FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Predikce vybraných událostí v basketbalovém utkání
Student:	Radim Křesťan
Vedoucí:	Ing. Karel Klouda, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

1) Proveďte rešerši zdrojů dat o zápasech a hráčích NBA. Zaměřte se také na data o událostech, která jsou aktualizována během daného utkání.
2) Proveďte rešerši známých metod a modelů používaných pro predikce výsledků a jiných událostí 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) Vaše predikce porovnejte také s kurzy sázkových kanceláří.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 31. prosince 2021 v Praze.

Bakalářská práce

PREDIKCE VYBRANÝCH UDÁLOSTÍ V BASKETBALOVÉM UTKÁNÍ

Radim Křesťan

Fakulta informačních technologií
Katedra aplikované matematiky
Vedoucí: Ing. Karel Klouda, Ph.D.
12. května 2022

Obsah

Poděkování vi

Prohlášení vii

Abstrakt viii

Seznam zkratek ix

1 Úvod 1
1.1 Cíle . . . 1

2 NBA a basketbal 3
2.1 Pravidla basketbalu . . . 3
2.2 Specifika NBA . . . 3
2.3 Rozdělení NBA týmů . . . 4
2.4 Rozdíl oproti ostatním týmovým sportům . . . 4
2.5 Pozice hráčů . . . 4

3 Práce na podobné téma 5
3.1 Artificial Intelligence in Sports Prediction . . . 5
3.2 NBA Game Result Prediction Using Feature Analysis and Machine Learning . . . 5
3.3 Sports Data Mining Technology Used in Basketball Outcome Prediction . . . 6
3.4 Predikce vybraných událostí v basketbalovém utkání . . . 6

4 Zdrojová data 7
4.1 Basketball-Reference . . . 7
4.1.1 Kvalita dat . . . 9
4.2 Rapid Api . . . 9
4.3 NBA stats . . . 9
4.3.1 Statistiky hráčů . . . 9
4.4 Data během utkání . . . 12
4.5 Legální dostupnost dat . . . 12

5 Proces získávání dat 13
5.1 Knihovna nba_api . . . 13
5.1.1 Statické moduly . . . 13
5.1.2 Endpoint moduly . . . 14

6 Experimenty a predikční model 15
6.1 Volba příznaků . . . 15
6.2 Výběr hyperparametrů . . . 17
6.3 Evaluace chyby . . . 17
6.4 Porovnání se kurzy sázkových kanceláří . . . 18

7 Závěr **19**
7.1 Možnosti pro navazující práce 19

Obsah přiloženého média **23**

Seznam obrázků

4.1 Giannis Antetokounmpo - BR 8
4.2 Luka Doncic statistiky 10
4.3 Boston Celtics vizualizace 11
4.4 Play By Play 11

5.1 players struktura 14
5.2 findPlayer funkce 14
5.3 PlayerGameLogs 14

6.1 hyperparametryAda 17
6.2 hyperparametryForest 17

Chtěl bych poděkovat mému vedoucímu, který se mnou pravidělně konzultoval výsledky mé práce. Dále bych chtěl poděkovat svým rodičům za poskytnutí zázemí a svým přátelům za neuvěřitelnou podporu při těžkých chvílích.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při 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 citovaného zákona.

V Praze dne 12. května 2022 .....

Abstrakt

Tato práce se zabývá problematikou živých předpovědí v oblasti basketbalu, konkrétně v NBA. Práce stručně popisuje doménu, které se predikce týkají a obsahuje analýzu experimentů, které již v minulosti byly provedeny. Dále detailně popisuje proces a možnosti získání dat, na kterých jsou následně experimentálně testovány jednotlivé metody. V praktické části této práce bylo použito několik modelů, mezi než patří například lineární regrese a náhodné lesy. Nejúspěšnější byla metoda náhodných lesů, která měla ve většině predikcí nejmenší odchylku. Predikovány byly statistiky hráčů na konci utkání s tím, že byla známá data z poloviny zápasu. Hlavním přínosem práce je, že umožňuje každému čtenáři se jasně a efektivně zorientovat v problematice basketbalových živých sázek a slouží jako teoretický základ pro jakékoliv další zkoumání.

Klíčová slova strojové učení, Jupyter Notebook, NBA data, predikce statistik, predikční modely, lineární regrese, analýza dat

Abstract

This thesis is focused on live predictions in basketball, specifically NBA. The thesis briefly describes the domain and includes an analysis of experiments that have been conducted in the past. It also describes the process and the possibilities of data mining. In the practical part of this thesis, several models have been experimentally tested, including but not limited to linear regression and random forests. The most successful method were random forests which had the lowest error in majority of predictions. Player stats at the end of the game were predicted with known mid-game data. The main contribution of this thesis is that it allows any reader to easily and effectively grasp the concept of basketball betting and serves as a theoretical foundation for any further investigation.

Keywords machine learning, Jupyter Notebook, NBA data, prediction of statistics, prediction models, linear regression, data analysis

Seznam zkratek

NBA	National Basketball Association
SF	Small Forward
PF	Power Forward
C	Center
SG	Shooting Guard
PG	Point Guard
BR	Basketball-Reference
API	Application Programming Interface
MAE	Mean Absolute Error
MSe	Mean Squared Error
RMSLE	Root Mean Squared Logaritmic Error

Kapitola 1
Úvod

Příchod digitálních technologií znamenal pro svět sázek a hazardu rapidní změnu. Už nějakou dobu každý z nás může pomocí několika kliknutí ze svého mobilnního telefonu vsázet na různorodé události a to nejen z oblasti sportu. Distanční sázky však nejsou jedinou změnou, která byla zapříčiněna informačními technologiemi. Pod povrchem aplikací můžeme najít poměrně sofistikované algoritmy, které jsou spoluzodpovědné za tvorbu kurzů. Dříve tuto práci vykonávali bookmakeři, jenže automatizované řešení přináší řadu výhod, jednou z nich je například live sázení, kterému by se zčásti měla věnovat i tato bakalářská práce. Procento lidského faktoru je v dnešní době těžké kvantifikovat, jelikož bychom museli vidět detailně do procesů jednotlivých firem. Co však můžeme tvrdit je, že každá firma se nachází v jiném stádiu digitalizace. Důkazem toho je například, že nejmenovaná česká společnost má jen velmi málo živých sázek na dílčí statistiky hráčů a druhá zahraniční společnost má live sázky o mnoho detailnější. Můžeme se tedy domnívat, že česká společnost ještě nemá perfektně vyřešenou danou problematiku, jelikož jinak by na sázkách mohla zvyšovat svůj zisk.

Tato práce je primárně určena pro společnosti, které ještě nemají perfektně dořešené predičkní modely a mohly by tak benefitovat z výsledků mé práce. Ovšem užitek z výsledků práce může mít kdokoliv, kdo se začíná zajímat o predikce sportovních utkání či jiných událostí. Problém pro individuální sázející je však ten, že sázkové kanceláře mají zhruba 15% rezervu, což jim umožnuje dosahovat zisku, i když nemají perfektní model.

Důvodem proč je práce psána na dané téma je, že proces automatizace sázek ještě není kompletní a je potřeba vytvořit a dodat dostatečně přesný predikční model, který bude předpovídat dílčí události v daném utkání.

