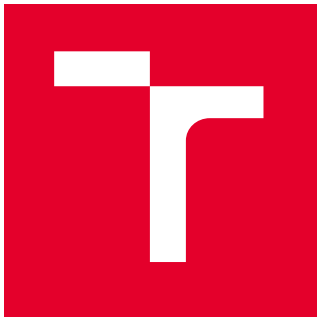

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ

FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY

INSTITUTE OF MATHEMATICS

MODELOVÁNÍ PRAVDĚPODOBNOSTI SKÓROVÁNÍ VE SPORTU

MODELING OF SCORING PROBABILITY IN SPORT

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **Ondřej Hilscher**

AUTHOR

VEDOUCÍ PRÁCE — **Ing. Pavel Hrabec, Ph.D.**

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání bakalářské práce

Ústav:	Ústav matematiky
Student:	**Ondřej Hilscher**
Studijní program:	Matematické inženýrství
Studijní obor:	bez specializace
Vedoucí práce:	**Ing. Pavel Hrabec, Ph.D.**
Akademický rok:	2021/22

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Modelování pravděpodobnosti skórování ve sportu

Stručná charakteristika problematiky úkolu:

Výsledné skóre v některých sportech není vždy objektivním měřítkem předvedeného výkonu. Zejména ve sportech, ve kterých je skórování "vzácnou" událostí (např. fotbal) mnohdy nereflektuje události na hřišti (velký vliv náhody). Pro objektivnější popis skutečného dění na hřišti lze použít některé statistické metody. Mezi příklady takových modelů lze zařadit např. tzv. "expected goals" ve fotbale či hokeji. Jedná se tedy v podstatě o převedení problému na modelování pravděpodobnosti skórování.

Cíle bakalářské práce:

Rešerše používaných modelů skórování ve zvoleném sportovním odvětví.
Seznámení s vhodnými nástroji matematické statistiky.
Analýza dat prostřednictvím vhodného softwarového nástroje (R, Python,).

Seznam doporučené literatury:

ANDĚL, Jiří. Základy matematické statistiky. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

AGRESTI, Alan. Categorical Data Analysis. 2nd ed. Hoboken: Wiley, 2002. Wiley series in probability and statistics. ISBN 04-713-6093-7.

Soccer analytics handbook (https://github.com/devinpleuler/analytics-handbook)

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2021/22

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt
Práce je zamìøena na modelování pravdìpodobnosti skórování ve fotbale. V práci je popsán nutný matematický aparát potøebný k sestavení modelu logistické regrese a základní testy statistických hypotéz. Popsaný matematický aparát je poté aplikován na volnì pøístupná data z profesionálních fotbalových utkání. Výsledný model pou¾ívá vysvìtlující promìnné jako zpùsob zakonèení, polohu na høi¹ti a zjednodu¹enì popsanou herní situaci.

Summary
This thesis aims for modelling of scoring probability in football. It describes necessary mathematical methods used in logistic regression model building and in basic statistical hypothesis tests. Afterwards the mathematical methods are used on available data from professional football matches. Resulting model uses shooting method, pitch location and simpli ed match situation as predictors.

Klíèová slova
logistická regrese, GLM, metoda maximální vìrohodnosti, Expected goals

Keywords
logistic regresion, GLM, maximum likelihood method, Expected goals

ONDØEJ HILSCHER Modelování pravdìpodobnosti skórování ve sportu. Brno: Vysoké uèení technické v Brnì, Fakulta strojního in¾enýrství, 2022. 23 s. Vedoucí diplomové práce Ing. Pavel Hrabec, Ph.D.

Prohla¹uji, ¾e jsem bakaláøskou práci Modelování pravdìpodobnosti skórování ve sportu napsal samostatnì pod vedením Ing. Pavla Hrabce Ph.D. a Ing. Martina Roseckého, v¹echny pou¾ité materiály jsou uvedeny v seznamu literatury.

Ondøej Hilscher

Dìkuji v¹em, kteøí mne bìhem vypracování podporovali a obohacovali mne svými zku¹enými a praktickými radami, zvlá¹tì Ing. Pavlu Hrabcovi Ph.D. a Ing. Martinu Roseckému, kteøí pomohli vytvoøit skvìlou atmosféru pro psaní práce, velké dík patøí také mé rodinì, její¾ podpora mì provází celé studium.

Ondøej Hilscher

Obsah

1. Sport a statistika

Pro mnoho lidí byly matematika a sport dlouhou dobu dva odli¹né svìty, a pro znaènou èást svìtové populace tomu tak stále je, ale ani¾ bychom si to uvìdomovali, matematika, konkrétnì statistika a statistické modely mohou mít a nìkde ji¾ mají zásadní vliv. Principy mìøení délky hodu, èasu, sèítání bodù i urèování poøadí jsou tisíce let staré. Samozøejmì byly vyu¾ívány jiné metody i jednotky, nicménì je vhodné zde uvést tvrzení, ¾e matematika sport provází ji¾ od jeho raných koøenù. Dne¹ní pohled na matematiku, statistiku a její modely ve sportu se zaèal utváøet v druhé polovinì minulého století, a» u¾ to byli nad¹enci z øad fanou¹kù, trenéøi, èi samotní sportovci, zaèaly se uchovávat informace o turnajích, utkáních nebo závodech, bez kterých si dnes nedoká¾eme pøedstavit ¾ádný televizní pøenos. U ka¾dé sportovní události je uvedena gra ka s mnoha statistickými údaji - nejdel¹í hod, poèet støel, es, úspì¹ných odpalù, dr¾ení míèe, úspì¹nost støelby, zákrokù gólmana, nejvìt¹í vyvinutá rychlost bìhem utkání, èas strávený na soupeøovì polovinì, tøetinì a takto by ¹lo vyjmenovat spoustu dal¹ích. Jistì si takovou tabulku ka¾dý, kdo vidìl nìjakou sportovní akci v televizi, vybaví.

Tato práce èteníøe uvede do prostøední datové analytiky sportu, uká¾e, jak mù¾e práce se sportovními daty (zde vyu¾ívány data rmy Stats Bomb[7]) vypadat v praxi a pøedstaví jejich vyu¾ití v oblasti sportu, a» u¾ z hlediska událostí uplynulých nebo tìch budoucích. K tomu jsou vyu¾ívány matematické resp. statistické nástroje, které jsou posléze aplikovány pøi vytváøení predikèních modelù pravdìpodobnosti skórování ve sportu. Modely jsou následnì testovány, tak aby bylo zjevné, který model by bylo vhodné pou¾ít, aby výsledky byly co nejpøesnìj¹í a zda jsou jeho promìnné relevantní.

2. Poèáteèní úvahy

Data zmiòovaná vý¹e (poèet støel, dr¾ení míèe. . .) uvádíme jako základní - popisná, kvatitativní, øeknou nám co a kolikrát se stalo, problém je, ¾e vzhledem k budoucnosti, pøí¹tím zápasùm, je jejich vyu¾ití omezené, èasto zkreslují situaci na høi¹ti, který tým byl lep¹í, kdo si zaslou¾il vyhrát. Kolikrát fanou¹ci po zápasech diskutují ve stylu,kdyby to tre l, vyhráli bychom\ nebo „sice nás pøehrávali, ale k ¾ádným vìt¹ím ¹ancím jsme je nepustili.\ Tato práce se vìnuje "pokroèilým"analytickým metodám, kromì kvantity vyu¾ívající i kvalitu, napøíklad kvalitu støelby, kvalitu samotných ¹ancí ke skórování nebo bodového zisku. Dal¹í výhodou je mo¾nost zpøesnìní jejich výsledkù pomocí parametrù vyskytujících se ve høe. Tyto pokroèilé, kvalitativní statistiky jsou èasto schovány za urèitou variací modelu "Oèekávané hodnoty"- Expected value model, který je modi kován speci cky pro ka¾dé sporotvní odvìtví. Zde je pozornost vìnována pøevá¾nì úvaze-tody Oèekávaných gólù\ { The Expected goals method, na základì které se nehodnotí èistì výsledky utkání a kvantitativní statistiky úèinkujících, ale výkony mu¾stev, jednotlivcù, díky nim¾ lze predikovat jejich pùsobení bìhem dal¹ích sportovních klání.

