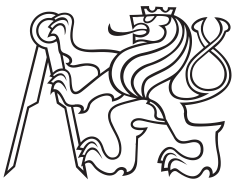
Bakalářská práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra řídicí techniky

Predikování hokejových zápasů pomocí neuronových sítí

Tomáš Grim

Vedoucí: Ing. Gustav Šír
Obor: Kybernetika a robotika
Srpen 2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Grim** Jméno: **Tomáš** Osobní číslo: **474565**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra řídicí techniky**

Studijní program: **Kybernetika a robotika**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Predikování hokejových zápasů pomocí neuronových sítí

Název bakalářské práce anglicky:

Predicting Ice-Hockey Matches with Neural Networks

Pokyny pro vypracování:

V posledních letech metody strojového učení rapidně mění odvětví prediktivní sportovní analýzy, avšak rigorózních a veřejně dostupných studií v této oblasti je stále málo. I když roste počet prací modelujících zápasy ve fotbale - nejpopulárnějším světovém sportu, ostatním sportům byla zatím věnována podstatně menší pozornost. Tato práce se zaměří na lední hokej, který je pravděpodobně nejoblíbenějším sportem v České republice. Očekává se, že student vyvine vlastní modely strojového učení pro problém predikcí výsledku hokejových zápasů, primárně založených na neuronových sítích. Zvláštní pozornost pak bude věnována modelování zápasů na různých úrovních datové granularity, aby se zjistilo, zda zvýšená granularita pomáhá zlepšit předpovědi, což je v této doméně zatím nevyjasněná hypotéza.
1) Analyzujte stávající modely v prediktivní sportovní analýze se zvláštním zaměřením na hokej.
2) Prozkoumejte dostupné zdroje hokejových záznamů s cílem získat statisticky významné množství dat s dostatečnou úrovní granularity.
3) Proveďte čištění dat, transformaci a extrakci prediktivních rysů.
4) Experimentujte s různými reprezentacemi dat, granularitou, a neurálními architekturami.
5) Porovnejte se SotA modely, případně i z jiných sportů.

Seznam doporučené literatury:

[1] Weissbock, Joshua. Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data. Diss. Université d'Ottawa/University of Ottawa, 2014.
[2] Pischedda, Gianni. "Predicting NHL match outcomes with ML models." International Journal of Computer Applications 101.9 (2014).
[3] Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.
[4] Gu, Wei, et al. "A game-predicting expert system using big data and machine learning." Expert Systems with Applications 130 (2019): 293-305.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

Ing. Gustav Šír, katedra počítačů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **05.01.2021** Termín odevzdání bakalářské práce: **13.08.2021**

Platnost zadání bakalářské práce:
do konce zimního semestru 2022/2023

Ing. Gustav Šír	prof. Ing. Michael Šebek, DrSc.	prof. Mgr. Petr Páta, Ph.D.
podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. _____
Datum převzetí zadání

_____
Podpis studenta

Poděkování

Tímto bych chtěl poděkovat svému vedoucímu Ing. Gustavu Šírovi, jehož vedení a rady byly velkým přínosem. Dále bych chtěl poděkovat za poskytnutí výpočetního výkonu projektem „e-Infrastruktura CZ“, jež byl využit při provádění experimentů. Na závěr bych také chtěl poděkovat své rodině, která mi nejen při psaní práce, ale i po celou dobu studia byla nesmírnou podporou.

Prohlášení

Tímto prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškerou použitou literaturu v souladu s metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 13. srpna 2021

Abstrakt

Cílem práce je vytvořit modely, využívající neuronové sítě, schopné predikce výsledků hokejových zápasů NHL. Na těchto modelech chceme následně ověřit několik doposud nepotvrzených hypotéz, zejména hypotézu tvrdící, že zvýšená granularita vstupních dat pomáhá ke zlepšení predikčních schopností modelu. Dílčími kroky jsou sběr a příprava statisticky významného množství hokejových dat pro učení modelů, tvorba experimentálního protokolu a ladění hyperparametrů.

Klíčová slova: strojové učení, neuronové sítě, predikce sportovních výsledků, NHL,lední hokej

Vedoucí: Ing. Gustav Šír
Katedra Počítačů, FEL, ČVUT v Praze
Karlovo náměstí 13, Praha 2

Abstract

The aim of this thesis is to create models using neural networks, capable of predicting the results of NHL hockey matches. With these models, we want to subsequently verify several hitherto unconfirmed hypotheses, especially the hypothesis that the increased granularity of the input data helps to improve the predictive capabilities of the models. Partial steps are the collection and preparation of a statistically significant amount of hockey data for learning models, creating an experimental protocol, and tuning hyperparameters.

Keywords: machine learning, neural networks, prediction of sports results, NHL, ice hockey

Title translation: Predicting Ice-Hockey Matches with Neural Networks

Obsah

Kapitola 1

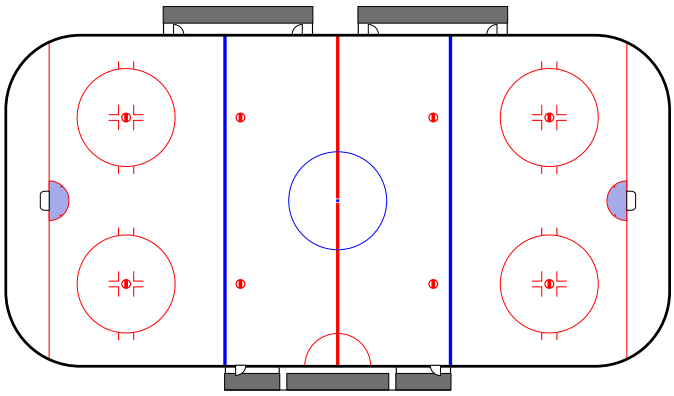
Úvod

Vývoj strojového učení se za posledních pár let výrazně posunul vpřed, zejména pak v oblasti neuronových sítí. Hlavním faktorem nebyl objev nových případně lepších algoritmů, nýbrž vyšší výkon počítačů. Dnešní počítače svým výkonem dalece převyšují počítače z konce 50. let, tedy doby kdy byly prováděny první experimenty s neuronovými sítěmi. V roce 1958 F. Rosenblatt [19] vytvořil první program (*Perceptron*) využívající neuronové sítě. Program tehdy běžel na počítači IBM 704 o výkonu 12 kFLOPS, což je zhruba milionkrát méně než měly hodinky Apple Watch 1. generace [6]. Právě díky výpočetnímu výkonu dnešních domácích počítačů se metody strojového učení, potažmo neuronových sítí, dostaly do rukou více lidí a staly se tak dostupným nástrojem pro řešení široké škály problémů. Dnes mohou lidé s trochou znalosti programování a s využitím veřejně dostupných knihoven jednoduše vytvářet modely využívající neuronové sítě. Například pro predikci výsledků jejich oblíbeného hokejového týmu.

1.1 Lední hokej

Lední hokej patří mezi nejpopulárnější sporty na světě a to Českou republiku nevyjímaje. Právě u nás se těší velké základně odborného i laického publika a můžeme jej pozorovat na všech výkonnostních úrovních. Vzhledem ke globálnímu rozsahu se od sebe pravidla[1] různých hokejových lig lehce liší, ale základy zůstávají stejné.

[1] Kompletní pravidla stanovené mezinárodní hokejovou federací je možné si přečíst zde.



Obrázek 1.1: Hokejové hřiště při pohledu shora

Hokej se řadí mezi týmové sporty, kde se střetávají dva týmy na ledové hrací ploše viz Obrázek 1.1 s cílem vstřelit puk do soupeřovy branky. Základní hrací doba je rozdělena na tři třetiny, kde každá čítá 20 minut. V případě nerozhodného stavu je hráno prodloužení formou „zlatého gólu“[2]. Pokud nedojde k rozhodnutí ani v prodloužení, přichází na řadu samostatné nájezdy. Do zápasu může nastoupit ±20 hráčů[3] za každý tým. Při normálním stavu má na hřišti každý tým 1 brankáře a 5 hráčů v poli, zbytek pak čeká na hráčské lavici. Na ledě se kromě 12 hráčů pohybují ještě 4 rozhodčí[4], kteří dohlížejí na férový průběh hry. Pokud některý z hráčů udělá faul, je poslán na trestnou lavici, kde stráví 2 až 10 minut podle závažnosti trestu.

Každý z hráčů na ledě zastává specifickou pozici, přičemž cíle jednotlivých pozic se liší:

- Útočník
  - vstřelit co největší počet gólů
  - Středový útočník
    - chodit na vhazování a snažit získat puk ve prospěch svého týmu
    - pomáhat obráncům s bráněním
  - Křídelní útočník
    - zavážet puk do útočného pásma
    - napadat soupeřovy obránce při rozehrávce
- Obránce
  - zabránit soupeřovým útočníkům ve vytváření šancí

[2] Tým, který vstřelí v prodloužení gól jako první, vyhrává a zápas tím končí

[3] toto číslo se liší v závislosti na pravidlech konkrétní soutěže, v NHL je to 18 hráčů do pole a 2 brankáři

[4] 2 hlavní rozhodčí a 2 čároví rozhodčí

- rozehrávat puk útočníkům

- Brankář
  - pochytat co nejvíce puků směřujících do branky

Vzhledem k těmto různým cílům jsou pro jednotlivé pozice zapotřebí odlišné dovednosti, proto většina profesionálních hráčů hraje a trénuje na jediné pozici celou svoji kariéru.

1.2 Popis problému

V hokeji lze vytvářet predikční modely pro spoustu různých událostí. Nejčastěji se však vytváří modely pro predikci výsledku zápasu. Hokejový zápas může skončit 3 různými výsledky. Prvním je výhra domácího týmu, druhým je remíza a třetím je výhra týmu hostů. V NHL, jíž se tato práce zabývá, se touto formou momentálně nehraje a po skončení základní hrací doby remízou se hraje prodloužení. V případně nerozhodného stavu i po prodloužení následují samostatné nájezdy do doby, než není znám vítěz. Tento formát je však v NHL, vzhledem ke své historii, zaveden pouze krátkou dobu. Aby se modely při trénování nemusely omezovat pouze na toto období, je základní predikční problém předefinován z binární klasifikace vítěze zápasu na 3 třídovou klasifikaci výsledku zápasu po základní hrací době. Tato změna umožňuje využívat jak data ze zápasů hraných novým formátem, tak i ze zápasů končících remízou. Cílem práce je pak vytvořit modely využívající neuronové sítě, schopné predikce tohoto předefinovaného problému.

1.3 Přehled kapitol

V této 1. Kapitole jsou popsány základní pravidla hokeje, je zde také definován problém a cíl práce. V následující Kapitole 2 jsou popsány vybrané předešlé práce, zabývající se tématikou strojového učení, sportovní analýzy a především práce spojené s využitím neuronových sítí pro predikci sportovních utkání. Ve 3. Kapitole je popsána práce s daty, od jejich výběru a sběru až po finální předzpracování. Ve 4. Kapitole jsou popsány všechny modely použité pro experimenty. V Kapitole 5 jsou popsány provedené experimenty a jejich výsledky. Poslední 6. Kapitola obsahuje závěr a návrhy na pokračování práce.

Kapitola 2

Rešerše

V této kapitole je stručný souhrn informací získaných při teoretické přípravě z prací souvisejících s problematikou strojového učení a predikce sportovních výsledků. Tyto poznatky jsou řazeny chronologicky dle data publikace jednotlivých článků.

V roce 1998 prováděly Deborah Feltz a Cathy Lirgg [7] experiment, při kterém se před každým hokejovým zápasem univerzitní ligy po dobu jedné sezóny ptali jednotlivých hráčů, zda si myslí, že soupeře porazí v 8 vybraných týmových statistikách. Dále se hráčů ptali na 3 otázky týkající se jejich osobního výkonu. Cílem experimentu bylo porovnat, jak moc jsou tyto subjektivní názory ukazatelem pro výsledek zápasu. Výsledkem jejich experimentu bylo, že zprůměrované subjektivní názory hráčů na kvality jejich týmu, jsou lepším ukazatelem než jejich zprůměrované názory ohledně vlastního výkonu.

V roce 2003 Joshua Kahn [11] vytvářel model pro predikci výsledků v NFL[1]. Jako model zvolil neuronovou síť s jednou skrytou vrstvou o 3 neuronech. Dále bylo vybráno 5 statistik popisující každý tým na základě kvalifikovaného posudku. Model byl učen na zápasech odehraných v prvních 12 týdnech sezóny 2003, testován pak byl na 14 a 15[2] týdnu kde dosahoval správné predikce v 57% a 85% zápasů. Závěry a postupy této práce jsou v některých ohledech diskutabilní.

V roce 2005 publikovali Patric Andersson, Jan Edman a Mattias Ekman článek [1], ve kterém se zaměřují na porovnání lidských predikčních schopností v závislosti na úrovni znalosti sportu. Článek je založen na datech posbíraných při mistrovství světa ve fotbale v roce 2002. Jednotliví účastníci tohoto průzkumu byly rozděleni do dvou skupin, první obdržela pouze dotazník s otázkami, zatímco druhá skupina měla k dotazníku přiložený světový žebříček týmů. Výsledkem tohoto průzkumu bylo, že fotbaloví experti měli nižší

[1] obdoba hokejové NHL pro americký fotbal

[2] celkem měl k dispozici ze 14 týdnů 208 zápasů

úspěšnost než účastníci, kteří neměli o sportu moc znalostí. Článek obsahuje dvě možné vysvětlení tohoto neočekávaného výsledku. Prvním z nich je výpadek 3 favorizovaných týmů hned v prvním kole a druhým je, že účastníci s nižšími znalostmi sportu využívali přiloženého světového žebříčku více než experti.

V roce 2007 Burak Aslan a Mustaf Inceoglu ve své práci [2] srovnávali dva různé LVQ[3] modely při predikci na fotbalových zápasech v sezóně 2001/02 (34 hracích týdnů) italské nejvyšší fotbalové ligy[4]. Jako vstupní vektor nejsou brány statistiky nýbrž RANKING, který je vytvářen při předzpracování dat z výsledků zápasů v jednotlivých kolech. Do trénování nejsou zahrnuty zápasy z prvních 6 týdnů, ale je na nich počítán RANKING. Pro trénink je využito trénovací okno, které bere první polovinu sezóny jako trénovací data a na následujícím týdnu testuje naučené vlastnosti. Po otestování vlastností je trénovací okno doplněno tento týden a pokračuje v učení. Takto naučené modely dosahují 51.13% a 53.25% úspěšnosti při testování na druhé polovině sezóny.

V roce 2008 Alan McCabe a Jarrod Trevathan publikovali článek [15] ve kterém se zabývají predikcí výsledků zápasů ve 4 ligách (NRL, AFL, EPL a Super Rugby). Pro predikci je využit MLP[5] model. Mezi vstupními hodnotami se objevili kromě numerických statistik[6] také údaje o chybějících hráčích a místu utkání[7]. Modely byly trénovány s pomocí zvětšujícího se okna stejně jako u [2], tentokrát však nebyly vynechány úvodní zápasy díky většímu počtu historických záznamů. Model dosahoval v průměru nejlepších výsledků v Super Rugby lize, kde měl úspěšnost 67.5%, nejhorších průměrných výsledků dosáhl v anglické EPL a to 54.6%. Autor toto připisoval formátu anglické ligy, kdy na konci každé sezóny sestoupí nejhorší 3 týmy do nižší ligy a opačným směrem jdou první 3 týmy z této nižší soutěže.