Moje bakalářská práce navazuje, respektive rozšiřuje práci studenta Ondřeje Schejbala, který psal na velmi podobné téma. Věnoval se však pouze predikci výsledků a navíc kvůli nastalé kovidové situaci nemohl práci dokončit. Podstatný rozdíl mezi touto a jeho prací je, že zde budou predikovány dílčí statistiky hráčů, jak jsou například doskoky či asistence, kdežto v práci studenta Schejbala je kladen důraz primárně na výsledné body týmů. Chtěl bych také poupravit část o získávání dat, jelikož mi přišla nedostatečná.

1.1 Cíle

Cílem této práce je vytvořit statistický model, který získá data v polovině zápasu a na základě toho bude schopen predikovat konečný počet doskoků, asistencí a bloků hráče v zápase. Práce bude obsahovat analýzu dostupnosti dat, jejich legální získatelnost a použití. Rešerše bude analyzovat používáné modely pro predikci sportovních výsledků a statistik. Tato analýza bude zaměřena nejen na basketbal, ale i na ostatní sporty.

Na získaných datech bude experimentálně oveřena vhodnost jednotlivých metod na daný problém a bude vybrán ten s největší přesnotí. Práce bude obsahovat ukázku jak získat data o právě probíhajícím utkání a aplikovat je do modelu, který se z experimentů ukáže jako nejvhodnější. Finální část práce pak bude obsahovat porovnávní mého výsledku s kurzy sázkových kanceláří. V ideálním případě by výsledek práce měl být schopný překonat modely sázkových kanceláří. Avšak implementace modelu, který bude dostatečně přesný, aby mohl být nasazen nějakou společností v reálném životě, je považováno za úspěch.

Kapitola 2
NBA a basketbal

Národní basketbalová asociace je profesionální basketbalová liga sídlící v New Yorku. Součástí NBA je 30 týmů, přičemž 29 z nich je z USA a jeden z Kanady. NBA má přidruženou organizaci s názvem NBA G League, což je vlastně organizace, kde se začínající hráči, trenéři a jiní pracovníci připravují na budoucí působení v NBA. Pro tuto práci je významná z důvodu, že se v lize často experimentuje s novými pravidly, které mohou být, v případné upravené podobě, implementovány v NBA. Jedním z těchto pravidel je změna systému házení trestných hodů, které se však do NBA zatím nepřeneslo. Před několika lety se diskutovalo i o nové čáře, která by umožňovala dát koš za 4 body. Jakákoliv taková změna by mohla mít potenciální dopad na případné predikční modely.

2.1 Pravidla basketbalu

Basketbal je týmová hra, kde se utkávají dvě družstva po pěti hráčích. Nastoupit do každého zápasu za družstvo může vždy maximálně 12 hráčů. Cílem každého týmu je získat co nejvíce bodů, které získávají tím, že vstřelí míč do soupeřova koše. Koš má hodnotu tří bodů pokud je vystřelen zpoza tříbodového oblouku, jinak má hodnotu dvou bodů. Jedno utkání je rozděleno do čtyř čtvrtin po deseti minutách čistého času. Družstvo má omezený čas na přenesení míče ze své poloviny hřiště na soupeřovu a také omezený čas na celkový útok, pokud tak neučiní, získává míč soupeř.

Ve chvíli, kdy je útočicí hráč faulován při střele, tak střílí nějaký počet trestných hodů. Pokud střelu mine, tak střílí takový počet trestných kolik by byla hodnota koše, kdyby střelu trefil, což je buď dva nebo tři body. Další možná situace je, kdy koš trefí a body jsou mu uznány, potom střílí pouze jeden trestný hod navíc. Poslední varianta, kdy hráč jde střílet trestné hody je, když někdo z protihráčského týmu provede nesportovní či technickou chybu. Existuje speciální pravidlo, že když má tým více než 4 týmové fauly za čtvrtinu, tak při každém obraném faulu jde faulovaný hráč střílet trestné hody. Každý hráč se může dopustit až 4 osobních chyb na zápas, pokud se dopustí páté, už nemůže dál hrát v daném utkání a musí opustit hrací plochu.

2.2 Specifika NBA

- Utkání trvá 4x12 minut namísto 4x10
- Hráč je vyloučen až při šestém faulu
- Na hřiště se nachází 3 rozhodčí namísto dvou
- Rozdílný systém oddechovým časů

2.3 Rozdělení NBA týmů

Každý tým je na základě své geografické polohy součástí jedné ze dvou konferencí, východní nebo západní. Týmy soupeří i napříč konferencemi, ale na konci běžné sezóny postupuje 8 nejlepších týmů z každé konference. Toto pravidlo bylo relativně nedávno nahrazeno takzvaným *play-in* turnajem, kdy na postup do play-off mají šanci i týmy umístěné na deváté a desáte příčce. Výsledných 16 týmů se pak utká mezi sebou v utkáních na 4 vítězné zápasy a úplný vítěz si odnese trofej a titul šampióna NBA.

2.4 Rozdíl oproti ostatním týmovým sportům

Zaměřímeli se na rozdíly mezi basketbalem a ostatními sporty, určitě nalezneme mnoho detailů, ve kterých se sporty budou lišit, ať už to bude v počtu hráčů v týmu, v délce zápasu, případně v náčiní, které ke sportu potřebujeme. Je však jeden specifický rozdíl, který je potřeba zmínit a to jsou pravidla střídání. V basketbalu totiž není nijak omezený počet střídání. Z tohoto důvodu i členové týmu, kteří nejsou ve startovní pětici hráčů, mají relativně konzistentní počet odehraných minut v zápasech. Tuto situaci můžeme ilustrovat například na českém hráči Tomáši Satoranském, který měl v posledních deseti zápasech sezóny 2021/2022 průměrně odehraných 23,1 minuty s výběrovou směrodatnou odchylkou 4,1. Tento systém je v tomto ohledu srovnatelný s hokejem, ovšem vezmemli například fotbal, tak zde už střídání a nepředvídatelná rozhodnutí trenérů mohou mít obrovský vliv na výslednou predikci.

2.5 Pozice hráčů

Každý hráč má v týmu nějakou roli, podobně jako tomu je u fotbalu, kde jsou útočníci, záloha a obránci. V americkém basketbalu většinou rozlišujeme 5 pozic, které jsou do jisté míry odstupňovány podle fyzických dispozic a můžeme jim tak přiřadit čísla od jedné do pěti. Tato čísla mohou být využívána při identifikaci konkrétní pozice hráče. V Čechii se můžeme setkat s pojmenováním rozehrávač, křídla a pivot. Týmová strategie není vždy stejná, takže nemůžeme přesně přiřadit naše lokální pojmenování k tomu mezinárodnímu.

1. PG - point guard
2. SG - shooting guard
3. SF - small forward
4. PF - power forward
5. C - center

Kapitola 3

Práce na podobné téma

Následující kapitola obsahuje souhrnný popis prací, které byly napsány na podobné téma, nebo se nějakým způsobem věnují tématice sportovních predikcí. Vzhledem k povaze této práce je však mnoho zdrojů ve formě blogů či krátkých tutoriálů, které popisují, jak si jednoduše vytvořit sportovní predikci. Tyto blogy však né vždy obsahují formální, respektive matematický popis, tudíž na ně kapitola neklade důraz.