Expected value mù¾e být pro skauty, trenéry, hráèe a vlastnì i pro fanou¹ky velmi u¾iteèná, její popularizace bìhem posledních let a hlavnì popularizace z ní odvozených modelù pomohla leckterému laikovi nahlédnout pod poklièku jednotlivým sportùm a ukázala spojitost mezi øí¹í èísel a pohybu. Zároveò týmy, které tyto modely dlouhodobì vyu¾ívají prokazatelnì dosahují, pro nì døív stì¾í dosa¾itelných, posunù v podobì vylep¹eného skautingu hráèù, který mnohdy pøiná¹í vy¹í ekonomické zisky, výkonostních posunù a s nimi souvisejícími pøípadnými postupy do vy¹ích soutì¾í.

2.1. Statistické modely v jednotlivých sportech

Ka¾dý sport je speci cký svou krásou, pravidly, nicménì mezi jejich statistickými modely lze najít èasto nìjaké podobnosti, samozøejmì, èím si jsou sporty podobnìj¹í, tím podobnìj¹í si budou i metody modelování pravdìpodobnosti skórování, na dal¹ích pár øádcích je krátce pøedstaveno, jak jsou predikovány události prùbìhu utkání.
V Baseballu jsou èasto zmiòovány "Oèekávané dobìhy"- Run expectancy (RE). Výsledky RE ukazují pravdìpodobnost poètu dobìhù (získaných bodù týmu) do konce smìny v závislosti zejména na poètu útoèících hráèù na metách a poètu outù, pøípadnì na historické úspì¹nosti nadhazovaèe nebo pálkaøe.
Baseball jako takový je pova¾ován za jeden z prùkopnických sportù, co se analytických modelù týèe, první velké úspìchy jsou nyní ji¾ 20 let staré, mù¾eme je vidìt zdokumentované napøíklad ve snímku "Moneyball"(re¾ie Bennett Miller, 2011).
Mezi nejznámìj¹í modely vyu¾ívaných v americkém fotbale patøí "Expected points added"(EPA), vychází z aktuální pozice míèe na høi¹ti, ukazují kolik bodù obvykle tým z dané akce získá pøi uskuteènìní urèitého jevu, jako je napøíklad zisk území díky pøebìhnutí urèité vzdálenosti s míèem nebo zatlaèení soupeøe pressingem zpìt ve høi¹ti.
Hokejoví analytici pou¾ívají pøevá¾nì nám ji¾ známé "Expected Goals", zji¹»ují pravdìpodonost, ¾e støela èi ¹ance skonèí v síti soupeøe, zohledòují pøitom historickou bilanci støel z dané pozice, mno¾ství hráèù v dráze støely, herní situaci, jedná-li se o nájezd, oslabení, pøesilovku 5 na 4, 5 na 3, hru v plném poètu hráèù a dal¹í parametry. Problém je, kdy¾ se

na základì této statistiky sna¾í srovnávat jednotlivé hráèe, jeliko¾ jejich èasy strávené na ledì jsou rùzné, proto se zavádí "Expected goals per 60"- Oèekávané góly na 60 minut, kde se nesrovnalosti s èasem na ledì vyrovnají a je poté sna¾ní porovnat výkony jednotlivcù.

2.1.1. Expected goals method

„Metoda Oèekávaných gólù\ {The Expected goals method (dále xG) je prosazována i ve fotbalové bublinì, je to také dáno tím, ¾e branek v hokeji nebo fotbale nepadá tolik jako tøeba v házené, samotná událost vstøelení branky je v obou tìchto sportech naprosto raritní vìc, která má zcela zásadní vliv na celou hru. V jedné z nejpopulárnìj¹ích fotbalových lig svìta, anglické Premier League v sezónì 2020/2021 bylo prùmìrnì vstøeleno 2.69 gólu za 90minutové fotbalové utkání[6]. Divák tak prùmìrnì mezi dvìma góly èekal déle ne¾ pùlhodinu. Právì fotbalu a metodì "Expected goals"je v textu vìnována nejvìt¹í pozornost.

Základní princip této metody je pøitom jednoduchý, støelám, kvantitativní informaci, je pøiøazována hodnota xG, kvalitativní infromace. Vezmeme-li v¹echny støely jako poèet pokusù a góly jako úspì¹né pokusy tak díky informaci, ¾e za sezóny 11/12 a¾ 15/16 bylo ve skotské lize a èetyøech nejvy¹¹ích anglických fotbalových soutì¾ích v prùmìru potøeba 8,4427 støel na gól[8], bude xG ka¾dé støely rovno $\frac{1}{8,4427} = 0,1184$.

Nyní vezmìme pevnì danou pozici na hrací plo¹e, ze které hráè vystøelil, v databázích statistik sportovních utkání lze nalézt tisíce støel pøesnì ze stejného místa na høi¹ti a hodnota xG (pravdìpodobnost ¾e padne branka) bude rovna:

$$xG = \frac{\text{celkový poèet gólù z dané pozice}}{\text{celkový poèet støel z dané pozice}} \qquad (2.1)$$

Takto mù¾eme „ohodnotit\ ka¾dou ¹anci v zápase a zjistit, jaké xG oba týmy nasbírali { kolik gólù by prùmìrnì padlo ze ¹ancí vytvoøených bìhem utkání.

Samotné modelování tak jednoduché není, jeliko¾ je hledána pravdìpodobnost uskuteènìní urèitého jevu, padne-li, èi nepadne branka. Je mo¾né vyu¾ít machine learning modely, díky jejich pøesnosti èastìji vyu¾ívané v bì¾né praxi, nebo také klasi kaèní modely jako je model logistické regrese, která oproti pøesnìj¹ím modelùm poskytuje lep¹í interpretovatelnost a proto je zde pøevá¾nì vìnována pozornost jí. Chceme vyu¾ít co nejvíce informací, které máme k dispozici a ovìøit, zda jejich zahrnutí do modelu je relevantní pro získané výsledky, aby hodnoty koe cientù regresních parametrù byly co nejpøesnìj¹í. Mezi parametry, se kterými pracujeme, patøí kromì pozice na hrací plo¹e také pod jakým úhlem støelec vidí branku, zakonèuje-li slab¹í, silnìj¹í nohou, kolik hráèù stojí v potenciální dráze støely, jedná-li se o volej, hlavièku a dal¹í. Z principu plyne, ¾e pravdìpodobnost vstøelení gólu do odkryté brány je daleko vy¹¹í, ne¾ kdy¾ hráè vystøelí z 25 metrù pøes chumel protihráèù. Co xG neuva¾uje, je informace, kdo danou støelu vykonal, xG øíká kolikrát ze 100 pokusù by se za daných podmínek tre l bì¾ný profesionální fotbalista.