V roce 2009 se John McCullagh a další [16] zabývali problematikou náboru[8] mladých hráčů do týmů hrajících AFL. Každý rok překročí minimální věkovou hranici pro vstup do AFL desítky elitních juniorských hráčů. Cílem rekrutujících manažerů týmů v AFL je co nejlépe odhadnout dovednosti těchto hráčů a s nejlepšími z nich při draftu uzavřít profesionální kontrakt. Jelikož je tento způsob nejjednodušším způsobem jak získat talentované hráče je do tzv. „scoutingu“ investováno velké množství peněz. Cílem autorů této práce, tak bylo vyvinout model s využitím neuronových sítí, který by byl na základě fyzických předpokladů, měřených před draftem, schopný odhalit nadcházející hvězdné hráče a ulehčit tak práci rekrutujícím manažerů. Model využíval kombinaci dvou neuronových sítí, kde každá vytvářela RATING hráčů. První síť hodnotila hráče se vzrůstem do 188 cm a druhá hráče vyšší. Pro trénování a validaci byly použity data 386 hráčů, kde každý hráč měl 58 měřených statistik. Výsledkem byl model schopný s 60.1% úspěšností rozhodnout o kvalitě hráče, což však bylo o 6.7% méně než naměřená úspěšnost rekrutujících manažerů.

[3] z anglického learning vector quantization, tato metoda je popsáná [20]
[4] Serie A TIM
[5] vícevrstvý perceptron
[6] i přes to, že se jedná o profesionální soutěže různých, ale velmi podobných sportů, jsou vybrané numerické statistiky pro všechny ligy stejné
[7] myšleno který tým je domácí a který na straně hostů
[8] anglicky draft

V roce 2011 se Samuel Buttrey, Alan Washburn a Wilson Price [5] věnovali tvorbě modelu využívající RATING jednotlivých týmů podle počtu vstřelený a obdržených branek. Tyto RATINGY jsou výsledkem modelu, který předpokládá, že jednotlivé góly jsou z Poissonova rozdělení. Porovnávány jsou pak varianty modelu, kdy jsou RATINGY počítány od začátku sezóny nebo jenom z posledních n dní. Výsledky experimentu ukazují, že ke konci sezóny jsou modely počítané z posledních n dní lepší než model, který je vytvářen z celé sezóny.

V roce 2012 Michael Mauboussin sepsal knihu [14] na téma štěstí ve sportu[9]. V knize popisuje, jak dovednosti jednotlivých hráčů nemají v týmových sportech takovou váhu. Jelikož velmi dobré schopnosti jednoho hráče mohou být velmi rychle vyrovnány nešťastnou chybou někoho ze spoluhráčů. Dále v knize popisuje, kde se jednotlivé sporty pohybují na ose štěstí. Na jedné straně osy jsou sporty/aktivity jejichž výsledek závisí pouze na štěstí např. ruleta, zatímco na straně druhé jsou výsledky, kde štěstí nehraje žádnou roli např. šachy. Většina sportů se pohybuje někde v druhé části osy. Jejich výsledky jsou tak z větší části založené na dovednostech. Jako příklad týmového sportu, který je hodně založený na dovednostech, uvádí autor basketbal, kde jednotliví hráči mají velký počet možností ovlivnit skóre zápasu a tím se tak projeví jejich dovednosti. Naopak jako týmový sport, kde je vliv hráčských dovedností menší, je uveden hokej, ve kterém ani ti nejlepší hráči nehrají více než 25 minut z hodinového zápasu a situací kdy má hráč možnost dát gól je výrazně méně než u zmiňovaného basketbalu.

V roce 2013 A. C. Thomas, Samuel L. Ventura, Shane Jensen a Stephen Ma [21] publikovali článek zabývající se tvorbou RATINGU jednotlivých hráčů. Stochastickou povahu zápasu modelují dvěma soupeřícími procesy o to „kdo dá gól“, kde každý proces obsahuje jak útočné informace tak i informace týkající se obraných schopností. Pro vytváření RATINGU jsou použita data ze sezón 2007/08 až 2011/12. Autoři upozorňují na kariérní výkonnostní křivku hráče, kterou při vytváření modelu neuvažovali. Dále zmiňují, že i velmi dobří hráči mohou mít ve špatných týmech statistiky podobné jako špatní hráči v týmech dobrých.

V roce 2014 Joshua Weissbock publikoval svou práci [22], ve které se věnuje třem hlavním tématům. Velikost trénovací a testovací množiny je celkem 517 zápasů ze sezóny 2012/13. V první části se Weissbock zabýval binární predikcí jednotlivých zápasů, pro každý tým vybral 12 statistik. Jako modely vyzkoušel mnoho různých algoritmů, neuronové sítě, naivní Bayesovský klasifikátor, SMO[10] a C4.5[11]. Pro zjištění přesnosti využil 10-násobnou křížovou validaci. Nejlepších výsledků dosáhla neuronová síť a to s 59.3%. Ve druhé části se věnuje predikci vítěze jednotlivých kol play-off, které se hrají na 4 vítězné zápasy. Jako trénovací data využil výsledky play-off od sezóny 2007/08, což činilo 90 odehraných kol. Autor zmiňuje, že větší vzorek dat by byl lepší, ale vzhledem k počtu kol v play-off v jedné sezóně by musel čekat přinejmenším jednu dekádu. Rozšířil tak alespoň počet statistik pro jednotlivé týmy z původních 12 na 34. I zde využil pro zjištění úspěšnosti modelu 10-násobnou křížovou validaci. Vyzkoušené modely byly stejné,

[9] s autorem byl natočen rozhovor o této problematice dostupný na YouTube
[10] sequential minimal optimization
[11] algoritmus využívající rozhodovací stromy

jako při predikci jednotlivých zápasů. Nejvyšší přesnosti dosahoval model využívající SMO a to 74% přesně. Ve třetí části se autor věnoval kombinaci tří různých klasifikátorů. První klasifikátor jako vstupní vektor využíval 3 statistiky, které nejvíce ovlivňovaly rozhodnutí klasifikátoru z první části. Druhý klasifikátor pracoval s celým textem, jež píší experti před jednotlivými zápasy. Třetí klasifikátor poté využíval četnost pozitivních a negativních slov v těchto textech. Pro rozhodnutí o výsledné třídě porovnal algoritmy největší důvěry, většinové hlasování a kaskádové klasifikace s využití SVM. Nejlepších výsledků dosáhlo většinové hlasování a to 60.25%, tedy oproti samostatné neuronové síti se zlepšila přesnost o téměř 1%.

V roce 2014 Gianni Pischedda ve svém článku [18] opakuje experimenty provedené Weissbockem [22]. Neopakuje však všechny experimenty, ale soustředí se pouze na binární predikci jednotlivých zápasů. Statistiky jednotlivých týmů byly do vstupního vektoru místo pouhého připojení „za sebe“, vloženy ve formě rozdílu statistik mezi týmy. Pro srovnání využívá modely s neuronovou sítí, s rozhodovacími stromy a přidává výsledky programu `ClusteR`, jež je vlastněn sázkovou společností a bližší informace nemohl poskytnout. Výsledky jsou lehce lepší než získal Weissbock [22], což jen potvrzuje jejich správnost a ukazuje, že různá reprezentace stejných dat může vést k rozdílným výsledkům. Nejlépe vychází program `ClusteR`, jenž dosahuje 61.54%. Dále se autor soustředí na využití natrénovaných modelů při sázení.

V roce 2015 Filip Šimsa publikoval svou magisterskou práci [23], ve které se zaměřuje na ekonomicky výhodné sázení na českou hokejovou extraligu. Jako zdroj dat bere statistiky od sezóny 1999/2000 až do sezóny 2014/15. Jako model pak využívá Kalmanův filtr. Velkou část práce věnoval autor popisu nalezených jevů v české extralize, mezi zajímavé jevy patří výsledky různých derby, případně únava ze zápasů hraných v krátkém intervalu.

V roce 2016 Wei Gu, Thomas Saaty a Rozann Whitaker publikovali článek [9], ve kterém se snažili predikovat výsledky jednotlivých zápasů v play-off. Pro predikci využili model ANP[12], který kombinoval 17 statistik, vybraných na základě jejich korelace s výsledky, s odhady expertů. Korelaci statistik zjišťovali na 1230 zápasech ze sezóny 2014/15 s využitím Wilcoxonova rank-sum testu. Výsledný model na 89 zápasech odehraných v play-off sezóny 2014/15 dosáhl úspěšnosti 77.5%.

V roce 2017 Rory P. Bunker a Fadi Thabtah napsali článek [4], ve kterém popisují jak by měl vypadat postup při vytváření modelu predikujícího výsledky kteréhokoliv sportu. V článku je sepsán návod o 6 důležitých bodech tohoto procesu. První bod je porozumění problému predikce a vybraného sportu. Druhý bod je porozumění datům a jejich sběr. Třetím bodem je zpracování dat do správného formátu a rozdělení mezi trénovací a testovací množinu. Zde autoři zdůrazňují, že je důležité si uvědomit, které statistiky jsou známé ještě před zápasem a pro které je znám pouze jejich průměr z předchozích zápasů. Čtvrtým bodem je výběr správného modelu na základě poznatků předchozích prací. Pátým bodem je trénink a hodnocení vybraného modelu. Zde autoři upozorňují na nevhodnost k-násobné křížové validace, jelikož nezachovává časové pořadí zápasů. Jako poslední bod zmiňují automatizaci sběru nových dat a následný trénink na

[12] anglická zkratka pro síť analytických procesů

těchto datech.

V roce 2018 Milton Leung ve svém internetovém článku [13] popisuje vytvoření modelu pro binární predikci výsledku hokejových zápasů. Autorovou hlavní motivací bylo získat model, se kterým by mohl úspěšně sázet. V článku popisuje postup, který se téměř shoduje s navrhovaným postupem Rory P. Bunkera a Fadi Thabtaha [4]. Výsledkem autorova snažení byl model založený na metodách SVM, který dosahoval F1 skóre 67.22% na testovací množině.

V roce 2019 Wei Gu, Krista Foster, Jennifer Shang a Lirong Wei publikovali článek [8], ve kterém popisují tvorbou systému vhodného pro manažery jednotlivých týmů NHL. V první části vytváří RATING hráčů a brankářů jak pro základní část tak i pro play-off. K tvorbě těchto RATINGŮ využívají statistickou metodu PCA, ze které vezmou pouze první 4 komponenty, ty za pomoci váženého průměru vytvoří výsledný RATING. Ve druhé části se zabývají predikcí zápasů nejdříve s modelem využívajícím SVM. Tento model trénuje na 1230 zápasech, ve kterých má vždy pro každý tým k dispozici 19 vybraných statistik. Autoři uvádějí pouze dosaženou přesnost na trénovacích datech, jež činí 94.05%. Dále aplikují Ensemble metody ADABOOST, BOOSTING, BAGGING a ROBUSTBOOST. Jako naivní klasifikátory používají rozhodovací stromy, naivní Bayesovský klasifikátor, SVM, KNN a LDA. Nejlepší kombinace je podle autorů LDA a ROBUSTBOOST, která dosahuje 91.84% na testovacích datech. Autoři bohužel neuvádí způsob výběru dat ani z jakých sezón tato data pocházejí.

Kapitola 3

Data

Tato kapitola se bude zabývat vhodným výběrem dat, následně jejich ziskem a zpracováním.

3.1 Výběr vhodných zdrojů dat

Vzhledem k popularitě hokeje se nabízí hned několik možných zdrojů dat pro učení. Je však nutné zvážit nejen jak moc do historie různé zdroje zasahují, ale také například jakou granularitu statistik poskytují, či v jakých formátech byly zápasy odehrány.

3.1.1 Extraliga ledního hokeje

Jako první se nabízí vybrat českou nejvyšší soutěž, Extraligu ledního hokeje, která vznikla z Československé hokejové ligy po rozdělení Československa v roce 1993. V prvních dvou letech hrálo extraligu pouze 12 týmů, ale od sezóny 1995/96 byl tento počet rozšířen na 14. Podmínkou pro působení v extralize je držení licence[1], kterou klub může získat postupem z nižší soutěže[2], nebo odkoupením od některého z týmu hrajících extraligu. Právě možnost získat licenci postupem z nižší soutěže zapříčinila, že extraligu hrálo již okolo 30 různých klubů, přičemž velká část z nich po postupu v následujícím roce z opět sestoupila.

[1] podmínky pro udělení licence jsou dostupné zde

[2] 1. české hokejové ligy, která je tak druhou nejvyšší hokejovou soutěží v české republice

Tento jev je velmi nežádoucí z hlediska statistik, protože takovéto kluby mají ve statistikách menší počet odehraných zápasů než kluby, které hrají extraligu od jejího založení.

Týmy v základní části mezi sebou hrají formou každý s každým a to celkem čtyřikrát[3]. Prvních 12 nejlépe umístěných týmů postupuje do rozšířeného play-off[4], kde týmy z 5. až 12. místa nejdříve hrají tzv. předkolo play-off na 3 vítězná utkání. Vítězové předkola spolu s prvními čtyřmi týmy ze základní části hrají klasické play-off na 4 vítězné zápasy.

Pravděpodobně nejlepší statistiky ohledně české extraligy poskytuje firma BPA sport marketing, která je výhradním marketingovým partnerem ELH. Ve spolupráci se společností eSports.cz, je také provozovatelem webových stránek hokej.cz, které jsou oficiálním zpravodajským serverem českého hokeje. Na těchto webových stránkách jsou k dispozici statistiky hráčů po jednotlivých zápasech. Statistické záznamy sahají až do roku 1993, kdy extraliga vznikla. Bohužel některé podrobnější statistiky hráčů, jako například zblokované střely nebo čas strávený na ledě, nejsou u starších zápasů dostupné.

Důvodem proč nakonec nebyl využit tento zdroj je kombinace obtížného zisku statistik hráčů v jednotlivých zápasech a fakt že podrobnější statistiky hráčů jsou vedeny až od roku 2013.

3.1.2 Kontinentální hokejová liga

Dalším možným zdrojem dat je Kontinentální hokejová liga, jenž vznikla v roce 2008. Hlavní myšlenkou KHL při jejím vzniku bylo posilnit Ruskou superligu[5] o týmy z Evropy i Asie a pokusit se tak svoji kvalitou vyrovnat severoamerické NHL. Vzhledem k tomu že se nejedná o ligu mistrů, jako například ve fotbale, ale pouze o evoluci Ruské superligy, tak se spousta evropských týmů rozhodla setrvat ve svých národních soutěžích. Počet týmů hrajících KHL se v různých sezónách dosti měnil, nejméně jich nastoupilo 23 v sezónách 2010/11 a 2011/12, naopak nejvíce jich bylo 29 v sezóně 2016/17. Od svého vzniku se již v KHL střetly týmy z více než 9 zemí, avšak stále nejpočetněji zastoupenou skupinou jsou týmy z Ruska. Vysoké zastoupení ruských týmů je z velké části způsobeno tím, že KHL je Ruskou hokejovou federací považována za nejvyšší hokejovou soutěž v Rusku. Ruské týmy mají také oproti ostatním týmům větší finanční jistotu[6], což je pravděpodobně další z důvodů jejich silného zastoupení, protože právě finanční potíže stojí za odchodem mnoha nových týmů zpět do svých národních lig. Peníze ruských týmů jsou také to co umožnilo vznik KHL a stojí za její vysokou kvalitou, spousta z těchto týmů na začátku nabízelo hráčům ze zámoří astronomické sumy peněz za přestup do KHL, vždyť i Jaromír Jágr hrál v KHL 3 roky za ruský Avangard Omsk.