3.1 Artificial Intelligence in Sports Prediction

Tato práce[1] byla napsána již v roce 2008, kde byl použit koncept neuronových sítí, konkrétně vícevrstvého perceptronu k tomu, aby predikoval výsledky zápasů 4 různých lig. Jednalo se zde o Australian Nationa Rugby League, Australian Fotball League, Super Rugby a Premiere League. Práce pracovala s několika vstupními parametry, kde se řešilo, jak tým hrál v posledních zápasech, ale například i lokace, kde se dané utkání utkání. V neposlední řadě se také počítalo s dostupnustí jednotlivých hráčů, jelikož absence kllíčového hráče týmu může mít signifikatně negativní dopad na daný tým. Průměrná úspěšnost experimentů se pohybovala okolo 63%, kromě EPL, kde to bylo skoro až o 10% méně, což podle autorů bylo z velké části způsobeno velkým zastoupením remíz v zápasech.

3.2 NBA Game Result Prediction Using Feature Analysis and Machine Learning

Na Novozélandské univerzitě v Aucklandu byla v roce 2019 napsána studie[2] zabývající se predikcí výsledků NBA zápasů. Jejich použité modely byly neuronové sítě a naivní Bayesův klasifikátor. V dané práci bylo vytvořeno 5 různých datasetů, na které se následně aplikovaly zvolené modely. Z tohoto důvodu se přesnost predikcí liší, ale na některých datasetech dosahovala až 83%, což je samo o sobě velmi působivý výsledek. Zajímavý závěr, který z práce můžeme vydedukovat, však pochází ze samotného výběru datasetů, kde bylo zjištěno, že paradoxně vynecháním určitých příznaků můžeme dojít k přesnějším predikcím. Přesnost se ale v tomto případě lišila pouze o 2-4%, což stále může znamenat jenom náhodnou chybu. Dalším dílčím výsledkem této práce je, že jedním z nejdůležitějších příznaků, které mají vliv na výsledek zápasu, je počet defenzivních doskoků hráče.

3.3 Sports Data Mining Technology Used in Basketball Outcome Prediction

Tato dizertační práce[3] se do detailu věnuje popisu, jak dolovat data z webu a vzhledem k tomu, že je z roku 2012, tak i systém získávání je už v dnešní době mnohem jednodušší. Co se týká predikcí, tak je zde vybrán průměr konkrétní statistiky z posledních deseti her a to slouží jako příznaky jednotlivých modelů. Práce experimentuje se čtyřmi různými modely, kterými jsou logistická regrese, neuronové sítě a naivní bayesův klasifikátor. Metody dosahují přesnosti mezi 65-67 procenty, přičemž nejlépe vycházela logistická regrese.

3.4 Predikce vybraných událostí v basketbalovém utkání

Jednou již zmíněná bakalářská práce[4] studenta fakulty informačních technologií ČVUT, kde student získává data z konce první čtvrtiny a snaží se na jejich základě odhadnout výsledek na konci třetí čtvrtiny. Dále také bere data z poloviny zápasu a na jejich základě se snaží predikovat výsledek třetí čtvrtiny. Přesnost modelů byla v prvním případě 65,7% a v druhém 74,8%. Práci je velmi relevatní z důvodu, že se jako jedna z mála zaobírá problematikou živých sázek, což jiné mezinárodní studie příliš neřeší. Příznaky jsou opět brány jako průměry daných statistik z posledních her týmu, v tomto případě z posledních 5 nebo 10 zápasů. Další důležitou metrikou, se kterou zde student Schejbal pracuje jen *win-share*[5], což je kvantifikovaný podíl hráče na výhře jeho týmu. Jinými slovy jakou část z jedné výhry hráč přinesl. Tato metrika má několik podob a může se lišit v několika ohledech. V této práci byla použita tak, že může být záporná a tým má za každou výhru jeden *win-share*, což znamená, že součet *win-sharů* všech hráčů v jednom zápase musí být jedna.

Kapitola 4
Zdrojová data

Jednou z nejdůležitějších částí každého modelu jsou data a jejich kvalita. Z tohoto důvodu je nutné provést detailní analýzu dostupných zdrojů a možnosti jejich využití. Srovnáme-li jednotlivé sporty a jejich zastřešující organizace, zjistíme, že zveřejňovaná data se velmi liší. Například u fotbalové organizace Premiere League najdeme jenom souhrné statisky hráčů a nemůžeme již zjišťovat, v kterém zápase je získal. NBA je v tomto ohledu poměrně napřed, uchovává a poskytuje veškerá data od roku 1996. Tato data obsahují velmi detailní analýzy zápasů a jsou dostupná veřejně a bez žádných poplatků. NBA také nyní experimentuje s tím[1], že by měla vizuální model celého zápasu, který by oproti kamerovému záznamu měl výhodu, že bychom hráče mohli pozorovat z různých úhlů a celou situaci detailně analyzovat. Na svém twitterovém účtu NBA zveřejnila svůj prototyp, který zatím připomíná starší počítačovou hru, ale je zde poměrně zřetelné, co se v danou situaci děje. Je nutné podotknout, že NBA má nějaké základní statistiky týmů již od jejího založení v roce 1946, tato data jsou však velmi základní a k predikcím příliš dobře sloužit nemohou.

Následující část kapitoly obsahuje srovnání jednotlivých zdrojů. Porovnávám zde dostupnost dat, zda-li je možnost získávat data pomocí nějakého API, případně pokud se musí data z webu dolovat/scrapovat. Další kritérium je zda jsou data kompletní, toto kritérium bude pravděpodbně nejlépe splňovat oficiální zdroj dat, avšak chceme-li porovnávat jiné dva zdroje mezi sebou, tak se může jednat o klíčovou vlastnost. V neposlední řade je potřeba zjistit, jaké jsou podmínky užití jednotlivých zdrojů a jestli je vůbec legální daná data získávat.

4.1 Basketball-Reference

Zdroj Baskeball-reference je součástí většího celku Sports-Reference, který nabízí statistiky různých sportů, mezi něž patří například baseball, lední hokej, americký fotbal nebo i klasický fotbal. Co se týče basketbalu, tak stránka sports-reference[2] nabízí statistiky univerzitního basketbalu, ale také, což je pro tuto práci určující, rozsáhlá nba data. Zaměřímeli se teď na stránku basketball-reference zjistíme, že nabízí široké spektrum různých statistik. První položkou jsou statistiky hráčů, které když rozklikneme získáme abecední seznam hráčů. Po rychlé analýze však zjišťujeme, že od každého písmena jsou zde vypsáni jenom nejvýznamější hráči, jejichž příjmení začíná na dané písmeno. Primárně jsou to hráči, kteří jsou stále aktivní nebo byli zapsáni do *Hall of Fame* NBA. U každého hráče můžeme najít souhrnné statistiky za jednotlivé sezóny, kdy hráči hráli. Zajímavostí je, že můžeme kliknout na jednotlivé řádky a získat tak souhrn za námi vybrané sezóny.

