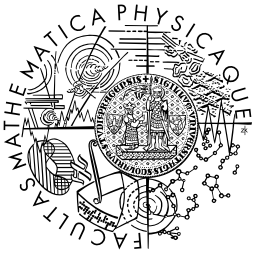


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

František Kožnar

Srovnání modelů pravděpodobností ve fotbalovém sázení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	RNDr. Jan Večeř, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2020

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Rád bych poděkoval profesoru Janu Večeřovi za cenné rady, věcné připomínky a vstřícnost při konzultacích a vypracování bakalářské práce.

Název práce: Srovnání modelů pravděpodobností ve fotbalovém sázení

Autor: František Kožnar

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jan Večeř, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Cílem práce je porovnat různé statistické modely pro fotbalové kurzy a předpovědět co nejlépe chování týmu na základě historických výkonů. Jsou zde nejméně dvě možnosti, jak odhadovat pravděpodobnost, a to konkrétně Poissonův model a metoda založená na strojovém učení. Myšlenkou je, že historické výkony týmů jsou dobré pro předpověď následujících zápasů. Můžeme tedy vzít všechny zápasy z celé sezóny Bundesligy (306 zápasů) a využít data pro předpověď pravděpodobností pro další sezónu. Výsledné pravděpodobnosti by měly být porovnány se skutečnými výsledky a určit nejlepší model.

Klíčová slova: sázkové kurzy, regresní analýza, strojové učení

Title: Comparison of Models for Probabilities in Football Betting

Author: František Kožnar

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jan Večeř, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of the thesis is to compare different statistical models for football betting odds and determine the best performing once based on the historical performance of sport teams. There are at least two possible approaches for computing the odds, namely Poisson regression and methods based on statistical machine learning. The idea is that the historical performance of teams is a good predictor of the future performance. Thus we can take the past performances, say all matches in the full season of the Bundesliga (306 matches), and use these data for predicting the odds for the following season. The resulting odds should be compared with the actual results using the scoring rules, which will identify the best performing model.

Keywords: betting odds, regression analysis, machine learning

Obsah

Úvod

Budeme se zabývat fotbalem v Německu, konkrétně sezónou 2015/2016 Bundesligy. Bundesligy se účastní 18 týmů. Během sezóny hraje každý tým s každým týmem přesně dvakrát, jednou v domácím prostředí a podruhé jako hostující. Tedy za celou sezónu se hraje přesně $18 \times 17 = 306$ zápasů. Za výhru jsou uděleny 3 body, za remízu 1 bod a za prohru 0 bodů. V sezóně 2015/2016 FC Bayern Mnichov vyhrál ligu s celkovým počtem 88 bodů. Můžeme si položit otázku, zda si zasloužili vyhrát, nebo jestli šlo jenom o štěstí.

Na jednu stranu si zasloužili vyhrát, protože hráli s každým týmem dvakrát a získali nejvíce bodů. Ale některé týmy jsou velmi vyrovnané, zápasy velmi těsné a konečný výsledek může být v podstatě rozhodnut nešťastnou chybou, náhodným gólem.

Situace je vlastně podobná ruletě. Předpokládejme, že hráč vyhraje sázku na lichá/sudá čísla. Tahle samotná hra nás nepřesvědčí, že hráč má větší šanci na výhru (je to lepší tým) než šance na výhru kasina. V dlouhodobé sérii her se ukáže pravděpodobnost výhry, která je velmi důležitá a favorizuje kasino, a ne hráče. Podobně můžeme uvažovat tým, který si zasloužil vyhrát, jako tým, který má největší pravděpodobnost na výhru celé sezóny. Toto se může lišit od týmu, který skutečně vyhrál.

V práci se podíváme na dva způsoby (Poissonův model a strojové učení), jakým můžeme spočítat pravděpodobnost, že daný tým vyhraje Bundesligu. V obou případech nejdříve budeme uvažovat výsledek jednoho zápasu dvou týmů. Například, když hraje FC Bayern Mnichov, jaká je pravděpodobnost výhry, remízy nebo prohry? Je zřejmé, že tyto pravděpodobnosti závisí proti jakému soupeři FC Bayern Mnichov hraje a také zda se hraje doma nebo venku (určitě je zde mnoho dalších relevantních faktorů, ale budeme je pro zjednodušení ignorovat).

Pokud tedy už máme odhadnuty pravděpodobnosti pro každý možný zápas v lize, už teoreticky můžeme spočítat pravděpodobnost, že daný tým vyhraje celou sezónu. Pro přesný výsledek je tento výpočet příliš složitý, tudíž mnohem jednodušší alternativou bude simulace a následně odhad pravděpodobnosti s libovolnou přesností. V podstatě můžeme simulovat libovolné množství sezón a odhadnout pravděpodobnost vítěze jako podíl simulovaných sezón, které FC Bayern Mnichov vyhrál. Pak týmy můžeme hodnotit podle pravděpodobnosti na výhru celé Bundesligy.

Nedílnou součástí této práce bude porovnat oba modely a určit jaký model nám dává lepší odhad pravděpodobností.

1. Data

K dispozici máme výsledky všech zápasů ze sezón 2011/2012-2015/2016, kde sezóny 2011/2012-2014/2015 budou použity jenom jako trénovací data pro strojové učení. Každý zápas disponuje konečným počtem vstřelených gólů na základě kterých budeme odhadovat pravděpodobnosti v obou modelech popsaných v kapitolách 2 a 3. Každý zápas ještě obsahuje další parametry, jako například průměrná rychlost hráčů, počet rohových kopů a mnoho dalších. Kompletní seznam parametrů je uveden v tabulce 1.2, tyto dodatečné informace budeme využívat až v kapitole 5. Veškeré data, se kterými budeme pracovat, jsou uvedeny v příloze A.1.

V tabulce 1.1 je souhrn statistik pro sezónu 2015/2016, kterou se budeme zabývat.

Tým	Průměr Vstřelených Gólů	Průměr Inkasovaných Gólů	W	D	L	Body
FC Bayern Mnichov	2.35	0.50	28	4	2	88
Borussia Dortmund	2.41	1.00	24	6	4	78
Bayer 04 Leverkusen	1.65	1.18	18	6	10	60
Borussia Mönchengladbach	1.97	1.47	17	4	13	55
FC Schalke 04	1.50	1.44	15	7	12	52
1. FSV Mainz 05	1.35	1.24	14	8	12	50
Hertha BSC	1.24	1.24	14	8	12	50
VfL Wolfsburg	1.38	1.44	12	9	13	45
1. FC Köln	1.12	1.24	10	13	11	43
Hamburger SV	1.18	1.35	11	8	15	41
FC Ingolstadt 04	0.97	1.24	10	10	14	40
SV Darmstadt 98	1.12	1.56	9	11	14	38
FC Augsburg	1.24	1.53	9	11	14	38
SV Werder Bremen	1.47	1.91	10	8	16	38
TSG 1899 Hoffenheim	1.15	1.59	9	10	15	37
Eintracht Frankfurt	1.00	1.53	9	9	16	36
VfB Stuttgart	1.47	2.21	9	6	19	33
Hannover 96	0.91	1.82	7	4	23	25

Tabulka 1.1: Statistiky jednotlivých týmů ze sezóny 2015/2016, Kde W - počet vyhraných zápasů, D - počet remíz a L - počet prohraných zápasů.

Počet žlutých karet
Počet faulů
Počet všech faulů (fauly obou týmů)
Počet ofsajdů
Počet levých rohových kopů
Počet pravých rohových kopů
Počet rohových kopů
Počet hlaviček
Počet střel za první poločas
Počet střel za druhý poločas
Počet střel za zápas
Počet vyhraných soubojů
Počet dobrých přihrávek
Dobré přihrávky procentuálně
Počet špatných přihrávek
Špatné přihrávky procentuálně
Počet kontaktů s míčem
Procentuální kontakt s míčem
Levé centry
Pravé centry
Centry celkem
Intenzivní běh [km]
Sprint [km]
Rychlý běh [km]
Počet intenzivního běhu
Počet sprintů
Počet rychlého běhu
Distance [km]
Průměrná rychlost
Počet červených karet

Tabulka 1.2: Rozšiřující parametry zápasu.

2. Poissonův model

2.1 Určení modelu

Jako první popíšeme Poissonův model. Text je založen na článku od A. J. Leeho (Lee, 1997). Vytvoříme model, kde vstup budou tvořit dva týmy. Řekněme například, že hraje v domácím prostředí FC Bayern Mnichov proti Borussia Dortmundu. Dále budeme předpokládat, že počet gólů vstřelených domácím týmem má Poissonovo rozdělení s parametrem λ_H (H z anglického home), podobně pro hostující tým předpokládáme Poissonovo rozdělení, ale s jiným parametrem λ_A (A z anglického away). Na základě výpočtu, který je uveden níže, můžeme předpokládat, že skóre dvou týmů je nezávislé. To znamená, že počet gólů vstřelených domácím týmem neovlivňuje rozdělení hostujícího týmu, tedy ani jeho skóre.

Poslední předpoklad může znít trochu nepřirozeně. Proto vytvoříme tabulku pro domácí a hostující skóre ze všech 306 her (ne pouze pro FC Bayern Mnichov a Borussia Dortmund), dostaneme následující tabulku:

		Skóre domácích				
		0	1	2	3	4+
	0	24	20	22	13	10
Skóre	1	25	31	29	15	10
hostujících	2	13	17	11	10	5
	3	8	23	3	5	1
	4+	4	5	2	0	0

Tabulka 2.1: Počet zápasů podle vstřelených gólů domácích a hostujících.