Uká¾eme si, jak mù¾e xG vypadat v praxi. Vezmeme zápas Sparta { Slavia s výsledkem 1:1, kde Sparta pøestøílí Slavii 9:5 a bude mít výraznì vy¹¹í procento dr¾ení míèe na svých kopaèkách. Na první pohled by se zdálo, ¾e leten¹tí byly lep¹ím týmem a remíza je

pro nì smolný výsledek. Støely spar»anských hráèù ov¹em byly z vìt¹ích vzdáleností od branky, a ne v¾dy z optimálních pozic, prùmìrnì mìla spar»anská støela xG = 0,1 (1 z 10 takovýchto støel rozvlní sí» soupeøe), Slavie si z protiútokù vytvoøila 5 støel s prùmìrným xG=0,3, tedy o nìco kvalitnìj¹í ¹ance. Sparta tedy za zápas získala 9 0,1 = 0,9 xG, zatímco se¹ívaní 5 0,3 = 1,5 xG, vidíme, ¾e Slavie byla v utkání nebezpeènìj¹í a mìla k výhøe blí¾e ne¾ její soupeø.

Pomocí xG lze sledovat jednotlivé hráèe, mù¾eme mít hráèe kteøí dávají hromadu branek, ov¹em jejich xG je daleko ni¾¹í, z toho usuzujeme, ¾e hráè má skvìlou formu nebo prostì jen ¹tìstí, ze statistického hlediska mu takové výkony, kdy pøekraèuje své xG, dlouhodobì nevydr¾í, jeliko¾ bìhem del¹ího èasového horizontu své xG pøekonává pouze jeden hráè na svìtì { Lionel Messi[4]. Naopak mù¾eme najít útoèníky, kteøí dávají výraznì ménì branek, ne¾ je jejich xG, to pak mù¾e kluby odradit od jejich anga¾ování, pøípadnì je donutit zapracovat na zakonèování hráèù. Dále bývá xG vyu¾íváno pro skautování skrytých talentù, èi pro pøedpovìï dal¹ích zápasù, co¾ je jeho veliká výhoda oproti èistì popisným statistikám.

3. Potøebný matematický aparát

V této kapitole jsou uvedeny matematické nástroje vyu¾íváné posléze pro vytvoøení modelu. Vychází ze zdrojù [1] [2] [3].

3.1. Zobecnìný lineární model

Základní lineární regresní modely pracují se závislými promìnnými s normálním rozdìlením, aby bylo mo¾né modelovat závislé promìnné, které mají rozdìlení jiné ne¾ normální, zavádíme Zobecnìné lineární modely (angl. Generalized linear models - GLMs). Zobecnìné lineární modely obsahují tøi èásti, náhodnou, systematickou èást a link function .

Náhodná èást Zobecnìného lineárního modelu obsahuje závislou promìnnou Y s nezávislými pozorováními $(y_1,\ldots,y_N)$, s hustotou pravdìpodobnosti:

$$f(y_i;\ _i) = a(\ _i)b(y_i)\exp[y_i Q(\ _i)] \qquad (3.1)$$

Hodnota parametru $_i$ je funkcí vysvìtlujících velièin pro ka¾dé $i = 1,\ldots, N$. Parametr Q_i oznaèujeme jako pøirozený parametr.

V systematické èásti je vyvìtlující promìnné pomocí lineárního modelu pøiøazen vektor $(\ _1;:::;\ _N)$. Nech» x_{ij} oznaèuje hodnotu vysvìtlující promìnné j $(j = 0;1;2;:::)$ pro subjekt i. Pak

$$_i = \sum_j\ _j x_{ij}\ ; \qquad i = 1;:::;N: \qquad (3.2)$$

Tato lineární kombinace vysvìtlujících promìnných se nazývá lineární prediktor. Pro koecient konstantního èlenu lineárního modelu $_0$ mimo vyjímky pro v¹echna i platí $x_{i0} = 1$. Aby byl model kompletní je nutné první dvì èásti propojit. Nech» $_i = E(Y_i); i = 1;:::;N$. Zavádíme monotónní diferencovatelnou funkci g, platí $_i = g(\ _i)$, kde g nazýváme link function. Díky g existuje vztah mezi $_i$ a vysvìtlujícími promìnnými

$$g(\ _i) = \sum_j\ _j x_{ij}\ ; \qquad i = 1;:::;N: \qquad (3.3)$$

3.1.1. Binomické Logit Modely

Uva¾ujme, ¾e chceme modelovat, zda urèitá událost nastane, takovéto pøípady pøedstavují binární závislé promìnné. Úspìch, uskuteènìní jevu a neúspìch, jeho neuskuteènìní reprezentujeme jako 1 a 0. Pøi binomickém rozdìlení pro závislou promìnnou Y platí: $P(Y = 1) =$, $P(Y = 0) = 1$ a $E(Y) =$. Jedná se o speciální pøípad binomického rozdìlení s n=1 s hustotu pravdìpodobnosti

$$f(y;\ ) = \ ^y(1\ )^{1\ y} = (1\ )\ \left(\frac{}{1\ }\right)^y = (1\ y)\exp\ y\ \log\frac{}{1\ } \qquad (3.4)$$

pro y = 0 nebo 1. Je zøejmé, ¾e se jedná o Zobecnìný lineární model s hustou pravdìpodobnosti (3.1). Parametru pøíslu¹í , $a(\ ) = 1$, $b(y) = 1$ a $Q(\ ) = \log[\ =(1\ )]$,

pøirozený parametrlog[= (1)] vyjadøuje log odds ("zlogaritmovaná ¹ance"), ¹ance, ¾e výstup modelu bude 1, dále v textu se setkáváme s oznaèenímlogit funkce. Modely vyu¾ívajícílogit jako¾tolink function jsou popsány v kapitole 3.4, vìnované logistické regresi a jejím modelùm, nebolilogit modelùm.

3.2. Zobecnìný lineární model pro binární data

Nech» Y je binární závislá promìnná, udávající "uskuteènìní"nebo "neuskuteènìní"události, ka¾dé pozorování uká¾e jeden z tìchto výsledkù, oznaèené 1 a 0. Støední hodnota Y je $E(Y) = P(Y = 1)$. Aby byla zøejmá závislost Y na hodnotách $x = (x_1; :::; x_p)$ oznaèíme $P(Y = 1)$ jako (x). Rozptyl Y je

$$var(Y) = (x)[1 (x)] \quad (3.5)$$

Jedná se o rozptyl binomického rozdìlení pøi = 1:

3.2.1. Lineární Pravdìpodobnostní Model

Regresní model binární závislé promìnné

$$(x) = + {}_1x_1 + ::: + {}_px_p \quad (3.6)$$

je nazýváníineární pravdìpodobnostní model.Lineární funkce pokrývají svými hodnotami celou reálnou osu, ov¹em pravdìpodnost nabývá hodnot v mezích pouze od 0 do 1. V modelu (3.6) mù¾e pro nìkteré hodnoty x nastat $(x) < 0$ nebo $(x) > 1$, ov¹em model mù¾e platit na omezené oblasti, pokud taková situace opravdu nastane je vyu¾ívána snadná iterpretace parametrù $_j$, které vyjadøují zmìnu (x) pøi jednotkové zmìnì x_j.

3.3. Metoda maximální vìrohodnosti

3.3.1. Pomocná tvrzení

De nice 3.1 Nech» f je reálná funkce de novaná na otevøeném intervalu I = (a,b), kde $1 < a < b < 1$. Funkce f se nazývákonvexní, platí-li

$$f[x + (1)y] f(x) + (1)f(y) \quad (3.7)$$

pro v¹echna x,y, taková, ¾e $a < x < y < b$ a $0 < < 1$.
Funkce f se nazývástriktnì konvexní, platí-li pro v¹echna uvedená x; y; v (3.7) ostrá nerovnost.
Funkce f se nazývá(striktnì) konkávní, je-li f (striktnì) konvexní.

