VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

VYHODNOCOVÁNÍ SÁZKOVÝCH KURZŮ NA ZÁPASY PREMIER LEAGUE

EVALUATION OF BETTING ODDS OF PREMIER LEAGUE'S MATCHES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **TOMÁŠ ZEJDA**
AUTHOR

VEDOUCÍ PRÁCE **Ing. JIŘÍ HYNEK, Ph.D.**
SUPERVISOR

BRNO 2020

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav informačních systémů (UIFS) Akademický rok 2019/2020

Zadání bakalářské práce

Student: **Zejda Tomáš**
Program: Informační technologie
Název: **Vyhodnocování sázkových kurzů na zápasy Premier League**
Evaluation of Betting Odds of Premier League's Matches
Kategorie: Umělá inteligence

Zadání:

1. Seznamte se s principy strojového učení založeného na technice učení s učitelem.
2. Proveďte průzkum nejvyšší anglické fotbalové ligy - Premier League. Prostudujte její pravidla, popište specifika ligy a vysvětlete, v čem je daná liga odlišná od ostatních světových fotbalových lig. Dále analyzujte atributy, které mají vliv na sázkové kurzy a výsledky této ligy.
3. Navrhněte model reprezentující vybrané atributy z bodu 2. Získejte reálná historická data.
4. Navrhněte aplikaci pro vyhodnocování kurzů a předvídání výsledků zápasů nejvyšší anglické fotbalové ligy Premier League využívající predikci výsledků na základě historických dat.
5. Navrženou aplikaci implementujte.
6. Otestujte funkcionalitu a použitelnost výsledné aplikace. Navrhněte možné rozšíření aplikace pro další světové fotbalové ligy.

Literatura:

- Raschka, S.: *Python machine learning*. Packt Publishing Ltd., 2015, ISBN: 978-1-78355-513-0.
- Coursera: *Machine Learning.* [online]. 2019 [cit. 2019-10-13]. Dostupné z: https://www.coursera.org/learn/machine-learning
- SciKit-Learn Developers: *SciKit-Learn User Guide* [online]. 2019 [cit. 2019-10-13]. Dostupné z: http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 až 4.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Hynek Jiří, Ing., Ph.D.**
Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2019
Datum odevzdání: 31. července 2020
Datum schválení: 22. října 2019

Abstrakt

Kurzové sázení je momentálně velkým fenoménem a tato práce se zaměřuje na vyhodnocení klíčových faktorů v sázení na nejvyšší anglickou fotbalovou ligu. Účelem této práce je dát sázejícímu odpověď na to, zda se s daným kurzem vyplatí vsadit si na určitý zápas při zvážení všech rizik, či nikoliv. V této práci jsou zanalyzována určitá data, na kterých by měl záviset výsledek zápasu. Data jsou posbíraná z několika předchozích ročníků dané soutěže a také z dalších soutěží, kterých se týmy z nejvyšší anglické fotbalové ligy účastní.

Abstract

Betting on sports is currently a trending phenomenon and this thesis is focusing on evaluating key factors for betting on the Premier League football matches. The aim of this thesis is to provide a better with an answer to whether it is a good idea to make a bet on a certain match considering current odds and risk factors or not. In this thesis we analysed certain amount of data on which the match results shall be depending on. We collected the data from previous seasons of this League as well as from other football competitions the Premier League teams take part in.

Klíčová slova

neuronové sítě, fotbal, anglický fotbal, Premier League, kurzy, kurzové sázení

Keywords

neural network, football, soccer, English, Premier League, betting, odds, English football

Citace

ZEJDA, Tomáš. *Vyhodnocování sázkových kurzů na zápasy Premier League*. Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Jiří Hynek, Ph.D.

Vyhodnocování sázkových kurzů na zápasy Premier League

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Jiřího Hynka, Ph.D.. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Tomáš Zejda
30. července 2020

Poděkování

Rád bych poděkoval Ing. Jiřímu Hynkovi Ph.D. za vedení při této práci, za rady a konzultace k problematice a obrovskou trpělivost, kterou se mnou měl.

Obsah

Kapitola 1

Úvod

Kurzové sázení je v současné době velmi oblíbená zábava. Nejpopulárnější jsou sázky na sport, zejména pak na fotbal. I přesto, že se nejedná o jeden z nejsnazších sportů na ziskové sázení, těší se obrovské oblibě po celém světě. O něco lépe se dá předpovídat konečné skóre například tenisových zápasů, kde mohou nastat pouze dva výsledky a je zde méně faktorů, které ovlivňují zápas. Přesto je hlavním cílem práce nalezení algoritmu, který by umožňoval dlouhodobý zisk ze sázek právě na fotbalová utkání. Konkrétně se tato práce zaměřuje na nejvyšší anglickou fotbalovou ligu Premier League.

Anglická nejvyšší fotbalová soutěž je považována za jednu z nejzajímavějších a nejvyrovnanějších na světě. Momentálně existuje pouze malé množství těžko dostupných programů a algoritmů, které se snaží předpovídat procentuální šanci na remízu či vítězství jednoho nebo druhého týmu. Většina z nich není příliš přesná. Cílem práce je vytvořit program, který dokáže zjistit, zda se na základě těchto šancí a vypsaných kurzů vyplatí sázet na daná utkání. Jedná se o propracovanější systém, který má navíc také odpovědět na otázku, zda se opravdu vyplatí investovat do dané konkrétní sázky. Je ale ovšem stále řeč o sportu, který je nevyzpytatelný a mohou nastávat výsledky proti všem předpokladům.

Ve druhé kapitole 2 této práce je popsána anglická fotbalová liga, její specifika a atributy od kterých se odvíjí výsledky zápasů. Třetí kapitola 3 se zabývá kurzovým sázením, základní charakteristikou a následně rozvedením sázení na sport a konkrétně na anglickou ligu. Poté jsou zde zmíněny další podobné algoritmy. Jejich účelem bylo předpovídání výsledků zápasů a snaha o finanční zisk z kurzového sázení. Ve čtvrté kapitole 4 je popsáno strojové učení, které bylo využito pro výpočet pravděpodobnosti vítězství v zápasech. V páté kapitole je vysvětleno, jakým způsobem byl algoritmus implementován, jak byla získávána data a jak probíhalo jejich vyhodnocení.

Kapitola 2

Premier League

Fotbal je týmový sport a řadí se mezi nejpopulárnější kolektivní sporty na světě. Takový, jaký ho dnes známe, vznikl v Anglii v 16. století. Hraje se s jedenácti hráči na každé straně, přičemž jedním z nich je brankář. Hrát můžou týmy v libovolném rozestavení, které si zvolí. V zápase se tedy kromě brankáře objevují v základním rozdělení také obránci, záložníci a útočníci. Minimální počet hráčů pro jeden tým je sedm. Ve většině světových lig a také v té anglické je maximální počet střídaných hráčů v zápase omezen na tři, což bylo ovšem v poslední době změněno a ovlivněno virovou krizí. Stejně tak je omezen počet hráčů na střídačce na sedm hráčů.

Zápas je rozdělen na dvě poloviny po čtyřiceti pěti minutách, kdy čas plyne bez přerušení. K základní hrací době se v daném poločase přidává čas, po který byl zápas přerušen z důvodu ošetřování či střídání. V anglické lize se každý zápas hraje o body, které určují pořadí v tabulce. Za vítězství tým získává tři body, za remízu jeden a za prohru nemá bod žádný. Pokud nastane nerozhodný výsledek, zápas se nikdy neprodlužuje. Důležité je si uvědomit, že nastavený čas není to stejné, co prodloužení. [4]

2.1 Nejvyšší anglická fotbalová liga

Anglie se právem označuje za kolébku fotbalu. Její nejvyšší liga je nejsledovanější fotbalovou ligou světa a tím pádem také ligou, na kterou se nejvíce sází. Každý zápas je velice ostře sledován nejen anglickým publikem. V současném formátu, v jakém ji známe dnes, byla založena roku 1992. V soutěži je 20 týmů, které se ve 38 kolech utkají o titul, o místa v pohárech a o sestup do nižší ligy. Jedná se o nejlepší celky Anglie a Walesu. První 3 týmy postupují přímo do nejvyšší evropské klubové soutěže, která se nazývá Liga mistrů, 4. celek ligy o ni hraje předkolo. Pátý celek postupuje do nižší Evropské ligy. Naopak poslední 3 týmy sestupují do nižší soutěže. [5]

Na anglické lize je specifická její špička. Do té se dá počítat až 6 týmů, což není úplně obvyklé. Tyto týmy mají silné postavení i ve světovém fotbalu a každým rokem se očekává, že ve finální tabulce v první šestici týmů této ligy budou právě Manchester United, Manchester City, Chelsea FC, Liverpool FC, Tottenham Hotspur a Arsenal London. Tyto týmy mají také k dispozici více finančních prostředků než zbytek ligy a tomu odpovídá i kvalita jejich kádru.

Stále se však bavíme o sportu, a tak to nemusí být pravidlem. U zbylých čtrnácti týmů není z počátku vůbec nic jisté a další vývoj se dá předpovídat až po několika odehraných kolech. Predikce před sezónou často končí fatálním neúspěchem. Příkladem může být sezóna

2015/2016, kdy celou ligu vyhrál tým Leicester City navzdory tomu, že před sezónou na jejich vítězství byl vypsán kurz pět tisíc ku jedné.[22]

Důležité jsou také další soutěže, do kterých se mužstva z anglické nejvyšší soutěže zapojují. Jedná se o FA Cup, Anglický ligový pohár, Ligu mistrů a Evropskou ligu. Zápasy z těchto soutěží nemohou být ignorovány, přestože se práce soustředí pouze na Premier League. Každá soutěž má jiný systém a nastupují v nich týmy zcela odlišných úrovní, a i s těmito faktory je potřeba při výpočtu počítat. Ku příkladu v FA Cupu můžou mužstva nastoupit proti týmu z páté nejvyšší anglické soutěže, naproti tomu v Lize mistrů mohou hrát proti nejlepším celkům ze Španělska, rozdíl v zápasech proti Stevenage FC s hodnotou týmu 400 tisíc euro [16] a Barceloně, která má momentálně hodnotu 897 milionů euro [13], je zřejmý, obě hodnoty jsou k datu 3. 7. 2020. I pro samotné týmy mají zápasy různou prioritu, a to se nejvíce odráží na sestavě, kterou do zápasu nasazují.

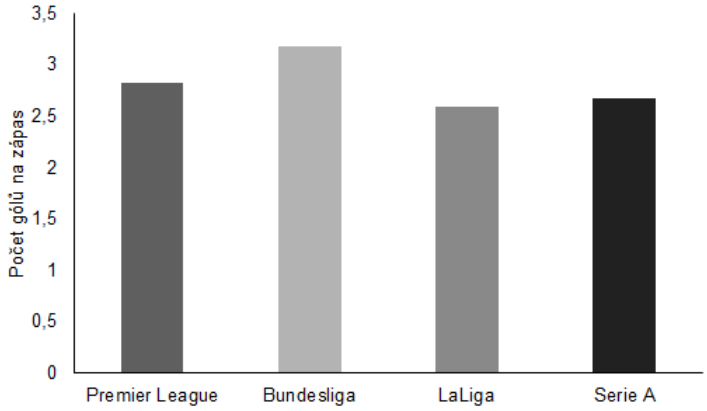
V průběhu sezóny, jak Leicester stále vyhrával, se kurz začal značně snižovat. Zatímco sázkaři stále nebrali celkové vítězství Leicesteru jako reálnou věc, v tomto okamžiku by už program, na rozdíl od lidského úsudku, chybu v podcenění týmu neudělal. Program jako takový nebere v potaz jméno týmu jako atribut, na kterém by mělo záležet vítězství. Něco takového, co se povedlo Leicesteru, se opakuje jednou za padesát let a ani zdaleka to není pravidlem. Tato skutečnost umožňuje pravidla nastavit a zjistit atributy, které mají na výsledku největší podíl.

2.2 Porovnání s ostatními ligami

Při srovnání nejlepších lig Evropy se často hovoří o top pěti soutěžích. Zahrnují německou Bundesligu, španělskou LaLigu, italskou Serii A a francouzskou Ligue 1. Pro svoje porovnání jsem se rozhodl vyloučit francouzskou nejvyšší ligu a to z toho důvodu, že v top třiceti klubech mají pouze dva zástupce podle hodnocení klubových koeficientů UEFA [12]. Proti tomu má Anglie mezi prvními třiceti kluby šest týmů stejně jako Španělsko, Německo tři a dva kluby těsně za hranicí třicítky a Itálie má také tři zástupce. Na následujících grafech jsou ukázány rozdíly v několika faktorech, které mohou mít vliv na vítěze zápasů.

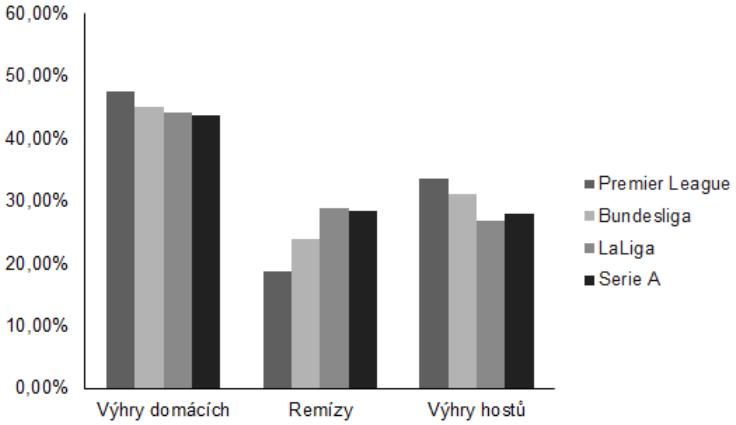
Všechny následující statistiky jsou vytvořeny z poslední uplynulé sezóny 2018/2019. Z grafu 2.1 je patrné, že co se týče vstřelených branek, vyčnívá nejvyšší německá soutěž. V porovnání s ostatními soutěžemi v ní soupeří pouze osmnáct klubů, ve zbylých zmiňovaných soutěžích klubů dvacet. Většina týmů v Německu se snaží hrát ofenzivní fotbal, a to i týmy ve spodních patrech tabulky, a tak často zápasy končí vysokými rozdíly. V ostatních ligách připisuji poměr nižších branek jiným faktorům. V Anglii je to dáno vyrovnaností a také tím, že se hraje už více na výsledek než pro zábavu fanoušků. Fotbal je tedy více zaměřený na taktiku a je opatrnější. V Itálii vždy převládal fotbal postavený na obraně, kdy stačí vstřelit jeden gól, když tým nedostane žádný. Ve španělské lize málo vstřelených branek způsobují nepříliš kvalitní celky ze spodní poloviny tabulky, které ve vzájemných zápasech nechtějí prohrát, a proti silnějším týmům se snaží neutrpět příliš velkou porážku. V posledních letech se i španělská liga o něco více vyrovnává, a i slabší celky jsou schopné uhrát zajímavé výsledky proti těm nejlepším.

U výhody domácího prostředí v grafu 2.2 je na tom nejlépe anglická liga, což se dá přičíst především návštěvnosti. Hned za ní je totiž s úspěšností německá liga. Na rozdíl od dalších dvou zmiňovaných soutěží je zvykem, že i zápasy týmů ze spodních příček německé a anglické ligy mají vysokou návštěvnost a často bývá vyprodáno. Ve Španělsku a Itálii na neatraktivní zápasy bývá stadion z větší části prázdný. V těchto ligách je také zvykem na



Obrázek 2.1: Průměr gólů vstřelených za zápas ve čtyřech nejlepší ligách Evropy [17].

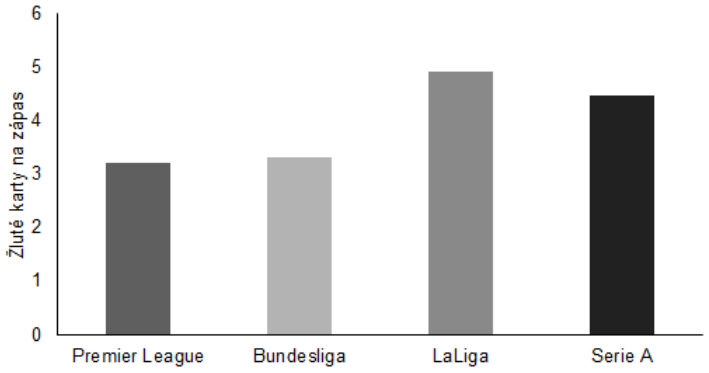
svůj tým pískat, pokud se nedaří. To je opět něco, co se v Premier League a v Bundeslize děje pouze ve velice výjimečných případech.



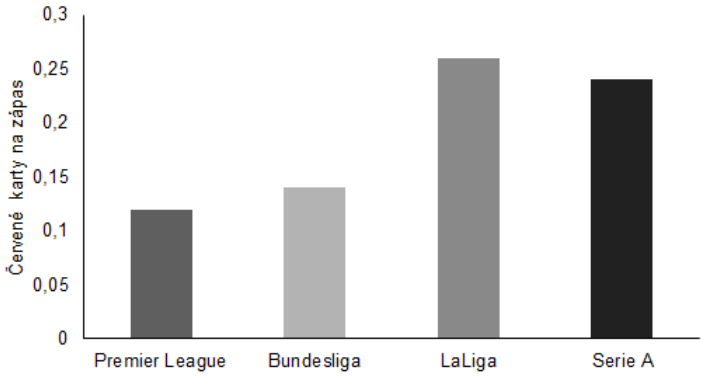
Obrázek 2.2: Poměr pro výsledky na domácích stadionech ve čtyřech nejlepší ligách Evropy [15].