Standardní χ^2 - test ukazuje, že data nevyvracejí předpoklad nezávislosti ($\chi^2 = 24.4$ o 16 stupních volnosti, p - hodnota $= 0.081$). V souladu s tím, budeme předpokládat nezávislost modelů.

V dalším kroku se budeme zabývat otázkou, na kterých faktorech by měl záviset parametr λ. Budeme uvažovat následující faktory:

- Jak silný útok má domácí tým? Například pravděpodobně můžeme předpokládat, že FC Bayern Mnichov bude silnější než Eintracht Frankfurt, který je na konci tabulky.
- Jak silná je obrana hostujícího týmu? Kvalitní oponent zabrání domácímu týmu hodně skórovat.
- Jak důležitá je domácí výhoda?

Můžeme se podívat, jak tyto faktory ovlivňují skóre týmu proti soupeři pomocí regrese, která využívá průměrné skóre všech týmů, schopnost útoku, schopnost obrany a výhodu domácího prostředí. Tento model se nazývá Poissonův, který je speciálním případem obecného lineárního modelu.

2.2 Využití Poissonova rozdělení k předpovědi skóre

Budeme předpokládat, že skóre X jednotlivého týmu v jednotlivém zápase má Poissonovo rozdělení:

$$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Chceme, aby střední hodnota λ tohoto rozdělení ukazovala sílu týmu, kvalitu protihráče a domácí výhodu. Vyjádříme tedy logaritmus střední hodnoty jako lineární kombinaci několika faktorů. Naše rovnice pro logaritmickou střední hodnotu pro domácí tým je (řekněme, že FC Bayern Mnichov hraje proti Borussia Dortmundu v domácím prostředí):

$$log(\lambda_H) = \beta + \beta_H + \beta_O(FC\ Bayern\ Mnichov) + \beta_D(Borussia\ Dortmund).$$

Podobně dostaneme rovnici pro Borussia Dortmund, který hraje jako hostující tým:

$$log(\lambda_A) = \beta + \beta_O(Borussia\ Dortmund) + \beta_D(FC\ Bayern\ Mnichov).$$

Máme tedy vyjádřené skóre λ_H a λ_A pomocí parametrů. Konstantní parametr β vyjadřuje průměrný počet vstřelených gólů v jedné hře, β_H vyjadřuje domácí výhodu. Nakonec následují parametry β_O a β_D, které po řadě vyjadřují sílu útoku a sílu obrany (O z anglického offence, D z anglického deffence). Parametr obrany β_D se taktéž přičítá, protože hodnoty β_D mohou nabývat záporných hodnot.

2.3 Určení parametrů metodou maximální věrohodnosti

Za předpokladu nezávislosti jednotlivých zápasů můžeme vypočítat parametry pomocí metody maximálně věrohodného odhadu.

Předpokládejme, že máme týmy T_i $i \in \{1,...,N\}$, kde $N = 18$ je počet týmů. Pak pravděpodobnost, že tým T_i dá x gólů a T_j dá y gólů, $i \neq j$, je:

$$\frac{e^{-\lambda_H}\lambda_H^x}{x!} \times \frac{e^{-\lambda_A}\lambda_A^y}{y!}, \tag{2.1}$$

kde λ_H a λ_A jsou:

$$\lambda_H = exp\left\{\beta + \beta_H + \beta_O(T_i) + \beta_D(T_j)\right\},$$
$$\lambda_A = exp\left\{\beta + \beta_O(T_j) + \beta_D(T_i)\right\}.$$

Tedy sdruženou hustotu můžeme vyjádřit jako:

$$L(X,Y,\boldsymbol{\theta}) = \prod_{k=1}^{n} \frac{e^{-\lambda_H}\lambda_H^{x_k}}{x_k!} \times \frac{e^{-\lambda_A}\lambda_A^{y_k}}{y_k!},$$

kde $n = 306$ je počet všech her všech týmů, λ_H a λ_A odpovídá týmu, který hrál k-tý zápas v domácím respektive hostujícím prostředí. Dále $X = (x_1,...,x_n)$,

$Y = (y_1,...,y_n)$, kde x_k a y_k značí počet vstřelených gólů domácím respektive hostujícím týmem v k-tém zápase a $\boldsymbol{\theta}$ je vektor všech neznámých parametrů:

$$\boldsymbol{\theta} = [\beta,\beta_H, \beta_O(T_1),...,\beta_O(T_N),\beta_D(T_1),...,\beta_D(T_N)]^T .$$

Dále ještě upravíme na logaritmickou sdruženou hustotu:

$$\begin{aligned} l(X,Y,\boldsymbol{\theta}) &= ln\left[L(X,Y,\boldsymbol{\theta})\right] = \\ &= -\sum_{k=1}^{n}(\lambda_H + \lambda_A) + \sum_{k=1}^{n}\left[x_k ln(\lambda_H) + y_k ln(\lambda_A)\right] - \sum_{k=1}^{n}\left[ln(x_k!) + ln(y_k!)\right]. \end{aligned}$$

Zavedeme ještě značení pro množiny zápasů odehrané týmem T_j:

M_j^H - Zápasy odehrané týmem T_j, hrající doma,
M_j^A - Zápasy odehrané týmem T_j, hrající jako hosté.

Nyní chceme maximalizovat logaritmickou sdruženou hustotu $l(X,Y,\boldsymbol{\theta})$, podíváme se proto na parciální derivace a položíme rovno nule:

$$\frac{\partial l(X,Y,\boldsymbol{\theta})}{\partial \beta} = -\sum_{k=1}^{n}(\lambda_H + \lambda_A) + \sum_{k=1}^{n}(x_k + y_k) = 0,$$

$$\frac{\partial l(X,Y,\boldsymbol{\theta})}{\partial \beta_H} = -\sum_{k=1}^{n}\lambda_H + \sum_{k=1}^{n}x_k = 0,$$

$$\frac{\partial l(X,Y,\boldsymbol{\theta})}{\partial \beta_O(T_j)} = \sum_{k=1}^{n}(x_k - \lambda_H)\mathbb{I}\{k \in M_j^H\} + \sum_{k=1}^{n}(y_k - \lambda_A)\mathbb{I}\{k \in M_j^A\} = 0,$$

$$\frac{\partial l(X,Y,\boldsymbol{\theta})}{\partial \beta_D(T_j)} = \sum_{k=1}^{n}(x_k - \lambda_H)\mathbb{I}\{k \in M_j^A\} + \sum_{k=1}^{n}(y_k - \lambda_A)\mathbb{I}\{k \in M_j^H\} = 0.$$

Tato soustava rovnic se nedá vyřešit explicitně, takže ji budeme muset vyřešit numericky. V příloze A.1 je uveden výpočet v Excelu pomocí nástroje Solver. V tabulce 2.2 můžeme vidět vypočtené parametry.

Tým	Parametr Útoku	Parametr Obrany
FC Bayern Mnichov	0.52	-0.96
Bayer 04 Leverkusen	0.19	-0.13
Borussia Dortmund	0.57	-0.26
SV Darmstadt 98	-0.18	0.13
1. FSV Mainz 05	-0.01	-0.09
FC Augsburg	-0.08	0.12
SV Werder Bremen	0.11	0.35
VfB Stuttgart	0.12	0.50
VfL Wolfsburg	0.02	0.06
Hertha BSC	-0.10	-0.10
Hamburger SV	-0.14	-0.01
1. FC Köln	-0.20	-0.10
Eintracht Frankfurt	-0.30	0.11
TSG 1899 Hoffenheim	-0.16	0.15
FC Schalke 04	0.11	0.07
Hannover 96	-0.38	0.28
FC Ingolstadt 04	-0.34	-0.11
Borussia Mönchengladbach	0.38	0.11
β		0.14
β_H		0.21

Tabulka 2.2: Výsledky vypočtených parametrů pro jednotlivé týmy a parametry β a β_H.

Parametr $\beta \doteq 0.1417$ nám říká, že typický hostující tým vstřelí $e^{0.1417} \doteq 1.152$ gólů. Parametr $\beta_H \doteq 0.2132$ nám říká, že se dá od domácího týmu očekávat $100 \times e^{0.2132} \doteq 123\%$ více gólů než vstřelených hostujícím týmem. To souhlasí s předchozím hrubým odhadem; 1.57 (průměr gólů vstřelených domácím týmem) je přibližně 123% z 1.26 (průměr gólů vstřelených hostujícím týmem).

Nyní se budeme zabývat parametry λ_H a λ_A. Vidíme, že nejvyšší hodnotu útočného parametru má Borussia Dortmund (0.5651) a FC Bayern Mnichov má nejnižší hodnotu obranného parametru (-0.9563).

Co je výhodou tohoto modelu než jenom jednoduché průměry? Model zohledňuje sílu útoku a obrany obou týmů, můžeme spočítat šanci výsledku zápasu a navíc můžeme vypočítat pravděpodobnost výhry, prohry a remízy.