[1] https://www.techspot.com/news/93823-nba-aired-entire-game-using-volumetric-video-tech.html
[2] https://www.sports-reference.com/

Regular Season | Playoffs

Season	Age	Tm	Lg	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
2013-14	19	MIL	NBA	SF	77	23	24.6	2.2	5.4	.414	0.5	1.5	.347	1.7	3.9	.440	.463	1.8	2.6	.683	1.0	3.4	4.4	1.9	0.8	0.8	1.6	2.2	6.8
2014-15	20	MIL	NBA	SG	81	71	31.4	4.7	9.6	.491	0.1	0.5	.159	4.6	9.1	.511	.496	3.2	4.3	.741	1.2	5.5	6.7	2.6	0.9	1.0	2.1	3.1	12.7
2015-16	21	MIL	NBA	PG	80	79	35.3	6.4	12.7	.506	0.4	1.4	.257	6.1	11.3	.537	.520	3.7	5.1	.724	1.4	6.2	7.7	4.3	1.2	1.4	2.6	3.2	16.9
2016-17 ★	22	MIL	NBA	SF	80	80	35.6	8.2	15.7	.521	0.6	2.3	.272	7.6	13.5	.563	.541	5.9	7.7	.770	1.8	7.0	8.8	5.4	1.6	1.9	2.9	3.1	22.9
2017-18 ★	23	MIL	NBA	PF	75	75	36.7	9.9	18.7	.529	0.6	1.9	.307	9.3	16.8	.554	.545	6.5	8.5	.760	2.1	8.0	10.0	4.8	1.5	1.4	3.0	3.1	26.9
2018-19 ★	24	MIL	NBA	PF	72	72	32.8	10.0	17.3	.578	0.7	2.8	.256	9.3	14.5	.641	.599	6.9	9.5	.729	2.2	10.3	12.5	5.9	1.3	1.5	3.7	3.2	27.7
2019-20 ★	25	MIL	NBA	PF	63	63	30.4	10.9	19.7	.553	1.4	4.7	.304	9.5	15.0	.631	.589	6.3	10.0	.633	2.2	11.4	13.6	5.6	1.0	1.0	3.7	3.1	29.5
2020-21 ★	26	MIL	NBA	PF	61	61	33.0	10.3	18.0	.569	1.1	3.6	.303	9.2	14.4	.636	.600	6.5	9.5	.685	1.6	9.4	11.0	5.9	1.2	1.2	3.4	2.8	28.1
2021-22 ★	27	MIL	NBA	PF	67	67	32.9	10.3	18.6	.553	1.1	3.6	.293	9.2	15.0	.616	.582	8.3	11.4	.722	2.0	9.6	11.6	5.8	1.1	1.4	3.3	3.2	29.9
Career			**NBA**		**656**	**591**	**32.6**	**7.9**	**14.8**	**.535**	**0.7**	**2.4**	**.288**	**7.2**	**12.4**	**.582**	**.558**	**5.3**	**7.4**	**.718**	**1.7**	**7.7**	**9.4**	**4.6**	**1.2**	**1.3**	**2.9**	**3.0**	**21.8**

Obrázek 4.1 Souhrnné statistiky hráče Giannis Antetokounmpo za všechny jeho odehrané sezóny Age - věk, Tm - tým za který v dané sezóně hrál, Lg - liga ve které hrál, Pos - pozice na které hrál, G - počet her v sezóně, které odehrál, GS - počet her ve kterých nastupoval v základní pětici, MP - průměrný počet odehraných minut za zápas, FG - průměrný počet úspěšných košů, FGA - průměrný počet vystřelených košů, FG% - průměrná úspěšnost střelby, 3P - průměrný počet úspěšných košů za tři body, 3PA - průměrný počet pokusů o koš za tři body, 3P% průměrná úspěšnost střelby za tři body, 2P - průměrný počet košů za dva body, 2PA průměrný počet pokusů o koš za tři body, 2P% průměrná úspěšnost střelby za dva body, eFG% - průměrná efektivní úspěšnost střelby, FT - průměrný počet trefených trestných hodů, FTA - průměrný počet vystřelených trestných hodů, FT% průměrná úspěšnost trestných hodů, ORB - průměrný počet útočných doskoků, DREB - průměrný počet obranných doskoků, TRB - průměrný počet všech doskoků, AST - průměrný počet asistencí, STL - průměrný počet ukradnutí míče soupeři, BLK - průměrný počet bloků, TOV - průměrný počet ztrát, PF - průměrný počet osobních chyb, PTS - průměrný počet bodů

4.1.1 Kvalita dat

Po prozkoumání BR zjistíme, že můžeme stahovat souhrnná data, jako je vidět na obrázku 4.1. Po hlubším zkoumání je možné nalézt přidruženou stránku stathead[3], kde můžeme získat pokročilá data za každou hru zvlášť. Tato stránka je ovšem zpoplatněna měsíčním předplatným, které stojí 8 euro. Avšak BR nabízí logy jednotlivých her, takže základní statistiky můžeme získat za všechny zápasy. Problémem však je, že data sice můžeme efektivně stáhnout a snadno exportovat do strojově čitelného formátu, avšak chybí zde možnost získat data nějakým api. Absence api, které bychom mohli využít v nějakém skriptu je možná obejít pomocí dolování dat, avšak v souboru robot.txt je nastaveno prodlení 3 vteřiny. Všechny doposud zmíněné problémy mají svá řešení, ovšem absence živých dat pro nás dělá stránku nepoužitelnou.

4.2 Rapid Api

Dalším možným zdrojem dat je rapid-api, kde je api knihovna API-NBA, která umožňuje získávat nba statistiky. Tato služba má však jednu podstatnou nevýhodu. Rapid-api je freemium služba, což znamená, že můžeme poslat maximálně 100 požadavků denně, jinak musíme mít pořízené jejich měsíční předplatné. Ovšem po provedené analýze jednotlivé endpointy umožňují relativně detailní sběr dat. Našeho případu by se nejpravděpodobněji týkal endpoint Statistics, kde můžeme získat statistiky hráče pro konkrétní hru.

4.3 NBA stats

Po provedené analýze možných zdrojů se oficiální stránky NBA stats zdají být jako nejobsáhlejší a nejsnáze přístupné. Dostupná data jsou kompletní a můžeme zde získávat statistiky ze všech her daného hráče. To jsme sice mohli i v případě BR, avšak zde jsou dostupná i data z poloviny zápasu, dokonce i z konců jednotlivých čtvrtin. Další obrovskou výhodou je, že je zde možnost sledovat aktuální stav hry v reálném čase a dokonce získávat informace o právě probíhající hře pomocí api. Statistiky jsou zde rozděleny podobně jako u BR na dva větší celky, kdy jeden se týká hráčů a druhý týmů. Vzhledem k povaze této práce jsou prioritní primárně statistiky týkající se hráčů.