Na grafech 2.3 a 2.4 je vidět, že italská a španělská liga mají největší počet žlutých a červených karet. Tento fakt připisuji dvěma faktorům. Prvním z nich je ten, že se během těchto zápasů více simuluje. Pokud rozhodčí pozná, že se jedná o simulaci, může hráče potrestat žlutou kartou. Pokud se mu to však nepodaří, tak některé zákroky vypadají hůře, než co se ve skutečnosti událo a opět rozhodčí může udělit kartu. V Anglii i v Německu se na simulování stále pohlíží s odporem a diváci si na to nezvykli. Hráči, kteří se toho dopustí jsou často veřejně kritizováni, a i na jeden takový incident se následně vzpomíná několik sezón, tak je simulování v těchto ligách stále spíše výjimka.

Druhým důvodem je, že zápasy jsou emotivnější. Hráči se snadněji nechají vyprovokovat a v zápalu hry následně oplácí předchozí zákrok. Také dochází k různým strkanicím, výměnám názorů, které jedna strana psychicky neunese a po těchto incidentech se také rozdávají žluté a někdy i červené karty.



Obrázek 2.3: Průměr žlutých karet na zápas ve čtyřech nejlepší ligách Evropy [17].



Obrázek 2.4: Průměr červených karet na zápas ve čtyřech nejlepší ligách Evropy [17].

Všechny tyto atributy budou použity pro výpočet šance na úspěch týmu. Některé faktory jsou v určitých ligách výraznější než v jiných. 5.1

Právě z těchto důvodů se práce nezabývá všeobecným algoritmem pro fotbal jako takový, ale zaměřuje se na jednotlivou ligu. Každá soutěž se liší, a proto se očekává, že model pro predikci by při zpracování ostatních soutěží změnila váhy pro jednotlivé parametry jinak, přestože samotné parametry, které ovlivňují zápasy se lišit nemusí.

Kapitola 3

Kurzové sázení

Kurzové sázení funguje na principu dohody dvou stran. Jedna strana, která na základě předpokladů vypisuje kurzy a druhá strana, která si na základě těchto kurzů vybírá sázku se zhodnocením rizik a následně vybere částku, kterou chce na danou příležitost vsadit. Nejčastější možností jsou sázky u sázkových kanceláří. Momentálně na trhu dominuje online sázení. Nedochází se do sázkových kanceláří s opisováním kódů zápasu a příležitostí, které chce sázející na zápas vsadit, ale vše se řeší přes web nebo mobilní aplikace. Na zápasy se dá sázet před jejich začátkem i živě během jejich průběhu.

3.1 Sázení mimo sázkové kanceláře

Kromě sázkových kanceláří kurzy vypisují i jednotliví bookmakeři. Bookmakerovi je možné zaslat i vlastní tipy a ten na jejich základě vypíše kurz. Mějme příklad zápasu Premier League, kde se utká Arsenal a Chelsea. Pokud by si sázkař chtěl vsadit na to, že Arsenal vyhraje přesně o jeden gól, v zápase padnou více než čtyři žluté karty a Chelsea bude mít alespoň tři rohové kopy, tuto možnost sázkové kanceláře neposkytují. Sázky jsou u nich omezené na jednu příležitost na zápas, to znamená možnost vsadit si pouze například na počet rohů nebo na to, že Arsenal vyhraje o jednu branku. Právě kvůli tomu poskytují svoje služby bookmakeři. U těchto sázek značně stoupá riziko, protože se jedná o kombinaci více událost, které všechny musejí vyjít. S vyšším rizikem se zvedá i kurz.

3.2 Možnosti sázek

Možné je sázet na společenské události například volby, reality show, závěry populárních seriálů, vývoj pandemie nebo ceny potravin. Obecně platí, že čím větší a populárnější sázková kancelář je, tím více možností pro sázení má. Nejčastější jsou samozřejmě stále sázky na sportovní utkání. Do kategorie sportů se v poslední době dostala i možnost sázení na e-sporty. Při sportovních událostech se jedná o široké pokrytí v podstatě všech sledovatelných sportů. U anglické fotbalové ligy je obvykle přes sto možností na co je možné si vsadit při každém utkání. Počínaje výsledkem zápasu, přes střelce gólů, počty žlutých karet, penalt, červených karet, handicapy pro jednu či druhou stanu až po vlastní góly. V bakalářské práci se budu zabývat sázkami na výsledek zápasu, tedy výhra domácích, remíza nebo výhra hostů.

3.3 Formy tiketů

V této podkapitole budu brát v potaz pouze sázení u sázkových kanceláří nikoli sázky u bookmakerů, protože pro práci nejsou s ohledem na jejich povahu relevantní. Stejně tak sázky živě nejsou cílem této bakalářské práce, protože k vyhodnocení všech faktorů dochází před zápasem.

Sázet se může několika způsoby. Jako nejlepší se pro dlouhodobé sázení ukazují sólo tikety, tedy tikety s jednou příležitostí, protože představují nejmenší riziko. Kurz, který byl na danou příležitost vypsán se vynásobí částkou, kterou sázkař podal. Další z možností je zápasy na tiketu kombinovat. V tomto případě můžeme tipovat více zápasů na jeden tiket s tím, že kurz se násobí. Pokud je například vsazeno na vítězství s kurzem dva a na remízu s kurzem tři, výsledný kurz bude šest a tímto kurzem se pak násobí vsazená částka. Kurz je sice lákavější, ale je zde vyšší riziko, že pokud jeden z tipů nevyjde, celá vsazená částka propadne sázkové kanceláři.

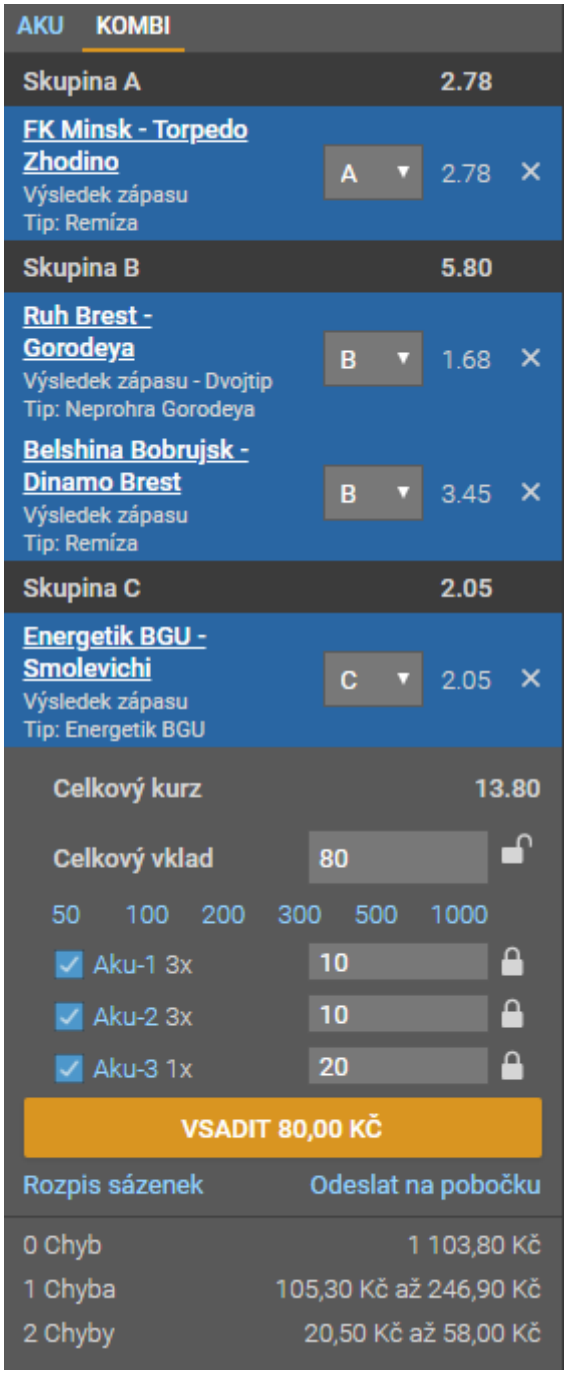
Zajímavé jsou kombinované tikety, kde si sázkař zápasy rozdělí do skupin a následně si určuje částku, kterou chce vsadit na každou z nich. Pro každou skupinu se počítá kurz jednotlivě. Čím více je skupin, tím více se částka, kterou chce sázející podat, násobí. Například u tří skupin se pro první dvě částka násobí číslem tři 3.1. Pokud bude vsazeno na první a druhou skupinu deset korun, tak musí být vsazeno třicet korun na každou z nich, stále se však do celkového zisku bude kurz násobit pouze deseti. Pokud je neúspěšný tip na jeden zápas ze skupiny, je skupina vyhodnocena celá jako nevýherní, ale stále může být úspěšný zbytek tiketu. Díky tomu může sázkař skončit v zisku. Nevýhodou je, že celkový vklad musí být větší a také, že celkový kurz nevzniká přímo násobením kurzu jednotlivých skupin a na rozdíl od předchozích příkladů je nižší. Ve skupinách může být jeden nebo více zápasů, limity na počet zápasů a celkový počet skupin se liší v různých sázkových kancelářích. Přesné výpočty s tím, jak je to s finálním určením kurzu stejně tak s číslem, kterým se násobí částka pro jednotlivé skupiny si kanceláře nechávají pro sebe.

3.4 Výpočet kurzů

Kurzy jsou počítány na základě pravděpodobnosti dané události. Nejtradičnější jsou dva zápisy, jeden z nich je typický pro dostihové závody. Pokud je kurz na vítězství koně v závodu například čtrnáct ku jedné, vsazená částka by byla násobena hodnotou 4,33 [6]. Pro tuzemské sázkové kanceláře jsou však častěji kurzy zapisovány už rovnou formou desetinných čísel.

Tato čísla jsou počítána tím způsobem, že se sto procent podělí pravděpodobností dané události. Vezměme příklad týmu A a týmu B s pravděpodobností výhry týmu A 46 %, remízy 28 % a výhry týmu B 26 %. Kurz na vítězství týmu A by byl 2,17, na remízu 3,57 a na úspěch týmu B by byl kurz 3,85. Takto se dopočítáme k čistým kurzům. Sázková kancelář následně tyto kurzy ještě přibližně o 10 % sníží.

Nejlépe je to pozorovatelné u naprosto vyrovnaných zápasů s dvěma možnými výsledky. U nejznámější české sázkové kanceláře Tipsport je kurz na oba účastníky 1,85, u druhé nejznámější kanceláře Fortuna je to ještě o setinu nižší. Část, o kterou byl kurz snížen, se označuje jako marže a slouží k zajištění zisku daných sázkových kanceláří.



Obrázek 3.1: Příklad kombinovaného tiketu [14]

3.4.1 Proměnlivost kurzů

Kurzy mohou být vypsány dlouhou dobu před zápasem. Opět záleží na jednotlivých kancelářích, ale nejčastěji jsou k dispozici týden až dva před utkáním.

Na obrázku 3.2 je možné pozorovat vývoj kurzu pro jedno z největších derby ve světovém fotbale, a to pro zápas Manchesteru United proti Liverpoolu, kdy první jmenovaný byl domácím celkem. Kurz byl vypsán po kole, kdy United porazili Chelsea poměrem 2 - 1, v té

době se klub Chelsea nacházel v první čtyřce tabulky. Liverpool vyhrál nad West Hamem ze středu tabulky 4-1. V této chvíli byl kurz, jak je patrné z grafu, na hodnotě 2,3 na United, 3,3 na remízu a 3,2 na výhru Liverpoolu. Další zápasy se odehráli 3.3, kdy Liverpool porazil Newcastle United 2-0, což byl očekávaný výsledek a na kurz neměl žádný vliv. Manchesteru United se však 5.3 při zápase na půdě Crystal Palace příliš nedařilo a až v posledních šesti minutách dokázali vývoj zápasu otočit a vyhrát poměrem 3:2 gólem v nastavení. Jejich výkon nebyl vůbec přesvědčivý a dlouhou dobu to vypadalo, že zápas prohrají. Kurz na Manchester United po zápase proto stoupl.

Naopak Liverpool měl ještě jeden zápas, a to čtyři dny před derby s United. Bylo to utkání Ligy mistrů, kde měl Liverpool již téměř jistý postup a celé utkání se odehrálo v tréninkovém tempu. Zápas tak skončil remízou, ale kurz na Liverpool přesto klesl, a to především z důvodu šetření hráčů, což byl jasný signál, že pro tým je důležitější víkendový zápas s Manchester United.

V grafu je možné pozorovat i další náhlé změny mimo hrané zápasy. Nejčastější jsou tři příčiny, jednou z nich je rozložení vkladů. Po vypsání kurzů bookmakerem se sázkařům může jeden z kurzů zdát výhodnější a většina z nich tak bude sázet právě na něj. V ten moment kurz klesne, aby sázková kancelář neměla příliš velké ztráty. Další příčinou můžou být špatné výroky ze strany manažerů nebo hráčů, které naznačují nepokoj v kabině nebo malou důvěru v mužstvo, stejně tak jako zranění klíčových hráčů při tréninku. Poslední příčinou je balancování kurzů s ohledem na ostatní sázkové kanceláře. Na toto téma se vyjádřil bookmaker Tipsportu s přezdívkou palo s odpovědí, že cílem asi všech kanceláří je mít "správný"kurz a ten se musí blížit trhovému průměru, pokud trh funguje správně [9].

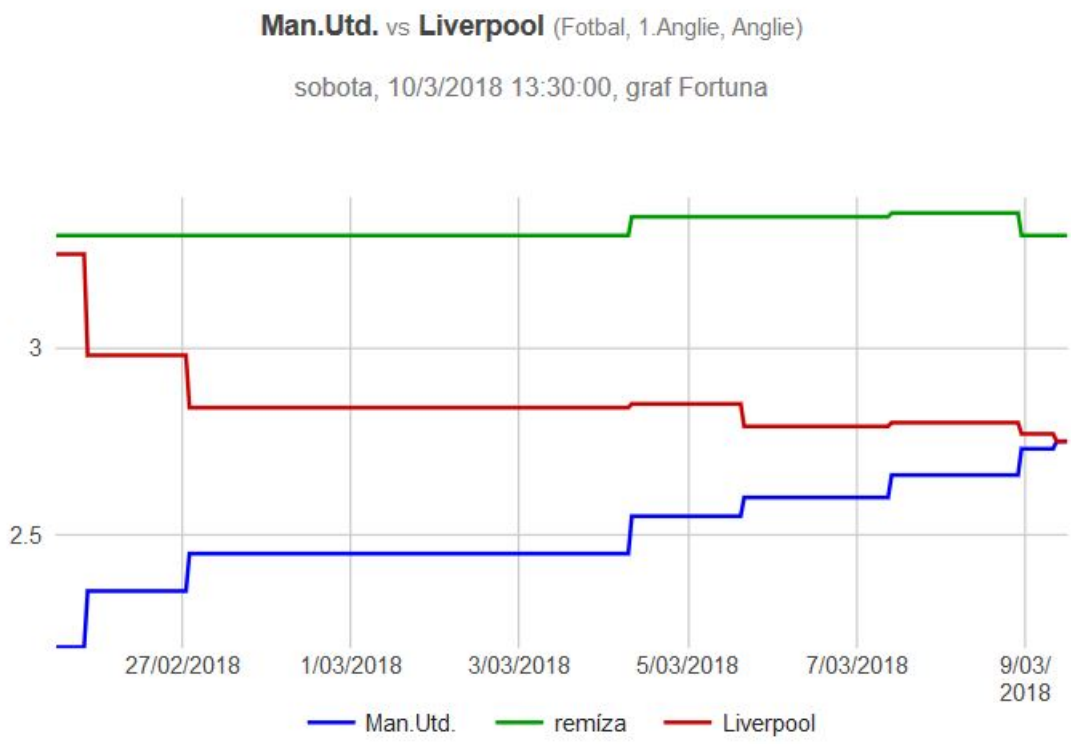
Nejpřesnější je kurz těsně před začátkem utkání, kdy jsou zveřejněny i soupisky, a právě proto se tímto směrem vydávám i v bakalářské práci. Soupisky jsou známy hodinu před utkáním, a to je většinou poslední moment, kdy se kurzy mohou nějak zásadněji změnit. I zde můžeme vidět, že se nakonec kurz posunul na vyrovnanou hodnotu přibližně 2,75 na oba celky.

3.5 Profit ze sázení

U profitu je důležitý zejména ten dlouhodobý. Není až tak zajímavé, že po jednom kole je sázkař v zisku nebo deficitu několika procent, ale důležitý je výsledek po pěti a více kolech. Při sázení na fotbal existuje nepřeberné množství strategií, kterými je možné se řídit. Jako příklad bych uvedl bakalářskou práci Kubeše [26] - Modelování a odhadování výsledků sportovních utkání za použití pravděpodobnosti a statistiky. V práci se zmiňuje o úspěšném použití modelu Dixona a Colese z roku 1997, model je ze statistického hlediska založen na Poissonově rozložení pravděpodobnosti. Do předpokladu bral v potaz výhodu domácího týmu, pravděpodobnost na dosavadních výsledcích s větší relevancí pro nejnovější výsledky a zohledňoval sílu útoku a obrany obou týmů, které se zápasu zúčasnily. Metodu testoval úspěšně právě na anglické nejvyšší fotbalové lize, na nižší anglické lize už takový úspěch neměla.