[3] 2x na domácím ledě a 2x na soupeřově ledě
[4] hraje se formou vyřazovacího pavouka
[5] do roku 2008 nejvyšší ruská hokejová liga
[6] za většinou ruských týmů stojí oligarchové

V KHL jsou týmy od její druhé sezóny rozděleny do dvou konferencí, každá konference je pak dále rozdělena na dvě divize.

- Východní konference
  - Černyševova divize
  - Charlamovova divize
- Západní konference
  - Tarasovova divize
  - Bobrovova divize

Počet zápasů, které týmy mezi sebou hrají, se liší podle toho do kterých divizí jsou zařazeny. Pokud jsou týmy ve stejné divizi, pak spolu hrají zápasy 4, pokud jsou z jiné divize, ale stejné konference, pak spolu hrají zápasy 3 a pokud jsou oba týmy z jiné konference, zápasy proti sobě hrají pouze 2. Po dohrání základní části se 8 nejlepších týmů z každé konference utká o Gagarinův pohár ve vyřazovací části.

Opět se jako nejlepší stránka se statistikami jeví oficiální web khl.ru spravovaný společností CHL LLC. Zde je možné dostat se ke statistikám hráčů v jednotlivých zápasech. Záznamy obsahují celou historii KHL, jen jsou bohužel některé podrobnější statistiky vedeny až od sezóny 2014/15.

I přes svoji vysokou kvalitu KHL nebyla vybrána jako zdroj dat. Hlavním důvodem je její relativně krátká historie a také vysoká úmrtnost nových týmů kvůli finanční náročnosti.

3.1.3 Mistrovství světa

Pokud bychom se ve sportu snažili najít místa, kde se střetávají pouze ti nejlepší sportovci z celého světa, pak by mezi tato místa určitě mělo patřit mistrovství světa. Na rozdíl od fotbalového mistrovství světa nebo zimních olympijských her, které se oboje konají jednou za 4 roky, se hokejové mistrovství světa koná každým rokem. Avšak v historii mistrovství světa tomu tak vždy nebylo, ve 20. letech 19. století by titul mistra světa udělován vítězi olympijských her, teprve až od roku 1930 se MS konalo každý rok[7]. V letech kdy se konaly olympijské hry, byl až do roku 1968 prohlašován za mistra světa právě vítěz olympijského turnaje.

[7] v letech 1940 až 1946 se MS nekonalo z důvodu 2. světové války

MS se svou formou velmi odlišuje od klasických ligových soutěží. Týmy jsou rozděleny do 4 základních skupin, ve kterých spolu hrají každý s každým právě jeden zápas. Z každé skupiny postupují 2 nejlepší týmy do vyřazovacího pavouka, kde spolu vždy hrají na 1 vítězný zápas.

Jako zdroj, ze kterého se dají čerpat statistiky ohledně MS, se nabízejí dvě stránky. První z nich je znovu hokej.cz, na kterých je možné dohledat i zápasy z MS 2000, bohužel však až do roku 2014 se jedná pouze o výsledky jednotlivých zápasů, od následujícího roku jsou to pak již základní statistiky na úrovni hráčů v jednotlivých zápasech. Druhou variantou je pak oficiální web archiv IIHF[8], kde jsou dostupné stejné statistiky jako na webu hokej.cz jen v méně uspořádané formě.

Důvodem proč ani mistrovství světa není vhodným zdrojem, je malý počet zápasů, který se každoročně odehraje[9]. Dále je to nízká granularita dat ze starších ročníků MS.

3.1.4 Národní hokejová liga

Rozhodně by ve výběru dat neměla chybět nejznámější a také nejstarší hokejová liga, severoamerická národní hokejová liga zkráceně NHL. Její vznik se datuje do roku 1917, od té doby se nehrála pouze v sezóně 2004/05[10]. NHL si za svou více než stoletou existenci prošla mnohými změnami, z původních 4 týmů se jich stalo 31. Jediný způsob jak se do NHL dostat, je zaplacením vstupního poplatku a schválením rady guvernérů NHL, ve které zasedají zástupci jednotlivých týmů společné s vedením NHL. Na historických záznamech je vidět, že přísné přijímací řízení a platový strop hráčů přináší týmům stabilitu. Za posledních 40 let tak NHL neopustil jediný klub. Pravdou zůstává, že několik klubů změnilo působiště, ale to není nijak závažná věc, protože hráči klubu zůstávají.

S tím jak V NHL přibývali týmy, měnil se i systém jakým spolu hráli, ale základní myšlenka konferencí a divizí zůstávala pořád stejná.

- Východní konference
  - Metropolitní divize
  - Atlantická divize
- Západní konference
  - Pacifická divize
  - Centrální divize

[8] webarchive.iihf.com

[9] při letošním mistrovství to bylo 63 zápasů a to včetně vyřazovacího pavouka, oproti tomu se v ELH odehrálo 364 zápasů a to jen v základní části.

[10] tehdy se zástupci klubů neshodli s představiteli hráčů na podmínkách pro hráče

Týmy které jsou ve stejné divizi, hrají vzájemně nejvíce zápasů. Týmy které spolu nesdílejí divizi, ale pouze konferenci, spolu hrají méně často a týmy které jsou z různých konferencí, hrají nejméně vzájemných zápasů. Tento systém umožňuje hrát každému týmu okolo 80 zápasů v základní části[11] při zachování spravedlnosti soutěže. Do play-off bojovat postupují vždy první 3 týmy z každé divize a další 2 nejlépe umístěné v každé konferenci. Každé kolo play-off se hraje na 4 vítězné zápasy, vítězi celého play-off je pak předán Stanly cup, pro mnohé hráče nejcennější hokejová trofej.

Statistiky o NHL nabízí například portál nhl.cz. Jejich hloubka sahá do sezóny 2014/15. Statistiky jsou ve formě několika webových tabulek. Jednotlivé zápasy se statistikami jsou přístupny přes adresu www.nhl.cz/zapas/*id*, zdali má *id* nějakou strukturu nebo pravidla nebylo zkoumáno. V Tabulce C.1 jsou vypsány statistiky[12], které stránky poskytují.

Další web který nabízí statistiky NHL je nhlportal.cz. Jejich záznamy zažínají sezónou 2005/06. Statistiky ze zápasu jsou opět v několika různých webových tabulkách. Jednotlivé zápasy lze procházet pomocí adresářové cesty /?page=zapasreview&id=*id_zápasu*, kde *id_zápasu* je identifikační číslo zápasu, které se s každým odehraným zápasem zvyšuje, první zápas sezóny 2005/06, tak má číslo 1. V Tabulce C.2 jsou s základní poskytované statistiky.

Statistiky NHL nabízejí také zahraniční portály, jedním z nich je hockey-reference.com. Z webu je možné získat statistiky sahající až na úplný začátek NHL tedy do roku 1917. Přístup k jednotlivým zápasům je možný přes www.hockey-reference.com/boxscores/*id_zápasu*.html, *id_zápasu* je tvořeno z data zápasu ve tvaru RRRRDDMM a třípísmenné zkratky domácího týmu. Statistiky hráčů v jednotlivých zápasech jsou rozdělené do několika webových tabulek. V Tabulce C.3 je pak vidět přehled statistik nabízených stránkami hockey-reference.com.

Pravděpodobně nejrozsáhlejší statistiky nabízí oficiální portál NHL. Na svých stránkách nhl.com nabízí kompletní historii statistik. K hráčským statistikám v jednotlivých zápasech je možné se dostat pomocí filtrů. Nevýhodou tohoto jinak velmi intuitivního a přehledného řešení je, že nelze zobrazit všechny hráčské statistiky do jedné tabulky. Dalším omezením je možnost zobrazit pouze 100 záznamů na jednu stránku a jestliže celkový počet odpovídajících záznamů přesáhne 10000 je možné procházet právě prvních 10000. V Tabulce C.4 jsou vidět některé statistiky, které lze ze stránek získat[13], celý výčet statistik je možné vidět na stránkách `www.nhl.com/stats/glossary`.

Jedním z mála zdrojů, kde lze najít statistiky v jiném formátu něž ve webových tabulkách, je statsapi.web.nhl.com. Na stránky tohoto API se nelze z oficiálních stránek nhl.com nikterak „proklikat“. Alespoň částečnou neoficiální dokumentaci k API sepsal Drew Hynes [10]. Statistiky, které lze získat s pomocí tohoto API, se svoji podrobností vyrovnají statistikám webu nhl.com. V Tabulce 3.1 jsou vidět statistiky, které lze z jednotlivých souborů získat.

[11] podle některých hráčů je tato porce zápasů příliš velká
[12] ve sloupci statistika jsou pouze základní statistiky (nelze dopočítat či odvodit)
[13] web nabízí spoustu odvozených statistik

Statistika	Hráči	Týmy
Čas na ledě	✓[1]	
Góly	✓	✓
Asistence	✓	
Střely	✓[2]	✓[2]
Zákroky	✓[3]	
Trestné minuty	✓	✓
+/-	✓[2]	
Vhazování	✓[1]	
Vyhraná vhazování	✓[1]	✓[7]
Zblokované střely	✓[4]	✓[8]
Získané puky	✓[4]	✓[8]
Ztracené puky	✓[4]	✓[8]
Hity	✓[4]	✓[8]
Střídání		
Počet přesilovek		✓[6]
Čas na ledě v přesilovkách	✓[1]	
Góly v přesilovkách	✓[5]	✓[5]
Asistence v přesilovkách	✓[5]	
Střely v přesilovkách		
Zákroky v přesilovkách	✓[1]	
Počet oslabení		✓[6]
Čas na ledě v oslabení	✓[1]	
Góly v oslabení	✓[5]	✓[5]
Asistence v oslabení	✓[5]	
Střely v oslabení		
Zákroky v oslabení	✓[1]	

[1] od sezóny 1997/98 [2] od sezóny 1959/60
[3] od sezóny 1955/56 [4] od sezóny 2010/11
[5] od sezóny 1933/34 [6] od sezóny 1987/89
[7] od sezóny 2010/11 v % [8] od sezóny 2002/03

Tabulka 3.1: Tabulka s přehledem některých statistik nabízených webem statsapi.web.nhl.com

NHL byla nakonec vybrána jako zdroj dat pro trénování všech modelů. Hlavní výhodou oproti ostatním soutěžím má NHL ve své dlouholeté historii, pečlivě vedeným statistikám a velkému množství zápasů za sezónu.

3.2 Předzpracování dat

Jako konkrétní zdroj dat bylo vybráno API, statsapi.web.nhl.com, vzhledem k jednoduchému přístupu ke statistikám jednotlivých zápasů bez nutnosti použití složitých „web scraperů“. Všechny použité programy pro práci s daty byly psány v jazyce Python.

3.2.1 Sběr dat

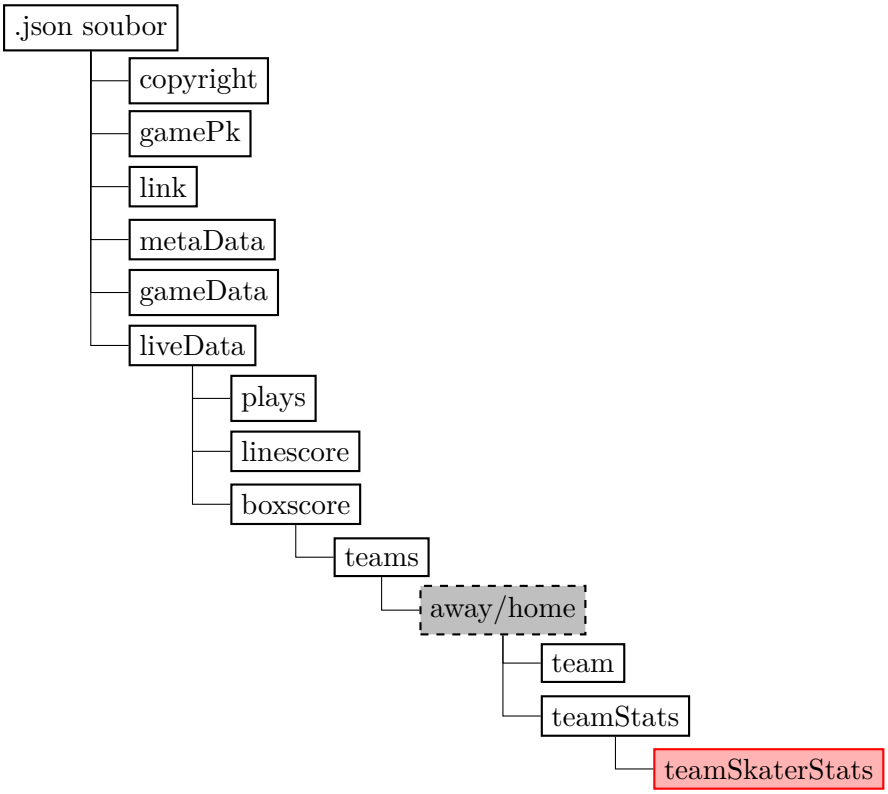
Data ze stránek statsapi.web.nhl.com byly automaticky staženy vlastním programem s využitím knihoven *urllib.request* a *os*. Stahovány byly pouze zápasy ze základní části, kvůli jinému hernímu formátu v play-off. Program si generoval jednotlivá URL zápasů na základě dokumentace od Drew Hynes [10] a následně celý obsah stránek stahoval a ukládal do souborů formátu JSON. V Programu 1 je naznačen jeho průběh. Tímto způsobem bylo staženo něco přes 58000 souborů.