4.3.1 Statistiky hráčů

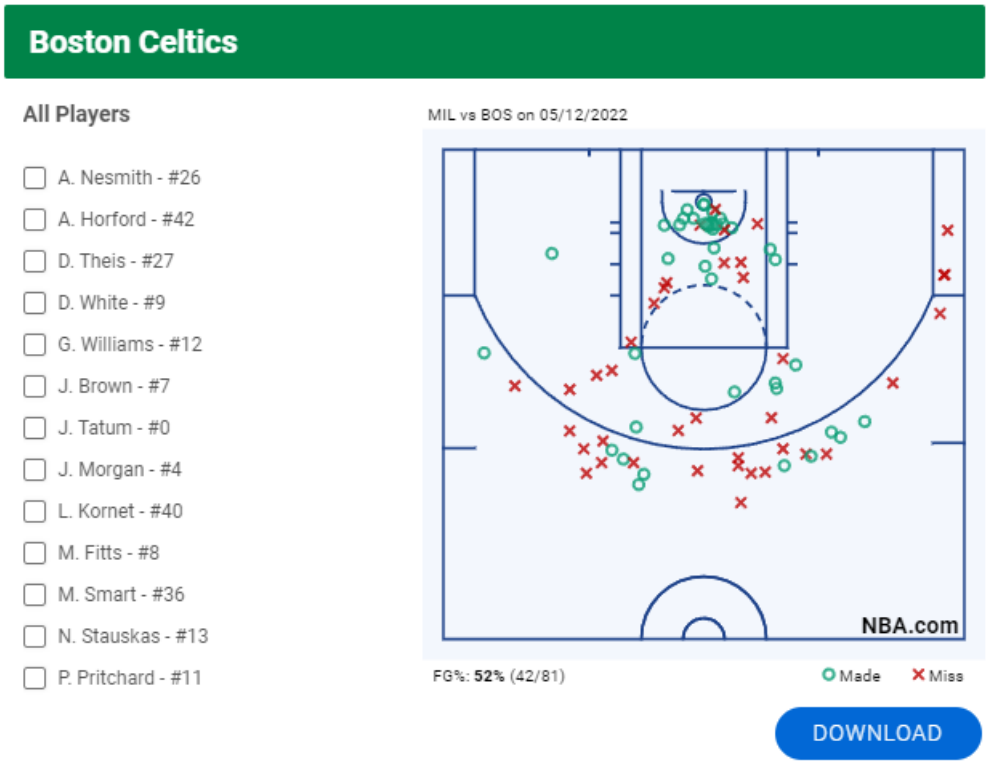
Podívámeli se na záložku players[4], ukáže se nám tabulka hráčů s pěti nejlepšími výkony dnešního dne a následně vidíme nejúspěšnější hráče v různých statistikách. Ve chvíli, kdy budeme chtít data o nějakém hráči, máme na výběr několik možností. Na profilu hráče[5] uvidíme opět souhrnné statistiky jako například u BR, které jsou zde rozděleny na pět různcých tabulek. První dataset je velmi podobný jako na obrázku 4.1, s tím rozdílem, že neobsahuje hodnoty počtu střel za dva body a úspěšnost střelby za dva body. Tyto statistiky jsou ovšem také dostupné, jenom v jiné z tabulek. Jednotlivá data můžeme zobrazit v několika režimech, prvním z nich je *per game*, což je vlastně jenom průměr za všechny hry, ovšem jsou zde možnosti i průměru za jednu minutu, případně statistiky za 100 herních akcí. Dále zde můžeme volit, jestli chceme data z období běžné sezóny nebo z playoffs.

Každý hráč má však své pokročilé statistiky, kde můžeme zjistit jeho výsledky během všech zápasů, do kterých kdy nastoupil. Volíme zde sezónu, zda-li se jedná o playoff zápas a případně i herní segment, ze kterého chceme data získat. Nevýhodou je, že zde nemůžeme vzít kompletní set všech her.

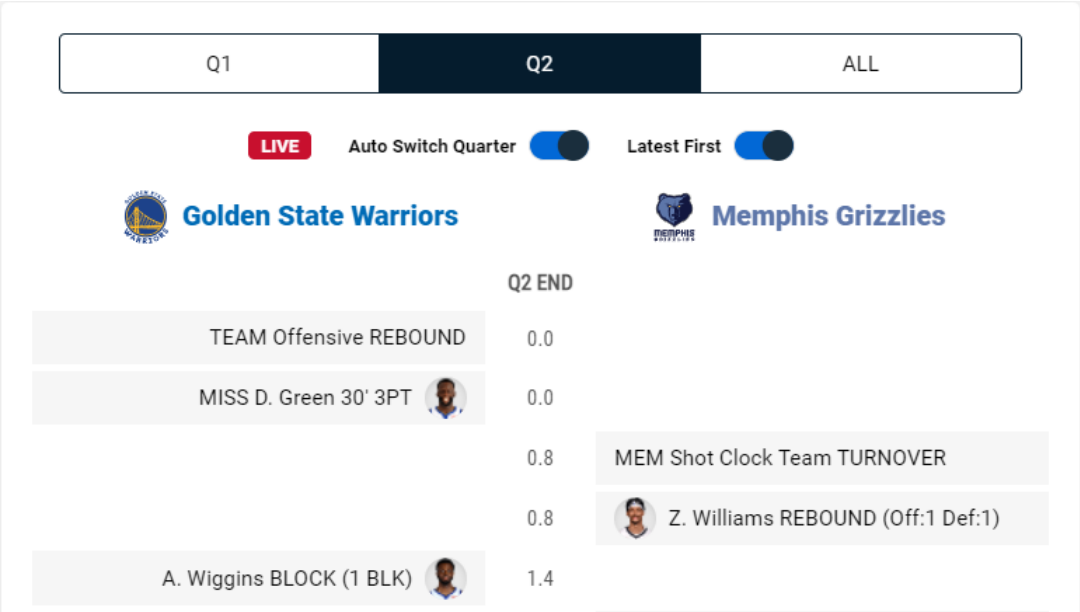
[3] https://stathead.com/basketball/
[4] https://www.nba.com/stats/players/
[5] https://www.nba.com/stats/player/1629029

MATCH UP	W/L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-
Apr 10, 2022 - DAL vs. SAS	W	29	26	8	18	44.4	2	5	40.0	8	8	100	1	7	8	9	2	1	4	1	19
Apr 08, 2022 - DAL vs. POR	W	30	39	12	21	57.1	7	14	50.0	8	11	72.7	0	11	11	7	0	1	5	0	41
Apr 06, 2022 - DAL @ DET	W	34	26	8	21	38.1	2	11	18.2	8	13	61.5	1	7	8	14	1	0	6	4	17
Apr 03, 2022 - DAL @ MIL	W	39	32	9	22	40.9	3	9	33.3	11	14	78.6	0	8	8	15	3	0	6	2	8
Apr 01, 2022 - DAL @ WAS	L	36	36	11	22	50.0	4	9	44.4	10	10	100	0	7	7	6	0	1	5	3	-22
Mar 30, 2022 - DAL @ CLE	W	38	35	14	26	53.8	3	6	50.0	4	4	100	0	9	9	13	1	2	3	1	14
Mar 29, 2022 - DAL vs. LAL	W	30	34	12	23	52.2	4	12	33.3	6	7	85.7	3	9	12	12	1	0	1	1	21

