FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Predikce vybraných událostí v basketbalovém utkání
Student:	Ondřej Schejbal
Vedoucí:	Ing. Karel Klouda, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	Do konce letního semestru 2020/21

Pokyny pro vypracování

1) Proveďte rešerši zdrojů dat o zápasech a hráčích NBA. Zaměřte se na dostupnost dat o jednotlivých akcích hráčů během utkání.
2) Proveďte rešerši známých metod používaných pro predikce výsledků a jiných událostí (např. počet bodů hráčů, výsledky jednotlivých čtvrtin) v utkáních kolektivních sportů zejm. basketbalu.
3) Ze získaných dat vytvořte vhodné příznaky a na nich experimentálně porovnejte vybrané metody predikce vybraných událostí. Zaměřte se na predikce využívající statistiky o probíhajícím utkání.
4) Výsledky porovnejte také s predikcemi sázkových kanceláří.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Karel Klouda, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 7. února 2020

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Predikce vybraných událostí v basketbalovém utkání

Ondřej Schejbal

Katedra aplikované matematiky
Vedoucí práce: Ing. Karel Klouda, Ph.D.

21. května 2020

Poděkování

V první řadě bych chtěl poděkovat vedoucímu mé práce Ing. Karlu Kloudovi, Ph.D za jeho čas a pomoc při tvorbě této bakalářské práce. Také bych chtěl poděkovat mé rodině a přítelkyni za jejich věčnou podporu a důvěru.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Ve Velkých Popovicích dne 21. května 2020 ……

Odkaz na tuto práci

Schejbal, Ondřej. *Predikce vybraných událostí v basketbalovém utkání.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2020.

Abstrakt

V rámci této bakalářské práce byl vytvořen model predikující celkový počet vstřelených bodů v následujícím vývoji basketbalového zápasu NBA. Predikce jsou založeny na datech z předchozích zápasů a statistik, které v daném zápase již byly zveřejněny. Za účelem získání dat byla provedena rešerše dostupných zdrojů, které se následně povedlo úspěšně využít pro vytvoření dostatečných materiálů k natrénování predikčního modelu. Také byl proveden průzkum již dokončených prací, zabývající se podobnou tématikou. Na základě nabytých poznatků byl zvolen pro predikci model lineární regrese a do výše zmíněných dat byly přidány zajímavé příznaky, které měly zlepšit predikci modelu. Model se povedlo natrénovat a jeho výsledky na testovacích datech se jevily jako příznivé. Avšak úplnou kvalitu výsledků by bylo možné získat pouze při testování na aktuálně hraných zápasech. To bohužel nebylo z důvodu pandemie COVID-19, která probíhala během tvorby bakalářské práce, možné.

Klíčová slova Lineární regrese, NBA, Predikční model, nba_api, Predikce nastřílených bodů, Strojové učení

Abstract

Within this bachelor's thesis, a model predicting the total number of points scored in future match development in NBA basketball match was created. Predictions are based on data from previous games and statistics, which were already published in the ongoing match. In order to obtain the data, a study of existing materials was made, which were then successfully used for the creation of sufficient materials for the training of the prediction model. Also, the research of already finished theses, which are focused on a similar topic, was made. Based on the gathered data, a linear regression prediction model was chosen, and interesting attributes were added to the data mentioned above, which were meant to improve the model's predictions. The model was trained successfully, and it's results on the testing set of data seemed to be favourable. Although the full quality of the results would be possible to obtain by testing the model on currently played matches. Unfortunately, this wasn't possible due to the ongoing COVID-19 pandemic, which took place during the creation of this bachelor's thesis.

Keywords Linear regression, NBA, Prediction model, nba_api, Prediction of goals made, Machine Learning

Obsah

Seznam obrázků

Kapitola 1

Úvod

Sázení na právě hraná utkání (tzv. živé sázky) se stává v poslední době velmi oblíbeným. Predikce událostí, které v právě hraném zápase mohou nastat v závislosti na událostech, které se v tomto rozehraném utkání již uskutečnily, je ovšem problematika, která na rozdíl od predikce před-zápasových událostí není dobře zmapována. Situace tedy otevírá prostor pro vyzkoušení, zda by šlo vytvořit model, který by s rozumnou úspěšností byl schopen takovéto události v právě hraných zápasech predikovat.

Pokud by se povedlo vytvořit dostatečně úspěšný model, otevírá to možnost pro sázkové kanceláře tento model dále rozvíjet a využít ho k lepšímu uzpůsobení svých nabízených kurzů. Aktuálně totiž změna nabízených kurzů reagujících na událost ve hře trvá i několik dlouhých minut, ve kterých sázkaři nemají možnost sázet, a tím sázkové kanceláře přicházejí o zisk.

Tato bakalářská práce se zabývá problematikou predikování vybraných událostí v basketbalovém utkání, přičemž důraz je kladen na predikci událostí v aktuálně hraném zápase. Téma této práce jsem si zvolil i z vlastního zájmu. Predikce různých sportovních statistik mi přijde velmi zajímavá a rád bych na základě získaných znalostí vytvořil dobrý predikční model.

1.1 Cíl

Cílem této bakalářské práce je získat dostupné basketbalové statistiky ve vhodné formě a nad získanými daty sestrojit model, který bude schopný predikovat další průběh aktuálně běžícího zápasu.

Prvním cílem je provést rešerši dostupných zdrojů basketbalových statistik a zvážit možnosti jejich užití v rámci bakalářské práce. Navazujícím cílem je přiblížit již dokončené a publikované práce, které mají podobné téma, přiblížit použité metody a efektivitu modelů, které byly v rámci těchto prací použity. Dalším cílem je z nalezených basketbalových statistik zvolit vhodné příznaky, pomocí nich sestavit datové tabulky a nad nimi sestavit model, který bude

predikovat budoucí události v probíhajícím utkání. Posledním cílem je porovnat výsledky tohoto modelu s kurzy sázkových kanceláří a určit, zda by natrénovaný model měl úspěch, pokud by se na základě jeho predikcí sázelo.

Kapitola 2

Analýza zdrojů dat

Při hledání vhodných zdrojů dat byl kladen důraz především na detaily, které data nabízejí, na jejich kvalitu, aktuálnost a v jaké frekvenci jsou aktualizována. Struktura a uchování dat byla také důležitou podmínkou, protože pokud by například neměl každý hráč svůj unikátní identifikátor (dále jen `Id`), které by ho jednoznačně identifikovalo napříč jednotlivými datovými tabulkami, tak by se orientace v datech velice ztížila a při zpracování dat by mohlo dojít k nechtěným kolizím, které by mohly nastat při zvolení nevhodné unifikační funkce jednotlivých hráčů. Také bylo důležité, aby v datech bylo dostupné větší množství různých příznaků a ne pouze základní statistiky.

Mezi nejzajímavější zdroje dat, na které jsem narazil, patří dvě stránky, které jsou detailněji popsány v následujících sekcích.

Jako hlavní zdroj byla nakonec použita data ze sekce 2.1, ke kterým se podařilo najít API, které umožňuje data získávat ve vhodném formátu. Toto API je blíže popsáno v sekci 4.1.

2.1 NBA stats

Oficiální stránky NBA statistik[1] obsahují nejdetailnější statistiky, které byly během průzkumu nalezeny. Jsou zde dostupné statistiky jednotlivých hráčů, jednotlivých týmů a dokonce jsou zde dostupné i detailní statistiky jednotlivých zápasů. Statistiky jednotlivých zápasů jsou doplňovány již v průběhu zápasu, vždy po každé odehrané čtvrtině.

U jednotlivých tabulek je vždy dostupný filtr obsahující rozmanité možnosti filtrace týkající se například časového období, či příslušnosti hráče k určitému týmu. Během zkoumání stránek jsem narazil na fakt, že lze jednotlivé tabulky filtrovat i strojově a to pomocí přidání dotazu na konec URL příslušné stránky. Například pro filtraci dat ze sezóny 2018/2019, které se

[1] `stats.nba.com`

zároveň týkají pouze hlavní části sezóny, se na konce URL adresy připojí `?Season=2018-19&SeasonType=Regular`.

Dostupná data (a jejich forma) jsou přiblížena v následujících podsekcích.

2.1.1 Statistiky jednotlivých hráčů

Statistiky jednotlivých hráčů jsou dostupné na `stats.nba.com/players`, kde je možné vidět jednotlivé druhy statistik, které se u hráčů sbírají. Je zde také možné dohledat a zobrazit si statistiky jednotlivých hráčů. Data každého hráče je možné vidět v tabulkách, kde u každé z nich je možnost data seřadit podle libovolného sloupce, či vyfiltrovat zobrazená data dle možností filtrů, které jsou dostupné u každé tabulky. Ukázku dat jedné z tabulek je možné vidět na obrázku 2.1.3.

