Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra matematiky

Bakalářská práce

Vybrané prediktivní modely pro výsledky zápasů NBA

Plzeň, 2018 Jan Ondříček

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných zdrojů informací.

V Plzni dne 23. května 2018

Jan Ondříček

Poděkování

Touto cestou bych rád poděkoval RNDr. Blance Šedivé, Ph.D. za vedení bakalářské práce, odborný dohled, ochotu a čas, který této práci věnovala. Poděkování patří také především celé mojí rodině za neustálou podporu během mého studia.

Abstrakt

Cílem této práce je pokusit se pomocí vybraných rankingových modelů, predikovat výsledky zápasů NBA a porovnat úspěšnost predikce těchto modelů mezi sebou a se sázkovou kanceláří. Vybrány byly tři modely, a to Keenerova metoda, Elo rating a PageRank. Pro každý z modelů byly použity pro predikci tři typy informací o výsledcích utkání, které metody zohledňují. Tyto typy jsou: pouze výsledek utkání (výhra/prohra), výsledné skóre utkání a *Four Factors* (čtyři faktory významně ovlivňující výsledky utkání) s použitím původně navržených vah a určením vah pomocí logistické regrese. Úspěšnost predikce je v této práci analyzována pro každou sezonu zvlášť a hodnotícím kritériem pro srovnání modelů a sázkové kanceláře je průměrná procentuální úspěšnost predikce ze všech sledovaných sezon.

Klíčová slova: ranking, Keenerova metoda, Elo rating, PageRank, Four Factors, predikce výsledků NBA

Abstract

The aim of this thesis is to try to predict NBA results using selected ranking models and to compare the predictive success of these models with each other and with a betting office. Three models were selected, namely Keener's method, Elo rating and PageRank. For each of these models, three types of information about game results were used. These types are: only match result (win/loss), final score and *Four Factors* (four factors significantly affecting match results) using originally designed weights and determination of weights by logistic regression. The predictive success is analyzed for each season separately and the criterion for the models and the betting office comparison is the average value of the predictive success during all studied seasons.

Keywords: ranking, Keener's method, Elo rating, PageRank, Four Factors, prediction of NBA results

Obsah

Seznam tabulek

Seznam obrázků

Kapitola 1

Úvod

Tato práce se zabývá predikcí výsledků basketbalových utkání, konkrétně nejvyšší profesionální ligy v Severní Americe – NBA (National Basketball Association), a to pomocí tzv. rankingových modelů. Ranking v našem kontextu představuje metodu sestavování pořadí týmů na základě určitých pravidel, díky kterým je všem týmům přidělen tzv. rating, podle kterého můžeme týmy seřadit a porovnávat je mezi sebou. Nemusí se samozřejmě jednat pouze o sportovní týmy (hráče) a aplikaci rankingu na sport, s metodami sestavování pořadí se setkáváme denně, například při vybírání nejlépe hodnocené restaurace nebo zboží při nakupování přes internet.

K aplikaci rankingových modelů a následné predikci výsledků zápasů je nejprve potřeba zajistit dostatečné množství historických dat, na která budeme modely aplikovat a testovat je. Zajištěny tak budou zápasové statistiky NBA z období od sezony 2002/2003 do sezony 2016/2017, a to ze všech utkání základních částí těchto sezon. Získáním těchto dat, jejich úpravou, zpracováním a základními statistickými charakteristikami se zabývá kapitola 2.

Princip vybraných rankingových modelů bude podrobně popsán a vysvětlen v kapitole 3. Díky této teoretické části budou získány postupy výpočtu ratingového hodnocení týmů podle všech tří vybraných modelů, které budeme moci použít k následné predikci výsledků zápasů.

Implementací predikce pomocí rankingových modelů se bude zabývat kapitola 4. Bude zde popsáno, jakým způsobem využijeme rankingové modely k predikci výsledků zápasů v praktické části této práce, přičemž hlavní myšlenkou bude průběžné přepočítávání resp. aktualizace ratingu. Informace využitá k výpočtu ratingu, kterou o sobě dva soupeřící týmy po vzájemném utkání předají, bude volena třemi resp. čtyřmi různými způsoby, a to pouze podle výsledku zápasu (výhra/prohra), podle výsledného skóre utkání a podle tzv. *Four Factors* (čtyři

faktory významně ovlivňující výsledky utkání) s použitím původně navržených vah a určenými váhami pomocí logistické regrese.

V kapitole 5 bude vyhodnocena úspěšnost predikce pomocí vybraných modelů, které mezi sebou budou porovnány na základě kritéria průměrné procentuální úspěšnosti predikce výsledků zápasů, za celé sledované období. Dojde zde rovněž ke srovnání analyzovaných modelů se sázkovou kanceláři na základě stejného kritéria. Analyzována bude také procentuální úspěšnost predikce modelů v závislosti na vzdálenosti ratingů a na prostředí.

Na závěr budou shrnuty výsledky celé této práce a prezentovány některé návrhy na možné vylepšení modelů. Všechny programy vytvořené za účelem získání výsledků této práce budou včetně všech použitých dat přiloženy v příloze A spolu s jejich stručným popisem.

Kapitola 2

Data z NBA

2.1 Získání dat

Historická data z výsledků utkání byla získána z oficiálních webových stránek NBA [20]. Jedná se tedy o nejdůvěryhodnější možný zdroj, jehož věrohodnost už z běžně dostupných zdrojů nelze lépe ověřit. Získány byly týmové statistky, a to jednotlivě pro každé utkání a každý tým zvlášť, obsahující:

- tým **(TEAM)**, soupeř **(OPP)**, údaj o domácím/venkovním prostředí **(H/A)**, datum utkání **(GAME DATE)**,
- výhra/prohra **(W/L)**, počet bodů **(PTS)**, rozdíl bodů v utkání **(+/-)**,
- počet proměněných střel z pole celkem **(FGM)**, počet pokusů z pole celkem **(FGA)**, procentuální úspěšnost střel z pole celkem **(FG%)**,
- počet proměněných 3bodových střel **(3PM)**, počet pokusů 3bodových střel **(3PA)**, procentuální úspěšnost 3bodových střel **(3P%)**,
- počet proměněných trestných hodů **(FTM)**, počet trestných hodů celkem **(FTA)**, procentuální úspěšnost trestných hodů **(FT%)**,
- počet útočných **(OREB)** a obranných doskoků **(DREB**), počet doskoků celkem **(REB)**, počet asistencí **(AST)**, počet zisků **(STL)**, počet bloků **(BLK)**, počet ztrát **(TOV)**, počet osobních faulů **(PF)**.

Data byla získána pro všechna utkání základních částí, a to od sezony 2002/2003 do sezony 2016/2017 (tedy 15 sezon). Tato doba byla považována za dostatečně dlouhou a reprezentativní. Během základní části sezony NBA hraje každý tým standardně 82 utkání. Prvních dvou zkoumaných sezon se účastnilo pouze 29 týmů, a proto byl celkový počet utkání v každé z těchto sezon 1189. Od sezony 2004/2005 se již soutěže účastnilo 30 týmů, tedy celkově 1230 utkání za sezonu. Výjimkou byla sezona 2011/2012, ve které se kvůli výluce odehrálo pouze 990

utkání (každý tým 55). V sezoně 2012/2013 se neodehrálo utkání v závěru sezony mezi Bostonem a Indianou (jediné neodehrané utkání v historii NBA), důvodem byl teroristický útok v Bostonu. Celkově tedy byly získány zápasové statistiky ze 18 127 vzájemných utkání. Data byla importována do softwaru Excel 2016, kde proběhly jejich úpravy, jako je rozdělení do příslušných sloupců, seřazení, nastavení formátu data, oddělovačů aj. Důležitou úpravou bylo přejmenování týmů na jejich aktuální jména, jelikož se v průběhu sledovaných sezon některá změnila. Seznam názvů všech aktuálních týmů a jejich oficiálních zkratek, pod kterými vystupují, je k nalezení v tabulce 2.1a. Následně byla získána data otvíracích decimálních kurzů, vypsaných na jednotlivá utkání ze serveru [19], shromažďující historické kurzy sázkových kanceláří, a to od sezony 2009/2010. Se staršími daty již nebylo pracováno, jelikož jsou špatně dostupná a pro účely této práce jsou získaná data dostačující. Primárně byly použity kurzy sázkové kanceláře *Bet365*, kterých bylo dostupné největší množství. Chybějící záznamy této kanceláře byly doplněny kurzy bookmakera *Pinnacle* nebo (podle dostupnosti) *Bwin* a *Unibet* ze serveru [21]. Z tohoto serveru byla rovněž namátkově zkontrolována věrohodnost získaných kurzů. V sezoně 2010/2011 chyběly záznamy z 5 utkání, a to v obou požitých zdrojích. Kurzy pro tato utkání byly získány ze zdroje [23] ve tvaru americké *moneyline* a následně přepočteny na používaný decimální tvar. Všechny získané kurzy byly získány na konečný výsledek utkání včetně prodloužení (tedy bez remízy). Následně proběhlo přiřazení získaných kurzů **(ODDS)** k zápasovým statistikám z jednotlivých utkání. V tabulce 2.2 je ukázka získaných a upravených dat pro 4 utkání. Kompletní upravená data jsou k nalezení v přiloženém souboru *NBAstats.xlsx* (příloha A.1), včetně kontingenční tabulky.

2.2 Základní statistické zpracování dat

Pro získané zápasové statistiky byly určeny základní statistické charakteristiky, které jsou k nalezení v tabulce 2.3a, a to pro celé sledované období. Následně byl určen aritmetický průměr všech zápasových statistik pro jednotlivé sezony, viz tabulka 2.3b. Pozoruhodným jevem je vývoj aritmetického průměru PTS, 3PM, 3PA a FGM. V průběhu sledovaného období totiž vzrostl průměrný počet 3bodových pokusů z 14,68 na 27,00 a průměrný počet proměněných 3bodových pokusů z 5,13 na 9,65, tedy o 88,11 %. To má za důsledek zvýšení celkového průměrného počtu proměněných střel, a tedy i počtu bodů na utkání o 10,51 bodů. Další zajímavostí je pokles průměru OREB, který ale není tak výrazný jako vývoje výše jmenovaných. Tyto vývoje aritmetických průměrů jsou zachyceny na obrázku 2.1 a jsou doplněny směrodatnými odchylkami.

Zkratka	Název týmu
ATL	Atlanta Hawks
BKN	Brooklyn Nets
BOS	Boston Celtics
CLE	Cleveland Cavaliers
DET	Detroit Pistons
CHA	Charlotte Hornets
CHI	Chicago Bulls
IND	Indiana Pacers
MIA	Miami Heat
MIL	Milwaukee Bucks
NYK	New York Knicks
ORL	Orlando Magic
PHI	Philadelphia 76ers
TOR	Toronto Raptors
WAS	Washington Wizards

(a) Východní konference

Zkratka	Název týmu
DAL	Dallas Mavericks
DEN	Denver Nuggets
GSW	Golden State Warriors
HOU	Houston Rockets
LAC	Los Angeles Clippers
LAL	Los Angeles Lakers
MEM	Memphis Grizzlies
MIN	Minnesota Timberwolves
NOP	New Orleans Pelicans
OKC	Oklahoma City Thunder
PHX	Phoenix Suns
POR	Portland Trail Blazers
SAC	Sacramento Kings
SAS	San Antonio Spurs
UTA	Utah Jazz

(b) Západní konference

Tabulka 2.1: Aktuální názvy všech současných týmů NBA

TEAM	H/A	OPP	GAME DATE	W/L	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM
ATL	H	DET	27. 10. 2015	L	94	37	82	45,1	8	27	29,6	12
DET	A	ATL	27. 10. 2015	W	106	37	96	38,5	12	29	41,4	20
CLE	A	CHI	27. 10. 2015	L	95	38	94	40,4	9	29	31,0	10
CHI	H	CLE	27. 10. 2015	W	97	37	87	42,5	7	19	36,8	16
GSW	H	NOP	27. 10. 2015	W	111	41	96	42,7	9	30	30,0	20
NOP	A	GSW	27. 10. 2015	L	95	35	83	42,2	6	18	33,3	19
DEN	H	HOU	28. 10. 2015	W	105	40	79	50,6	13	27	48,1	12
HOU	A	DEN	28. 10. 2015	L	85	30	87	34,5	8	35	22,9	17

FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-	SEASON	ODDS
15	80,0	7	33	40	22	9	4	15	25	-12	2015/2016	1,33
26	76,9	23	36	59	23	5	3	15	15	12	2015/2016	3,40
17	58,8	11	39	50	26	5	7	11	21	-2	2015/2016	2,30
23	69,6	7	40	47	13	6	10	13	22	2	2015/2016	1,66
22	90,9	21	35	56	29	8	7	20	29	16	2015/2016	1,18
27	70,4	8	25	33	21	9	3	19	26	-16	2015/2016	5,25
18	66,7	9	40	49	26	9	10	21	26	20	2015/2016	5,50
26	65,4	15	29	44	17	13	8	17	19	-20	2015/2016	1,16

Tabulka 2.2: Ukázka vstupních dat, 4 utkání

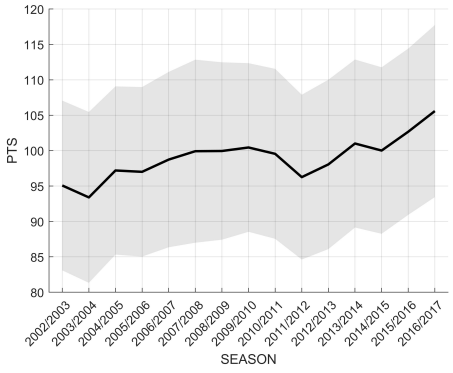
	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF
Aritmetický průměr	99,06	36,98	81,67	45,38	6,75	18,96	35,24	18,34	24,24	75,72	11,16	31,00	42,16	21,58	7,55	4,86	14,46	14,46
Rozptyl	153,35	26,58	57,06	31,92	11,27	48,17	142,49	39,17	59,36	102,37	15,83	28,11	42,25	25,79	8,45	6,64	15,46	15,46
Směrodatná odchylka	12,38	5,16	7,55	5,65	3,36	6,94	11,94	6,26	7,70	10,12	3,98	5,30	6,50	5,08	2,91	2,58	3,93	3,93
Modus	98	37	80	50,00	6	16	33,30	18	21	75,00	11	30	41	21	7	4	13	13
Medián	99	37	81	45,20	6	18	35,30	18	24	76,30	11	31	42	21	7	5	14	14
Horní kvartil	107	40	86	49,30	9	23	42,90	22	29	82,60	14	35	46	25	9	6	17	17
Dolní kvartil	91	33	77	41,60	4	14	27,30	14	19	69,20	8	27	38	18	6	3	12	12
Kvartilové rozpětí	16	7	9	7,70	5	9	15,60	8	10	13,40	6	8	8	7	3	3	5	5
Minimum	53	16	55	24,40	0	1	0,00	1	1	14,30	1	12	17	4	0	0	2	2
Maximum	168	63	129	68,40	25	61	100,00	52	64	100,00	38	55	81	47	22	19	33	33
Variační rozpětí	115	47	74	44,00	25	60	100,00	51	63	85,70	37	43	64	43	22	19	31	31
Variační koeficient	0,13	0,14	0,09	0,12	0,50	0,37	0,34	0,34	0,32	0,13	0,36	0,17	0,15	0,24	0,39	0,53	0,27	0,27
Šikmost	0,15	0,14	0,31	0,10	0,61	0,49	0,07	0,43	0,40	-0,38	0,47	0,17	0,20	0,26	0,42	0,66	0,28	0,28
Špičatost	3,17	3,14	3,51	2,98	3,40	3,31	3,43	3,21	3,21	3,41	3,34	3,04	3,13	3,12	3,25	3,53	3,11	3,11