Dalším příkladem je diplomová práce na téma Statistické a rozhodovací postupy při sázení[27], ve které se zabýval sázením na tři možnosti a to výhru, remízu či prohru. Vycházel z dat nejvyšších fotbalových lig z Česka, Německa a Anglie. Jako největší problém pro dosažení profitu uvádí marže sázkových kanceláří.

Práce z VUT se zaměřila na použití neuronových sítích při snaze o ziskové sázení na tenisové zápasy. V závěru práce bylo zmíněno, že samotná neuronová síť nebyla dostačující pro dosáhnutí správných kurzů. Hlavní rozdíl byl také ve snaze predikovat výsledky teni-



Obrázek 3.2: Vývoj kurzu před zápasem Manchester United vs. Liverpool [28]

sových, a ne fotbalových zápasů [21]. Nutné je také zmínit, že sázkové kanceláře mohou snižovat maximální částku vkladů pro jednotlivého uživatele, což také značně znesnadňuje dlouhodobý profit.

3.6 Získávání kurzů

V Česku se na trhu nachází několik kanceláří, od kterých lze získávat kurzy. Nejznámější z nich je Tipsport. Poskytují jak kurzy před zápasem, tak live v průběhu utkání, navíc se k nim dá dostat i bez registrace a přihlášení uživatele. API pro stahování kurzů poskytují pouze svým partnerům. Podobně funguje i dceřiná společnost Chance, která patří pod Tipsport. Jejich web se v podstatě neliší a vypisují i stejné kurzy.

Další sázková kancelář, která stojí za zmínění, je Fortuna. Co se týká přístupu na stránku a ke kurzům, neliší se od Tipsportu. Jelikož už se jedná o jinou společnost, tak se liší ve výšce kurzů vypsaných na zápasy. U nižších kurzů jsou rozdíly obvykle v řádu několika setin, výjimečně se vyšplhají přes desetiny, naopak u těch nejvyšších jsou už rozdíly v desetinách a v některých případech jdou i do jednotek. Nevýhodou je, že Fortuna má větší marže, jenž se většinou pohybují od 1.05 do 1.09 proti tomu Tipsport, u kterého je obvykle marže od 1.04 do 1.07 [28].

Poslední dvě větší sázkové kanceláře na našem území jsou SYNOT TIP a Sazkabet. Obě z nich se neliší od dříve zmíněných, co se týká nabídky kurzů a API. Liší se ve vypsaných kurzech, protože každá kancelář si je vypisuje sama. Co se týká marží, mají blíže k sázkové kanceláři Fortuna.

Pro kompletní srovnání kurzů, co se českých kanceláří týká, se mi nepodařilo najít bezplatnou platformu. Proto pro vyhledávání kurzů v této práci bude sloužit skript, který bude kurzy stahovat v reálném čase přímo z webů kanceláří.

Kapitola 4

Strojové učení

Jedná se o oblast v umělé inteligenci, která umožňuje strojům se učit pomocí programů. Strojové učení nemá stanovenou svoji přesnou definici. Arthur Samuel poprvé použil termín strojové učení v roce 1959 a popsal ho tak že: Počítač může být naprogramovaný takovým způsobe, že se naučí hrát dámu lépe, než ten kdo ho programoval [36]. Modernější definici z roku 1998 sestavil Tom Mitchell, který strojové učení definoval jako: Počítačovému programu je dáno, ať se učí ze zkušenosti E s braním ohledu na úkol T a výkonnostním opatřením P, pokud se jeho výsledky na T, jak jsou opatřeny pomocí P, zlepšují se zkušenostmi z E [29]. Tato definice se dá přenést na tuto bakalářskou práci s tím, že úkol T je zjistit, kdo má největší šanci vyhrát následující zápas. Šance na správný výsledek je P a zkušenosti E by byly nasbírané z výsledků a znalostí o předchozích zápasech.

Strojové učení nabízí spoustu možností, jak se vypořádávat s velkým množstvím dříve pro počítač neřešitelných problémů. Využití je široké, ať už se jedná o použití při zobrazovaní příspěvků nebo reklam na webu, či filtrování e-mailu při zjišťování, zda se jedná o nevyžádanou poštu. Samotné strojové učení už je známé poměrně dlouho, jak lze vyčíst z dříve zmíněné definice Arthura Samuela. Ten vytvořil program pro hraní šachů, přestože v nich sám nebyl zkušený hráč. Program se však učil z toho, jaké pozice figurek na hrací ploše vedly k vítězství. Dokázal simulovat obrovské množství zápasů a situací, mnohem větší objem zápasů, než by zvládnul běžný člověk. Z těch pak vybíral ty úspěšnější a snažil se je opakovat a zlepšovat [31].

4.1 Algoritmy strojového učení

Tato práce bude vycházet z dat z předešlých let a vhodné algoritmy se tak omezí pouze na učení s učitelem, ty vycházejí z předchozích zkušeností. Mají k dispozici informace o tom, které vstupy odpovídaly kterým výstupům. Na těchto informacích se postupně učí, na základě zkušeností, které postupně získávají, upravují výsledky. Tomuto procesu se také říká trénování. Například u odhadu ceny nemovitostí, jako vstup bude použita velikost v metrech čtverečních a jako výstup odhad ceny. Algoritmus by dostal množinu dat, u kterých by měl předchozí informace, že sto metrů čtverečních stojí tři milióny, sto dvacet metrů tři a půl miliónu a spoustu dalších informací ve stejném formátu, na kterých by měl možnost se natrénovat. Po trénování je určena ještě množina dat pro testování. Pro to, jak dobře se algoritmus natrénoval, už se používají pouze vstupy a na jejich základu program odhaduje výstupy. Ty se pak porovnají s těmi skutečnými, aby se zjistilo, jak dobře je algoritmus natrénován. Reálně je samozřejmě pro cenu nemovitosti více kritérií, které by bylo nutno

zohlednit. Proti tomu učení bez učitele začíná bez jakýchkoliv předešlých znalostí. Používá se například pro shlukování dat, kdy k sobě přiřazuje vstupy s podobnými vlastnostmi.

Od algoritmu se očekává navracení reálných hodnot, a to v podobě procentuálních šancí na jednu z možností. Proto problematika spadá pod regresní řešení úloh, kdy je úkolem navrátit hodnotu na rozdíl od klasifikačních úloh, kdy je cílem algoritmu zařadit nový prvek mezi předem rozpoznané.

Z nejznámějších algoritmů se nabízí lineární regrese, která slouží k získání předpisu, kdy pokud mezi dvěma proměnnými existuje přímá souvislost, tak umožní předpovědět hodnotu jedné ze znalosti druhé proměnné. Mezi těmito body lze proložit přímku 4.1 i v N rozměrném prostoru. Základ pro lineární regresi je přímka s následujícím předpisem:

$$y = ax + b \tag{4.1}$$

Kde y je odezva nebo také závislá proměnná na x, kterému se říká prediktor. Koeficient b je konstantní člen, jehož hodnota udává posunutí přímky po ose y. Druhý koeficient a určuje sklon přímky. Oba se odhadují takovým způsobem, aby co nejvíce popisovali datový soubor. Pro výpočet těchto koeficientů se používá metoda nejmenších čtverců. Přímka 4.2:

$$y = ax + b \tag{4.2}$$

je přímka, proložená metodou nejmenších čtverců souborem bodů

$$[x_1, y_1], [x_2, y_2], ..., [x_n, y_n],$$

jestliže pro koeficienty a, b platí 4.3, 4.4

$$a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i, y_i, \tag{4.3}$$

$$a \sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \tag{4.4}$$

Pro jednoduchý příklad dat při ceně za velikost pozemku 4.1, kde x reprezentuje rozměry a y reprezentuje cenu.

x	y
100	50 000
200	130 000
300	150 000

Tabulka 4.1: Hodnoty pro příklad s lineární regresí

Podle rovnic, které musí platit, mohou být sestaveny rovnice 4.5, 4.6 pro tento příklad. Po spočítání všech sum se hodnoty dosadí do soustavy lineárních rovnic

$$140000 * a + 600 * b = 76000000 \tag{4.5}$$

$$600 * a + 3 * b = 330000 \tag{4.6}$$

Po výpočtu těchto rovnic je koeficient a = 500 a koeficient b = 1000, proto by byl nejlepší předpis přímky y = 1000 + 500 x, přímka by sice žádný z bodů přímo neprotnula, ale blížila by se k řešení v každém z jejích bodů [38].

U sportu se nemůže počítat s lineárním průběhem. O něco lepší by mohla být polynomionální regrese, kde se místo přímky data snaží proložit polynomem o n stupních, přičemž první stupeň je lineární. Čím vyšší stupeň polynomu je vybrán, tím komplexnější křivce se může přizpůsobit.

Nejlepší pro řešení regresních problémů s větším množstvím atributů jsou neuronové sítě, které budou použity i v této práci. Za úvahu by stálo použití neuronové sítě i pro výpočet formy mužstva. Síť by mohla vyprofilovat jiné atributy než ty, které budou použity analyticky [20].

4.2 Neuronové sítě

Velké popularitě se sítě těšily v osmdesátých a na počátku devadesátých let. Poté zájem o neuronové sítě poměrně opadl a vrátil se opět v posledních letech. Ze značné části kvůli dostatečné kapacitě výpočetního výkonu.

Neuronové sítě vznikly za účelem kopírování funkce mozku. Místo toho, aby se mozek programoval jako několik tisíc funkcí, které by spolu spolupracovaly, jsou využívány neuronové sítě. Do vstupní vrstvy se vyberou vhodné vstupní parametry. Pro tuto práci se jedná o jednotlivé atributy, které ovlivňují zápasy. Každý z těchto atributů bude mít nastavenou úvodní váhu, ta se může určit podle očekávání, jaká jsou pro trénování sítě nebo se mohou vybrat náhodně. Ke každému neuronu v první skryté vrstvě poté budou připojeny všechny neurony ze vstupní vrstvy ve vztahu N:1, kde N je počet vstupních proměnných. Další vrstvy fungují také ve vztahu N:1, kdy N je počet neuronů v předchozí vrstvě [31].

4.2.1 Model neuronu

Model neuronu ze skutečného neuronu kopíruje používání výstupů ostatních neuronů jako svoje vstupy, nastavuje si váhy pro jednotlivé vstupy, samotné tělo neuronu se chová jako výpočetní jednotka, která sčítá všechny vstupy a neuron má jeden výstup. U komplikovanějších problémů v neuronových sítích využívají skryté vrstvy, které mohou a nemusí být použity. Díky tomu může postupným řešením jednodušších problémů vyřešit ty složitější. Při příkladu rozpoznávání obrazu by se očekávalo, že jednotlivé pixely na vstupech v další vrstvě mohou tvořit složitější obrazce a obraz by prošel přes několik dalších vrstev až do té finální, kde by byl rozpoznán jako celek. Momentu, kdy je neuronu přiřazena nějaká hodnota se říká aktivace, a to jak u vstupní vrstvy, kde je hodnota implicitní, tak u vrstev skrytých. [30]

Proti reálu modelový neuron může také vytvořit více výstupů. Pro bakalářskou práci je to důležité s ohledem na tři výstupy, které chceme pro každý zápas, a to šanci na výhru domácích, remízu a výhru hostů. V opačném případě by bylo nutné neuronovou síť používat pro každý zápas třikrát [31].

Výpočet hodnoty pro každý neuron probíhá pomocí vzorce 4.7, který vyjadřuje kterýkoliv z neuronů z výstupní vrstvy na obrázku 4.1. Pomocí l je vyjádřena vrstva, ve které se nový neuron počítá, j vyjadřuje o kolikátý neuron ve vrstvě l se jedná, k je počet všech neuronů v l-1 vrstvě, w vyjadřuje váhu spojení, b zastupuje aktivaci neuronů a p je použit

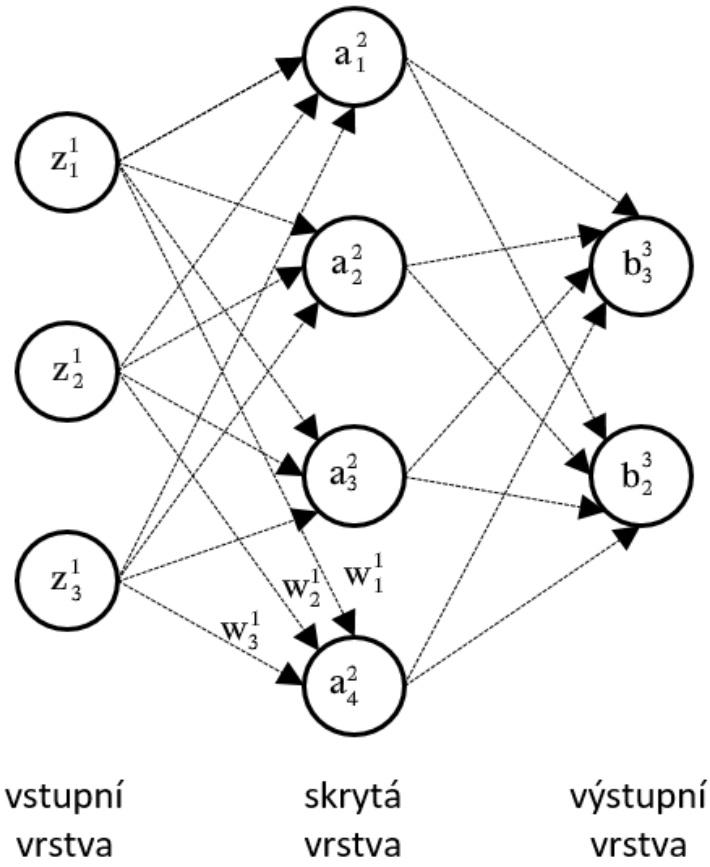
místo hodnoty prahu, který se používá pro udržení hodnoty, která spustí aktivační funkci.

$$b_l^j = \sigma(\sum_k w_{jk}^l b_k^{l-1} + p_j^l) \tag{4.7}$$

Výpočet pro čtvrtý neuron v druhé vrstvě na obrázku 4.1 by byl 4.8:

$$a_2^4 = \sigma(w_1^1 z_1^1 + w_2^1 z_2^1 + w_3^1 z_3^1 + p_4^2) \tag{4.8}$$

kde p na konci značí práh [32, 39].



Obrázek 4.1: Model neuronu se skrytou vrstvou [23]

Jako aktivační funkce je zde využita funkce sigmoid, který má předpis 4.9:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.9}$$

Hlavním důvodem pro její použití je její existence na rozsahu od nuly do jedné, což se skvěle hodí pro předpovídání pravděpodobnosti. Její mírně stoupající průběh také zamezuje skokům ve výsledcích. Její nevýhodou je, že při větších hodnotách na vstupech nedělá tak velké rozdíly ve výstupu.

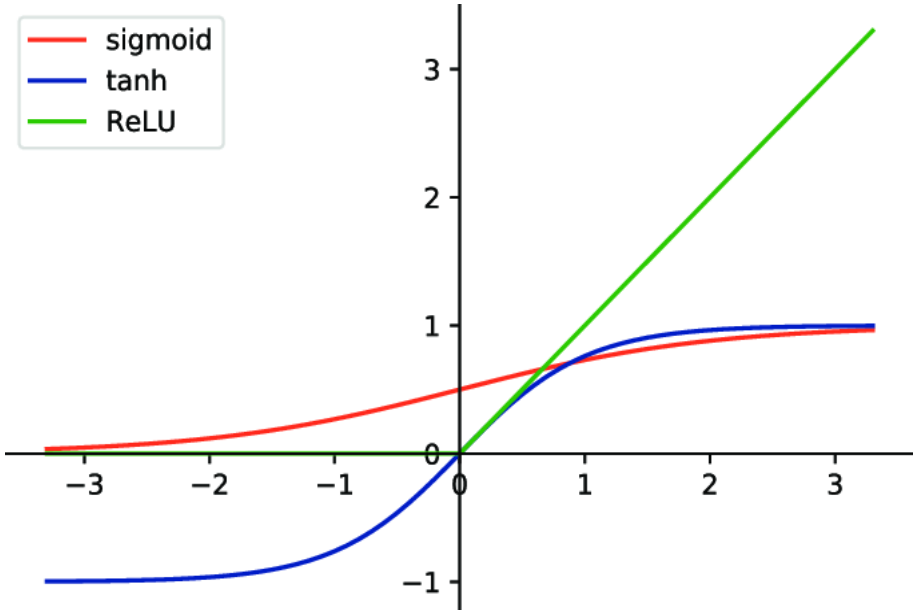
Dalšími dvěma často používanými aktivačními funkcemi jsou ReLU a její různé obměny a TanH. Pro druhou zmíněnou platí předpis 4.10:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4.10}$$

Je lepší variantou pro modelování s hodnotami pohybujícími se blízko okolo nuly. Její nevýhodou je podobně jako u sigmoidu, transformace vyšších hodnot. Pro ReLU platí předpis 4.11:

$$f(x) = max(0, x) \tag{4.11}$$

transformuje negativní hodnoty na hodnotu nula, z čehož vyplývá její nedostatek, kdy v momentě, kdy se vstupy příliš přiblíží nule nebo jsou negativní, tak se nedokáže síť dále učit pomocí zpětné propagace. Její výhodou je, že velice rychle dokáže konvergovat. K zabránění tomuto problému právě vznikají její modifikace, ale pro formát vstupů v této práci jsou negativní hodnoty na vstupech irelevantní, a proto nejsou zmíněny. Na následujícím grafu (obrázek 4.2) lze vidět jejich podobu .



