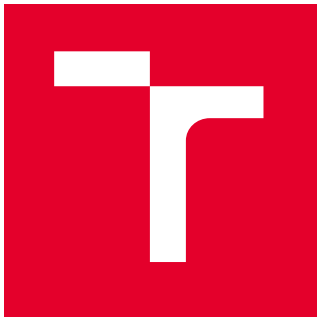


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

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 zaměřena na modelování pravděpodobnosti skórování ve fotbale. V práci je popsán nutný matematický aparát potřebný k sestavení modelu logistické regrese a základní testy statistických hypotéz. Popsaný matematický aparát je poté aplikován na volně přístupná data z profesionálních fotbalových utkání. Výsledný model používá vysvětlující proměnné jako způsob zakončení, polohu na hřišti a zjednodušeně 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 simplified match situation as predictors.

Klíčová slova
logistická regrese, GLM, metoda maximální věrohodnosti, Expected goals

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

ONDŘEJ HILSCHER *Modelování pravděpodobnosti skórování ve sportu.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2022. 23 s. Vedoucí diplomové práce Ing. Pavel Hrabec, Ph.D.

Prohlašuji, že jsem bakalářskou práci *Modelování pravděpodobnosti skórování ve sportu* napsal samostatně pod vedením Ing. Pavla Hrabce Ph.D. a Ing. Martina Roseckého, všechny použité materiály jsou uvedeny v seznamu literatury.

Ondřej Hilscher

Děkuji všem, kteří mne během vypracování podporovali a obohacovali mne svými zkušenými a praktickými radami, zvláště Ing. Pavlu Hrabcovi Ph.D. a Ing. Martinu Roseckému, kteří pomohli vytvořit skvělou atmosféru pro psaní práce, velké dík patří také mé rodině, jejíž podpora mě provází celé studium.

Ondřej Hilscher

Obsah

1. Sport a statistika

Pro mnoho lidí byly matematika a sport dlouhou dobu dva odlišné světy, a pro značnou část světové populace tomu tak stále je, ale aniž bychom si to uvědomovali, matematika, konkrétně statistika a statistické modely mohou mít a někde již mají zásadní vliv. Principy měření délky hodu, času, sčítání bodů i určování pořadí jsou tisíce let staré. Samozřejmě byly využívány jiné metody i jednotky, nicméně je vhodné zde uvést tvrzení, že matematika sport provází již od jeho raných kořenů. Dnešní pohled na matematiku, statistiku a její modely ve sportu se začal utvářet v druhé polovině minulého století, ať už to byli nadšenci z řad fanoušků, trenéři, či samotní sportovci, začaly se uchovávat informace o turnajích, utkáních nebo závodech, bez kterých si dnes nedokážeme představit žádný televizní přenos. U každé sportovní události je uvedena grafika s mnoha statistickými údaji - nejdelší hod, počet střel, es, úspěšných odpalů, držení míče, úspěšnost střelby, zákroků gólmana, největší vyvinutá rychlost během utkání, čas strávený na soupeřově polovině, třetině a takto by šlo vyjmenovat spoustu dalších. Jistě si takovou tabulku každý, kdo viděl nějakou sportovní akci v televizi, vybaví.

Tato práce čtenáře uvede do prostřední datové analytiky sportu, ukáže, jak může práce se sportovními daty (zde využívány data firmy Stats Bomb[7]) vypadat v praxi a představí jejich využití v oblasti sportu, ať už z hlediska událostí uplynulých nebo těch budoucích. K tomu jsou využívány matematické resp. statistické nástroje, které jsou posléze aplikovány při vytváření predikčních modelů pravděpodobnosti skórování ve sportu. Modely jsou následně testovány, tak aby bylo zjevné, který model by bylo vhodné použít, aby výsledky byly co nejpřesnější a zda jsou jeho proměnné relevantní.

2. Počáteční úvahy

Data zmiňovaná výše (počet střel, držení míče...) uvádíme jako základní - popisná, kvatitativní, řeknou nám co a kolikrát se stalo, problém je, že vzhledem k budoucnosti, příštím zápasům, je jejich využití omezené, často zkreslují situaci na hřišti, který tým byl lepší, kdo si zasloužil vyhrát. Kolikrát fanoušci po zápasech diskutují ve stylu „kdyby to trefil, vyhráli bychom“ nebo „sice nás přehrávali, ale k žádným větším šancím jsme je nepustili.“ Tato práce se věnuje "pokročilým"analytickým metodám, kromě kvantity využívající i kvalitu, například kvalitu střelby, kvalitu samotných šancí ke skórování nebo bodového zisku. Další výhodou je možnost zpřesnění jejich výsledků pomocí parametrů vyskytujících se ve hře. Tyto pokročilé, kvalitativní statistiky jsou často schovány za určitou variací modelu "Očekávané hodnoty"- Expected value model, který je modifikován specificky pro každé sporotvní odvětví. Zde je pozornost věnována převážně úvaze „metody Očekávaných gólů“ – The Expected goals method, na základě které se nehodnotí čistě výsledky utkání a kvantitativní statistiky účinkujících, ale výkony mužstev, jednotlivců, díky nimž lze predikovat jejich působení během dalších sportovních klání.

Expected value může být pro skauty, trenéry, hráče a vlastně i pro fanoušky velmi užitečná, její popularizace během posledních let a hlavně popularizace z ní odvozených modelů pomohla leckterému laikovi nahlédnout pod pokličku jednotlivým sportům a ukázala spojitost mezi říší čísel a pohybu. Zároveň týmy, které tyto modely dlouhodobě využívají prokazatelně dosahují, pro ně dřív stěží dosažitelných, posunů v podobě vylepšeného skautingu hráčů, který mnohdy přináší vyšší ekonomické zisky, výkonostních posunů a s nimi souvisejícími případnými postupy do vyšších soutěží.

2.1. Statistické modely v jednotlivých sportech

Každý sport je specifický svou krásou, pravidly, nicméně mezi jejich statistickými modely lze najít často nějaké podobnosti, samozřejmě, čím si jsou sporty podobnější, tím podobnější si budou i metody modelování pravděpodobnosti skórování, na dalších pár řádcích je krátce představeno, jak jsou predikovány události průběhu utkání.
V Baseballu jsou často zmiňovány "Očekávané doběhy"- Run expectancy (RE). Výsledky RE ukazují pravděpodobnost počtu doběhů (získaných bodů týmu) do konce směny v závislosti zejména na počtu útočících hráčů na metách a počtu outů, případně na historické úspěšnosti nadhazovače nebo pálkaře.
Baseball jako takový je považován za jeden z průkopnických sportů, co se analytických modelů týče, první velké úspěchy jsou nyní již 20 let staré, můžeme je vidět zdokumentované například ve snímku "Moneyball"(režie Bennett Miller, 2011).
Mezi nejznámější modely využívaných v americkém fotbale patří "Expected points added"(EPA), vychází z aktuální pozice míče na hřišti, ukazují kolik bodů obvykle tým z dané akce získá při uskutečnění určitého jevu, jako je například zisk území díky přeběhnutí určité vzdálenosti s míčem nebo zatlačení soupeře pressingem zpět ve hřišti.
Hokejoví analytici používají převážně nám již známé "Expected Goals", zjišťují pravděpodonost, že střela či šance skončí v síti soupeře, zohledňují přitom historickou bilanci střel z dané pozice, množství hráčů v dráze střely, herní situaci, jedná-li se o nájezd, oslabení, přesilovku 5 na 4, 5 na 3, hru v plném počtu hráčů a další parametry. Problém je, když se

na základě této statistiky snaží srovnávat jednotlivé hráče, jelikož jejich časy strávené na ledě jsou různé, proto se zavádí "Expected goals per 60"- Očekávané góly na 60 minut, kde se nesrovnalosti s časem na ledě vyrovnají a je poté snažší porovnat výkony jednotlivců.