Vìta 3.2 Nech» f je de nována na I a nech» je tam spojitá. Nech» existuje na I spojitá f^0 a nech» f^{00} existuje a je koneèná na I. Pak funkce f je konvexní právì tehdy, platí-li $f^{00} 0$ pro v¹echna x2 I:

Dùkaz. Viz Fichtengolc I (1958), odst. 143, str. 299, vìta 2.

Vìta 3.3 (Jensenova nerovnost) Nechf je konvexní funkce de novaná na I. NechX je náhodná velièina s koneènou støední hodnotou taková,P(X 2 I) = 1. Pak platí

$$f(EX) \quad Ef(X): \tag{3.8}$$

Je-li f striktnì konvexní, pak nerovnost(3.8) je ostrá s výjimkou pøípadu, kdy velièina X je rovna konstantì s pravdìpodobností 1. Nerovnost(3.8) se nazývá Jensenova.

Dùkaz. Viz Lehmann (1991), str. 50, vìta 6.3, nebo Rao (1978), kap. 1e.5.

3.3.2. Vìrohodnostní funkce a maximálnì vìrohodné odhady

Uva¾ujme náhodný vektorX = (X 1; : : : ; X n)0 se sdru¾enou hustotoufu(x,), kde 2 , kde je parametriký prostor parametrù : Nechx má pevnì danou hodnotu, pak se funkcef (x;) jako¾to funkce nazývá vìrohodnostní funkce
Hodnota ^ parametru , která maximalizujeme pro danéX = x vìrohodnostní funkci p(x;), se nazývámaximálnì vìrohodný odhad parametru :

Nech» X je náhodný vektor se sdru¾enou hustotoufu(x;), kde 2 ! Rm. Uva¾ujme funkciu : ! , která zobrazuje na Rk. Pøedpisem = u( ) ka¾dému 2 pøiøadíme 2 . Nech»G( ) = : 2 ; u( ) = . Oznaème

$$M(x; \quad) = \sup_{2G(\ )} f(x; \ )$$

M jako¾to funkci nazýváme vìrohodnostní funkcí indukovanou parametrickou funkcí u. Hodnotu ^ , která maximalizuje M(X ;) oznaèujeme jako¾to maximálnì vìrohodný odhad parametrické funkce u.

Nech» je jednorozmìrný parametr a platí následující pøedpoklady.

P1 : Nech» je parametrický prostor, který obsahuje takový neprázdný otevøený interval ! , ¾e skuteèná hodnota parametru 0 patøí do! .

P2 : Nech»X = (X 1; : : : ; X 2)0, kde X i jsou stejnì rozdìlené, nezávislé velièiny s hustotou f (x;) vzhledem k nìjaké -koneèné míøe.

P3 : Nech»M = x : f (x;) > 0 nezávisí na .

P4 : Nech» 1; 2 2 . Pak f (x; 1) = f (x; 2) skoro v¹ude (vzhledem k míøe) právì tehdy, je-li 1 = 2.

Tedy vzhledem k míøe = je sdru¾ená hustota náhodného vektoruX
f (x;) = f (x1;) f (xn;).

Vìta 3.4 Jestli¾e n ! 1 , pak pro ka¾dé takové pevné2 , ¾e 6 0, platí

$$P_0\{f(X; \ 0) > f(X; \ )\} ! \ 1 \tag{3.9}$$

Dùkaz. Viz [3] kpt. 7.6, str. 149.

Pokraèujme v úvaze, ¾eje jednorozmìrný parametr. Uva¾ujme funkci promìnné $f(x;\ )$ pro pevné x. Funkce

$$L(x;\ ) = \ln f(x;\ ) \tag{3.10}$$

se nazývá logaritmická vìrohodnostní funkce. Èasto $L(x;\ )$ znaèíme jen jako $L(\ )$.

Vìta 3.5 Nech» jsou splnìny pøedpoklady P_1 – P_4. Nech» na intervalu I existuje $f'(x;\ ) = \frac{\partial f(x;\ )}{\partial}$ pro skoro v¹echna x. Pak pro ka¾dé > 0 pøi $n \to 1$ platí, ¾e s pravdìpodobností konvergující k jedné má vìrohodnostní rovnice

$$\frac{\partial L(X;\ )}{\partial} = 0 \tag{3.11}$$

takový koøen $\hat{}_n = \hat{}_n(X)$, ¾e $|\hat{}_n - {}_0| < $, kde ${}_0$ je skuteèná hodnota parametru, v ní¾ $p(x;\ )$ nabývá svého maxima.

Dùkaz. Nech» > 0 je tak malé, ¾e $[{}_0 - ;\ {}_0 +] \subset$. De nujme

$$S_n = \{x : L(x;\ {}_0) > L(x;\ {}_0 -)\ a\ L(x;\ {}_0) > L(x;\ {}_0 +)\}. \tag{3.12}$$

Dle (3.9) platí $P_{\ {}_0}(X \in S_n) \to 1$. Pro ka¾dé $X \in S_n$ tedy s pravdìpodobností blí¾ící se jedné existuje $\hat{}_n$ takové, ¾e ${}_0 - < \hat{}_n < {}_0 +$ a ¾e funkce $L(X;\ )$ má lokální maximum v bodì $= \hat{}_n$. Pak $L'(X;\ \hat{}_n) = 0$.

3.3.3. Vìrohodnostní funkce velièiny s binomickým rozdìlením

Uva¾ujme náhodnou velièinu $X \sim Bi(n;\ ); 0 < < 1$, pro kterou platí

$$f(x) = P_p(X = x) = \binom{n}{x}\ ^x(1 -)^{n-x}. \tag{3.13}$$

Hledejme maximálnì vìrohodný odhad , tedy pro X=x jde o maximalizaci funkce $g(\ ) = \ ^x(1 -)^{n-x}$. Nech» $x \in \{0; \ldots; n\}$. Pak

$$g'(\ ) = [x(1 -) - (n - x)\ ]\ ^{x-1}(1 -)^{n-x-1} \tag{3.14}$$

nulovým bodem funkce $g'(\ )$ je $\hat{} = \frac{x}{n}$. Nyní si uká¾eme, ¾e v tomto bodì skuteènì funkce $g'(\ )$ nabývá svého maxima. De nujme

$$z_i = \begin{cases} \frac{n}{x}; & i = 1; \ldots; x; \\ \frac{n(1 -)}{n - x}; & i = x + 1; \ldots; n. \end{cases} \tag{3.15}$$

Aritmetický prùmìr tìchto èísel je

$$z = \frac{1}{n}\ x\frac{n}{x} + (n - x)\frac{n(1 -)}{n - x} = 1; \tag{3.16}$$

geometrický prùmìr je

$$z_G = \left(\frac{n}{x}^{x} \frac{n(1 \quad)}{n \quad x}^{n \quad x} \right)^{\frac{1}{n}} \tag{3.17}$$

z nerovnosti z_G z zjistíme, ¾e

$$^{x}(1 \quad)^{n \quad x} \quad \frac{x}{n}^{x} \frac{n \quad x}{n}^{n \quad x} \tag{3.18}$$

zjevnì pøi $= \hat{} = \frac{x}{n}$ nastává rovnost a tedy$\hat{}$ maximalizuje funkci $g( )$:
Nastane-li jeden z pøípadù $x = 0$ nebo $x = n$, pak maximálnì vìrohodný odhad parametru na inetrvalu $(0; 1)$ neexistuje, nastal by spor s pøedpokladem P_1.