Tedy, pokud hraje FC Bayern Mnichov v domácím prostředí proti Borussia Dortmundu, tak pravděpodobnost, že FC Bayern Mnichov vstřelí x gólů a Borussia Dortmund vstřelí y gólů, je jak už víme z (2.1):

$$\frac{e^{-\lambda_H}\lambda_H^x}{x!} \times \frac{e^{-\lambda_A}\lambda_A^y}{y!},$$

kde λ_H a λ_A jsou:

$$\lambda_H \doteq exp\{0.1417 + 0.2132 + 0.5192 - 0.2586\} \doteq 1.8505,$$
$$\lambda_A \doteq exp\{0.1417 + 0.5651 - 0.9563\} \doteq 0.7791.$$

Tedy tyto hodnoty nám říkají, že střední hodnota vstřelených gólů FC Bayern Mnichov je 1.8505 a Borussia Dortmund 0.7791. Pro výpočet výhry domácího týmu nám stačí sečíst všechny kombinace, kde $x > y$, pravděpodobnost remízy $x = y$ a pravděpodobnost prohry domácích $x < y$. Tedy:

$$P(\text{Výhra domácích}) = \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} \frac{e^{-\lambda_H} \lambda_H^x}{x!} \times \frac{e^{-\lambda_A} \lambda_A^y}{y!},$$

$$P(\text{Remíza}) = \sum_{x=0}^{\infty} \frac{e^{-\lambda_H} \lambda_H^x}{x!} \times \frac{e^{-\lambda_A} \lambda_A^x}{x!},$$

$$P(\text{Prohra domácích}) = \sum_{x=0}^{\infty} \sum_{y=x+1}^{\infty} \frac{e^{-\lambda_H} \lambda_H^x}{x!} \times \frac{e^{-\lambda_A} \lambda_A^y}{y!}.$$

V tabulce 2.3 můžeme vidět pravděpodobnosti výhry, prohry a remízy domácích pro několik vybraných týmů.

Domácí Tým	Hostující Tým	Výhra	Remíza	Prohra
FC Bayern Mnichov	Borussia Dortmund	0.628	0.220	0.152
Borussia Dortmund	FC Bayern Mnichov	0.244	0.261	0.495
FC Bayern Mnichov	Hannover 96	0.915	0.069	0.017
Hannover 96	FC Bayern Mnichov	0.037	0.117	0.846
Borussia Dortmund	Hannover 96	0.878	0.085	0.037
Hannover 96	Borussia Dortmund	0.080	0.138	0.781
1.FSV Mainz 05	FC Augsburg	0.521	0.250	0.229
FC Augsburg	1.FSV Mainz 05	0.342	0.271	0.387
VfB Stuttgart	Hertha BSC	0.328	0.233	0.439
Hertha BSC	VfB Stuttgart	0.590	0.205	0.204

Tabulka 2.3: Pravděpodobnosti výhry, prohry a remízy domácích pro několik vybraných dvojic týmů na základě Poissonova modelu.

2.4 Simulování sezóny

Teď můžeme spekulovat, zda FC Bayern Mnichov měl pouze štěstí v sezóně. Jak už bylo zmíněno, Poissonův model nám umožňuje spočítat pravděpodobnost výhry, prohry a remízy. Tohle nám v podstatě umožňuje spočítat celkové umístění v lize. Přesný výpočet by byl příliš složitý, tedy provedeme jenom simulaci. Zdrojový kód programu pro simulaci je uveden v příloze A.2.
Pro každých 306 odehraných zápasů budeme simulovat výsledek. Počet simulovaných sezón je 1 000 000.

Tým	Získané Body	Očekávané Body	Pravděpodobnost Výhry [%]
FC Bayern Mnichov	88	84.38	88.20
Borussia Dortmund	78	74.20	11.58
Bayer 04 Leverkusen	60	57.42	0.11
Borussia Mönchengladbach	55	57.56	0.10
1. FSV Mainz 05	50	49.75	0.00
FC Schalke 04	52	48.77	0.00
VfL Wolfsburg	45	46.18	0.00
Hertha BSC	50	46.97	0.00
1. FC Köln	43	44.07	0.00
Hamburger SV	41	43.06	0.00
FC Augsburg	38	41.06	0.00
SV Darmstadt 98	38	37.69	0.00
SV Werder Bremen	38	39.61	0.00
VfB Stuttgart	33	35.04	0.00
Eintracht Frankfurt	36	35.31	0.00
TSG 1899 Hoffenheim	37	37.89	0.00
Hannover 96	25	28.25	0.00
FC Ingolstadt 04	40	40.22	0.00

Tabulka 2.4: Pravděpodobnost výhry a očekávané body celé sezóny pro jednotlivé týmy pomocí simulace.

Na základě tabulky 2.4 vidíme, že FC Bayern Mnichov měl jednoznačně nejvyšší pravděpodobnost výhry celé sezóny.

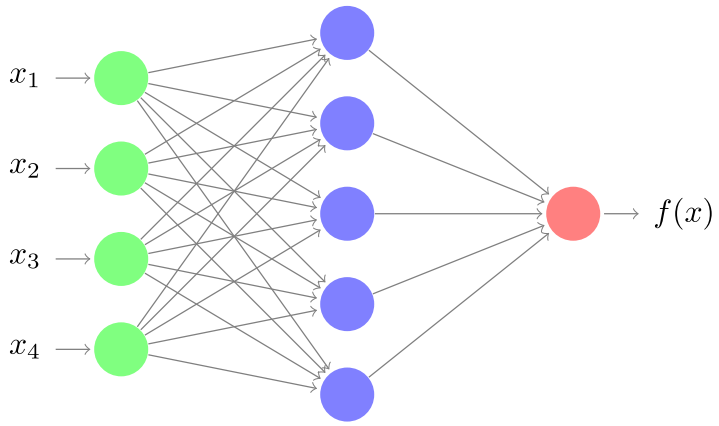
3. Strojové učení

Nyní se podíváme na druhou metodu jak počítat pravděpodobnosti a to pomocí strojového učení. Teorie je převzatá z knih od Müllera, Guida (Müller a Guido, 2016) a Efrona, Hastie (Efron a Hastie, 2016). Nejdříve popíšeme potřebnou teorii a potom ji použijeme na naše data, která jsou stejná jako v předchozí kapitole (tj. k výpočtu pravděpodobností využijeme jenom počty gólů vstřelených v jednotlivých zápasech).

Strojové učení je druh umělé inteligence, která je stále více využívaná během posledních několika let. Pro mnoho problémů je totiž téměř nemožné navrhnout pravidla, která by pomohla najít řešení například kvůli omezení nebo složitosti, pro tyto případy se právě využívá strojové učení. Nejvíce využívaný typ algoritmu pro strojové učení jsou takové, které automatizují rozhodovací proces na základě známých dat. Uživatel tedy musí poskytnou vstup s příslušnými výstupy a algoritmus najde příslušný výstup pro zadaná data. Konkrétně algoritmus je schopný vytvořit výsledek na základě vstupu, který nikdy předtím neviděl, bez jakékoli pomoci. Tento typ strojového učení se nazývá učení s učitelem (supervised learning), druhým typem strojového učení se nazývá učení bez učitele (unsupervised learning), kde ke vstupním datům není známý výstup. V této kapitole popíšeme více do hloubky učení s učitelem a pokusíme se teorii použít na naše data a předpovědět pravděpodobnost výhry, remízy a prohry.

3.1 Neuronové sítě

Umělé neuronové sítě ANNs (z anglického Artificial neural networks) je výpočetní technika založená na chování biologických struktur. Lidský mozek se skládá z nervových buněk nazývané neurony, které jsou vzájemně propojeny axony. ANNs se skládají z uzlů, které napodobují biologické neurony. Uzly jsou spojeny přenosovou funkcí, která napodobuje axony. Každý uzel přijímá vstupní data a provádí jednoduchou operaci a výsledek posílá do dalších uzlů.



Obrázek 3.1: Ukázka neuronové sítě se čtyřmi vstupy a jednou skrytou vrstvou.

Na obrázku 3.1 je jednoduchý příklad neuronové sítě. Jsou tam čtyři vstupy

x_j, jedna skrytá vrstva s pěti uzly $a_l = g(w_{l0}^{(1)} + \sum_{j=1}^{4} w_{lj}^{(1)} x_j)$ a jeden výstup $o = h(w_0^{(2)} + \sum_{l=1}^{5} w_l^{(2)} a_l)$. Každý neuron a_j je spojen se vstupní vrstvou pomocí vektorem váhových parametrů $\left\{w_{lj}^{(1)}\right\}_1^p$ ((1) odpovídá první vrstvě, lj odpovídá j-té proměnné u l-tého neuronu). Velikost vah $w_{lj}^{(1)}$ vyjadřuje uložení zkušeností do neuronu. Čím je vyšší hodnota, tím je daný vstup důležitější. Funkce g je nelineární, často využívaná je sigmoidální přenosová funkce $g(t) = 1/(1 + e^{-t})$. Myšlenkou je, aby se každý neuron naučil odpovídat ANO/NE; kompromisem je naše sigmoidální přenosová funkce, která je hladká a diferencovatelná. Poslední výstupní vrstva má taky váhy a přenosovou funkci h. Pro kvantitativní regresi je běžně využívaná identická funkce a pro binární výstup zase sigmoidální přenosová funkce. Je důležité si uvědomit, že pokud vynecháme skryté vrstvy, tak neuronové sítě jsou vlastně obecný lineární model. Pro určení vah využijeme teorii maximální věrohodnosti. Nejdříve ale zavedeme následující značení.