Program 1: Stažení souborů

```
1  for year ← 1917 to 2020 :
2      gameNumber ← 0
3      gamesMissed ← 0
4      while gamesMissed < 4 :              // tolerance na chybějící zápasy
5          gameNumber = gameNumber + 1
6          gameId = createGameId(year, gameNumber)
7          url = createURL(gameId)
8          try:
9              data = URLopen(url)
10             file = createFile(gameId)
11             write(file, data)
12             gamesMissed ← 0
13         except:
14             gamesMissed = gamesMissed + 1
15         end
16     end
17 end
```

3.2.2 Extrakce statistik

Jednotlivé soubory obsahují kromě statistik také informace o jednotlivých týmech, hráčích, ale i událostech na ledě v průběhu zápasu. To způsobuje, že některé soubory obsahují



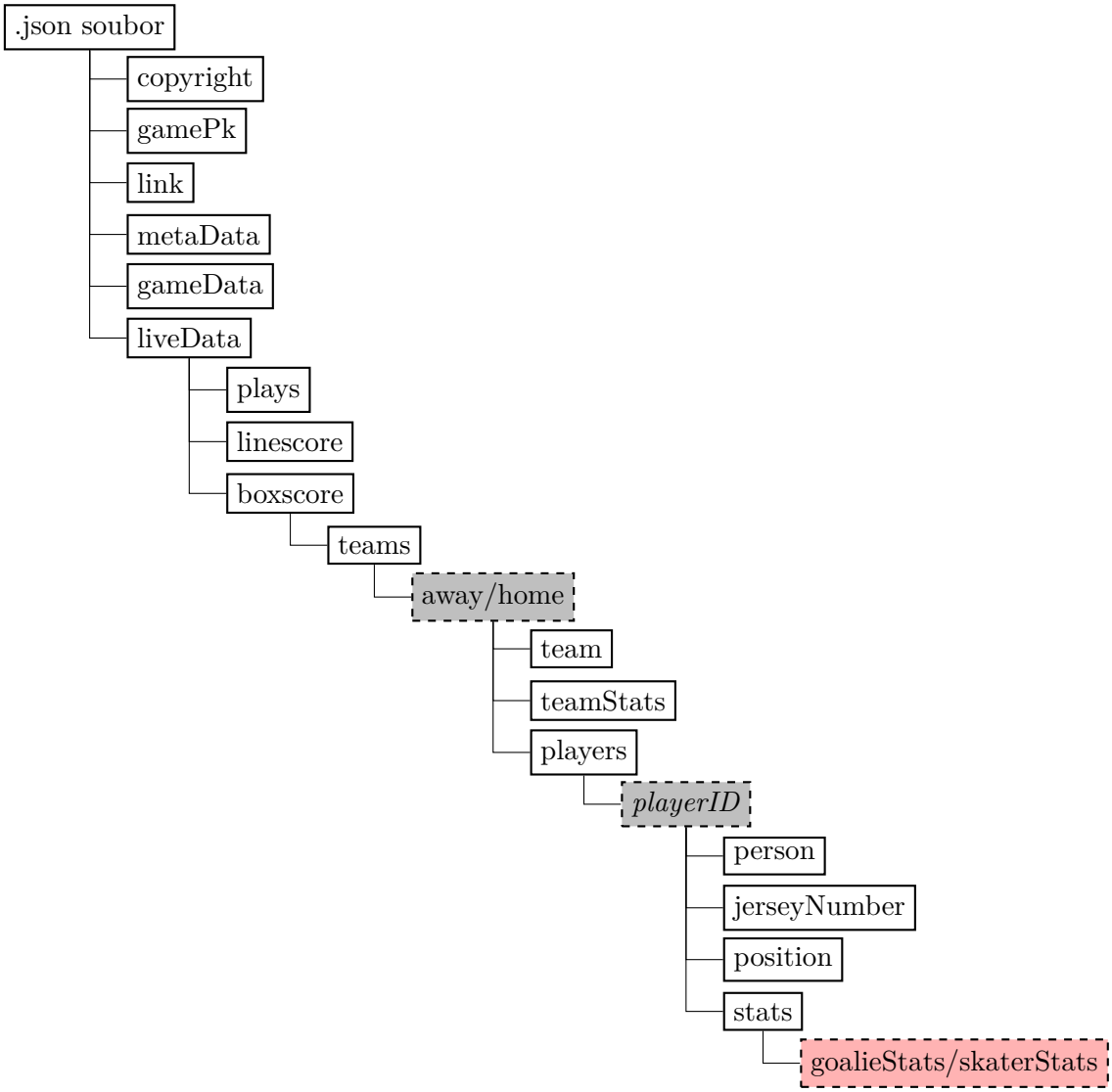
Obrázek 3.1: Cesta k týmovým statistikám

klidně i 20000 řádků a zabírají stovky kB. Celkem stažené soubory zabírají přes 9,6 GB. Pracovat s takovýmito daty je nanejvýš nešikovné a proto bylo učiněno rozhodnutí extrahovat statistiky do rozumnějšího formátu.

Prvním úkolem bylo lokalizovat jednotlivá místa, na kterých se požadované statistiky nachází. Soubory sice mají složitější strukturu, která se v průběhu historie rozrůstala o nové větve a koncové body, ale nikdy se nestalo, že by některé větve nebo koncové body byly v následujících letech přesunuty nebo aby zmizeli. Základní kostra se statistikami tak zůstala po celou historii NHL stejná a proto byla místa se statistikami nalezena ručně. Na Obrázku 3.1 je naznačena cesta ke statistikám domácího a hostujícího týmu. Na Obrázku 3.2 je pak cesta k hráčským statistikám.

S pomocí knihovny *json* byl každý soubor nejprve převeden do slovníku a následně z něj byly vyextrahovány požadované statistiky. Statistiky byly po průchodu každého souboru ukládány do před připraveného objektu `DataFrame` z knihovny *Pandas*. Tento způsob fungoval velmi dobře pro zisk informací o jednotlivých týmech, hráčích a také pro zisk týmových statistik v jednotlivých zápasech. Při zisku statistik hráčů z jednotlivých zápasů, však nastaly komplikace[14], vzhledem k více než 2 milionovému počtu unikátních záznamů došla na výpočetním zařízení paměť RAM. Proto bylo učiněno rozhodnutí,

[14] všechno programování bylo prováděno na notebooku Dell s 8 GB operační paměti RAM bez možnosti dalšího rozšíření



Obrázek 3.2: Cesta k hráčským statistikám

použít objekt `DataFrame` pouze jako mezistupeň a pro výsledné uložení statistik využít databázi *PostgreSQL*. Jednotlivé objekty `DataFrame` byly ze zálohovacích důvodů vždy po 10 letech ukládány a teprve až po uložení byly s pomocí knihovny *sqlalchemy* převáděny do SQL databáze. V Programu 2 je naznačen běh výsledného kódu. Struktura databáze se získanými statistikami je vidět na Obrázku 3.3.

3.2.3 Čištění a úpravy dat

Velké množství numerických záznamů bylo v souborech typu JSON uloženo jako datový typ `string` a bylo je potřeba převést na správný číselný formát. Naštěstí díky mezikroku

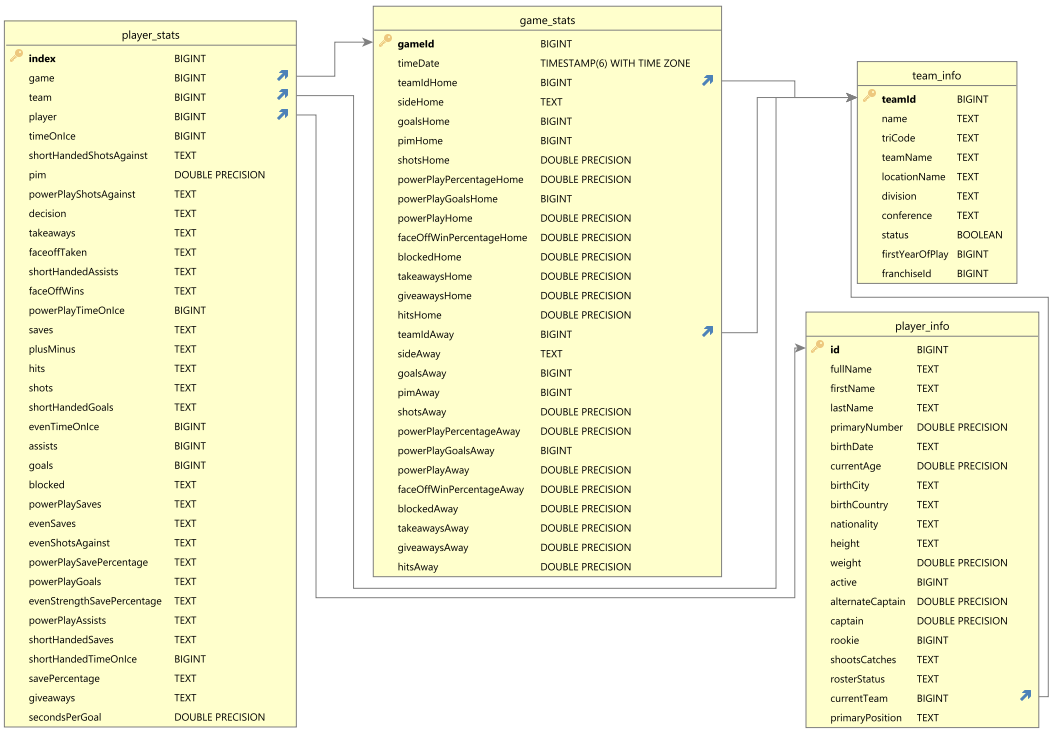
Program 2: Extrakce statistik

```
1  engine = createConnectionToDatabase()
2  keys = [identifiers, wantedStats, ...]
3  df = DataFrame
4  for year ← 2020 to 1917 :
5      gameFiles = listAllGamesInYear(year)
6      for file ← gameFiles :
7          jsonData = openFile(file)
8          try:
9              dictData = convertJson(jsonData)
10         except:
11             reportError
12         end
13         for each instance :                          // team or player
14             stats = {}
15             for key ← keys :
16                 try:
17                     stats[key] = getStat(key)
18                 except:
19                     stats[key] = None
20                 end
21             end
22             df.add(stats)
23         end
24     end
25     if year%10 = 0 then
26         df.toPKL(fineName)
27         df.toSQL(engine)
28         df = DataFrame
29     end
30 end
```

ukládání jednotlivých `DataFramů` byla možnost přistupovat ke statistikám s pomocí knihovny *Pandas*, která nabízí jednoduchou manipulaci s daty. Jednotlivé soubory s uloženými `DataFramy` byly postupně procházeny a upravovány datové typy. Tabulky *player_stats* a *game_stats* byly v databázi nahrazeny jejich obdobami se správnými datovými typy.

Vzhledem k povaze úlohy predikce budoucích zápasů na základě již odehraných utkání, jsou statistiky uložené po jednotlivých zápasech nevhodné. Statistiky z daného zápasu přirozeně nejsou v době jeho predikce ještě k dispozici. Vhodným formátem pro takto časově závislá data[15] je ke každému zápasu kumulovat hráčské i týmové statistiky z již odehraných zápasů. Proto byly do databáze přidány další dvě tabulky, *cumulative_player_stats*

[15] myšleno zápasy



Obrázek 3.3: Struktura databáze s RAW daty

a *cumulative_game_stats*.

V databázové tabulce *cumulative_game_stats* 3.2 však není kumulováno 11 základních statistik, nýbrž se zde nachází jejich průměrné hodnoty z předchozích zápasů doplněné o 5 pokročilých statistik. Každá z těchto 16 průměrových statistik byla rozdělena ještě na 3 další kategorie a to podle toho jestli statistiky byla získána na domácím ledě, na ledě soupeře nebo na obou dohromady.

V databázové tabulce *cumulative_player_stats* 3.3 se nachází 4 typy statistik. Prvním typem jsou přes předchozí zápasy nakumulované základní numerické statistiky z tabulky *player_stats* zobrazené na Obrázku 3.3. Oproti týmovým statistikám zde má cenu tyto kumulované statistiky ponechat, důvodem je, že zkušenosti jednotlivých hráčů nelze jednoznačně vyjádřit. Kumulované statistiky jsou tak tím nejbližším možným vyjádřením zkušeností hráčů. Druhým typem statistik jsou statistiky průměrované vzhledem k odehraným zápasům. Třetí typ statistik jsou také průměrované, ale tentokrát vzhledem k odehraným minutám. Průměr vzhledem k minutám dává o něco podrobnější náhled na schopnosti hráčů. Posledním typem jsou statistiky procentuální. Hráčské statistiky lze rozdělit také podle jiného kritéria a to podle toho jestli se jedná o statistiky, které může získat brankář nebo hráč v poli. Tímto bychom mohli vytvořit dvě nové tabulky, ale vzhledem k částečnému překryvu brankářských statistik s hráčskými bylo toto kroku upuštěno. K rozdělení podle postu hráče docházelo až při načítání dat před trénovací smyčkou.

Celkově[1]	**Doma**[2]	**Venku**[3]
GF_A	GF_A	GF_A
GA_A	GA_A	GA_A
S_A	S_A	S_A
SA_A	SA_A	SA_A
S/G_A	S/G_A	S/G_A
SA/GA_A	SA/GA_A	SA/GA_A
PIM_A	PIM_A	PIM_A
$PPGF_A$	$PPGF_A$	$PPGF_A$
PPO_A	PPO_A	PPO_A
$PP\%_A$	$PP\%_A$	$PP\%_A$
$PKGA_A$	$PKGA_A$	$PKGA_A$
PKO_A	PKO_A	PKO_A
$PK\%_A$	$PK\%_A$	$PK\%_A$
$FOW\%_A$	$FOW\%_A$	$FOW\%_A$
BS_A	BS_A	BS_A
$BS\%_A$	$BS\%_A$	$BS\%_A$
$Hits_A$	$Hits_A$	$Hits_A$
GvA_A	GvA_A	GvA_A
TkA_A	TkA_A	TkA_A

$_A$ značí statistiky průměrné
[1] statistiky získané bez rozdílu hřiště
[2] statistiky získané na domácím hřišti
[3] statistiky získané soupeřivě hřišti

Tabulka 3.2: Statistiky v tabulce *cumulative_game_stats*, jejich vysvětlivky jsou součástí přílohy B

Program 3 zajišťuje tvorbu databázových tabulek *cumulative_player_stats* 3.3 a *cumulative_game_stats* 3.2. Princip programu je pro obě dvě tabulky stejný, liší se pouze ve funkcích, které počítají konkrétní statistiky. Program postupně procházel všechny záznamy a statistiky z nich sčítal do připraveného slovníku s hráči/týmy. Ve slovníku byly kromě kumulovaných statistik ještě vedeny záznamy o počtu výskytu konkrétních statistik pro jednotlivé hráče/týmy. V Programu 3 je naznačen běh programu.

Program 3: Agregace a tvorba pokročilých statistik

```
1  engine = createConnectionToDatabase()
2  cumulativeDict = {}
3  finalDict = { games : [ ], identifiers : [ ], finalStats : [ ] }
4  for data ←loadFromDatabase(engine, player_stats/team_stats) :
5      game, identifier, stats = parseData(data)
6      if identifier not in cumulativeStats then
7          cumulativeDict[identifier] = {cumulativeStats : [zeros],
8                                        frequencyStats : [zeros],
9                                        advancedStats : [zeros]}
10     end
11     finalDict[games].append(game)
12     finalDict[identifiers].append(identifier)
13     cumulativeStats = cumulativeDict[identifier][cumulativeStats]
14     advancedStats = cumulativeDict[identifier][advancedStats]
15     finalStats = [ ]
16     for stat ← cumulativeStats, advancedStats :
17         finalStats.append(stat)
18     end
19     finalDict[finalStats].append(finalStats)
20     i = 0
21     for stat ← stats :
22         if stat != None then
23             cumulativeDict[identifier][cumulativeStats][i] += stat
24             cumulativeDict[identifier][frequencyStats][i] += 1
25         end
26         i += 1
27     end
28     advancedStats = createAdvancedStats(cumulativeDict)
29     cumulativeDict[identifier][advancedStats] = advancedStats
30 end
31 df = DataFrame(finalDict)
32 df.toSQL(engine,cumulative_player_stats/cumulative_team_stats)
```

Hráči				Brankáři			
Celkem	**Za minutu**	**Za hru**	**%**	**Celkem**	**Za minutu**	**Za hru**	**%**
S	S	S	S[1]	SV	SV	SV	SV
A	A	A		S[3]	S[3]	S[3]	
G	G	G	G[2]	SHSA			
PM				SHSV		SHSV	SHSV
TOI		TOI		PPSA			
PIM	PIM	PIM		PPSV		PPSV	PPSV
FOW			FOW	EVSA			
FO				EVSV		EVSV	EVSV
Hits	Hits	Hits	Hits[1]	PIM	PIM	PIM	
GvA	GvA	GvA		TOI		TOI	
TkA	TkA	TkA	TkA[1]	A	A	A	
SHA	SHA	SHA		G	G	G	
SHG	SHG	SHG					
SHTOI		SHTOI					
PPA	PPA	PPA					
PPG	PPG	PPG					
PPTOI		PPTOI					
EVTOI							
BS			BS[1]				

[1] % z týmové statistiky
[2] jako % z týmové statistiky ale také jako G/S
[3] myšleno jako střely na brankáře

Tabulka 3.3: Průměrné statistiky tabulky *cumulative_game_stats*, jejich vysvětlivky jsou součástí přílohy B

Kapitola 4

Modely

Pro vyvození správných závěrů z daných dat je vždy potřeba mít alespoň nějaké základní srovnání. Proto byly vytvořeny kromě modelů využívajících neuronové sítě ještě modely základní.

