #20 CARTER(3), Wrist, Off. Zone, 26 ft. Assists: #28 CHIPCHURA(6); #50 BODIE(2)**	78 28 36 6 77 38 C L L D D G	20 28 50 4 21 1 C C R D D G
8	1	EV	2:04 17:56	FAC	EDM won Neu. Zone - EDM #46 STORTINI vs ANA #28 CHIPCHURA	19 91 46 6 77 38 C C R D D G	28 13 16 4 21 1 C R R D D G

4. ábra. NHL mérkőzésstatisztika

A statisztikák (4. ábra) rögzítenek minden, a játék szempontjából érdekes információt, pl. a játékmegszakítások időpontját, azok okát (gól, büntetés, les stb.), az adott pillanatban jégen tartózkodó játékosok mezszámát. Átlagosan 20 másodpercenként kapunk új információt, azaz minden meccshez 180–200 adatsor tartozik.

A tanulási folyamat megkezdése előtt az adatok előfeldolgozása szükséges. A nyers adatokból a mérkőzés minden játékmegszakítását egy $\mathbb{R}^{d+1}$-beli vektorral reprezentáljuk, amely tartalmazza a $d > 300$ darab független változó és a célváltozó aktuális játékmegszakításhoz tartozó értékét. Így minden buli előtt van egy adatsorunk, amely aggregált információt tartalmaz az eddig történt játékeseményekről.

Ez után az adatbázist kettéválasztjuk: az adatok 80%-án tanítjuk, a maradékon teszteljük a modellt.

Néhány példa független változókra:

– Kategorikus változók:

  – a buli előtti játékesemény (harmad kezdete, harmad vége, kiállítás, gól stb.),

  - a buli előtt jégen tartózkodó játékosok mezszáma az egyes pozíciókban (balszélső, jobbszélső, center, balhátvéd, jobbhátvéd, kapus).

- Numerikus változók:
  - a mérkőzésen eddig eltelt másodpercek száma,
  - a legutóbbi játékmegszakítás óta eltelt másodpercek száma,
  - egy játékos átlagosan mennyit játszott az előző 1/5/10 cserében.

- Bináris változók:
  - minden játékosra bevezetünk egy változót, amely megadja, hogy az adott játékos jégen volt-e az előző 1/30/60/90 másodpercben, illetve az előző mérkőzésen,
  - emberhátrány – értéke 1, ha a csapat emberhátrányban van, és 0, ha nem.

A végső modellben több mint 300 független változóval dolgozunk. A túl sok változó alkalmazása lelassítja a futásidőt, valamint pontatlan eredményhez vezethet. Ezt kiküszöbölendő, feature selection [5] módszerrel kiválogattuk a számunkra leghasznosabb független változókat. Az eljárás minden változóhoz hozzárendel egy úgynevezett hasznossági értéket, ami azt hivatott tükrözni, hogy az adott változó mennyi információt hordoz. A modell minden olyan változót megtart, amelynek a hasznossági értéke nagyobb, mint az összes érték mediánjának k-szorosa. A modell alapbeállítása mellett $k = 1$ értékkel dolgozunk.

3. A modell

A modellünk három részből áll. Az első rész minden játékosra előre jelzi, hogy a következő bulinál jégen lesz-e vagy sem, valamint tárolja a játékosokhoz tartozó játékbakerülési valószínűséget is. A második rész az első eredményei alapján minden pozícióra meghatározza a két legmagasabb valószínűséggel játékba kerülő játékost, valamint visszatér a legmagasabb valószínűségű játékos mezszámával – azaz előre jelzi, hogy pontosan kik alkotják majd a sort a korongbedobásnál. A harmadik rész a második modell eredményeit felhasználva megvizsgálja, hogy a két legnagyobb valószínűséggel rendelkező játékosok mely kombinációja adja a legnagyobb gólszerzési valószínűséget, és visszatér azok számával.

3.1. Első modell – játékbakerülési valószínűségek

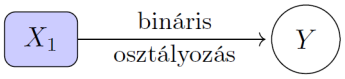
Az első modellben minden játékoshoz hozzárendelünk egy bináris változót, amely reprezentálja, hogy a játékos jégen lesz-e a most következő bulinál, vagy sem. Ezt egy osztályozó algoritmus célváltozójaként alkalmazzuk. A független változók

aggregált információt tartalmaznak a $(j-1)$-edik játékmegszakításig bekövetkezett játékeseményekről.

Alapesetben az algoritmus becslést ad a célváltozó lehetséges értékeinek valószínűségére. Amennyiben a valószínűség értéke legalább $0,5$, úgy az algoritmus kimeneti értéke 1 lesz, egyébként 0.