(a) Základní charakteristiky

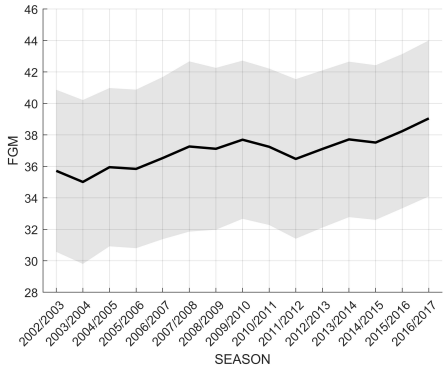
	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF
2002/2003	95,08	35,72	80,79	44,31	5,13	14,68	34,35	18,51	24,43	75,82	12,05	30,26	42,31	21,50	7,94	5,02	14,92	21,75
2003/2004	93,40	35,01	79,82	43,94	5,18	14,93	34,30	18,20	24,21	75,22	12,09	30,11	42,20	21,30	7,93	5,06	14,92	21,45
2004/2005	97,20	35,95	80,34	44,86	5,60	15,75	35,16	19,70	26,05	75,61	12,01	29,85	41,86	21,28	7,52	4,86	14,50	22,63
2005/2006	97,01	35,84	78,99	45,49	5,73	15,98	35,47	19,60	26,30	74,42	11,18	29,78	40,96	20,61	7,17	4,70	14,41	22,76
2006/2007	98,74	36,53	79,70	45,93	6,07	16,94	35,33	19,62	26,08	75,30	11,12	29,93	41,05	21,29	7,24	4,61	15,13	22,22
2007/2008	99,92	37,26	81,50	45,80	6,55	18,11	35,75	18,84	24,94	75,64	11,20	30,78	41,98	21,75	7,28	4,74	14,11	21,02
2008/2009	99,95	37,12	80,92	45,96	6,65	18,12	36,29	19,07	24,74	77,04	11,04	30,26	41,30	20,98	7,27	4,80	14,03	21,04
2009/2010	100,45	37,70	81,70	46,26	6,43	18,14	35,12	18,63	24,54	75,78	10,96	30,77	41,72	21,24	7,22	4,86	14,22	20,86
2010/2011	99,55	37,25	81,22	45,98	6,46	18,01	35,36	18,60	24,36	76,24	10,91	30,48	41,39	21,50	7,33	4,86	14,25	20,71
2011/2012	96,26	36,47	81,43	44,89	6,41	18,38	34,40	16,90	22,46	75,36	11,37	30,81	42,18	20,98	7,68	5,09	14,58	19,57
2012/2013	98,06	37,11	81,95	45,40	7,16	19,95	35,53	16,69	22,17	75,41	11,16	30,94	42,10	22,12	7,80	5,13	14,54	19,83
2013/2014	101,01	37,72	83,00	45,58	7,75	21,53	35,72	17,83	23,59	75,71	10,91	31,84	42,75	22,00	7,68	4,71	14,65	20,70
2014/2015	100,01	37,52	83,57	45,01	7,85	22,41	34,78	17,14	22,84	75,05	10,89	32,41	43,29	22,03	7,74	4,80	14,35	20,21
2015/2016	102,67	38,24	84,57	45,33	8,52	24,08	35,20	17,68	23,36	75,95	10,42	33,34	43,76	22,29	7,85	4,96	14,38	20,27
2016/2017	105,59	39,05	85,41	45,82	9,65	27,00	35,66	17,84	23,11	77,11	10,14	33,38	43,51	22,63	7,70	4,74	13,95	19,90

(b) Průměrné sezonní hodnoty

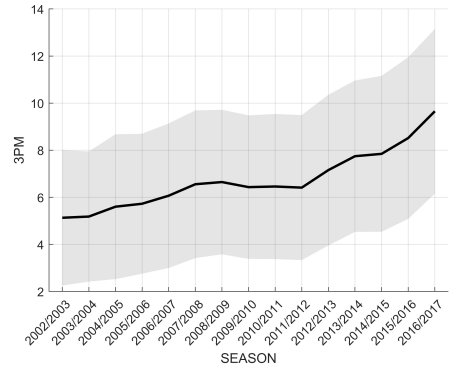
Tabulka 2.3: Základní statistické charakteristiky zápasových statistik



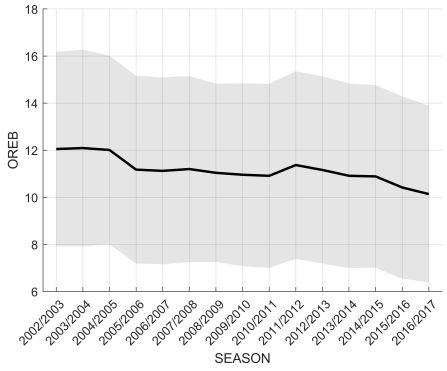
(a) PTS



(b) FGM



(c) 3PM



(d) OREB

Obrázek 2.1: Vývoj průměrů a směrodatných odchylek vybraných zápasových statistik

Kapitola 3

Teoretická část rankingových modelů

Ranking je obecně metoda sestavování pořadí na základě určitých pravidel. Setkáváme se s ním v běžném životě, například při nakupování přes internet, ve snaze vyhledat nejlépe hodnocený produkt a prodejce, při používání internetového vyhledávače (více v kapitole 3.3), nebo při hledání ubytování na dovolené. Ve sportu se s rankingem setkáváme při vytváření například žebříčku nejlepších světových hráčů a týmů (ATP Rankings, FIFA World Ranking, IIHF World Ranking, aj.), ale také při hodnocení týmů a hráčů v soutěžních sezonách. Jednotlivé týmy (hráči) jsou na základě nějakého modelu ohodnoceny tzv. *ratingem* a seřazeny, čímž získají určitý *rank* (pořadí). Tímto způsobem tak můžeme porovnávat, který tým (hráč) je lepší resp. nejlepší v určité oblasti. V této práci budeme využívat metodu rankingu pro predikci výsledků budoucích utkání. Na námi získaná data z NBA budeme aplikovat tři pokročilejší rankingové modely, díky kterým budeme postupně získávat rating jednotlivých týmů na základě odehraných utkání, a získáme tak možnost odhadnout, který z dvojice týmů má větší šanci vyhrát nadcházející zápas. Jedná se o modely: Keenerova metoda, Elo rating a PageRank. Tyto tři modely byly vybrány především ve snaze srovnat modely s pokročilejším matematickým pozadím, existuje však celá řada různých rankingových modelů a dalších přístupů k jejich využití. Všechny námi vybrané modely zahrnují tzv. *sílu rozpisu*, což znamená, že rozlišují sílu soupeřů, se kterými tým hraje (v NBA nehrají všechny týmy stejný počet utkání se stejnými týmy, záleží na divizi a konferenci, kde může být síla týmů v jednotlivých oblastech rozdílná).

3.1 Keenerova metoda

Tato metoda, pojmenovaná po americkém matematikovi J. P. Keenerovi, který ji poprvé použil v roce 1993 [6] na fotbalovou ligu, vychází z tzv. **přímé metody**. Předpokládejme, že vektor $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, kde n je celkový počet

týmů, je ratingový vektor, jehož všechny složky jsou kladné a odpovídají silám jednotlivých týmů. Dále předpokládejme, že a_{ij} jsou nezáporná čísla závislá na výsledcích utkání mezi týmem i a týmem j (např. konečný počet vstřelených bodů). Síla týmu i vzhledem k týmu j se nazývá *relativní síla* a je vyjádřena jako $f_{ij} = a_{ij}r_j$. Potom definujeme *absolutní sílu* týmu i jako součet relativních sil, tedy

$$f_i = \sum_{j=1}^{n} a_{ij}r_j. \tag{3.1}$$

Vektor $\mathbf{f} = [f_1, f_2, \ldots, f_n]$ je pak vektorem absolutních sil všech týmů. Matice $\mathbf{A}$ s prvky a_{ij} se nazývá *preferenční matice*. Maticově zapsáno, zjistíme vektor obsahující absolutní sílu všech týmů z rovnosti $\mathbf{f} = \mathbf{Ar}$, přičemž $\mathbf{f}$ a $\mathbf{r}$ jsou neznámé. Víme však, že rating každého týmu, by měl být přímo úměrný absolutní síle týmu [7], tedy $\mathbf{f} = \lambda\mathbf{r}, \lambda \in \mathbb{R}$. To znamená, že vektor ratingů $\mathbf{r}$ můžeme určit ze vztahu

$$\mathbf{Ar} = \lambda\mathbf{r}, \tag{3.2}$$

tedy λ je vlastním číslem matice $\mathbf{A}$ a vektor $\mathbf{r}$ je (pravý) vlastní vektor k němu příslušný. K určení požadovaného ratingového vektoru $\mathbf{r}$ využijeme následující větu, kterou lze najít ve zdroji [6] včetně důkazu.

Věta 3.1.1 (Perronova-Frobeniova) *Nechť (netriviální) matice $\mathbf{A}$ má všechny prvky nezáporné. Potom existuje kladné vlastní číslo λ matice $\mathbf{A}$ a k němu příslušný nezáporný vlastní vektor $\mathbf{r}$. Navíc pokud matice $\mathbf{A}$ je nerozložitelná, existuje právě jedno vlastní číslo λ takové, že je spektrálním poloměrem[1] matice $\mathbf{A}$ a k němu příslušný vlastní vektor $\mathbf{r}$ je určen jednoznačně a je kladný.*

Tento vlastní vektor se nazývá (pravý) **Perronův vektor** a jedná se právě o námi hledaný vektor ratingů[2] $\mathbf{r}$. Určit ho lze pomocí numerických metod, více např. [10]. Pro upřesnění, matice $\mathbf{A}$ se nazývá nerozložitelná, pokud neexistuje její permutace taková, že ji lze zapsat ve tvaru

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}, \tag{3.3}$$

kde $\mathbf{A}_{11}$ a $\mathbf{A}_{22}$ jsou čtvercové matice a matice $\mathbf{0}$ je matice samých nul. Nerozložitelnost při aplikaci Keenerovy metody lze zaručit například přičtením vždy

[1] Spektrálním poloměrem matice $\mathbf{A}$ se nazývá číslo $\rho(\mathbf{A}) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$, kde λ_i jsou vlastní čísla matice $\mathbf{A}$.

[2] K výpočtu ratingového vektoru byla napsána funkce *get_rating_keener.m* (viz příloha A.5) a použita funkce *perron.m* (viz příloha A.9).

nějaké malé hodnoty (např. 0,001) ke všem prvkům matice $\mathbf{A}$. Otázkou zůstává, jak zvolit matici $\mathbf{A}$ resp. její prvky a_{ij}. Jednou z možností je

$$a_{ij} = \begin{cases} 1, & \text{pokud tým } i \text{ porazil tým } j \\ 0, & \text{jinak} \end{cases}. \tag{3.4}$$

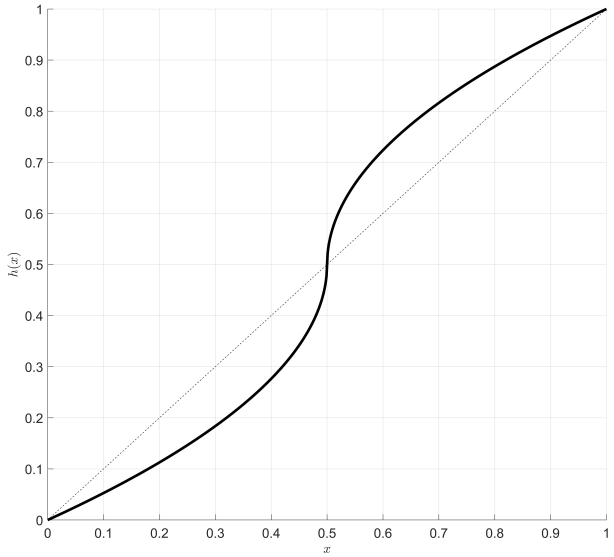
Pokud spolu dva týmy sehrají více utkání (v NBA během základní části sezony vždy minimálně 2), a_{ij} je pak počet výher týmu i, proti týmu j. Nevýhodou této volby je to, že bod za výhru dostává pouze vítěz a není tak zohledněna hra poraženého. Týmy bez výhry mají pak rating 0 a nepřispívají k hodnocení jejich soupeřů, navíc $\mathbf{A}$ může být rozložitelná. Další možností volby a_{ij} je rozdělení jednoho bodu mezi oba týmy soupeřící v daném utkání, a to podle výsledného skóre. Označme S_{ij} jako výsledný počet bodů, které tým i vstřelil týmu j, potom $a_{ij} = \frac{S_{ij}}{S_{ij}+S_{ji}}$, a tedy $a_{ij} + a_{ji} = 1$. Nevýhodou je to, že pokud některý tým nevstřelí žádný bod, dostaneme se opět do situace jako v předchozí uvažované volbě. Aby se této situaci předešlo, a_{ij} je konstruováno jako $a_{ij} = \frac{S_{ij}+1}{S_{ij}+S_{ji}+2}$. V basketbalu se nestává, že by některý z týmu nezískal žádný bod, hodnota S_{ij} bude ovšem určována i jinými způsoby, viz kapitola 4.2. Pokud spolu dva týmy sehrají více utkání, S_{ij} je součtem bodů, které získal tým i proti týmu j, ze všech vzájemných utkáních. Aplikace nelineární funkce $h(x)$ na obrázku 3.1 zaručí, že těsnější výhry mají větší váhu. To napomáhá předcházet situaci, kdy výrazně silnější tým může ukazovat svou převahu a rozdíl skóre je velký, což může mít dopad na výsledný rating. Prvky matice $\mathbf{A}$ jsou pak ve tvaru

$$a_{ij} = h\left(\frac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right), \tag{3.5}$$

kde

$$h(x) = \frac{1}{2} + \frac{1}{2}\operatorname{sgn}\left(x - \frac{1}{2}\right)\sqrt{|2x-1|}. \tag{3.6}$$

Poslední možnou úpravou prvků a_{ij} je jejich normalizace, tedy $a_{ij} = \frac{a_{ij}}{n_i}$, kde n_i je počet odehraných utkání týmem i. Touto úpravou můžeme eliminovat rozdíly způsobené různým počtem utkání každého týmu.