2.1.2 Statistiky jednotlivých týmů

Statistiky jednotlivých týmů jsou dostupné na adrese `stats.nba.com/teams`, kde je možné vidět žebříčky různých statistik, a které týmy v nich vítězí. Je zde také možné si zobrazit statistiky jednotlivých týmů. Statistiky každého týmu jsou dostupné v tabulkách, kde se každá tabulka specializuje na jinou kategorii dat. V každé tabulce je data možné seřadit podle libovolného sloupce, či vyfiltrovat zobrazená data dle možností filtrů, které jsou dostupné u každé tabulky. Ukázku dat je možné vidět na obrázku 2.1.3.

2.1.3 Detailní statistiky konkrétních her

Ke každé odehrané hře je možné nalézt tabulky s detailními statistikami týkajících se dané hry. Mezi ty nejdůležitější patří tabulka Play By Play. Ta obsahuje kompletní výpis všech událostí, které v zápase proběhly. Každá událost vždy obsahuje i časový údaj, podle kterého jsou i data v tabulce seřazena, aby bylo možné vidět postupný vývoj událostí, které se v zápase udály. V datech jsou vždy oddělené jednotlivé čtvrtiny a tyto data jsou aktualizována i během zápasu, na konci každé čtvrtiny. Tato tabulka je jedna z nejdůležitějších pro cíle této bakalářské práce a její podoba je zachycena na obrázku 2.1.3.

BY YEAR	TEAM	GP	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	FP	DD2	TD3	+/-
2019-20	OKC	58	1564	633	262	443	59.1	1	3	33.3	108	183	59.0	196	347	543	141	86	50	65	111	1755.1	21	0	144
2018-19	OKC	80	2669	1108	481	809	59.5	0	2	0.0	146	292	50.0	391	369	760	124	135	117	76	204	2650.0	29	0	389
2017-18	OKC	76	2487	1056	448	712	62.9	0	2	0.0	160	286	55.9	384	301	685	88	128	92	78	215	2392.0	28	0	321
2016-17	OKC	80	2389	905	374	655	57.1	0	1	0.0	157	257	61.1	281	332	613	86	146	89	78	195	2124.6	16	0	195
2015-16	OKC	80	2014	636	261	426	61.3	0	0	0.0	114	196	58.2	219	314	533	62	84	42	89	223	1677.6	6	0	478
2014-15	OKC	70	1771	537	217	399	54.4	0	2	0.0	103	205	50.2	199	324	523	66	99	38	86	222	1536.6	10	0	46
2013-14	OKC	81	1197	265	93	185	50.3	0	0	0.0	79	136	58.1	142	190	332	43	71	40	57	203	947.9	1	0	57

Obrázek 2.1: Celkové statistiky her z základní části sezóny hráče Steven Adams v každé sezóně, ve které působil [1]
BY YEAR - sezóna, TEAM - tým, za který hrál, GP - odehrané hry, MIN - odehrané minuty, PTS - získané body, FGM - proměněné hody, FGA - celkový počet hodů na koš, FG% - (FG/FGA)*100, 3PM - proměněné hody za 3 body, 3PA - celkový počet pokusů na koš, které by byly za 3 body, 3P% - (3PM/3PA)*100, FTM - proměněné trestné hody, FTA - celkový počet trestných hodů, FT% - (FTM/FTA)*100, OREB - získané odražené míče v útočné fázi, DREB - získané odražené míče v obranné fázi, AST - asistence, TOV - turnovers, STL - počet sebrání míče, BLK - počet zablokovaných střel, PF - osobní fauly, DD2 - double doubles, TD3 - triple doubles, +/- - rozdíl bodů týmu když hráč je a není na hřišti

OVERALL	GP	MIN	PTS	W	L	WIN%	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
2018-19	82	3981	8605	22	60	.268	3266	7205	45.3	745	2123	35.1	1328	1695	78.3	718	2799	3517	1796	1159	603	351	1663	-690

LOCATION	GP	MIN	PTS	W	L	WIN%	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
Home	41	1983	4247	9	32	.220	1623	3615	44.9	359	1049	34.2	642	824	77.9	378	1375	1753	927	554	289	156	809	-402
Road	41	1998	4358	13	28	.317	1643	3590	45.8	386	1074	35.9	686	871	78.8	340	1424	1764	869	605	314	195	854	-288

WINS/LOSSES	GP	MIN	PTS	W	L	WIN%	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
Wins	22	1091	2497	22	0	1.000	928	1894	49.0	215	559	38.5	426	532	80.1	189	822	1011	511	325	178	103	464	176
Losses	60	2890	6108	0	60	.000	2338	5311	44.0	530	1564	33.9	902	1163	77.6	529	1977	2506	1285	834	425	248	1199	-866

Obrázek 2.2: Přehled 3 tabulek týmu Chicago Bulls v sezóně 2018/19. První shora zachycuje statistiky týmu za celou sezónu, druhá shora zachycuje statistiky v závislosti na lokaci utkání a třetí zachycuje statistiky ve vyhraných a prohraných utkáních[2] OVERALL - sezóna, LOCATION - místo konání zápasu, GP - odehrané hry, MIN - odehrané minuty, PTS - získané body, W - počet výher, L - počet proher, WIN% - (W)/(GP), FGM - proměněné hody, FGA - celkový počet hodů na koš, FG% - (FG/FGA)*100, 3PM - proměněné hody za 3 body, 3PA - celkový počet pokusů na koš, které by byly za 3 body, 3P% - (3PM/3PA)*100, FTM - proměněné trestné hody, FTA - celkový počet trestných hodů, FT% - (FTM/FTA)*100, OREB - získané odražené míče v útočné fázi, DREB - získané odražené míče v obranné fázi, AST - asistence, TOV - turnovers, STL - počet sebrání míče, BLK - počet zablokovaných střel, PF - osobní fauly, +/- - rozdíl bodů týmu, když hráč je a není na hřišti

Milwaukee Bucks		Charlotte Hornets
	Start of Q1	
	12:00	Jump Ball Biyombo vs. Lopez: Tip to Antetokounmpo
MISS Matthews 27' 3PT Jump Shot	11:45	
Matthews REBOUND (Off:1 Def:0)	11:39	
Matthews 26' 3PT Jump Shot (3 PTS) (Antetokounmpo 1 AST)	11:31 **3 - 0**	
	11:16 **3 - 2**	**Rozier 3' Driving Finger Roll Layup (2 PTS)**
Bledsoe Bad Pass Turnover (P1.T1)	10:56	Rozier STEAL (1 STL)
	10:50 **3 - 4**	**Biyombo 3' Hook Shot (2 PTS) (Washington 1 AST)**
Antetokounmpo Violation:Defensive Goaltending (K.Lane)	10:50	
Antetokounmpo 7' Turnaround Fadeaway (2 PTS)	10:26 **5 - 4**	
	10:13	MISS Graham 7' Driving Floating Jump Shot
Bucks Rebound	10:13	
Antetokounmpo 2' Finger Roll Layup (4 PTS)	9:53 **7 - 4**	
	9:53	Biyombo S.FOUL (P1.T1) (S.Wall)
Antetokounmpo Free Throw 1 of 1 (5 PTS)	9:53 **8 - 4**	

Obrázek 2.3: Prvních 14 událostí z první čtvrtiny zápasu Milwaukee Bucks proti Charlotte Hornets ze dne 1.3.2020. Každá z událostí se vždy týká konkrétního hráče, a je tedy zapsána v sloupci týmu, ke kterému daný hráč patří. Pokud událost změnila aktuální skóre, je nový bodový stav zachycen v prostředním sloupci.

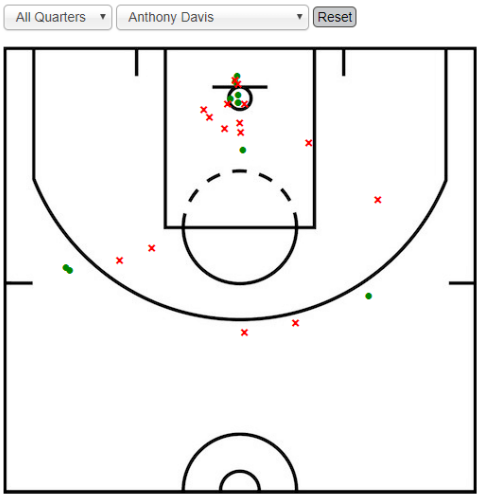
2.1.4 Legální dostupnost dat

Dostupná data jsou vlastněna, případně licencována, majitelem stránek. Jejich stáhnutí v rámci osobního, případně nekomerčního využití a zpracování je povoleno. Stahování obrázků zachycujících hráče, či některá místa je zakázáno[3].

2.2 Basketball reference

Basketball reference sídlící na stránce `basketball-reference.com` je web věnující se statistikám basketbalu, který se snaží statistiky nejen zprostředkovat, ale nabízí nad nimi možnost i provádět jednoduché operace a vizualizace. Například je možné si z dostupných dat nechat vykreslit mapu míst, ze kterých vybraný hráč střílel během zápasu, jak je možné vidět na obrázku 2.2.