3.4. Logistická regrese

Pro práci s binárními závislými promìnnými, je vhodný nástroj logistická regrese, která je ¹iroce vyu¾ívána napøíè obory, a» u¾ se jedná o medicínu, ekonomii nebo právì sport.

3.4.1. Interpretace parametrù

Pøedpokládejme binární závislou prmìnnou Y a a promìnnou X, na které je Y závislá, nech» $(x) = P(Y = 1 j X = x) = 1 \quad P(Y = 0 j X = x)$. Model Logistické regrese je

$$(x) = \frac{\exp(+ x)}{1 + \exp(+ x)}: \tag{3.19}$$

Je tedy zjevné, ¾e logit je lineární

$$\text{logit}[(x)] = \log \frac{(x)}{1 \quad (x)} = + x \tag{3.20}$$

Interpretace parametru

Nyní se zamìøíme na z rovnice (3.20), znaménko tohoto parametru urèuje, zda (x) bude rostoucí èi klesající s rostoucím x. Sklon logistické køivky roste s nárùstem j j, pro $! \quad 0$ se logistická køivka pøibli¾uje k horizontální rovné pøímce. V pøípadì nezávislosti Y na X platí $= 0$.
Aplikací exponenciály na obì strany rovnice (3.20) ukazujeme, ¾e "ance"(odds) jsou exponenciální funkcí x, odds se zvì t¹íe -krát s ka¾dým jednotkovým navý¹ením promìnné x, e je pomìr odds pøi $X = x + 1$ a odds pøi $X = x$. Vzhledem k nelineární závislosti (x) na x je zmìna (x) pro rozdílná x v logistické regresi "nerovnomìrná".

3.4.2. Fitování modelù logistické regrese

Uva¾ujme n binárních nezávislých pozorování, nech» $x_i = (x_{i1}; : : : ; x_{ip})$ udává nastavení hodnot p vyvìtlujících promìnných, i=1,. . . ,N . Nabývají-li v¹echna pozorování rùzných hodnot, pak N=n . Model logistické regrese (3.19), uva¾ující regresní parametr jako konstantní, je

$$(x_i) = \frac{\exp(\sum_{j=1}^{p} {}_j x_{ij})}{1 + \exp(\sum_{j=1}^{p} {}_j x_{ij})} \tag{3.21}$$

Vìrohodnostní rovnice

Kdy¾ více ne¾ jedno pozorování nastane pøi pevné hodnotì x_i, je vhodné zaznamenat poèet pozorování n_i a poèet úspì¹ných pokusù. Poté uva¾ujeme y_i jako ¾to poèet úspì¹ných pokusù, namísto odezvy jednotlivých binárních promìnných. Pak $Y_1; : : : ; Y_N$ jsou nezávislé binomické promìnné se støední hodnotou $E(Y_i) = n_i \ (x_i)$, kde $n_1 + : : : n_n = n$: Jejich marginální hustota pravdìpodobnosti je úmìrná souèinu N binomických funkcí

$$\begin{aligned} &\prod_{i=1}^{N} (x_i)^{y_i} [1 \quad (x_i)]^{n_i \quad y_i} = \\ &= \left(\prod_{i=1}^{N} \exp \log \frac{(x_i)}{1 \quad (x_i)} \right)^{y_i} \left(\prod_{i=1}^{N} [1 \quad (x_i)]^{n_i} \right) \\ &= \exp \left(\sum_i y_i \log \frac{(x_i)}{1 \quad (x_i)} \right) \left(\prod_{i=1}^{N} [1 \quad (x_i)]^{n_i} \right) : \end{aligned} \tag{3.22}$$

Pro model (3.21), jeho i-tý logit $\sum_j \ _j x_{ij}$, tak¾e exponenciální èlen v posledním výraze je roven $\exp[\sum_i y_i (\sum_j \ _j x_{ij})] = \exp[\sum_i (\sum_j y_i x_{ij}) \ _j]$. Jeliko¾ $[1 \quad (x_i)] = [1 + \exp(\sum_j \ _j x_{ij}]^{\ 1}$, tak vìrohodnostní funkce je rovna

$$L(\ ) = \sum_j \left(\sum_i y_i x_{ij} \right) \ _j \quad \sum_i n_i \log \left(1 + \exp \sum_j \ _j x_{ij} \right) : \tag{3.23}$$

Vìrohodnostní rovnice získáme, kdy¾ polo¾íme $@L(\ )=@ \ = 0$: Platí

$$\frac{@L(\ )}{@ \ } = \sum_i y_i x_{ij} \quad \sum_i n_i x_{ij} \frac{\exp(\sum_k \ _k x_{ik})}{1 + \exp(\sum_k \ _k x_{ik})}; \tag{3.24}$$

vìrohodnostní rovnice jsou

$$\sum_i y_i x_{ij} \quad \sum_i n_i \hat{} _i x_{ij} = 0; \qquad j = 1; : : : ; p; \tag{3.25}$$

kde $\hat{} _i = \exp(\sum_k \hat{} _k x_{ik})=[1 + \exp(\sum_k \hat{} _k x_{ik})]$ je maximální vìrohodný odhad (x_i): Tyto rovnice jsou nelineární a vy¾adují numerické øe¹ení. Pøi modelování je v programovacím jazyce Python bìhem vytváøení modelu vyu¾ívána metoda IRLS - Iteratively reweighted least squares [10].

3.4.3. Testování podmodelu

K testování podmodelu je vyu¾it test pomìrem vìrohodnosti (Likelyhood ratio test), který vyu¾ívá logaritmickou vìrohodnostní funkci, která pro binomické závislé promìnné vypadá následovnì:

$$\begin{aligned} L(\ ) &= \log \prod_{i=1}^{n} \ _i^{Y_i} (1 \quad _i)^{1 \ Y_i} \\ &= \sum_{i=1}^{n} (Y_i \log \ _i + (1 \quad Y_i) \log(1 \quad _i) \\ &= \sum_{i=1}^{n} Y_i \log \frac{i}{1 \quad _i} + \sum_{i=1}^{n} \log(1 \quad _i) \end{aligned} \tag{3.26}$$

3.4. LOGISTICKÁ REGRESE

Pozorované náhodné velièiny se v logaritmické vìrohodnostní funkci vyskytují v souèinech s výrazy $\log(\ _i=(1\ \ _i))$.