Obrázek 3.1: Nelineární funkce $h(x)$

3.2 Elo rating

Elo rating systém je další metodou, která slouží k porovnávání jednotlivých týmů (hráčů) na základě jejich relativní síly. Jejím autorem byl fyzik maďarského původu, Arpad Elo. Původně byl tento systém zaveden pro hodnocení hráčů šachu, kde nahradil starší systémy a používá se ve většině šachových federací (např. FIDE) dodnes. Postupně se však začal aplikovat i v různých jiných sportech, kde se střetávají dva týmy (hráči), a existuje tak možnost spolu tyto týmy porovnat na základě rozdílu jejich „bodů Elo“ (ratingu). Mimo jiné je tento systém hodnocení používán v mnoha videohrách. Výhodou tohoto modelu je především jeho jednoduchost. Systém Elo přepočítává rating průběžně po každém odehraném utkání pouze pro dvojici soupeřů, kteří spolu právě sehráli utkání, a to s využitím jejich předchozích ratingů. Základní myšlenkou této metody je zvýšit rating vítěznému týmu a naopak týmu, který prohrál, o stejnou hodnotu rating snížit. Výhodou je to, že hodnota, o kterou se týmu rating zvýší resp. sníží, závisí právě na ratingu soupeře. Jednoduše řečeno, vítězstvím nad silnějším týmem se zvýší rating daného týmu více, než kdyby zvítězil nad slabším týmem, a naopak prohrou se slabším týmem se rating sníží více, než prohrou s týmem silnějším [18] [4].

Na začátku je potřeba stanovit počáteční rating pro všechny týmy v soutěži,

tedy stanovit vektor $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, kde n je celkový počet týmů. Po odehraném utkání týmu i proti týmu j se nový rating[3] $i$tého týmu vypočítá následujícím způsobem

$$r_{i_{new}} = r_{i_{old}} + K(W_{ij} - E_{ij}), \tag{3.7}$$

kde $r_{i_{old}}$ je předchozí rating týmu i, $r_{i_{new}}$ je nově získaný rating týmu i, W_{ij} je získaný počet bodů týmu i závislý na výsledku utkání (například výsledné skóre utkání), E_{ij} je očekávaný počet bodů, který tým i musí získat v utkání s týmem j, aby jeho rating zůstal zachován a K je tzv. koeficient rozvoje [25].

Získaný počet bodů W_{ij} je možné volit různými způsoby, základní volbou je

$$W_{ij} = \begin{cases} 1, & \text{pokud tým } i \text{ porazil tým } j \\ 0, & \text{jinak} \end{cases}. \tag{3.8}$$

Při této volbě není zohledněno výsledné skóre utkání, ale pouze definován vítěz a poražený. Proto opět zaveďme druhou možnost volby [2] ve stejném tvaru jako v Keenerově metodě (kapitola 3.1), tedy

$$W_{ij} = \frac{S_{ij} + 1}{S_{ij} + S_{ji} + 2}, \tag{3.9}$$

kde S_{ij} je konečný počet bodů, které tým i vstřelil týmu j. V tomto případě je však každé utkání vyhodnocováno individuálně, nejde tak o součet bodů ze všech vzájemných utkání, tak jak tomu bylo u Keenerovy metody. Jedná se tedy o upravený poměr výsledného skóre, proto $W_{ij} \in (0, 1)$ a $W_{ij} + W_{ji} = 1$. Kromě výsledného skóre je možné použít i jakýkoliv jiný způsob bodového ohodnocení výsledku utkání, viz kapitola 4.2.

Očekávaný počet bodů E_{ij} z rovnice (3.7) vyjadřuje v tomto případě pravděpodobnost výhry hráče i nad hráčem j. K určení této hodnoty je v metodě Elo využívána logistická funkce $L(x) = \frac{1}{1+10^{-x}}$ v závislosti na rozdílu ratingů $d_{ij} = r_{i_{old}} - r_{j_{old}}$ dvou soupeřících týmů. Očekávaný počet získaných bodů resp. pravděpodobnosti výhry v utkání pak lze odhadnout jako

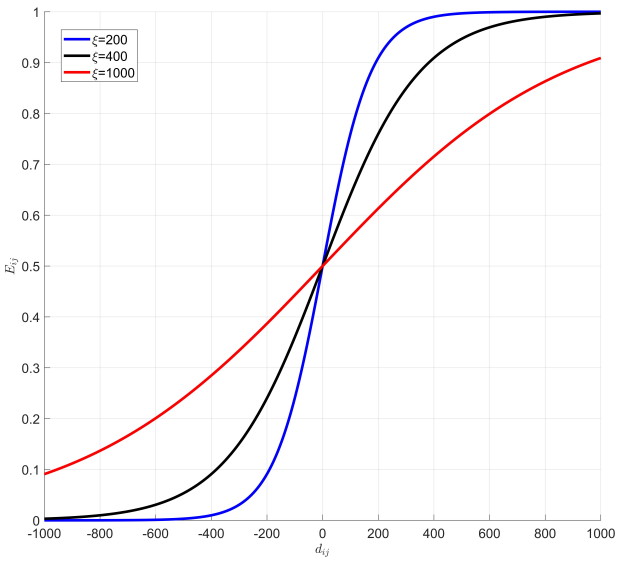
$$E_{ij} = \frac{1}{1 + 10^{\frac{-d_{ij}}{\xi}}}, \tag{3.10}$$

kde ξ je parametr, který určuje rozpětí ratingu. Pro každých ξ ratingových bodů, o kterých má tým i více než tým j, je očekávaný počet získaných bodů týmem i z daného utkání (pravděpodobnost výhry) desetkrát větší než očekávaný

[3] K výpočtu ratingového vektoru byla napsána funkce *get_rating_elo.m* (viz příloha A.6).

počet získaných bodů týmem j. Vhodná hodnota tohoto parametru se může lišit pro různé sporty. Často doporučovanou hodnotou např. pro šachové turnaje je $\xi = 400$ [18]. Na obrázku 3.2 vidíme vliv parametru ξ na výslednou hodnotu E_{ij}. Poznamenejme, že $E_{ij} \in (0, 1)$ a $E_{ij} + E_{ji} = 1$.

Koeficient rozvoje K z rovnice (3.7) určuje maximální možnou hodnotu zvýšení resp. snížení ratingu během jednoho odehraného utkání. Rozhoduje o tom, do jaké míry ovlivňuje výsledek utkání změnu z předchozího ratingu týmu ($r_{i_{old}}$) na nový rating ($r_{i_{new}}$). Pokud je K příliš velké, rating je na jednotlivá utkání příliš citlivý. To znamená, že pokud se týmu nevydaří jedno utkání tak, jak je jeho zvykem, může toto utkání příliš ovlivnit jeho rating. Naopak pokud K je příliš malé, model nebude schopný reagovat na zlepšení resp. zhoršení síly týmu a ratingy se tak nebudou příliš měnit. Vhodná volba koeficientu K je tedy důležitá, ale je odlišná pro různé sporty a situace. Tento koeficient nemusí být jenom konstantní, ale také se může v čase měnit. Na začátku sezony může být K vyšší, týmy tak mají větší šanci změnit své počáteční ratingy, než se jejich hodnota stabilizuje a více odpovídá jejich reálnému výkonu [5]. Tento koeficient může ale záviset i na jiných faktorech. Například v ČR se pro národní Elo šachových hráčů mění K v závislosti na jejich věku, protože se předpokládá, že u mladších hráčů nastávají větší změny výkonnosti [25]. Další možnost proměnlivosti K je např. v závislosti na výhodě domácího prostředí, důležitosti utkání (turnaje), aktuální hodnotě ratingu, nebo na výsledku daného utkání [15].



Obrázek 3.2: Vliv parametru ξ na výslednou hodnotu E_{ij}

3.3 PageRank

PageRank je metoda pojmenovaná po jednom ze svých několika spoluzakladatelů Larry Pageovi. Vznikla jako výzkumný projekt a poprvé byla představena v roce 1998 v pracích [3] a [9] spolu s webovým vyhledávačem sloužícím k testování této metody, který autoři Page a Brin nazvali *Google*. Tuto metodu používá *Google* dodnes k ohodnocení důležitosti webových stránek. Algoritmus přiřadí všem webovým stránkám rating (PageRank), přičemž se náhodný člověk surfující na internetu s největší pravděpodobností dostane po čase na webové stránky s vyšším ratingem. Díky tomu může vyhledávač zobrazit nejprve nejpopulárnější a nejnavštěvovanější webové stránky. Nejedná se však o jedinou metodu, podle které *Google* aktuálně řadí své vyhledávání. Díky efektivnosti tohoto algoritmu se začal používat i v jiných odvětvích, například právě ve sportu [22] [16].

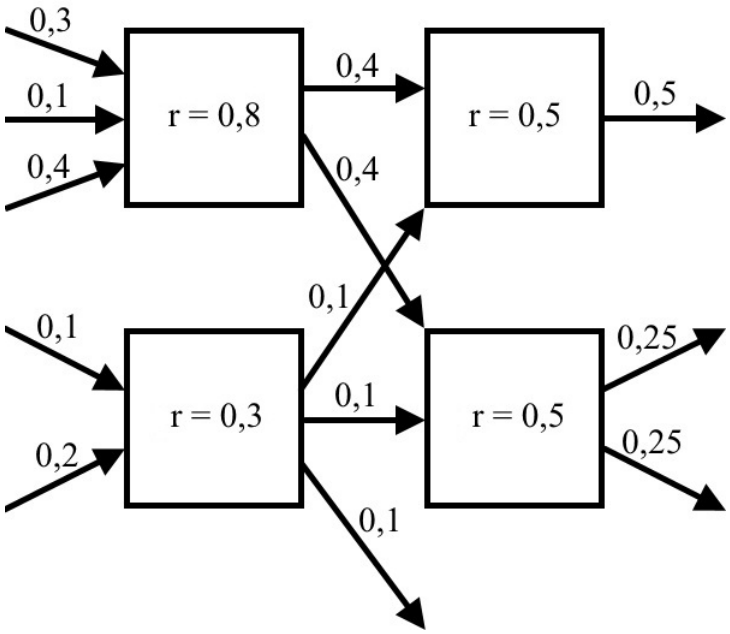
Metoda vychází z citační analýzy[4] s využitím hypertextových odkazů. V této kapitole bude PageRank vysvětlen na jeho původní aplikaci, tedy hodnocení webových stránek. Ve sportu je však celý princip této metody totožný s tím, že místo webových stránek si představme sportovní týmy a místo hypertextových odkazů informaci, kterou o sobě týmy předají na základě vzájemně odehraného utkání, čemuž budou věnovány pouze vybrané části této kapitoly. Základní myšlenkou je to, že webová stránka má vysoký rating, pokud součet ratingů ze zpětných odkazů je vysoký. To platí pro případ, kdy se jedná o stránku, která má mnoho zpětných odkazů, ale také pro případ, kdy stránka má jen několik zpětných odkazů, ale zato s vysokým ratingem. Zpětným odkazem je myšlen odkaz na jiné webové stránce, který vede právě na sledovanou stránku. Z této myšlenky vychází základní zjednodušená definice PageRanku [9].

Definice 3.3.1 (PageRank) *Nechť i je webová stránka. F_i je množina stánek, na které odkazuje stránka i a B_i je množina stránek, odkazujících na stránku i. Nechť $N_i = |F_i|$ je počet odkazů vedoucí ze stránky i. **PageRank** stránky i označený jako r_i, je potom vyjádřen jako*

$$r_i = \sum_{j \in B_i} \frac{r_j}{N_j}. \tag{3.11}$$

Web si lze představit jako orientovaný graf, kde jednotlivé vrcholy reprezentují webové stránky a orientované hrany jsou hypertextové odkazy z jedné stránky na jinou. Na obrázku 3.3 je ilustrováno, jak výpočet pomocí vzorce 3.11 funguje.

[4] Citační analýza je metoda, která poměřuje vztahy mezi autory a dokumenty, na základě bibliografických citací. Více např. [17].



Obrázek 3.3: Ilustrační ukázka výpočtu PageRanku

Při realizaci výpočtu je třeba si uvědomit, že výchozí hodnoty PageRanku webových stránek nejsou známy. Proto jsou tyto hodnoty voleny libovolně a výpočet probíhá **iteračně**, dokud nezačnou hodnoty konvergovat. Jedním z problémů výpočtu je tzv. **rank sink**. Tento problém vzniká ve chvíli, kdy se vyskytnou dvě stránky (nebo skupina stránek), které odkazují pouze na sebe navzájem, ne na žádnou jinou stránku. Představme si, že existuje stránka, která odkazuje na jednu z nich. Potom v průběhu iterací budou zmíněné dvě stránky resp. skupina kumulovat rating, ale žádný nebudou předávat ostatním. Dalším problémem, který zmiňují autoři, jsou tzv. **dangling links**, tedy odkazy, které odkazují na stránku, ze které už nevedou další odkazy nikam jinam (dangling stránka). K vyřešení těchto problémů byl zaveden pojem **random surfer**, tedy uživatel, který surfuje po internetu klikáním na odkazy, nikdy se nevrací zpět, ale po několik krocích se „začne nudit“ a přeskočí na jinou, náhodnou stránkou, kde začne se surfováním znova [12]. Pokud uživatel dorazí na stránku, ze které nevede žádný odkaz (dangling stránku), vybere si jinou, náhodnou. Pravděpodobnost, že v každém kroku bude uživatel pokračovat na další odkaz je tzv. **damping faktor** d, naopak pravděpodobnost, že přeskočí na novou stránku je $1 - d$. Hodnota této konstanty je obvykle volena okolo $d = 0,85$ [3]. Dále se předpokládá, že stránky bez odchozích odkazů (dangling stránky) odkazují na všechny ostatní stránky (random surfer na ně může náhodně přejít), ale aby došlo ke spravedlivému rozdělení, jsou náhodné přechody přičteny ke všem stránkám na webu [22].