Obrázek 4.2: Model neuronu se skrytou vrstvou [35]

Skrytých vrstev může být využito více, nemusí být pouze jedna. Stejně tak záleží na implementaci pro počet neuronů v každé z těchto vrstev. Podobně je na tom i výstup z neuronové sítě, který nemusí být pouze jeden, ale může jich být několik. V této práci by výstupem mohla být hodnota pravděpodobnosti tří možných výsledků a to výhry, remízy nebo prohry. [31]

4.2.2 Trénování sítě

Bylo již zmíněno, že bakalářská práce se bude zabývat trénováním sítě při učení s učitelem. Na vstupy budou přivedeny atributy zmíněné v první kapitole převedené do hodnot srozumitelných pro neuronovou síť.

Kvalita trénování sítě se určuje pomocí ztrátové funkce. Po zjištění její hodnoty se síť pomocí postupných kroků snaží najít její lokální minimum a tím dosáhnout co nejpřesnějších výsledků. Pro úpravu vah a prahů se bude používat zpětná propagace. Tedy jejich úprava proběhne po zjištění ztráty. Ztrátová funkce se opět dá počítat několika způsoby. Pro menší počet vstupů je využívána log-ztrátová funkce LBFGS. Její výpočet je pomocí vzorce 4.12, kde n je počet trénovacích dat, suma je počítána přes jednotlivé trénovací

případy, y reprezentuje požadované výstupy, a je výstup neuronu. [32]

$$C = -\frac{1}{n}\sum_{x}[ylna + (1-y)ln(1-a)] \tag{4.12}$$

Další z možností pro optimalizaci neuronové sítě je algoritmus Adam, jehož původní účel sice nebyl určený pro vícevrstvé neuronové sítě, nicméně se ukázalo, že na větším počtu dat prokazuje lepší výsledky než právě zde zmíněný LBFGS [25]. Autoři algoritmu ho popisují jako kombinaci výhod dvou dalších rozšíření stochastického klesání gradientu. A to Adaptive Gradient Algorithm, který si udržuje míru učení pro každý parametr, který vylepšuje výkon u problému s rozptýleným gradientem a Root Mean Square Propagation, který si stejně jako předchozí zmíněný uchovává míru učení pro každý parametr, jejíž adaptace probíhá v závislosti na průměru posledních relevancí gradientu pro váhy. Místo přizpůsobení parametru pro míru učení založeném na prvním průběhu Adam také využívá průměr druhého průběhu gradientů. Algoritmus počítá pohybující se exponenciální průměr gradientu a jeho druhou mocninu. Dva parametry poté kontrolují míru klesání těchto pohybujících se průměrů. Inicializační hodnota pro tyto parametry by se měla blížit jedné [19]. Tradičně se pro úpravu vah používá také stochastické snižování gradientu neboli sgd. Jedná se o zjednodušení pro výpočetní náročnost u algoritmu snižování gradientu. Děje se to během vybíraní datových bodů při každém kroku za účelem vypočítání derivátů. Sgd náhodně vybere jeden datový bod z kompletní množiny dat při každé iteraci pro zredukování výpočetní náročnosti [41].

Při zpětné propagaci se upravují váhy a prahy na základě spočítané ceny pro všechny tři neurony ve výstupní vrstvě, která nese označení *l*. Po jejich korekci se přistupuje k vrstvě *l-1*, kde dochází k dalším změnám vah pro neurony v *l-1* vrstvě. Změny vah by se analogicky propisovaly až do *l-n* vrstvy, kde n je počet vrstev v síti. Váhy a prahy se upravují s určitou velikostí kroku, která by neměla být příliš velká, aby lokální minimum nepřekročila, ale zároveň ani příliš malá, aby se do minima dostala co nejdřív. Celý tento proces je výpočetně náročný i pro dnešní počítače, proto se množiny dat rozdělí na náhodně množiny pouze o několika prvcích. Opět je důležité být opatrný při výběru počtu prvků z důvodu přesnosti, ale s ohledem na výpočetní výkon.

4.2.3 Problémy pří trénování

Je nutné zjistit, jaký je optimální počet dat pro trénování sítě. Při příliš krátkém trénování nebude mít síť dostatečnou znalost a při následném testování bude odhadovat špatné výsledky. Podobný problém může ale nastat i při trénování na příliš velkém množství dat. V tomto případě je možné síť přetrénovat do bodu, kdy bude věnovat pozornost statistickému šumu a už nebude podávat validní výsledky. Jak už bylo zmíněno dříve, výkonnost sítě se měří pomocí ztrátové funkce. Jejím sledováním se zjišťuje, kdy síť dosáhla lokálního minima, tedy nejpřesnějších výsledků, které bylo pro síť možné dosáhnout. Přetrénování sítě lze tedy odhalit pomocí zvyšující hodnoty ztrátové funkce.

Předejít podtrénování nebo přetrénování lze několika způsoby. Pravděpodobně nejsnazší je spustit síť několikrát, pokaždé s jiným počtem vzorků a odladit počet dat metodou pokusomyl, což ovšem může stát velké množství času. Nejpoužívanější metodou je metoda brzkého zastavení, během které se sleduje proces trénování. Trénování ovšem není vhodné zastavit v momentě, kdy výkonnost poprvé klesne, protože problémy řešené pomocí neuronových sítí jsou stochastické. Znamená to, že přesnost může několikrát klesat a následně zase stoupat [18].

4.3 Nástroje pro strojové učení

Nejsnazší cestou k naprogramování algoritmu, který se bude sám učit, je využití knihoven, které jsou k pythonu nabízeny. Samozřejmě je možnost při komplexní znalosti problematiky knihovny nepoužívat a nějaký algoritmus pro strojové učení si vytvořit od základu, nicméně by to bylo zbytečné přidělávání práce.

Pro řešení problémů je nejpoužívanějším nástrojem na strojové učení Jupyter Notebook. Mezi jeho hlavní výhody patří rychlé zpracování dat, efektivní platforma a podpora tří jazyků a to Julia, R a Python. Další možností je framework TensorFlow, který je používán pro zpracování velkého množství dat. Poskytuje modely pro neuronové sítě a podporuje jazyk Python. TensorFlow běží zároveň na CPU a GPU, nejčastěji se používá na zpracování přirozeného jazyka a klasifikaci obrazu. Stejně jako TensorFlow i Azure machine learning studio od Microsoftu využívá zároveň CPU i GPU a poskytuje jednoduchý způsob pro propojování modelů a datových množin [40].

Základem pro většinu těchto algoritmů je knihovna NumPy. Tato knihovna sama stačí k vybudování jednoduché neuronové sítě při využití její práce s vícerozměrnými poli, možnosti generovat pseudonáhodné proměnné nebo výpočtu exponenciálu pro sigmoid funkci. Přestože je knihovna navržena pro vědecké účely, sama o sobě není až tak často využívána pro strojové učení. Dobrou alternativu proti Numpy představuje Pytorch, je také založen na jazyku Python a má flexibilní použití. Jedná se o framework pro hluboké učení, je velmi rychlý, protože dobře využívá GPU. Je využíván pro stavění hlubokých neuronových sítí a výpočty tenzoru [33].

Pro strojové učení se používají knihovny scikit a SciPy. Knihovny jsou vzájemně propojené, kdy SciPy využivá algoritmy první jmenované knihovny pro zpracování obrazu a nástroje pro strojové učení. Knihovna scikit je postavená na již dříve zmíněném NumPy, SciPy a matplotlib. Může být využívána na nejpoužívanějších operačních systémech a to Windows, macOS a Linux. Instalace balíčku probíhá pomocí *pip*, který slouží k instalaci balíčků pro python, může to umožňovat samotné IDE, které ale pouze použije *pip* příkazy. Poskytuje funkce a algoritmy pro řešení klasifikačních, regresních a shlukovacích problémů, pro redukování náhodných proměnných pro zvážení, vybírání modelů a předzpracování. Stejně tak zvládne zpracovat data pro učení s učitelem i pro učení bez učitele.

Již dříve zde bylo zmíněno, že problém řešení výsledků zápasů je regresní a že bude řešen pomocí učení s učitelem, protože jsou k dispozici historická data a k řešení bude použita neuronová síť. Pro tuto možnost poskytuje scikit třídu MLPClassifier, která implementuje vícevrstvý model perceptronu, který trénuje síť pomocí zpětné propagace a nabízí více možností pro výpočet ztrátové funkce. Umožňuje nastavit počet skrytých vrstev bez vrstvy vstupní a výstupní a také počet prvků v nich. Třída poskytuje výběr z několika aktivačních funkcí, zahrnující již zde zmíněné funkce sigmoid, TanH a ReLU funkci. Stejně tak poskytuje více možností pro optimalizaci vah. Atributy této třídy zahrnují výpočetní funkci ke kontrole úspěšnosti sítě a možnost nastavit si počet výstupů [34].

Kapitola 5

Návrh

V této kapitole bude popsán návrh práce pro implementaci na základě teoretických znalostí z předchozí kapitoly. V první sekci bude vysvětleno proč a jaké atributy 5.1 byly vybrány do neuronové sítě, dále bude nastíněno stahování a ukládání dat do lokální databáze 5.2, poté vysvětlen návrh modelu, který byl vybrán pro určování výsledků zápasů 5.2.4 a nakonec popsaná podoba finální aplikace 5.2.5.

5.1 Atributy ovlivňující zápasy

Následující atributy v tabulce 5.1 jsou kombinací mého vlastního úsudku a souhrnu průzkumu z uvedeného zdroje [2]. Jsou vybrány do první iterace výpočtů a budou se na základě výsledků měnit. Atributy, které berou v potaz bookmakeři sázkových kanceláří nejsou z pochopitelných důvodů nikde zveřejněny.

Atribut	**Priorita**	**Časový horizont**
Forma	Vysoká	Posledních pět zápasů
Domácí hřiště	Vysoká	Předchozí dvě sezóny včetně té současné.
Pozice v tabulce	Nadprůměrná	Současná sezóna
Sestava	Průměrná	Ze zápasů z aktuální sezóny se bude kalkulovat, jak se s hráči, kteří mají nastoupit do utkání, týmu dařilo.
Přestávky v sezóně	Průměrná	Od posledního hraného utkání
Důležitost utkání	Průměrná	Jedná se o vývoj současné sezóny.
Změna trenéra	Průměrná	Pokud v předchozím zápase tým vedl někdo jiný, bere se tento aspekt v potaz.
Únava	Nízká	Kdy se hrál poslední zápas.

Tabulka 5.1: Atributy které mají vliv na průběh utkání

5.1.1 Forma

Největší vliv na výsledek zápasu má současná forma. To, jak tým hrál několik předchozích utkání má největší podíl na stanovení kurzu. U tohoto ukazatele je důležité brát zřetel i na to, s kým tým hrál, protože je velký rozdíl porazit v zápase poslední nebo první tým

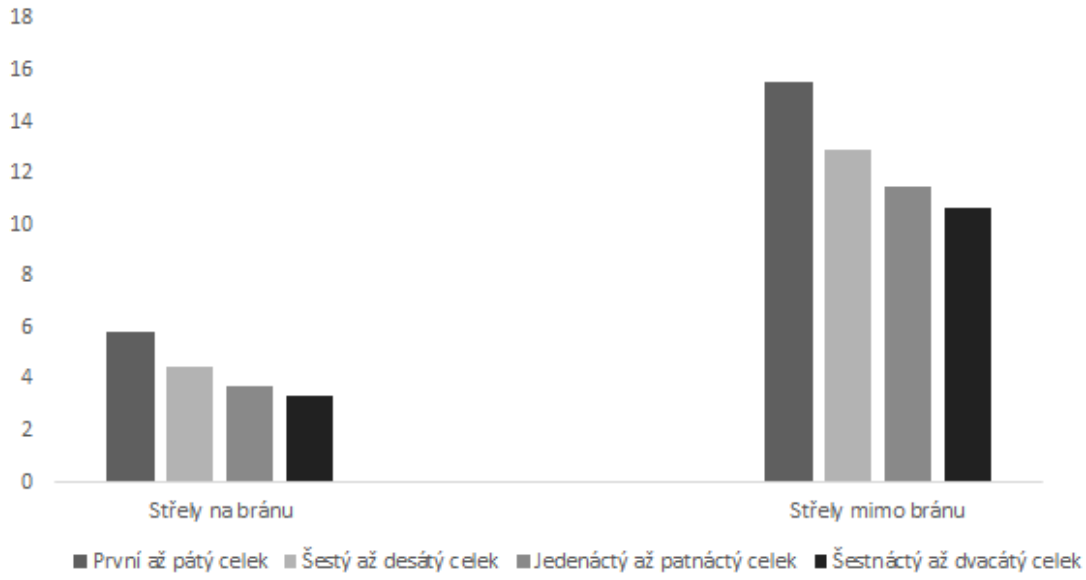
tabulky. Proto výsledky z předchozích zápasů budou násobeny sílou soupeře. Soupeři jsou rozděleni podle jejich kvality. V lize podle dosaženého množství bodů ze všech získatelných v ostatních soutěžích, ze kterých se sbírá forma, podle žebříčku UEFA [12] pro kvalitnější celky v evropských soutěžích, případně podle kvality soutěže o kterou se jedná.

Do současné formy však nespadá pouze výsledek, ale i to, jak se týmu v posledních jednotlivých zápasech dařilo herně. Pod tento atribut spadá počet střel, a to jak na bránu, tak mimo. Podle článku [1] se však ukázal velice důležitým faktorem úspěšných týmu poměr střel k počtu vstřelených gólů. Data nasbíraná z posledních čtyř sezón v nejvyšší anglické soutěži ukazují velmi nízké rozdíly, co se týká střel na bránu v poměru ke vstřeleným gólům. Zatímco u celkových střel už jsou rozdíly větší, ve středu tabulky se liší minimálně (tabulka 5.2) [8].

	1 - 5 tým	**6 - 10 tým**	**11 - 15 tým**	**16 - 20 tým**
Celkový počet střel na gól	7,94	9,37	9,75	11,91
Počet střel na bránu na gól	2,85	3,2	3,13	3,73

Tabulka 5.2: Statistika kolik je potřeba střel na gól ze sezón 15/16 - 18/19

U střel samotných, jak při celkovém počtu, tak na bránu, jde opět pozorovat rozdíly mezi nejlepšími celky, středem tabulky a nejslabšími týmy. Jsou vidět na grafu 5.1.



Obrázek 5.1: Statistika střel ze sezón 15/16 - 18/19 [8]

Z těchto hodnot je možné spočítat, kolik střel mají úspěšnější celky proti těm neúspěšným. Díky tomu lze určit formu, která bude přiřazena na základě těchto hodnot. Je důležité započítat i střely, které byly proti celku, pro který se zrovna počítá forma. Jejich hodnota bude určena stejným způsobem, jako střely vystřelené týmem, pro který se počítá forma.

Podobným způsobem jsou určeny i vstřelené a obdržené góly s vypočítaným průměrem za uplynulé sezóny pro určené zóny v tabulce. Zprůměrované statistiky jsou uvedeny v tabulce 5.3 [8].

	1 - 5 tým	6 - 10 tým	11 - 15 tým	16 - 20 tým
Góly vstřelené za zápas	1,99	1,4	1,19	0,92
Góly obdržené za zápas	0,93	1,32	1,55	1,70

Tabulka 5.3: Statistika průměru vstřelených a obdržených gólů za sezóny 15/16 - 18/19

Ještě komplikovanější je to s držením míče, přestože se dá předpokládat, že týmy s větším držením míče jsou celkově úspěšnější. Na statistiku samotnou má dost často velký vliv momentální průběh zápasu. Mějme například zápas třetího týmu tabulky, který hraje doma proti patnáctému týmu tabulky. V momentě, kdy je stav vyrovnaný dvě třetiny zápasu, silnější tým na míči bude po tuto dobu dominovat za účelem vstřelení branky. V momentě, kdy se jim to však podaří a budou vést pouze o gól, tak přibližně v závěrečné desetiminutovce, ať předpoklady před zápasem hovoří jakkoliv, bude na míči dominovat slabší soupeř, a tak se statistika může značně vyrovnat. Proto je u výpočtu formy nutné s tímto údajem pracovat obezřetněji. Nicméně vetší držení míče znamená lepší výkon.

Naopak naběhaná vzdálenost hráčů nebyla prokázána jako faktor ovlivňující úspěšnost týmu. Je to dáno tím, že slabší týmy proti silnějšímu soupeři hrají většinou méně s míčem a tím pádem toho více naběhají. Stejně tak není prokázáno, že by počet odebrání míče soupeři měl dopad na celkový výkon týmu.

Na základě údajů z předchozích sezón můžeme sestavit vzorec 5.1, 5.2, který spočítá formu pro daný celek. S relevancí založenou na chronologii zápasů, přičemž poslední odehraný zápas bude mít největší relevanci. Naopak první z pěti zápasů, které jsou počítány jako důležité pro výpočet formy, bude mít relevanci nejmenší.