A modell kimeneti mátrixának oszlopvektorai a beállítástól függően reprezentálhatják az egyes játékosokhoz tartozó előrejelzéseket, vagy a nyers valószínűségeket.

Legyen X_1 a független változók mátrixa, és Y a kimeneti mátrix. A modell foyamatábrája a következő:



3.1.1. Célváltozók

Legyen $y_p \in \mathbb{R}^d$ a p játékoshoz tartozó célváltozó, ahol d az adatsorok (korongbedobások) száma. Jelölje j a buli sorszámát az adatsorunkban, $j \geq 1$. Ekkor:

$$y_{pj} = \begin{cases} 1, & \text{ha a } p \text{ játékos jégen van a } j\text{-edik bulinál,} \\ 0 & \text{egyébként.} \end{cases} \tag{1}$$

A vizsgált három szezon alatt 55 játékos játszott az Anaheim Ducks csapatában, tehát 55 bináris célváltozónk van: $p \in \{1, \ldots, 55\} =: P$. A modell minden (p, j) párra előrejelzi az y_{pj}, $j \geq 2$ értéket a j-edik játékmegszakítás előtt bekövetkezett események alapján.
A célmátrix:

$$Y = \{y_1, y_2, \ldots, y_{55}\} \in \mathbb{R}^{d \times 55}. \tag{2}$$

Megjegyzés. A kapussal együtt a játék minden pillanatában 4–6 játékos van a jégen, ezért

$$4 \leq \sum_{p \in P} y_{pj} \leq 6, \ \ \forall\, j \in J = \{1, 2, \ldots, d\}\,.$$

3.1.2. Független változók

Jelölje $\hat{y}_{ij} = \{0, y_{i2}, y_{i3} \ldots, y_{id}\} \in \mathbb{R}^d$ azt a vektort, amely megmondja, hogy az i-edik játékos a jégen volt-e a j-edik játékmegszakítás pillanatában.
Az alapmodell független változói:

$$x_{ij} = \hat{y}_{ij}, i \in P, j \in \{1, 2, \ldots, d\}\,, \tag{3}$$

azaz a sorösszeállítás a játékmegszakítás pillanatában.

Megjegyzés. Természetesen csak $j \geq 2$ esetén van információnk a mérkőzésről, ezért $\hat{y}_{i1} = 0$, $\forall i \in P$.

Az alapmodell teljesítményét javítandó újabb változókat adtunk a modellhez:

– SH/PP bináris változókat, amelyek értéke 1, ha a csapat emberhátrányban/emberelőnyben játszik, és 0, ha nem,

– az előző játékosmegszakítás óta eltelt másodpercek számát,

– a jégen töltött másodpercek átlagos számát az utolsó 5/10 cserében, illetve az elmúlt 30/60/90 másodpercben minden játékosra.

Jelölje $X = \{\hat{y}_1, \ldots, \hat{y}_{55}, x_1 \ldots, x_m\} \in \mathbb{R}^{d\times(m+55)}$ az összes független változót oszlopvektorként tartalmazó mátrixot, ahol d a bulik száma, és m a hozzáadott független változók száma, $m \geq 0$.

3.1.3. A modell kimenete

A kimeneti mátrix oszlopvektorai a modell beállításától függően reprezentálhatják az egyes játékosokhoz tartozó nyers valószínűségeket, vagy az előrejelzett osztályokat. A döntési határ $0,5$.

$$out_1 = \overline{Y_1} = \left(P\left(y_1 = 1\right), \ldots, P\left(y_{55} = 1\right)\right), \tag{4}$$

$$out_2 = \overline{Y_2} = \left(\overline{y}_1, \ldots, \overline{y}_{55}\right), \tag{5}$$

$$\overline{y}_i = \begin{cases} 1, & \text{ha } P\left(y_i = 1\right) \geq 0,5, \\ 0 & \text{egyébként.} \end{cases}$$

3.1.4. Az eredmények kiértékelése

Az alapmodell a (3) egyenletben megadott független változókkal, logisztikus regresszióval és a feature selection alapbeállításával ($k = 1$) próbálja előrejelezni, hogy egy játékos jégre lép-e a korongbedobásnál.

A modell meglehetősen gyenge teljesítményt nyújt, a ROC-görbe alatti terület a legtöbb játékos esetén $0,65$ alatt van, ami azt jelenti, hogy a modell alig teljesít jobban, mintha véletlenszerűen találgatnánk.