Data, která jsou na webu k dispozici obsahují statistiky jednotlivých hráčů, týmů i jednotlivých her. Stránka působí dojmem, že pouze kopíruje data z oficiálních NBA stats, zmíněných v sekci 2.1 Po bližším zkoumání je možné si všimnout, že data jsou vždy obohacena o nějakou informaci navíc, která



Obrázek 2.4: Souhrn střeleckých pokusů Anthonyho Davise za Los Angeles Lakers ve hře proti Boston Celtics dne 23.února 2020. Zelené body zachycují proměněné hody. Červené kříže zachycují neproměněné hody.

je ve všech případech spíše nepodstatného významu z pohledu cíle a zájmu této práce. Byly to například již zmíněné vykreslovací funkce nad daty, nebo informace o historii změn názvu u jednotlivých týmů. Stránka také obsahuje odkazy na různé články obsahující analýzu vztahů v týmech, informace o aktuálních zraněních apod. Bohužel tyto informace jsou uloženy v textech na externích stránkách, tedy v těžko získatelné podobě.

Jednotlivé tabulky a data v nich jsou aktualizovány denně. To znamená, že oficiální stránky NBA statistik mají oproti této stránce velkou výhodu v aktuálnosti dat. Aktuální výsledky právě hraných zápasů tedy nejsou na stránkách k dispozici.

2.2.1 Legální dostupnost dat

Zachycení, ukládání a jakékoliv zpracování dat je bez souhlasu majitele stránky zakázáno [4]. Vlastník stránky vydal oficiální oznámení, že o data byl velký zájem ze strany studentů i akademických organizací. Nejprve se snažili jim vycházet vstříc, ale vzhledem ke kontinuálnímu velkému množství žádostí o udělení práv na zpracování dat se rozhodli oficiálně oznámit, že pokud kdokoliv chce data užívat, bude se muset s firmou domluvit na částce, jejíž minimální hodnota je stanovena na 1 000 dolarů [5]. Nedostupnost těchto dat byl jeden z hlavních důvodů, proč hlavním zdrojem dat pro tuto bakalářskou práci byla vybrána data, zmiňovaná v sekci 2.1.

Kapitola 3

Rešerše prací na podobné téma

V této kapitole jsou rozebrány odborné práce, které byly nalezeny při průzkumu dokončených a zveřejněných prací na stejné, či podobné téma s touto bakalářskou prací. U každé práce je zmíněno její téma a jsou přiblíženy použité metody a výsledky získané během práce.

3.1 Predikce celkového počtu bodů v jednotlivých hrách v NBA

Ve své práci [6] se Alameda-Basora z IOWA State University věnuje predikci celkového počtu bodů v zápasech NBA pomocí bayesovských sítí. Jeho cílem bylo postavit kvalitní model bayesovské sítě, pomocí kterého by na základě dat z NBA databáze a hlavně dostupných dat z aktuálně hraného zápasu byl schopný získávat pravděpodobnosti dosažení počtu bodů, z právě hraného zápasu, z intervalů, na které je možné v sázkových kancelářích sázet. Na základě těchto skutečností pak nechat svůj model rozhodnout zda vsadit 100 dolarů, či ne. Poté porovnal výsledky sázení svého modelu s amatérskými sázecími strategiemi, informovanými pomocí jednodušších prediktivních modelů. Pro naučení a zkonstruování sítě použil data ze zápasů z posledních 5 sezón, vyjma zápasů, které šly do prodloužení.

Natrénovaný model se ukázal být efektivní a sázky dle jeho predikcí generovaly profit v intervalu od 6 do 10 %. Basor poukázal na to, že jednodušší metody jako např. Naivní Bayesův klasifikátor nebyly schopny generovat jakýkoliv profit.

3.2 Predikce výsledků z hlavní sezóny týmů z NBA

V této práci [7] se Yang zaměřuje na analýzu korelace mezi statistikami individuálních hráčů a celkovým výkonem jejich týmu. Na základě této analýzy pak sestavil model, který je možné použít k odhadu výsledků jednotlivých zápasů

v NBA na základě běžných statistik hráčů. Pro naučení svého modelu použil data z posledních 20 sezón, které analyzoval v jazyce R za použití metody nejmenších čtverců pro regresní data.

V závěru své práce zmiňuje, že většina lidí své predikce zakládá pouze na statistikách jednotlivých hráčů a přitom přehlíží korelaci individuálních výsledků s kolektivními výkony týmu jako celku. Ve svém modelu použil také zajímavou metriku Win Share. Tato metrika by mohla být užitečná pro tuto bakalářskou práci.

3.2.1 Win Share

Win Share vyvinul sportovní expert Bill James. Jedná se o metriku, která zachycuje podíl hráče na výhře svého týmu na základě celosezónních statistik jednotlivých hráčů a týmů. James zkonstruoval tuto techniku pro baseball, ale dá se použít i na jiné sporty a objevuje se i v různých modifikacích. Obecně se metrika Win Share skládá ze dvou částí – Offensive Win Share a Defensive Win Share, které se vypočítávají z individuálních statistik hráčů. Finální hodnota Win Share metriky pak vznikne součtem hodnot Offensive Win Share a Defensive Win Share. V Jamesově originální verzi byla jedna výhra týmu rovna 3 Win Share bodům a byla navžena tak, že hodnota Win Share nemohla nabývat záporných hodnot.[8]

Pro basketbalové statistiky existuje upravená verze, která se od původní Jamesovy verze liší ve dvou věcech. Jednou z nich je, že připouští negativní hodnotu Win Share metriky. Další je, že jedna výhra týmu je rovna 1 Win Share bodu.[9]

Výpočet Win Share metriky vychází z mnoha statistik. Detaily výpočtu jsou popsány na stránce věnující se popisu této modifikované verze výpočtu Win Shares pro basketbal.[9]

3.3 Predikce výsledků v NBA

Cílem práce [10] Parka bylo postavit užitečné modely pro předzápasové predikce, ale i predikce právě probíhajících zápasů v NBA. Také se ve své práci věnuje výpočtu potencionálního zisku, pokud by podle svého modelu sázel.

Jedním z modelů, který popisuje, je Elo rating model, který vymyslel Arpard Elo původně pro výpočet pravděpodobností výhry u šachových hráčů. Pro basketbal se Park rozhodl tento model rozšířil na 3 podmodely. Podmodel pro domácí hry, venkovní hry a pro posledních x her, protože je obecně známo, že výkony týmů se liší v domácích a venkovních zápasech, a proto je třeba tyto hry odlišit. Posledních pár her je také třeba počítat zvlášť, protože zachycují aktuální formu týmu.

Park v závěru své práce zdůrazňuje, že pokud by sázel dle svého modelu, tak by dle jeho výpočtů, nedosáhl skoro žádného profitu. Proto se rozhodl zúžit používaná data na posledních 5 sezón a nasadit optimalizační techniku

@RISK[11]. @RISK je optimalizační technika, která provádí analýzu dat, ve kterých se vyskytuje rizikový faktor. @RISK je software vyvinutý pro práci s buňkami v Excelových tabulkách. Tuto techniku Park použil k zlepšení tzv. testů výdělečnosti. Tyto testy zakládal na kurzech sázkových kanceláří a používal je ve své práci jako pomocnou techniku k určení, na který zápas vsadit. Po aplikaci těchto změn se Parkovi výsledky jeho predikčního modelu zlepšily a drobný zisk již zaznamenal.

Kapitola 4

Proces získávání dat

Pro co nejefektivnější získání dat ze stránek NBA statistik byl proveden průzkum dostupných API, přes která by bylo možné data co nejjednodušeji získávat. Bohužel pro stránku NBA statistik žádné oficiální API neexistuje. Bylo tedy nutné hledat API od externích zdrojů. Takových existuje větší množství, ale většina není volně dostupná. Jako nejlepší řešení z externích poskytovatelů byl nalezen RapidAPI, který nabízí přístup k databázi statistik NBA na endpointu `api-nba`[2]. Ten je dostupný pod tzv. freemium licencí, což znamená, že k němu má uživatel přístup po bezplatné registraci a přístup má v omezeném množství, a to 1 000 requestů za den.[12] Tento zdroj se zdál být jako velmi zajímavý, ale nakonec nebyl využit a to hlavně z důvodu, že se povedlo najít zdroj s lepším formátem výstupních dat.

Během hledání dostupných poskytovatelů endpointů byla nalezena volně dostupná knihovna napsána v jazyce Python s názvem `nba_api`[3], která stahuje data přímo z oficiálních stránek NBA a nabízí přístup k datům v přehledné a dobře zpracovatelné formě. Tento zdroj se díky svému zpracování, dobré dokumentaci a ideálnímu formátu výstupních dat ukázal být jako nejvhodnější způsob získávání dat pro tuto bakalářskou práci a je detailněji popsán níže. Knihovna sice obsahuje vhodné endpointy potřebné pro tuto práci, ale pro správné sestavení výsledných tabulek bylo potřeba ještě tato data zpracovat. Důvod proč bylo nutné data ještě dále zpracovávat a jak to bylo provedeno, je popsán níže v sekci 4.2.