Po zavedení damping faktoru vypadá vzorec pro výpočet PageRanku stránky i následně

$$r_i = (1-d) + d \sum_{j \in B_i} \frac{r_j}{N_j}. \tag{3.12}$$

Tento vzorec však obsahoval chybu [12], protože Page a Brin ve své práci uvedli, že součet PageRanků všech webových stránek je roven jedné [3], což by s použitím vzorce (3.12) neplatilo. Z tohoto důvodu, by měl tento vztah vypadat následně

$$r_i = \frac{1-d}{n} + d \sum_{j \in B_i} \frac{r_j}{N_j}, \tag{3.13}$$

kde n je celkový počet všech hodnocených stránek.

Celý tento model lze chápat jako **Markovův řetězec**[5], kde stavy jsou jednotlivé webové stránky. Vyjádřeme proto vztahy vedoucí k výpočtu PageRanku maticově, z čehož nejenom že bude aplikace Markovových řetězců zřejmá, ale výrazně se tím zefektivní celý výpočet, protože bude možné počítat PageRank pro všechny stránky (v případě aplikace na sport – pro všechny týmy) najednou. Algoritmus se tak stane jednodušší, rychlejší a méně náročný na výpočetní zdroje [12]. Uvažujme čtvercovou matici $\mathbf{A}$ o rozměrech $n \times n$ s prvky

$$a_{ij} = \begin{cases} 1, & \text{pokud existuje odkaz ze stránky } i \text{ na stránku } j \\ 0, & \text{jinak} \end{cases}. \tag{3.14}$$

Při aplikaci této metody na basketbal, by prvky matice $\mathbf{A}$ vypadaly následujícím způsobem [16] (pozor na indexy prvků)

$$a_{ji} = \begin{cases} 1, & \text{pokud tým } i \text{ porazil tým } j \\ 0, & \text{jinak} \end{cases}, \tag{3.15}$$

což by stejně jako v Keenerově metodě 3.1 a v Elo ratingu 3.2 určovalo pouze vítěze daného utkání resp. počet vítězství týmu i proti týmu j. Prvky této matice lze tedy opět definovat i jinými způsoby, podobně jako v předchozích metodách, a to

$$a_{ji} = S_{ij}, \tag{3.16}$$

[5]Markovovy řetězce jsou náhodné procesy, pro které platí, že pravděpodobnost přechodu do dalšího stavu závisí pouze na stavu současném, nikoliv na stavech předchozích. Více např. [11].

kde S_{ij} může být například počet vstřelených bodů v daném utkání resp. součet skóre ze všech vzájemných zápasů, které tým i získal proti týmu j. Způsoby volby S_{ij} nesoucí informaci o výsledku vzájemného utkání mezi týmy i a j mohou být různé, nemusí se jednat pouze o výsledné skóre utkání, viz kapitola 4.2.

Matici $\mathbf{A}$ ještě upravme následujícím způsobem do tvaru matice $\mathbf{T}$ s prvky

$$t_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}. \tag{3.17}$$

Matice $\mathbf{T}$ má tu vlastnost, že její řádkové součty jsou rovny jedné. Prvky této matice vyjadřují pravděpodobnosti přechodu ze stránky i na stránku j. Jedinou výjimkou jsou dangling stránky, tedy stránky, ze kterých nevede žádný odkaz. V tomto případě jsou v řádcích příslušných dangling stránkám samé nuly, tedy i jejich řádkový součet je roven nule. Proto všechny prvky nulového řádku nahraďme $\frac{1}{n}$, jinými slovy rovnoměrně rozložme pravděpodobnost přechodu na náhodnou webovou stránku mezi všechny stránky. Získáme tak matici $\mathbf{S}$. Formálně zapsáno

$$\mathbf{S} = \mathbf{T} + \frac{1}{n}\mathbf{V}, \tag{3.18}$$

kde $\mathbf{V}$ je matice velikosti $n \times n$, pro jejíž řádky v_i platí

$$v_i = \begin{cases} [1, 1, \ldots, 1], & \text{pokud } i \text{ je dangling stránka} \\ [0, 0, \ldots, 0], & \text{jinak} \end{cases}. \tag{3.19}$$

Takto definovaná matice $\mathbf{S}$ má již tu vlastnost, že je **stochastická**, což znamená, že všechny její prvky jsou nezáporné a zároveň všechny řádkové součty jsou rovny jedné. Tato vlastnost je nezbytná pro Markovovy řetězce. V matici $\mathbf{S}$ však ještě není zahrnut výše zmíněný damping faktor, proto je potřeba provést ještě její poslední úpravu, a to na tvar tzv. Google matice $\mathbf{G}$ [1],

$$\mathbf{G} = (1 - d)\frac{1}{n}\mathbf{I} + d\ \mathbf{S}, \tag{3.20}$$

kde $\mathbf{I}$ je matice samých jedniček velikosti $n \times n$. Lze jednoduše dokázat, že matice $\mathbf{G}$ je rovněž stochastická a navíc nerozložitelná (viz vztah (3.3)). Její prvky vyjadřují pravděpodobnosti přechodu ze stránky i na stránku j už se zahrnutým damping faktorem. Jak již bylo zmíněno výše, jedná se o aplikaci Markovova řetězce. $\mathbf{G}$ je matice pravděpodobností přechodu tohoto řetězce a množina všech webových stránek (resp. všech týmů v soutěži) je množinou stavů. Vektor PageRanků[6] webových stránek $\mathbf{r}$ je pak **stacionárním rozdělením** tohoto Markovova

[6] K výpočtu ratingového vektoru byla napsána funkce *get_rating_pagerank.m* (viz příloha A.7).

řetězce, neboli platí rovnost

$$\mathbf{r} = \mathbf{r}\mathbf{G}, \tag{3.21}$$

kde $\mathbf{r} = [r_1, r_2, \ldots, r_n]$ je řádkový vektor. Stacionární rozdělení popisuje chování Markovova řetězce pro čas jdoucí do nekonečna neboli ustálené rozdělení pravděpodobností. Jelikož se jedná o nerozložitelný řetězec s konečnou množinou stavů, znamená to, že všechny stavy jsou trvalé nenulové, a je tak zaručena nejen existence, ale dokonce jednoznačnost stacionárního rozdělení [11].

Pokud budeme řešit tuto rovnost iteračně s libovolným počátečním rozdělením $\mathbf{r}_0$, tedy $\mathbf{r}_{k+1} = \mathbf{r}_k\mathbf{G}$, posloupnost vektorů $\mathbf{r}_k$ bude konvergovat právě ke stacionárnímu rozdělení $\mathbf{r}$, tedy k vektoru PageRanků. Jinými slovy nám PageRank stránky i říká, jaká je pravděpodobnost příchodu random surfera na stránku i (resp. kolik procent času tam stráví) po velkém množství kliknutí, přičemž jak již bylo zmíněno $\sum_{i=1}^{n} r_i = 1$ (jedná se o pravděpodobnosti).

K hledání stacionárního rozdělení lze rovněž přistupovat také jako k řešení úlohy na vlastní čísla. Podíváme-li se na rovnost (3.21), z definice vlastních čísel můžeme říct, že vektor $\mathbf{r}$ je levým vlastním vektorem matice $\mathbf{G}$ odpovídající vlastnímu číslu $\lambda = 1$. Jelikož matice $\mathbf{G}$ je stochastická, víme, že $\lambda = 1$ je vlastním číslem této matice a zároveň je jejím spektrálním poloměrem [24]. Lze tedy na matici $\mathbf{G}$ aplikovat Perron-Frobeniovu větu (věta 3.1.1) a vektor $\mathbf{r}$ je pak levý Perronův vektor této matice.

Kapitola 4

Implementace

Všechny tři vybrané rankingové modely byly popsány v teoretické části (viz kapitola 3), díky čemuž byl získán postup výpočtu ratingových vektorů. Modely byly aplikovány na reálná data z NBA (viz kapitola 2.1) a získané ratingové vektory byly použity k predikci výsledků utkání s následným srovnáním se skutečnými výsledky odehraných zápasu ve sledovaném období. Kompletní data ze všech patnácti analyzovaných sezon byla importována[7] do softwaru MATLAB R2017b, ve kterém proběhly veškeré výpočty, implementace námi analyzovaných rankingových modelů, predikce výsledků zápasů a vyhodnocení úspěšnosti predikce. Touto implementací se celá tato kapitola zabývá.

4.1 Predikce pomocí rankingových modelů

Jak již bylo uvedeno, rankingové modely jsou obecně používány k sestavování pořadí týmů resp. žebříčků, na jejichž základě pak můžeme hodnotit, který tým si vedl podle daného modelu lépe. Cílem této práce však není sestavení žebříčků a vyhodnocování např. nejlepších týmů jednotlivých sezon nebo určitých období, ale snaha predikovat na základě odehraných utkání výsledky nadcházejících zápasů a zhodnotit úspěšnost všech vybraných modelů v této predikci, pro různé nastavení vstupních parametrů.

4.1.1 Způsob predikce

Hlavní myšlenkou navrženého způsobu predikce výsledků zápasů pomocí rankingových modelů, je průběžné přepočítávání (aktualizace) rankingu resp. ratingových vektorů během sezony a následné porovnání nejaktuálnějšího ratingu dvojice týmů, jejichž utkání chceme predikovat. Tým s vyšším aktuálním ratingem byl pak odhadnut jako vítěz daného utkání. K přepočítávání ratingu byla vždy použita data **pouze z aktuálně hrané sezony**, a to vždy ze všech dosavadně

[7] Data jsou k nalezení v přiloženém souboru *NBAstats.mat* (viz příloha A.2).

odehraných zápasů. Tento způsob výpočtu ratingu nemusí být na začátku sezony příliš vypovídající o síle týmů, neboť je odehráno málo utkání a výkony týmů se nestihnou stabilizovat. Výhodou je fakt, že byla hodnocena pouze aktuální síla resp. síla v dané sezoně a nebyla do výpočtů zahrnuta utkání z předchozích sezon, která už nemusí o aktuální síle týmu vůbec vypovídat.

4.1.2 Frekvence přepočítávání ratingu

Vhodnou frekvencí aktualizace resp. přepočítávání ratingového vektoru by byl zvolený počet soutěžních kol tak, aby všechny týmy měly odehraný vždy stejný počet utkání. V NBA se však nehrají pravidelná soutěžní kola, kterých by se účastnily všechny týmy. Není tak výjimkou, že týmy mají v průběhu sezony rozdílný počet odehraných utkání. Z tohoto důvodu byla **frekvence** přepočítávání ratingu volena vzhledem k celkovému počtu utkání dané sezony (obvykle 1230 utkání). Všechny predikční modely pak byly vyhodnocovány pro různé hodnoty frekvence F. Jako příklad uveďme $F = 30$ utkání. Tato frekvence odpovídá průměrně dvěma odehraným zápasům každého týmů a obvykle znamená 41 přepočtů ratingů za sezonu. Nově vypočítaný vektor ratingů je tedy vždy používán pro predikci nadcházejících F utkání. Problém však nastává pro predikci prvních F utkání sezony, jelikož ještě není vypočítaný žádný ratingový vektor. Proto byl pro predikci těchto prvních utkání každé sezony používán poslední ratingový vektor z předešlé sezony. V první pozorované sezoně, tedy 2002/2003, bylo těchto prvních F utkání použito pouze pro výpočet prvního ratingového vektoru, z hodnocení predikce však byla vynechána (vektor ratingů z minulé sezony není dostupný). V sezoně 2004/2005 přibyl do soutěže tým CHA, pro predikci prvních F utkání byl tomuto týmu přidělen průměrný rating všech ostatních týmů z konce předešlé sezony.

4.2 Způsoby rozdělování bodů

V kapitole 3 bylo u jednotlivých metod uvedeno, že informace použitá k výpočtu ratingu, kterou o sobě dva týmy po vzájemném utkání předají, může být určena **různými způsoby**. Jinými slovy se jedná o rozdělení počtu bodů (bodového ohodnocení) mezi týmy, na základě výsledku a průběhu vzájemných utkání neboli vytvoření matice $\mathbf{A}$ v Keenerově metodě (kapitola 3.1) a v PageRanku (kapitola 3.3) resp. určení prvků W_{ij} v Elo ratingu (kapitola 3.2). V této práci byly použity tři resp. čtyři způsoby rozdělování bodů, a to pro všechny zkoumané modely. Úspěšnost predikce pak byla analyzována pro jednotlivé typy volby.

4.2.1 Rozdělování bodů podle výsledku utkání

Prvním z použitých způsobů je přidělení 1 bodu vítězi utkání a 0 bodů poraženému týmu, viz výrazy (3.4), (3.8) a (3.15). Jelikož v NBA spolu sehrávají všechny týmy v průběhu sezony minimálně 2 utkání, v Keneerově metodě a v PageRanku se tak při určování prvků matice $\mathbf{A}$ jednalo o součet vítězství mezi dvojicemi týmů, v Elo ratingu bylo každé utkání vyhodnocováno individuálně, viz výraz (3.7). Tento způsob ohodnocení sehraného utkání však nepřináší žádnou detailnější informaci o průběhu utkání, pouze definuje vítěze daného utkání a hra poraženého není zohledněna. Nemusí se však nutně jednat o nevýhodu.

4.2.2 Rozdělování bodů podle výsledného skóre

Dalším ze způsobů rozdělování bodů, tedy sestavení matice $\mathbf{A}$ resp. určení prvků W_{ij}, který byl u všech zkoumaných metod v kapitole 3 zmíněn, je použití výsledného skóre utkání. Jedná se o hodnoty S_{ij} z výrazů (3.5), (3.9) a (3.16), kde S_{ij} je celkový počet bodů, které tým i vstřelil týmu j. V Keenerově a Elově metodě byl vždy rozdělen 1 bod v upraveném poměru výsledného skóre utkání resp. součtu všech vzájemných utkání, v PageRanku se body poměrově nepřepočítávaly, pouze sčítaly, ale došlo zde k upravení matice $\mathbf{A}$ na matici stochastickou, viz výrazy (3.17) a (3.18).