A végső modellben a 3.1.2 részben leírtak szerint kibővítettük a független változók halmazát, valamint a feature selection eljárás döntési határát $k = 2,5$-re változtattuk. Az osztályozó algoritmus továbbra is logisztikus regresszió.

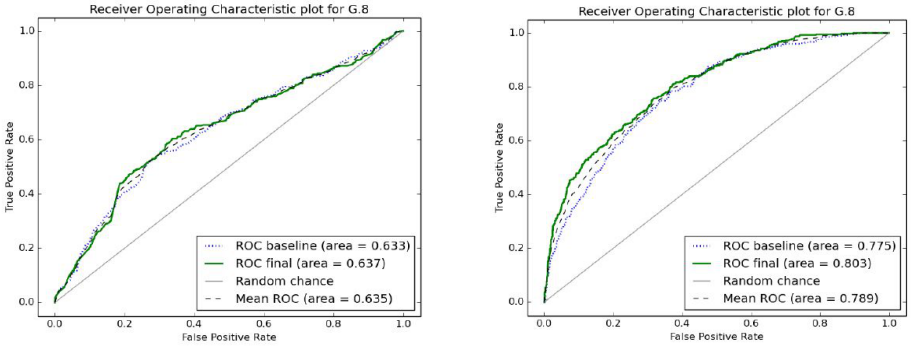
A modell teljesítménye a fenti beállítások mellett szignifikáns javulást mutatott, a játékosokhoz tartozó ROC-görbe alatti terület $0,8$ körüli értéket vett fel, több esetben meg is haladta azt. Az osztályozás érzékenysége is jelentősen javult.

Az 5(a). és 5(b). ábra az Anaheim legendás jobbszélsőjéhez, Teemu Selännehoz tartozó ROC-görbéket mutatja. A pontozott vonal mindkét esetben az aktuális modell feature selection nélküli teljesítményét reprezentálja, a vastag vonal

ugyanezt feature selection alkalmazásával. A 2. táblázat tartalmazza a részletes teljesítmény-értékelést.

Teemu Selänne #8, jobbszélső	**A modell teljesítménye**			
	Tanító adatok		**Tesztadatok**	
	Alapmodell	**Végső modell**	**Alapmodell**	**Végső modell**
Tévesztési mátrix	4031 56 1413 69	3868 219 757 471	878 30 369 30	828 80 207 192
Pontosság	0, 7714	0, 8163	0, 6947	0, 7804
Érzékenység	0, 0561	0, 3835	0, 0751	0, 4812
ROC AUC	0, 63	0, 77	**0, 64**	**0, 80**

2. táblázat. A modell teljesítménye Selänne esetében



(a) Alapmodell (b) Végső modell

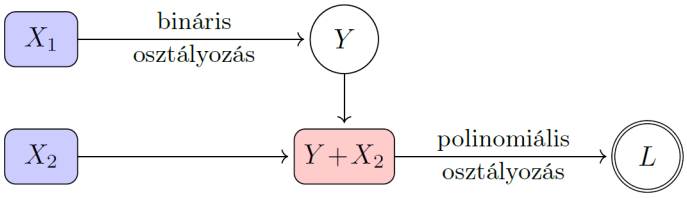
5. ábra.

3.2. Második modell – sorprediktor

A második modell polinomiális osztályozással minden játékmegszakításnál minden pozícióhoz hozzárendel két játékost, akik a legnagyobb valószínűséggel jégre lépnek a következő bulinál. A modell tárolja a játékosok azonosítóját és a hozzájuk tartozó valószínűségeket, valamint alapbeállításban visszatér a magasabb valószínűséggel rendelkező játékos mezszámával. Ezáltal előrejelzést ad a következő sorösszeállításra.

Megjegyzés. Amennyiben a csapat emberhátrányban játszik, a hiányzó játékos pozíciójában a 0 érték szerepel.

Jelölje X_1 az első modell prediktor mátrixát, Y a (4) egyenletben megadott output mátrixát. Legyen X_2 a második modellhez tartozó prediktormátrix. A kimenetet (ami az előrejelzett sor) jelöljük L-lel. A modell folyamatábrája a következő:



3.2.1. Célváltozók

Jelölje $l_{i,j} \in \mathbb{Z}$ a j-edik korongbedobásnál az i-edik pozícióban jégre lépő játékos mezszámát, ahol $i \in \{1, 2, \ldots, 6\}$, $j \in \{1, \ldots, d\}$.