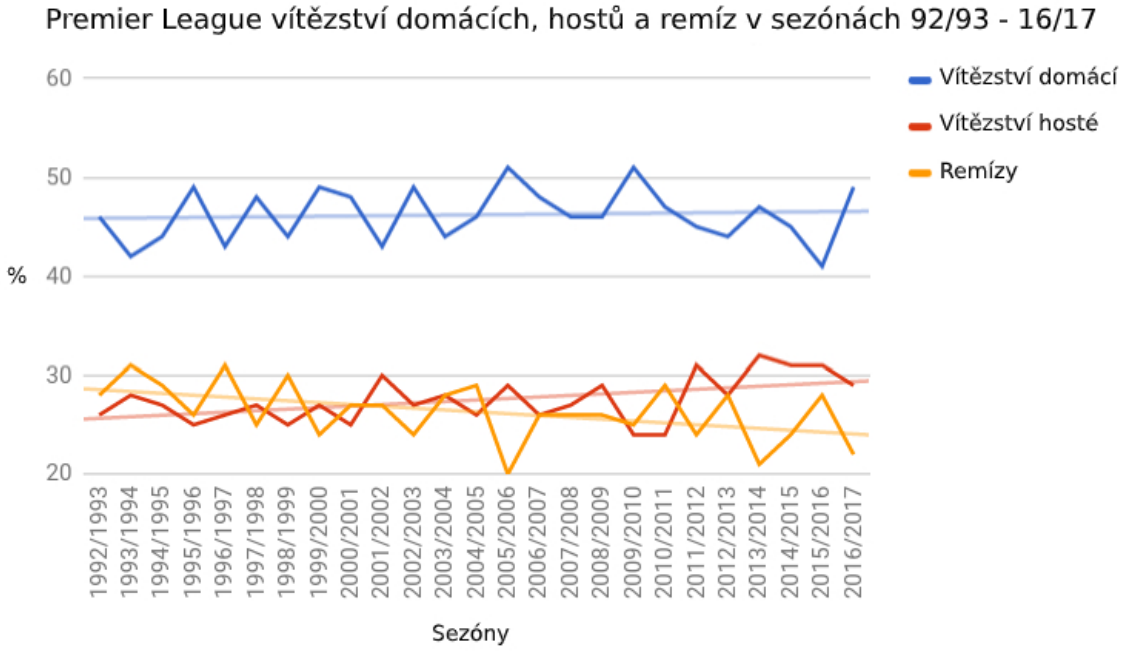
$$f = \sum_{i=1}^{5} c * (v * k) + z \tag{5.1}$$

$$z = d + s_t + s_p + s_g \tag{5.2}$$

Kde f je výsledná forma týmu před zápasem, c značí relevanci zápasu ve smyslu, kdy byl odehrán, v zastupuje výsledek zápasu, který počítá i s gólovým rozdílem, k je kvalita soupeře a z jsou statistiky posbírané z dění na hřišti. V druhém vzorci d značí držení míče, s s indexem t jsou střely, jak na bránu tak mimo, týmu v daném zápase naopak s indexem p se jedná o střely, které byly zaznamenány proti týmu, kterému se počítá forma a s s indexem g je statistika, kolik střel tým potřeboval ke vstřelení gólu. Opět se počítají jak střely celkové, tak střely na bránu.

5.1.2 Domácí hřiště

Zajímavým aspektem je také výhoda domácího hřiště. Domácí týmy vyhrávají přes 46 procent utkání, remízy nastávají v přibližně 28 procentech případů a výhry hostů ve zbývajících 26 procentech zápasů (obrázek 5.2). Tyto údaje jsou zdůvodňovány především domácími fanoušky a známým prostředím. První polovina atributů pro neuronovou síť budou atributy domácího týmu, proto je očekáváno, že tyto atributy budou mít po vytrénování sítě větší váhu.



Obrázek 5.2: Poměr vítězství [3]

5.1.3 Pozice v tabulce

Velice zajímavý atribut je postavení týmu v tabulce. Příkladem by mohl být tým, který se v tabulce nachází na prvním místě. Z posledních pěti zápasů má dvě výhry, dvě prohry a dvě remízy a hraje proti týmu, který je na dvanácté pozici. Tento tým se tam však dostal z posledního místa po pěti výhrách v řadě. Pouze na základě posledních odehraných zápasů by bylo logické vsadit na tým na dvanácté pozici, přestože by favoritem rozhodně nebyl. V případě, že by se zápas odehrával na hřišti hůře postaveného týmu, je utkání velice rizikové a jedním z možných řešení je na zápas raději vůbec nesázet. Druhý pohled je však ten, že kurz na první tým v tabulce, který je zpravidla nízký, by mohl značně narůst. V tomto případě jsou klíčové všechny zbylé faktory, které by napomohly k rozhodnutí.

Není však správné hodnotit týmy pouze na základě jejich celkového pořadí v tabulce. Některé týmy mezi sebou mají pouze velice malý bodový rozdíl, a proto není klíčová pozice v tabulce ale především bodový rozestup.

5.1.4 Sestava a důležitost utkání

V Anglické lize je pravidlem, že hodinu před začátkem fotbalového utkání je známá sestava pro daný zápas. Většinou se značná část sestavy dá odhadnout ještě dřív, z předchozích zápasů. Ke konci sezóny, kdy některé týmy už nemají v lize o co hrát, a naopak mohou ještě v jiných pohárových soutěžích dosáhnout větších úspěchů, mohou nasazovat slabší či juniorské hráče. Proti tomuto však má nejvyšší anglická liga pravidlo, které říká, že týmy do zápasu musí nastupovat v plné síle, případně budou čelit finanční pokutě [37]. Ovšem i když sestava musí být nejsilnější, tak se většinou jedná o subjektivní pohled a stále může být pár hráčů ze sestavy obměněno. Potíže u ligové komise nastávají až při očividném vypouštění zápasů. Síla sestavy bude určená na základě předchozích výkonů, takže do určité míry

bude kopírovat formu týmů. Bude zde však záležet na tom, jak často je hráč nasazován do zápasů, do kterých mohl zasáhnout. Vyloučena budou tedy utkání, během kterých byl zraněný. Bude také přihlíženo na to, do kolika gólů vstřelených vlastním týmem se zapojil. Zde bude nahlíženo na poměr vstřelených gólů týmem proti těm, ve kterých figuroval sám hráč. Naopak karetní tresty ovlivňují hráčovy statistiky negativně. Síla sestavy poté bude vyjádřena jako součet jednotlivých hráčů, kteří do utkání zasáhnou.

Důležitost utkání se tak dá částečně vyčíst ze sestavy nasazené do zápasu. Pokud tým například nasadí tři a více nových hráče ke konci sezóny, je jasné, že spíše zkouší nové věci, případně chce, aby si odpočinuly opory týmu. Důležitost utkání se zvedá pro týmy v určitých zónách tabulky s tím, jak sezóna pokračuje, a to v sestupové zóně, zóně bojující o Evropskou ligu, zóně bojujících o Ligu mistrů a zóně bojující o titul.

Na obrázku 5.3 jsou dané zóny ukázány z probíhajícího ročníku. Zóna o titul už je tento rok irelevantní, protože i když zbývá ještě deset zápasů do konce sezóny, bodový náskok (B) lídra je příliš velký na to, aby na titul mohl pomýšlet někdo další. Mnohem zajímavější je zóna boje o Ligu mistrů, kde je rozdíl od čtvrtého do sedmého místa pět bodů. To je deset utkání před koncem soutěže stále málo, a tak o postup hraje ještě spoustu celků. Podobně je to i u hranice pro sestup, kdy sestupují všechny tři celky, i zde je rozdíl mezi patnáctým a devatenáctým místem malý, a tak se dá očekávat, že tyto týmy budou s přibývajícím počtem kol podávat výkony nad očekávání. Tento faktor tedy bude zvažován v poměru bodů od důležitých umístění v tabulce. Kolo, kdy tento faktor začne být relevantní, bylo stanoveno na dvacáté osmé, tedy deset kol před koncem sezóny, kdy už je tabulka jasně vyprofilovaná a začíná se hrát o vše.

5.1.5 Přestávky v sezóně

Přestávek je několik a liší se soutěž od soutěže. V nejvyšší anglické lize se jedná především o reprezentační přestávky na dva týdny. Během nich se můžou zranit klíčoví hráči. Naopak někteří si odpočinou a může dojít k mentálnímu restartu mužstva po nevydařeném období, takže k zápasům po pauzách je třeba přistupovat obezřetněji. V Anglii na rozdíl od jiných soutěží není zimní pauza. Koncem prosince je velký sled zápasů, které po sobě následují někdy pouze s dvou denní přestávkou, a tak je třeba brát ohled na rotace sestav. Největší pauzou je letní, která nastává mezi dvěma sezónami, a proto je lepší se několika úvodním zápasům pro sázení vyvarovat, protože výsledky z přátelských utkání nemohou být brány jako relevantní. Proto pokud před zápasem nastala pauza delší než deset dní, tak se počet relevantních zápasů snižuje na tři. Pokud nastala pauza delší než padesát dní, což značí hlavně konce sezón nebo také globální pandemii, která může přerušit průběh ligy, tak se začíná od znovu a každý z týmů musí odehrát alespoň tři zápasy, aby byla data pro výpočet relevantní.

5.1.6 Změna trenéra

Zasáhnout do vývoje ročníku může událost, jako je změna trenéra. Dost často se mužstvo minimálně na několik prvních zápasů pod vedením nového trenéra psychicky i výsledkově zvedne. Skvělou ukázkou může být Manchester United v ročníku 2018/2019, kdy po nepovedeném úvodu byl nahrazen trenér Jose Mourinho. V jeho posledních deseti zápasech v Premier League získal pro tým šestnáct bodů, a to za čtyři výhry, čtyři remízy a dvě prohry. Po jeho odchodu a příchodu nového manažera 19. 12. 2018 [7] nastala neočekávaná změna, kdy jeho stále ještě málo zkušený nástupce a bývalý hráč týmu Ole Gunnar Solskjaer získal z úvodních deseti zápasů dvacet šest bodů za osm výher a dvě remízy. Výměna nastala v

#	Tým	Z	V	R	P	G	B
1.	Liverpool	29	27	1	1	66:21	82
2.	Manchester City	28	18	3	7	68:31	57
3.	Leicester	29	16	5	8	58:28	53
4.	Chelsea	29	14	6	9	51:39	48
5.	Manchester Utd	29	12	9	8	44:30	45
6.	Wolves	29	10	13	6	41:34	43
7.	Sheffield Utd	28	11	10	7	30:25	43
8.	Tottenham	29	11	8	10	47:40	41
9.	Arsenal	28	9	13	6	40:36	40
10.	Burnley	29	11	6	12	34:40	39
11.	Crystal Palace	29	10	9	10	26:32	39
12.	Everton	29	10	7	12	37:46	37
13.	Newcastle	29	9	8	12	25:41	35
14.	Southampton	29	10	4	15	35:52	34
15.	Brighton	29	6	11	12	32:40	29
16.	West Ham	29	7	6	16	35:50	27
17.	Watford	29	6	9	14	27:44	27
18.	Bournemouth	29	7	6	16	29:47	27
19.	Aston Villa	28	7	4	17	34:56	25
20.	Norwich	29	5	6	18	25:52	21

Postup - Liga mistrů (Skupinová fáze)
Postup - Evropská liga (Skupinová fáze)
Sestup - Championship

Obrázek 5.3: Vývoj současného ročníku k 01. 05. 2020 se zvýrazněním důležitých pozic [11]

polovině sezóny, takže se neudála žádná důležitá změna v kádru. Jakmile však manažer dostal 28. 3. 2019 [7] trvalý kontrakt, tak tým upadl do své staré formy. A ve zbývajících sedmi zápasech tým nasbíral pouze osm bodů za dvě výhry, dvě remízy a tři porážky [10]. Na příkladu jde skvěle vidět, jak velký psychický rozdíl může pro tým znamenat výměna trenéra a jaký celkový vliv má psychika na vývoj zápasů.

5.1.7 Minoritní faktory

Je důležité také zvážit, zda má na zápas vliv něco jako zhoršené počasí a zda je nutné tento atribut brát v potaz. Přilepšoval by samozřejmě horšímu celku, protože na absolutně podmáčeném, či zasněženém trávníku se mnohem hůř hraje tvořivý fotbal za účelem vstřelení branky. Zrovna v Anglii, kde je počasí značně proměnlivé a hraje se celou zimu i během vánočních svátků, je vhodné tento faktor zvážit. Nicméně za sezóny 13/14 - 15/16, údaje prozradily, že počet branek v poměru letních ku zimním měsícům se liší pouze o sedm desetin, proto tedy počasí není bráno jako faktor ovlivňující zápasy.

Zvlášť opatrně je třeba přistupovat k vyhrocenějším zápasům, jako jsou například derby. Jedná se o zápasy odvěkých rivalů, zpravidla ze stejného města nebo ze sousedících měst. Ty mohou končit divokými a nečekanými výsledky. Zajímavým atributem může být únava z předchozích utkání. Celky v polovině tabulky nehrají evropské soutěže a nejsou tím pádem tak vytížené, nicméně na nejvyšší úrovni, na které se anglická liga hraje, pokud hráč nebude naplno připravený na zápas, tak k němu ani nenastoupí, proto to není faktor, který bude zahrnut.

Zamyslet se dá například i nad výkonností brankáře daného týmu. V hokeji jsou procenta chycených zákroků důležitá a dost často se na výkonu brankáře odráží výkon celého týmu. Ve fotbale, především v tom špičkovém, je to jinak. Rozdíly v brankářích na nejvyšší úrovni jsou absolutně minimální. Střel na bránu je navíc v zápasech mnohem méně. Průměr je okolo devíti střel za zápas, a to při průměru tří gólů při utkání. Vzorek střel tím pádem není dostatečný na porovnání. Navíc je velmi těžké odhadnout náročnost střel a záleží také na obraně, která před brankářem stojí.

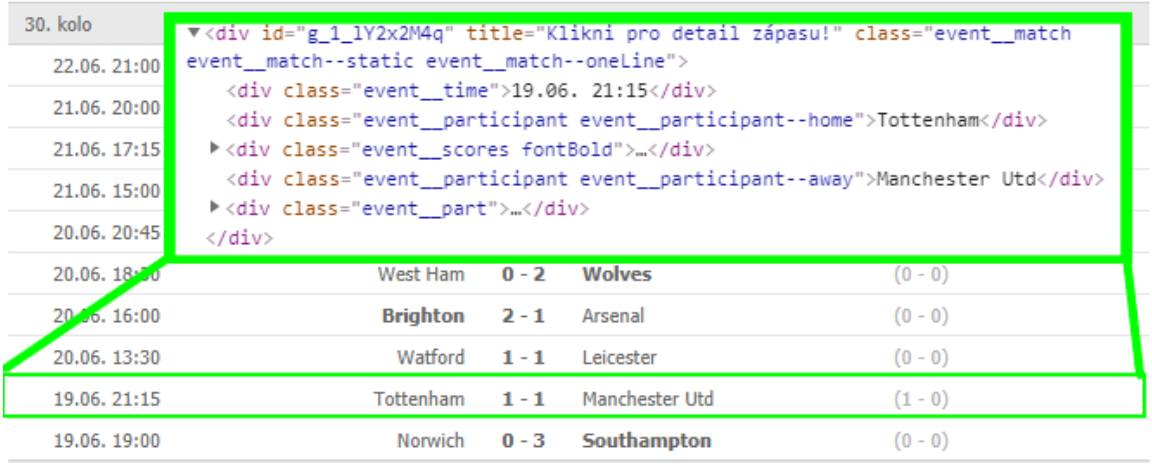
Často se také zmiňují historické statistky ze vzájemných zápasů obou týmů. Tento atribut však neberu jako relevantní, protože fakt, že tým na nějakém stadionu nevyhrál několik let, nemá moc společného s týmem, který hraje ten den. Každý rok se v každém týmu mění několik hráčů, podobně je to i s trenéry, kteří jsou měněni i několikrát během sezóny, a proto si nemyslím, že je průběh zápasu, který se odehrál půl roku nebo několik let zpět relevantní pro ten současný [8].

5.2 Trénovací množina dat

Potřebná data budou vybíraná podle atributů zmíněných v předchozí sekci 5.1. Z každého zápasu tedy bude vybraná množina dat, která je určena pro trénování sítě nebo pro pozdější možná rozšíření těchto atributů, či jejich úpravě.

5.2.1 Získávání dat ze zápasů

Pro zdroj informací byl vybrán web Livesport. Jedná se o českou společnost a pro účely práce bylo umožněno data exportovat. Pro zápasy všech lig mají webové stránky Livesportu standartizovaný zápis (obrázek 5.4), který lze vidět na obrázku. Z těchto informací se následně dá získat informace o datu zápasu a názvy týmů účastnících se zápasu. Při rozkliknutí zápasu se získávají další informace. Zápisy se neliší ani napříč soutěžemi, proto je snadné skrz zápasy iterovat a data si ze stránky stáhnout. Procházet se bude pěti soutěžemi, kterých se týmy účastní. Procházet se nimi bude chronologicky, tedy od prvního zápasu sezóny po poslední vždy v rámci soutěže. Následně se stáhnou další ročníky soutěže k současnému datu.



Obrázek 5.4: Pohled na výpis zápasů 30. kola Premier League z ročníku 19/20 [10]

Iterovat se bude zápas po zápasu, od začátku soutěže až po její konec, každý zápas se otevře a začnou se z něj extrahovat data s tím, že nejdřív bude nutné stáhnout soupisky a hráče, protože se později budou objevovat v událostech ze zápasu například gólech, žlutých kartách a podobně. Poté se ze zápasu vyjmou statistiky, které jsou zajímavé pro predikci výsledků, a nakonec se vyberou události ze zápasu, ke kterým se přiřadí už dříve stažení hráči. Tím se informace pro zápas uzavřou a přejde se na další. Tímto způsobem se bude iterovat skrz celou sezónu, po jejím dokončení se stránka načte s další sezónou a celý proces se opakuje. Stejný postup se bude aplikovat pro všechny soutěže. Jedinou výjimkou bude EFL Cup dříve nazývaný Carling Cup, ke kterému nejsou občas k dispozici sestavy a statistiky ze zápasu, nicméně se jedná o soutěž, které je přikládaná nejmenší váha ze strany klubů i veřejnosti a spíše slouží týmům k vyzkoušení mladíků, proto detailnější informace o těchto zápasech nejsou nutné.