4.2.3 Four Factors – váhy podle Olivera

Jak již bylo řečeno, hodnoty S_{ij} nemusí reprezentovat pouze výsledné skóre utkání. V této práci byly k určení těchto hodnot využity tzv. *Four Factors* (čtyři faktory). V roce 2004 navrhl statistik Dean Oliver ve své knize [8] kombinaci čtyř faktorů, které úzce souvisejí s vítězstvím týmu v basketbalovém zápase a přidělil jim váhy. Tyto faktory a jejich váhy podle Olivera určují, proč týmy vyhrávají nebo prohrávají. Jedná se o následující faktory a váhy:

1) Efektivita procentuální úspěšnosti střelby z pole (eFG%), váha **40 %**.
   Tento ukazatel, na rozdíl od standardní procentuální úspěšnosti střelby z pole, zahrnuje fakt, že hodnota střeleckého pokusu za 3 body je o 50 % vyšší než u 2bodového pokusu. Výpočet je ve tvaru

$$\text{eFG\%} = \frac{\text{FGM} + 0{,}5 \cdot \text{3PM}}{\text{FGA}}. \tag{4.1}$$

2) Ztrátové procento (TOV%), váha **25 %**.
   Tato statistika je odhadem počtu ztrát na 100 útoků jednoho týmu. Vy-

počítána je jako

$$\mathrm{TOV\%} = \frac{\mathrm{TOV}}{\mathrm{FGA} + 0{,}44 \cdot \mathrm{FTA} + \mathrm{TOV}}. \tag{4.2}$$

3) Procento útočných doskoků (OREB%), váha **20 %**.
Jedná se o ukazatel, který udává poměr útočných doskoků a počet všech dostupných doskoků po neproměněné střele z pole, což znamená

$$\mathrm{OREB\%} = \frac{\mathrm{OREB}}{\mathrm{OREB} + \mathrm{oppDREB}}, \tag{4.3}$$

kde oppDREB je počet obranných doskoků soupeře.

4) Hodnota trestných hodů (FTrate), váha **15 %**.
Tento faktor je odhadem toho, jak často se tým dostává k trestným hodům a jak často je proměňuje vůči střelám z pole. Výpočet je následující

$$\mathrm{FTrate} = \frac{\mathrm{FTM}}{\mathrm{FGA}}. \tag{4.4}$$

Všechny tyto faktory úzce souvisí s ukončením držení míče během útoku týmu. Obecně jsou tyto faktory používané ke srovnání týmů mezi sebou v každé z těchto kategorií zvlášť. NBA tyto statistické kategorie oficiálně vede na svých webových stránkách.

V této práci byly čtyři faktory počítány pro každé utkání individuálně, a to vždy pro oba soupeřící týmy. Každý z faktorů byl po odehraném utkání zvlášť porovnán a týmu, který v dané kategorii „zvítězil“, byl přidělen příslušný počet bodů získaných za vítězství v této kategorii. Jedinou kategorií, jejíž vítěz je tým s nižší hodnotou, je TOV%. Celkem bylo tímto způsobem rozděleno[8] v každém utkání 10 bodů, a to podle Oliverových vah příslušného faktoru, viz výše. Získané body resp. jejich součet ze všech vzájemných utkání týmů i a j, pak nahradily ve vztazích (3.5), (3.9) a (3.16), hodnoty S_{ij}.

4.2.4 Four Factors – váhy určené pomocí logistické regrese

Posledním ze způsobů rozdělování bodů bylo opět použití *Four Factors* (viz kapitola 4.2.3), tentokrát však s jinými váhami, než které navrhl Dean Oliver. Váhy

[8] Pro celý proces rozdělování bodů podle *Four Factors*, byla vytvořena funkce *four_factors.m* (viz příloha A.8).

čtyř faktorů byly v tomto případě určeny pomocí tzv. logistické regrese. Následující text vychází ze zdroje [13]. Předpokladem logistické regrese je, že náhodná veličina y se řídí alternativním rozdělením pravděpodobnosti s parametrem P. V našem případě náhodná veličina y nabývá hodnot *výhra* (1) nebo *prohra* (0) týmu v daném utkání. Parametr $P \in (0, 1)$ pak vyjadřuje pravděpodobnost výhry a je závislý na veličinách x_1, x_1, x_3, x_4, které reprezentují právě čtyři faktory. Na proměnnou P lze aplikovat logitovou transformaci a za předpokladu závislosti transformované hodnoty na veličinách x_1, x_1, x_3, x_4 ve tvaru lineární regrese s absolutním členem ji lze vyjádřit jako

$$P \approx \frac{1}{1 + e^{-[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4]}}. \tag{4.5}$$

Koeficienty $\beta_0, \ldots, \beta_4$ se obvykle odhadují metodou maximální věrohodnosti, v této práci byly odhadnuty pomocí funkce *glmfit*. Koeficienty $\beta_1, \beta_2, \beta_3$ a β_4 určují váhy čtyř faktorů, které byly vypočítány pro každou analyzovanou sezonu zvlášť a jsou k nahlédnutí v tabulce 4.1. V každé sezoně pak byly pro určení hodnot S_{ij} použity váhy získané z předešlé sezony[9]. V první sledované sezoně (2002/2003) byly použity váhy navržené Oliverem, jelikož data z předešlé sezony nejsou dostupná. Následný postup přidělení bodů je stejný jako v kapitole 4.2.3.

	eFG%	TOV%	OREB%	FTrate
2002/2003	46,33	29,56	9,55	14,57
2003/2004	44,92	30,10	11,65	13,32
2004/2005	45,67	28,37	12,03	13,94
2005/2006	43,29	31,33	12,84	12,55
2006/2007	45,34	31,83	11,65	11,18
2007/2008	42,26	31,89	11,59	14,26
2008/2009	43,66	31,06	13,15	12,12
2009/2010	44,09	26,05	13,93	15,93
2010/2011	45,88	30,30	10,08	13,73
2011/2012	43,08	29,84	14,50	12,58
2012/2013	42,96	35,14	10,57	11,33
2013/2014	45,44	31,38	10,78	12,40
2014/2015	47,09	30,78	9,85	12,28
2015/2016	45,64	31,12	12,18	11,06
2016/2017	46,90	30,91	12,45	9,74

Tabulka 4.1: Získané váhy čtyř faktorů v jednotlivých sezonách

[9] Získané váhy pro všechny sezony jsou součástí funkce *four_factors.m* (viz příloha A.8).

4.3 Vyhodnocení úspěšnosti predikce

K vyhodnocení úspěšnosti predikce výsledků utkání pomocí rankingových modelů bylo napsáno a použito několik funkcí v MATLABu, které jsou přiloženy k této práci[10]. Výstupem však není pouze průměrná procentuální úspěšnost predikce výsledků utkání daných modelů s různým nastavením parametrů (viz kapitola 4.1), ale také podrobnější výstupy zohledňující doplňkové informace (například vliv domácího prostředí na úspěšnost predikce), které byly zkoumány z důvodu návrhu na možné vylepšení modelů. Jednotlivé analyzované výstupy budou popsány v následujících kapitolách.

4.3.1 Procentuální úspěšnost predikce rankingových modelů

Procentuální úspěšnost predikce výsledků zápasů je zásadním výstupem celé této práce. V každé z pozorovaných sezon proběhla predikce podle pravidel popsaných v kapitole 4.1, s nastavením požadovaných vstupních parametrů frekvence přepočítávání ratingu a způsobu rozdělování bodů. Následně byla spočítána procentuální úspěšnost správně odhadnutých výsledků utkání v jednotlivých analyzovaných sezonách a tyto hodnoty pak byly zprůměrovány pro účel srovnávání modelů (kritérium). V případě rovnosti ratingů dvou týmu, které se spolu utkaly, nebyl výsledek tohoto utkání odhadován a ani zařazen do výpočtu procentuální úspěšnosti modelů. Situace rovnosti ratingu však v námi analyzovaných modelech nastává jen výjimečně.

4.3.2 Procentuální úspěšnost predikce sázkové kanceláře

Dalším z výstupů je procentuální úspěšnost predikce výsledků zápasů sázkové kanceláře, která byla určena za účelem srovnání s úspěšností predikce pomocí vybraných modelů v této práci, viz kapitola 4.3.1. Tato úspěšnost byla opět vypočítána pro každou sezonu zvlášť a tyto hodnoty byly následně zprůměrovány. Jak již bylo uvedeno v kapitole 2.1, kurzy sázkových kanceláři jsou dostupné pouze od sezony 2008/2009, pro prvních 7 sledovaných sezon proto nebyla úspěšnost predikce sázkové kanceláře počítána, tedy ani zahrnuta do průměrné úspěšnosti sázkové kanceláře. Pro potřeby této práce stačilo pouze určit, koho sázková

[10] Funkce *season_prediction.m* (viz příloha A.3) a *evaluate_prediction.m* (viz příloha A.4), jsou končné vyhodnocující funkce.

kancelář odhadovala jako vítěze utkání. Při tomto rozhodování nebyl brán ohled na různé možné faktory, které sázkové kanceláře zohledňují (například snaha odhadnout chování sázejících, což může ovlivnit výsledné kurzy), ale pouze bylo posouzeno, na který tým v daném utkání vypsala sázková kancelář **nižší decimální kurz**. Tento tým byl pak brán jako vítěz předpokládaný sázkovou kanceláří. Pokud byly vypsané kurzy na oba týmy shodné, toto utkání nebylo zařazeno do výpočtu procentuální úspěšnosti predikce sázkové kanceláře.

4.3.3 Procentuální úspěšnost predikce v závislosti na vzdálenosti ratingů

Podrobnějším výstupem, který hodnotí úspěšnost predikce, je procentuální úspěšnost v závislosti na vzdálenosti ratingů. Hlavní myšlenkou bylo zjistit, jestli se potvrdí předpoklad, že modely predikují výsledek utkání dvou týmů lépe v případě, že je rozdíl jejich ratingů vyšší, než když je nižší. Všechny zápasy proto byly rozděleny do **pěti skupin** podle vzdálenosti ratingů dvou soupeřících týmů v jednotlivých utkání. Procentuální úspěšnost predikce pak byla počítána pro každou z těchto skupin zvlášť a na závěr byl opět vypočítán k vyhodnocení úspěšnosti modelů průměr ze všech sledovaných sezon v jednotlivých skupinách.

Při každé aktualizaci (viz kapitola 4.1.2) ratingového vektoru $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, kde n je počet všech týmů, byl interval všech možných vzdáleností ratingů $I = \langle 0, \max\{r_1, r_2, \ldots, r_n\} - \min\{r_1, r_2, \ldots, r_n\}\rangle$, rozdělen na 1000 *intervalů vzdálenosti* s ekvidistantním krokem $step$, kde

$$step = \frac{\max\{r_1, r_2, \ldots, r_n\} - \min\{r_1, r_2, \ldots, r_n\}}{1000}. \tag{4.6}$$

Následně byla během porovnávání ratingu dvou soupeřících týmů, tedy predikci výsledku tohoto utkání, vždy zaznamenána absolutní hodnota rozdílu (vzdálenost) jejich ratingů a podle této vzdálenosti bylo utkání zařazeno do příslušného intervalu z celkově 1000 intervalů vzdálenosti. Tímto způsobem proběhlo přiřazení intervalů vzdálenostem vždy během celé sezony. Tyto intervaly vzdálenosti nazvěme lépe ***kategoriemi vzdálenosti***. Maximální možná vzdálenost dvou ratingů je totiž v průběhu sezony při každé aktualizaci ratingového vektoru různá, navíc rozpětí ratingů se s přibývajícím množstvím utkání zvyšuje. Z tohoto důvodu jsou při každé aktualizaci intervaly různě velké a utkání tak byla na základě vzdálenosti ratingů přiřazena vždy do jedné z 1000 kategorií, které jsou relativní vzhledem k aktuálnímu rozpětí ratingů. Na konci hodnocené sezony tak byla výsledkem přiřazená hodnota kategorie vzdálenosti všem predikovaným utkáním

v rozmezí kategorií od 1 do 1000. Při rovnosti ratingů dvou týmů v utkání nebyl opět tento zápas zařazen do dalších výpočtů.

Kategorie vzdálenosti byly rozděleny do pěti skupin, jejichž meze byly voleny tak, aby četnosti zápasů v jednotlivých skupinách byly přibližně stejné. Tento požadavek splňuje rozložení kategorií do skupin zachycené v tabulce 4.2, které odpovídá obvykle 200 až 300 utkáním v každé skupině, což bylo zjištěno na základě testování a zvoleno jako optimální rozložení.

Skupina	Kategorie vzdálenosti
1	$\{1, 2, \ldots, 99\}$
2	$\{100, 101, \ldots, 199\}$
3	$\{200, 201, \ldots, 299\}$
4	$\{300, 301, \ldots, 449\}$
5	$\{450, 451, \ldots, 1000\}$

Tabulka 4.2: Rozdělení do pěti skupin podle kategorií vzdálenosti

Bylo zjištěno, že čím větší vzdálenost v ratingu mezi sebou mají dva týmy, tím méně utkání spolu sehrávají. To vyplývá z logiky věci, jelikož je zřejmé, že možný počet utkání velmi vzdálených (ratingově) soupeřů je malý (např. možné utkání týmů s nejvyšším a nejnižším ratingem je pouze jedno, pokud nepředpokládáme rovnost ratingů) a naopak. Na konci každé sezony byla spočítána procentuální úspěšnost predikce výsledku všech zápasů v jednotlivých skupinách.

4.3.4 Procentuální úspěšnost predikce v závislosti na prostředí

Posledním výstupem hodnotící úspěšnost predikce výsledků zápasů byla matice zachycující procentuální úspěšnost predikce výsledků v závislosti na prostředí. Hlavní myšlenkou bylo zjistit, zda se potvrdí předpoklad, že existuje rozdíl v úspěšnosti predikce v případech, kdy

- bylo predikováno vítězství domácího týmu a zároveň domácí tým skutečně zvítězil,
- bylo predikováno vítězství hostujícího týmu a zároveň hostující tým skutečně zvítězil.

Příklad této matice je zobrazen v tabulce 4.3.

	predikce vítězství domácích	predikce vítězství hostů
vítězství domácích	70 %	40 %
vítězství hostů	30 %	60 %

Tabulka 4.3: Matice úspěšnosti predikce výsledků v závislosti na prostředí

V této konkrétní ukázce lze interpretovat hodnoty tabulky (matice 2×2) tak, že pokud bylo predikováno vítězství domácího týmu, v 70 % případů tento tým skutečně zvítězil, a pokud bylo predikováno vítězství hostujícího týmu, v 60 % případů hosté skutečně vyhráli. V případech, kdy došlo k rovnosti ratingů dvou soupeřících týmů, nebyl opět výsledek tohoto utkání predikován a byl vyřazen ze všech výpočtů (jedná se o výjimečné případy).