2.1.1. Expected goals method

„Metoda Očekávaných gólů" –The Expected goals method (dále xG) je prosazována i ve fotbalové bublině, je to také dáno tím, že branek v hokeji nebo fotbale nepadá tolik jako třeba v házené, samotná událost vstřelení branky je v obou těchto sportech naprosto raritní věc, která má zcela zásadní vliv na celou hru. V jedné z nejpopulárnějších fotbalových lig světa, anglické Premier League v sezóně 2020/2021 bylo průměrně vstřeleno 2.69 gólu za 90minutové fotbalové utkání[6]. Divák tak průměrně mezi dvěma góly čekal déle než půlhodinu. Právě fotbalu a metodě "Expected goals"je v textu věnována největší pozornost.
Základní princip této metody je přitom jednoduchý, střelám, kvantitativní informaci, je přiřazována hodnota xG, kvalitativní infromace. Vezmeme-li všechny střely jako počet pokusů a góly jako úspěšné pokusy tak díky informaci, že za sezóny 11/12 až 15/16 bylo ve skotské lize a čtyřech nejvyšších anglických fotbalových soutěžích v průměru potřeba 8,4427 střel na gól[8], bude xG každé střely rovno $\frac{1}{8{,}4427} = 0{,}1184$.
Nyní vezměme pevně danou pozici na hrací ploše, ze které hráč vystřelil, v databázích statistik sportovních utkání lze nalézt tisíce střel přesně ze stejného místa na hřišti a hodnota xG (pravděpodobnost že padne branka) bude rovna:

$$xG = \frac{\text{celkový počet gólů z dané pozice}}{\text{celkový počet střel z dané pozice}} \tag{2.1}$$

Takto můžeme „ohodnotit" každou šanci v zápase a zjistit, jaké xG oba týmy nasbírali – kolik gólů by průměrně padlo ze šancí vytvořených během utkání.

Samotné modelování tak jednoduché není, jelikož je hledána pravděpodobnost uskutečnění určitého jevu, padne-li, či nepadne branka. Je možné využít machine learning modely, díky jejich přesnosti častěji využívané v běžné praxi, nebo také klasifikační modely jako je model logistické regrese, která oproti přesnějším modelům poskytuje lepší interpretovatelnost a proto je zde převážně věnována pozornost jí. Chceme využít co nejvíce informací, které máme k dispozici a ověřit, zda jejich zahrnutí do modelu je relevantní pro získané výsledky, aby hodnoty koeficientů regresních parametrů byly co nejpřesnější. Mezi parametry, se kterými pracujeme, patří kromě pozice na hrací ploše také pod jakým úhlem střelec vidí branku, zakončuje-li slabší, silnější nohou, kolik hráčů stojí v potenciální dráze střely, jedná-li se o volej, hlavičku a další. Z principu plyne, že pravděpodobnost vstřelení gólu do odkryté brány je daleko vyšší, než když hráč vystřelí z 25 metrů přes chumel protihráčů. Co xG neuvažuje, je informace, kdo danou střelu vykonal, xG říká kolikrát ze 100 pokusů by se za daných podmínek trefil běžný profesionální fotbalista.

Ukážeme si, jak může xG vypadat v praxi. Vezmeme zápas Sparta – Slavia s výsledkem 1:1, kde Sparta přestřílí Slavii 9:5 a bude mít výrazně vyšší procento držení míče na svých kopačkách. Na první pohled by se zdálo, že letenští byly lepším týmem a remíza je

pro ně smolný výsledek. Střely sparťanských hráčů ovšem byly z větších vzdáleností od branky, a ne vždy z optimálních pozic, průměrně měla sparťanská střela xG = 0,1 (1 z 10 takovýchto střel rozvlní síť soupeře), Slavie si z protiútoků vytvořila 5 střel s průměrným xG=0,3, tedy o něco kvalitnější šance. Sparta tedy za zápas získala $9 \cdot 0{,}1 = 0{,}9$ xG, zatímco sešívaní $5 \cdot 0{,}3 = 1{,}5$ xG, vidíme, že Slavie byla v utkání nebezpečnější a měla k výhře blíže než její soupeř.

Pomocí xG lze sledovat jednotlivé hráče, můžeme mít hráče kteří dávají hromadu branek, ovšem jejich xG je daleko nižší, z toho usuzujeme, že hráč má skvělou formu nebo prostě jen štěstí, ze statistického hlediska mu takové výkony, kdy překračuje své xG, dlouhodobě nevydrží, jelikož během delšího časového horizontu své xG překonává pouze jeden hráč na světě – Lionel Messi[4]. Naopak můžeme najít útočníky, kteří dávají výrazně méně branek, než je jejich xG, to pak může kluby odradit od jejich angažování, případně je donutit zapracovat na zakončování hráčů. Dále bývá xG využíváno pro skautování skrytých talentů, či pro předpověď dalších zápasů, což je jeho veliká výhoda oproti čistě popisným statistikám.

3. Potřebný matematický aparát

V této kapitole jsou uvedeny matematické nástroje využíváné posléze pro vytvoření modelu. Vychází ze zdrojů [1] [2] [3].

3.1. Zobecněný lineární model

Základní lineární regresní modely pracují se závislými proměnnými s normálním rozdělením, aby bylo možné modelovat závislé proměnné, které mají rozdělení jiné než normální, zavádíme Zobecněné lineární modely (angl. *Generalized linear models* - GLMs). Zobecněné lineární modely obsahují tři části, náhodnou, systematickou část a *link function*.

Náhodná část Zobecněného lineárního modelu obsahuje závislou proměnnou Y s nezávislými pozorováními $(y_1,\ldots,y_N)$, s hustotou pravděpodobnosti:

$$f(y_i, \phi_i) = a(\phi_i)b(y_i)\exp[y_i Q(\phi_i)] \tag{3.1}$$

Hodnota parametru ϕ_i je funkcí vysvětlujících veličin pro každé $i = 1,\ldots, N$. Parametr Q_i označujeme jako *přirozený parametr*.

V systematické části je vyvětlující proměnné pomocí lineárního modelu přiřazen vektor $(\eta_1, ..., \eta_N)$. Nechť x_{ij} označuje hodnotu vysvětlující proměnné j $(j = 0, 1, 2, ...)$ pro subjekt i. Pak

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., N. \tag{3.2}$$

Tato lineární kombinace vysvětlujících proměnných se nazývá *lineární prediktor*. Pro koeficient konstantního členu lineárního modelu β_0 mimo vyjímky pro všechna i platí $x_{i0} = 1$. Aby byl model kompletní je nutné první dvě části propojit. Nechť $\mu_i = E(Y_i), i = 1, ..., N$. Zavádíme monotónní diferencovatelnou funkci g, platí $\eta_i = g(\mu_i)$, kde g nazýváme *link function*. Díky g existuje vztah mezi μ_i a vysvětlujícími proměnnými

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., N. \tag{3.3}$$

3.1.1. Binomické Logit Modely

Uvažujme, že chceme modelovat, zda určitá událost nastane, takovéto případy představují binární závislé proměnné. Úspěch, uskutečnění jevu a neúspěch, jeho neuskutečnění reprezentujeme jako 1 a 0. Při binomickém rozdělení pro závislou proměnnou Y platí: $P(Y = 1) = \pi$, $P(Y = 0) = 1 - \pi$ a $E(Y) = \pi$. Jedná se o speciální případ binomického rozdělení s n=1 s hustotu pravděpodobnosti

$$f(y;\pi) = \pi^y(1-\pi)^{1-y} = (1-\pi)\left(\frac{\pi}{1-\pi}\right)^y = (1-y)\exp\left[y\left(\log\frac{\pi}{1-\pi}\right)\right] \tag{3.4}$$

pro y = 0 nebo 1. Je zřejmé, že se jedná o Zobecněný lineární model s hustou pravděpodobnosti (3.1). Parametru ϕ přísluší π, $a(\pi) = 1 - \pi$, $b(y) = 1$ a $Q(\pi) = log[\pi/(1 - \pi)]$,

přirozený parametr $log[\pi/(1-\pi)]$ vyjadřuje *log odds ("zlogaritmovaná šance")*, šance, že výstup modelu bude 1, dále v textu se setkáváme s označením *logit* funkce. Modely využívající *logit* jakožto *link function* jsou popsány v kapitole 3.4, věnované *logistické regresi* a jejím modelům, neboli *logit modelům.*

3.2. Zobecněný lineární model pro binární data

Nechť Y je binární závislá proměnná, udávající "uskutečnění"nebo "neuskutečnění"události, každé pozorování ukáže jeden z těchto výsledků, označené 1 a 0. Střední hodnota Y je $E(Y) = P(Y = 1)$. Aby byla zřejmá závislost Y na hodnotách $\boldsymbol{x} = (x_1, ..., x_p)$ označíme $P(Y = 1)$ jako $\pi(\boldsymbol{x})$. Rozptyl Y je

$$\text{var}(Y) = \pi(\boldsymbol{x})[1 - \pi(\boldsymbol{x})] \tag{3.5}$$

Jedná se o rozptyl binomického rozdělení při $n = 1$.