4.1 Základní modely

Za základní modely byly vybrány modely na principu logistické regrese a modely využívající rozhodovacích stromů, konkrétně metoda náhodného lesa[1].

4.1.1 Logistická regrese

Logistická regrese je binární klasifikační metoda, která určuje pravděpodobnost výskytu dané třídy na základě vstupních dat [3]. Pro určení pravděpodobnosti výskytu nějaké třídy se používá logistická funkce ve tvaru

$$f(x) = \frac{1}{1 + e^{-x}}^{2}. \tag{4.1}$$

[1] známá také pod svým anglickým názvem Random forest

[2] častěji se tomuto tvaru funkce říká sigmoida

Při logistické regresi se Rovnice 4.1 používá ve tvaru s kladnými exponenty a za x se dosazuje skalární součin vektoru vstupních dat $\mathbf{x}$ a vektoru vah $\mathbf{w}$[3]. Ukázka pro příklad binární klasifikace výherce hokejového zápasu:

$$f(\mathbf{x}, \mathbf{w}) = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1 + e^{\mathbf{w}^T\mathbf{x}}} = P(Y(\mathbf{x}) = 1|\mathbf{w}) \in \langle 0, 1\rangle, \tag{4.2}$$

$$\mathbf{x} = [x_1, ..., x_k, x_{k+1}, ..., x_n],$$

kde výraz na levé straně Rovnice 4.2 je pravděpodobnost jevu $Y(\mathbf{x}) = 1$, tedy pravděpodobnost že zvítězí tým domácích při takto nastavených vahách $\mathbf{w}$, x_1 až x_k jsou statistiky domácího týmu a x_{k+1} až a_n jsou statistiky týmu hostů. Pravděpodobnost prohry domácího týmu by byla spočítána jako

$$P(Y(\mathbf{x}) = 0|\mathbf{w}) = 1 - P(Y(\mathbf{x}) = 1|\mathbf{w}) \tag{4.3}$$

Predikovaná třída do které vektor $\mathbf{x}$ patří je pak třída s vyšší pravděpodobností.

Logistickou regresi lze upravit i pro použití při klasifikaci K tříd [3]. Místo toho aby výsledkem byla pravděpodobnost jednoho jevu, bude výsledkem vektor pravděpodobností, kde k-tý prvek výsledného vektoru bude pravděpodobnost že vektor $\mathbf{x}$ je třídy k.

$$\begin{aligned}\mathbf{f}(\mathbf{x}, \mathbf{W}) &= \frac{e^{\mathbf{w}_k^T\mathbf{x}}}{\sum_{i=1}^{K} e^{\mathbf{w}_i^T\mathbf{x}}} = \\ &= \mathbf{P}(y = k|\mathbf{x}, \mathbf{w}_k) = \begin{bmatrix} P(Y(x) = 1|\mathbf{w}_1) \\ P(Y(x) = 2|\mathbf{w}_2) \\ \vdots \\ P(Y(x) = K|\mathbf{w}_K) \end{bmatrix},\end{aligned} \tag{4.4}$$

$$\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]; \quad \mathbf{x}, \mathbf{w}_i \in \mathbb{R}^n.$$

Funkce $\mathbf{f}(\mathbf{x}, \mathbf{W})$ se často nazývá $\texttt{softmax}(\mathbf{x}, \mathbf{W})$. Výslednou třídu $\hat{y}$ pak určí funkce

$$\hat{y} = \arg\max_{\mathbf{x}} \mathbf{f}(\mathbf{x}, \mathbf{W}) \tag{4.5}$$

Pro správnou klasifikaci je nutné aby matice vah $\mathbf{W}$ měla správné hodnoty. Pro odhad těchto hodnot se může použít metoda maximalizace věrohodnostní funkce $\mathcal{L}$ [3].

$$\mathcal{L}(\mathbf{W}) = \prod_{k=1}^{K} P(Y(x) = k|\mathbf{w}_k) \tag{4.6}$$

Častěji se ale jako ztrátová funkce[4] používá záporný logaritmus věrohodnostní funkce.

$$L(\mathbf{W}) = -\log \mathcal{L}(\mathbf{W}) \tag{4.7}$$

[3] někdy se těmto hodnotám říká parametry modelu

[4] Hodnotě $L(\mathbf{W})$ se po anglicku říká Loss

K implementaci logistické regrese byly použity funkce `Linear` a `CrossEntropyLoss` z knihovny *PyTorch*.

4.1.2 Random Forest

Metoda náhodného lesa pro klasifikaci používá výsledky několika různých rozhodovacích stromů, ze kterých je za finální třídu vybrána třída s nejčastějším výskytem [3]. Samotné rozhodovací stromy nejsou příliš dobrými modely, protože jsou pouhým souborem jednoduchých pravidel, které se vytváří při trénování a tím tak mají silnou tendenci k přeučení.

Jednotlivé stromy jsou vytvářeny pomocí velmi jednoduchého iteračního algoritmu.

1. Začni u kořene stromu
2. Pro každý list m, každý atribut x_i a možné rozdělení s, spočítej hodnotu kriteriální funkce
3. Vyber nejlepší možnou kombinaci listu m a rozdělení s podle atributu x_i
4. Opakuj od bodu 2 dokud není splněna ukončující podmínka

Jako kriteriální funkce se velmi často používá Gini index [3]

$$G_n = \sum_{k=1}^{K} \hat{p}_{n,k}(1 - \hat{p}_{n,k}), \tag{4.8}$$

$$\hat{p}_{n,k} = \frac{\#\text{počet instancí třídy } k \text{ v oblasti } n}{\#\text{počet všech instancí v oblasti } n}$$

Ve snaze předejít přeučení modelu jsou stromy vytvářeny na náhodně vybraném vzorku dat[5] z trénovací množiny[6]. Každý strom by tak měl mít unikátní sadu dat na které je vytvářen.

Pokud se však ukáže nějaký atribut jako velmi vhodný indikátor pro rozdělování do tříd, bude mezi predikcemi jednotlivých stromů panovat silná korelace. Pro snížení této korelace využívá metoda náhodného lesa při vytváření stromů ještě náhodného výběru z

[5] opakování je povoleno

[6] anglicky se toto nazývá "bootstrap aggregating"nebo také "bagging"

množiny atributů [3]. Tedy při kroku 2 se nevezmou v potaz všechny atributy $\mathbf{x}$, ale jen některé náhodně vybrané.

Pro implementaci byla zvolena knihovna *scikit-learn*, která mimo jiné nabízí přímo funkci `RandomForestClassifier`.

4.2 Modely využívající neuronové sítě

Hlavními modely byly modely využívající hlubokých neuronových sítí, které lze rozdělit dle typu statistik a granularity vstupních dat. Všechny modely využívající neuronové sítě používali pro výpočet hodnoty ztrátové funkce cross-entropii [3].

4.2.1 Modely s týmovými statistikami

Modelem s nejnižší granularitou dat který se v této kategorii nachází je model, který obdržel pouze identifikátory různých týmů a z nich se snažil predikovat výsledek zápasu. Prvním krokem tohoto modelu bylo vytvoření vektoru atributů z identifikátoru jednotlivých týmů. K tomu se využívá tzv. *embedding* metoda , která není vlastně ničím jiným než součinem maticí učitelných parametrů a tzv. one-hot vektoru [12]. One-hot vektor je pouhým převedením identifikátoru týmu na unikátní vektor, kde i-tý člen vektoru je 1 a ostatní jsou 0.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,I} \\ x_{2,1} & x_{2,2} & \dots & x_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,I} \end{bmatrix} \tag{4.9}$$

$$\mathbf{1}_2 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \end{bmatrix} \tag{4.10}$$

Matice $\mathbf{X} \in \mathbb{R}^{I \times N}$, kde N je počet vytvořených atributů a I je celkový počet identifikátorů. V $\mathbf{1}_2$ je one-hot vektor pro identifikátor týmu s číslem 2. Výstupní vytvořené atributy pro oba týmy jsou spojeny do jednoho vektoru který je vstupem do neuronové sítě.

Další dva týmové modely pracovaly již se statistikami jednotlivých týmů a lišili se od sebe pouze zvolenou podrobností statistik. První z těchto dvou modelů obsahoval pouze statistiky celkové, zatímco druhý kromě celkových statistik obsahoval i statistiky rozdělené podle toho jestli byly domácím nebo hostujícím mužstvem.

Samotná neuronová síť pak je tvořena několika nelineárními vrstvami [3], přičemž pro

každou vrstvu platí že má méně neuronů než ta předcházející:

$$\begin{aligned} H_{l,out} &= \left\lfloor \frac{M-K}{L} \cdot (L-l) + K \right\rceil, \\ H_{l,in} &= H_{l-1,out}, \end{aligned} \tag{4.11}$$

kde $H_{l,in/out}$ je počet vstupních/výstupních atributů l-té vrstvy, K je počet klasifikačních tříd, L je celkový počet lineárních vrstev modelu a M je počet atributů vstupního vektoru. Mezi jednotlivými lineárními vrstvami je na každý výstup aplikována aktivační funkce tanh

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4.12}$$

4.2.2 Modely s hráčskými statistikami

Zatímco modely založené na týmových statistikách byly svou architekturou velmi jednoduché, u modelů s hráčskými statistikami je to o něco složitější. Klasická metoda spojení vektorů atributů „za sebe“ by zde nefungovala už jen z důvodu, že ke každému zápasu nemusí nastoupit stejný počet hráčů. Dále by velkou roli hrálo pořadí, ve kterém by jednotliví hráči byli řazeni do tohoto vektoru.

Proto byl zvolen způsob, kdy se nejdříve jednotlivé hráčské atributy přes hráče v týmu zprůměrují. Tímto tak vznikne jeden týmový vektor $\mathbf{t}$, kde každý element vektoru $\mathbf{t}$ je průměrem jednoho hráčského atributu n.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,N} \\ x_{2,1} & x_{2,2} & \dots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{P,1} & x_{P,2} & \dots & x_{P,N} \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_1, \dots, t_N \end{bmatrix},$$

$$t_n = \frac{\sum_{p=1}^{P} x_{p,n}}{P} \tag{4.13}$$

Matice $\mathbf{X}$ obsahuje atributy hráčů jednoho týmu pro jeden zápas, P je tedy počet hráčů a N je počet jednotlivých atributů. Malá p a n značí konkrétního hráče a konkrétní atribut. Oba vzniklé týmové vektory se teprv až teď mohou spojit a vytvořit tak vstupní vektor do neuronové sítě.

I přesto, že takovýto postup nepotřebuje mít stejný počet hráčů v každém týmu a zápase, je jeho implementace bez využití knihoven *numpy* nebo *pytorch* na uvažovaném množství dat velmi pomalá. Aby se pro tento postup daly využít zmíněné knihovny a jejich velmi rychlé operace s maticemi je nutné doplnit[7] matice $\mathbf{X}$ na stejnou dimenzi.

[7] anglicky se tomuto říká padding

Aby při průměrování nedocházelo k zanášení chyby při doplňování matice $\mathbf{X}$, bylo místo nulových vektorů využito vektoru s průměrnými atributy hráčů daného týmu pro konkrétní zápas. Tyto vektory byly počítány při načítání dat z databáze.

Hráčské modely stejně jako týmové modely mají 3 varianty lišící se v granularitě dat. První model používá pouze identifikátory jednotlivých hráčů a pro vytvoření vektoru atributů využívá embedding. Od embeddingu v týmovém modelu se liší pouze dimenzí matice 4.9. Po embeddingu následovalo klasické průměrování 4.13.

Druhý model využíval pouze statistiky součtové, zde na rozdíl od týmových statistik dávají smysl, jelikož v sobě nesou informaci o zkušenostech hráče. Hráč který strávil v NHL dlouho dobu, bude mít přirozeně tyto statistiky vyšší než nový hráč. Přestože už nemusí dosahovat stejných kvalit jako na vrcholu své kariéry, můžou jeho zkušenosti být klíčovým faktorem ve vyrovnaných zápasech.

Třetí model pak obsahuje součtové statistiky, procentuální statistiky, ale také statistiky průměrné a to jak vzhledem k počtu odehraných zápasů, tak k počtu minut strávených na ledě.

Konvoluční model

Jako alternativní model k výše popisovaným hráčským modelům je vytvořen model, který pomocí parametrizovatelné transformace určuje důležitost jednotlivých statistik hráčů ještě před jejich průměrováním. K tomu se využívá konvoluční vrstvy s M konvolučními jádry o rozměrech $(1, N)$. Při těchto parametrech se konvoluční jádra posouvají maticí $\mathbf{X}$ pouze ve směru hráčů. Každý vektor hráče $\mathbf{x}_p$ je tak vynásoben M konvolučními jádry,čímž vznikne M nových atributů.

$$\hat{\mathbf{x}}_p = \mathbf{x}_p \cdot \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_M \end{bmatrix}^T \tag{4.14}$$

$$\hat{\mathbf{x}}_p \in \mathbb{R}^M, \mathbf{x}_p \in \mathbb{R}^N, \mathbf{w}_m \in \mathbb{R}^N, M < N$$

Konvoluční vrstva se nachází ještě před průměrováním. Jednotlivé parametry konvolučních jader se pak aktualizují se zbytkem sítě. Důležitou podmínkou je využití funkcí z jedné knihovny např. *pytorch* při všech krocích (konvoluce, průměrování, neuronová síť). V opačném případě by automatický výpočet gradientů, použitých algoritmem pro zpětné šíření chyby[8], byl přerušen.

[8] anglicky backpropagation

Kapitola 5

Experimenty

V této kapitole jsou popsány a diskutovány výsledky experimentů.

5.1 Experimentální protokol

Jednotlivé experimenty jsou prováděny pouze na datech ze sezón 2002/03 až 2019/20. Za tímto rozhodnutím stojí výskyt některých podrobnějších statistik. Sice nejpodrobnější nabízené statistiky jsou až od sezóny 2010/11, ale ty byly obětovány ve prospěch většího množství použitelných dat. Sezóna 2020/21 v době sběru dat nebyla ještě dohraná, což by mohlo ovlivnit „férovost“ dat, proto nebyla sezóna 2020/21 zahrnuta.

Vzhledem k časové závislosti jednotlivých zápasů bylo nutné při rozdělování dat, dodržet jejich chronologické uspořádání. Proto byly jako testovací data zvoleny zápasy z poslední uvažované sezóny 2019/20. Testovací data byly po celou dobu trénování a ladění hyperparametrů drženy stranou. Zbylých 16[1] sezón bylo využíváno jako trénovací a validační data.