Kapitola 5

Analýza výsledků

Tato kapitola se zaměřuje na analýzu a vyhodnocení úspěšnosti predikce výsledků zápasů pomocí všech tří analyzovaných modelů v této práci (Keenerova metoda, Elo rating, PageRank), aplikovaných na získaná data z NBA (viz kapitola 2.1). Implementace predikce pomocí vybraných modelů je popsána v kapitole 4. Zdůrazněme, že všechny hodnoty úspěšnosti predikce byly počítány vždy pro každou analyzovanou sezonu zvlášť (viz příloha A.3) a modely byly následně porovnávány na základě (aritmetického) průměru ze všech sezon (viz příloha A.4).

5.1 Analýza parametrů jednotlivých modelů

Jak již bylo uvedeno v kapitole 4.1, predikční schopnost námi analyzovaných metod byla zkoumána s různými vstupními parametry (frekvence přepočítávání ratingu, způsob rozdělování bodů). Tento odstavec se zaměřuje na vhodnou volbu těchto parametrů resp. určení, pro kterou kombinaci testovaných vstupních parametrů byla získána nejvyšší hodnota průměrné procentuální úspěšnosti predikce výsledků zápasů ze všech analyzovaných sezon pro jednotlivé modely. Zaveďme v této kapitole značení parametrů

- F, pro frekvenci přepočítávaní ratingu (kapitola 4.1.2),
- $0/1$, pro rozdělování bodů podle výsledku utkání (kapitola 4.2.1),
- score, pro rozdělování bodů podle výsledného skóre (kapitola 4.2.2),
- $4F_{logist}$, pro rozdělování bodů podle čtyř faktorů s váhami určenými pomocí logistické regrese (kapitola 4.2.4),
- $4F_{Oliver}$, pro rozdělování bodů podle čtyř faktorů s váhami navrženými Oliverem (kapitola 4.2.3).

Testované hodnoty frekvence F byly zvoleny $F = 1$, tedy přepočet ratingového vektoru po každém odehraném utkání, $F = 15$, $F = 30$ a $F = 60$, což odpovídá

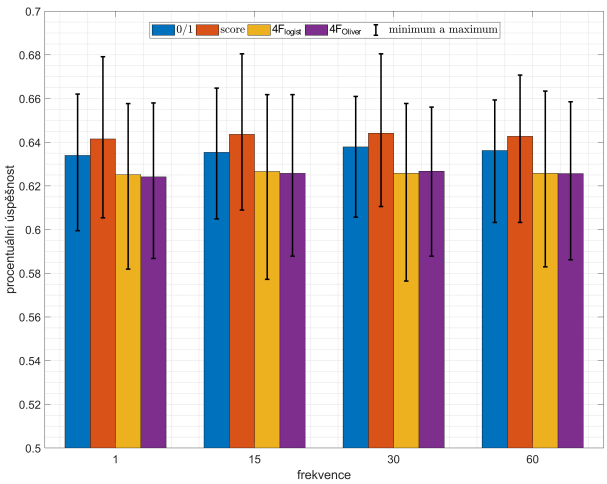
průměrně jednomu, dvěma a čtyřem utkáním jednoho týmu před aktualizací ratingového vektoru. Vyšší hodnoty frekvence již nebyly testovány, jelikož byly považovány za příliš velké na to, aby aktualizace ratingových vektorů stihla včas reagovat na změny sil týmů.

5.1.1 Keenerova metoda

Při výpočtu procentuální úspěšnosti predikce pomocí tohoto modelu bylo vždy dosaženo lepších výsledků bez normalizace prvků a_{ij}, jak popisuje poslední odstavec kapitoly 3.1. Z tohoto důvodu nebyla normalizace nadále vůbec prováděna. Průměrná procentuální úspěšnost predikce podle Keenerovy metody v závislosti na frekvenci a způsobu rozdělování bodů je k vidění v tabulce 5.1. Graficky je znázorněna skupinovým sloupcovým grafem na obrázku 5.1, včetně zachycení vždy nejhůře (minimum) a nejlépe (maximum) predikované sezony, pro každou kombinaci parametrů.

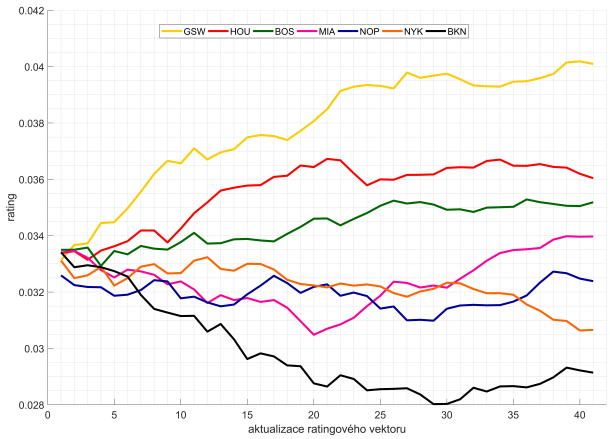
	1/0	score	$4F_{logist}$	$4F_{Oliver}$
$F = 1$	63,40 %	64,15 %	62,52 %	62,42 %
$F = 15$	63,54 %	64,36 %	62,66 %	62,58 %
$F = 30$	63,78 %	**64,41** %	62,58 %	62,67 %
$F = 60$	63,62 %	64,27 %	62,57 %	62,56 %

Tabulka 5.1: Průměrná procentuální úspěšnost predikce podle Keenerova modelu



Obrázek 5.1: Průměrná procentuální úspěšnost predikce podle Keenerova modelu se znázorněním minima a maxima

Z tabulky a grafu je zřejmé, přestože rozdíly nejsou v závislosti na parametrech příliš velké, že nejvyšších hodnot průměrné procentuální úspěšnosti během analyzovaných sezon dosáhly modely s rozdělováním bodů podle výsledného skóre, a to pro všechny testované frekvence, když jako jediné přesáhly všechny tyto modely hranici 64 % průměrné úspěšnosti. Nejvyšší hodnoty pak dosáhl model s parametrem frekvence $F = 30$, a to hodnoty průměrné úspěšnosti 64,41 %. Nadále se budeme zabývat pouze Keenerovo modelem s těmito parametry, které byly posouzeny jako optimální. Pro ilustraci bylo vybráno několik týmů, včetně týmů s nejvyšším a nejnižším ratingem na konci poslední analyzované sezony 2016/2017, jejichž průběh vývoje ratingů v této sezoně je zachycen v grafu na obrázku 5.2. V grafu lze pozorovat rostoucí rozpětí ratingů s přibývajícím počtem odehraných zápasů během sezony.



Obrázek 5.2: Průběh vývoje ratingu vybraných týmů v sezoně 2016/2017, podle optimálního Keenerova modelu

5.1.2 Elo rating

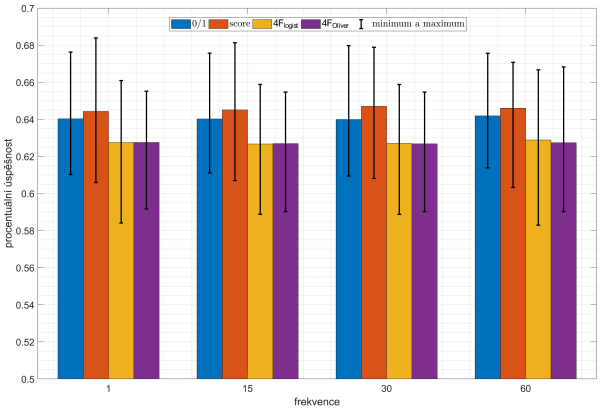
V tomto modelu vystupují kromě parametrů frekvence a způsobu rozdělování bodů také parametry K a ξ, viz rovnosti (3.7) a (3.10). Koeficient rozvoje byl zvolen konstantní, a to $K = 20$. Z této hodnoty vychází statistik Nate Silver ve svém algoritmu pro výpočet Elo ratingu a uvádí ji jako optimální zjištěnou hodnotu K pro NBA [15], nevolí ji však konstantně. Parametr ξ, který ovlivňuje rozpětí ratingu, byl zvolen vyšší, než používá Nate Silver ve svém algoritmu, nebo který je používán pro hodnocení šachových turnajů ($\xi = 400$). Důvodem je to, že v této práci jsou analyzovány sezony jednotlivě, a rozpětí by za tak krátké období bylo příliš malé. Proto byla zvolena hodnota $\xi = 1000$, která rovněž

koresponduje s hodnotou používanou ve zdrojích [2] a [7]. Jako počáteční rating v každé sezoně byla zvolena hodnota 0 pro všechny týmy, přičemž pro naše účely predikce výsledků může být tato hodnota libovolná (stejná pro všechny týmy).

Vhodná volba parametrů frekvence a způsobu rozdělování bodů byla analyzována stejným způsoben jako u Keenerovo metody (kapitola 5.1.1). Průměrná procentuální úspěšnost predikce výsledků utkání podle Elo ratingu je pro různé kombinace těchto parametrů k vidění v tabulce 5.2. Grafické znázornění je pak zachyceno na obrázku 5.3, včetně zobrazení vždy nejhůře a nejlépe predikované sezony pro každou kombinaci parametrů.

	1/0	score	$4F_{logist}$	$4F_{Oliver}$
$F = 1$	64,03 %	64,43 %	62,75 %	62,76 %
$F = 15$	64,02 %	64,51 %	62,67 %	62,69 %
$F = 30$	63,99 %	**64,71** %	62,71 %	62,68 %
$F = 60$	64,19 %	64,60 %	62,88 %	62,74 %

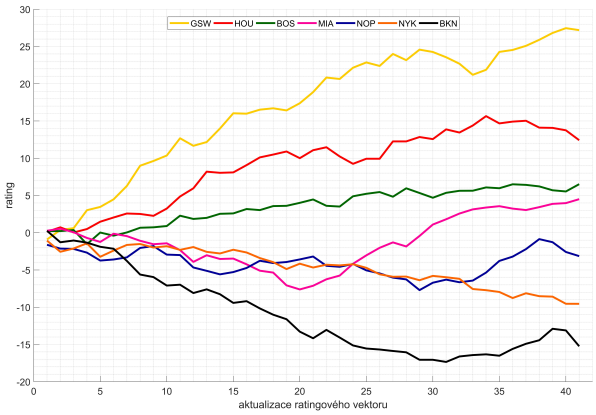
Tabulka 5.2: Průměrná procentuální úspěšnost predikce podle modelu Elo rating



Obrázek 5.3: Průměrná procentuální úspěšnost predikce podle modelu Elo rating se znázorněním minima a maxima

Výsledné hodnoty průměrné procentuální úspěšnosti predikce jsou podobné jako při použití Keenerovy metody. Nejvyšších hodnot dosáhly opět modely s parametrem rozdělování bodů podle výsledného skóre, ze kterých nabývá nejvyšší hodnoty průměrné procentuální úspěšnosti predikce model s parametrem frekvence $F = 30$, a to hodnoty 64,71 %. Modelem s touto optimální kombinací

parametrů se budeme nadále zabývat. Pro ilustraci vývoje ratingu v průběhu sezony při použití tohoto modelu bylo opět vybráno stejných šest týmů a poslední z analyzovaných sezon. Na obrázku 5.4 lze pozorovat, že průběh vývoje ratingů se velmi podobá průběhu vývoje ratingů u Keenerovy metody (obrázek 5.2).



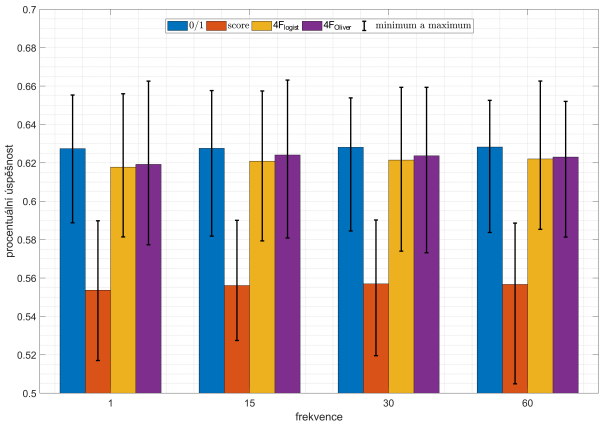
Obrázek 5.4: Průběh vývoje ratingu vybraných týmů v sezoně 2016/2017, podle optimálního modelu Elo rating

5.1.3 PageRank

V tomto modelu se vyskytuje také parametr d, tedy damping faktor (viz kapitola 3.3). Na základně testování byla tato hodnota zvolena $d = 0,75$, což je o 0,1 méně než obvykle volená hodnota při původní aplikaci této metody na hodnocení webových stránek [3]. Průměrná procentuální úspěšnost predikce výsledků podle metody PageRank, v závislosti na parametrech frekvence a způsobu rozdělování bodů, se výrazně lišila od předchozích dvou modelů. Hodnoty průměrné procentuální úspěšnosti jsou vidět v tabulce 5.3. Graficky jsou hodnoty znázorněny na obrázku 5.5.

	1/0	score	$4F_{logist}$	$4F_{Oliver}$
$F = 1$	62,74 %	55,37 %	61,78 %	61,92 %
$F = 15$	62,76 %	55,61 %	62,09 %	62,41 %
$F = 30$	62,81 %	55,70 %	62,15 %	62,37 %
$F = 60$	**62,82** %	55,67 %	62,20 %	62,30 %

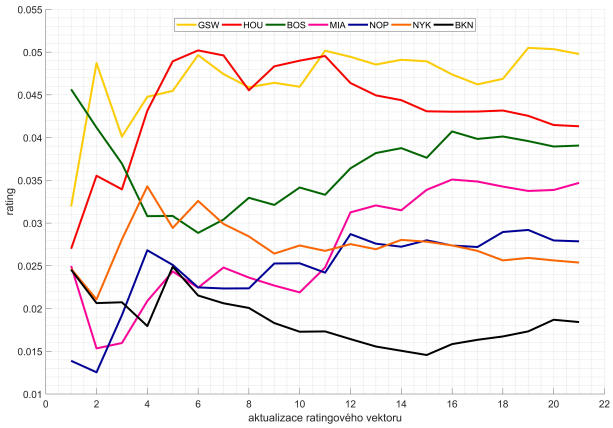
Tabulka 5.3: Průměrná procentuální úspěšnost predikce podle modelu PageRank



Obrázek 5.5: Průměrná procentuální úspěšnost predikce podle modelu PageRank se znázorněním minima a maxima

Zatímco u předchozích dvou metod byly nejúspěšnější modely se způsobem rozdělování bodů podle výsledného skóre, v tomto případě je to právě naopak. Důvodem je přidělování velkého počtu bodů poraženému i vítěznému týmu. Při konstruování stochastické matice (viz kapitola 3.3), je pak oběma týmům (vítězi i poraženému) přidělena téměř stejná hodnota (záleží také na počtu odehraných utkání). Tento model tak není schopen při tak velkém počtu bodů dobře ohodnotit vítěze utkání, z čehož plynou velmi nízké hodnoty průměrné procentuální úspěšnosti predikce. Naopak nejvyšších hodnot dosáhly modely s rozdělováním bodů podle výsledku utkání, tedy pouze s přidělením 1 bodu vítězi utkání. Absolutně nejvyšší hodnoty 62,82 % dosáhl model s kombinací parametrů frekvence $F = 60$ a rozdělováním bodů podle výsledku utkání $(1/0)$, proto se tímto modelem budeme nadále zabývat.