3.2.1. Lineární Pravděpodobnostní Model

Regresní model binární závislé proměnné

$$\pi(\boldsymbol{x}) = \alpha + \beta_1 x_1 + ... + \beta_p x_p \tag{3.6}$$

je nazýván *lineární pravděpodobnostní model.* Lineární funkce pokrývají svými hodnotami celou reálnou osu, ovšem pravděpodnost nabývá hodnot v mezích pouze od 0 do 1. V modelu (3.6) může pro některé hodnoty $\boldsymbol{x}$ nastat $\pi(\boldsymbol{x}) < 0$ nebo $\pi(\boldsymbol{x}) > 1$, ovšem model může platit na omezené oblasti $\boldsymbol{x}$, pokud taková situace opravdu nastane je využívána snadná iterpretace parametrů β_j, které vyjadřují změnu $\pi(\boldsymbol{x})$ při jednotkové změně x_j.

3.3. Metoda maximální věrohodnosti

3.3.1. Pomocná tvrzení

Definice 3.1 *Nechť f je reálná funkce definovaná na otevřeném intervalu I = (a,b), kde* $-\infty < a < b < \infty$. *Funkce f se nazývá* konvexní, *platí-li*

$$f[\gamma x + (1-\gamma)y] \leq \gamma f(x) + (1-\gamma) f(y) \tag{3.7}$$

pro všechna x,y,γ *taková, že* $a < x < y < b$ *a* $0 < \gamma < 1$.
Funkce f se nazývá striktně konvexní, *platí-li pro všechna uvedená* x, y, γ *v* (3.7) *ostrá nerovnost.*
Funkce f se nazývá (striktně) konkávní, *je-li* $-f$ *(striktně) konvexní.*

Věta 3.2 *Nechť f je definována na I a nechť je tam spojitá. Nechť existuje na I spojitá* f' *a nechť* f'' *existuje a je konečná na I. Pak funkce f je konvexní právě tehdy, platí-li* $f'' \geq 0$ *pro všechna* $x \in I$.

Důkaz. Viz Fichtengolc I (1958), odst. 143, str. 299, věta 2.

Věta 3.3 (Jensenova nerovnost) *Nechť f je konvexní funkce definovaná na I. Nechť X je náhodná veličina s konečnou střední hodnotou taková, že $P(X \in I) = 1$. Pak platí*

$$f(\mathbb{E}\, X) \leq \mathbb{E}\, f(X). \tag{3.8}$$

Je-li f striktně konvexní, pak nerovnost (3.8) je ostrá s výjimkou případu, kdy veličina X je rovna konstantě s pravděpodobností 1. Nerovnost (3.8) se nazývá Jensenova.

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

3.3.2. Věrohodnostní funkce a maximálně věrohodné odhady

Uvažujme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$ se sdruženou hustotou $f(\boldsymbol{x},\boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \Omega$, kde Ω je parametriký prostor parametrů $\boldsymbol{\theta}$. Nechť $\boldsymbol{x}$ má pevně danou hodnotu, pak se funkce $f(\boldsymbol{x}, \boldsymbol{\theta})$ jakožto funkce $\boldsymbol{\theta}$ nazývá *věrohodnostní funkce.*
Hodnota $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$, která maximalizujeme pro dané $\boldsymbol{X} = \boldsymbol{x}$ věrohodnostní funkci $p(\boldsymbol{x}, \boldsymbol{\theta})$, se nazývá *maximálně věrohodný odhad* parametru $\boldsymbol{\theta}$.

Nechť $\boldsymbol{X}$ je náhodný vektor se sdruženou hustotou $f(\boldsymbol{x}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \omega \subset \mathbb{R}_m$. Uvažujme funkci $u : \Omega \to \Omega^*$, která zobrazuje Ω na $\Omega^* \subset \mathbb{R}_k$. Předpisem $\boldsymbol{\theta}^* = u(\boldsymbol{\theta})$ každému $\boldsymbol{\theta} \in \Omega$ přiřadíme $\boldsymbol{\theta}^* \in \Omega^*$. Nechť $G(\boldsymbol{\theta}^*) = \boldsymbol{\theta} : \boldsymbol{\theta} \in \Omega, u(\boldsymbol{\theta}) = \boldsymbol{\theta^*}$. Označme

$$M(\boldsymbol{x}, \boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta} \in G(\boldsymbol{\theta}^*)} f(\boldsymbol{x}, \boldsymbol{\theta})$$

M jakožto funkci $\boldsymbol{\theta}^*$ nazýváme *věrohodnostní funkcí indukovanou parametrickou funkcí u.* Hodnotu $\hat{\boldsymbol{\theta}}^*$, která maximalizuje $M(\boldsymbol{X}, \boldsymbol{\theta}^*)$ označujeme jakožto *maximálně věrohodný odhad parametrické funkce u.*

Nechť θ je jednorozměrný parametr a platí následující předpoklady.

$\mathbf{P}_1$: Nechť Ω je parametrický prostor, který obsahuje takový neprázdný otevřený interval ω, že skutečná hodnota parametru θ_0 patří do ω.

$\mathbf{P}_2$: Nechť $\boldsymbol{X} = (X_1, \ldots, X_2)'$, kde X_i jsou stejně rozdělené, nezávislé veličiny s hustotou $f(x, \theta)$ vzhledem k nějaké σ-konečné míře μ.

$\mathbf{P}_3$: Nechť $M = x : f(x, \theta) > 0$ nezávisí na θ.

$\mathbf{P}_4$: Nechť $\theta_1, \theta_2 \in \Omega$. Pak $f(x, \theta_1) = f(x, \theta_2)$ skoro všude (vzhledem k míře μ) právě tehdy, je-li $\theta_1 = \theta_2$

Tedy vzhledem k míře $\nu = \mu \times \cdots \times \mu$ je sdružená hustota náhodného vektoru $\boldsymbol{X}$ $f(\boldsymbol{x}, \theta) = f(x_1, \theta) \times \cdots \times f(x_n, \theta)$.

Věta 3.4 *Jestliže $n \to \infty$, pak pro každé takové pevné $\theta \in \Omega$, že $\theta \neq \theta_0$, platí*

$$P_{\theta_0}\{f(\boldsymbol{X}, \theta_0) > f(\boldsymbol{X}, \theta)\} \to 1 \tag{3.9}$$

Důkaz. Viz [3] kpt. 7.6, str. 149.

Pokračujme v úvaze, že θ je jednorozměrný parametr. Uvažujme funkci proměnné θ $f(\boldsymbol{x}, \theta)$ pro pevné $\boldsymbol{x}$. Funkce

$$L(\boldsymbol{x}, \theta) = \ln f(\boldsymbol{x}, \theta) \tag{3.10}$$

se nazývá *logaritmická věrohodnostní funkce*. Často $L(\boldsymbol{x}, \theta)$ značíme jen jako $L(\theta)$.

Věta 3.5 *Nechť jsou splněny předpoklady $P_1 - P_4$. Nechť na intervalu ω existuje $f'(x, \theta) = \frac{\partial f(x,\theta)}{\partial \theta}$ pro skoro všechna x. Pak pro každé $\epsilon > 0$ při $n \to \infty$ platí, že s pravděpodobností konvergující k jedné má věrohodnostní rovnice*

$$\frac{\partial L(\boldsymbol{X}, \theta)}{\partial \theta} = 0 \tag{3.11}$$

takový kořen $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X})$, že $|\hat{\theta}_n - \theta_0| < \epsilon$, kde θ_0 je skutečná hodnota parametru θ, v níž $p(\boldsymbol{x}, \theta)$ nabývá svého maxima.

Důkaz. Nechť $\epsilon > 0$ je tak malé, že $[\theta_0 - \epsilon, \theta_0 + \epsilon] \subset \omega$. Definujme

$$S_n = \boldsymbol{x} : L(\boldsymbol{x}, \theta_0) > L(\boldsymbol{x}, \theta_0 - \epsilon) \text{ a } L(\boldsymbol{x}, \theta_0) > L(\boldsymbol{x}, \theta_0 + \epsilon). \tag{3.12}$$

Dle (3.9) platí $P_{\theta_0}(\boldsymbol{X} \in S_n) \to 1$. Pro každé $\boldsymbol{X} \in S_n$ tedy s pravděpodobností blížící se jedné existuje $\hat{\theta}_n$ takové, že $\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon$ a že funkce $L(\boldsymbol{X}, \theta)$ má lokální maximum v bodě $\theta = \hat{\theta}_n$. Pak $L'(\boldsymbol{X}, \hat{\theta}_n) = 0$.