4.1 Python knihovna nba_api

Knihovna nba_api má dle svého popisu sloužit jako API client pro stránky NBA statistik, tedy jejím účelem je zjednodušit přístup k endpointům dané stránky a zprostředkovat jejich širší dokumentaci. Knihovna je volně dostupná

[2] `rapidapi.com/api-sports/api/api-nba`
[3] `github.com/swar/nba_api`

z repozitáře na GitHubu. Zde je možné najít veškerou dokumentaci včetně krátkých příkladů použití knihovny. Knihovna je otevřena jakékoliv pomoci ve vývoji a opravování případných chyb a díky tomu je přibližně jednou měsíčně aktualizována.

Díky zabudovaným metodám jsou nabízeny celkem 3 typy návratového formátu dat, a to `JSON`, `dictionary`[4] a `DataFrame`[5] . Na kteroukoliv odpověď ze serveru je možné zavolat jednu z metod zachycenou na obrázku 4.1, pomocí které se vrácená odpověď převede na požadovaný formát.

```
# Returns data in a JSON string.
player_info.available_seasons.get_json()

# Returns data in a dictionary.
player_info.available_seasons.get_dict()

# Returns the data set in a pandas DataFrame.
player_info.available_seasons.get_data_frame()
```

Obrázek 4.1: Příklad konverze odpovědi serveru na jiný formát

Knihovna má také zabudované hledání jednotlivých hráčů a týmu pomocí regulárního výrazu. Mezi nejdůležitější části nacházející se v knihovně patří moduly `players.py`, `teams.py` a třída `Endpoint` obsahující mnoho užitečných metod na posílání requestů a následnou práci s nimi. Tyto části jsou popsány v sekcích níže. Pro lepší představu použití knihovny je možné na obrázku 4.2 vidět ukázku kódu pro získání všech odehraných her týmu Los Angeles Lakers a část dat, které dotaz vrátil. Vrácená data obsahují stejné sloupce, které je možné najít v tabulkách na stránkách NBA statistik.[13]

4.1.1 Modul players.py

Tento modul slouží k přístupu k informacím o hráčích evidovaných na stránce `stats.nba.com` bez nutnosti zasílání explicitních requestů. Modul nabízí několik funkcí, z nichž většina je zaměřená na vyhledávání hráčů pomocí regulárního výrazu[6]. Z tohoto modulu byla v této práci využita jen funkce `get_active_players()`, která vrací seznam všech aktuálně aktivních hráčů v soutěži NBA.

Každý hráč je reprezentován v seznamu jako dictionary se strukturou, kterou je možné vidět na obrázku 4.3.

[4] Dictionary je kolekce dvojic (klíč-hodnota) objektů v pythonu
[5] Objekt z knihovny pandas sloužící na uložení dat
[6] `regular-expressions.info`

In [1]:
```
from nba_api.stats.static import teams
from nba_api.stats.endpoints import leaguegamefinder

nba_teams = teams.get_teams()
# Select the dictionary for the Lakers, which contains their team ID
lakers = [team for team in nba_teams if team['abbreviation'] == 'LAL'][0]
lakers_id = lakers['id']

# Query for games where the Lakers were playing
gamefinder = leaguegamefinder.LeagueGameFinder(team_id_nullable=lakers_id)
# The first DataFrame contains relevant data for our task
games = gamefinder.get_data_frames()[0]
games.head()
```

Out[1]:

	SEASON_ID	TEAM_ID	TEAM_ABBREVIATION	TEAM_NAME	GAME_ID	GAME_DATE	MATCHUP	WL
0	22019	1610612747	LAL	Los Angeles Lakers	0021900861	2020-02-25	LAL vs. NOP	W
1	22019	1610612747	LAL	Los Angeles Lakers	0021900842	2020-02-23	LAL vs. BOS	W
2	22019	1610612747	LAL	Los Angeles Lakers	0021900833	2020-02-21	LAL vs. MEM	W
3	22019	1610612747	LAL	Los Angeles Lakers	0021900817	2020-02-12	LAL @ DEN	W
4	22019	1610612747	LAL	Los Angeles Lakers	0021900801	2020-02-10	LAL vs. PHX	W

5 rows × 28 columns

Obrázek 4.2: Příklad kódu pro získání všech odehraných her týmu Los Angeles Lakers a náhled části získaných dat

```
{'id': 203500,
 'full_name': 'Steven Adams',
 'first_name': 'Steven',
 'last_name': 'Adams',
 'is_active': True}
```

Obrázek 4.3: Příklad dictionary, který zachycuje informace o hráči Stevenu Adamsovi

4.1.2 Modul teams.py

`Teams.py` poskytuje přístup k informacím o jednotlivých týmech, které se účastní hlavní soutěže NBA bez nutnosti zasílání explicitních requestů. Je zde definováno několik funkcí, z nichž většina je, podobně jako v modulu `player.py`, zaměřená na vyhledávání týmů pomocí regulárního výrazu. Z tohoto modulu byla v této práci využita jen funkce, která je definovaná jako `get_teams(regex_pattern, row_id)`. Tato funkce se dá zavolat i bez parametrů a v takovém případě vrátí seznam všech týmů v NBA, což bylo přesně to, co bylo v této práci explicitně využito.

Každý tým je reprezentován v seznamu jako dictionary se strukturou, kte-

rou je možné vidět na obrázku 4.4.

```
{'id': 1610612737,
 'full_name': 'Atlanta Hawks',
 'abbreviation': 'ATL',
 'nickname': 'Hawks',
 'city': 'Atlanta',
 'state': 'Atlanta',
 'year_founded': 1949}
```

Obrázek 4.4: Příklad dictionary, který zachycuje informace o týmu Atlanta Hawks

4.1.3 Třída Endpoint

Třída `Endpoint` obsahuje soubor funkcí, kde každá funkce realizuje volání endpointu na serveru `stats.nba.com` a vrací požadovanou tabulku závisející na volbě funkce. Detaily těchto endpointů, možnosti parametrů, které jdou při volání použít a formáty odpovědí je možné dohledat v repozitáři[7]. V následujících odstavcích jsou popsány endpointy, které byly v této bakalářské práci použity.

Endpoint `teamgamelogs`, který zpřístupní týmové statistiky v jednotlivých zápasech v sezóně, kterou dostane specifikovanou jako jeden z argumentů. U tohoto endpointu je třeba kromě sezóny předat jako argument i `Id` týmu, jehož záznamy se mají získat. Statistiky, které jsou v těchto datech obsažené, zahrnují všechny týmové statistiky, které se v současné době v basketbalu zaznamenávají a mezi ně patří například počet hodů na koš, počet trestných hodů týmu a jiné.

Endpoint `playergamelogs` se používá k zpřístupnění detailních statistik hráče v zápasech, které odehrál. Argumenty k vyspecifikování dat, které se v této práci používají, jsou `Id` hráče, `Id` sezóny a číslo čtvrtiny zápasu, z jejíhož konce hráčovy statistiky chceme. Pokud předáme jako číslo čtvrtiny zápasu nulu, získáme kompletní hráčovy statistiky včetně statistik z případného prodloužení, pokud k němu ve hře došlo. V datech jsou obsažené všechny statistiky, které se u hráčů v basketbalu zaznamenávájí, což je například počet strávených minut na hřišti, počet střeleckých pokusů a mnoho dalších.

Endpoint `boxscoretraditionalv2` se používá k získání seznamu hráčů, kteří se účastnili konkrétního utkání. Je tedy třeba předat jako argument `Id` zápasu.

Endpoint `playbyplay` se používá k získání tabulky `Play By Play` pro jeden konkrétní zápas. Tato tabulka (a co je možné v ní najít) je přiblížena v sekci 2.1.3. Tomuto endpointu je nutné předat jako argument `Id` zápasu.

[7] `github.com/swar/nba_api/tree/master/docs/nba_api/stats/endpoints`

V získané tabulce jsou zachyceny veškeré události, které se v zápase udály. Je to také jedna z mála tabulek, u které je možné zjistit průběžné skóre na konci jednotlivých čtvrtin zápasu.

4.1.4 Legální užití knihovny

Knihovna je volně dostupná a její užití, šíření a případné vlastní úpravy jsou povoleny. Pokud by se některá část kódu z knihovny monetizovala, je třeba zahrnout do licence produktu také část licence, která je vyznačena v podmínkách užití knihovny.[14]

4.2 Zpracování dat dostupných z nba_api

Knihovna `nba_api` popsaná v sekci 4.1 obsahuje vhodné endpointy pro tuto práci, ale pro správné sestavení výsledných tabulek s potřebnými údaji bylo třeba získat data z několika různých endpointů a bylo potřeba enpointy volat opakovaně, např. pro každý tým minimálně 3krát. Při prvotním zpracování dat se ukázalo, že tento zvýšený počet volání způsobuje nežádoucí problémy. Hlavním problémem byla reakce serveru na velké množství dotazů během krátkého časového úseku. Server po chvilce zablokuje zasílání veškerých odpovědí, a to klidně i na dobu delší než 5 minut. Toto omezení by značně omezovalo rychlost programu na vytváření výsledných tabulek určených přímo pro predikční model. Tato omezená rychlost vytváření by byla překážkou hlavně v části experimentování s jednotlivými příznaky za účelem zlepšení predikovaných výsledků.