Z první vrstvy L_1 do druhé L_2 označíme:

$$z_l^{(2)} = w_{l0}^{(1)} + \sum_{j=1}^{p} w_{lj}^{(1)} x_j,$$
$$a_j^{(2)} = g^{(2)}(z_l^{(2)}).$$

Obecně můžeme označit z vrstvy L_{k-1} do vrstvy L_k:

$$z_l^{(k)} = w_{l0}^{(k-1)} + \sum_{j=1}^{p_{k-1}} w_{lj}^{(k-1)} a_j^{(k-1)},$$
$$a_j^{(k)} = g^{(k)}(z_l^{(k)}).$$

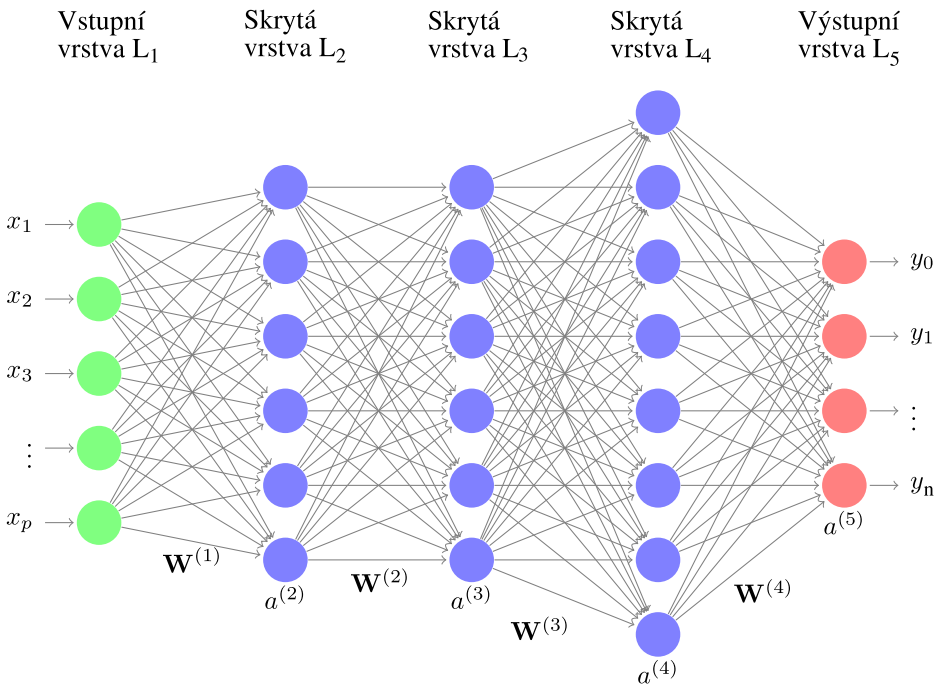
V tomto případě značení odpovídá $a_l^{(1)} = x_l$ a $p_1 = p$, což je počet vstupních parametrů. Pro zjednodušení budeme používat vektorový zápis:

$$z^{(k)} = \mathbf{W}^{(k-1)} a^{(k-1)},$$
$$a^{(k)} = g^{(k)}(z^{(k)}),$$

kde $\mathbf{W}^{(k-1)}$ je matice vah z vrstvy L_{k-1} do vrstvy L_k, $a^{(k)}$ je vektor uzlů ve vrstvě L_k. Parametry $w_{l0}^{(k-1)}$ jsou taktéž uloženy v matici $\mathbf{W}^{(k-1)}$, tudíž tedy předpokládáme, že jsme rozšířili každý vektor $a^{(k)}$ o konstantu 1. Funkce $g^{(k)}$ jsou většinou stejné, ale nutno zmínit, že to tak být nemusí.

Finální funkce $g^{(K)}$ (kde K je počet vrstev) je většinou speciální. Například pro M klasifikační problém je typicky používaná $softmax$ funkce, která počítá pravděpodobnost pro každou třídu (v našem případě pokud zvolíme tři výstupní třídy jako $Výhra, Remíza$ a $Prohra$, tak funkce $softmax$ je přesně to co potřebujeme)

$$g^{(K)}(z_m^{(K)}, z^{(K)}) = \frac{e^{z_m^{(K)}}}{\sum_{l=1}^{M} e^{z_l^{(K)}}}. \tag{3.1}$$



Obrázek 3.2: Ukázka neuronové sítě s p vstupy, čtyřmi skrytými vrstvy a n výstupy. Na ukázku jsou zobrazeny váhy $W^{(k)}$ a uzly $a^{(k)}$.

3.2 Odhad parametrů neuronové sítě

Protože všechny vrstvy jsou funkcí předešlých vrstev a tedy také funkcí vstupního vektoru x, tak síť představuje obecnou funkci $f(x,\mathcal{W})$, kde $\mathcal{W}$ jsou všechny váhy. Pro dvojici dat, které známe, $\{x_i, y_i\}_{i=1}^n$ a ztrátovou funkci $L\left[y,f(x)\right]$ řešíme následující úlohu:

$$min_{\mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L\left[y_i,f(x_i,\mathcal{W})\right] + \lambda J(\mathcal{W}) \right\}, \tag{3.2}$$

kde $\lambda \geq 0$ a $J(\mathcal{W})$ je nezáporná regulace vah $\mathcal{W}$ (v praxi může být více regulací s vlastní λ). Příklad regulace J:

$$J(\mathcal{W}) = \frac{1}{2} \sum_{k=1}^{K-1} \sum_{j=1}^{p_k} \sum_{l=1}^{p_{k+1}} \{w_{lj}^{(k)}\}^2.$$

Tato regulace penalizuje váhy a přibližuje je k 0 a typicky váhy $w_{l0}^{(k)}$ nejsou penalizovány. Ztrátová funkce je obvykle konvexní v f, ale ne na prvcích $\mathcal{W}$, proto řešení 3.2 je složité.

3.3 Výpočet gradientu

Existuje mnoho metod pro výpočet gradientu a mnoho metod je založena na základě největšího spádu. Zde si uvedeme jednu metodu. Budeme uvažovat výpočet derivací funkce $L\left[y,f(x,\mathcal{W})\right]$ podle všech prvků $\mathcal{W}$ pro obecnou dvojici

vstupu $\{x_i, y_i\}$. Protože funkce daná v (3.2) je tvořena sumou, výsledný gradient bude součtem jednotlivých gradientů ztrátové funkce.

Myšlenka výpočtu je následující. Nejdříve se spočítáme uzly $a_l^{(k)}$ ve všech vrstvách včetně výstupní vrstvy. Poté budeme chtít pro každý uzel $a_l^{(k)}$ spočítat $\delta_l^{(k)}$, který měří rozdíl mezi predikcí a správným výstupem y. Pro $a_l^{(K)}$ je to jednoduché, protože to je ztrátová funkce. Pro ostatní $\delta_l^{(k)}$ to bude vážená suma residuí, kde uzly $a_l^{(k)}$ jsou brány jako vstup.

Algoritmus.

1. Pro daný pár $\{x_i, y_i\}$ provedeme výpočet všech $a_l^{(k)}$ ve všech vrstvách L_2,...,L_K. To znamená, že vypočteme $f(x,\mathcal{W})$ v bodě x pro daný $\mathcal{W}$ a uložíme všechny hodnoty.

2. Pro každý výstup l ve vrstvě L_K spočítáme:

$$\delta_l^{(K)} = \frac{\partial L\,[x,f(x,\mathcal{W})]}{\partial z_l^{(K)}} = \frac{\partial L\,[x,f(x,\mathcal{W})]}{\partial a_l^{(K)}} \dot{g}^{(K)}(z_l^{(K)}), \tag{3.3}$$

   kde $\dot{g}$ značí derivaci $g(z)$ podle z. Například pro $L\,[y,f] = \frac{1}{2}\|y-f\|_2^2$ potom z 3.3 dostaneme $-(y_l - f_l)\dot{g}^{(K)}(z_l^{(K)})$.

3. Pro vrstvy $k = K-1, K-2$,...,2 a pro každý uzel l ve vrstvě k položíme:

$$\delta_l^{(k)} = \left(\sum_{j=1}^{p_{k+1}} w_{jl}^{(k)} \delta_j^{(k+1)} \right) \dot{g}^{(k)}(z_l^{(k)}).$$

4. Potom parciální derivace jsou dány vztahem:

$$\frac{\partial L\,[x,f(x,\mathcal{W})]}{\partial w_{lj}^{(k)}} = a_j^{(k)} \delta_l^{(k+1)}.$$

3.4 Využití strojového učení pro výpočet pravděpodobností

Nyní použijeme vytvořenou teorii na naše data. Jako trénovací data použijeme čtyři předchozí sezóny a naším vstupem bude 8 parametrů:

x_1 = Průměr vstřelených gólů v domácím prostředí,
x_2 = Průměr vstřelených gólů hrající jako hosté,
x_3 = Průměr inkasovaných gólů v domácím prostředí,
x_4 = Průměr inkasovaných gólů hrající jako hosté,
x_5 = Průměr vstřelených gólů v domácím prostředí,
x_6 = Průměr vstřelených gólů hrající jako hosté,
x_7 = Průměr inkasovaných gólů v domácím prostředí,
x_8 = Průměr inkasovaných gólů hrající jako hosté.

Parametry x_1,...,x_4 patří týmu hrající v domácím prostředí a parametry x_5,...,x_8 patří týmu hrající jako hosté. Výstup bude tvořit pravděpodobnost výhry, prohry a remízy domácích, kde k výpočtu využíváme (3.1). V tabulce 3.1 můžeme vidět obdobu tabulky 2.3 kde jsou uvedeny pravděpodobnosti výhry, remízy a prohry domácích.

Domácí Tým	Hostující Tým	Výhra	Remíza	Prohra
FC Bayern Mnichov	Borussia Dortmund	0.755	0.108	0.137
Borussia Dortmund	FC Bayern Mnichov	0.580	0.106	0.314
FC Bayern Mnichov	Hannover 96	0.827	0.153	0.020
Hannover 96	FC Bayern Mnichov	0.048	0.211	0.741
Borussia Dortmund	Hannover 96	0.813	0.153	0.034
Hannover 96	Borussia Dortmund	0.069	0.222	0.708
1.FSV Mainz 05	FC Augsburg	0.359	0.307	0.334
FC Augsburg	1.FSV Mainz 05	0.190	0.263	0.546
VfB Stuttgart	Hertha BSC	0.282	0.234	0.484
Hertha BSC	VfB Stuttgart	0.584	0.288	0.128

Tabulka 3.1: Pravděpodobnost výhry, prohry a remízy pro několik vybraných dvojic týmů na základě strojového učení.