Při načítání byla data rozdělována do skupin po sezónách, načež se sezóny řadily do posuvného okna o velikosti w, kde prvních $w-1$ sezón bylo bráno jako trénovací a poslední sezóna w jako validační. Metoda s posuvným oknem byla použita z důvodu nemožnosti výběru náhodného vzorku pro validaci, při rozdělování na trénovací a validační data. Pokud by byly za validační data vybrány náhodné zápasy, tak by docházelo k predikci „minulosti“ na znalostech z „budoucnosti“. Trénink na jednotlivých iteracích posuvného okna probíhal sekvenčně. Začínalo se na okně, které obsahovalo nejstarší zápasy a ve chvíli,

[1] i přes to že se jedná o 17 let tak v sezóně 2004/05 byla v NHL výluka a neodehrál se jediný zápas

kdy byly splněny podmínky pro ukončení tréninku se okno posunulo o jeden rok dále. Výsledky z jednotlivých epoch a iterací posuvného okna byly zaznamenávány do objektu `DataFrame`, ten byl uložen do souboru ve formátu Pickle a společně s natrénovaným modelem uložen ve vlastním adresáři.

O zastavení trénovací smyčky se starala automatická funkce[2], jejímž cílem bylo správně vystihnou okamžik, kdy je nalezeno „dobré“ lokální minimum ztrátové funkce L a zabránit tak přeučení na trénovacích datech. Funkce za tímto účelem využívala frontu o velikosti c, do které po projití všech trénovacích a validačních zápasů v dané iteraci posuvného okna ukládala hodnoty ztrátové funkce L získané na validačních datech. Po každé takovéto epoše spočítala průměrnou hodnotu fronty a první člen fronty byl odebrán. Funkce porovnává tento průměr s nejnižším zaznamenaným průměrem pro dané posuvné okno a pokud se tato hodnota po s epochách nezlepší, je trénování na dané iteraci posuvného okna ukončeno. Následuje uložení všech nasbíraných statistik a posunutí okna o jednu sezónu vpřed. Výsledná kombinace parametrů s a c zastavující funkce byla zvolena po obsáhlém testování na různých kombinacích metaparametrů. Parametry s a c byly kompromisem mezi nízkými hodnotami, které zajišťovaly ukončení trénování před nějakým vyšším nárůstem validační chyby, ale zase nedávaly moc velkou šanci k překročení lokálních minim, s hodnotami vyššími, které umožňovaly některým modelům dosáhnout lepších výsledků, ale u jiných modelů naopak nezastavili trénování včas a došlo k přeučení modelu.

Ještě před spuštěním trénovací smyčky byla data standardizována. Na každou statistiku n byla puštěna standardizační funkce [17].

$$\mathbf{x}'_n = \frac{\mathbf{x}_n - \mu_n}{\sigma_n}, \tag{5.1}$$

kde μ_n je aritmetický průměr statistiky n přes celé posuvné okno a σ_n je její směrodatná odchylka.

$$\mu_n = \sum_{p=1}^{P} x_{p,n}, \tag{5.2}$$

$$\sigma_n = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(x_{p,n} - \mu_n)^2}, \tag{5.3}$$

kde malé p značí jednotlivé výskyty atributu n a velké P je celkový počet výskytů.

Pro hledání ideálních metaparametrů, bylo využito výpočetních serverů MetaCentrum, na kterých byly spuštěny nejrůznější kombinace[3] parametrů. Celkem takto bylo spuštěno 1360 různých kombinací. Na servery byla také kromě výpočtů přesunuta i databáze se statistikami.

[2] anglicky se tato funkce nazývá early stopping

[3] pomocí klasické metody grid-search

Sezóna	Zápasy	Domácí		Hosté		Remízy	
		Počet	%	Počet	%	Počet	%
2002/03	1230	597	48,5	476	38,6	157	12,7
2003/04	1230	590	47,9	470	38,2	170	13,8
2005/06	1230	631	51,3	454	36,9	145	11,7
2006/07	1230	598	48,6	468	38,0	164	13,3
2007/08	1230	584	47,4	490	39,8	156	12,6
2008/09	1230	610	49,5	461	37,4	159	12,9
2009/10	1230	599	48,6	447	36,3	184	14,9
2010/11	1230	580	47,1	501	40,7	149	12,1
2011/12	1230	605	49,1	444	36,0	181	14,7
2012/13*	720	358	49,7	265	36,8	97	13,4
2013/14	1230	578	46,9	474	38,5	178	14,4
2014/15	1230	581	47,2	479	38,9	170	13,8
2015/16	1230	592	48,1	531	43,1	107	8,6
2016/17	1230	636	51,7	495	40,2	99	8,0
2017/18	1271	650	51,1	518	40,7	103	8,1
2018/19	1271	637	50,1	547	43,0	87	6,8
2019/20**	1082	540	49,9	456	42,1	86	7,9

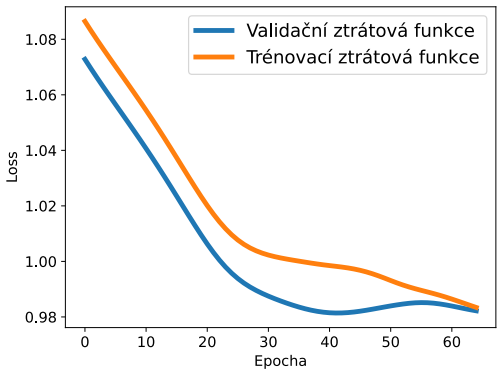
* Kvůli výluce začala sezóna až 19. ledna

** Kvůli viru Covid-19 byla sezóna 12. března ukončena

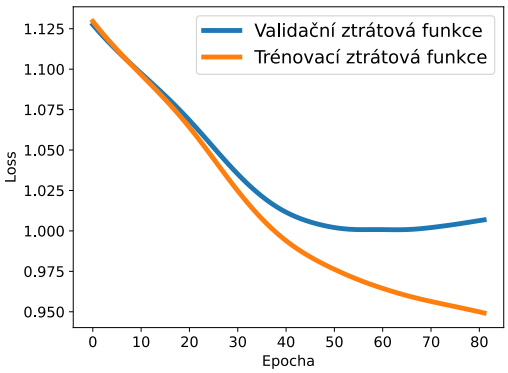
Tabulka 5.1: Tabulka četnosti výskytu jednotlivých tříd.

5.2 Analýza

Při pohledu na zastoupení jednotlivých tříd v uvažovaných sezónách viz Tabulka 5.1 je vidět vývoj NHL. Prvního čeho si je potřeba všimnout, je nerovnoměrné zastoupení jednotlivých tříd. Jako druhé je možné si povšimnout snižujícího se procenta remíz v jednotlivých sezónách. Tyto dva jevy mají za následek, že při některých iteracích posuvného okna jsou validační data výrazně jednoduší než data trénovací. Obsahují totiž nižší procento remíz. Predikce této třídy je pro modely, díky svému nízkému zastoupení v celé množině dat, obtížná. Nastane tak situace, kdy se ztrátová funkce na validačních datech po celou dobu trénování pohybuje pod hodnotou ztrátové funkce trénovacích dat. Tento jev je možno vidět na Obrázku 5.1a, kde posuvné okno o velikosti 2 trénuje na sezóně 2009/10 a validuje na sezóně 2010/11. Pokud však okno posuneme o rok dál, graf se ztrátovými funkcemi vypadá tak, jak bychom jej čekali viz Obrázek 5.1b.



(a) : trénovací sezóna 2009/10, validační sezóna 2010/11



(b) : trénovací sezóna 2010/11, validační sezóna 2011/12

Obrázek 5.1: Validační a trénovací ztrátová funkce na posuvném okně velikosti 2.

Parametr	**Hodnoty**
Typ statistik	Hráčské, Týmové
Granularita dat	Nízká, Střední, Vysoká
Embedding*	10
Velikost posuvného okna	2
Počet vrstev modelu	5
Zvýhodnění novějších zápasů	0
Konvoluční model	Ne
Resetování parametrů**	Ano

* Použito pouze pro model s nejnižší granularitou

** Resetování parametrů modelu při posunu posuvného okna

Tabulka 5.2: Tabulka s hodnotami metaparametrů pro experiment zaměřený na granularitu dat.

5.3 Granularita

Cílem tohoto experimentu je rozhodnout, zda vyšší granularita statistik zlepšuje predikci hokejových zápasů. Při tomto experimentu se mění pouze metaparametry granularita dat a typ statistik. Ostatní metaparametry jsou zafixovány na svých základních hodnotách. Hodnoty je možno vidět v Tabulce 5.2.

5.3.1 Porovnání týmových modelů

Modelem využívající statistiky s nejnižší granularitou celkově je právě týmový model používající pouze identifikátory týmů. Model se střední granularitou dat využíval statistiky celkové a model s vysokou granularitou využíval všechny dostupné týmové statistiky. Na Obrázku 5.2 je graf porovnávající tyto tři modely. Sytěji je vynesena přesnost modelů a bledě je vynesena hodnota ztrátové funkce.[4]

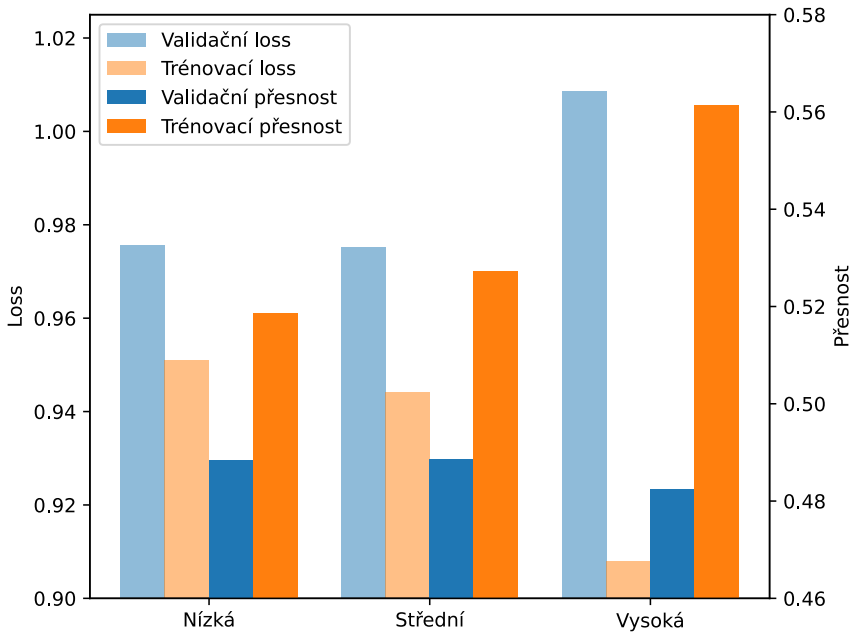
Z grafu na Obrázku 5.2 je vidět, že model s vysokou granularitou dat je oproti ostatním modelům horší. Toto může být způsobeno stylem vytváření jednotlivých modelů, kdy počet statistik, které vstupují do modelu, ovlivňuje celkový počet neuronů[5] v jednotlivých vrstvách modelu. U modelu s vysokou granularitou dat tak dochází přeučení. Zastavující funkce, která tomuto jevu měla předcházet, tak bohužel „nestihne“ zastavit proces učení včas. Modely s nízkou a střední granularitou se zde od sebe nijak výrazněji neliší, a dodané statistiky zde tedy zjevně nepřinášejí novou informaci oproti samotným identifikátorům týmů.

5.3.2 Porovnání hráčských modelů

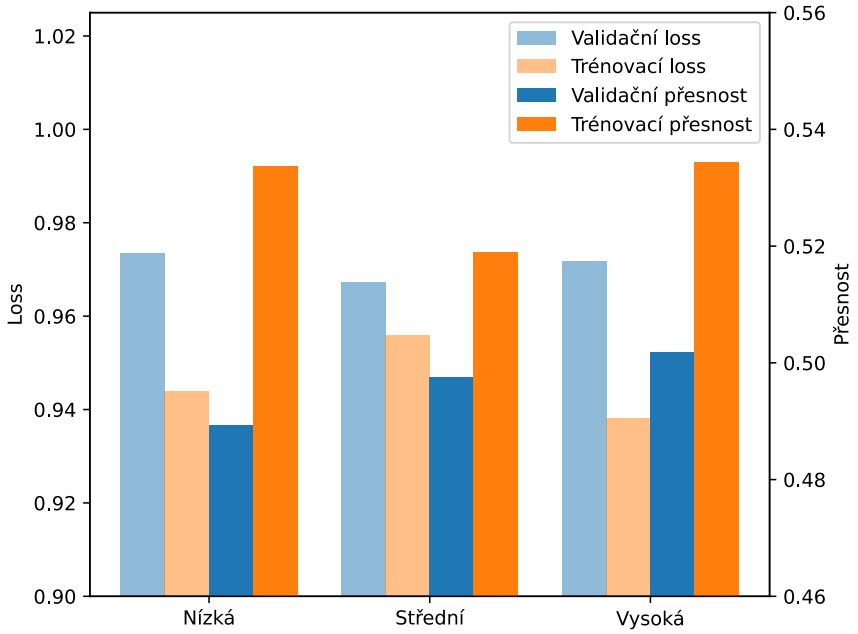
O něco vyšší granularitu statistik přináší hráčské modely, i zde model s nízkou granularitou využívá metod embeddingu. Model se střední granularitou dat využívá statistiky součtové a model s vysokou granularitou využívá všechny dostupné hráčské statistiky. Na Obrázku 5.3 je graf srovnávající hráčské modely. Sytěji je opět vynesena přesnost a bledě hodnota ztrátové funkce.

Z grafu na Obrázku 5.3 je vidět, že zde model s nízkou granularitou na ostatní modely začíná trochu ztrácet. Zajímavé porovnání nabízí model se střední a vysokou granularitou, kdy první zmíněný model dosahuje nejlepší validační chyby, ale má horší validační přesnost než model s vysokou granularitou. Toto může být zapříčiněno způsobem výpočtu validační loss a validační přesnosti. Model se střední granularitou, tak mohl sice chybně určit více zápasů, ale v těchto zápasech si svou předpovědí nebyl moc jistý[6], zatímco model s vysokou granularitou chyboval méně často, ale ve svých chybných predikcích si byl svým výsledkem docela jist. Model s vysokou granularitou navíc projevuje známky lehkého přeučení, ale rozhodně ne takového jakého dosáhl týmový model s vysokou granularitou.

[4] v grafu použit anglický termín loss
[5] podrobněji popsáno v 4.11
[6] výstupní hodnoty poslední vrstvy neuronů si byly navzájem podobné



Obrázek 5.2: Přesnost a hodnota ztrátové funkce pro modely s týmovými statistikami.



Obrázek 5.3: Přesnost a hodnota ztrátové funkce pro modely s hráčskými statistikami.