3.3.3. Věrohodnostní funkce veličiny s binomickým rozdělením

Uvažujme náhodnou veličinu $X \sim Bi(n, \pi), 0 < \pi < 1$, pro kterou platí

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

Hledejme maximálně věrohodný odhad π, tedy pro X=x jde o maximalizaci funkce $g(\pi) = \pi^x (1 - \pi)^{n-x}$. Nechť $x \neq 0, x \neq n$. Pak

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

nulovým bodem funkce $g'(\pi)$ je $\hat{\pi} = \frac{x}{n}$. Nyní si ukážeme, že v tomto bodě skutečně funkce $g'(\pi)$ nabývá svého maxima. Definujme

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

Aritmetický průměr těchto čísel je

$$\bar{z} = \frac{1}{n}\left[x\frac{n\pi}{x} + (n - x)\frac{n(1 - \pi)}{n - x}\right] = 1, \tag{3.16}$$

geometrický průměr je

$$\bar{z_G} = \left\{ \left(\frac{n\pi}{x}\right)^x \left[\frac{n(1-\pi)}{n-x}\right]^{n-x} \right\}^{\frac{1}{n}} \tag{3.17}$$

z nerovnosti $\bar{z_G} \leq \bar{z}$ zjistíme, že

$$\pi^x(1-\pi)^{n-x} \leq \left(\frac{x}{n}\right)^x \left(\frac{n-x}{n}\right)^{n-x} \tag{3.18}$$

zjevně při $\pi = \hat{\pi} = \frac{x}{n}$ nastává rovnost a tedy $\hat{\pi}$ maximalizuje funkci $g(\pi)$.
Nastane-li jeden z případů $x = 0$ nebo $x = n$, pak maximálně věrohodný odhad parametru π na inetrvalu $(0, 1)$ neexistuje, nastal by spor s předpokladem $\boldsymbol{P_1}$.

3.4. Logistická regrese

Pro práci s binárními závislými proměnnými, je vhodný nástroj logistická regrese, která je široce využívána napříč obory, ať už se jedná o medicínu, ekonomii nebo právě sport.

3.4.1. Interpretace parametrů

Předpokládejme binární závislou prměnnou Y a a proměnnou X, na které je Y závislá, nechť $\pi(\boldsymbol{x}) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. Model Logistické regrese je

$$\pi(\boldsymbol{x}) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{3.19}$$

Je tedy zjevné, že logit je lineární

$$\text{logit}[\pi(\boldsymbol{x})] = \log\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \alpha + \beta x \tag{3.20}$$

Interpretace parametru β

Nyní se zaměříme na β z rovnice (3.20), znaménko tohoto parametru určuje, zda $\pi(\boldsymbol{x})$ bude rostoucí či klesající s rostoucím $\boldsymbol{x}$. Sklon logistické křivky roste s nárůstem $|\beta|$, pro $\beta \to 0$ se logistická křivka přibližuje k horizontální rovné přímce. V případě nezávislosti Y na X platí $\beta = 0$.
Aplikací exponenciály na obě strany rovnice (3.20) ukazujeme, že *"šance"(odds)* jsou exponenciální funkcí $\boldsymbol{x}$, odds se zvětší e^{β}-krát s každým jednotkovým navýšením proměnné $\boldsymbol{x}$, e^{β} je poměr odds při $X = x + 1$ a odds při $X = x$. Vzhledem k nelineární závislosti $\pi(\boldsymbol{x})$ na $\boldsymbol{x}$ je změna $\pi(\boldsymbol{x})$ pro rozdílná $\boldsymbol{x}$ v logistické regresi "nerovnoměrná".

3.4.2. Fitování modelů logistické regrese

Uvažujme n binárních nezávislých pozorování, nechť $x_i = (x_{i1}, \ldots, x_{ip})$ udává nastavení i hodnot p vyvětlujících proměnných, *i=1,...,N*. Nabývají-li všechna pozorování různých hodnot, pak *N=n*. Model logistické regrese (3.19), uvažující regresní parametr α jako konstantní, je

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

Věrohodnostní rovnice

Když více než jedno pozorování nastane při pevné hodnotě x_i, je vhodné zaznamenat počet pozorování n_i a počet úspěšných pokusů. Poté uvažujeme y_i jakožto počet úspěšných pokusů, namísto odezvy jednotlivých binárních proměnných. Pak $Y_1, \ldots, Y_N$ jsou nezávislé binomické proměnné se střední hodnotou $E(Y_i) = n_i\pi(x_i)$, kde $n_1 + \ldots n_n = n$. Jejich marginální hustota pravděpodobnosti je úměrná součinu N binomických funkcí

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

Pro model (3.21),je i-tý logit $\sum_j \beta_j x_{ij}$, takže exponenciální člen v posledním výraze je roven $\exp[\sum_i y_i(\sum_j \beta_j x_{ij})] = \exp[\sum_i(\sum_j y_i x_{ij})\beta_j]$. Jelikož $[1-\pi(x_i)] = [1+exp(\sum_j \beta_j x_{ij}]^{-1}$, tak věrohodnostní funkce je rovna

$$L(\boldsymbol{\beta}) = \sum_j \left(\sum_i y_i x_{ij}\right)\beta_j - \sum_i n_i \log\left[1 + exp\left(\sum_j \beta_j x_{ij}\right)\right]. \tag{3.23}$$

Věrohodnostní rovnice získáme, když položíme $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 0$. Platí

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij}\frac{\exp(\sum_k \beta_k x_{ik})}{1+\exp(\sum_k \beta_k x_{ik})}, \tag{3.24}$$

věrohodnostní rovnice jsou

$$\sum_i y_i x_{ij} - \sum_i n_i \hat{\pi}_i x_{ij} = 0, \qquad j = 1, \ldots, p, \tag{3.25}$$

kde $\hat{\pi}_i = \exp(\sum_k \hat{\beta}_k x_{ik})/[1+\exp(\sum_k \hat{\beta}_k x_{ik})]$ je maximálně věrohodný odhad $\pi(x_i)$. Tyto rovnice jsou nelineární a vyžadují numerické řešení. Při modelování je v programovacím jazyce Python během vytváření modelu využívána metoda *IRLS - Iteratively reweighted least squares*[10].

3.4.3. Testování podmodelu

K testování podmodelu je využit *test poměrem věrohodnosti (Likelyhood ratio test)*, který využívá logaritmickou věrohodnostní funkci, která pro binomické závislé proměnné vypadá následovně:

$$\begin{aligned}L(\boldsymbol{\mu}) &= \log\prod_{i=1}^{n} \mu_i^{Y_i}(1-\mu_i)^{1-Y_i} \\ &= \sum_{i=1}^{n}(Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i) \\ &= \sum_{i=1}^{n} Y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \sum_{i=1}^{n}\log(1-\mu_i)\end{aligned} \tag{3.26}$$

Pozorované náhodné veličiny se v logaritmické věrohodnostní funkci vyskytují v součinech s výrazy log $(\mu_i/(1-\mu_i))$.

Uvažujme nejbohatší možný model, model s větší hodnotou věrohodnostní funkce nelze vytvořit, takový model se nazývá *saturovaný*. Saturovaný model má právě tolik parametrů, kolik je různých hodnot vektorů x_i. Maximální hodnotu věrohodnostní funkce v saturovaném modelu označíme L_{max}. Každý další model je podmodelem saturovaného modelu. Pomocí *deviance* posoudíme přiléhavost běžného modelu

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

Čím je model přiléhavější, hodnota deviance D klesá. Dále předpokládáme, že všechny vektory x_i jsou různé, pak má saturovaný model k parametrů $\mu_1, ..., \mu_k$. Odhadem střední hodnoty μ_i je Y_i, dle (3.26) platí

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

Označme odhad pravděpodobnosti jedničky $\hat{\mu}_i = \mu(\boldsymbol{x_i})$, dosazením do (3.28) vyjádříme devianci v modelu logistické regrese jako

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

Test poměrem věrohodnosti

Uvažujme model M_1 s odhadem parametru $\boldsymbol{b} = \hat{\boldsymbol{b}}$ a jeho podmodel M_2, který vznikl odebrání části regresorů modelu M_1, s odhadem partametru $\boldsymbol{b} = \tilde{\boldsymbol{b}}$. Při testu poměrem věrohodnosti testujeme, zda všechny parametry obsažené v modelu M_1 a zároveň vynechané v podmodelu M_2 jsou rovny nule, porovnáváme hodnoty logaritmické věrohodnostní funkce pro $\hat{\boldsymbol{b}}$ a $\tilde{\boldsymbol{b}}$ pomocí statistiky

$$LR = 2\left(L(\hat{\boldsymbol{b}}) - L(\tilde{\boldsymbol{b}})\right). \tag{3.30}$$

LR statistiku lze také vyjádřit pomocí deviance modelu a podmodelu

$$2\left(L(\hat{\boldsymbol{b}}) - L(\tilde{\boldsymbol{b}})\right) = \left(2(L_{max} - L(\tilde{\boldsymbol{b}}))\right) - \left(2(L_{max} - L(\hat{\boldsymbol{b}}))\right) = D(\tilde{\boldsymbol{b}}) - D(\hat{\boldsymbol{b}}). \tag{3.31}$$

Platí-li testovaný podmodel a jsou-li zároveň splněny pomínky regularity[2], pak má statistika LR asymptoticky rozdělení χ^2_q, kde q je rozdíl počtu nezávislých parametrů v modelech, které porovnáváme.