Uva¾ujme nejbohat¹í mo¾ný model, model s vìt¹í hodnotou vìrohodnostní funkce nelze vytvoøit, takový model se nazývásaturovaný. Saturovaný model má právì tolik parametrù, kolik je rùzných hodnot vektorù x_i: Maximální hodnotu vìrohodnostní funkce v saturovaném modelu oznaèíme L_{max}. Ka¾dý dal¹í model je podmodelem saturovaného modelu. Pomocídevianceposoudíme pøiléhavost bì¾ného modelu

$$D(b) = 2(L_{max} \quad L(b)). \tag{3.27}$$

Èím je model pøiléhavìj¹í, hodnota deviance D klesá. Dále pøedpokládáme, ¾e v¹echny vektory x_i jsou rùzné, pak má saturovaný model$k$ parametrù $\ _1;:::;\ _k$. Odhadem støední hodnoty $\ _i$ je Y_i, dle (3.26) platí

$$L_{max} = \sum_{i=1}^{k} (Y_i \log Y_i + (1 \quad Y_i)\log(1 \quad Y_i)) = 0 \tag{3.28}$$

Oznaème odhad pravdìpodobnosti jednièky$\hat{\ }_i = \ (x_i)$, dosazením do (3.28) vyjádøíme devianci v modelu logistické regrese jako

$$D(b) = \quad 2L(b) = \quad 2\sum_{i=1}^{k} (Y_i \log \hat{\ }_i + (1 \quad Y_i)\log(1 \quad \hat{\ }_i)) \tag{3.29}$$

Test pomìrem vìrohodnosti

Uva¾ujme modelM_1 s odhadem parametru$b = \hat{b}$ a jeho podmodelM_2, který vznikl odebrání èásti regresorù modeluM_1, s odhadem partametru$b = \tilde{b}$. Pøi testu pomìrem vìrohodnosti testujeme, zda v¹echny parametry obsa¾ené v modeluM_1 a zároveò vynechané v podmodeluM_2 jsou rovny nule, porovnáváme hodnoty logaritmické vìrohodnostní funkce pro$\hat{b}$ a$\tilde{b}$ pomocí statistiky

$$LR = 2 \quad L(\hat{b}) \quad L(\tilde{b}) \ : \tag{3.30}$$

LR statistiku lze také vyjádøit pomocí deviance modelu a podmodelu

$$2 \quad L(\hat{b}) \quad L(\tilde{b}) = 2(L_{max} \quad L(\tilde{b})) \quad 2(L_{max} \quad L(\hat{b})) = D(\tilde{b}) \quad D(\hat{b}): \tag{3.31}$$

Platí-li testovaný podmodel a jsou-li zároveò splnìny pomínky regularity[2], pak má statistika LR asymptoticky rozdìlení χ^2_q, kde q je rozdíl poètu nezávislých parametrù v modelech, které porovnáváme.

4. Modelování pravdìpodobnosti skórování

V této kapitole jsou uvedeny jednotlivé modely pro modelování pravdìpodnosti skórování ve fotbale, modelování Expected goals - xG. Data byla získána od rmy Statsbomb, která je volnì distribuuje, jejich následné zpracování probìhlo v programovacím jazyce Python, s vyu¾itím knihoven jako pandas, numpy, statsmodels nebo FCPython. Zpracováno bylo 568 zápasù, ve kterých dohromady padlo 14775 støel, branka padla v 1753 pøípadech.
Postupnì budou uvádìny modely od nejjednodu11ího po modely komplexnìj1í, které pak lze otestovat proti jednodu11ím modelùm, a zjistit tak relevanci jejich parametrù.

4.1. Triválnì modely

Chceme.li znát pravdìpodobnost, ¾e ze støely padne branka, tak se samozøejmì jako nejjednodu11í úvaha nabízí pomìr

$$xG = \frac{\text{celkový poèet gólù}}{\text{celkový poèet støel}}.$$

Výsledkem tohoto triviálního modelu, je ¾e ka¾dá støela má $xG = 0;1184$, jeliko¾ z principu víme, ¾e pravdìpodobnost padnutí branky ze 2 metrù je vìt1í, ne¾, ¾e hráè rozvlní sí» z poloviny høi1tì, je zjevné, ¾e daný model lze sice vyu¾ít, spoléhat na jeho jakoukoliv pøesnost je ov1em na pová¾enou.
Proto dále uva¾ujeme opìt pomìr $\frac{\text{celkový poèet gólù}}{\text{celkový poèet støel}}$, nicménì ji¾ jej poèítáme pro ka¾dý "ètverec"metr na metr na hrací plo1e. Celou hrací plochu, kterou uva¾ujeme délky 100 a 1iøky 70 metrù, jsme rozdìlili na sí» tìchto metrových ètvercù, pøièem¾ jejich hranice jsou celoèíselné. Tak¾e pøesnìji poèítáme $\frac{\text{celkový poèet gólù z jednoho ètverce na høi1ti}}{\text{celkový poèet støel z toho samého ètverce na høi1ti}}$. Tento model je vyobrazen na 4.3. Pro sna¾1í interpretaci byly na obrázcích 4.1 respektive 4.2, vyobrazeny v1echny støely a góly z pou¾ité databáze.

Obrázek 4.1:
lokace jednotlivých støel

Obrázek 4.2:
lokace v¹ech gólù

Na obrázku 4.3 ji¾ vidíme vykreslené jednoduché xG, tmavá oblast okolo brány potvrzuje, ¾e nejvy¹¹í pravdìpodobnost vstøelení branky je, kdy¾ jsme k ní co nejblí¾e, tmav¹í body u stran høi¹tì jsou zpùsobené malým poètem støel z tìchto pozic v databázi. Vzhledem k nerovnomìrnému poètu støel z jednotlivých èástí høi¹tì, je ov¹em tento model znaènì nepøesný.

Obrázek 4.3: jednoduché xG

4.2. Pokroèilé modely

Nyní ji¾ zaèneme vyu¾ívat matematické nástroje popsané v kapitole 3. Binární data na výstupu modelu, padne/nepadne branka, modelujeme pomocí logistické regrese, kdy budeme postupnì vytváøet modely s rùznými promìnnými, jejich¾ poèet budeme postupnì navy¹ovat a vytváøet tak modely komplexnìj¹í.

4.2.1. Základní modely logistické regrese

xG v závislosti na vzdálenosti

Zaèneme modelem jedné promìnnéX, bereme za ni vzdálenost od branky, nikoliv ov¹em od jejího støedu, ale kolmou vzdálenost od brankové èáry, takto uva¾ovanou vzdálenost znaèíme$X$ i v dal¹ích modelech. Program z dat získal následující hodnoty koe cientù a jejich intervalový odhad pøi hladinì významnosti $= 0:05$.

Tabulka 4.1: xG v závislosti na$X$

Promìnná	koe cinet	[0,025	0,975]
konstanta	0,7193	0,615	0,824
X	0,0924	0,085	0,100

Deviance modelu je 10057

Tak¾e funkce xG jako taková vypadala následovnì

$$xG = \frac{1}{1 + e^{0;7193+0;0924x}}; \tag{4.1}$$

mù¾eme ji vidìt vykreslenou ná následujícím obrázku.

Obrázek 4.4: xG v závislosti na vzdáloenosti

Z obrázku 4.4 je zøetelnì pozorovatelný nelineární vztah pravdìpodobnosti padnutí branky a støelcovì vzdálenosti od brankové èáry, co ov¹em model neuva¾uje je, kde pøesnì se hráè nachází, mù¾e být ve stejné vzdálenosti$X$ od brankové èáry, nicménì mù¾e stát nìkde u autové èáry nebo nìkde uprostøed, obìma tìmto pøípadùm pøiøadí model stejné xG,

proto kromì kolmé vzdálenostiX , pøidáme do modelu úhel, pod kterým hráè brankovou konstrukci vidí, viz 4.5.

Obrázek 4.5: Úhel, pod kterým støelec vidí branku

xG v závislosti na vzdálenosti a úhlu

Nyní ji¾ v modelu kromX uva¾ujeme i úhel popsaný vý¹e. Problematika stejných hodnot xG pro situace, kdy¾ je hráè v rohu høi¹tì a kdy¾ je tìsnì pøed brankou, tímto zaniká (jejich vzdálenostX od brankové èáry je v takovém pøípadì stejná). Pøi hladinì hladinì významonsti = 0:05 vy¹ly koe cienty promìnných a jejich intervalové odhady následovnì.