A célváltozók mátrixa:

$$line = (l_{1,j}, l_{2,j} \ldots, l_{6,j}) \in \mathbb{Z}^{d \times 6}. \tag{6}$$

3.2.2. Független változók

Az alapmodellhez tartozó független változók:

$$x_{ij} = prob_{ij}, \text{ ahol}$$

$prob_{ij} = P\left(y_{ij} = 1\right), i \in P$ az első modellből kapott valószínűségek, y_{ij} pedig az (1) egyenlettel megadott változó. Tehát $prob_{ij}$ jelöli annak a valószínűségét, hogy a i-edik játékos jégre lép a j-edik bulinál, $\forall i \in P$.

A modell továbbfejlesztéséhez újabb változókat adtunk a modellhez:

- SH/PP, a 3.1.2-ben megadott változók,
- az előző játékmegszakítás óta eltelt másodpercek számát,
- a játékosokhoz tartozó jégrelépési valószínűségek az első modell különböző beállításaival (logisztikus regresszió helyett döntési fák, véletlen erdők, valamint adaboost [6, 657-663. oldal] algoritmus alkalmazása; a feature selection döntési határa minden esetben $k = 1, 5$).

A teljes prediktormátrix:

$$X = \{prob_1, \ldots, prob_{55}, x_1, \ldots, x_m\} = \left(\hat{P}, \hat{X}\right) \in \mathbb{R}^{d \times (m+55)},$$

ahol d a bulik száma, m a hozzádott változók száma, $m \geq 0$ és $\hat{P}$ a (4) egyenletben megadott mátrix.

3.2.3. A modell kimenete

A modell kimenete egy olyan mátrix, amely megadja az egyes játékosok jégrelépési valószínűségét minden pozícióban; vagy a legmagasabb valószínűséggel rendelkező játékosok mezszámait.
Jelölje $\#_i$ az i-edik játékos számát, $i \in \{1, \ldots, 55\}$:

$$out_1 = \overline{P} = \left(P\left(l_1 = \#_i\right), \ldots, P\left(l_6 = \#_i\right)\right) \in \mathbb{Z}^{d\times 6}, \forall i \tag{7}$$

$$out_2 = line^* = \left(l_1^*, \ldots, l_6^*\right), \text{ ahol} \tag{8}$$

$$l_k^* = \max_i \left(P\left(l_k = \#_i\right)\right), \quad k \in \{1, \ldots, 6\}.$$

3.2.4. Az eredmények kiértékelése

Az alapmodell polinomiális logisztikus regresszióval, feature selection nélkül tesz előrejelzést a sorösszeállításra. Akárcsak az előző esetben, a modell teljesítménye meglehetősen gyenge. A tanító adatokon $49,2\%$-os hibával jósolja meg a sorösszeállítást, a tesztadatokon a hiba közel 53%-os. Itt megjegyeznénk, hogy a sorösszeállítás helyes előrejelzése jóval komplexebb feladat, mint az egyes játékosok jégrelépési valószínűségének kiszámítása, ám ez az eredmény ennek ellenére sem elfogadható.

A végső modell a kibővített prediktormátrix mellett továbbra is polinomiális logisztikus regresszióval dolgozik. A feature selection eljárás döntési határa $k = 1,5$. A modell hibája a tanító adatokon $33,7\%$-ra, a tesztadatokon $43,4\%$-ra csökkent.

A $43,4\%$-os hibaarány elfogadható, de nem kiemelkedően jó eredmény. Jelenleg is dolgozunk a modell javításán, pl. további független változók bevezetésével (ellenfél sorösszeállítása a játékmegszakítás pillanatában; egyéni játékos-statisztikák stb.), illetve különböző osztályozási algoritmusok alkalmazásával.

A 3. és 6. táblázat részletes leírást ad az egyes modellékosztályozási hibákról, a 4., 5., 7. és 8. táblázat pedig példát ad néhány sor-előrejelzésre az alapmodell és a végső modell segítségevel a tanító, illetve tesztadatokon.

3.3. Harmadik modell – alternatív sorok

A harmadik modell célja, hogy olyan alternatív sorösszeállítást találjon, amely legalább akkora valószínűséggel szerez gólt a következő cserében, mint az előző modellben előrejelzett sor.