4. Modelování pravděpodobnosti skórování

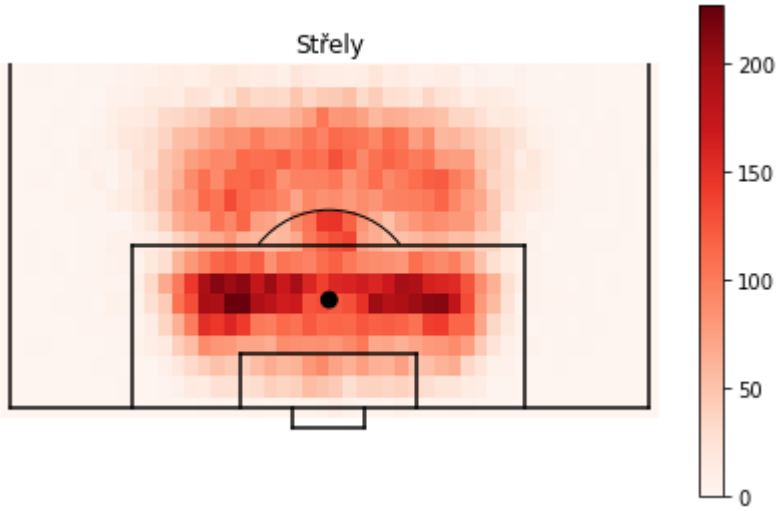
V této kapitole jsou uvedeny jednotlivé modely pro modelování pravděpodnosti skórování ve fotbale, modelování Expected goals - xG. Data byla získána od firmy Statsbomb, která je volně distribuuje, jejich následné zpracování proběhlo v programovacím jazyce Python, s využitím knihoven jako *pandas, numpy, statsmodels* nebo *FCPython*. Zpracováno bylo 568 zápasů, ve kterých dohromady padlo 14775 střel, branka padla v 1753 případech.
Postupně budou uváděny modely od nejjednoduššího po modely komplexnější, které pak lze otestovat proti jednodušším modelům, a zjistit tak relevanci jejich parametrů.

4.1. Triválnί modely

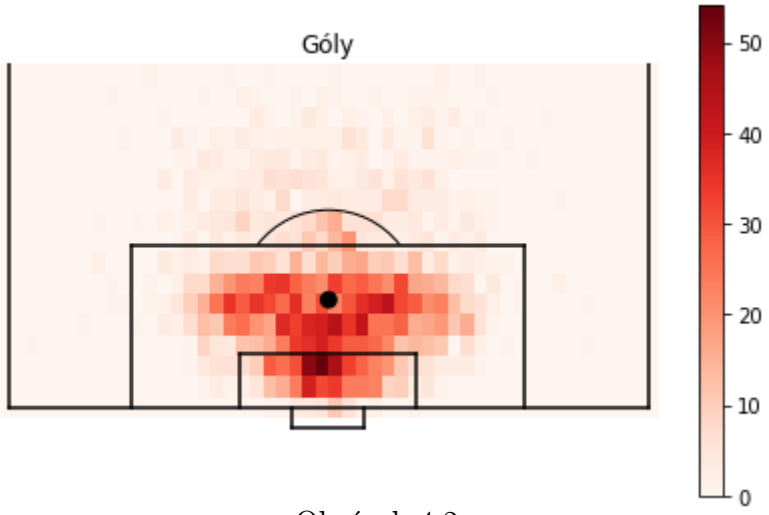
Chceme.li znát pravděpodobnost, že ze střely padne branka, tak se samozřejmě jako nejjednodušší úvaha nabízí poměr

$$xG = \frac{\text{celkový počet gólů}}{\text{celkový počet střel}}.$$

Výsledkem tohoto triviálního modelu, je že každá střela má $xG = 0,1184$, jelikož z principu víme, že pravděpodobnost padnutí branky ze 2 metrů je větší, než, že hráč rozvlní síť z poloviny hřiště, je zjevné, že daný model lze sice využít, spoléhat na jeho jakoukoliv přesnost je ovšem na pováženou.
Proto dále uvažujeme opět poměr $\frac{\text{celkový počet gólů}}{\text{celkový počet střel}}$, nicméně již jej počítáme pro každý "čtverec"metr na metr na hrací ploše. Celou hrací plochu, kterou uvažujeme délky 100 a šířky 70 metrů, jsme rozdělili na síť těchto metrových čtverců, přičemž jejich hranice jsou celočíselné. Takže přesněji počítáme $\frac{\text{celkový počet gólů z jednoho čtverce na hřišti}}{\text{celkový počet střel z toho samého čtverce na hřišti}}$. Tento model je vyobrazen na 4.3. Pro snažší interpretaci byly na obrázcích 4.1 respektive 4.2, vyobrazeny všechny střely a góly z použité databáze.

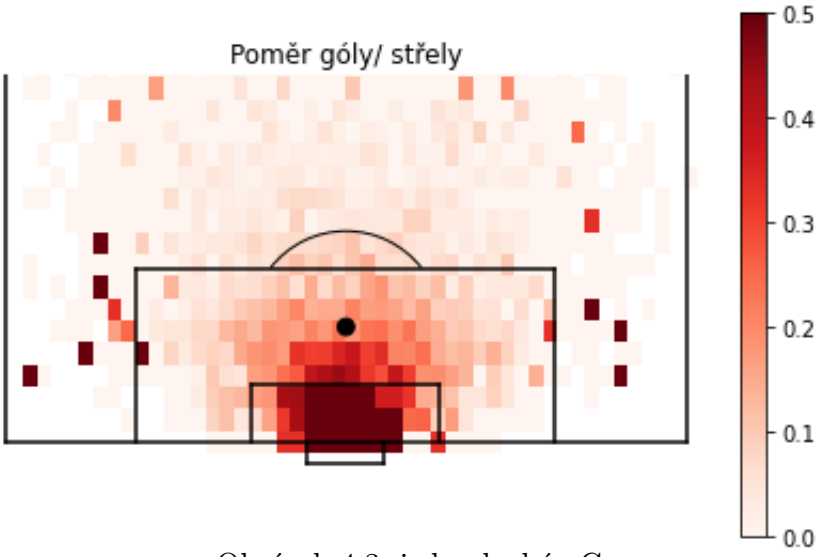


Obrázek 4.1:
lokace jednotlivých střel



Obrázek 4.2:
lokace všech gólů

Na obrázku 4.3 již vidíme vykreslené jednoduché xG, tmavá oblast okolo brány potvrzuje, že nejvyšší pravděpodobnost vstřelení branky je, když jsme k ní co nejblíže, tmavší body u stran hřiště jsou způsobené malým počtem střel z těchto pozic v databázi. Vzhledem k nerovnoměrnému počtu střel z jednotlivých částí hřiště, je ovšem tento model značně nepřesný.



Obrázek 4.3: jednoduché xG

4.2. Pokročilé modely

Nyní již začneme využívat matematické nástroje popsané v kapitole 3. Binární data na výstupu modelu, padne/nepadne branka, modelujeme pomocí logistické regrese, kdy budeme postupně vytvářet modely s různými proměnnými, jejichž počet budeme postupně navyšovat a vytvářet tak modely komplexnější.

4.2.1. Základní modely logistické regrese

xG v závislosti na vzdálenosti

Začneme modelem jedné proměnné - X, bereme za ni vzdálenost od branky, nikoliv ovšem od jejího středu, ale kolmou vzdálenost od brankové čáry, takto uvažovanou vzdálenost značíme X i v dalších modelech. Program z dat získal následující hodnoty koeficientů a jejich intervalový odhad při hladině významnosti $\alpha = 0.05$.