Tento problém se podařilo vyřešit tím, že byly vytvořeny 3 Jupyter Notebooky zaměřující se pouze na stažení dat, která byla nezbytná pro vytvoření konečných tabulek, a tyto data uložily ve vhodném formátu na lokální disk. Po stažení dat bylo nejen možné s daty pracovat bez připojení k internetu, čímž se předešlo nechtěným výpadkům, které by mohly přerušit proces vytváření v jeho průběhu, ale také se tím vyřeší mnoho problémů, které vznikaly s větším množstvím requestů na server.

Jedním z problému, který se podařilo odstranit, byla blokace ze strany serveru. Předešlo se jí tím, že se mezi každým odeslaným requestem vždy počkalo 1 vteřinu, což se ukázalo být dostatečným rozestupem mezi requesty, pro to, aby server nezablokoval posílání odpovědí našemu programu. Toto řešení bylo sice časově náročné, ale vzhledem k tomu, že data bylo potřeba stáhnout pouze jednou, ukázalo se to být vhodným řešením. Kdyby se tyto notebooky na stahování dat nevytvořily a použila by se technika časových rozestupů mezi requesty rovnou při procesu vytváření tabulek pro model, bylo by časové zdržení mnohonásobně větší.

Všechna data se týkala období od sezóny 2014/15 do současné sezóny 2019/20 a notebooky byly rozděleny dle druhu dat, která stahovaly. Tyto notebooky a soubory, do kterých se data ukládají jsou popsány v podsekcích

níže. Data jsou vždy ukládána pro každou sezónu zvlášť a jsou ukládána ve formátu `JSON`.

4.2.1 download_playByPlay_data.ipynb

Tento notebook se věnuje stáhnutí všech tabulek `Play By Play` ze všech zápasů. Data jsou ukládána do složky `PlayByPlays`, kde se soubor reprezentující data ze sezóny 2014/15 jmenuje `playByPlays_by_gameID_2014-15.json`. Soubor představuje `JSON` reprezentaci dictionary, který má jako klíč vždy `Id` zápasu a jako hodnotu `DataFrame` reprezentující výsledek volání endpointu `playbyplay`, který je přiblížen v sekci 4.1.3.

4.2.2 download_player_game_info.ipynb

Notebook se věnuje stáhnutí dat týkajících se statistik jednotlivých hráčů na konci každé čtvrtiny jednotlivé hry ze sledovaného období a současně využívá získaných dat z použitých endpointů k uložení soupisek hráčů v každém utkání, a také seznam hráčů, kteří v jednotlivých zápasech byli v základní sestavě.

Data obsahující statistiky hráčů jsou ukládána do složky `PlayerGameLogs`, kde je soubor reprezentující data například ze sezóny 2014/15 pojmenován `player_game_logs_2014-15.json`. `JSON` zachycuje dictionary, který má jako klíč hráčovo `Id`. Jako hodnotu má vždy další vnořený dictionary, jehož klíče jsou čísla od 0 do 4. Toto číslo vždy odkazuje na `DataFrame`, který je výsledkem volání endpointu `playergamelogs`, který je přiblížen v sekci 4.1.3. Tento `DataFrame` reprezentuje hráčovy statistiky, které jsou relevantní pro konec čtvrtiny zápasu reprezentované číslem klíče. Nula reprezentuje zápasové statistiky včetně případného prodloužení.

Zároveň notebook stahuje data obsahující soupisky týmů pro každý zápas. Ta jsou ukládána do složky `GameRosters`, kde se soubor reprezentující data např. ze sezóny 2014/15 jmenuje `gameIdToTeamIdToRoster2014-15.json`. `JSON` zde zachycuje vícevrstvý dictionary, který má jako klíč `Id` zápasu a jako hodnotu má vnořený dictionary. Vnořený dictionary má jako klíč `Id` zúčastněného týmu v daném zápase a jako hodnotu má seznam `Id` hráčů, kteří byli na soupisce týmu v daném zápase a zápasu se zúčastnili. Na získání těchto dat a dat níže popsaných byl využit endpoint `boxscoretraditionalv2`, který je přiblížen v sekci 4.1.3.

V notebooku se stahují i seznamy hráčů, kteří v jednotlivých zápasech nastupovali v základní sestavě. Ta jsou ukládána do složky `Starters`, kde se soubor reprezentující sezónu 2014/15 jmenuje `list_of_starters2014-15.json`. `JSON` zde zachycuje dictionary, který má jako klíč `Id` zápasu a jako hodnotu seznam `Id` hráčů z obou týmů, kteří nastoupili v základní sestavě.

4.2.3 download_team_game_logs.ipynb

Notebook `download_team_game_logs` se věnuje stáhnutí dat týkajících se statistik jednotlivých týmů z každé hry, kterou ve vybraných sezonách odehrály. Data jsou ukládána do složky `TeamGameLogs`, kde se soubor reprezentující sezónu 2014/15 jmenuje `game_logs_by_team_id2014-15.json`. Soubor představuje `JSON` reprezentaci dictionary, který má jako klíč vždy `Id` daného týmu a jako hodnotu `DataFrame` reprezentující výsledek volání endpointu `teamgamelogs`, který je přiblížen v sekci 4.1.3.

Kapitola 5

Výsledný predikční model

V této kapitole je přiblížen postup při tvorbě finálních tabulek, které model využívá a jejich popis. Je zde také popsán závěrečný predikční model, pro který byly vylepšovány příznaky vstupních dat s cílem dosáhnout při predikci celkového počtu bodů na konci třetí čtvrtiny zápasu co nejpřesnějšího výsledku, a to na základě dvou tabulek, jejichž struktura je popsána v sekci níže. Tabulky byly za účelem správného natrénování modelu rozděleny do dvou množin – trénovací a testovací.

5.1 Struktura připravených tabulek

Predikce modelu pro odhad celkového počtu nastřílených bodů na konci třetí čtvrtiny je díky `Play By Play` tabulkám (získaných prostřednictvím knihovny `nba_api`) možné založit na dvou tabulkách. První tabulka obsahuje statistiky po odehrání pouze první čtvrtiny zápasu společně se statistikami z konce třetí čtvrtiny zápasu, které chceme predikovat. Druhá tabulka je stejná, ale je obohacena o statistiky z konce druhé čtvrtiny zápasu. Obě tabulky také obsahují kompletní statistiky zápasů z posledních 5 sezón.

Díky vytvoření těchto dvou různých tabulek se shodnými příznaky, odlišných pouze tím, z jaké čtvrtiny zápasu jsou, je možné porovnat predikce modelu, který predikuje na základě dat pouze z první čtvrtiny zápasu a predikce modelu, který své predikce zakládá na výsledcích i druhé třetiny zápasu, neboli v době, kdy je zápas již v pokročilejší fázi. Ve výsledcích, které jsou popsány níže je možné vidět, že data z konce druhé čtvrtiny zápasu se přímo podílela na kvalitě predikcí, které i výrazně zlepšila. Tento fakt jasně indikuje, že dostupnost a množství statistik ze zápasu, ve kterém události predikujeme, se přímo podílí na kvalitě predikcí.

Data jsou nejprve zpracována pro každou sezónu zvlášť a následně jsou ve formě `DataFramu` uložena do složky `TeamDataTableByYear`. Poté jsou obsahy

všech takto vzniklých souborů spojeny a uloženy do jednoho velkého csv[8] souboru. Výsledný soubor byl koncipován tak, aby každý řádek ve vytvořeném `DataFramu` reprezentoval statistiky z jednoho konkrétního zápasu, tedy obsahoval jak statistiky týmu domácího, tak hostujícího týmu. Při vytváření byly použity nejen obecné statistiky jednotlivých týmů, ale také příznaky vytvořené na základě zpracování dostupných dat. Dobrým příkladem je například průměrný počet bodů z posledních 10 zápasů domácího týmu.

Také se povedlo zakomponovat Win Share metriku, jejíž princip je popsán v sekci 3.2.1. Tato metrika se vypočítává pro každého hráče zvlášť, aby pasovala na formát dat, který se používá pro výsledný model, bylo třeba tuto metriku tomu uzpůsobit. Win Share se tedy vypočítal pro každého hráče, který v daném zápase byl na soupisce týmu a hry se zúčastnil. Poté se vypočítal průměrný Win Share pro daný tým z jednotlivých hodnot Win Sharu hráčů. Hodnota Win Sharu byla u těchto dat omezena na statistiky z maximálně 15 předchozích her pro konkrétního hráče v dané sezóně. Seznam všech příznaků s vysvětlivkami je možné vidět v následující sekci.