Obrázek 4.2 Statistiky jednotlivých zápasů hráče Luka Doncic za tým Dallas Mavericks. Data jsou z konce zápasu



Obrázek 4.3 Vizualizace střel na koš týmu Boston Celtics. Zelená kolečka značí trefené střely, červené křížky ty netrefené



Obrázek 4.4 Ukázka po sobě jdoucích akcí, které se v zápasu staly(play-by-play)

4.4 Data během utkání

Součástí této práce je analýza dat, která jsou aktualizována během samotného utkání. I tyto informace jsou na oficiálních stránkách nba k dispozici a to ve třech různých podobách. První, klasická, udává základní statistiky hráčů, podobné, jaké můžeme vidět na obrázku 4.2. Druhou formou infromací, které můžeme získat o právě probíhajícím zápase jsou *shot charts*, kde vidíme odkud jednotliví hráči střílí, pro ilustraci slouží obrázek 4.3. V neposlední řadě máme možnost získávat informace o jednotlivých akcích, které se během utkání dějí. Tento zdroj obsahuje všechny události, které se během daného zápasu odehrály a jsou následně vypsané v chronologickém pořadí.

4.5 Legální dostupnost dat

Na svých stránkách NBA stat jasně definují, pro jaké účely mohou být jejich data používána. Tedy tím, že data používám se zavazuji k tomu, že jakékoliv využití těchto statistik bude doprovázet odkaz na oficiální stránky. Všechna stažená data musí být použita výhradně pro osobní nekomerční účely. Statistiky nesmí být použity pro žádná sponzorství ani jiné komerční účely a nesmí být použita ve spojení s jakýmkoliv hazardem.

Poslední bod může být v našem případě trochu problematický, jelikož ve výsledku porovnáváme úspěšnost modelu se sázkovými kancelářemi. Ale jelikož se nijak na sázení nepodílíme a kurzy využíváme pouze k porovnání kvality modelu, nemělo by se jednat o nelegální aktivitu.

Kapitola 5
Proces získávání dat

Během analýzy dostupných zdrojů bylo zjištěno, že oficiální stránka obsahuje všechna potřebná data, které jsou potřeba na vytvoření predičkního modelu. Jak již bylo zmíněno k oficiálním stránkám existuje knihovna nba-api, která usnadňuje použití endpointů oficiálních statistik NBA. Výhodou knihovny je, že obsahuje rozsáhlou dokumentaci k jednotlivým endpointům, avšak přestože usnadňuje zisk dat nepřidává nějakou úroveň abstrakce navíc, která by zjednodušovala stahování dat. Z tohoto důvodu je k dispozici pouze přesně to, co bylo popsáno v předchozí kapitole a jakoukoliv logickou nadstavbu je potřeba provést ručně. Nutné je také podotknout, že poslední commit v knihovně je starý již víc než půl roku, což pravděpodobně znamená, že se dále nijak nerozvíjí.

Během práce s knihovnou byl nalezen problém, že když je posíláno na stránku příliš mnoho requestů v rychlém sledu tak uživatele stránka dočasně zablokuje. Toto chování se těžko předpovídá, jelikož nba api nemá v robot.txt nastavený žádný delay. Pro účely jednodušší manipulace se zdrojem dat je doporučeno přidat časovou pauzu mezi požadavky, které by byly posílany v rychlém sledu. Tato funkce však implementována nebyla, jelikož problémy s blokováním se primárně vyskytovaly při implementaci stahovacího skriptu.

5.1 Knihovna nba_api

Knihovna[1] napsána v jazyce python, obsahující api, které je napojeno na oficiální stránky nba. Je volně dostpuná na platformě GitHub a licencována pod MIT licencí. Knihovna obsahuje rozsáhlou škálu různých modulů, ale všechny se dají rozdělit do jedné ze tří kategorií. Může se jednat o statická data, která obsahují základní informace o týmech nebo hráčích. Jak je psáno na oficiální dokumentaci[2] knihovny, význam tohoto modulu je aby uživatel mohl pomocí regulárního výrazu zjistit nejzákladnější informace ať už o hráčích či týmech a nemusel k tomu odesílat požadavek. Druhým typem dat je třída *Endpoints*, kde se dá získat většina dostupných dat. Poslední možností, kterou můžeme získávat data je modul live, ze kterého můžeme čerpat statistiky, které jsou vyobrazeny na obrázcích 4.4 a 4.3

5.1.1 Statické moduly

Mezi statické moduly patří již zmiňované *players.py* a *teams.py*, v práci byl sice primárně použit ten první, ale zde je popis obou modulů.

[1] https://github.com/swar/nba_api
[2] https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/static/players.md

Co se týče modulu players.py, ten má metodu, která nám umožňuje vrátit seznam *dictionaries*, které obsahují strukturu, kterou můžeme vidět na obrázku 5.1. V práci je tento modul použit při získávání id konkrétního hráče, které je potom využito na získání dat o hrách. Použití ilustruje obrázek 5.2.

```
player = {
    'id': player_id,
    'full_name': full_name,
    'first_name': first_name,
    'last_name': last_name,
    'is_active': True or False
}
```

Obrázek 5.1 Ukázkový formát dat, který získáme při volání dat z modulu players.py

```
value = "Jrue Holiday"
playerId = 0
all_Players = players.get_players()
for item in all_Players:
    if value in item.values():
        print(item['id'])
```

Obrázek 5.2 Ilustrační použití statického modulu pro získání id hráče, jehož jméno je specifikováno v proměnné value

5.1.2 Endpoint moduly

V modulu *Endpoint* je v této práci primárně využívána funkce *playergamelogs.PlayerGameLogs*, díky které můžeme získat historická data z poloviny zápasů jednotlivých hráčů. Tato třída má již výše zmíněnou nevýhodu a to, že musíme data získávat po sezónách, jelikož třída neumožňuje hromadné stažení všech dat. Navíc ještě odděluje data z běžné sezóny a playoff, což může být výhoda, ale ve chvíli, kdy potřebujeme všechna data musíme vytvořit funkci, která bude spojovat dílčí data do jednoho velkého datasetu. To je v práci řešeno tak, že se nejdříve vypočítá rok, kdy končí aktuální sezóna a následně postupně voláme požadavky na playoff a běžnou sezónu. Ve chvíli kdy v tomto pořadí získáme data snížíme rok o 1 a opakujeme do té doby než narazíme na prázdnou sezónu, což znamená že hráč v tu dobu ještě nehrál, případně nebyl celou sezónu aktivní. Na obrázku 5.3 je vidět konrkétní použití pro zisk dat z playoff.

```
player_games = playergamelogs.PlayerGameLogs(player_id_nullable=player_id,
                                game_segment_nullable=game_segment,
                                measure_type_player_game_logs_nullable="Base",
                                per_mode_simple_nullable="Totals",
                                season_type_nullable="Playoffs",
                                season_nullable=season)
```

Obrázek 5.3 Použití endpointu PlayerGameLogs, který umožňuje vzít jako parametry id hráče, herní segment a typ sezóny a vráti všechny údaje, které splňují daná kritéria.