Tabulka 4.1: xG v závislosti na X

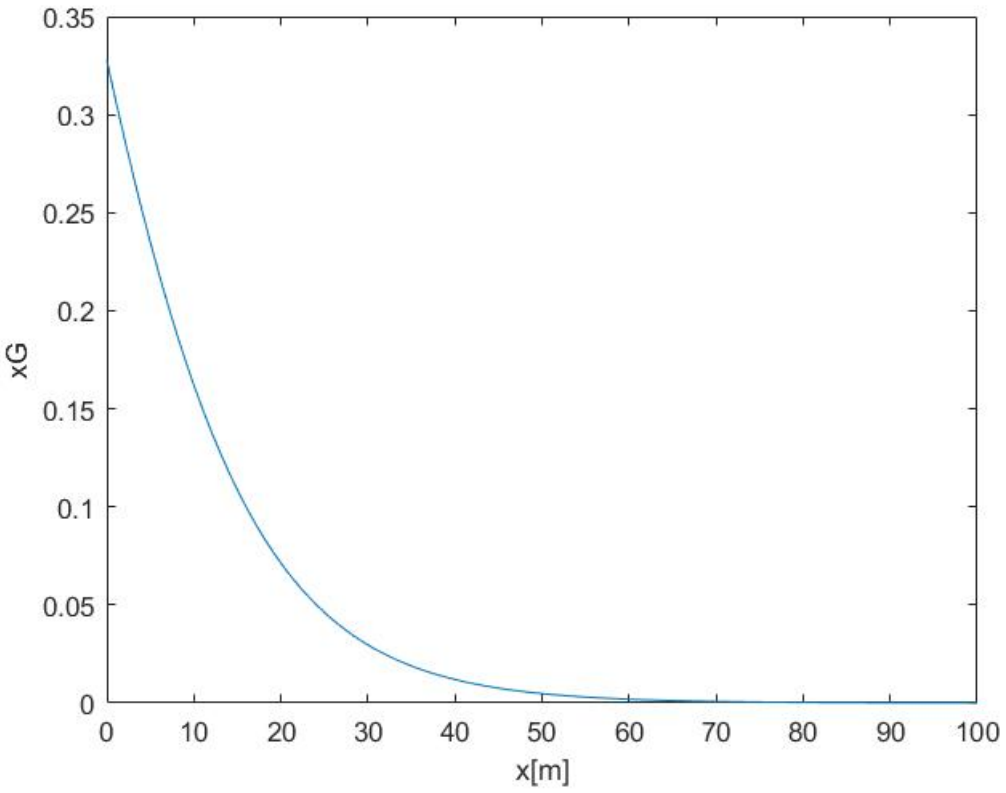
Proměnná	**koeficinet**	**[0,025**	**0,975]**
konstanta	0,7193	0,615	0,824
X	0,0924	0,085	0,100

Deviance modelu je 10057

Takže funkce xG jako taková vypadala následovně

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

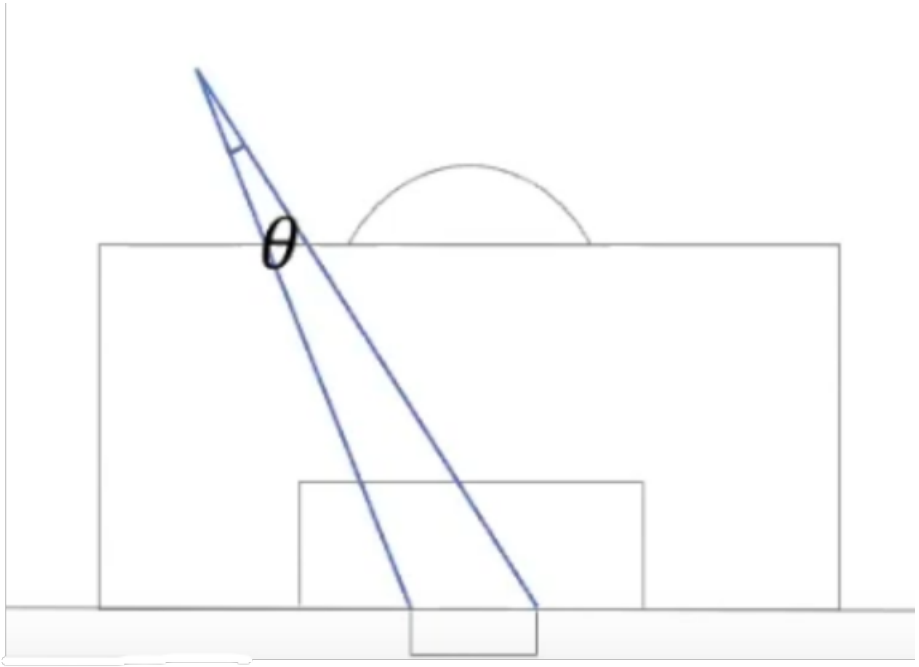
můžeme ji vidět vykreslenou ná následujícím obrázku.



Obrázek 4.4: xG v závislosti na vzdáloenosti

Z obrázku 4.4 je zřetelně pozorovatelný nelineární vztah pravděpodobnosti padnutí branky a střelcově vzdálenosti od brankové čáry, co ovšem model neuvažuje je, kde přesně se hráč nachází, může být ve stejné vzdálenosti X od brankové čáry, nicméně může stát někde u autové čáry nebo někde uprostřed, oběma těmto případům přiřadí model stejné xG,

proto kromě kolmé vzdálenosti X, přidáme do modelu úhel θ, pod kterým hráč brankovou konstrukci vidí, viz 4.5.



Obrázek 4.5: Úhel, pod kterým střelec vidí branku

xG v závislosti na vzdálenosti a úhlu

Nyní již v modelu krom X uvažujeme i úhel θ popsaný výše. Problematika stejných hodnot xG pro situace, když je hráč v rohu hřiště a když je těsně před brankou, tímto zaniká (jejich vzdálenost X od brankové čáry je v takovém případě stejná). Při hladině hladině významnosti $\alpha = 0.05$ vyšly koeficienty proměnných a jejich intervalové odhady následovně.

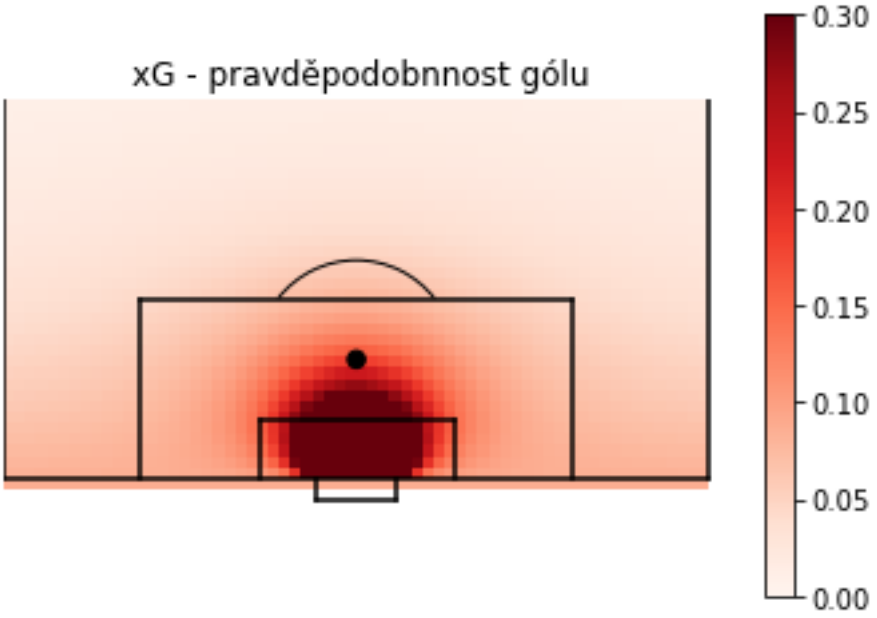
Tabulka 4.2: xG v závislosti na X a θ

Proměnná	**koeficinet**	**[0,025**	**0,975]**
konstanta	2,0586	1,857	2,260
X	0,0503	0,041	0,059
θ	-1,7412	-1,957	-1,525

Deviance modelu je 9778

Známenko koeficientu jednoduše interpretuje, zdali se s rostoucí proměnnou hodnota xG zvyšuje, resp. snižuje. Je-li koeficient proměnné kladný, tak s rostoucí proměnnou xG klesá, lze vidět u vzdálenosti X, čím dál je hráč od brankové čáry, tím bude menší pravděpodobnost, že skóruje. Z logiky věci plyne, bude-li znaménko koeficientu proměnné mínus, bude s rostoucí proměnnou růst zároveň hodnota xG. Tento stav pozorujeme u úhlu θ, pravděpodobnost střelení branky roste s velikostí úhlu, pod kterým hráč vidí brankovou konstrukci. Někteří trenéři mládeže tento fakt využívají při výchově mladých fotbalových talentů, snaží se jim předat informaci, pokud mají dobrý výhled na branku