Vyšší frekvence se celkově v modelu PageRank ukázala jako výhodná, a to hlavně díky výpočtům ratingových vektorů ze začátku sezony. Pokud je totiž odehráno v dané sezoně málo utkání, vypočítané ratingy můžou být zkreslené a může tak dojít k nepřiměřeně velkým rozdílům v ratingu mezi týmy s rozdílným počtem odehraných utkání. Vyšší hodnota frekvence zajistí více odehraných utkání každého týmu mezi aktualizacemi ratingového vektoru, což může tento problém eliminovat, a to především na začátku sezony. Průběh vývoje ratingu v sezoně 2016/2017 stejných týmů jako u předchozích modelů je zachycen na obrázku 5.6.



Obrázek 5.6: Průběh vývoje ratingu vybraných týmů v sezoně 2016/2017, podle optimálního modelu PageRank

5.2 Srovnání modelů

V kapitole 5.1 byly vybrány pro každý ze tří námi analyzovaných modelů parametry, pro které vyšla nejvyšší průměrná procentuální úspěšnost predikce výsledků zápasů z námi sledovaných sezon. Tyto tři modely nazveme jako **optimální modely** a pokusíme se je v této kapitole porovnat. Kromě srovnání modelů mezi sebou bude jejich úspěšnost predikce porovnána také s úspěšností predikce **sázkové kanceláře**, viz kapitola 4.3.2.

5.2.1 Úspěšnost predikce modelů a sázkové kanceláře

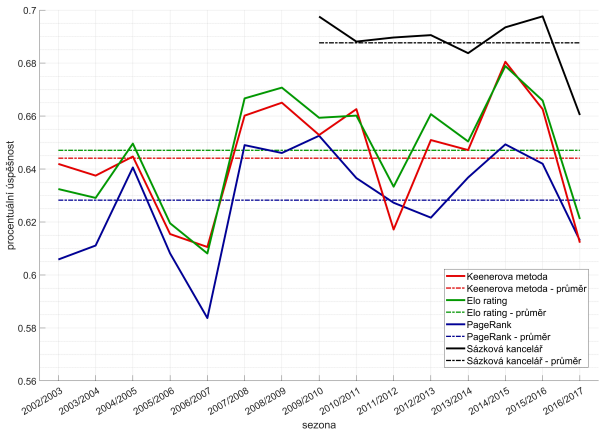
Mezi získanými kurzy sázkových kanceláří (tedy od sezony 2009/2010) se vyskytlo pouze 18 zápasů, kdy došlo k rovnosti kurzů vypsaných na oba soupeřící týmy, tedy došlo k vynechání těchto utkání při hodnocení procentuální úspěšnosti predikce sázkové kanceláře. Situace rovnosti ratingu dvou soupeřících týmů, tedy vynechání tohoto zápasu při hodnocení úspěšnosti predikce modelů (viz kapitola 4.3.1), nenastala při predikci pomocí optimálních modelů ani jednou. V tabulce 5.4 jsou pro porovnání shrnuty výsledné průměrné procentuální úspěšnosti predikce pomocí všech tří optimálních modelů a také průměrná procentuální úspěšnost predikce sázkové kanceláře.

	Všechny dostupné sezony	Od sezony 2009/2010
Sázková kancelář	68,76 %	68,76 %
Elo rating	64,71 %	65,37 %
Keenerova metoda	64,41 %	64,82 %
PageRank	62,82 %	63,49 %

Tabulka 5.4: Průměrné procentuální úspěšnosti predikce podle modelů s optimálními parametry a sázkové kanceláře

Z tabulky 5.4 je zřejmé, že **průměrně nejúspěšnějším** z námi analyzovaných (optimálních) modelů je model Elo rating s průměrnou procentuální úspěšností 64,71 %, což je o 4,05 procentních bodů méně, než je průměrná procentuální úspěšnost sázkové kanceláře, tedy 68,76 %. Jen o 0,30 procentního bodu méně než Elo rating dosáhla průměrná úspěšnost predikce pomocí Keenerovy metody s hodnotou 64,41 %. Nejhůře skončil model PageRank s hodnotou 62,82 %. Poznamenejme, že pokud bychom počítali průměrné hodnoty úspěšnosti predikce pouze ze stejného období, z jakého jsou dostupné kurzy sázkové kanceláře (od sezony 2009/2010), výsledné průměrné hodnoty by byly vyšší, viz tabulka 5.4 (tyto hodnoty slouží pouze k porovnání v této části práce, nadále budeme uvažovat pouze průměrné hodnoty ze všech dostupných sezon).

Graficky je zachycen průměr a vývoj procentuální úspěšnosti predikce výsledků zápasů v průběhu sledovaného období, na obrázku 5.7.



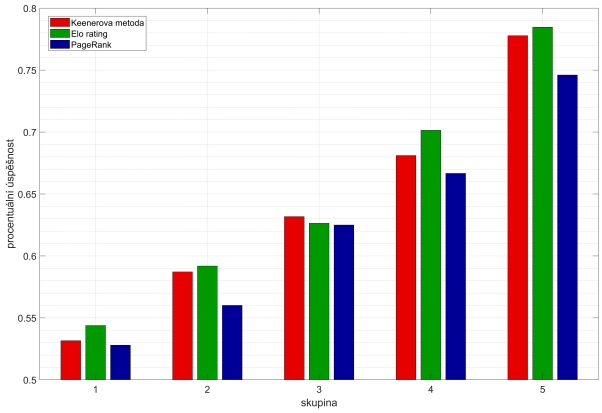
Obrázek 5.7: Srovnání průměru a vývoje procentuální úspěšnosti predikce pomocí optimálních modelů a sázkové kanceláře

Na obrázku 5.7 lze pozorovat podobnost vývoje procentuální úspěšnosti všech

modelů i sázkové kanceláře, nejvíce pak Elo ratingu a Keenerovy metody. Podobnost těchto dvou modelů potvrzují podobné hodnoty průměrné úspěšnosti predikce a podobné vývoje ratingů v jednotlivých sezonách, viz obrázky 5.2 a 5.4. Sázková kancelář byla ve všech sezonách úspěšnější než všechny tři optimální modely. V deseti z celkově patnácti sezon byl nejúspěšnějším ze sledovaných optimálních modelů Elo rating, ve zbylých pěti sezonách Keenerova metoda. Pouze v sezonách 2011/2012 a 2016/2017 nebyl PageRank nejméně úspěšným modelem, ale v obou případech předčil Keenerovu metodu.

5.2.2 Úspěšnost predikce v závislosti na vzdálenosti ratingů

V této části budou vyhodnoceny procentuální úspěšnosti predikce pomocí optimálních modelů v závislosti na vzdálenosti aktuálního ratingu mezi dvěma soupeřícími týmy, viz kapitola 4.3.3. Ve skupinovém sloupcovém grafu na obrázku 5.8 je zobrazena pro všechny tři optimální modely průměrná procentuální úspěšnost predikce během celého sledovaného období v každé z pěti vzdálenostních skupin (viz tabulka 4.2).



Obrázek 5.8: Průměrná procentuální úspěšnost predikce v jednotlivých vzdálenostních skupinách

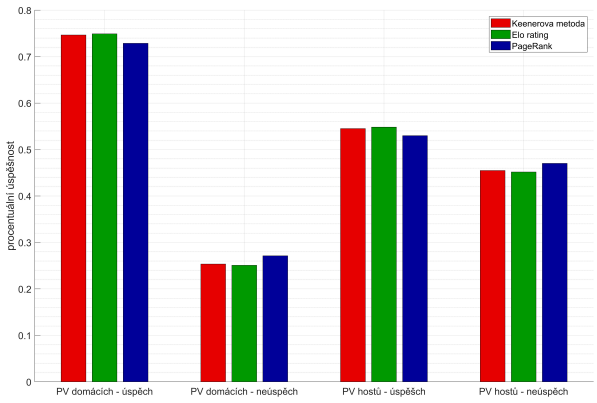
Na první pohled je z grafu zřejmé, že existuje závislost mezi úspěšností predikce a vzdáleností ratingů. Všechny tři modely potvrdili předpoklad, že čím vetší je vzdálenost mezi ratingy dvou soupeřících týmu (tzn. čím větší je rozdíl v jejich síle), tím vyšší je pravděpodobnost správně predikovaného utkání. Konkrétní hodnoty průměrné procentuální úspěšnosti predikce v jednotlivých vzdálenostních skupinách vidíme v tabulce 5.5.

	skupina 1	skupina 2	skupina 3	skupina 4	skupina 5
Keenerova metoda	53,16 %	58,72 %	63,17 %	68,10 %	77,78 %
Elo rating	54,38 %	59,18 %	62,64 %	70,14 %	78,47 %
PageRank	52,80 %	56,01 %	62,50 %	66,67 %	74,60 %

Tabulka 5.5: Průměrná procentuální úspěšnost predikce v jednotlivých vzdálenostních skupinách

5.2.3 Úspěšnost predikce v závislosti na prostředí

Tento odstavec se zaměřuje na vyhodnocení průměrné úspěšnosti predikce pomocí optimálních modelů v závislosti na prostředí, viz kapitola 4.3.4. Na obrázku 5.9 je ve skupinovém sloupcovém grafu zobrazeno shrnutí průměrných procentuálních úspěšností resp. neúspěšností všech tří optimálních modelů v jednotlivých situacích.



Obrázek 5.9: Průměrná procentuální úspěšnost resp. neúspěšnost predikce v závislosti na prostředí (PV = predikce vítězství)

Již z grafu je zřejmá závislost úspěšnosti predikce na prostředí, což potvrzuje náš předpoklad, že v případě predikce vítězství domácího týmu je úspěšnost predikce výrazně vyšší, než v situaci predikce vítězství týmu hostujícího, a to ve všech třech analyzovaných modelech. Konkrétní hodnoty resp. matice úspěšnosti predikce (viz tabulka 4.3) vidíme pro všechny tři optimální modely v tabulce 5.6.

	predikce vítězství domácích	predikce vítězství hostů
vítězství domácích	74,65 %	45,48 %
vítězství hostů	25,35 %	54,52 %

(a) Matice úspěšnosti predikce výsledků v závislosti na prostředí – Keenerova metoda

	predikce vítězství domácích	predikce vítězství hostů
vítězství domácích	74,91 %	45,17 %
vítězství hostů	25,09 %	54,83 %

(b) Matice úspěšnosti predikce výsledků v závislosti na prostředí – Elo rating

	predikce vítězství domácích	predikce vítězství hostů
vítězství domácích	72,87 %	47,02 %
vítězství hostů	27,13 %	52,98 %

(c) Matice úspěšnosti predikce v závislosti na prostředí – PageRank

Tabulka 5.6: Matice úspěšnosti predikce jednotlivých optimálních modelů v závislosti na prostředí

Kapitola 6

Závěr

Tato práce byla zaměřena na rankingové modely a jejich použití k predikci výsledků zápasů NBA. Jako vstupní data byly získány zápasové statistiky všech utkání základních částí od sezony 2002/2003 do sezony 2016/2017. Tato objemná data bylo nejprve potřeba vhodně upravit a zpracovat, čímž se zabývala kapitola 2. Poté proběhlo jejich základní statistické zpracování, ze kterého vyplynul především růst počtu střel na koš (zejména trojkových pokusů) ve sledovaném období, což má za následek zvyšující se počet bodů v utkáních.

Ranking je obecně metoda sestavování pořadí resp. žebříčků na základě stanovených pravidel. V kapitole 3 byly vybrány tři rankingové modely, a to Keenerova metoda, Elo rating a PageRank. Jednotlivé modely byly v této kapitole podrobně popsány a teoreticky bylo vysvětleno jejich matematické pozadí. Objasněny tak byly algoritmy výpočtu, vedoucí k získání ratingových vektorů všech tří vybraných modelů. Tyto algoritmy pak byly použity v praktické části této práce.

V kapitole 4 byl popsán způsob, podle kterého probíhala predikce výsledků utkání pomocí zmíněných tří modelů. Zásadní myšlenkou byla průběžná aktualizace ratingových vektorů během sezony a predikce založená na porovnání nejaktuálnějších hodnot ratingu dvou soupeřících týmů, výsledek jejichž utkání jsme chtěli predikovat. Informace využitá k výpočtu ratingu, kterou o sobě dva soupeřící týmy po vzájemném utkání předají, byla volena třemi resp. čtyřmi způsoby, a to pouze podle výsledku zápasu, pode výsledného skóre zápasu a podle *Four Factors* s váhami podle Deana Olivera a váhami určenými pomocí logistické regrese. Na základě jednoho z těchto čtyř způsobů byly po utkání rozděleny body mezi dva soupeřící týmy.

Úspěšnost predikce výsledků zápasů podle všech tří vybraných modelů byla vyhodnocena pro různé kombinace parametrů frekvence aktualizace ratingu F a způsobu rozdělování bodů. Analýzou výsledků se zabývala kapitola 5. Jako

srovnávací kritérium byla zvolena průměrná hodnota procentuální úspěšnosti predikce výsledků zápasů z celého sledovaného období. Nejvyšší hodnoty tohoto kritéria dosáhl model Elo rating s parametry $F = 30$ a rozdělováním bodů podle výsledného skóre, a to hodnoty 64,71 % průměrné úspěšnosti predikce. V predikci podle Keenerovy metody si nejlépe vedl model se stejnými parametry a výslednou hodnotou 64,41 % průměrné úspěšnosti predikce. Nejvyšší hodnoty průměrné procentuální úspěšnosti predikce s využitím metody PageRank, dosáhl model s parametry $F = 60$ a způsobem rozdělování bodů pouze podle výsledku utkání, a to hodnoty 62,82 %. Žádný z těchto modelů, které byly vybrány jako optimální, nepředčil průměrnou úspěšnost predikce sázkové kanceláře, tedy 68,76 %, a to ani v jedné ze sledovaných sezon (viz obrázek 5.7).