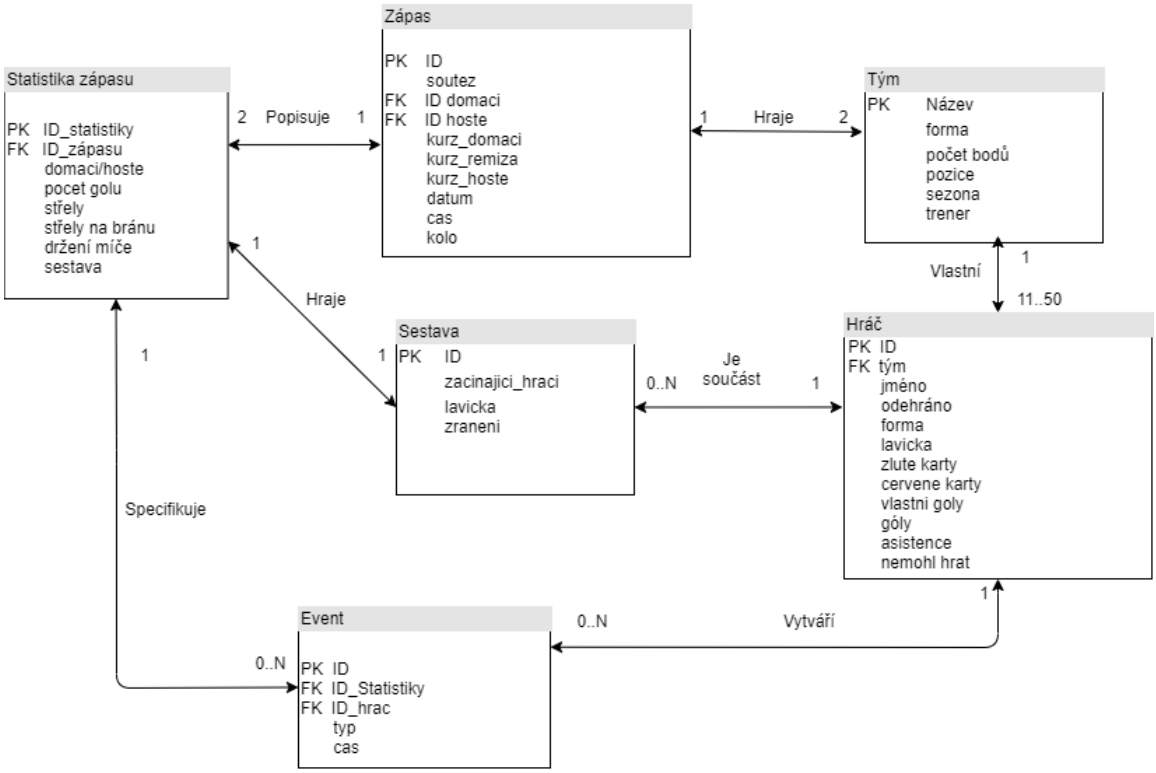
5.2.2 Získávání kurzů

Pro trénovací a testovací množiny dat byly kurzy získané ze stránek Livesportu pomocí web scrapingu zároveň s ostatními údaji o zápasu z detailu utkání, kde jsou pro každý zápas vypsané kurzy od sázkové kanceláře Tipsport. Ze sázkových kanceláří vybírám pouze ty české, kde je díky jednoduchému ověření plnoletosti snazší založit si účet. Pro zahraniční sázkové kanceláře zase mluví jednoduchost v získávání kurzu pomocí různých nabízených API, které jsou ale většinou placené při použití ve větším počtu. Proto i kurzy na nadcházející zápasy budou získávané pomocí dolování dat z webu Livesportu. Použity budou kurzy z největších společností u nás Tipsport, Sazkabet a Fortuna. Kancelář Chance není brána v potaz, jelikož se jedná o dceřinou společnost Tipsportu a jejich kurzy se neliší.

5.2.3 Ukládání dat

Pro data bude vytvořena databáze, jejíž model popisuje obrázek 5.5.

Je potřeba vytvořit tabulku pro týmy, které se budou účastnit jednotlivých zápasů. Za tyto týmy poté nastupují hráči, kteří jsou součástí sestavy pro zápas. Pro každý zápas jsou vytvořeny dvě podrobné statistiky, jedna pro domácí tým a jedna pro hosty. Na tyto statistiky jsou následně navázány sestavy účastnící se zápasu a také události, které se při



Obrázek 5.5: Diagram pro databázi zápasů

zápasu odehrály, jako góly, karty a asistence. Pro tyto události je také uvedeno, jaký hráč se jich účastnil. Většina sloupců v databázi vychází z dříve zmíněných atributů. Jiné sloupce v tabulkách jsou naopak pro možnosti dalších rozšíření, například u tabulky *Event* může *čas* sloužit pro sázky na události v prvním nebo druhém poločasu. V tabulce *Zápas* jsou také zapsány kurzy, které jsou určeny pro výpočet ziskovosti ze sázek na zápasy, protože samotná úspěšnost tipů nezaručí zisk z dlouhodobého hlediska. Pokud by vycházely sázky pouze na malé kurzy je těžší se dostat do kladných čísel.

5.2.4 Návrh modelu

Jako model bude použitý vícevrstvý perceptron, který se učí funkcí 5.3

$$f(\cdot) : R^m \rightarrow R^o \tag{5.3}$$

Kde m je počet dimenzí na vstupu a o je počet výstupů. Pro jeho implementaci bude využita třída z knihovny scikit-learn, která se nazývá *MLPClassifier*, ta se učí pomocí zpětné propagace vysvětlené v kapitole 4.2.2. Jeho učení se zakládá na historických datech, řeší tedy úlohy pro učení s učitelem [34].

První možnost byla využít místo klasifikačního řešení možnosti regrese s tím, že by model navracel pravděpodobnosti pro možné výsledky. Problém by byl ovšem s tím, že pro neuronovou síť by byly jako ukazatele správnosti použity přepočty kurzu sázkových kanceláří na pravděpodobnost. Ve výsledku by tedy síť konvergovala ke správnému řešení, ale bylo by to řešení, které poskytují sázkové kanceláře a bylo by tedy v podstatě nemožné dostat se k zisku.

Proto byla vybraná možnost klasifikace, kde se síť jasně vyjádří k tomu, na kterou příležitost vsadit. Případně jaký zápas vynechat v případě, že se nerozhodne ani pro jednu z možností. Na vstup je přivedeno dohromady jedenáct atributů, pro oba týmy zahrnující formu, sílu základní sestavy, důležitost utkání, a body na které tým dosáhl před zápasem. Nakonec se k těmto faktorům ještě přidaly kurzy na výsledek zápasu, které vypisují sázkové kanceláře. Model má tři výstupy, kdy každý z nich pokrývá jednu variantu pro výsledek zápasu, tedy vítězství domácích, remízu a vítězství hostů.

Po testech na historických datech a několika vyzkoušených iteracích byla z aktivačních funkcí vybránu funkce *relu*, kdy její výsledky byly lepší než výsledky funkcí *sigmoid* a *tanh*. Síť má tři skryté vrstvy, kdy první dvě obsahují po šesti neuronech a třetí skrytá vrstva obsahuje osm neuronů. Jiné možnosti byly zkoušeny na několika iteracích, kdy se měnil počet vrstev i počet neuronů pro každou vrstvu, kdy počet neuronů se pohyboval od jednoho do dvaceti na vrstvu a počet vrstev od žádné do pěti. Finální podoba sítě byla vybraná jako nejúspěšnější z iterací.

5.2.5 Aplikační rozhraní

Rozhraní aplikace bude konzolové, kdy po spuštění skriptu se zobrazí informace o aktualizaci dat, která začne probíhat, pokud nebyly atributy při spuštění nastaveny jinak. Tuto aktualizaci tedy bude možné vypnout pomocí parametrů, a to pro některé soutěže nebo pro všechny. Pokud tento parametr nebude vybrán, spustí se automaticky kontrola. Jestli se v nějaké z relevantních soutěží neodehrál zápas. Pokud se tak stane, zobrazí se uživateli upozornění.

Následně se zobrazí zápasy aktuálního dne. Pro zápasy bude možné kvůli soupiskám dělat predikce pouze hodinu před zápasem, dříve nejsou zveřejněny. Proto po spuštění skriptu bude následovat výpis se zápasy, u kterých bude vypsaná predikce s kurzy pro zápasy, pro které už je to možné a dále upozornění pro zápasy jež ještě nejdou predikovat s informací, kdy se pro ně možnost nejspíše odemkne. Výstup bude jednoduchý výpis kola, času, data a týmů s nabídkou kurzů a následným výběrem pro sázku, který bude barevně označen, kdy modrá pro všechny kurzy bude znamenat, že není výhodné podat sázku na daný zápas. Zeleně vypsaný kurz bude označovat zápas na který vsadit a červené kurzy potom to na co nevsázet. Tyto informace budou dostupné také v nápovědě samotného programu.

Kapitola 6

Implementace

V kapitole bude popsáno, jakým způsobem byla získána a uložena data pro neuronovou síť. Následovat bude vysvětlení způsobu jejich modifikace, a nakonec implementace samotné neuronové sítě.

6.1 Využité technologie

Všechny části byly programované v jazyce **Python** ve verzi 3.7.3. Tento jazyk byl zvolen pro svoji jednoduchost a rozsáhlé možnosti knihoven, které se hodily pro využití na neuronové sítě a zpracování dat. K uložení dat byl vybrán databázový systém **PostgreSQL**, který umožňuje snadné použití v rozhraní **PyCharm**, které bylo použito jako IDE pro tuto práci. **PostgreSQL** je otevřený a objektově-relačně orientovaný databázový systém. Pro zálohování a verzování práce byl využit **GitHub**.

6.2 Získávání dat

Pro získání aktuálního času je využívána knihovna *datetime*. Ta pracuje se systémovým časem a vytvoří proměnnou typu *date*. Podle ní se následně orientuje v tom, jaké zápasy se v daný den mají odehrát a jaké se do té doby odehrály. Podle této proměnné se následně stahují potřebná data. Data jsou extrahována ze stránek Livesport.cz pomocí knihovny *Selenium*. Přestože je tato knihovna těžkopádnější než ostatní, které je možné pro stahování dat z webů využít, tak je potřeba ji použít na načtení kompletní stránky. Načítání výpisu všech zápasů se provádí dynamicky přes tlačítko *Zobrazit další zápasy*, knihovna umožní prvek lokalizovat pomocí vyhledání konkrétního textu a následně na něj klikat do té doby, než budou načteny všechny zápasy. Následně je potřeba knihovnu využít i při iteraci přes každý zápas, kdy se odkaz sice dá odvodit z DOM stránky, kde je množné informaci o id každého zápasu použít pro vytvoření odkazu pro konkrétní zápas, ale při přístupu na stránku se nenačtou všechny hodnoty potřebné pro extrakci. Stránka obsahuje skripty v jazyce JavaScript, které je nutné interpretovat. Je tedy opět potřeba použít knihovnu *Selenium*, která se obvykle spíše používá při testování, protože kopíruje přístup na stránku klasického uživatele.

Při používání knihovny *Selenium* se nejdříve vytvoří objekt *webdriver*, který pro vytvoření potřebuje znát umístění spustitelného souboru prohlížeče. Dále má volitelné parametry jako nastavení pří spuštění, například možnost spustit prohlížeč v anonymním okně anebo spuštění prohlížeče uživateli skrýt. Po vytvoření instance se volají její metody, metoda *get*

spustí v prohlížeči stránku, která je předána jako parametr a předá jí řízení. Následně se ze stránky dají extrahovat data pomocí selektorů. Za tímto účelem je využit HTML kód, kdy se prvek dá lokalizovat pomocí absolutní cesty, relativní cesty nebo třeba text, který je v daném prvku uložen. Na jednotlivé zápasy se tedy nepřistupuje pomocí jejich odkazu, ale opět klikáním na ně. Tím se otevře další okno a pomocí metody na instanci *driver.window_handles* se předá řízení právě poslední stránce. Následně se na stránce najde nejdříve tlačítko s nápisem Sestavy, který je hledán pomocí funkce *find_element_by_link_text*. U sestav začíná dolování dat, kvůli tomu že hráči pak mohou být používáni v dalších tabulkách a ti kteří dosud nebyli zaneseni do databáze, musí být nejdříve vytvořeni. Pro získání textu z webových elementů je použita funkce *text.strip()*, které navrací hodnotu typu *string* z daného webového prvku, na který byla tato funkce zavolána. Tímto způsobem se prohledají následně statistiky zápasů a poté i události (góly, střídání apod.), které se v zápasech staly. Takto se prochází všechny ročníky a všechny soutěže. Odkazy pro soutěže a jednotlivé ročníky se vytvářejí automaticky, protože jediné, co se v nich mění, je samotný ročník. Stahování jednoho zápasu zabere přibližně devatenáct sekund, aby se zamezilo pádům z důvodu pomalého načítání stránky. Aktualizace všech dat s prázdnou databází tedy zabere několik hodin.

6.3 Uchování dat

Data jsou uchovávána v lokální *PostgreSQL* databázi. Pro připojení k databázi je využit adaptér *psycopg 2.8.5*, který je implementovaný v jazyce C [24]. Pro připojení využívá parametry z inicializačního souboru, ve kterém jsou zapsány informace o hostu, názvu databáze a přihlašovací údaje uživatele. Pomocí funkcí knihovny *psycopg2* se inicializuje spojení a po vytvoření kurzoru se přes něj posílají příkazy o vložení, výběru nebo editaci dat do databáze. Na konci transakce se změny musí potvrdit pomocí příkazu *commit* a spojení uzavřít.

Následné plnění tabulek je rozdělené do tří částí. Nejdříve se naplní tabulka týmů, protože na *id* jednotlivých týmů jsou navázány ostatní informace. Poté se plní informace získávané přímo ze zápasů, jak je popsáno v předchozí podkapitole 6.2. Nakonec, když jsou všechna historická data uložená v databázi, se nimi prochází a postupně se doplňují ostatní informace, jak je popsáno v následující kapitole 6.4. Jedná se o data, která musí být aktuální k daným zápasům. Z webu nejsou dostupné konkrétní statistiky před každým zápasem několik let zpět, proto se tyto sloupce v tabulkách doplňují později.

Každý záznam jako primární klíč používá automaticky generované a inkrementující se ID. Záznamy jsou většinou typu *smallint*, který dostačuje pro sezónní statistiky. Výjimky tvoří sloupce pro *id* a také některé informace jako kurzy, které musí mít formát *real*. Pro názvy týmů a záznamy formy z několika zápasů je použit datový typ *text*. V tabulce se nachází několik sloupců, které v momentální podobě skriptu nejsou používány, ale jsou určeny pro možná budoucí rozšíření, například časy u událostí ze zápasů.

6.4 Implementace neuronové sítě

Poté co, jsou k dispozici data z několika předešlých let, je nutné je upravit pro finální atributy, které se budou moci ve vhodném formátu předat na zpracování neuronové síti. Trénovací data jsou rozdělena na dvě poloviny, a to na vstupní atributy a druhá tabulka je tvořena výsledky zápasů, pro určení správnosti predikcí neuronové sítě. Pro úpravu do

finální podoby se z databáze vyjmou všechny dosud nezpracované zápasy a seřadí se podle jejich data. Následně se jeden po druhém prochází a doplňují se informace o vstřelených gólech, odehraných zápasech a další informace ke každému z týmů a hráčů. Z těchto dat jsou počítány hodnoty pro finální tabulku atributů, ze které je trénována neuronová síť. V tabulce se kromě určených atributů popsaných v sekci 5.1 nachází dodatečné informace o tom, k jakému zápasu se dané statistiky vztahují.

Pro trénování bude využita třída z knihovny *scikit* a to *MLPClassifier*. Tato třída disponuje parametry pro nastavení neuronové sítě. Nejdůležitější je parametr počtu skrytých vrstev a neuronů v nich obsažených, který nese název *hidden_layer_sizes* a přijímá proměnou typu *tuple*, kde i-tý element představuje počet neuronů v i-té vrstvě. Třída poskytuje výběr ze čtyř aktivačních funkcí popsaných v kapitole 4.2.1. Poslední aktivační funkce, která nebyla vysvětlena je *identity*, která vrací na výstup x, které bylo vloženo na vstupu.

Aktivační funkce jsou vkládány do parametru *activation*. Pokud tento parametr není vybrán, je implicitně zvolena aktivační funkce ReLU. Dalším z důležitých parametrů je nastavení optimalizátoru vah, který pracuje se ztrátovou funkcí. Na výběr je ze tří možností popsaných v kapitole 4.2.2, pokud parametr není určen, tak má implicitní hodnotu algoritmu *adam*, jenž je vysvětlený v sekci 4.2.2. Dalším nastavitelným parametrem je L2 regulace, která slouží k vyhnutí se přeučení. V základu má tento parametr hodnotu jedné desetitisíciny a jako parametr se nastavuje do proměnné *alpha*.