Az alternatív sorokat az előzőleg minden pozícióra meghatározott két legmagasabb valószínűségű játékos összes kombinációi között keressük. Ez $2^5 = 32$ alternatív sort jelent. A modell minden alternatív sorra, mint a modellhez tartozó prediktormátrix részére lefuttat egy bináris osztályozást. Az osztályozás célváltozója reprezentálja azt, hogy a következő játékmegszakításig lőnek-e gólt a

Alapmodell	**Osztályozási hiba**		
	Teljes adathalmaz	**Tanító adatok**	**Tesztadatok**
Támadó sorok	0,63	0,625	0,65
Védő sorok	0,547	0,541	0,572
Kapus	0,078	0,072	0,106
Összegzés	0,499	0,492	**0,529**

3. táblázat. Alapmodell

FO	**Valódi sorösszeállítás**						**Predikció**					
	C	**JSz**	**BSz**	**JH**	**BH**	**K**	**C**	**JSz**	**BSz**	**JH**	**BH**	**K**
1.	11	8	14	19	34	35	15	9	14	7	27	35
2.	15	9	10	7	27	35	15	9	14	19	34	35
3.	11	8	14	19	34	35	15	9	0	7	27	35
4.	15	9	10	7	27	35	15	22	10	19	34	35
5.	15	9	10	7	27	35	15	9	10	7	27	35

4. táblázat. Tanító adatok

FO	**Valódi sorösszeállítás**						**Predikció**					
	C	**JSz**	**BSz**	**JH**	**BH**	**K**	**C**	**JSz**	**BSz**	**JH**	**BH**	**K**
1.	7	41	39	5	21	1	7	8	10	4	23	1
2.	63	8	9	17	32	1	15	8	10	17	32	1
3.	14	16	19	5	21	1	15	9	10	5	21	1
4.	14	8	19	5	21	1	15	16	10	5	21	1
5.	14	8	19	5	21	1	15	9	10	5	21	1

5. táblázat. Tesztadatok

Végső modell	Osztályozási hiba		
	Teljes adathalmaz	Tanító adatok	Tesztadatok
Támadó sorok	0,468	0,449	0,547
Védő sorok	0,400	0,384	0,467
Kapus	0,018	0,005	0,073
Összegzés	0,37	0,337	**0,434**

6. táblázat. Végső modell

FO	Valódi sorösszeállítás						Predikció					
	C	JSz	BSz	JH	BH	K	C	JSz	BSz	JH	BH	K
1.	11	8	14	19	34	35	11	8	14	7	34	35
2.	15	9	10	7	27	35	15	9	10	7	27	35
3.	11	8	14	19	34	35	11	8	10	7	34	35
4.	15	9	10	7	27	35	22	9	10	7	27	35
5.	15	9	10	7	27	35	15	9	10	7	27	35

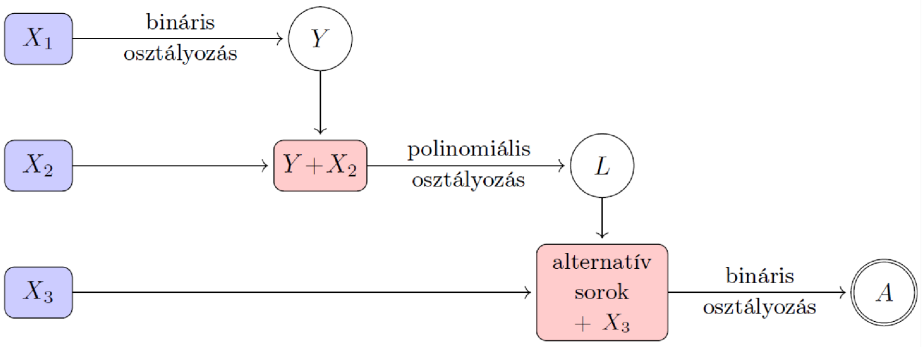
7. táblázat. Tanító adatok

FO	Valódi sorösszeállítás						Predikció					
	C	JSz	BSz	JH	BH	K	C	JSz	BSz	JH	BH	K
1.	7	41	39	5	21	1	7	8	10	4	23	1
2.	63	8	9	17	32	1	63	8	9	17	32	1
3.	14	16	19	5	21	1	15	9	10	5	23	1
4.	14	8	19	5	21	1	14	16	19	5	21	1
5.	14	8	19	5	21	1	15	16	19	5	21	1

8. táblázat. Tesztadatok

játékosok. Eredményként minden sorhoz kapunk egy gólszerzési valószínűséget, amelyből ezután kiválasztjuk a legnagyobb valószínűségű sort.