Kromě prosté průměrné procentuální úspěšnosti predikce výsledků zápasů byla rovněž analyzována průměrná úspěšnost predikce v závislosti na vzdálenosti aktuálních ratingů dvou soupeřících týmů (viz kapitoly 4.3.3 a 5.2.2). Potvrdil se předpoklad, že čím vzdálenější (ratingově) jsou týmy od sebe, tím vyšší je úspěšnost predikce. Nejúspěšnější model (Elo rating) dosáhl průměrné hodnoty úspěšnosti predikce v nejvyšší skupině vzdálenosti 78,47 % (viz tabulka 5.5). Analyzována byla také průměrná úspěšnost predikce v závislosti na prostředí (viz kapitoly 4.3.4 a 5.2.3). Opět se potvrdil předpoklad, že pokud je predikováno vítězství domácího týmu, úspěšnost predikce je průměrně vyšší (nejúspěšnější model Elo rating dosáhl hodnoty 74,91 %), než pokud je predikováno vítězství hostujícího týmu (nejúspěšnější model Elo rating dosáhl hodnoty 54,83 %), viz tabulka 5.6.

Díky potvrzení předpokladů o existence závislosti úspěšnosti predikce na vzdálenosti ratingů a prostředí, lze tuto informaci využít jako návrh na vylepšení všech tří rankingových modelů. Naše aplikace vybraných modelů totiž zohledňuje pouze některé výsledné statistiky z utkání a *sílu rozpisu*, nezahrnuje však spoustu dalších situací, které výsledky zápasů ovlivňují a jejich zahrnutí do modelů by tak mohlo výrazně zkvalitnit predikční schopnosti modelů, například právě výhoda domácího prostředí, kdy se nabízí možnost vždy podle určitých pravidel zvýšit rating domácího týmu. Další situace, které ovlivňují utkání a bylo by možné je do modelů zahrnout, jsou například série vítězství/porážek, série velkého počtu utkání v málo dnech (únava), rozhodující zápasy na konci sezony, tankování (úmyslná snaha být posledním týmem z důvodu lepší pozice při draftu), zohlednění ratingu z předešlých sezon, derby, nebo zranění klíčových hráčů. Predikční schopnost modelů by rovněž mohlo vylepšit zaměření se více na volbu vhodných parametrů jednotlivých metod.

Literatura

[1] Austin, D. How Google finds your needle in the web's haystack [online]. *American Mathematical Society Feature Column*. 2006, 10, s. 12. [cit. 2018/03/30]. Dostupné z: `http://www.ams.org/publicoutreach/feature-column/fcarc-pagerank`.

[2] Barrow, D. et al. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*. 2013, 9, 2, s. 187–202. ISSN 1559-0410.

[3] Brin, S. – Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN systems*. 1998, 30, 1, s. 107–117. ISSN 0169-7552.

[4] Elo, A. E. *The rating of chessplayers, past and present*. Arco Pub., 1978. ISBN 0-668-04721-6.

[5] Hoppenbrouwer, J. – Winkels, M. – Schaffner, C. *Sport Ratings* [online]. University of Amsterdam, 2013. [cit. 2018/03/20]. Dostupné z: `https://homepages.cwi.nl/~schaffne/projects/BscAI/2013/`.

[6] Keener, J. P. The Perron-Frobenius Theorem and the Ranking of Football Teams. *SIAM Review*. 1993, 35, 1, s. 80–93. ISSN 00361445.

[7] Langville, A. N. – Meyer, C. D. *Who's# 1?: The Science of Rating and Ranking*. Princeton University Press, 2012. ISBN 978-0691162317.

[8] Oliver, D. *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc., 2004. ISBN 978-1-57488-688-7.

[9] Page, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab, 1999. Dostupné z: `http://ilpubs.stanford.edu:8090/422/`.

[10] Press, W. H. et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007. ISBN 978-0521880688.

[11] Prášková, Z. – Lachout, P. *Základy náhodných procesů I.* MATFYZPRESS, Praha, 2012. Vyd. 2., V Matfyzpressu 1. vyd. ISBN 978-80-7378-210-8.

[12] Raška, O. *Komplexní teoretické i praktické vysvětlení PageRanku* [online]. 2009. [cit. 2018/03/06]. Dostupné z: https://www.zdrojak.cz/clanky/komplexni-teoreticke-i-prakticke-vysvetleni-pageranku/.

[13] Reif, J. *Metody matematické statistiky*. ZČU v Plzni, 2004. ISBN 80-7043-302-7.

[14] Seshadri, A. *Perron root computation – MathWorks* [online]. 2009. [cit. 2018/03/19]. Dostupné z: https://www.mathworks.com/matlabcentral/fileexchange/22763-perron-root-computation.

[15] Silver, N. – Fischer-Baum, R. *How We Calculate NBA Elo Ratings* [online]. 2015. [cit. 2018/03/20]. Dostupné z: https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

[16] Stanek, J. *Ranking NCAA Basketball Teams Using the Google PageRank Algorithm*. 2015. DOI: 10.13140/RG.2.1.1233.0404.

[17] *Citation analysis — Wikipedia, The Free Encyclopedia* [online]. 2018. [cit. 2018/03/25]. Dostupné z: https://en.wikipedia.org/wiki/Citation_analysis.

[18] *Elo rating system — Wikipedia, The Free Encyclopedia* [online]. 2018. [cit. 2018/03/20]. Dostupné z: https://en.wikipedia.org/wiki/Elo_rating_system.

[19] *NBA 3in1 database* [online]. 2018. [cit. 2018/02/22]. Opening odds. Dostupné z: https://www.indatabet.com/bskb-free-3in1.html.

[20] *NBA Advanced Stats* [online]. 2018. [cit. 2018/02/20]. Box Scores. Dostupné z: http://stats.nba.com/teams/boxscores/.

[21] *Odds Portal* [online]. 2018. [cit. 2018/02/22]. NBA Results and Historical Odds. Dostupné z: http://www.oddsportal.com/basketball/usa/nba/results/.

[22] *PageRank — Wikipedia, The Free Encyclopedia* [online]. 2018. [cit. 2018/03/24]. Dostupné z: https://en.wikipedia.org/wiki/PageRank.

[23] *Sports Books* [online]. 2018. [cit. 2018/03/01]. NBA scores and odds archives. Dostupné z: http://www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm.

[24] *Stochastic matrix — Encyclopedia of Mathematics* [online]. 2014. [cit. 2018/03/26]. Dostupné z: http://www.encyclopediaofmath.org/index.php?title=Stochastic_matrix&oldid=35214.

[25] *Šachový ELO systém* [online]. 2007. [cit. 2018/03/20]. Dostupné z: http://www.juniorbb.iprofil.cz/ELO_CZ.pdf.

Příloha A

Zde je uveden seznam souborů, které se nacházejí na přiloženém CD (ve složce *Příloha A*) a jejich stručný popis. Jedná se o upravená vstupní data a funkce použité pro získání všech výsledků této práce. Jak již bylo řečeno, všechny funkce byly napsány a použity v softwaru MATLAB R2017b.

A.1 *NBAstats.xlsx*

Soubor obsahující přehledně upravená vstupní data včetně decimálních kurzů. Data z každé jednotlivé sezony jsou na samostatném listu vždy v jedné tabulce. Soubor obsahuje rovněž list s tabulkou dat ze všech dostupných 15 sezon dohromady a list s kontingenční tabulkou. Soubor byl vytvořen v softwaru Excel 2016.

A.2 *NBAstats.mat*

Soubor obsahující tabulku se všemi získanými daty, v přesném formátu požadovaném pro použití dat v dalších funkcích.

A.3 *season_prediction.m*

Funkce, která vyhodnocuje úspěšnost predikce (viz kapitola 4.3) zvoleného modelu v jedné sezoně, se zadanými parametry frekvence a způsob rozdělování bodů.

[uspesnost_modelu, uspesnost_sazkovky, prctg , matice_HA] =
= season_prediction(NBAstats, sezona, frekvence, typ, model)

- *NBAstats*, je tabulka všech získaných dat v přesném požadovaném formátu (viz příloha A.2)
- *sezona*, je hodnocená sezona, např. {'2016/2017'}
- *frekvence*, je frekvence zápasů, po kterých je ratingový vektor aktualizován (viz kapitola 4.1.2)
- *typ*, je způsob rozdělování bodů (viz kapitola 4.2)

  - *1*... rozdělování bodů podle výsledku utkání (1/0)
  - *2*... rozdělování bodů podle výsledného skóre
  - *3*... rozdělování bodů podle *Four Factors* – váhy určené pomocí logistické regrese
  - *4*... rozdělování bodů podle *Four Factors* – váhy podle Olivera
- *model*, je zvolený model, podle kterého jsou získávány ratingové vektory a predikovány výsledky zápasů
  - *"keener"*
  - *"elo"*
  - *"pagerank"*
- *uspesnost_modelu*, je procentuální úspěšnost predikce výsledků zápasů, podle zvoleného modelu s vybranými parametry v dané sezoně
- *uspesnost_sazkovky*, je procentuální úspěšnost predikce výsledků zápasů podle sázkové kanceláře v dané sezoně
- *prctg*, je procentuální úspěšnost predikce výsledků zápasů, podle zvoleného modelu s vybranými parametry v pěti vzdálenostních skupinách ratingu v dané sezoně, viz kapitola 4.3.3
- *matice_HA*, je matice o rozměrech 2×2, zachycující procentuální úspěšnost predikce výsledků zápasů v závislosti na prostředí, podle zvoleného modelu s vybranými parametry v dané sezoně, viz kapitola 4.3.4
- při odkomentování vyznačené části v kódu této funkce, dojde k vykreslení grafu průběhu ratingu všech týmů v dané sezoně (ve výchozím nastavení je tato část zakomentovaná, aby nedocházelo ke zbytečnému a zpomalujícímu vykreslování grafů průběhu ratingu všech sezon, při použití funkce *evaluate_prediction.m*, viz příloha A.4)

A.4 *evaluate_prediction.m*

Funkce, která vyhodnocuje průměrnou úspěšnost predikce (viz kapitola 4.3) zvoleného modelu ve všech sledovaných sezonách, se zadanými parametry frekvence a způsob rozdělování bodů. Tato funkce v zásadě především průměruje výstupy z funkce *season_prediction.m* (viz příloha A.3) a vykresluje grafy průběhu úspěšnosti predikce ve sledovaném období.

[prumer_uspesnost_modelu, prumer_uspesnost_sazkovka, prumer_prctg, prumer_matice_HA] = evaluate_prediction(model, frekvence, typ)

- výstupy i vstupy mají stejný význam jako ve funkci *season_prediction.m* (příloha A.3) s tím rozdílem, že výstupy nejsou hodnoty pro zadanou sezonu (vstup *sezony* se zde vůbec nenachází), ale jedná se o průměrné hodnoty za celé sledované období, pro které byla získána data
- výstupem této funkce jsou také tři obrázky (grafy), a to vývoj procentuální úspěšnosti predikce zvoleného modelu a sázkové kanceláře, vývoj procentuální úspěšnosti predikce zvoleného modelu v jednotlivých vzdálenostních skupinách a vývoj procentuální úspěšnosti predikce zvoleného modelu v závislosti na prostředí

A.5 *get_rating_keener.m*

Funkce sloužící k získání všech (aktualizovaných) ratingových vektorů během jedné sezony pomocí Keenerovy metody, pro zvolené parametry frekvence a způsob rozdělování bodů.

res_rating = get_rating_keener(NBAstats, sezona, typ, frekvence)

- *NBAstats*, *sezona*, *typ* a *frekvence*, mají stejný význam jako ve funkci *season_prediction.m*, viz příloha A.3.
- *res_rating*, je matice, jejíž sloupce odpovídají jednotlivým aktualizovaným ratingovým vektorům v průběhu celé sezony (poslední sloupec je ratingový vektor zahrnující všechna odehraná utkání v dané sezoně) a řádky odpovídají jednotlivým týmům

A.6 *get_rating_elo.m*

Funkce sloužící k získání všech (aktualizovaných) ratingových vektorů během jedné sezony pomocí Elo ratingu, pro zvolené parametry frekvence a způsob rozdělování bodů.

res_rating = get_rating_elo(NBAstats, sezona, typ, frekvence)

- všechny vstupy i výstupy jsou shodné s funkcí *get_rating_keener.m*, viz příloha A.5

A.7 *get_rating_pagerank.m*

Funkce sloužící k získání všech (aktualizovaných) ratingových vektorů během jedné sezony pomocí PageRanku, pro zvolené parametry frekvence a způsob rozdělování bodů.

res_rating = get_rating_pagerank(NBAstats, sezona, typ, frekvence)

- všechny vstupy i výstupy jsou shodné s funkcí *get_rating_keener.m*, viz příloha A.5

A.8 *four_factors.m*

Funkce, která pro konkrétní utkání rozděluje body (celkově 10) mezi dva soupeřící týmy, a to podle vah *Four Factors* (čtyř faktorů), viz kapitoly 4.2.3 a 4.2.4.

[ff_points_i, ff_points_ii] = four_factors(match, sezona, typ, FGM, 3PM, FGA, TOV, FTA, OREB, DREB, FTM, PTS)

- *match*, je pořadí zápasu ve zvolené sezoně
- *sezona*, je hodnocená sezona, např. {'2016/2017'}
- *typ*, je způsob rozdělování bodů
  - *3*... rozdělování bodů podle *Four Factors* – váhy určené pomocí logistické regrese
  - *4*... rozdělování bodů podle *Four Factors* – váhy podle Olivera
- *FGM, 3PM, FGA, TOV, FTA, OREB, DREB, FTM, PTS*, jsou vektory zápasových statistiky pro celou hodnocenou sezonu, viz kapitola 2.1
- *ff_points_i*, je získaný počet bodů (z 10) jednoho ze soupeřících týmů, podle vah *Four Factors*
- *ff_points_ii*, je získaný počet bodů (z 10) druhého ze soupeřících týmů, podle vah *Four Factors*

A.9 *perron.m*

Jedná se o open source funkci, která vypočítá Perronův, více [14].