5.1.1 Popis příznaků ve sloupcích tabulek

Výsledná tabulka se statistikami z konce první čtvrtiny zápasu, která se následně dělí na trénovací a testovací data, obsahuje sloupce s hodnotami:

- **game_id**
- **home_team_id**
- **away_team_id**
- **season_end_year** = rok, kdy sezóna končí
- **nth_game** = označuje číslo hry domácího týmu v aktuální sezóně
- **match_result** = 1 pro výhru domácího týmu ; 0 pro remízu ; -1 pro výhru týmu hostů
- **total_scored_points**
- **Q1_result** = 1 pro výhru domácího týmu ; 0 pro remízu ; -1 pro výhru týmu hostů
- **Q3_result** = 1 pro výhru domácího týmu ; 0 pro remízu ; -1 pro výhru týmu hostů
- **Q1_total_points** = celkový počet bodů na konci 1. čtvrtiny
- **Q3_total_points** = celkový počet bodů na konci 3. čtvrtiny

[8] comma-separated values, tabulkový formát uložení dat, kde jsou hodnoty oddělené pomocí čárek

- **home_team_scored_points** = celkový počet bodů domácího týmu
- **away_team_scored_points** = celkový počet bodů týmu hostů
- **home_team_Q1_points** = celkový počet bodů domácího týmu na konci 1. čtvrtiny
- **away_team_Q1_points** = celkový počet bodů týmu hostů na konci 1. čtvrtiny
- **home_team_Q3_points** = celkový počet bodů domácího týmu na konci 3. čtvrtiny
- **away_team_Q3_points** = celkový počet bodů týmu hostů na konci 3. čtvrtiny

Příznaky níže zachycují průměr bodů z maximálně 10 předchozích zápasů v sezóně

- **home_team_win_share** = win share domácího týmu
- **away_team_win_share** = win share týmu hostů
- **home_team_avg_points_Q1** = průměr bodů z konce 1. čtvrtiny
- **home_team_avg_points_Q2** = průměr bodů z konce 2. čtvrtiny
- **home_team_avg_points_Q3** = průměr bodů z konce 3. čtvrtiny
- **away_team_avg_points_Q1** = průměr bodů z konce 1. čtvrtiny
- **away_team_avg_points_Q2** = průměr bodů z konce 2. čtvrtiny
- **away_team_avg_points_Q3** = průměr bodů z konce 3. čtvrtiny

Příznaky níže zachycují průměr bodů z maximálně 5 předchozích zápasů

- **home_team_avg_points_in_home_games_Q1** = průměr bodů v domácích zápasech domácího týmu z konce 1. čtvrtiny
- **home_team_avg_points_in_home_games_Q2** = průměr bodů v domácích zápasech domácího týmu z konce 2. čtvrtiny
- **home_team_avg_points_in_home_games_Q3** = průměr bodů v domácích zápasech domácího týmu z konce 3. čtvrtiny
- **away_team_avg_points_in_home_games_Q1** = průměr bodů ve venkovních zápasech týmu hostů z konce 1. čtvrtiny
- **away_team_avg_points_in_home_games_Q2** = průměr bodů ve venkovních zápasech týmu hostů z konce 2. čtvrtiny

- **away_team_avg_points_in_home_games_Q3** = průměr bodů ve venkovních zápasech týmu hostů z konce 3. čtvrtiny

Druhá tabulka obsahuje identické sloupce, ale je obohacena o sloupce obsahující statistiky z konce druhé čtvrtiny zápasu. Hodnoty, které jsou zde oproti první tabulce navíc jsou:

- **Q2_result** = 1 pro výhru domácího týmu ; 0 pro remízu ; -1 pro výhru týmu hostů
- **Q2_total_points** = celkový počet bodů na konci 1. čtvrtiny
- **home_team_Q2_points** = celkový počet bodů domácího týmu na konci 2. čtvrtiny
- **away_team_Q2_points** = celkový počet bodů týmu hostů na konci 2. čtvrtiny

Před použitím dat pro predikční model byly vždy z tabulek odebrány sloupce obsahující statistiky z konce třetí čtvrtiny a z konce zápasu. Odebrán byl také sloupec obsahující `Id` zápasu.

5.2 Trénovací a testovací data

Data bylo třeba jednoznačně rozdělit na trénovací a testovací z důvodu, aby bylo možné ve fázi vylepšování modelu porovnávat výsledky založené na stejných datech. V době vytváření bakalářské práce ve světě probíhala pandemie COVID-19 [15]. Jako následek této události se přestaly hrát zápasy NBA. Této nepředvídatelné a omezující situaci se přizpůsobilo i rozdělování dat. V datech byl zaveden sloupec `nth_game`, který reprezentuje číslo hry domácího týmu v aktuální sezóně, tedy jakési virtuální *kolo* zápasů. Následně se v datech identifikovala hodnota v tomto sloupci, která se vyskytovala u všech týmů (tzn. všechny týmy měly kolo, reprezentované daným číslem, odehrané) a zároveň byla nejvyšší možná v nejaktuálnější sezóně. Řádky, které mají v nejaktuálnější sezóně toto číslo vyšší, byly umístěny do testovací množiny a zbytek dat byl umístěn do trénovací množiny. Tímto vznikla data v poměru velikostí trénovací ku testovacím 6759 ku 175. Z těchto čísel je možné vidět, že celkový počet dat v tabulce byl poměrně malý. Běžně predikční modely pracují s mnohem větším objemem dat a díky tomu jsou i predikce přesnější. Data jsou ovšem vázána na počet odehraných zápasů ve vybraném časovém období, a to je důvod, proč celkový počet dat nemůže být vyšší. Počet dat by se dal zvýšit tím, že by se zvětšilo vybrané období, tedy přidaly by se statistiky ze starších zápasů. Ovšem na základě informací, které vyplývají ze závěrů prostudovaných prací[10], které se věnovaly podobné problematice, je zřejmé, že použití dat starších více než 5 sezón vede ke zhoršení kvality predikcí modelu.

Za účelem otestování chování výsledného modelu na jiných datech byla připravena ještě jedna sada trénovacích a testovacích dat. Pro toto rozdělení byla použita data ze sezóny 2018/19 a předchozích 3 sezón. V testovací množině dat bylo posledních 200 zápasů ze sezóny 2018/19 a zbytek dat byl umístěn do trénovací množiny. Výsledky zvoleného predikčního modelu na množině těchto dat jsou popsány v sekci 5.5.

5.3 Experimenty s příznaky

Za účelem získání co nejlepších výsledků predikčního modelu bylo třeba provést různé experimenty s hodnotami ve sloupcích tabulky, které model zpracovával a na nichž zakládal své predikce. Aby bylo zlepšování predikce modelu efektivní, bylo potřeba definovat míru chyby v predikci modelu a také provádět predikce vždy na stejných datech. Konzistentnímu rozdělení dat se věnuje sekce 5.2.

Pro výpočet velikosti chyby se používá odhad *MSE*. Vzorec pro výpočet této chyby je $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, kde n reprezentuje počet řádků v tabulce, Y_i skutečnou hodnotu vysvětlované proměnné, kterou se snažíme predikovat, a $\hat{Y}_i$ je hodnota, kterou pro daný řádek v tabulce predikoval náš model.

Experimentování s příznaky v této problematice často probíhá tak, že se přidávají nové sloupce hodnot do tabulky, kterou model zpracovává, a zkoumá se hodnota chyby predikcí. Přidávané hodnoty mohou být například nové statistiky získané z jiného, ještě nepoužitého zdroje, nebo hodnoty vzniklé kombinací některých, v tabulce již přítomných statistik, či jiné hodnoty, které jsou ve vztahu k obsahu tabulky relevantní. V určitých případech se některé sloupce, či řádky, které se vyznačující nějakou nevhodnou vlastností, odeberou a sleduje se, zda i tato změna nepomůže predikcím modelu. V této práci byly sloupce s příznaky spíše přidávány a jejich popis je v následujících odstavcích. Při experimentech byly použity i jiné změny, než které jsou popsány níže, ovšem jejich aplikace nenesla výrazné proměny v hodnotě vypočítávané chyby predikcí. Níže jsou tedy popsány pouze ty změny, které zlepšily kvalitu predikcí modelu nejvíce.

Kombinací statistik z připravených dat bylo možné přidat hodnotu Win Sharu do tabulky. Toto byla jedna z nejnáročnějších změn na implementaci, protože Win Share reprezentuje zásluhy hráče na výhře svého týmu a jelikož výsledná tabulka obsahuje v řádcích týmové statistiky k jednotlivým hrám, bylo třeba Win Share upravit, aby byla jeho hodnota relevantní. Při výpočtu Win Sharu se vycházelo ze soupisek týmů, které obsahují seznamy zúčastněných hráčů. Pro každý tým se na základě tohoto seznamu a s využitím připraveného souboru popsaného v sekci 4.2.2 obsahující statistiky jednotlivých hráčů, vypočítala hodnota Win Share pro každého hráče. Výsledný Win Share týmu byl průměr vypočítaných hodnot Win Share u jeho hráčů. Win Share se v tabulce počítá z posledních maximálně 10 zápasů v dané sezóně, ve

které se zápas odehrává, protože experimenty ukázaly, že vyšší počet těchto zápasů již predikci nezlepšoval.