Tabulka 4.2: xG v závislosti naX a

Promìnná	koe cinet	[0,025	0,975]
konstanta	2,0586	1,857	2,260
X	0,0503	0,041	0,059
	-1,7412	-1,957	-1,525

Deviance modelu je 9778

Známenko koe cientu jednodu¹e interpretuje, zdali se s rostoucí promìnnou hodnota xG zvy¹uje, resp. sni¾uje. Je-li koe cient promìnné kladný, tak s rostoucí promìnnou xG klesá, lze vidìt u vzdálenostiX , èím dál je hráè od brankové èáry, tím bude men¹í pravdìpodobnost, ¾e skóruje. Z logiky vìci plyne, bude-li znaménko koe cientu promìnné mínus, bude s rostoucí promìnnou rùst zároveò hodnota xG. Tento stav pozorujeme u úhlu , pravdìpodobnost støelení branky roste s velikostí úhlu, pod kterým hráè vidí brankovou konstrukci. Nìkteøí trenéøi mláde¾e tento fakt vyu¾ívají pøi výchovì mladých fotbalových talentù, sna¾í se jim pøedat informaci, pokud mají dobrý výhled na branku

stojí za to akci zakonèit, jak si tuto informaci pøebírají hráèi je u¾ na nich.
Obrázky 4.6 a 4.7 ukazují rozlo¾ení pravdìpodobnosti skórování na hrací plo¹e, zlep¹ení oproti pøedchozím modelùm je ihned pozorovatelné.

Obrázek 4.6: xG v závislosti na X a

Následující obrázek pøevádí obr. 4.6 do prostoru, jedná se o graf logistické funkce promìnných z tabulky.

$$xG = \frac{1}{1 + e^{2;0586+0;0503X \quad 1;7412}} \tag{4.2}$$

Obrázek 4.7: xG v závislosti na X a

4.2.2. Pokroèilé modely logistické regrese

V pøedchozích modelech byly uva¾ovány pouze spojité promìnné, a» u¾ se jedná o vzdálenost èi úhel. V této èásti budeme model s promìnnými, X postupnì roz¹iøovat o

promìnné diskrétní, mù¾e se jednat napøíklad o zpùsob zakonèení a dal¹í. Zároveò budou modely testovány, jestli nám jejich roz¹íøení pøidá na jejich "kvalitì"nebo jestli pøidaná promìnná není pro modelování relevantní.

xG v závislosti na vzdálenosti, úhlu a "tlaku"

Zahrnování diskrétních promìnných zaène pøidáním promìnné, která speci kuje, zda støílející hráè byl bìhem zakonèení pod tlakem èi nikoliv. Mù¾eme ji speci kovat, jako¾to okolnosti, které mohly ovlivnit hráèe bìhem støelby. Podstupoval zároveò souboj s protihráèem? Byl protihráèem napadán, ¾e musel se støelou spìchat jinak by o míè pøi¹el nebo by ji¾ ke støele nemìl prostor? Tyto a podobné otázky jsou zahrnuté v na¹í promìnné pojmenované "tlak", jeliko¾ se jedná o diskrétní promìnnou nabývá hodnot 1 - hráè byl pøi støele pod tlakem a 0 - hráè nebyl pøi støele pod tlakem, mìl na ni dostatek èasu a prostoru.

Tabulka 4.3: xG v závislosti na$X$, a tlaku

Promìnná	koe cinet	[0,025	0,975]
konstanta	1,9392	1,736	2,142
X	0,0528	0,044	0,062
	-1,7850	-2,003	-1,567
tlak	0,6291	0,479	0,779

Deviance modelu je 9702,5

Vidíme, ¾e pokud je hráè pod tlakem, ¹ance na skórování se sni¾uje.
Nyní pomocí testu pomìrem vìrohodnosti zjistíme, zda je pøidaná promìnná, tedy tlak, pro model významná. Porovnáme pøedchozí model $_{n1}$: xG = f(X,) - model "ni¾¹í"a poslední "rozsáhlej¹í"model $_{v1}$, xG = f(X, , tlak) - model "vy¹¹í".
Pro oba modely pomocí programu získáme logaritmickou vìrohodnostní funkci $L(\ )$, bude urèena hodnota ^{2} testu pomìrem vìrohodnosti, kterým je porovnáváme a výsledná p-hodnota bude porovnána s hladinou vìrohodnosti = 0.05.

$$L(\ _{v1}) = 4851{,}257934716189$$
$$L(\ _{n1}) = 4888{,}9960593471619$$
výsledek ^{2} testu: 75,47624959132008
$$\text{p-hodnota} = 4{,}0788773255253521 \cdot 10^{-17}$$

Vidíme, p-hodnota < , odebráním promìnné tlaku, se sní¾í kvalita modelu, tato promìnná je pro model významná a proto budeme dále pracovat s modelem xG = f(X, , tlak).

xG v závislosti na vzdálenosti, úhlu, tlaku a zpùsobu zakonèení

Nyní zahrneme zpùsob zakonèení, jako¾to promìnné modelu. Uva¾ujeme zakonèení levou (lf) a pravou nohou (rf), tøetí zpùsob zakonèení, hlavou je zahrnut v konstantním èlenu.

Tabulka 4.4: xG v závislosti naX , , tlaku a zpùsobu zakonèení

Promìnná	koe cinet	[0.025	0.975]
konstanta	3,1537	2,877	3,431
X	0,0600	0,051	0,069
	-2,3270	-2,569	-2,085
tlak	0,2884	0,133	0,444
levá noha (lf)	-1,2302	-1,418	-1,042
pravá noha (rf)	-1,2242	-1,403	-1,046

Deviance modelu je 9477,2

Koe cienty u zakonèení levou èi pravou jsou záporné a podobné. Jeliko¾ v modelu neuva¾ujeme kdo støílel, mají pro nás pravá i levá noha stejnou váhu, záporné znaménko souvisí s poslední mo¾ností zakonèení a to hlavou, pravdìpoodobnost gólu vstøeleným nohou je vy¹íí oproti zakonèení hlavou, proto po dosazení "jednièky"za jednu z promìnných lf, rf , co¾ se v realitì rovná zakonèení spodní konèetinou, zvý¹íme hodnotu xG støely. Test pomìrem vìrohodnosti vy¹el následovnì, za "vy¹íí"model $_{v2}$ pova¾ujemexG = f(X, , tlak, rf, lf), za "ni¾¹í"model $_{n2}$ beremexG = f(X, , tlak).

$$L(_{v2}) = 4738{,}609961574685$$
$$L(_{n2}) = 4851{,}257934716189$$

výsledek ^{2} testu: 225,295946

p-hodnota= $1{:}1956577\ 10^{-49}$

Opìt uva¾ujeme hladinu významnosti =0,05, p-hodnota< . Odebrání promìnnýchlf, rf znaènì ovlivnilo model, jsou pro nìj významné. Tudí¾ dále budeme pracovat s modelem xG = f(X, , tlak, rf, lf) .

xG v závislosti na vzdálenosti, úhlu, tlaku, zpùsobu zakonèení a stranì høi¹tì, ze které bylo zakonèeno

Nyní budeme zkoumat, zdali ovlivníme model, budeme-li jako¾to promìnnou uva¾ovat stranu, ze které hráè zakonèil- pravá, levá strana høi¹tì. Jako¾to promìnnou bereme napøíklad pravou stranu (rs), kterou stranu vezmeme model neovlivní, druhá strana se promítne do konstantního èlenu. Program vypoèítal koe cienty a jejich intervalové odhady na hladinì významnosti $= 0{;}05$ následovnì:

Tabulka 4.5: xG v závislosti naX , , tlaku, zpùsobu zskonèení a stranì høi¹tì, ze které se støílelo

Promìnná	koe cinet	[0,025	0,975]
konstanta	3,1980	2,915	3,481
X	0,0601	0,051	0,069
	-2,3232	-2,565	-2,081
tlak	0,2899	0,135	0,445
levá noha (lf)	-1,2236	-1,412	-1,036
pravá noha (rf)	-1,2295	-1,408	-1,051
pravá strana (rs)	-0,0866	-0,194	0,021

Deviance modelu je 9474,7

4.2. POKROÈILÉ MODELY

Testem pomìrem vìrohodnosti nyní bude zkontrolován vliv strany høi¹tì na pravdìpodobnost skàování. "Vy¹¹í model" $_{v3}$ - xG=f(X, , tlak, rf, lf, rs) , porovnáme s modelem "ni¾¹ím" $_{n3}$ - xG=f(X, , tlak, rf, lf).

$$L(\ _{v3}) = 4737{,}364539455062$$
$$L(\ _{n3}) = 4738{,}609961574685$$
$$\text{výsledek } ^2 \text{ testu: } 2{,}490844239246144$$
$$\text{p-hodnota= } 0{,}28781938827292586$$

Nyní ji¾ p-hodnota pøekroèila hladinu významnosti = 0,05, vliv strany, ze které bylo zakonèeno je pøi této hladinì zanedbatelný, dostateènì pøesné výsledky poskytuje "ni¾¹í model" $_{n3}$ - xG=f(X, , tlak, rf, lf). Pravdìpodonost skàování, xG bude tedy poèítáno v závislosti na vzdálenosti, úhlu, tlaku a zpùsobu zakonèení.

Tabulka 4.6: xG v závislosti naX , , tlaku a zpùsobu zakonèení

Promìnná	koe cinet	[0.025	0.975]
konstanta	3,1537	2,877	3,431
X	0,0600	0,051	0,069
	-2,3270	-2,569	-2,085
tlak	0,2884	0,133	0,444
levá noha (lf)	-1,2302	-1,418	-1,042
pravá noha (rf)	-1,2242	-1,403	-1,046

Deviance modelu je 9477,2

Samotná funkcexG s danými koe cienty z tabulky vypadáa následovnì:

$$xG = \frac{1}{1 + e^{3{,}1537+0{,}06X \quad 2{,}3270 \ +0{,}2884\,\text{tlak} \quad 1{,}2302\,\text{lf} \quad 1{,}2242\,\text{rf}}} \tag{4.3}$$

Po testování jednotlivých modelù, zùstaly tyto promìnné jako¾to pro model relevantní, tohle tvrzení lze podpoøit i shlédnutím jejich intervalových odhadù, kde ¾ádný neobsahuje nulu.
Následující obrázky 4.8 ukazují nìkteré pøípady výsledného modelu. Nejpatrnìj¹í je zmìna pøi zakonèování hlavièkou, kde je ¹ance skórování ni¾¹í, co¾ je na grafu vidìt jeho "zplo¹tìním"oproti dvoum grafùm nad ním. Zároveò je patrný vliv "tlaku"na hráèe, kdy graf potvrzuje, ¾e hráè který není pod tlakem má z dané pozice vy¹¹í ¹anci na vstøelení branky.

Obrázek 4.8: xG pro rùzné varianty výsledného modelu.

Obrázek 4.9: Porovnání dvou situací z obrázku 4.8

.

Na obrázku 4.9 jsou vyobrazeny situace, kdy model nabývá nejvy¹¹ích hodnot, pravdìpodobnost padnutí gólu pøi støelbì levou nohou, kdy¾ hráè je vklidu - modrá plocha, a situace, kby nabývá hodnot nejni¾¹ích, pravdìpodobnost padnutí branky pøi zakonèení hlavou pod tlakem - zelená plocha. Rozdíl, jako¾to vý¹e zmínìné zplo¹tìní grafu je u zelené plochy oproti plo¹e modré na první pohled zjevný. Graf dokazuje, ¾e zakonèení nohou v klidu má vy¹¹í pravdìpodonost, ¾e skonèí brankou, oproti zakonèení hlavou pod tlakem, uva¾ujeme-li zakonèení ze stejné pozice na høi¹ti. Zelená plocha nabývá v ka¾dém bodì hrací plochy (souøadnice X,Y) ni¾¹í hodnoty ne¾ plocha modrá.

5. Závìr

Práce popisuje problematiku pravdìpodobnosti skórování ve sportu, pøedstavuje jednotlivé metody pou¾ívaných modelù ve sportech, zvlá¹tì podrobnì pak ve fotbale. Zavádí nejznámìj¹í metodu "oèekávaných gólù", pøedstaví její výhody i nevýhody. Problematika je následnì popsána ze statistického hlediska, zavedením zobecnìných lineárních modelù, pøesnìji pak modelù logistické regrese, které jsou k interpretaci binomických závislých promìnných, jako je výskyt nebo nevýskyt urèitého jevu, zde padnutí branky, vhodné. Pøedstavena je metoda maximální vìrohodnosti, která je vyu¾ita pøi výpoètu koe cientù jednotlivých promìnných modelu. Jednotlivé modely jsou poté charakterizovány pomocí jejich deviance a porovnány na základì vìrohodnostních fukncí.
Pøedstavený matematický aparát je následnì aplikován v programovacím jazyce Python, kde z open source dat rmy Statsbomb [7] byly získány koe cienty jednotlivých spojitých i diskrétných vysvìtlujících promìnných, pomocí kterých byly následnì sestaveny modely pravdìpodobnosti skórování v profesionálním fotbalu. Vypoètené modely s rùznými poèty vysvìtlujících promìnných byly následnì pomocí testu vìrohodnostním pomìrem porovnávány. Bylo zji¹tìno, které vysvìtlující promìnné mají statisticky významný vliv na pravdìpodobnost vstøelení branky. Po testování nìkolika modelù dochází práce k závìru, ¾e z dostupných dat pravdìpodobnost skórování ve fotbale nejvhodnìji predikuje model vyu¾ívající promìnné: zpùsob zakonèení, herní situace a poloha na hrací plo¹e. Test vìrohodnostním pomìrem ukázal, ¾e v profesionálním fotbale nehraje roli, jestlo hráè zakonèuje z levé nebo pravé strany hrací plochy.

Literatura

[1] AGRESTI, Alan. Categorical data analysis. 2nd ed. Hoboken: Wiley, 2002, xv, 710 s. : grafy, tab. ISBN 0-471-36093-7.

[2] ZVÁRA, Karel. Regrese. Praha: Matfyzpress, 2008, 253 s. : il. ; 24 cm. ISBN 978-80-7378-041-8.

[3] AND Ì L, Jiøí. Základy matematické statistiky. Vyd. 3. Praha: Matfyzpress, 2011, 358 s. : grafy, tab. ISBN 978-80-7378-162-0.

[4] TIPPETT, James. The Expected Goals Philosophy. 1. Velká Británie: vydáno nezávisle, 2019. ISBN 978-1-08988-318-0.

[5] Friends of Tracking, https://www.youtube.com/channel/UCUBFJYcag8j2rm$_9$*HkrrA*7*w*

[6] Premier League o cial website, https://www.premierleague.com/

[7] Stats Bomb soccer data, https://statsbomb.com/what-we-do/soccer-data/

[8] http://allanderek.github.io/football-analysis/posts/shots-per-goal/

[9] Soccer analytics handbook https://github.com/devinpleuler/analytics-handbook

[10] https://www.statsmodels.org/