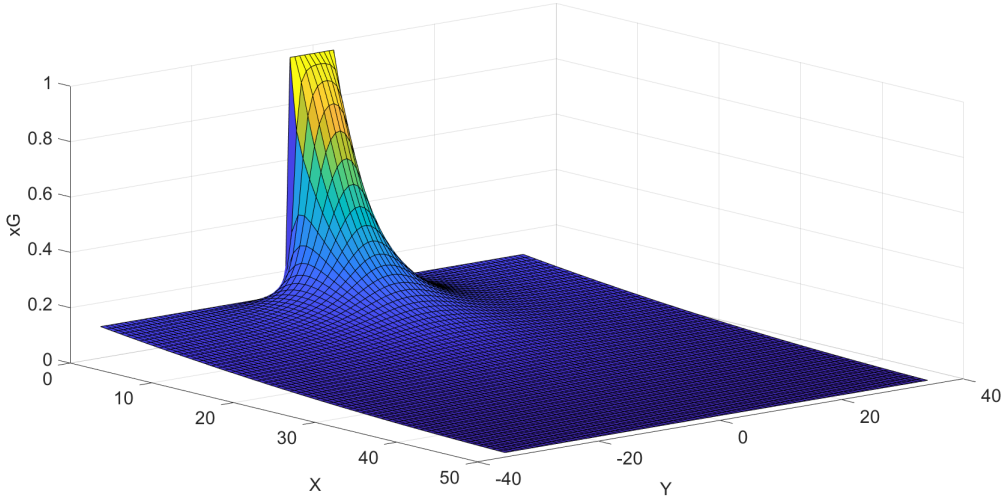
stojí za to akci zakončit, jak si tuto informaci přebírají hráči je už na nich.
Obrázky 4.6 a 4.7 ukazují rozložení pravděpodobnosti skórování na hrací ploše, zlepšení oproti předchozím modelům je ihned pozorovatelné.



Obrázek 4.6: xG v závislosti na X a θ

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

$$xG = \frac{1}{1 + e^{2{,}0586 + 0{,}0503X - 1{,}7412\theta}} \tag{4.2}$$



Obrázek 4.7: xG v závislosti na X a θ

4.2.2. Pokročilé modely logistické regrese

V předchozích modelech byly uvažovány pouze spojité proměnné, ať už se jedná o vzdálenost či úhel. V této části budeme model s proměnnými θ, X postupně rozšiřovat o

proměnné diskrétní, může se jednat například o způsob zakončení a další. Zároveň budou modely testovány, jestli nám jejich rozšíření přidá na jejich "kvalitě"nebo jestli přidaná proměnná není pro modelování relevantní.

xG v závislosti na vzdálenosti, úhlu a "tlaku"

Zahrnování diskrétních proměnných začne přidáním proměnné, která specifikuje, zda střílející hráč byl během zakončení pod tlakem či nikoliv. Můžeme ji specifikovat, jakožto okolnosti, které mohly ovlivnit hráče během střelby. Podstupoval zároveň souboj s protihráčem? Byl protihráčem napadán, že musel se střelou spěchat jinak by o míč přišel nebo by již ke střele neměl prostor? Tyto a podobné otázky jsou zahrnuté v naší proměnné pojmenované "tlak", jelikož se jedná o diskrétní proměnnou nabývá hodnot 1 - hráč byl při střele pod tlakem a 0 - hráč nebyl při střele pod tlakem, měl na ni dostatek času a prostoru.

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

Proměnná	**koeficinet**	**[0,025**	**0,975]**
konstanta	1,9392	1,736	2,142
X	0,0528	0,044	0,062
θ	-1,7850	-2,003	-1,567
tlak	0,6291	0,479	0,779

Deviance modelu je 9702,5

Vidíme, že pokud je hráč pod tlakem, šance na skórování se snižuje.
Nyní pomocí testu poměrem věrohodnosti zjistíme, zda je přidaná proměnná, tedy *tlak*, pro model významná. Porovnáme předchozí model $\boldsymbol{\mu_{n1}}$: *xG = f(X, θ)* - model "nižší"a poslední "rozsáhlejší"model $\boldsymbol{\mu_{v1}}$, *xG = f(X, θ, tlak)* - model "vyšší".
Pro oba modely pomocí programu získáme logaritmickou věrohodnostní funkci $L(\boldsymbol{\mu})$, bude určena hodnota χ^2 testu poměrem věrohodnosti, kterým je porovnáváme a výsledná p-hodnota bude porovnána s hladinou věrohodnosti $\alpha = 0.05$.

$$L(\boldsymbol{\mu_{v1}}) = -4851,257934716189$$
$$L(\boldsymbol{\mu_{n1}}) = -4888,9960593471619$$
výsledek χ^2 testu: 75,47624959132008
p-hodnota= $4,078877325525352 \cdot 10^{-17}$

Vidíme, p-hodnota $< \alpha$, odebráním proměnné tlaku, se sníží kvalita modelu, tato proměnná je pro model významná a proto budeme dále pracovat s modelem *xG = f(X, θ, tlak)*.

xG v závislosti na vzdálenosti, úhlu, tlaku a způsobu zakončení

Nyní zahrneme způsob zakončení, jakožto proměnné modelu. Uvažujeme zakončení levou (lf) a pravou nohou (rf), třetí způsob zakončení, hlavou je zahrnut v konstantním členu.

Tabulka 4.4: xG v závislosti na X,θ, tlaku a způsobu zakončení

Proměnná	**koeficinet**	**[0.025**	**0.975]**
konstanta	3,1537	2,877	3,431
X	0,0600	0,051	0,069
θ	-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

Koeficienty u zakončení levou či pravou jsou záporné a podobné. Jelikož v modelu neuvažujeme kdo střílel, mají pro nás pravá i levá noha stejnou váhu, záporné znaménko souvisí s poslední možností zakončení a to hlavou, pravděpoodobnost gólu vstřeleným nohou je vyšší oproti zakončení hlavou, proto po dosazení "jedničky"za jednu z proměnných *lf, rf*, což se v realitě rovná zakončení spodní končetinou, zvýšíme hodnotu xG střely.
Test poměrem věrohodnosti vyšel následovně, za "vyšší"model $\boldsymbol{\mu_{v2}}$ považujeme $xG = f(X, \theta, tlak, rf, lf)$, za "nižší"model $\boldsymbol{\mu_{n2}}$ bereme $xG = f(X,\theta, tlak)$.

$$L(\boldsymbol{\mu_{v2}}) = -4738,609961574685$$
$$L(\boldsymbol{\mu_{n2}}) = -4851,257934716189$$
$$\text{výsledek } \chi^2 \text{ testu: } 225{,}295946$$
$$\text{p-hodnota= } 1.1956577 \cdot 10^{-49}$$

Opět uvažujeme hladinu významnosti α =0,05, p-hodnota $< \alpha$. Odebrání proměnných *lf, rf* značně ovlivnilo model, jsou pro něj významné. Tudíž dále budeme pracovat s modelem $xG = f(X, \theta, tlak, rf, lf)$.

xG v závislosti na vzdálenosti, úhlu, tlaku, způsobu zakončení a straně hřiště, ze které bylo zakončeno

Nyní budeme zkoumat, zdali ovlivníme model, budeme-li jakožto proměnnou uvažovat stranu, ze které hráč zakončil- pravá, levá strana hřiště. Jakožto proměnnou bereme například pravou stranu (rs), kterou stranu vezmeme model neovlivní, druhá strana se promítne do konstantního členu. Program vypočítal koeficienty a jejich intervalové odhady na hladině významnosti $\alpha = 0,05$ následovně:

Tabulka 4.5: xG v závislosti na X, θ, tlaku, způsobu zskončení a straně hřiště, ze které se střílelo

Proměnná	**koeficinet**	**[0,025**	**0,975]**
konstanta	3,1980	2,915	3,481
X	0,0601	0,051	0,069
θ	-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

Testem poměrem věrohodnosti nyní bude zkontrolován vliv strany hřiště na pravděpodobnost skŕování. "Vyšší model" $\boldsymbol{\mu_{v3}}$ - *xG=f(X, θ, tlak, rf, lf, rs)*, porovnáme s modelem "nižším" $\boldsymbol{\mu_{n3}}$ - *xG=f(X, θ, tlak, rf, lf)*.