Podobně jako v předchozí kapitole můžeme simulovat průběh celé sezóny a spočítat pravděpodobnost výhry celé sezóny a odpovídající očekávané body. Pro každých 306 zápasů budeme simulovat výsledek. Počet simulovaných sezón je 1000000.

Tým	Získané Body	Očekávané Body	Pravděpodobnost Výhry [%]
FC Bayern Mnichov	88	80.11	73.69
Borussia Dortmund	78	74.69	25.64
Bayer 04 Leverkusen	60	57.72	0.35
Borussia Mönchengladbach	55	58.00	0.28
1. FSV Mainz 05	50	49.87	0.02
FC Schalke 04	52	48.17	0.01
Hertha BSC	50	48.06	0.01
VfL Wolfsburg	45	47.12	0.00
1. FC Köln	43	44.43	0.00
Hamburger SV	41	42.74	0.00
FC Ingolstadt 04	40	42.16	0.00
FC Augsburg	38	39.62	0.00
SV Darmstadt 98	38	36.49	0.00
SV Werder Bremen	38	37.39	0.00
VfB Stuttgart	33	31.28	0.00
Eintracht Frankfurt	36	36.39	0.00
TSG 1899 Hoffenheim	37	37.74	0.00
Hannover 96	25	28.37	0.00

Tabulka 3.2: Pravděpodobnost výhry a očekávané body celé sezóny pro jednotlivé týmy podle strojového učení.

4. Srovnání

4.1 Motivace

V této kapitole popíšeme metodu, jakým způsobem proti sobě porovnat modely pravděpodobností a rozhodnout, jaký model je lepší. Následující text je založen na článcích (Večeř (2018)) a (Večeř (2020)). Myšlenka je taková, že pokud máme dva různé odhady pravděpodobností, otevírá se možnost porovnat je pomocí hypotetických sázek. Popíšeme jakým způsobem konstruovat sázky, kde sázející maximalizuje nějakou užitkovou funkci, která určuje optimální chování sázejícího při daných pravděpodobnostech. Nejdříve popíšeme obecnou teorii a následně ji použijeme pro náš specifický případ, kde jsou možné tři výstupy: výhra, remíza a prohra domácích.

Sázející, každý se svým různým odhadem pravděpodobností, tvoří sázkový trh. Sázkový trh bude tvořit za určitých podmínek equilibrium (rovnováhu), kde se sázející budou obírat navzájem. U lepšího modelu lze očekávat, že bude vydělávat na úkor horšího modelu. Ukážeme, že realizovaný výnos jednotlivých sázejících vede k Bayesovskému updatu, což je známý výsledek z Bayesovské statistiky. Nejdříve popíšeme optimální chování jednoho sázejícího, najdeme tedy analytické vyjádření optimální velikosti sázek při daných kurzech. Následně na to navážeme s přidáním dalších sázejících s jejich odhady pravděpodobností. Výsledkem sázky bude výnos P_k, který odpovídá výhře sázejícího, pokud nastane výstup x_k pro $k \in \{1,...,n\}$. K výpočtu budeme využívat užitkovou funkci $U(x)$, která je rostoucí a konkávní. Obvyklé volby jsou:

$$\text{Logaritmická: } U(x) = log\left(1 + \frac{x}{B}\right),$$
$$\text{Exponenciální: } U(x) = 1 - exp\left(-\frac{x}{B}\right),$$
$$\text{Mocninná: } U(x) = \frac{\left(1 + \frac{x}{B}\right)^{1-a} - 1}{1 - a}, a > 0.$$

B je volný parametr a je interpretován jako bankroll, to znamená, že ztráta sázejícího nemůže přesáhnout B. V našem případě budeme používat logaritmickou užitkovou funkci, která právě vede k Bayesovskému updatu.

4.2 Optimální sázky jednoho sázejícího

Začneme s diskrétní náhodnou veličinu X, která může nabývat n hodnot $x_1,...,x_n$. Rozdělení náhodné veličiny X z pohledu prvního sázejícího je:

$$\begin{aligned} p_1 &= \mathbb{P}(X = x_1), \\ p_2 &= \mathbb{P}(X = x_2), \\ &\vdots \\ p_n &= \mathbb{P}(X = x_n). \end{aligned}$$

Hypotetický trh vyplatí $A_j = \mathbb{I}(X = x_j)$ pro $j = 1,2,...,n$. To znamená, že sázkový trh vyplatí jednotku, pokud $X = x_j$ a nic jinak. Uvažujme, že sázkový trh nabízí ceny q_j, tedy z pohledu sázkového trhu má X rozdělení:

$$\begin{aligned} q_1 &= \mathbb{Q}(X = x_1), \\ q_2 &= \mathbb{Q}(X = x_2), \\ &\vdots \\ q_n &= \mathbb{Q}(X = x_n). \end{aligned}$$

Buď V_j velikost sázky vsazená na j-tý výstup. Celkový výnos je pak reprezentovaný náhodnou veličinou:

$$P = \sum_{j=1}^{n} V_j(A_j - q_j).$$

Sázející se snaží najít optimální hodnoty sázek V_j, které maximalizují užitkovou funkci výnosu P podle subjektivního pohledu $\mathbb{P}$ sázejícího na rozdělení náhodné veličiny X:

$$\max_{V_j, j\in\{1,2,...,n\}} \mathbb{E}^{\mathbb{P}} U(P) = \max_{V_j, j\in\{1,2,...,n\}} \mathbb{E}^{\mathbb{P}} U\left(\sum_{j=1}^{n} V_j(A_j - q_j)\right).$$

Optimální hodnoty sázek označíme jako $V_j^{(p,q)}$. Všimněme si lineární závislosti:

$$\sum_{j=1}^{n}(A_j - q_j) = 0,$$

tedy hodnoty V_j mají pouze $n - 1$ stupňů volnosti. Dále platí:

$$P = \sum_{j=1}^{n} V_j(A_j - q_j) = \sum_{j=1}^{n}(V_j + C)(A_j - q_j),$$

to znamená, že přičtení nějaké konstanty C ke všem hodnotám V_j nemění realizovaný výnos. Realizovaný výnos P je invariantní vůči posunutí, tedy je více přirozené se zabývat náhodnou veličinou výnosu k-tého výstupu:

$$P_k = V_k - \sum_{j=1}^{n} V_j q_j.$$

V našem případě budeme uvažovat logaritmickou užitkovou funkci $U(x) = log(1+\frac{x}{B})$, sázející tedy maximalizuje:

$$u^{\mathbb{P}}(P) := \mathbb{E}^{\mathbb{P}} U(P) = \sum_{k=1}^{n} p_k \cdot log\left(1 + \frac{P}{B}\right). \tag{4.1}$$

Věta 1. *Maximum očekávaného realizovaného výnosu dané v (4.1) nabývá v:*

$$V_k^{(p,q)} = B \cdot \frac{p_k q_n - p_n q_k}{q_k q_n},$$

optimální výnos $P_k^{(p,q)}$ pro k-tý výstup je daný:

$$P_k^{(p,q)} = B \cdot \left(\frac{p_k}{q_k} - 1\right).$$

Důkaz. Chceme hledat maximum funkce danou vzorcem v (4.1) tedy po rozepsání dostaneme:

$$u^{\mathbb{P}}(P) = \sum_{k=1}^{n} p_k \cdot log\left(1 + \frac{P}{B}\right) = \sum_{k=1}^{n} p_k \cdot log\left(1 + \sum_{j=1}^{n} \frac{V_j}{B}(A_j - q_j)\right).$$

Zderivujeme podle V_m a dosadíme hodnoty $V_k^{(p,q)}$:

$$\begin{aligned}
\frac{\partial u^{\mathbb{P}}(P)}{\partial V_m} &= \sum_{k=1}^{n} p_k \frac{1}{B}\left[A_m - q_m\right] / \left[1 + \sum_{j=1}^{n} \frac{V_j}{B}(A_j - q_j)\right] \\
&= \sum_{k=1}^{n} p_k \frac{1}{B}\left[A_m - q_m\right] / \left[1 + \frac{V_k}{B} - \sum_{j=1}^{n} \frac{V_j}{B} q_j\right] \\
&= \sum_{k=1}^{n} p_k \frac{1}{B}\left[A_m - q_m\right] / \left[1 + \frac{p_k q_n - p_n q_k}{q_k q_n} - \sum_{j=1}^{n} \frac{p_j q_n - p_n q_j}{q_j q_n} q_j\right] \\
&= \sum_{k=1}^{n} p_k \frac{1}{B}\left[A_m - q_m\right] / \left[1 + \frac{p_k q_n - p_n q_k}{q_k q_n} - \left(1 - \frac{p_n}{q_n}\right)\right] \\
&= \sum_{k=1}^{n} p_k \frac{1}{B}\left[A_m - q_m\right] / \left[\frac{p_k}{q_k}\right] \\
&= \sum_{k=1}^{n} q_k \frac{1}{B}\left[A_m - q_m\right] \\
&= \frac{q_m}{B} - \sum_{k=1}^{n} \frac{q_k q_m}{B} = \frac{q_m}{B} - \frac{q_m}{B} = 0.
\end{aligned}$$

Kde využíváme rovnosti $\sum_{k=1}^{n} q_k = 1$. Dostali jsme tedy, že parciální derivace funkce $u^{\mathbb{P}}(P)$ jsou rovny nule. Druhá parciální derivace funkce $u^{\mathbb{P}}(P)$ vypadá následovně:

$$\frac{\partial^2 u^{\mathbb{P}}(P)}{\partial P^2} = -\sum_{k=1}^{n} p_k \frac{1}{(B+P)^2} < 0.$$