Legyen X_1 az első modell prediktormátrixa, Y a kimeneti mátrixa. Jelölje X_2 és X_3 a második, illetve harmadik modell prediktormátrixát. Jelöljük L-lel a különböző pozíciókhoz tartozó két legmagasabb valószínűségű játékost tartalmazó listát, és legyen A a harmadik modell kimeneti mátrixa. A modell folyamatábrája a következő:



3.3.1. Célváltozó

Legyen $g \in \mathbb{R}^d$ az osztályozás célváltozója, ahol d a bulik száma, $j \in \{1 \dots d\}$. Ekkor:

$$g_j = \begin{cases} 1, & \text{ha a csapat gólt szerez a } j\text{-edik cserében} \\ 0, & \text{ha nem.} \end{cases} \tag{9}$$

3.3.2. Független változók

Az alapmodell független változói:

$$x_{i1} = 0, \tag{10}$$

$$x_{ij} = line_{j-1}, j \in \{2, 3, \dots, d\}, \tag{11}$$

ahol $line_j = (l_{1j}, l_{2j} \dots, l_{6j}), \forall j \in \{1 \dots d\}$, ahol l_{ij} az i-edik pozícióhoz tartozó játékos mezszáma (azaz a sorösszeállítás a j-edik buli előtt).

A modell továbbfejlesztéséhez hozzádott változók:

- SH/PP, a 3.1.2-ben megadott változók,
- az előző játékmegszakítás óta eltelt másodpercek száma,
- a játékosok lövési hatékonysága, pontjaik száma (gól + gólpassz), valamint a büntetésperceik száma.

A teljes prediktormátrix:

$$X = \{l_1, l_2, \ldots, l_6, x_1, \ldots, x_m\} = \left(line, \hat{X}\right) \in \mathbb{R}^{d\times(m+6)},$$

ahol d a bulik száma, m a hozzádott független változók száma, $m \geq 0$, és $line$ a (6) egyenlettel megadott változó.

3.3.3. A modell kimenete

Az **osztályozáshoz tartozó kimenet** lehet a gólszerzés valószínűsége, vagy az előrejelzett osztály a c döntési határ mellett, ahol $c \in \mathbb{R}$, $0 \leq c \leq 1$:

$$out_1 = \overline{P} = \left(P\left(g_1 = 1\right), \ldots, P\left(g_d = 1\right)\right), \tag{12}$$

$$out_2 = \overline{G} = \left(\overline{g}_1, \ldots, \overline{g}_d\right), \text{ ahol} \tag{13}$$

$$\overline{g}_j = \begin{cases} 1, & \text{ha } P\left(g_j = 1\right) \geq c, \\ 0 & \text{egyébként.} \end{cases}, j \in \{1, \ldots, d\}$$

Jelölje Λ a lehetséges sorok halmazát, Γ a prediktormátrixok halmazát, Π a lehetséges gólszerzési valószínűségek halmazát. Jelölje $F : \Lambda \times \Gamma \to \Pi$ a (12) kimenettel rendelkező osztályozást.

Az alternatív sort előrejelző **modellhez tartozó kimenet** valamely $X' \in \Gamma$ adott prediktormátrix esetén:

$$line^* = \underset{line\in\Lambda}{\operatorname{argmax}} F\left(line, X'\right). \tag{14}$$

3.3.4. Példa

Vegyük a 4. bulit a 8. táblázatból. A bulihoz a valóságban a következő sor áll fel:

Valódi sor					
C	**JSz**	**BSz**	**JH**	**BH**	**K**
14	8	19	5	21	1

9. táblázat. A játékosok mezszámai

A második modell a következő játékosokat rendeli hozzá a pozíciókhoz, mint két legvalószínűbb jégrelépőt:

	Valószínűségek			
	#	p_1	#	p_2
Center	14	0, 669	15	0, 185
Jobb szélső	16	0, 329	51	0, 148
Bal szélső	19	0, 395	12	0, 143
Jobb hátvéd	5	0, 814	4	0, 114
Bal hátvéd	21	0, 772	32	0, 130
Kapus	1	0, 972	31	0, 028

10. táblázat. A második modell eredménye

A modell ezután generálja a játékosok összes lehetséges kombinációját. Az alternatív sorokat tartalmazó mátrixot jelölje $line_{12} \in \mathbb{Z}^{32\times 6}$:

$$line_{1,2} = \begin{bmatrix} 14 & 16 & 19 & 5 & 21 & 1 \\ 14 & 16 & 19 & 5 & 32 & 1 \\ 14 & 16 & 19 & 4 & 21 & 1 \\ 14 & 16 & 19 & 4 & 32 & 1 \\ 14 & 16 & 12 & 5 & 21 & 1 \\ 14 & 16 & 12 & 5 & 32 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