Další je dvojice spolu úzce souvisejících parametrů a to *learning_rate* a *learning_rate_init*. První ze zmiňovaných slouží k určení způsobu, jakým se budou upravovat váhy. Druhý parametr slouží k nastavení hodnoty, o jakou se váhy budou upravovat. Jedna z možností pro způsob úpravy vah je konstantní, kdy úprava probíhá podle hodnoty nastavené v parametru *learning_rate_init*. Pokud nastavená není, tak je v parametru uložena hodnota jedna tisícina. Další z možností je *invscaling*, který postupně snižuje míru učení po každém kroku v čase t za používání dalšího parametru *power_t* jako exponentu pro inverzní škálování míry učení a to pomocí vzorce 6.1:

$$\text{effective\_learning\_rate} = \frac{\text{learning\_rate\_init}}{\text{pow(t, power\_t)}} \tag{6.1}$$

Poslední z možností pro nastavení míry učení je *adaptive*, kdy je učení drženo konstantně na nastavené hodnotě *learning_rate_init* po dobu, co ztráta při tréninku klesá. V momentě, kdy dvě po sobě jdoucí iterace nezaznamenají zlepšení, se míra učení sníží. Nastavování hodnoty pro *learning_rate* má význam pouze, pokud je jako optimalizátor vybraná možnost *sgd* popsaná v kapitole 4.2.2.

Pro náhodné počáteční nastavení vah a prahů slouží parametr *random_state*, který se nastavuje pomocí typu *integer* a je implicitně nastaven na *None*. Pořadí vzorků lze změnit v každé iteraci pomocí atributu *shuffle*, který se nastavuje pomocí Boolean hodnot a pokud je nechán beze změny, je nastaven na *True*. Parametr *max_iter* je určen pro optimalizaci pomocí *sgd* a *adam*. Slouží k určení maximálního počtu iterací, které zkusí k tomu, aby konvergoval. Implicitně je nastavený na 200 iterací a je definovaný hodnotou integer. Konvergence je určena parametrem *tol* a jedná se o toleranci pro optimalizaci. Implicitní hodnota je jedna desetitisícina a hodnota je určena jako datový typ *float*. Pokud se nezlepšuje ztráta nebo hodnocení sítě o hodnotu *tol* po *n_iter_no_change* a *learning_rate* není adaptivní, tak je konvergence považována za úspěšnou a trénování se zastaví.

Pokud je parametr *warm_start* nastaven na logickou hodnotu *True*, bude použito předchozí řešení, které vrátila metoda *fit*. Dalším parametrem je *momentum*, jenž se dá použít, pokud je jako optimalizátor zvoleno *sgd*. Slouží pro úpravu spádu gradientu. S tímto souvisí

i další parametr a tím je *nesterovs_momentum*, který slouží k určení toho, zda bude použito Nesterovovo momentum. Může být použito pouze v kombinaci s *sgd* a *momentum* musí být větší než nula.

Další parametr k využití je *early_stopping*, který je možné nastavit na logické hodnoty, kdy implicitně je nastavený na *False* a slouží k ukončení trénování v momentě, kdy se už nezlepšuje validační skóre neuronové sítě. Tento atribut je efektivní pouze při využití řešení *sgd* nebo *adam*. S tímto parametrem úzce souvisí *validation_fraction*, který slouží k určení množství dat, která se dají stranou pro případné brzké zastavení. Z toho vyplývá, že jeho hodnota musí být mezi 0 a 1.

Další dva parametry a to *beta_1* a *beta_2* 4.2.1 jsou určeny pro řešení používající *adam*, jejich hodnota by se měla pohybovat od 0 do 1 a implicitně jsou nastaveny na hodnoty blížící se k jedné. Pro *adam* slouží i parametr *epsilon*, který je hodnotou pro numerickou stabilitu. Posledním z parametrů je *max_fun* a používá se pouze s řešením *lbfgs* a určuje maximální počet volání ztrátové funkce. Pokud by síť nekonvergovala, tak může být zastavena právě pomocí tohoto parametru, místo parametru *max_iter*. Hodnota by měla být větší nebo rovna hodnotě maximálního počtů iterací. Parametr *verbose* je využíván k výpisu zpráv o progresu na standardní výstup. Automaticky je nastaven na hodnotu *False*, pro výpisky je tedy nutné ho změnit.

Co se týká atributů *MLPClassifier*, v nich si uchovává informace o označení každého výstupu, současnou ztrátu počítanou pomocí ztrátové funkce, počet iterací, které už proběhly, počet vrstev, počet výstupů, jméno výstupní aktivační funkce, i-tý element v listu reprezentujícím váhovou matici odpovídající i-té vrstvě a i-tý element v listu reprezentující prahový vektor odpovídající vrstvě i+1.

Po vytvoření instance z jejich parametrů se zavolá metoda *fit* pro vytrénování sítě, která jako výstup přijímá dvě proměnné, a to vstupní data do sítě a výsledné hodnoty. Podle toho, kolik výstupních hodnot se předá jako parametr, si síť nastaví počet výstupů. Data předávaná do sítě jako vstupy a výstupy jsou upravena do datové struktury typu *ndarray*, kterou lze vytvořit pomocí knihovny *numpy*. Metoda *fit* navrací vytrénovanou neuronovou síť, na kterou lze volat několik dalších metod. Například *partial_fit*, která slouží pro aktualizaci modelu na základě nových předložených hodnot. Zamezuje tak při přidání nových dat potřebě trénovat celou síť znovu, což je užitečné u rozsáhlých sítí s několika tisíci vzorky.

Nejdůležitější metodou je *predict*, která na základě vytrénované sítě predikuje hodnoty pro další vstupy a navrací proměnnou typu *ndarray*, kde každá položka představuje jednu predikci. Metodou *score* se dá zjistit, jakou má úspěšnost síť na datech, která jí jsou předložena. Opět má dva povinné parametry typu *ndarray* zahrnující vstupy a správné výstupy. Třída má také metody pro předpovídání pravděpodobnosti a logaritmické pravděpodobnosti – *predict_proba* a *predict_log_proba*. Obě si jako první atribut berou vytrénovanou neuronovou síť a vstupní data ve formátu *ndarray* a vrací stejný typ s pravděpodobnostmi. Zároveň třída umožňuje zavolat metody *set_params* a *get_params* pro nastavení a zjištění parametrů.

Při testování na historických datech byla z knihovny *scikit-learn* je využita její součást *train_test_split*, která slouží pro rozdělení dat pomocí poměru, a to na trénovací a testovací množinu. Zároveň parametr *randomstate* umožňuje přeházení dat do takové podoby, aby pro každou iteraci byla v jiném pořadí vybrána jak trénovací, tak testovací data a bylo možné nasbírat více validních výsledků. Pro tyto účely má i třída *MLPClassifier* metodu pod názvem *score*. Tato metoda přijme na vstup atributy pro neuronovou síť a jejich správné výsledky a na výstupu vrátí skóre neuronové sítě, což pro klasifikaci znamená procentuální úspěšnost trefených výsledků.

Kapitola 7

Testování

Testování bylo prováděno na historických datech, která následně slouží pro trénování sítě. Byla použita data zahrnující výsledky a statistiky od sezóny 2015/2016 ze všech pěti soutěží, kterých se účastní týmy anglické nejvyšší ligy, a to Premier League, FA Cupu, Ligy mistrů, EFL Cupu a Evropské ligy. Predikce jsou tvořeny pouze pro zápasy Premier League. Stejně tak probíhá i testování na historických datech. Informace z ostatních zápasů jsou pouze podpůrné a jsou sbírány pouze ze zápasů celků, které hrají v nejvyšší anglické soutěži. Dohromady se jedná o 1729 zápasů, které jsou určeny pro trénování a testování. Tato testovací kolekce je rozdělená na trénovací a testovací část v testovacím skriptu. Ve finální aplikaci probíhá trénink na všech předchozích zápasech a predikce probíhá pouze pro zápasy z daného hracího dne. Data jsou rozdělena poměrem 8,5 ku 1,5. Tedy osmdesát pět procent je určeno na trénování neuronové sítě a zbytek je určen pro testování. Trénuje se na 1469 zápasech a testuje na 260.

Celá sezóna anglické nejvyšší soutěže představuje 380 utkání. Podíl testovacích dat z celé sezóny je tedy téměř sedmdesát procent, což už by měl být dostatečný údaj pro prokazatelnost ziskovosti sítě. Z dlouhodobého hlediska by bylo vhodné, aby neuronová síť po pěti iteracích byla schopna generovat zisk. Budou provedeny testy pro 500 iterací za účelem zjištění nejvýhodnějších parametrů pro nastavení neuronové sítě. Cílem je skončit v zisku po testování 1300 zápasů, což představuje přibližně tři a půl sezóny.

Každá iterace probíhala nad stejnými 1729 zápasy jako ta předchozí, ale obě množiny testovacích i trénovacích vstupů se měnily. Nejdůležitějším kritériem je zisk a dále je to počet utkání, která byla správně vsazena. Nejdříve bylo nutné vybrat, které atributy budou do finální iterace patřit a poté se vybírala finální podoba sítě.

7.1 Testování atributů

Bylo stanoveno deset počátečních atributů zahrnujících formu, body, důležitost utkání, sílu sestavy a změnu trenéra. Všechny tyto atributy byly zahrnuty pro každý tým. První iterace byla spuštěna na neuronové síti bez skrytých vrstev, s aktivační funkcí sigmoid a optimalizací vah *lgbtfs*. Pro atributy bylo klíčové zjisti výsledné skóre sítě, tedy kolik výsledků byla schopna předpovědět na základě atributů, jak je ukázáno v tabulce 7.1.

První iterace je deset již zmíněných atributů. V druhé iteraci z nich byla odebrána změna trenéra a ve třetí jsou do nich jako další atributy přidány kurzy sázkových kanceláří. Co se týká odstranění atributu změny trenéra, ukázalo se, že úspěchy po změně většinou zaznamenávají pouze lepší týmy, které mají větší potenciál. Změna trenéra u týmů ze spod-

aktivační funkce	solver	počet iterací	atributy	skóre
sigmoid	lbfgs	1000	10	39 %
sigmoid	lbfgs	1000	8	40 %
sigmoid	lbfgs	1000	11	45 %

Tabulka 7.1: Tabulka s nastavením parametrů neuronové sítě pro vyhodnocení vhodných atributů

ního patra tabulky většinou nepřinese okamžitý efekt a týmu se ani po výměně okamžitě dařit nezačne. Rozhodl jsem se o odebrání atributu a ukázalo se, že pro celkovou úspěšnost sítě byl poměrně irelevantní, jak ukazuje druhé měření. Další myšlenkou bylo přidání kurzů sázkových kanceláří za účelem zpřesnění predikce, protože pro předzápasové predikce pravděpodobnosti jsou kurzy nejbližším a nejpřesnějším odhadem. Ukázalo se, že to opravdu predikci pomohlo a po 1000 iteracích to zvedlo úspěšnost odhadování výsledků o celých pět procent. Proto pro první iterace, ve kterých se bude testovat výsledek, bylo jako vstup použito jedenáct atributů.

7.2 Testování sítě

Jakmile byly vybrány atributy pro neuronovou síť, začal jsem testovat ideální podobu sítě. Pro výběr nejlepšího formátu byly provedeny testy nad 500 iteracemi s výpisem toho, jak si která síť vedla na kterých kurzech. Na základě těchto informací poskytuje možnost upravovat vklady pro jednotlivé příležitosti a tím se dostat do většího zisku. Nejdříve bylo cílem určit podobu neuronové sítě, co se týká počtu vrstev a neuronů v nich. Cílem je najít počet vrstev pro co největší zisk. Obecně není přesný návod kolik neuronů a kolik vrstev se má používat. Smyslem skrytých vrstev je simplifikace problému, takže by mělo být neuronů v každé vrstvě méně a méně. Často jsou sítě používány pro problémy s mnohonásobně větším počtem vstupů, než jak zde bylo určeno jedenáct. Pro zjednodušení takového problému by proto mělo být dostatečné využití dvou skrytých vrstev. Tímto postupem však nebylo nalezeno řešení, které by vykazovalo lepší výsledky než původní testování bez skrytých vrstev, což nabízelo dvě možnosti, a to nechat síť bez skrytých vrstev nebo více experimentovat. Bylo zkoušeno více skrytých vrstev, testy probíhaly na sto iteracích, protože bylo nutné udělat mnohonásobně větší množství testů. Nejlépe se z testů ukázala neuronová síť s třemi skrytými vrstvami, kdy v první je šest neuronů, stejně tak v druhé a ve třetí vrstvě je jich osm s tím, že tři výstupy zůstávají.

V následující tabulce (tabulka 7.2) je popsáno nastavení sítě. V dalších tabulkách se na číselné označení odkážou výsledky jednotlivých sítí. Výpis není z každé iterace, která byla provedena, ale jedná se o nejlepší výsledky, co se dané aktivační funkce a optimalizátoru týká. Není zde využita solver sgd ani aktivační funkce tanh, protože ani jeden nepodával v ničem dostatečně lepší výkony, ani v zisku, v úspěšnosti sázek nebo nějakém specifickém kurzu. Byla testována i varianta bez kurzových sázek za účelem zjištění, zda se neuronová síť nepřizpůsobila kurzům příliš a nyní v závislosti na tom nevynechává vyšší kurzy. V následující tabulce 7.3 se nacházejí sloupce pro deficit a zisk, jenž jsou počty iterací, které byly celkově v plusu nebo mínusu. Sloupec celkový zisk nese informace o výsledku po 500 iteracích, kdy na každou příležitost, kterou neuronová síť doporučí, je vsazeno 50 jednotek.

číslo	aktivační funkce	solver	počet iterací	rychlost učení	atributy
1	sigmoid	adam	500	0.15	8
2	sigmoid	lbfgs	500	0.1	8
3	relu	adam	500	0.001	8
4	relu	lbfgs	500	0.001	8
5	sigmoid	adam	500	0.1	11
6	sigmoid	lbfgs	500	0.1	11
7	relu	adam	500	0.001	11
8	relu	lbfgs	500	0.001	11

Tabulka 7.2: Tabulka s informacemi o nastavení atributů pro neuronové sítě

číslo	skóre	zisk	deficit	celkový zisk	vyhráno	prohráno	nevsazeno
1	40 %	309	191	151 390	52 050	51 108	26 482
2	0	0	0	0	0	0	130 000
3	50,9 %	249	251	-5 542	66 245	60 917	2 838
4	43 %	176	324	-458 334	55 938	69 053	5 009
5	47,5 %	349	151	240 892	61 817	62 622	5 561
6	0 %	0	0	0	0	0	130 000
7	47,8 %	323	177	184 867	62 152	55 721	12 127
8	36,9 %	199	301	-713 123	48 001	79 695	2 304

Tabulka 7.3: Tabulka s výsledky na základě parametrů neuronové sítě z tabulka 7.2

Pro nastavení číslo dva i při nastavení různé velikosti míry učení síť nebyla schopná nalézt řešení pro žádnou sázku v žádné z 500 iterací. Krok byl nastaven od jedné tisíciny do jedné patnáctiny, ani v jednom bodě síť nebyla schopná se učit. Stejně dopadlo i učení pro stejné nastavení s přidanými kurzy sázkových kanceláří. U osmého nastavení byly potíže s konvergováním lbfgs.

Kurzy v tabulce (tabulka 7.4) jsou v rozsahu od o desetinu menší než je hodnota kurzu sázkové kanceláře do hodnoty o desetinu větší tedy od 1 do 1,1, od 1,1 do 1,2 a tak dále. V druhé tabulce (tabulka 7.5) je rozsah od předchozí hodnoty do té další, tedy od 1,9 do 2,0 a 3,0 do 3,5. Pro všechny tyto možnosti byl vypočítán i průměrný kurz, na který se v dané hladině vsázelo. Ten se ve většině pohyboval lehce pod hranicí poloviny. Pro hodnoty mezi 1,1 a 1,2 to bylo 1,14. To, že průměr je nižší než polovina je dáno tím, že menší kurz znamená větší pravděpodobnost na vítězství a tak se může předpokládat, že z dané kurzové hladiny na stupnici od 1,1 do 1,2 si algoritmus bude vybírat menší kurzy.

Z výsledků se dají odfiltrovat neúspěšné sítě a u těch co prokázaly zisk z dlouhodobého hlediska zjistit, u jakých kurzů slavily největší úspěšnost. Na těchto údajích se následně dá postavit strategie pro sázení, která nebude mít takové výkyvy a bude mít minimální šanci na prodělečnou sezónu. Vypočítaná procenta z úspěšnosti se převedou na kurzy. V následujících tabulkách 7.6, 7.7 tedy menší kurz znamená menší riziko a lze vyčíst, že na některé kurzy, které napovídají jednotlivé neuronové sítě, se vyplatí vsázet více než jiné. Zeleně zvýrazněná jsou pole, která jsou výhodnější proti průměru, který byl sázen.

Ukázalo se, že obecně lepší neuronové sítě pro posbíraná data produkuje optimalizátor adam ve srovnání s ostatními. V samotných aktivačních funkcích už takové rozdíly nebyly.