Tedy funkce $u^{\mathbb{P}}(P)$ je negativně definitní pro všechny body P. Z negativní definitnosti plyne, že funkce $u^{\mathbb{P}}(P)$ je konkávní a tedy v bodech $V_k^{(p,q)}$ nabývá výnos P globální maximum. Druhá část věty plyne z dosazení a upravení:

$$\begin{aligned}
P_k^{(p,q)} &= V_k - \sum_{j=1}^{n} V_j q_j \\
&= B \cdot \frac{p_k q_n - p_n q_k}{q_k q_n} - \sum_{j=1}^{n} B \cdot \frac{p_j q_n - p_n q_j}{q_j q_n} q_j \\
&= B \cdot \frac{p_k q_n - p_n q_k}{q_k q_n} - \sum_{j=1}^{n} B \cdot \left(p_j - \frac{p_n}{q_n} q_j\right) \\
&= B \cdot \frac{p_k q_n - p_n q_k}{q_k q_n} - B \cdot \left(1 - \frac{p_n}{q_n}\right) \\
&= B \cdot \left(\frac{p_k}{q_k} - 1\right).
\end{aligned}$$

□

4.3 Equilibrium

Uvažujme trh s $u \in \{1,2,...,N\}$ sázejícími. Každý sázející maximalizuje svojí logaritmickou užitkovou funkci s bankrollem B^u a svým názorem p^u na rozdělení náhodné veličiny X. Trh nabývá equilibrium (rovnováhu) m, jestliže celkový realizovaný výnos pro každý výstup x_k $k \in \{1,...,n\}$ splňuje:

$$\sum_{u=1}^{N} P_k^{(p^u,m)} = 0.$$

Věta 2. *Rozdělení m je dáno vztahem:*

$$m_k = \sum_{u=1}^{N} \left(\frac{B^u}{B^M}\right) \cdot p_k^u \qquad k \in \{1,...,n\},$$

kde

$$B^M = \sum_{u=1}^{N} B^u$$

je celkový bankroll trhu.

Důkaz.

$$\begin{aligned}\sum_{u=1}^{N} P_k^{(p^u,m)} &= \sum_{u=1}^{N} B^u \left(\frac{p_k^u}{m_k} - 1\right) = \frac{1}{m_k} \sum_{u=1}^{N} (B^u p_k^u) - B^M \\ &= \frac{B^M}{m_k} \sum_{u=1}^{N} \left(\frac{B^u}{B^M} p_k^u\right) - B^M = \frac{B^M}{m_k} m_k - B^M = 0.\end{aligned}$$

□

Rozdělení m je vlastně vážený průměr jednotlivých rozdělení p^u s vahami B^u. Uvažujme nyní specifický případ, kdy máme $N = 2$ sázející s bankrollem $B^1 = B^2$. Equilibrium je jednoduše:

$$m = \frac{1}{2}(p^1 + p^2)$$

a výsledný výnos je dán vztahem:

$$P_k^{(p^1,m)} = B^1 \cdot \left(\frac{p_k^1}{m_k} - 1\right) = B^1 \cdot \frac{p_k^1 - p_k^2}{p_k^1 + p_k^2}. \tag{4.2}$$

Všimněme si, že $P_k^{(p^1,m)} = -P_k^{(p^2,m)}$, to znamená, že ztráta z pohledu druhého sázejícího je zisk z pohledu prvního sázejícího.

4.4 Bayesovský update

Ve statistice je použití Bayesova faktoru alternativa ke klasickému testování hypotéz. Bayesův faktor popíšeme v následující sekci 4.5. Cílem je na základě naměřených dat vybrat z předem daných modelů ten, který je nejvíce v souladu s daty oproti ostatním modelům. Předpokládejme tedy N různých modelů

a označme je M_u $u \in \{1,...,N\}$. Pravděpodobnost, že model M_u je správný, označíme jako $\mathbb{P}(M_u)$ a platí $\sum_{u=1}^{N} \mathbb{P}(M_u) = 1$. Každý model má různý pohled na pravděpodobnost události $D_k := [X = x_k]$:

$$p_k^u = \mathbb{P}(D_k|M_u) \qquad u \in \{1,...,N\},\ k \in \{1,...,n\}.$$

Každému modelu přiřadíme bankroll $B^u = \mathbb{P}(M_u)$, který zprvu iniciujeme $\frac{1}{N}$, potom $B^M = 1$. Z věty 2 máme:

$$m_k = \sum_{u=1}^{N} \left(\frac{B^u}{B^M}\right) \cdot p_k^u = \sum_{u=1}^{N} \mathbb{P}(D_k|M_u) \cdot \mathbb{P}(M_u) = \mathbb{P}(D_k),$$

což je nepodmíněná pravděpodobnost události D_k. Výnos $P_k^{(p^u,m)}$ je podle věty 1:

$$P_k^{(p^u,m)} = B^u \cdot \left(\frac{p_k^u}{m_k} - 1\right) = \mathbb{P}(M_u) \cdot \left(\frac{\mathbb{P}(D_k|M_u)}{\mathbb{P}(D_k)} - 1\right).$$

Pokud nastane událost D_k, můžeme udělat update bankrollu $B^u = \mathbb{P}(M_u)$ na $\mathbb{P}(M_u|D_k)$, protože:

$$\begin{aligned}\mathbb{P}(M_u) + P_k^{(p^u,m)} &= \mathbb{P}(M_u) + \mathbb{P}(M_u) \cdot \left(\frac{\mathbb{P}(D_k|M_u)}{\mathbb{P}(D_k)} - 1\right) \\ &= \frac{\mathbb{P}(D_k \cap M_u)}{\mathbb{P}(D_k)} = \mathbb{P}(M_u|D_k).\end{aligned}$$

Tento update bankrollu udává pravděpodobnost, že model M_u je správný za předpokladu, že nastala událost D_k. Update bankrollu můžeme sekvenčně opakovat.
Algoritmus.

1. Iniciujeme bankroll modelu M_u jako $\mathbb{P}(M_u) = \frac{1}{N}$.
2. Po první událost D_{k_1} spočítáme bankroll:

$$\mathbb{P}(M_u|D_{k_1}) = \mathbb{P}(M_u) + P_{k_1}^{(p^u,m)}.$$

3. Obecně po r-té události D_{k_r} spočítáme bankroll:

$$\mathbb{P}(M_u|D_{k_1},...,D_{k_r}) = \mathbb{P}(M_u|D_{k_1},...,D_{k_{r-1}}) + P_{k_r}^{(p^u,m)},$$

kde hustota equilibria m je dána vztahem:

$$m_{k_r} = \sum_{u=1}^{N} \mathbb{P}(D_{k_r}|M_u) \cdot \mathbb{P}(M_u|D_{k_1},...,D_{k_{r-1}})$$

a výnos je dán vztahem:

$$P_{k_r}^{(p^u,m)} = \mathbb{P}(M_u|D_{k_1},...,D_{k_{r-1}}) \cdot \left(\frac{p_{k_r}^u}{m_{k_r}} - 1\right).$$

Konečný bankroll $\mathbb{P}(M_u|D_{k_1},...,D_{k_r})$ nám udává jaká je pravděpodobnost, že model M_u je správný, bereme-li v úvahu naměřená data.

Příklad. Uvažujme náhodnou veličinu X, která může nabývat $n = 3$ hodnot $\{1,2,3\}$. Můžeme si X představit jako výsledek zápasu s možnými výsledky: výhra, remíza a prohra domácích. Uvažujme $N = 2$ modely: M_1 a M_2 s distribucemi $\mathbb{P}^1$, respektive $\mathbb{P}^2$ a distribucí equilibria $\mathbb{M}$. Bankroll modelů je $B^1 = B^2 = \frac{1}{2}$.

X	$\mathbb{P}^1$	$\mathbb{P}^2$	$\mathbb{M}$	$P^{(p^1,m)}$	$P^{(p^2,m)}$
0	1/3	1/4	7/24	1/14	-1/14
1	1/3	1/4	7/24	1/14	-1/14
2	1/3	1/2	5/12	-1/10	1/10

Tabulka 4.1: Pravděpodobnosti a výnos.