A modell ezután meghatározza a sorokhoz tartozó gólszerzési valószínűségeket:

$$\begin{bmatrix} 14 & 16 & 19 & 5 & 21 & 1 \\ 14 & 16 & 19 & 5 & 32 & 1 \\ 14 & 16 & 19 & 4 & 21 & 1 \\ 14 & 16 & 19 & 4 & 32 & 1 \\ 14 & 16 & 12 & 5 & 21 & 1 \\ 14 & 16 & 12 & 5 & 32 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 15 & 16 & 12 & 5 & 21 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \xrightarrow[\text{alkalmazása}]{\text{Osztályozó modell}} \begin{bmatrix} 0.0438 \\ 0.0327 \\ 0.0435 \\ 0.0325 \\ 0.0454 \\ 0.0339 \\ \vdots \\ 0.0457 \\ \vdots \end{bmatrix}$$

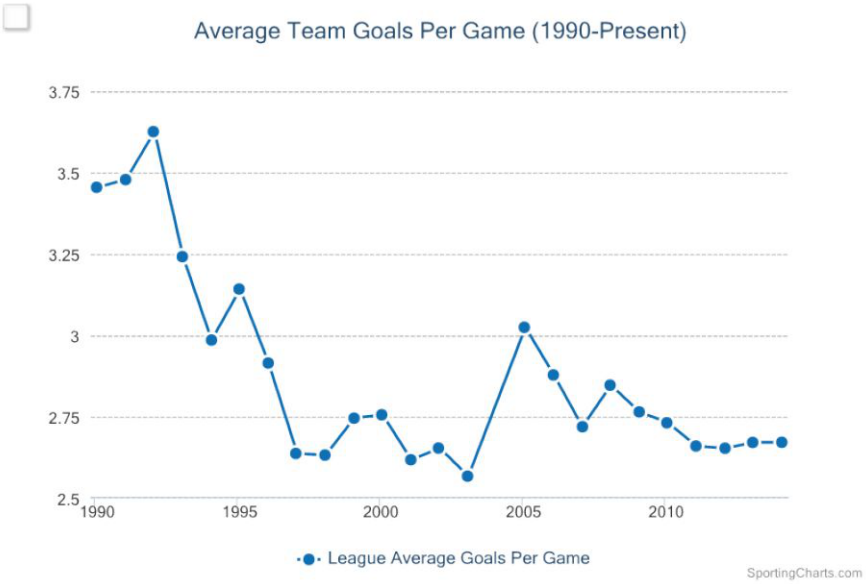
A modell megadja azt a sorösszeállítást, amelyhez a legnagyobb gólszerzési valószínűség tartozik:

Alternatív sor					
C	**JSz**	**BSz**	**JH**	**BH**	**K**
15	16	12	5	21	1

Megjegyzés. A valódi sorhoz tartozó valószínűség 0.0448 volt, tehát az alternatív sor valóban nagyobb valószínűséggel szerez gólt a következő cserében.

3.3.5. A gólprediktor modell kiértékelése

Az osztályozást hasonlóképpen értékeltük, mint az előző két esetben. Az alapmodellből kiindulva fejlesztettük ki a legjobb eredményt adó modellt. A gólok alacsony száma miatt (6. ábra, [9]) nehéz megjósolni, hogy a csapat szerez-e gólt a következő cserében, vagy sem. A gólszerzés valószínűsége nagyon alacsony, ezért az osztályozás a $0,5$-ös döntési határ mellett nagy valószínűséggel minden esetet negatívként osztályozna – azaz a csapat nem lő gólt a következő cserében – ezáltal a modell pontossága nagyon magas lesz, a szenzitivitása viszont nagyon alacsony. Ezt a döntési határ $0,5$-ről 0.06-ra változtatásával próbáltuk kiküszöbölni. Ez ugyan alacsonyabb pontossághoz vezetett, de a szenzitivitzás jelentősen nőtt, így az alapmodell és a végső modell eredményei összehasonlíthatóvá váltak.



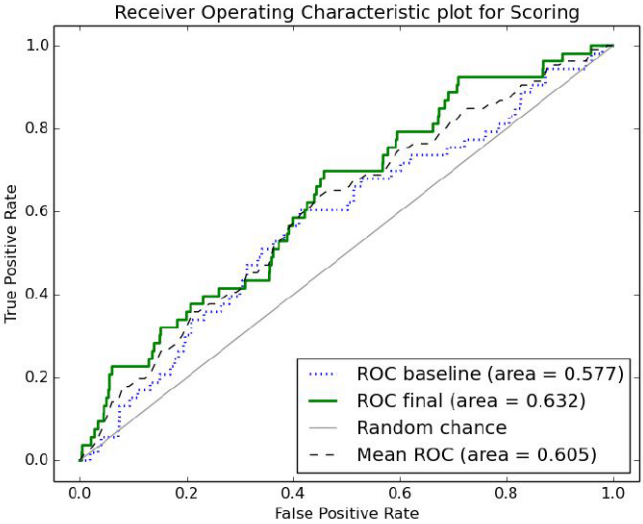
6. ábra. Gól/meccs átlag az NHL-ben

Az alap– és a végső modell esetében is logisztikus regressziót használtunk, a végső modell független változóit a 3.3.2 részben leírtak szerint kibővítettük. A feature selection határát 1-ről 2-re változtattuk.