číslo	do 1.1	do 1.2	do 1.3	do 1.4	do 1.5	do 1.6	do 1.7	do 1.8	do 1.9
1	83 %	91 %	76 %	75 %	67 %	70 %	60 %	54 %	56 %
2	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
3	80 %	90 %	77 %	74 %	68 %	67 %	58 %	55 %	55 %
4	81 %	88 %	74 %	73 %	66 %	66 %	57 %	55 %	55 %
5	80 %	91 %	75 %	76 %	67 %	70 %	59 %	54 %	59 %
6	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
7	79 %	90 %	77 %	75 %	68 %	69 %	60 %	56 %	59 %
8	70 %	83 %	69 %	69 %	61 %	64 %	56 %	51 %	53 %

Tabulka 7.4: Tabulka s procentuálním vyhodnocením úspěšnosti neuronových sítí

číslo	do 2.0	do 2.2	do 2.5	do 3.0	do 3.5	do 4.0	nad 4.0
1	56 %	47 %	44 %	33 %	35 %	32 %	21 %
2	0 %	0 %	0 %	0 %	0 %	0 %	0 %
3	54 %	48 %	45 %	33 %	32 %	29 %	22 %
4	53 %	47 %	42 %	31 %	26 %	21 %	15 %
5	55 %	48 %	44 %	34 %	36 %	31 %	22 %
6	0 %	0 %	0 %	0 %	0 %	0 %	0 %
7	56 %	48 %	44 %	34 %	35 %	32 %	25 %
8	52 %	45 %	40 %	31 %	24 %	19 %	15 %

Tabulka 7.5: Tabulka s procentuálním vyhodnocením úspěšnosti neuronových sítí

Zároveň je zajímavé srovnaní neuronové sítě číslo tři, která procentuálně dosáhla nejvyššího úspěchu, kdy dokázala odhadnout přes polovinu utkání. Přesto se nedostala do plusových čísel, protože trefila málo výsledků v rozmezí 1,2 až 1,9, což je větší část sázek, přestože zaznamenala poměrně dobré výsledky u vyšších kurzů.

Nejméně kurzů se vyskytuje v hladině pod 1,1. Anglická liga je velice vyrovnaná a podobné kurzy se objevují minimálně a z výsledků vyplývá, že podávat na ně sázky se nevyplatí. Za necelých pět sezón se těchto kurzů nahromadilo pouze 15. Například v hladině od 1,1 do 1,2 jich bylo 59 a testovaný vzorek je větší. Co se týká vynechávání sázení u hladiny pod 1,1 nebylo doporučeno vyhnout se sázení ani jednou, což je pochopitelné na základě okolností, které k vypsání tak nízkého kurzu vedou. Na zbytku kurzů od 1,1 do 2,0 jde vidět proměnlivost, co se téměř každé hladiny týká. Nejvíce sázek v každé iteraci neuronové sítě umístily na hladinu mezi 1,2 a 1,3, ale z dlouhodobého hlediska by se jednalo o ztrátovou variantu. Nejvíce výsledků, na které bylo doporučeno nevsázet, se pohybovalo od kurzu 2,2 výše, ale nakonec vyšly. Vysokých kurzů vychází poměrně velké množství, ale jedná se většinou o vyrovnané zápasy, u kterých je těžké něco předvídat. Zde by bylo vhodnější regresní řešení, kdy by se určila pravděpodobnost pro všechny možnosti a na základě toho by se rozložil vklad. S odebráním marže z každého kurzu od sázkové kanceláře by ale bylo ještě těžší z dlouhodobého hlediska získat profit.

Pro výběr finální sítě je důležité se zamyslet nad stylem, jakému se chce sázkař věnovat. Z dlouhodobého hlediska se mi osvědčilo vybírání menších kurzů. To znamená kurzy od 1 do 1,5, nicméně je nutné se podívat do tabulek ze kterých vyplývá, že se v žádném z případů nevyplatilo podávat sázky na kurzy menší než 1,1. S ohledem na výsledky byla

číslo	do 1.1	do 1.2	do 1.3	do 1.4	do 1.5	do 1.6	do 1.7	do 1.8	do 1.9
1	1,19	1,1	1,32	1,33	1,5	1,43	1,68	1,84	1,78
2	0	0	0	0	0	0	0	0	0
3	1,25	1,11	1,3	1,35	1,47	1,49	1,72	1,81	1,81
4	1,23	1,14	1,35	1,37	1,51	1,51	1,75	1,81	1,81
5	1,25	1,1	1,33	1,32	1,49	1,43	1,69	1,85	1,69
6	0	0	0	0	0	0	0	0	0
7	1,27	1,11	1,3	1,33	1,47	1,45	1,67	1,79	1,69
8	1,43	1,2	1,45	1,45	1,64	1,57	1,79	1,96	1,89

Tabulka 7.6: Tabulka s kurzovým vyhodnocením úspěšnosti neuronových sítí

číslo	do 2.0	do 2.2	do 2.5	do 3.0	do 3.5	do 4.0	nad 4.0
1	1,79	2,13	2,27	3,04	2,86	3,125	4,76
2	0	0	0	0	0	0	0
3	1,85	2,08	2,22	3,03	3,13	3,45	4,55
4	1,89	2,13	2,38	3,23	3,85	4,76	6,66
5	1,81	2,08	2,27	2,94	2,78	3,23	4,55
6	0	0	0	0	0	0	0
7	1,79	2,08	2,27	2,94	2,86	3,13	4
8	1,92	2,22	2,5	3,23	4,16	5,26	6,66

Tabulka 7.7: Tabulka s kurzovým vyhodnocením úspěšnosti neuronových sítí

do finálního skriptu vybrána neuronová síť číslo pět. Nedosáhla sice nejlepšího hodnocení, co se skóre týká, ale měla největší podíl iterací v zisku, kdy iterace je necelých sedmdesát procent jedné sezóny. V sedmdesáti procentech tato síť skončí v kladných hodnotách. Za 342 odsimulovaných sezón se dostala do zisku průměrně 704 jednotek za sezónu při sázce 50 jednotek na každý doporučený zápas. Výkony sítě by mohly být ještě vylepšené co se čistého zisku týká, a to rozložením větších vkladů na příležitosti, které se z dlouhodobého hlediska ukázaly jako více profitující. Zde je znovu na uvážení sázkaře, do jakých kurzů se chce pouštět, protože kurzy nad 2,5 a více jsou už poměrně rizikové. Jedná se o pravděpodobnost čtyřiceti procent a nižší. Nicméně více možností ukázalo, že z dlouhodobého hlediska se vyšší kurzy mohou vyplatit. Ovšem návratnost je rozhodně pomalejší než u těch nižších, protože i když po čtyřech sázkách na kurz čtyři a vyšší sázka vyjde, tak návratnost je pomalejší, než když vyjde u kurzu na hranici 1,15 devadesát procent zápasů. Na začátku testování bylo uvedeno, že cílem je, aby po dobu pěti iterací byla síť schopna zaručit zisk, její výsledky v několika běžích jsou ukázány v následující tabulce 7.8.

Je nutné zmínit dobrý výběr atributů, protože i přes odebrání marže, kterou si berou sázkové kanceláře, se v několika případech podařilo snížit riziko pod kurzy, které sázková kancelář nabízí. V ostatních případech lze pozorovat, jak kurzy a pravděpodobnosti kopírují kurzy sázkových kanceláří. Což je u nejsledovanější fotbalové ligy světa a pravděpodobně také nejlépe odhadnuté, co se týká hladiny kurzů, úspěch. Zároveň jde vidět, že i dobře postavená neuronová síť, do které nebyly přidány kurzy kanceláře jako atribut, dokázala dosáhnout zisku.

běh	počet iterací	zisk
1	5	1 500
2	5	2 611
3	5	2 557
4	5	- 257
5	5	2 451
6	5	1 132
7	5	1 461
8	5	4 748
9	5	3 165
10	5	1 062

Tabulka 7.8: Tabulka se ziskovým vyhodnocením během menšího počtu iterací

Co se týká pandemie, která ovlivnila svět fotbalu, z hlediska zpracování zápasů na ni bylo nahlíženo jako na restart sezóny. Aby bylo možné předpovídat, musely mít oba týmy odehrány alespoň tři zápasy. Výjimka byla zapracovaná tím způsobem, že pokud nastane více než dvacetidenní rozestup mezi zápasy, tak se to počítá jako restart sezóny. Reprezentační pauzy trvají kratší dobu, a tak se na ně vztahují jiná pravidla.

Kapitola 8

Závěr

Cílem práce bylo odhadnout kurzy a z dlouhodobého hlediska dosáhnout zisku při sázení u sázkových kanceláří. Zaměření bylo na nejvyšší anglickou fotbalovou ligu a jednalo se o vyhodnocení na základě historických dat.

Cíl pro dosažení profitu byl nastaven na necelé tři a půl fotbalové sezóny. Za tuto dobu vybraná neuronová síť dokázala skončit v zisku v devíti případech z deseti. Dokázala to při sázení padesáti jednotek na jedno utkání. V jednom případě ve kterém se to nepodařilo, byla částka v deficitu vzhledem k vsazené částce minimální. Tato možnost je rozebrána v sekci 7.2, kde je zmíněna možnost přizpůsobit částky v závislosti na úspěšnosti na určitých kurzech. Touto úpravou by bylo možné dojít k ještě větším ziskům. Marže sázkových kanceláří jsou také problematické, ale pochopitelné a je to něco, s čím se musí počítat, pokud má někdo v úmyslu vyhrávat.

Práce by se dále dala rozšířit dvěma směry. Jedním z nich by byl pokus získat data z jiných lig a vytrénovat na ně neuronovou síť podobným způsobem. Druhou možností rozšíření jsou rozšiřující sázky pro jednotlivé zápasy, například na vstřelené branky v utkání, na průběhy poločasů a další možnosti, které se u zápasů nacházejí. S ohledem na to, že nebyly brány atributy, které by byly nějak specifické pouze pro anglický fotbal, tak by neuronová síť měla fungovat i pro ostatní ligy.

Co se týká snahy o poražení sázkových kanceláří, byl by vhodný výběr méně sledované ligy, případně i jiného sportovního odvětví. Sporty, které nejsou tolik sledované a nabízí pouze dvě možnosti výsledku, by byly ideální příležitostí. Možná nejlepší volbou by byly e-sporty, které ještě nejsou tak známé a kurzy na ně nejsou tak dobře vypracované. Dále se nabízí tenis, ve kterém se však na nižších úrovních objevilo velké množství problémů s korupcí a podplácením, protože se jedná o sport pro jednotlivce.

Literatura

[1] Applications of logistic regression to shots at goal in association football: calculation of shot probabilities, quantification of factors and player/team. 2004.

[2] *Performance analysis*. 2015, [Online; navštíveno 29.10.2019].
URL https://www.footballscience.net/special-topics/performance-analysis/?fbclid=IwAR0aUcLeDth1WCWTwOXAZ6FKGDQN4MyvXXkeulH2YNA0PKGgN-Ts_PT2es8

[3] *Why you should consider home advantage for football trading*. 2017, [Online; navštíveno 30.10.2019].
URL https://help.smarkets.com/hc/en-gb/articles/115000647291-Why-you-should-consider-home-advantage-for-football-trading

[4] *Fotbal*. 2018, [Online; navštíveno 18.10.2018].
URL https://www.britannica.com/sports/football-soccer

[5] *History: Discover the origins and history of the top tier of English football*. 2018, [Online; navštíveno 17.10.2018].
URL https://www.premierleague.com/history

[6] *How Sports Betting Odds are calculated*. 2019, [Online; navštíveno 16.05.2020].
URL https://wetten.com/ca/info/guide/how-betting-odds-are-calculated/

[7] *Ole Gunnar Solskjaer*. 2019, [Online; navštíveno 16.05.2020].
URL https://www.transfermarkt.com/ole-gunnar-solskjaer/stationen/trainer/7286

[8] *Premier League - All time stats*. 2019, [Online; navštíveno 28.10.2019].
URL https://www.premierleague.com/stats/top/clubs/

[9] *Ptali jste se bookmakera. Přečtěte si odpovědi*. 07 2019, [Online; navštíveno 16.05.2020].
URL https://www.tipsport.cz/clanek/12671-ptejte-se-bookmakera-

[10] *Výslekdy zápasů sezóny 2018/2019*. 2019, [Online; navštíveno 16.05.2020].
URL https://www.livesport.cz/fotbal/anglie/premier-league-2018-2019/vysledky/

[11] *Aktuální tabulka Premier League*. 2020, [Online; navštíveno 16.05.2020].
URL https://www.livesport.cz/fotbal/anglie/premier-league/tabulka/

[12] *Club coefficients*. 2020, [Online; navštíveno 03.07.2020].
URL https://www.uefa.com/memberassociations/uefarankings/club/#/yr/2020

[13] *FC Barcelona*. 2020, [Online; navštíveno 03.07.2020].
URL https://www.transfermarkt.com/fc-barcelona/startseite/verein/131

[14] *Kombi tiket ze sázkové kanceláře Tipsport*. 2020, [Online; navštíveno 01.05.2020].
URL Tipsport.net

[15] *Livesport*. 2020, [Online; navštíveno 03.07.2020].
URL https://www.livesport.cz/

[16] *Stevenage FC*. 2020, [Online; navštíveno 03.07.2020].
URL https://www.transfermarkt.com/fc-stevenage/startseite/verein/3684/saison_id/2019

[17] *Team Statistics*. 2020, [Online; navštíveno 03.07.2020].
URL https://www.whoscored.com/

[18] Brownlee, J.: *A Gentle Introduction to Early Stopping to Avoid Overtraining Neural Networks*. 06. 08. 2019, [Online; navštíveno 17.05.2020].
URL https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/

[19] Brownlee, J.: *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*. 14. 11. 2019, [Online; navštíveno 17.07.2020].
URL https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

[20] Brownlee, J.: Supervised and Unsupervised Machine Learning Algorithms [online]. 16. 03. 2016 [cit. 2020-05-17].
URL https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[21] Bělohlávek, J.: Agent pro kurzové sázení [online]. 2008 [cit. 2020-05-17].
URL http://hdl.handle.net/11012/53223

[22] Carr, P.: *How Leicester City's 5,000-1 odds compare to other long shots. 2016, [Online; navštíveno 17.10.2018].*
URL http://www.espn.com/chalk/story/_/id/15447878/putting-leicester-city-5000-1-odds-perspective-other-long-shots-espn-chalk

[23] Durčák, I. P.: Neuronové sítě a princip jejich fungování. 08. 09. 2017, [Online; navštíveno 17.05.2020].
URL https://www.napocitaci.cz/33/neuronove-site-a-princip-jejich-fungovani-uniqueidgOkE4NvrWuNY54vrLeM670eFNQh552VdDDulZX7UDBY/

[24] Gregorio, F. D.: psycopg2. 6. 4. 2020, [Online; navštíveno 17.07.2020].
URL https://pypi.org/project/psycopg2/

[25] Kingma, D. P.; Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980*, 2014.*

[26] KUBEŠ, J.: Modelování a odhadování výsledků sportovních utkání za použití pravděpodobnosti a statistiky [online]. 2011 [cit. 2020-05-14].
URL https://theses.cz/id/23p3st/

[27] *Le Van, T.: Statistické a rozhodovací postupy při sázení [online]. 2015 [cit. 2020-05-17].*
URL http://hdl.handle.net/11025/17974

[28] *Milan: Zvládne Manchester United bitvu s Liverpoolem? 03 2018, [Online; navštíveno 17.05.2020].*
URL https://www.livekurzy.cz/premier-league-zvladne-manchester-united-bitvu-s-liverpoolem/

[29] *Mitchell, T. M.:* Machine Learning. *USA: McGraw-Hill, Inc., první vydání, 1997, ISBN 0070428077.*

[30] *Nagyfi, R.: The differences between Artificial and Biological Neural Networks. 04. 09. 2018, [Online; navštíveno 17.06.2020].*
URL https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7

[31] *Ng, A.: Machine Learning by Stanford University [online]. 2015 [cit. 2020-05-17].*
URL https://www.coursera.org/learn/machine-learning/home/

[32] *Nielsen, M. A.:* Neural Networks and Deep learning. *Determination Press, 2015.*

[33] *Oliphant, T. E.:* A guide to NumPy, *ročník 1. Trelgol Publishing USA, 2006.*

[34] *Pedregosa, F.; Varoquaux, G.; Gramfort, A.; aj.: Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, *ročník 12, 2011: s. 2825–2830.*

[35] *Rizvi, Z.; Akhtar, S.; Sabeeh, W.; aj.: Effective thermal conductivity of unsaturated soils based on deep learning algorithm. 09 2020.*

[36] *Samuel, A. L.: Some studies in machine learning using the game of checkers.* IBM Journal of Research and Development, *ročník 44, č. 1.2, 2000: s. 206–226.*

[37] *Scott, M.: Premier League set to loosen 'weakened teams' rule that hit Holloway. 07 2011, [Online; navštíveno 16.05.2020].*
URL https://www.theguardian.com/football/2011/jul/15/premier-league-weakened-teams-holloway

[38] *Seber, G. A.; Lee, A. J.:* Linear regression analysis, *ročník 329. John Wiley & Sons, 2012.*

[39] *Sharma, S.: Activation Functions in Neural Networks. 06. 09. 2017, [Online; navštíveno 17.06.2020].*
URL https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

[40] *Shivashish: Top 15 Most Used Machine Learning Tools By Experts. 13. 03. 2020, [Online; navštíveno 18.06.2020].*
URL https://towardsdatascience.com/top-15-most-used-machine-learning-tools-by-experts-d6602f1ac14c

[41] *Srinivasan, A. V.: Stochastic Gradient Descent — Clearly Explained. 7. 9. 2019, [Online; navštíveno 17.07.2020].*
URL https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31/

Příloha A

Obsah CD

- Soubor README.txt s popisem adresářů
- Návod pro spuštění aplikace a testů
- Zdrojové kódy aplikace
- Text bakalářské práce v PDF
- Zdrojový kód bakalářské práce v LaTeX