Kapitola 6

Experimenty a predikční model

Tato kapitola shrnuje provedené postupy, které vedly k vytvoření výsledného modelu. Jedná se o tvorbu nových příznaků ze získaných dat a následné ladění a volba modelů. Součástí této kapitoly jsou také různé druhy metrik, které jsou používány při měření chyby regresních problémů. Jsou zde porovnávány tři různé algoritmy, přičemž na každém z nich jsou testovány predikce čtyř různých statistik.

Po procesu získání dat bychom v souboru měli mít uložené všechny hry daného hráče. Soubory dodržují jmennou konvenci *Pidhrace* a můžeme tak uchovat data několika hráčů bez nutnosti je znovu stahovat, což ušetří podstatné množství času a requestů. Název souboru je jednoznačně určen, jelikož id hráče je unikátní identifikátor. Data v souboru však vždy potřebují úpravu podle toho na jakou predikci se budou využívat.

Jak již bylo zmíněno v této práci se počítá s tím, že se mohou predikovat až čtyři statistiky, kterými jsou asistence, zisk míče, bloky a doskoky. Z toho vyplývá, že máme 4 různé vektory vysvětlované proměnné, které je potřeba odstranit, pokud je zrovna nebudeme používat, jelikož informace o konci zápasu nemáme v jeho půlce k dispozici.

6.1 Volba příznaků

Pomocí Endpointu *PlayerGameLogs* jsme získali velké množství různých příznaků, přičemž velkou část z nich nepotřebujeme, respektive by mohly narušovat přesnost našeho modelu. Kromě níže uvedených příznaků bychom ještě mohli přidat informace o počtu vystřelených košů za dva body, jelikož tyto hodnoty se vyskytují v live datech. Příznaky, které byly ponechány:

- GAME_ID
- GAME_DATE
- MATCHUP
- WL
- MIN - Počet odehraných minut
- FGM - Počet daných košů
- FGA - Počet vystřelených střel
- FG_PCT - Procentuální úspěšnost střelby
- FTM - Počet daných trestných hodů

- FTA - Počet vystřelených trestných hodů
- FT_PCT - Procentuální úspěšnost střelby trestných hodů
- OREB - Počet útočných doskoků
- DREB - Počet obranných doskoků
- AST - Počet Asistencí
- TOV - Počet ztrát míče
- STL - Počet zisků míče
- BLK - Počet bloků
- PTS - Počet bodů

Kromě těchto příznaků byly manuálně vytvořeny nové. Tyto nové příznaky jsou výrazně inspirovány příznaky, které byly použity v pracích na podobné téma. Jedná se o aritmetický průměr za posledních 5 nebo 10 her. Ve chvíli, kdy se dostáváme na začátek kariéry hráče bereme průměr maximálních možných her, případně prvních několik her z datasetu odebereme, aby nevytvářely falešné korelace. Tyto průměry jsou vypočítávány z hodnot jak v polovině, tak i na konci zápasu. Umožňuje to tak predikčnímu modelu lépe odhadnout výsledek na základě výkonnosti hráčů v předchozích utkáních. Pokud toto neuděláme a zobrazíme si jednotlivé hodnoty, zjistíme, že po sobě jdoucí hodnoty v grafu jsou velmi náhodné, což modelu spíše škodí. Výčet přidaných příznaků:

- last_5_at_half_ast_avg - průměrný počet asistencí v polovině zápasu v posledních pěti utkáních
- last_5_at_half_reb_avg - průměrný počet doskoků v polovině zápasu v posledních pěti utkáních
- last_5_at_half_blk_avg - průměrný počet bloků v polovině zápasu v posledních pěti utkáních
- last_5_at_half_stl_avg - průměrný počet zisků míče v polovině zápasu v posledních pěti utkáních
- last_10_at_half_ast_avg - průměrný počet asistencí v polovině zápasu v posledních deseti utkáních
- last_10_at_half_reb_avg - průměrný počet doskoků v polovině zápasu v posledních deseti utkáních
- last_10_at_half_blk_avg - průměrný počet bloků v polovině zápasu v posledních deseti utkáních
- last_10_at_half_stl_avg - průměrný počet zisků míče v polovině zápasu v posledních deseti utkáních
- last_5_at_end_ast_avg - průměrný počet asistencí na konci zápasu v posledních pěti utkáních
- last_5_at_end_reb_avg - průměrný počet doskoků na konci zápasu v posledních pěti utkáních
- last_5_at_end_blk_avg - průměrný počet bloků na konci zápasu v posledních pěti utkáních
- last_5_at_end_stl_avg - průměrný počet zisků míče na konci zápasu v posledních pěti utkáních
- last_10_at_end_ast_avg - průměrný počet asistencí na konci zápasu v posledních deseti utkáních
- last_10_at_end_reb_avg - průměrný počet doskoků na konci zápasu v posledních deseti utkáních
- last_10_at_end_blk_avg - průměrný počet bloků na konci zápasu v posledních deseti utkáních
- last_10_at_end_stl_avg - průměrný počet zisků míče na konci zápasu v posledních deseti utkáních

6.2 Výběr hyperparametrů

Modely jako jsou náhodné lesy případně adaBoost mají v *scikit-learn* implementacích možnost určit si hyperparametry. To jsou parametry, které jsou optimalizovány v závislosti na výsledcích ve validační množině. Výběr samotných hyperparametrů a rozmezí hodnot, které můžou nabývat byl řešen z velké části odhadem a experimentováním metodou pokus omyl.

```
parameter_grid = {
    'n_estimators': range(1,120, 6),
    'learning_rate': [0.01, 0.05, 0.1, 0.3, 0.5, 1, 1.5, 2]
}
```

Obrázek 6.1 Zde vidíme výsledné hyperparametry, pro které model AdaBoost produkoval nejlepší výsledky

```
parameter_grid = {
    'max_depth': range(2, 5),
    'n_estimators': range(20, 40),
    'min_samples_split': range(2, 5),
    'criterion': ["squared_error", "absolute_error"]
}
```

Obrázek 6.2 Zde vidíme výsledné hyperparametry, pro které model náhodných lesů produkoval nejlepší výsledky

6.3 Evaluace chyby

Ve chvíli pokud máme spojitou vysvětlovanou proměnnou je potřeba využít nějakou metriku, která nám bude zachycovat chybovost našeho modelu. Pro tyto účely se používají převážně tři různé výpočty. Kde Y_i je skutečná reálná hodnota a $\hat{Y}_i$ je predikovaná hodnota.

- $MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- $RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(Y_i + 1) - log(\hat{Y}_i + 1))^2}$

V této práci pro výpočet chybovosti modelu byl použit MAE, jelikož srozumitelně ukazuje jakou máme průměrně chybu v modelu. Průměrná chybovost se samozřejmě liší v závislosti na tom kterou hodnotu predikujeme, avšak také na hráči ke kterému se daný model vztahuje. Predikujemeli hráče hrajícího na pozici SG je velmi pravděpodobné, že bude mít malý počet bloků, jelikož na této pozici hrají převážně nižší hráčí.