Další změnou, která výrazně zlepšila predikci modelu, bylo přidání příznaků, které zachycovaly průměrný počet bodů domácího, respektive venkovního týmu v předchozích maximálně 10 zápasech dané sezóny v první, druhé a třetí čtvrtině zápasu. Zlepšení predikcí po přidání těchto hodnot ukázalo, že výsledky týmu jsou přímo ovlivněny výkony v zápasech, které danému utkání předcházely.

Poslední nejužitečnější změnou bylo přidání příznaků, které zachycovaly průměrný počet bodů domácího týmu v předchozích domácích zápasech na konci první, druhé a třetí čtvrtiny zápasu. Obdobně byl přidán i příznak pro průměrný počet bodů venkovního týmu z jeho předešlých venkovních zápasů. Tyto výpočty byly omezeny na průměr z maximálně 5 předchozích zápasů v relevantní sezóně. Důvodem přidání těchto příznaků byla snaha pozorovat, zda výkon v domácím, respektive cizím prostředí, má na tým vliv a ovlivňuje celkový počet bodů, které daný tým získá. Tato domněnka se vzhledem k zlepšení kvality predikcí potvrdila.

V této práci bylo experimentováno i s redukováním dat. Při redukování dat se dosáhlo zlepšení predikce pouze, když se z tabulky odebraly řádky reprezentující vždy první tři zápasy každého týmu v každé sezóně. Důvodem, proč se odebralo prvních pár her bylo, že přidané příznaky popsané v předchozích odstavcích používaly statistiky z minulých utkání. Tyto hodnoty byly ovšem v prvních hrách sezóny vždy značně nevypovídající, a proto byly odebrány. Větší množství odebraných her již predikci nadále nezlepšovalo. Data obsahují hry z posledních 5 sezón a pokus o snížení tohoto počtu na 4 nevedlo ke zlepšení predikcí modelu, naopak tyto predikce zhoršilo.

Po dokončení experimentů s příznaky byla nad výslednou tabulkou použita technika pro identifikaci nejlepších příznaků `SelectKBest` z knihovny `scikit-learn`. Pomocí této techniky bylo nalezeno 5 sloupců v tabulce, jejichž hodnoty nejvíce pozitivně ovlivňovaly kvalitu predikcí. Sloupce, které dle výsledků této metody byly nejdůležitější pro predikce jsou:

- Q1_total_points
- away_team_Q1_points
- home_team_Q1_points
- away_team_avg_points_Q3
- home_team_avg_points_Q3

5.4 Experimenty s predikčními modely

Pro predikci bylo nutné vybrat model, který pro vstupní data bude mít nejlepší predikce. Kromě výsledného modelu, který je popsán v následující sekci, byly vyzkoušeny ještě další modely, které jsou popsány v odstavcích níže.

Jedna z vhodných technik pro práci s regresními daty je technika `AdaBoost` neboli Adaptive Boosting. Jako její základní stavební kámen byl vybrán model rozhodovacího stromu `DecisionTreeRegressor`. Pro vybrání nejvhodnějších parametrů pro rozhodovací strom a následně i pro `AdaBoost` byla použita technika křížové validace. Výsledky predikcí tohoto modelu měly vyšší hodnotu chyby pouze o necelých 10 bodů, než vybraný model, a to když pracovaly s daty obsahující příznaky pouze z konce první čtvrtiny zápasu. Při predikci na základě tabulek obsahující data i z konce druhé čtvrtiny, byla hodnota chyby srovnatelná s výsledným modelem.

Technika, která je Adaptive Boostingu podobná, a také je jejím základem rozhodovací strom se jmenuje `Random Forest Reggression`. Pro tento model bylo potřeba nalézt pouze optimální hloubku dílčích stromů a byl připraven k zpracování připravených dat. Výsledky tohoto modelu byly srovnatelné s výsledky `AdaBoostu`. Hodnota chyby byla při predikcích na základě obou tabulek větší vždy pouze o 3 body, což je zanedbatelný rozdíl.

Při pokusech s modely byla připravená data využita k predikci i modelem logistické regrese. Tento model ovšem pracuje s klasifikačními druhy problémů, neboli problémy, jejichž predikovaná proměnná nabývá pouze několika málo hodnot. Na základě toho se vytvořily kopie dvou vstupních tabulek, které se lišily pouze v tom, že ve sloupci `match_result` měly pouze hodnoty symbolizující výhru a prohru. Tedy v případě, že zápas na konci čtvrté čtvrtiny zápasu skončil remízou a tím pádem by hodnota v tomto sloupci správně měla být nula, hodnota v tomto sloupci reprezentovala výsledek zápasu po prodloužení. Tento sloupec se také v modelu logistické regrese predikoval. Výsledky těchto predikcí byly následující:

- Predikce výsledku třetí čtvrtiny zápasu na základě dat z konce první čtvrtiny:
  - správně určená prohra v 43 případech,
  - špatně určená výhra v 39 případech,
  - špatně určená prohra v 21 případech,
  - správně určená výhra v 72 případech,
- Predikce výsledku třetí čtvrtiny zápasu na základě dat z konce druhé čtvrtiny:
  - správně určená prohra v 58 případech,
  - špatně určená výhra v 24 případech,

- špatně určená prohra v 20 případech,
- správně určená výhra v 73 případech.

Další model, s kterým bylo experimentováno byl model neuronové sítě, konkrétně `MLPRegressor` z knihovny `scikit-learn`. Predikcím tohoto modelu vycházela hodnota chyby mírně vyšší než u vybraného modelu, a to jak u predikce založené na datech z konce první čtvrtiny zápasu, tak u predikce založené na datech z konce druhé čtvrtiny. Ani za použití křížové validace na získání nejvhodnějších parametrů pro tento model se nepovedlo výsledky lineární regrese překonat. Důvod, proč neuronová síť nepřinesla lepší výsledky než lineární regrese, je nejspíš ten, že dat bylo na tento typ modelu poměrně malé množství.

Nejblíže se výsledným hodnotám vybraného modelu blížily výsledky modelu hřebenové regrese. Její výsledky se lišily v hodnotě chyby pouze o desetiny bodu. Pro získání správných hyperparametrů pro tento model byla použita křížová validace. Výsledky tohoto modelu byly srovnatelné i na úrovni hodnot chyby v predikcích jednotlivých zápasů.

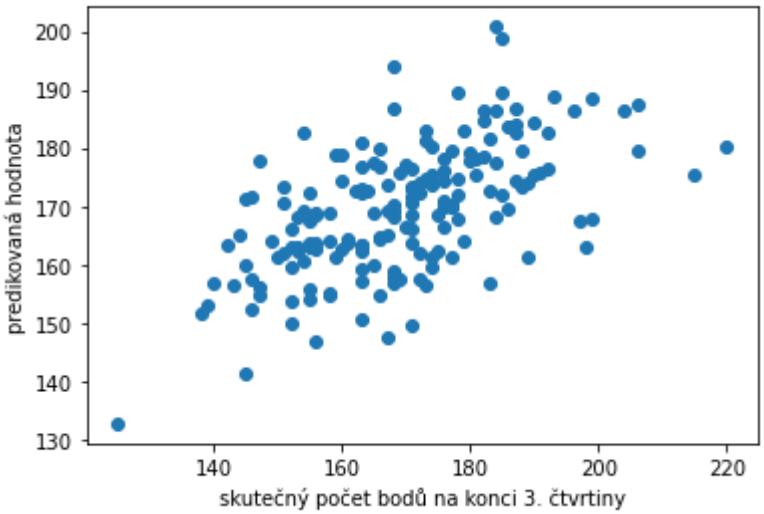
5.5 Zvolený predikční model a jeho výsledky

Jako model s nejmenší hodnotou chyby a nejlepšími predikcemi se ukázal být model lineární regrese. Využit byl model pro lineární regresi z knihovny `scikit-learn`.

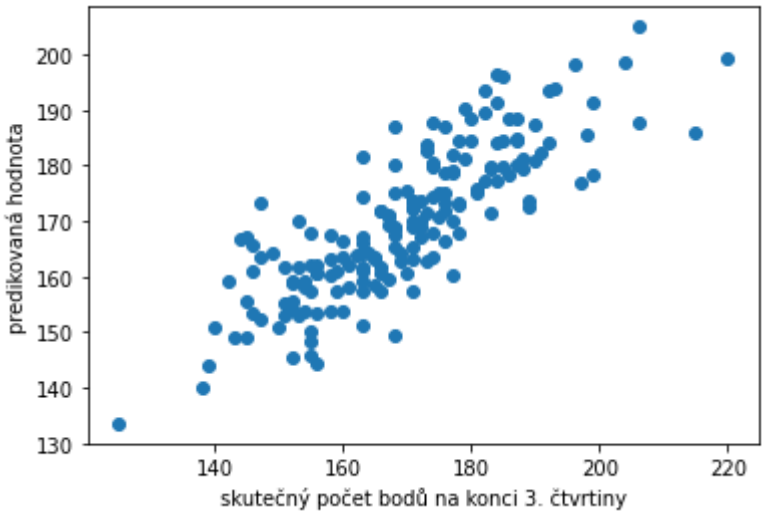
Pro predikci založenou na datech z konce první čtvrtiny zápasu vycházela nejmenší hodnota chyby v predikci při použití výchozí hodnoty nastavení parametrů. Při tomto nastavení byla hodnota chyby přibližně 166.28. Nejhůře předpovězené body z konce třetí čtvrtiny zápasu se lišily v průměru o 35 bodů od skutečné hodnoty. Skutečný počet vstřelených bodů v zápasech s nejlepšími predikcemi se lišil v průměru o 15 bodů. Porovnání predikcí s reálnými hodnotami je možné vidět na obrázku 5.1.