Az alapmodell ismét nagyon gyenge teljesítményt nyújtott: a tanító adatokon 225-ből mindössze 4 gólt jelzett előre sikeresen, a tesztadatokon 53-ból egyet sem. A végső modellnél a sikeres előrejelzések száma 79 – 16-ra módosult, ezzel a érzékenységet sikerült 0%-ról 30%-ra emelni. Ez továbbra sem számít kiemelkedően jó eredménynek (7. ábra), ám úgy véljük, hogy alkalmasabb prediktor változók segítségével a teljesítmény javítható.

Gólprediktor modell	**A modell teljesítménye**			
	Tanító adatok		**Tesztadatok**	
	Alapmodell	**Végső modell**	**Alapmodell**	**Végső modell**
Tévesztési mátrix	5090 0 221 4	4367 723 146 79	1247 7 53 0	1058 196 37 16
Pontosság	0,9584	0,8365	0,9540	0,8217
Érzékenység	0,0177	0,3511	0	0,3018
ROC AUC	0,59	0,63	0,57	**0,62**

11. táblázat.



7. ábra. A gólprediktor modell ROC-görbéje

3.4. A modell teljesítménye - összegzés

A modell az esetek 74%-ában olyan sort választ ki, amelynek nagyobb a gólszerzési valószínűsége, mint a valóságban jégre küldött sornak. A sorprediktor modell javításával ez az arány megközelíthetné a 100%-ot, mivel ekkor a valóságban felküldött sor gyakrabban szerepelne a 32 alternatív sor között. Így az alternatív sorösszeállítás közelebb állna az edző eredeti elképzeléséhez, aki így komolyabb strukturális változtatások nélkül tudja optimalizálni a csapatot játék közben.

Természetesen a javítások sem garantálják, hogy a modell hosszú távon valóban növeli a csapat által lőtt gólok számát, ezt kizárólag élő tesztekkel igazolhatnánk.

Szeretném kifejezni őszinte hálámat dr. Tóth Jánosnak, aki már a kezdetektől rengeteg támogatást nyújtott, és aki nélkül ez a cikk nem valósulhatott volna meg, továbbá szeretném megköszönni Kangyal Balázsnak, a Magyar Jégkorong Szövetség sportigazgatójának, korábbi válogatott játékosnak és jelenleg aktív edzőnek a szakmai segítséget, és a modellfejlesztést elősegítő ötleteket. A dolgozat részben a K 84060 sz. OTKA pályázat támogatásával készült.

Hivatkozások

[1] T. Fawcett: *An Introduction to ROC Analysis, Pattern Recognition Letters*, **27** (2006)

[2] D.W. Hosmer: *Applied Logistic Regression* (2nd edition), Lemeshow, Stanley (2000).

[3] J.R. Quinlan: *C4.5: Programs for Machine Learning*, M. Kaufmann, (1993).

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone: *Classification and Regression Trees*, Wadsworth, Belmont, CA, (1984).

[5] T. Hastie, R. Tibshirani, J. Friedman: *Elements of Statistical Learning*, Springer, (2009).

[6] C.M. Bishop: *Pattern Recognition and Machine Learning*, 657–663.

[7] R.P. Schumaker, O.K. Solieman, H. Chen: *Sports Data Mining*, Springer US, (2010).

[8] `http://en.wikipedia.org/wiki/Receiver_operating_characteristic`

[9] `http://www.sportingcharts.com/nhl/`

(Beérkezett: 2015. április 21.)

SÜDY BARBARA
BME Analízis Tanszék
1111 Budapest, Egry J. u. 1.

REAL-TIME OPTIMIZATION OF ICE HOCKEY TEAMS

Barbara Südy

The goal of this paper is to introduce a novel data-driven approach for in-game decision making in ice hockey. Using predictive data mining techniques we build a model which attempts to determine the optimal team structure of an ice hockey team for the upcoming shift. The model gives the coach feedback on the optimal line combination in real time. The predictions of the model are based on the time series data that arises from the past game events.