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

Nyní již p-hodnota překročila hladinu významnosti $\alpha = 0,05$, vliv strany, ze které bylo zakončeno je při této hladině zanedbatelný, dostatečně přesné výsledky poskytuje "nižší model" $\boldsymbol{\mu_{n3}}$ - *xG=f(X, θ, tlak, rf, lf)*. Pravděpodonost skŕování, xG bude tedy počítáno v závislosti na vzdálenosti, úhlu, tlaku a způsobu zakončení.

Tabulka 4.6: xG v závislosti na X,θ, tlaku a způsobu zakončení

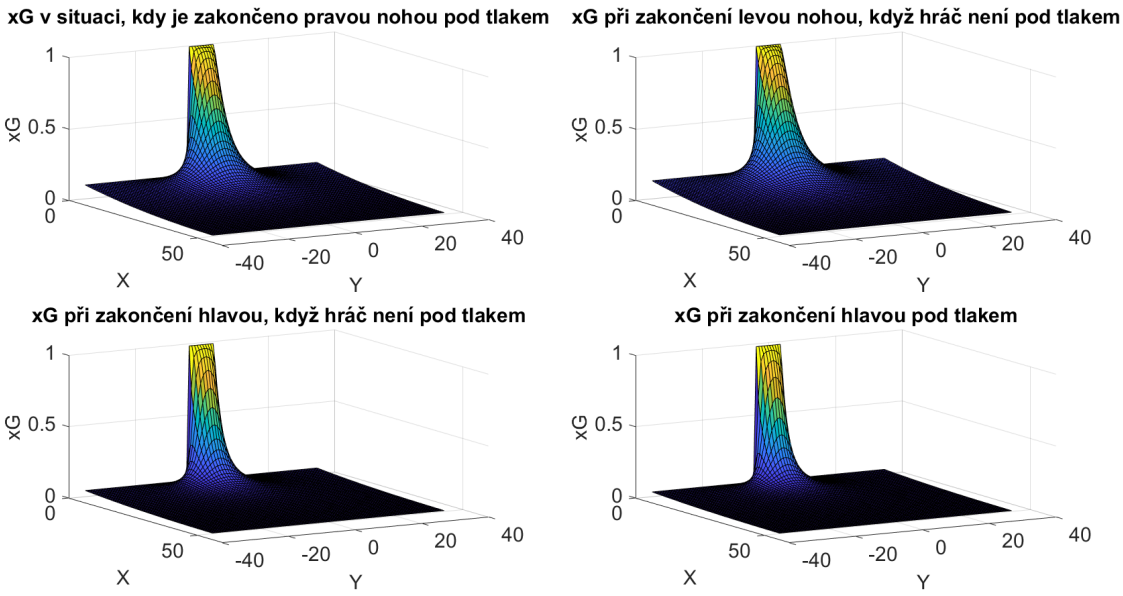
Proměnná	**koeficinet**	**[0.025**	**0.975]**
konstanta	3,1537	2,877	3,431
X	0,0600	0,051	0,069
θ	-2,3270	-2,569	-2,085
tlak	0,2884	0,133	0,444
levá noha (lf)	-1,2302	-1,418	-1,042
pravá noha (rf)	-1,2242	-1,403	-1,046

Deviance modelu je 9477,2

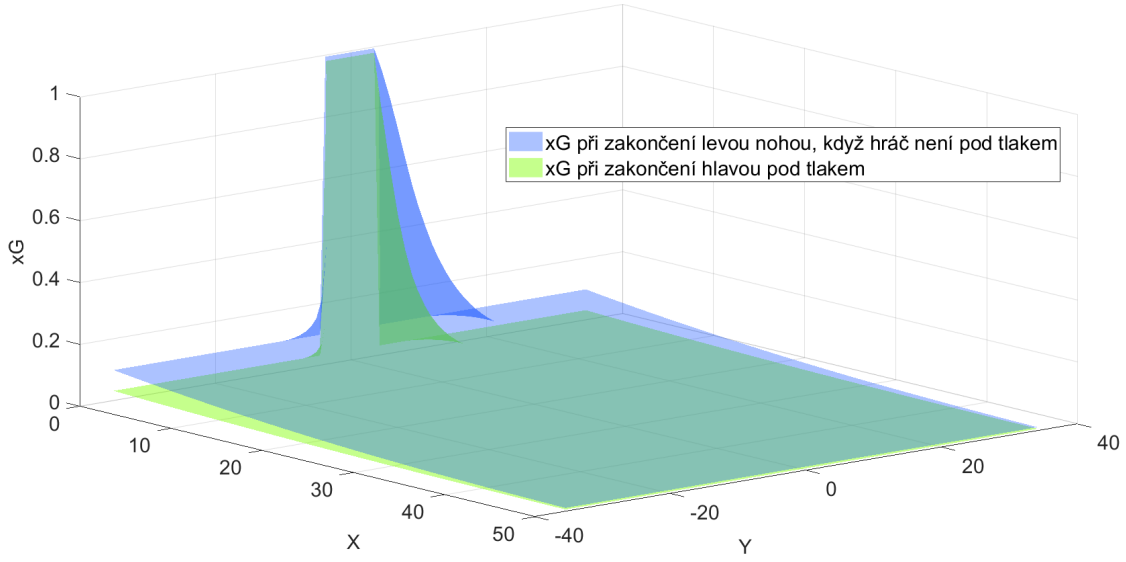
Samotná funkce xG s danými koeficienty z tabulky vypadáa následovně:

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

Po testování jednotlivých modelů, zůstaly tyto proměnné jakožto pro model relevantní, tohle tvrzení lze podpořit i shlédnutím jejich intervalových odhadů, kde žádný neobsahuje nulu.
Následující obrázky 4.8 ukazují některé případy výsledného modelu. Nejpatrnější je změna při zakončování hlavičkou, kde je šance skórování nižší, což je na grafu vidět jeho "zploštěním"oproti dvoum grafům nad ním. Zároveň je patrný vliv "tlaku"na hráče, kdy graf potvrzuje, že hráč který není pod tlakem má z dané pozice vyšší šanci na vstřelení branky.



Obrázek 4.8: xG pro různé varianty výsledného modelu.



Obrázek 4.9: Porovnání dvou situací z obrázku 4.8

.

Na obrázku 4.9 jsou vyobrazeny situace, kdy model nabývá nejvyšších hodnot, pravděpodobnost padnutí gólu při střelbě levou nohou, když hráč je vklidu - modrá plocha, a situace, kby nabývá hodnot nejnižších, pravděpodobnost padnutí branky při zakončení hlavou pod tlakem - zelená plocha. Rozdíl, jakožto výše zmíněné zploštění grafu je u zelené plochy oproti ploše modré na první pohled zjevný. Graf dokazuje, že zakončení nohou v klidu má vyšší pravděpodonost, že skončí brankou, oproti zakončení hlavou pod tlakem, uvažujeme-li zakončení ze stejné pozice na hřišti. Zelená plocha nabývá v každém bodě hrací plochy (souřadnice X,Y) nižší hodnoty než plocha modrá.

5. Závěr

Práce popisuje problematiku pravděpodobnosti skórování ve sportu, představuje jednotlivé metody používaných modelů ve sportech, zvláště podrobně pak ve fotbale. Zavádí nejznámější metodu "očekávaných gólů", představí její výhody i nevýhody. Problematika je následně popsána ze statistického hlediska, zavedením zobecněných lineárních modelů, přesněji pak modelů logistické regrese, které jsou k interpretaci binomických závislých proměnných, jako je výskyt nebo nevýskyt určitého jevu, zde padnutí branky, vhodné. Představena je metoda maximální věrohodnosti, která je využita při výpočtu koeficientů jednotlivých proměnných modelu. Jednotlivé modely jsou poté charakterizovány pomocí jejich deviance a porovnány na základě věrohodnostních fukncí.
Představený matematický aparát je následně aplikován v programovacím jazyce Python, kde z open source dat firmy Statsbomb[7] byly získány koeficienty jednotlivých spojitých i diskrétných vysvětlujících proměnných, pomocí kterých byly následně sestaveny modely pravděpodobnosti skórování v profesionálním fotbalu. Vypočtené modely s různými počty vysvětlujících proměnných byly následně pomocí testu věrohodnostním poměrem porovnávány. Bylo zjištěno, které vysvětlující proměnné mají statisticky významný vliv na pravděpodobnost vstřelení branky. Po testování několika modelů dochází práce k závěru, že z dostupných dat pravděpodobnost skórování ve fotbale nejvhodněji predikuje model využívající proměnné: způsob zakončení, herní situace a poloha na hrací ploše. Test věrohodnostním poměrem ukázal, že v profesionálním fotbale nehraje roli, jestlo hráč zakončuje 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] ANDĚL, Jiří. 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 official 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/