Očekávaný výnos z pohledu $\mathbb{P}^1$, $\mathbb{P}^2$ a $\mathbb{M}$ je:

$$\mathbb{E}^{\mathbb{P}^1}\left[P^{(p^1,m)}\right] = \mathbb{E}^{\mathbb{P}^2}\left[P^{(p^2,m)}\right] = \frac{1}{70}, \qquad \mathbb{E}^{\mathbb{M}}\left[P^{(p^1,m)}\right] = \mathbb{E}^{\mathbb{M}}\left[P^{(p^2,m)}\right] = 0.$$

Update bankrollu modelu M_1 bude následovný:

$$B^1 = \begin{cases} 4/7 \doteq 0.571, & \text{pokud } X = 0, \\ 4/7 \doteq 0.571, & \text{pokud } X = 1, \\ 2/5 = 0.400 & \text{pokud } X = 2 \end{cases}$$

a update bankrollu modelu M_2 bude:

$$B^2 = \begin{cases} 3/7 \doteq 0.429, & \text{pokud } X = 0, \\ 3/7 \doteq 0.429, & \text{pokud } X = 1, \\ 3/5 = 0.600 & \text{pokud } X = 2. \end{cases}$$

4.5 Interpretace Bayesovských faktorů

Tato sekce je založená na článku od Kass a Raftery (1995). Mějme data $\mathbf{D} = (D_{k_1},...,D_{k_r})$, dva modely M_1 a M_2 a jejich pravděpodobnosti $\mathbb{P}(M_1)$ a $\mathbb{P}(M_2)$, které představují, že model je správný. Před testováním předpokládáme, že tyto pravděpodobnosti jsou stejné, tedy $\mathbb{P}(M_1) = \mathbb{P}(M_2) = \frac{1}{2}$. Podle předchozí sekce nám Bayesovský update vytvoří pravděpodobnosti $\mathbb{P}(M_1|\mathbf{D})$ a $\mathbb{P}(M_2|\mathbf{D})$, která berou v úvahu už naměřená data. Dále platí z Bayesovy věty:

$$\mathbb{P}(M_u|\mathbf{D}) = \frac{\mathbb{P}(\mathbf{D}|M_u)\cdot\mathbb{P}(M_u)}{\mathbb{P}(\mathbf{D})} \qquad u \in \{1,2\},$$

kde $\mathbb{P}(\mathbf{D}|M_u)$ představuje pravděpodobnost, že data $\mathbf{D}$ pocházejí z modelu M_u. Věrohodnost, že data pocházejí z modelu M_1 nebo M_2 je hodnocena Bayesovským faktorem K:

$$K = \frac{\mathbb{P}(\mathbf{D}|M_1)}{\mathbb{P}(\mathbf{D}|M_2)} = \frac{\frac{\mathbb{P}(M_1|\mathbf{D})\cdot\mathbb{P}(\mathbf{D})}{\mathbb{P}(M_1)}}{\frac{\mathbb{P}(M_2|\mathbf{D})\cdot\mathbb{P}(\mathbf{D})}{\mathbb{P}(M_2)}} = \frac{\mathbb{P}(M_1|\mathbf{D})}{\mathbb{P}(M_2|\mathbf{D})}\frac{\mathbb{P}(M_2)}{\mathbb{P}(M_1)} = \frac{\mathbb{P}(M_1|\mathbf{D})}{\mathbb{P}(M_2|\mathbf{D})}. \tag{4.3}$$

Pokud je hodnota $K > 1$, znamená to, že naměřená data jsou více v souladu s modelem M_1 než s modelem M_2. Harold Jeffreys vytvořil škálu pro interpretaci hodnoty K:

K	Síla
<1	Negativní (data podporují M_2)
1 do 3.16	Slabá
3.16 do 10	Podstatná
10 do 31.6	Silná
31.6 do 100	Velmi silná
>100	Rozhodující

Tabulka 4.2: Škála pro interpretaci Bayesovského faktoru K.

4.6 Srovnání Poissonova modelu a Strojového učení

V první části práce jsme odhadli pravděpodobnosti výhry, prohry a remízy pomocí Poissonova modelu pomocí dat, které tvořily vstřelené góly. V druhé části jsme pravděpodobnosti odhadli pomocí strojového učení pomocí stejných dat. Nyní je budeme chtít porovnat a rozhodnout, zda je mezi nimi nějaký statisticky významný rozdíl. V tabulce 4.3 můžeme vidět dva různé odhady pro několik vybraných zápasů.

		Poissonův Model			Strojové Učení		
Domácí Tým	Hostující Tým	W	D	L	W	D	L
FC Bayern Mnichov	Borussia Dortmund	0.628	0.220	0.152	0.755	0.108	0.137
Borussia Dortmund	FC Bayern Mnichov	0.244	0.261	0.495	0.580	0.106	0.314
FC Bayern Mnichov	Hannover 96	0.915	0.069	0.017	0.827	0.153	0.020
Hannover 96	FC Bayern Mnichov	0.037	0.117	0.846	0.048	0.211	0.741
Borussia Dortmund	Hannover 96	0.878	0.085	0.037	0.813	0.153	0.034
Hannover 96	Borussia Dortmund	0.080	0.138	0.781	0.069	0.222	0.708
1.FSV Mainz 05	FC Augsburg	0.521	0.250	0.229	0.359	0.307	0.334
FC Augsburg	1.FSV Mainz 05	0.342	0.271	0.387	0.190	0.263	0.546
VfB Stuttgart	Hertha BSC	0.328	0.233	0.439	0.282	0.234	0.484
Hertha BSC	VfB Stuttgart	0.590	0.205	0.204	0.584	0.288	0.128

Tabulka 4.3: Srovnání pravděpodobností pro Poissonův model a strojové učení, kde W - pravděpodobnost výhry, D - pravděpodobnost remízy a L - pravděpodobnost prohry.

Z tabulky 4.3 vidíme, že některé pravděpodobnosti jsou velmi podobné, ale například ve druhém řádku je pravděpodobnost výhry velmi rozdílná. Nyní už přejdeme k samotnému porovnání Poissonova modelu a strojového učení. Počet zápasů za sezónu je $r = 306$ a dále máme $N = 2$ modely, které označíme jako:

$$M_1 = \text{Strojového učení},$$
$$M_2 = \text{Poissonův model}.$$

K porovnání použijeme Bayesův faktor K daný vzorcem (4.3). Data $\mathbf{D}$ reprezentují výsledky jednotlivých zápasů, náhodná veličina X v tomto konkrétním případě může nabývá $n = 3$ hodnot: výhra, remíza nebo prohra domácích. Samotný výpočet je uveden v příloze A.1. Výsledné pravděpodobnosti jsou následující:

$$\mathbb{P}(M_1|\mathbf{D}) \doteq 0.9999863,$$
$$\mathbb{P}(M_2|\mathbf{D}) \doteq 0.0000137.$$

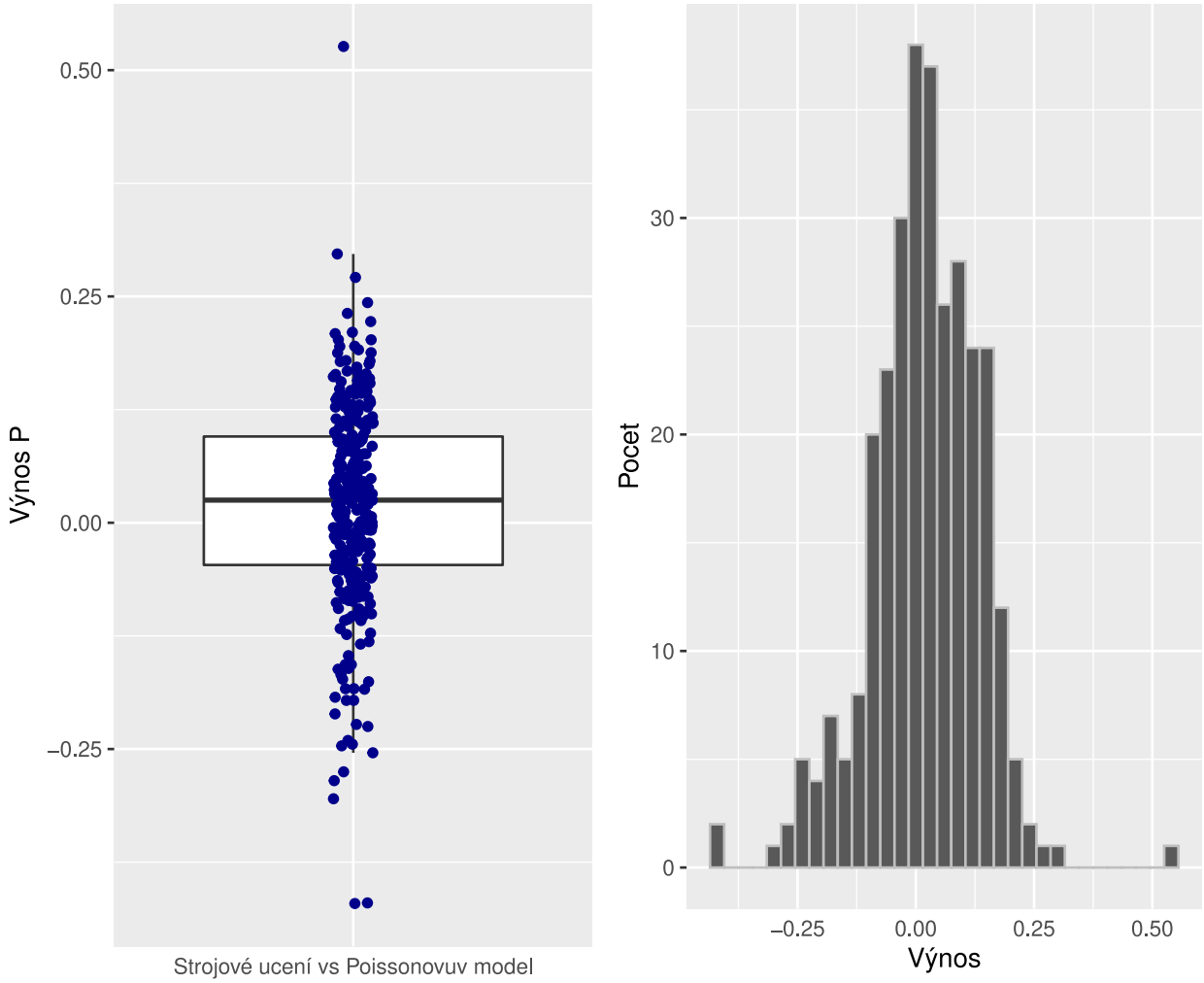
Bayesův faktor je tedy $K \doteq 73168.573$. Na základě tabulky 4.2 je hodnota K rozhodující ve prospěch strojového učení.

Nyní se podíváme, kde Poissonův model nejvíce prohrál. V tabulce 4.4 jsou zobrazeny největší ztráty Poissonova modelu ze všech her. Tyto zápasy nejvíce ovlivňují výsledek „který model je lepší“. Většinou jde o zápasy, kde nastal méně očekávaný výstup, čili došlo k překvapení a vyhrál tým s malou pravděpodobností.