Predikce vycházející z tabulky se statistikami z konce druhé čtvrtiny zápasu vycházely mnohem lépe, což je pochopitelné vzhledem k tomu, že tabulka obsahuje navíc statistiky z pokročilejší fáze zápasu. Pro tyto data vycházely predikce lépe, když byla data v tabulce nejdříve znormalizována. Tato normalizace byla provedena pomocí normalizační techniky zprostředkované pomocí knihovny `scikit-learn` přímo v modelu lineární regrese. Hodnota chyby predikce byla přibližně 79. Nejhorší predikce bodů z konce třetí čtvrtiny zápasu se lišily v průměru o 24 bodů a ty nejlepší v průměru pouze o 11 bodů. Predikce modelu a jejich porovnání s reálnými hodnotami je možné vidět na obrázku 5.2.

Pro otestování výsledků finální podoby tabulek byly vyzkoušeny i upravené verze trénovací a testovací množiny. Tyto množiny nepracovaly s daty aktuální sezóny 2019/20 a jsou popsány na konci sekce 5.2. Výsledky na těchto



Obrázek 5.1: Porovnání predikovaných hodnot s hodnotami skutečnými na základě dat z konce první čtvrtiny zápasu



Obrázek 5.2: Porovnání predikovaných hodnot s hodnotami skutečnými na základě dat z konce druhé čtvrtiny zápasu

množinách měly hodnotu chyby predikcí dokonce nižší, než byla hodnota na datech, pro které byly příznaky zlepšovány, což indikuje dobře strukturované tabulky s vhodnými příznaky. Hodnota chyby byla, jak v případě dat obsahujících pouze statistiky z konce první čtvrtiny zápasu, tak u dat obsahujících statistiky i z konce druhé čtvrtiny zápasu, přibližně o 25 bodů nižší.

Na závěr byl proveden pokus o predikce celkového počtu bodů na konci čtvrté čtvrtiny zápasu. Predikce měly vyšší hodnotu chyby oproti predikcím počtu bodů na konci třetí čtvrtiny. To je způsobeno tím, že je pro model obtížnější predikovat události v zápasu čím více jsou časově vzdáleny od statistik, které jsou v tabulce dostupné. Na těchto výsledcích bylo také možné vidět, jak moc velkou roli hrají data z průběhu zápasu na kvalitu predikcí.

Závěr

V této bakalářské práci se zabýváme predikcí vybraných událostí v basketbalovém utkání. Jedním z hlavních cílů rešeršní části bylo vhodně zmapovat dostupné zdroje basketbalových statistik. Mezi dostupnými zdroji se povedlo nalézt a přiblížit ty nejužitečnější a hlavně popsat `nba_api` knihovnu, která se ukázala být nejvhodnějším zdrojem dat pro vytvořený model, protože velmi usnadnila přípravu dat do vhodného formátu.

Další cíl rešeršní části se týkal průzkumu prací věnujících se tématům, která se zabývají predikcí podobných událostí jako tato práce. Tyto práce byly představeny, bylo přiblíženo jejich zaměření, použité predikční modely, a také popsány závěry, ke kterým autoři daných prací dospěli. Při průzkumu zmíněných prací byla objevena metrika Win Share, která má přispívat k zlepšení výsledků predikčního modelu, a proto byla blíže popsána a vysvětlen její princip.

Hlavní cíl praktické části se týkal vytvoření predikčního modelu, vytvoření vhodných příznaků pro tabulky, které tento model zpracovává a následně pomocí experimentů s příznaky predikce zlepšovat. Model se vytvořit povedlo a experimenty s příznaky vedly ke zlepšení predikcí až do stavu, kdy se výsledky dají považovat za dobré. V této problematice je ovšem vždy prostor pro zlepšení, kterého by bylo možné dosáhnout dlouhodobými experimenty, což by již bylo nad rámec této práce. V rámci této části bylo také experimentováno s různými predikčními modely a z nich byl vybrán ten s nejlepšími výsledky.

Poslední cíl bohužel nebylo možné zrealizovat a to kvůli celosvětové pandemii COVID-19, která probíhala po dobu vytváření této bakalářské práce. Zejména kvůli tomu, že jeden z důsledků pandemie bylo přerušení celé sezóny NBA. Nebylo tedy možné ověřit kvalitu predikcí v průběhu hry a porovnat ji s vývojem tzv. live sázek u sázkových kanceláří.

Literatura

[1] *NBA player stats: Steven Adams* [online]. [cit. 2020-05-08]. Dostupné z: https://stats.nba.com/player/203500

[2] *NBA team stats: Chicago Bulls* [online]. [cit. 2020-05-08]. Dostupné z: https://stats.nba.com/team/1610612741

[3] *NBA: Terms of Use* [online]. [cit. 2020-03-27]. Dostupné z: https://nba.com/news/termsofuse

[4] *Sports Reference: Terms of Use* [online]. [cit. 2020-03-27]. Dostupné z: https://sports-reference.com/termsofuse.html

[5] *Sports Reference: Data Use* [online]. [cit. 2020-03-27]. Dostupné z: https://sports-reference.com/data_use.html

[6] ALAMEDA-BASORA, Enrique Marcos. *Dynamic Bayesian network to predict the total points scored in national basketball association games* [online]. 2019 [cit. 2020-03-27]. Dostupné z: https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=7962&context=etd. Iowa State University.

[7] YANG, Yuanhao (Stanley). *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics* [online]. 2015 [cit. 2020-03-27]. Dostupné z: https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=7962&context=etd. University of California at Berkeley.

[8] JAMES, Bill a Jim HENZLER. *Win Shares*. 1. 2002. ISBN STATS Publishing Inc. part 2. ISBN 9781931584036.

[9] *Sports Reference: Calculating Win Shares* [online]. [cit. 2020-03-26]. Dostupné z: https://sports-reference.com/cbb/about/ws.html

[10] PARK, Jong-Ho. *The Prediction of Outcomes in the National Basketball Association* [online]. 2014 [cit. 2020-03-27]. Dostupné z: `https://researchbank.rmit.edu.au/eserv/rmit:161885/Park.pdf`. RMIT University.

[11] *Guide to using @RISK* [online]. 2004 [cit. 2020-05-15]. Dostupné z: `http://risk.ef.jcu.cz/data/risk45.pdf`

[12] *RapidAPI: Subscription Plans & Pricing* [online]. [cit. 2020-03-27]. Dostupné z: `https://docs.rapidapi.com/docs/api-pricing`

[13] *Nba_api: An API Client package to access the APIs for NBA.com* [online]. [cit. 2020-03-26]. Dostupné z: `https://github.com/swar/nba_api`

[14] *Nba_api: license* [online]. [cit. 2020-05-16]. Dostupné z: `github.com/swar/nba_api/blob/master/LICENSE`

[15] *World Health Organization: Coronavirus disease (COVID-19) Pandemic* [online]. [cit. 2020-05-15]. Dostupné z: `who.int/emergencies/diseases/novel-coronavirus-2019`

Příloha A

Zkratky

JSON JavaScript Object Notation

URL Uniform Resource Locator

csv Comma-separated values

NBA National Basketball League

API Application program interface

MSE Mean squared error

Příloha **B**

Obsah přiloženého CD

Níže je přiblížena struktura přiloženého CD, které obsahuje všechny soubory týkající se této bakalářské práce.

- `README.md` ..... obecné informace k bakalářské práci
- `vysledkyPredikci.txt` ..... výsledky experimentů s příznaky a modely
- `BP_Schejbal_Ondrej_2020.pdf` ... text bakalářské práce v PDF formátu
- `BP_Schejbal_Task.png` ..... zadání bakalářské práce
- `Code` ..... složka s Jupyter Notebooky a vytvořenými datovými soubory
  - `GameRosters` ..... složka se soupiskami týmů pro každou hru
  - `PlayerDataTableByYear` ..... složka se statistikami hráčů
  - `PlayerGameLogs` ..... složka se zpracovanými statistikami hráčů
  - `Starters` ..... složka se seznamy začínajících hráčů v každé hře
  - `TeamDataTableByYear` ..... složka se statistikami týmů
  - `TeamGameLogs` ..... složka se seznamy zápasů každého týmu
  - `playByPlays` ..... složka s tabulkami Play By Play pro každou hru
- `Documentation-latex` ..... složka se zdrojovými LaTeX soubory
- `Research` ..... složka s poznámkami z analýzy zdrojů