5.4 Porovnání týmových a hráčských modelů s modely základními

5.4.1 Nízká granularita

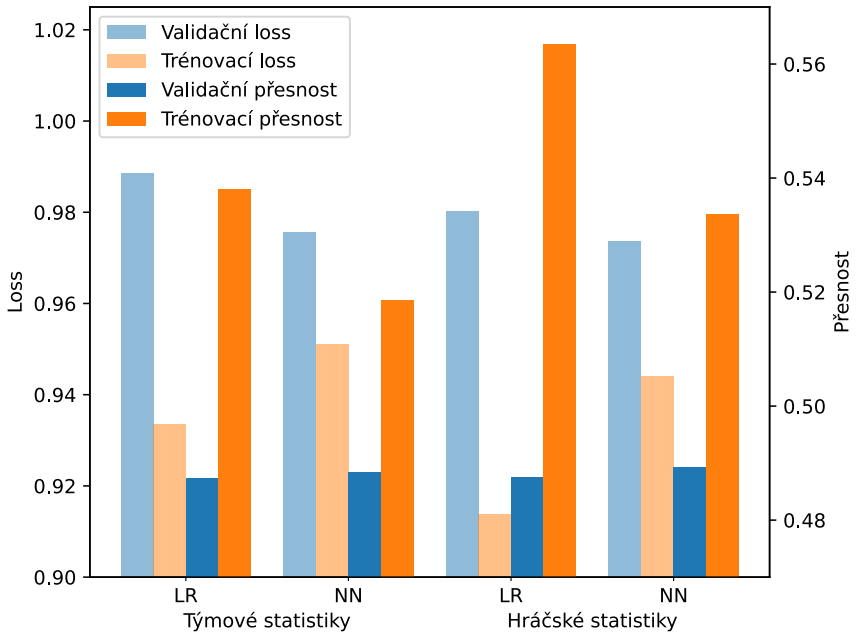
Ve srovnání modelů s nízkou granularitou chybí modely využívající metodu náhodného lesa a to z důvodu způsobu učení, který neumožňuje využít metod embeddingu. Na Obrázku 5.4 je graf porovnávající jednotlivé modely. Z grafu je vidět, že neuronová síť dosahuje na datech s nízkou granularitou o něco lepších výsledků než logistická regrese. Při porovnání modelů neuronových sítí mezi sebou není vidět téměř žádný rozdíl, toto ukazuje že při vytváření embeddingu pro neuronové sítě nehraje roli, zda jsou k dispozici pouze týmové identifikátory či identifikátory jednotlivých hráčů.

5.4.2 Střední granularita

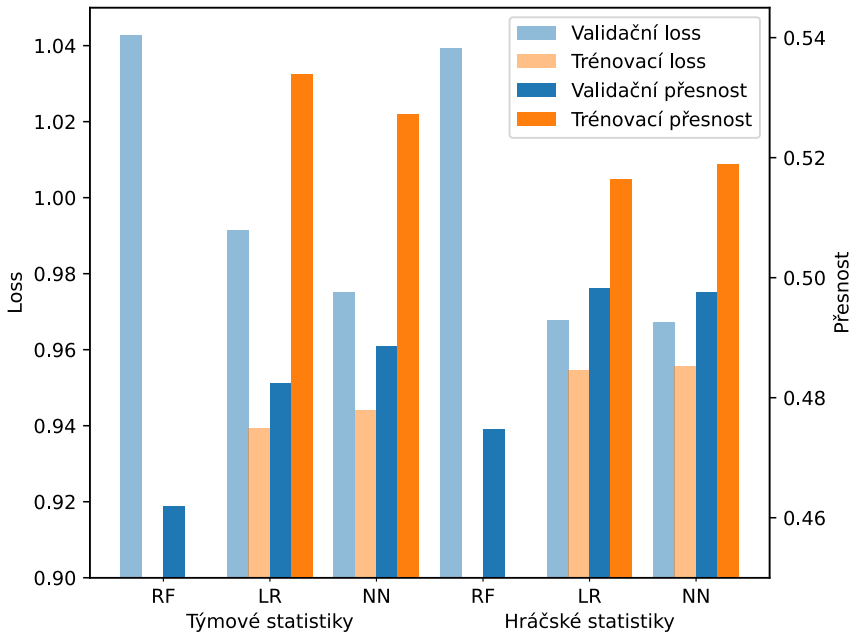
V tomto porovnání se již nacházejí všechny modely. Obrázek 5.5 zobrazuje graf s naměřenými hodnotami přesnosti a loss. V grafu chybí hodnoty trénovací loss a trénovací přesnosti pro metodu náhodného lesa, u které nemají vzhledem k jejímu tréninku tyto veličiny žádnou vypovídající hodnotu. S predikcí si nejhůře vedla právě metoda náhodného lesa, nejlépe pak logistická regrese a neuronoví síť pro hráčské statistiky.

5.4.3 Vysoká granularita

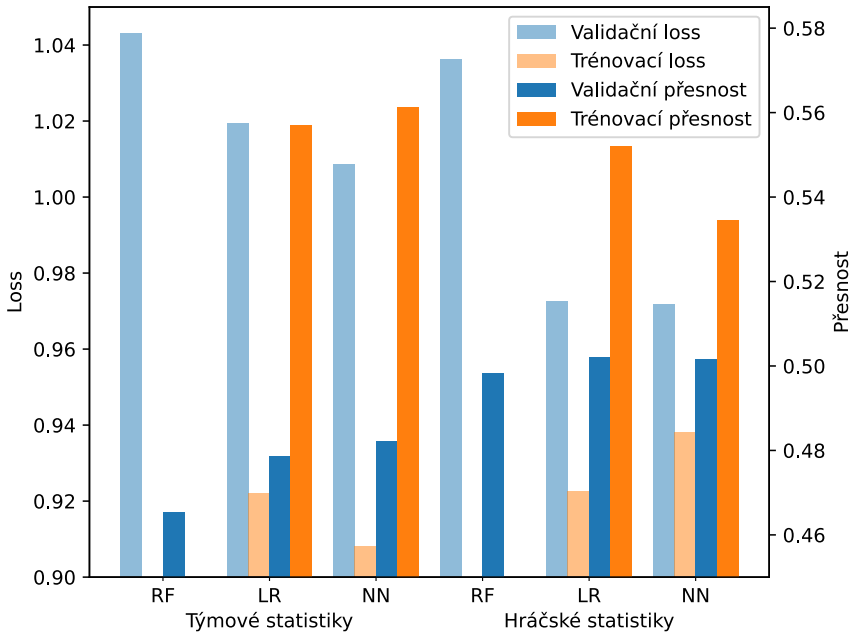
V porovnání modelů s vysokou granularitou dat jsou opět všechny vytvořené modely. Obrázek 5.6 zobrazuje graf tohoto porovnání. Zde nejlepších výsledků opět dosahuje neuronová sít a logistická regrese s hráčskými statistikami. Na tomto porovnání je dobře vidět, jak u modelu logistické regrese a neuronové sítě využívající týmové statistiky dochází k výraznému přeučení. To je opět způsobeno formou vytváření modelu viz Kapitola 4 a faktem že jednotlivých týmových statistik bylo vytvořeno více než těch hráčských.



Obrázek 5.4: Srovnání modelů s nízkou granularitou dat. LR - logistická regrese. NN - neuronová síť.



Obrázek 5.5: Srovnání modelů se střední granularitou dat. RF - random forest. LR - logistická regrese. NN - neuronová síť.



Obrázek 5.6: Srovnání modelů s vysokou granularitou dat. RF - random forest. LR - logistická regrese. NN - neuronová síť.

5.4.4 Diskuze

V Zabulce 5.3 jsou vidět výsledky provedených experimentů se zafixovanými hodnotami z tabulky 5.2. I přesto, že rozdíly v některých případech jsou pouze v řádech desetin procenta, vždy lépe vyšly modely využívající hráčské statistiky viz Obrázky 5.5 a 5.6. V porovnání mezi jednoltivými typy modelů vyšla metoda random forest nejhůře, to může být způsobeno nevhodným naladěním hyperparametrů této metody, kterému nebylo z časových důvodů věnováno více pozornosti. Vítěze mezi logistickou regresí a neuronovou sítí z takto těsných výsledků nelze s jistotou určit. Díky velmi podobnému principu těchto modelů, můžou tyto výsledky naznačovat, že využití hlubokých neuronových sítí nemusí vždy přinášet lepší výsledky.

5.5 Velikost modelu

Jak již naznačil experiment s granularitou dat, mohlo by se zdát, že velké neuronové sítě nejsou tím správným řešením pro predikci hokejových zápasů. Proto jsou v rámci tohoto experimentu porovnávány různé hodnoty metaparametrů ovlivňující velikost sítě.

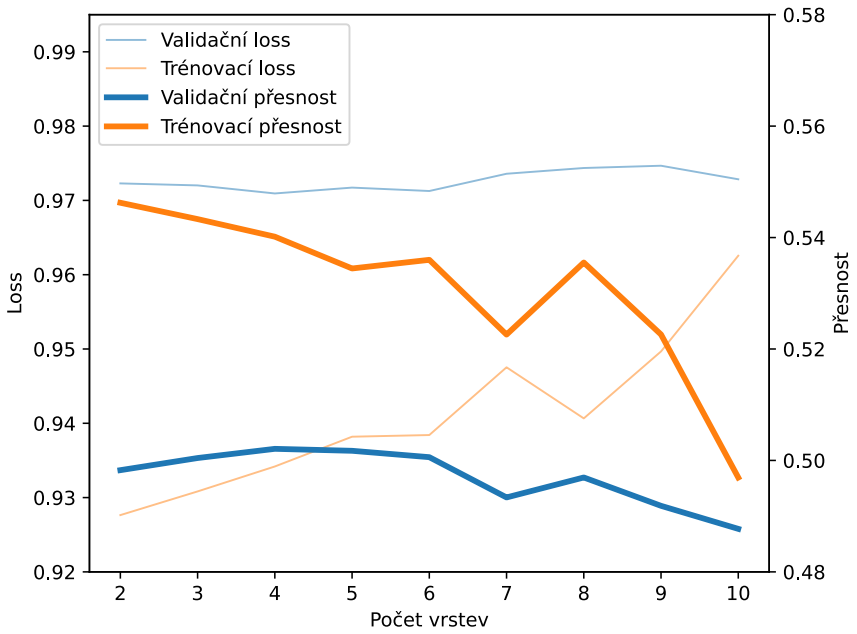
Typ statistik	Granularita	Model	Loss		Přesnost	
			trn	val	trn	val
Týmové	Vysoká	RF		1.043		46.5%
Hráčské	Vysoká	RF		1.036		49.8%
Týmové	Střední	RF		1.043		46.2%
Hráčské	Střední	RF		1.039		47.5%
Týmové	Vysoká	LR	0.922	1.019	55,7%	47.9%
Hráčské	Vysoká	LR	0.923	0.973	55.2%	**50.2%**
Týmové	Střední	LR	0.939	0.991	53.4%	48.2%
Hráčské	Střední	LR	0.955	0.968	51.6%	49.8%
Týmové	Nízká	LR	0.933	0.989	53.8%	48.7%
Hráčské	Nízká	LR	0.914	0.980	**56.4%**	48.8%
Týmové	Vysoká	NN	**0.908**	1.009	56.1%	48.2%
Hráčské	Vysoká	NN	0.938	0.972	53.4%	**50.2%**
Týmové	Střední	NN	0.944	0.975	52.7%	48.9%
Hráčské	Střední	NN	0.956	**0.967**	51.9%	49.8%
Týmové	Nízká	NN	0.951	0.976	51.9%	48.8%
Hráčské	Nízká	NN	0.944	0.974	53.4%	48.9%

Tabulka 5.3: Tabulka s naměřenými údaji při experimentech s granularitou. RF - random forest. LR - logistická regrese. NN - neuronová síť

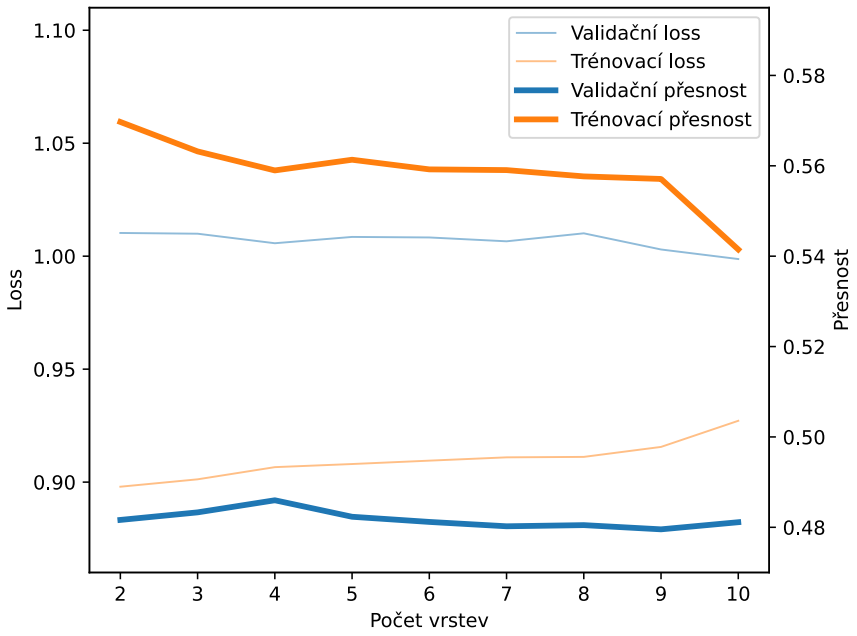
5.5.1 Počet vrstev

První metaparametr ovlivňující velikost sítě je počet jednotlivých vrstev. Jelikož se v Experimentu 5.3 ukázalo jako nejlepší využívat statistiky s vysokou granularitou, je tento experiment prováděn pouze na modelech této granularity. Ostatní metaparametry budou zafixovány na základních hodnotách viz Tabulka 5.4.

Z grafu na Obrázku 5.7 je vidět, že pro model využívající hráčské statistiky s vysokou granularitou je nejlepších výsledků dosaženo při použití 4 vrstvev. Dále je z grafu vidět, jak se s přibývajícími vrstvami přesnost modelu snižuje a to jak validační, tak trénovací. Někoho by mohl tento výsledek zarazit, protože očekával, že na modelu s více vrstvami se bude projevovat přeučení, tedy zvyšující se trénovací a snižující se validační přesnost, k tomu však nedochází z důvodů využití zastavovací funkce viz její popis v Sekci 5.1. Na Obrázku 5.8 je graf zobrazující výsledky tentokrát pro model s týmovými statistikami. I zde model dosahoval nejlepších výsledků při použití 4 vrstev.



Obrázek 5.7: Vývoj přesnosti modelu hráčských statistik s vysokou granularitou v závislosti na počtu vrstev.



Obrázek 5.8: Vývoj přesnosti modelu týmových statistik s vysokou granularitou v závislosti na počtu vrstev.

Parametr	Hodnoty
Typ statistik	Hráčské, Týmové
Granularita	Vysoká
Velikost posuvného okna	2
Počet vrstev modelu	2, 3, 4, 5, 6, 7, 8, 9, 10
Zvýhodnění novějších zápasů	0
Konvoluční model	Ne
Resetování parametrů*	Ano
* Resetování parametrů modelu při posunu posuvného okna	

Tabulka 5.4: Tabulka s hodnotami metaparametrů pro experiment zaměřený na počet vrstev.