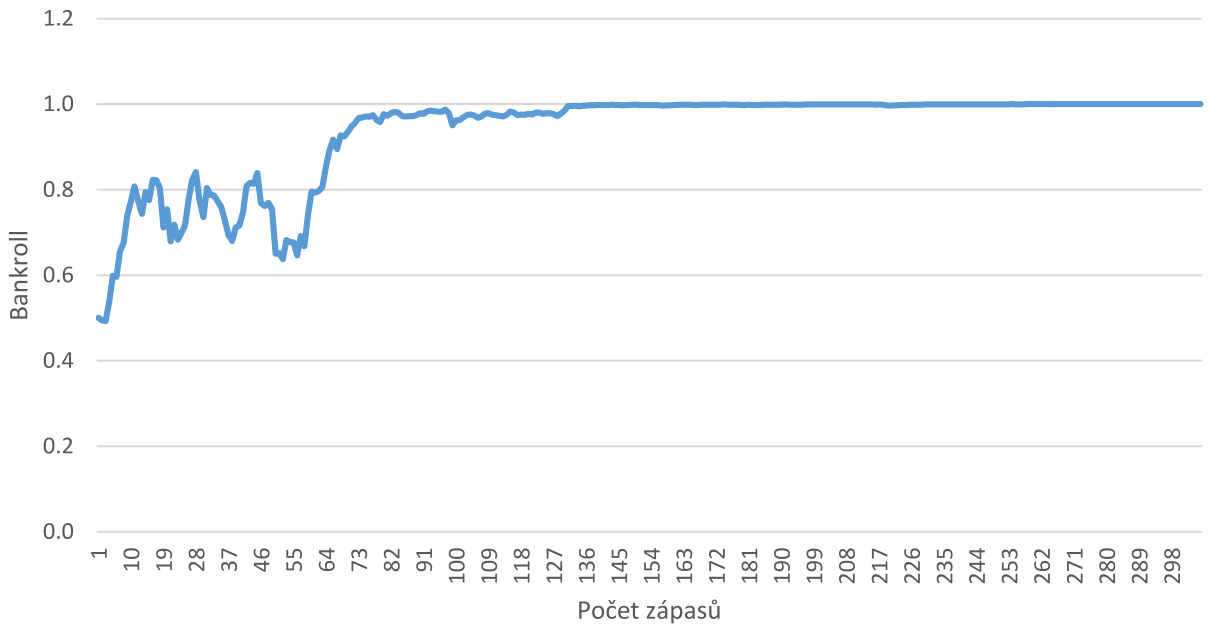
Domácí Tým	Hostující Tým	Výsledek	PM	SU	Výnos
Borussia Mönchengladbach	FC Bayern Mnichov	W	0.126	0.407	-0.26
FC Augsburg	SV Darmstadt 98	L	0.273	0.504	-0.15
VfL Wolfsburg	Borussia Mönchengladbach	W	0.353	0.615	-0.14
Hamburger SV	SV Darmstadt 98	L	0.253	0.416	-0.12
FC Ingolstadt 04	Borussia Mönchengladbach	W	0.284	0.455	-0.12
FC Augsburg	Hamburger SV	L	0.321	0.505	-0.11
Eintracht Frankfurt	VfL Wolfsburg	W	0.320	0.490	-0.11
1. FC Köln	FC Augsburg	L	0.271	0.415	-0.10
Hamburger SV	FC Augsburg	L	0.284	0.427	-0.10

Tabulka 4.4: Největší ztráty z pohledu Poissonova modelu, kde W - výhra domácích, D - remíza a L - prohra domácích. PM - Pravděpodobnosti pomocí Poissonova modelu, SU - Pravděpodobnosti pomocí strojového učení. Zobrazeny jsou pouze pravděpodobnosti výsledku zápasu, který skutečně nastal. Výnos je dán vztahem podle (4.2).

Na obrázku 4.1 jsou pro znázornění vykreslené všechny výnosy ve prospěch strojového učení podle vzorce (4.2) s konstantním bankrollem 1/2. Graf 4.2 zobrazuje vývoj bankrollu B^1 strojového učení. Z grafu je patrné, že už přibližně po 100 zápasech je hodnota velmi blízko jedné a tedy strojové učení lépe odhaduje pravděpodobnosti výsledku zápasu.



Obrázek 4.1: Vykreslené hodnoty výnosu z (4.2) pro všech 306 zápasů.



Obrázek 4.2: Vývoj bankrollu B^1 Strojového učení.

5. Využití všech údajů ze zápasu

V poslední kapitole se pro zajímavost podíváme na pravděpodobnosti, které odhadneme pomocí všech parametrů (viz tabulka 1.2). Vstupy do neuronové sítě v tomto případě bude tvořit rozdíl parametrů domácího týmu a hostujícího týmu (*domácí* − *hostující*). Teď máme mnohem více informací o tom, jak zápasy probíhají, proto by jsme mohli očekávat, že odhady budou přesnější.

Výsledky simulování sezóny je uvedeno v tabulce 5.1.

Tým	Body	Očekávané Body	Pravděpodobnost Výhry [%]
FC Bayern Mnichov	88	69.45	57.56
Borussia Dortmund	78	68.03	40.29
Borussia Mönchengladbach	55	55.11	0.90
Bayer 04 Leverkusen	60	52.72	0.68
Hamburger SV	41	49.46	0.18
1. FSV Mainz 05	50	49.24	0.13
FC Schalke 04	52	46.95	0.07
1. FC Köln	43	46.34	0.06
FC Ingolstadt 04	40	45.49	0.05
Hertha BSC	50	48.18	0.05
SV Darmstadt 98	38	43.62	0.01
FC Augsburg	38	44.70	0.01
TSG 1899 Hoffenheim	37	42.54	0.00
VfL Wolfsburg	45	39.35	0.00
Eintracht Frankfurt	36	37.06	0.00
SV Werder Bremen	38	38.30	0.00
VfB Stuttgart	33	37.98	0.00
Hannover 96	25	27.69	0.00

Tabulka 5.1: Simulování výsledků sezóny 2015/2016 pomocí strojového učení s použitím všech parametrů.

V kapitole 4 vyšlo lépe strojové učení, budeme tedy testovat, zda odhady založené na všech parametrech dávají lepší odhad než odhady založené pouze na vstřelených gólech. Označme:

$$M_1 = \text{Strojové učení založené na všech parametrech,}$$
$$M_2 = \text{Strojové učení založené na vstřelených gólech.}$$

Využijeme teorii v kapitole 4 a k porovnání použijeme znovu Bayesův faktor K daný vzorcem (4.3). Pravděpodobnosti jsou následující:

$$\mathbb{P}(M_1|\mathbf{D}) \doteq 1,$$
$$\mathbb{P}(M_2|\mathbf{D}) \doteq 0.$$

Bayesův faktor je $K > 10^{10}$. Na základě tabulky 4.2 je hodnota K rozhodující ve prospěch strojového učení založená na všech parametrech.

Závěr

V této práci jsme se věnovali dvěma metodám, jak odhadovat pravděpodobnosti výsledku fotbalového zápasu. Nejdříve jsme popsali Poissonův model, kde jsme využívali vstřelené góly v jednotlivých zápasech. Jako druhý model jsme popsali teorii pro strojové učení a následně využili na stejná data, jako v Poissonově modelu.

V kapitole 4 jsme popsali jakým způsobem porovnávat modely. Pozorovali jsme lepší výsledky pro strojové učení na úkor Poissonova modelu. Nicméně na rozdíl od strojového učení, v Poissonově modelu vidíme jasnou strukturu, což při rozhodování výběru modelu může hrát taktéž velkou roli.

Na závěr jsme pro zajímavost uvedli odhad pravděpodobností s využitím všech údajů z jednotlivých zápasů. Viděli jsme, že tyto odhady jsou lepší něž odhady pomocí strojového učení založené pouze na vstřelených gólech.

Seznam použité literatury

Efron, B. a Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, New York, NY, USA, 1st edition. ISBN 1107149894, 9781107149892.

Kass, R. E. a Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, vol. 90, no. 430, 1995, pp. 773–795. JSTOR*. URL www.jstor.org/stable/2291091.

Lee, A. J. (1997). Modeling scores in the premier league: Is manchester united really the best? *CHANCE*, **10**(1), 15–19. doi: 10.1080/09332480.1997.10554791. URL https://doi.org/10.1080/09332480.1997.10554791.

Müller, A. a Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media. ISBN 9781449369897. URL https://books.google.cz/books?id=vbQlDQAAQBAJ.

Večeř, J. (2018). Dynamic scoring: Probabilistic model selection based on utility maximization. URL https://ssrn.com/abstract=3276544.

Večeř, J. (2020). State price density equilibrium and bayesian statistics. URL https://ssrn.com/abstract=3532237.

Seznam tabulek

A. Přílohy

Přiložená příloha obsahuje dvě složky: Data a Zdrojové Kódy.

A.1 Data

Příloha Data obsahuje veškerá data ze zápasů (trénovací data a testovací data). Dále obsahuje souhrnné statistiky ze sezón, výpočet parametrů pro Poissonův model a výpočet pro srovnání modelů.

A.2 Zdrojové Kódy

Příloha Zdrojové Kódy obsahuje veškeré zdrojové kódy. Zdrojový kód SezonSimulation je naprogramovaný v jazyce *C#* a simuluje průběh sezóny. Zdrojový kód MachineLearning je naprogramovaný v jazyce Python a pomocí knihovny scikit-learn počítá pravděpodobnosti v kapitole 3.