Po provedení experimentů se dostáváme do situace, kdy nejlepší predikce asistencí, bloků a zisků míče vytváří náhodné lesy a pro predikci doskoků je nejlepší použít lineární regresi. Pokud vezmeme například známého hráče ze Slovinka Luku Doncice a aplikujeme na něj modely a vypočítáváme MAE, zjistíme že průměrná chyba je u jednotlivých predikcí následující:

- AST = 1.6
- BLK = 0.16
- REB = 1.99
- STL = 0.67

6.4 Porovnání se kurzy sázkových kanceláří

Jak už zde bylo řečeno, data, která jsme stáhli z oficiálních stránek NBA nemůže používat k žádnému hazardu, byť legálnímu. Ovšem pro akademické účely je dobré otestovat jestli je náš model konkurence schopný. Po prozkoumání dvou největších sázkových společností v Čechii bylo zjištěno, že ani jedna z nich nenabízí živé sázky na hráče při utkáních NBA. Tato skutečnost může být způsobena časovým posunem, avšak některá utkání se hrají okolo deváté hodiny večerní středoevropského času. Z tohoto důvodu předpokládejme, že nemají dostatečně přesný model na to aby kurzy mohli vypsat. V tomto ohledu se dá předpokládat, že náš model by po nějakých úpravách mohl být nasazen do některých sázkových kanceláří, jelikož standartně sázkové společnosti mají několika procentní rezervu, aby měli zisk i při nepřesném odhadu.

Byl však proveden hypotetický test, kdybychom sázeli u jedné mimoevropské společnosti. V příkladu by bylo vsazeno celkem 1500 korun na celkem 15 různých událostí. Sázkové kanceláře ale vypisují sázky ve formátu, kdy je předem dané nějaké predikované číslo a my si můžeme vsázat, jestli hráč nasbírá do konce utkání dané statistiky více či méně. Naše hypotetická sázecí strategie byla přímočará, věřili jsme vždy více našemu modelu a ve chvíli, kdy jsme se trefilo blízko celočíselné hodnotě, která se shodou okolností shodovala s hodnotou, kterou vypsala sázková kancelář, vsadili jsme na lepší kurz. Tímto způsobem bychom z hypotetických 1500 vytvořili 1876, což je zhruba 125% původní částky. Samozřejmě aby naše hypotéza byla relevantnější bylo by zapotřebí provést více sázek, ale i tak je to velmi působivý výsledek.

Kapitola 7
Závěr

Cílem práce bylo vytvořit statistický model, který bude schopen na základě dat z poloviny zápasu predikovat dílčí statistiky hráčů na konci utkání.

V této práci bylo shrnuto základní fungování NBA a stručně popsána pravidla a princip fungování basketbalu. Byly zde popsány rozdíly mezi evropským basketbalem a NBA a nastíněna některá potenciálně důležitá fakta, která by mohla mít vliv na výsledek predikcí, jedním takovým příkladem je analýza střídání v basketbalovém utkání.

Načež bylo navázáno rešerší o pracích, které byly napsány na podobné téma. V tomto případě se vysktl problém, jelikož většina prací se zaměřovala na neživé sázení, což ve výsledku znamenalo, že byly analyzovány postupy, které byly v pracích používány. Tyto postupy byly pak následně poupraveny a aplikovány na náš specifický případ. Avšak konkrétní práci, která by řešila kompletně to samé jsem nenašel. Tento fakt může být brán relativně pozitivně, jelikož má práce přináší něco nového.

Následně byla sepsána analýza dostupných dat, kde byly zkoumány tři různé zdroje, přičemž do většího detailu pouze dva. V případě BR zde byl popsán způsob jak se k datům dostat a kde konkrétní data najít. Následně byly pops ány jednotlivé přiznaky, tak aby čtenář, který se příliš nevyzná v basketbalové tématice porozuměl všem použitým zkratkám. Co se týče oficiálních stránek NBA, zde byla popsána a provedene nejdůkladnější analýza, jelikož se jedná o zdroj, ze kterého bylo ve finále čerpáno. Důraz zde byl kladen i na data, která se aktualizují během utkání a byly zde ukázýny vizualizace zajímavých dat.

V následující kapitole bylo popsáno, jak jednotlivá data získat pomocí knihovny nba_api a co vše je potřeba udělat, abychom získali kompletní a ucelený dataset. Popsány zde byly primárně dva *Endpointy*, které jsou základním stavebním kamenem praktické části této práce.

Ve neposlední řadě zde byl proveden důkladný popis všech příznaků, které byly pro model vytvořeny, jedná se primárně o příznaky, které získávají průměr nějaké statistiky z předchozích her. Dále zde byl krátce rozebrán vůběr hyperparametrů a proces evaluace výsledné chyby. Ve finálé byl vykonán krátky teoretický experiment, kde byl analyzován teoretický zisk, kdybychom výsledky práci zkusili otestovat proti sázkovým kancelářím.

Výsledkem práce je, že byl vytvořen poměrně přesný model, který může konkurovat reálným společnost. Dále byl také položen základ, který říká jaké příznaky jsou pro predikci tohoto typu důležité.

7.1 Možnosti pro navazující práce

Vytvořená práce nezahrnuje všechny možné příznaky, které jsem schopni z aktuálních utkání získat, proto je možné predikce pravděpodobně vylepšit. Dále je určitě možné rozšířit analýzu

použitých metod například o Naivní Bayesův Klasifikátor, případně o hřebenovou regresi. Co se týče jednotlivých skriptů, je pravděpodobné, že proces sběru dat by šel zjednodušit, že se budou stahovat pouze data, která ještě nebyla stažena, což urychlý celý proces o značnou část, což může mít v konečném důsledku podstatný vliv, jelikož poločas v utkání trvá jenom omezenou dobu.

Bibliografie

1. MCCABE, Alan; TREVATHAN, Jarrod. Artificial intelligence in sports prediction. In: *Fifth International Conference on Information Technology: New Generations (itng 2008)*. 2008, s. 1194–1197.
2. THABTAH, Fadi; ZHANG, Li; ABDELHAMID, Neda. NBA game result prediction using feature analysis and machine learning. *Annals of Data Science*. 2019, roč. 6, č. 1, s. 103–116.
3. CAO, Chenjie. Sports data mining technology used in basketball outcome prediction. 2012.
4. ONDŘEJ, Schejbal. *Predikce vybranych událostí v basketbalovém utkání*. 2020. B.S. thesis. České vysoké učení technické.
5. *NBA Win Shares*. [B.r.]. Dostupné také z: `https://www.basketball-reference.com/about/ws.html`.

Obsah přiloženého média

- `readme.md` ..... stručný popis obsahu média
- `live_games` ..... Složka obsahující statistiky z právě probíhajícího utkání
- `player_stats` ..... Složka obsahující statistiky hráčů
- `data_downloader.ipynb` ..... Notebook který stáhne všechna data zadaného hráče
- `live_data_downloader.ipynb` ..... Notebook který umožňuje stažení dat z aktuálně probíhajícího zápasu
- `train_ntb.ipynb` ..... Notebook na natrénování modelů
- `environment.yml` ..... Soubor s conda prostředím
- `BAP` ..... Zdrojové kódy práce
- `bakalarska_prace.pdf` ..... text práce ve formátu PDF