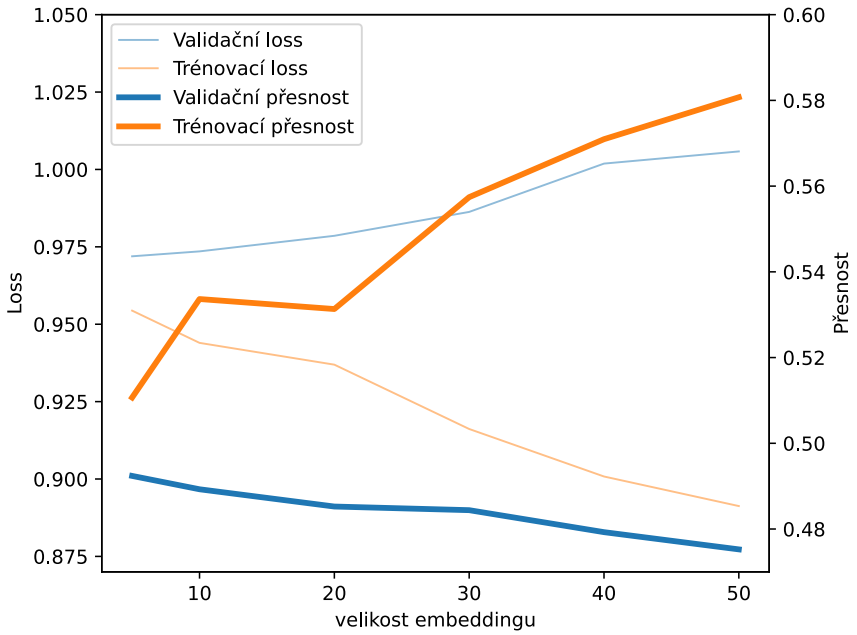
Parametr	Hodnoty
Typ statistik	Hráčské, Týmové
Granularita	Nízká
Embedding	5, 10, 20, 30, 40, 50
Velikost posuvného okna	2
Počet vrstev modelu	5
Zvýhodnění novějších zápasů	0
Konvoluční model	Ne
Resetování parametrů*	Ano
* Resetování parametrů modelu při posunu posuvného okna	

Tabulka 5.5: Tabulka s hodnotami metaparametrů pro experiment zaměřený na velikost embeddingu.

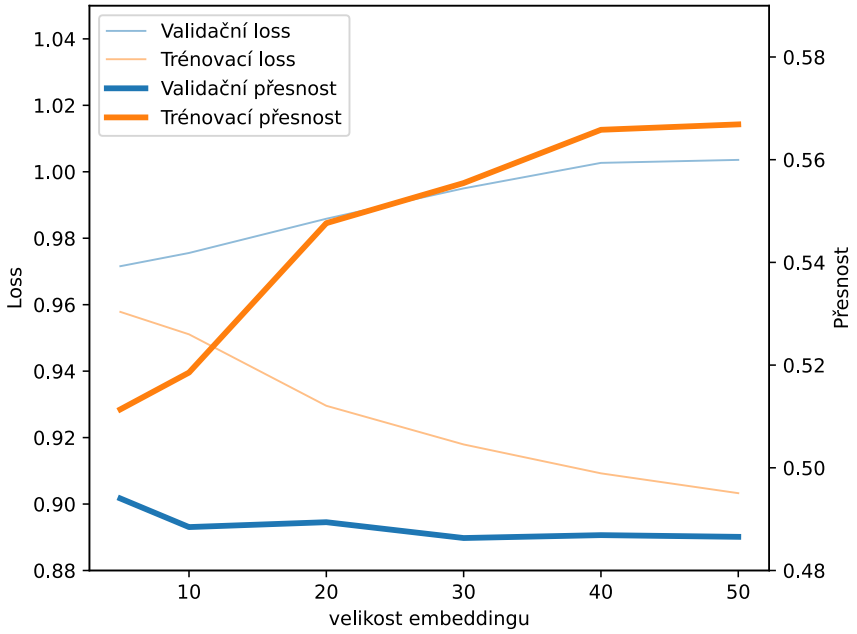
5.5.2 Velikost embeddingu

Velikost embeddingu je velmi důležitý parametr pro případy, kdy nejsou dostupné žádné statistiky a je tak nutné vycházet pouze z jednotlivých identifikátorů. Při tomto experimentu je využito modelu s nízkou granularitou dat. Cílem tohoto experimentu je zjistit, zdali při těchto podmínkách, hraje nastavení velikosti embeddingu nějakou výraznější roli. V průběhu experimentu jsou další metaparametry zafixovány na hodnotách z Tabulky 5.5.

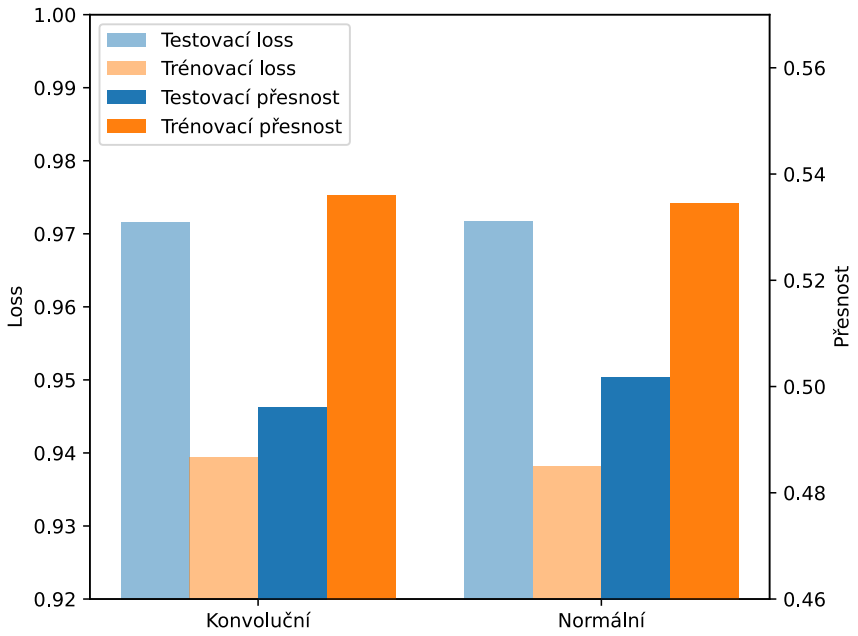
Na Obrázku 5.9 je graf získaných hodnot pro model, který využívá hráčské identifikátory. Z grafu je vidět, že se zvyšujícím se počtem vytvořených atributů, pomocí embeddingu, dochází k většímu přeučení. Validační loss je stále větší zatímco trénovací loss se daří modelu snižovat. Na Obrázku 5.10 jsou na graf vyneseny naměřené hodnoty pro model s týmovými identifikátory. I zde docházelo k většímu přeučení modelu s větším počtem vytvořených atributů.



Obrázek 5.9: Vývoj přesnosti hráčského modelu s nízkou granularitou v závislosti na velikost embeddingu.



Obrázek 5.10: Vývoj přesnosti týmového modelu s nízkou granularitou v závislosti na velikost embeddingu.



Obrázek 5.11: Graf přesnosti hráčského normálního a konvolučního modelu s vysokou granularitou.

5.5.3 Konvoluční hráčský model

Jelikož konvoluční model pro hráčské statistiky mění jeho velikost, byl experiment s tímto modelem zařazen do Sekce 5.5. Při tomto experimentu se redukoval počet hráčských statistik ještě před průměrováním, ale zároveň tím přibyly parametry konvolučních filtrů. Tento pokus nebyl původní součástí plánovaných experimentů a proto byl z časových důvodů natrénován pouze jeden model. Počet hráčských statistik byl v modelu konvoluční sítí zredukován na třetinu. Hodnoty ostatních fixních metaparametrů jsou v Tabulce 5.6.

Na obrázku 5.11 jsou zobrazeny výsledky tohoto experimentu. Bohužel výsledky jsou si příliš podobné a více kombinací metaparametrů by bylo potřeba pro rozhodnutí o výhodnosti využití tohoto konvolučního modelu.

Parametr	Hodnoty
Typ statistik	Hráčské
Granularita	Vysoká
Velikost posuvného okna	2
Počet vrstev modelu	5
Zvýhodnění novějších zápasů	0
Konvoluční model	Ano, Ne
Resetování parametrů*	Ano

* Resetování parametrů modelu při posunu posuvného okna

Tabulka 5.6: Tabulka s hodnotami metaparametrů pro experiment s Konvolučním hráčským modelem

5.5.4 Diskuze

Při tomto experimentu se ukázalo, že prvotní nastavení počtu vrstev, jenž bylo pouze kvalifikovaným odhadem, na hodnotu 5 jako velmi dobré. Modely s hráčskými i týmovými statistikami totiž dosahovali na této hodnotě skoro nejlepších výsledků. Nejlepší výsledky pak byly dosaženy na hodnotě 4 viz grafy na Obrázcích 5.7 a 5.8. Dále se ukázalo, že tvar neuronové sítě může mít vliv na kvality zastavovací funkce. U modelů s týmovými statistikami vysoké granularity, jež mají na svém vstupu velký počet statistik, funkce early stopping nebyla schopna přeučení[7], s měnící se hloubkou sítě, nijak zabránit. Zatímco u hráčského modelu s vysokou granularitou[8] bylo přeučení lehce vidět na grafu 5.3, early stopping funkce ho byla schopna se zvyšujícím se počtem vrstev výrazně snížit. Pro určení velikost embeddingu bylo dosaženo jednoznačného výsledku, že s větším počet atributů vytvářených embeddingem se vlasnosti modelu zhoršují. Protože u tohoto experimentu nebyly testovány hodnoty[9] < 5, nelze o ideální velikosti embeddingu říct více, než že se nachází mezi hodnotou 1 a 5.

5.6 Vliv časového vývoje

Experimenty provedené v teto části zkoumají vliv času na trénování modelů. Jde o reakci na vypozorovaný trend snižujícího se počtu remíz v novějších sezónách. viz Tabulka 5.1.

[7] je pěkně vidět na 5.2
[8] má méně statistik vstupujících do první vrstvy modelu
[9] na znovuspuštění experimentu s nízkými hodnotami nevyzbyl čas

Parametr	Hodnoty
Typ statistik	Hráčské, Týmové
Granularita	Vysoká
Velikost posuvného okna	1, 2, 4, 6, 8, 10, 12, 14, 16
Počet vrstev modelu	5
Zvýhodnění novějších zápasů	0
Konvoluční model	Ne
Resetování parametrů*	Ano

* Resetování parametrů modelu při posunu posuvného okna

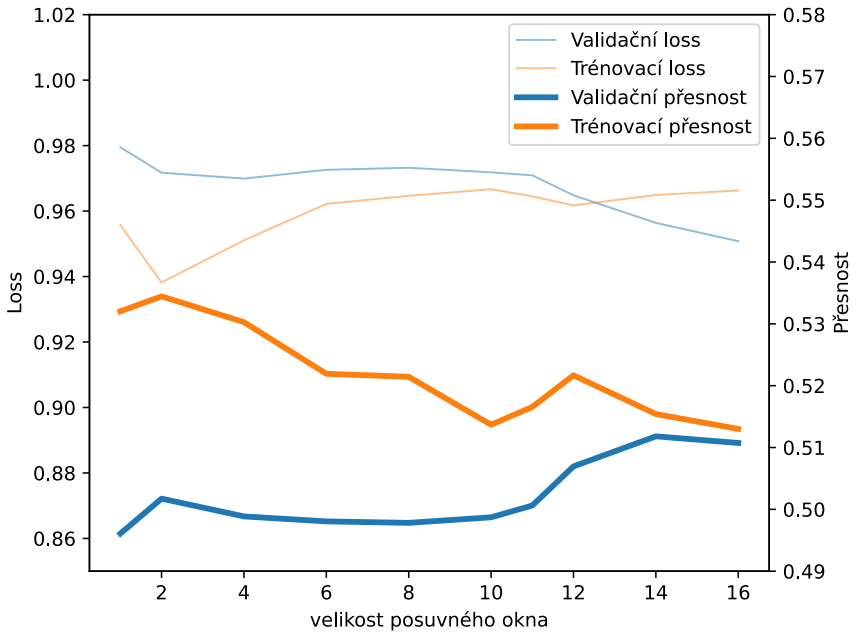
Tabulka 5.7: Tabulka s hodnotami metaparametrů pro experiment zaměřený na posuvné okno.

5.6.1 Velikost posuvného okna

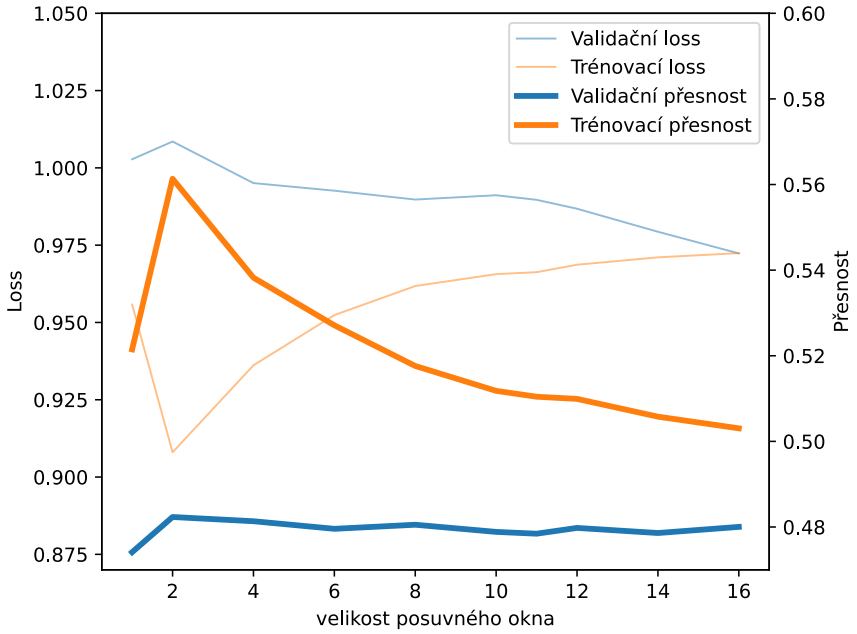
První experiment zkoumá velikost posuvného okna, které bylo do trénovací smyčky zavedeno z důvodů omezení vlivu historie při trénování. Experiment se snaží dokázat, že myšlenka posuvného okna napomáhá vytvářet obecnější modely a potlačovat iregularity mezi jednotlivými sezónami a tím tak zlepšit predikční schopnosti. U tohoto experimentu byly všechny ostatní metaparametry zafixovány na hodnotách vypsaných v Tabulce 5.7.

Velikost posuvného okna 1 značí okno, které nikam neposouvalo svůj začátek, ale pouze přidávalo následující roky do tréninkové množiny, podobný způsob využili při trénování i [2], [15] nebo [5]. Toto okno tak na rozdíl od ostatních nemělo v jednotlivých iteracích posunu stejnou velikost. Z grafu na Obrázku 5.12 je vidět, že při větším trénovacím okně dosahuje hráčský model lepších výsledků. Může to ale být způsobeno tím, že pro velká okna 12+ jsou validační data z hlediska predikce pro model jednoduší[10]. Tomuto vysvětlení by nahrával i fakt, že výraznější zlepšení je možné pozorovat od 12. okna, které koresponduje se skokovým poklesem počtu remíz mezi sezónami 2014/15 a 2015/16. Na Obrázku 5.13 je graf zobrazující experiment pro model s týmovými statistikami. Zde je vidět, že model byl na malých oknech přeučený a zvětšení mu pomohlo k lepší generalizaci. Důvodem snižující se validační loss, ale nikterak se nezlepšující validační přesnosti může být to, že model si byl při svých správných predikcí více jist, než u těch kde chyboval.

[10] vysvětleno v Sekci 5.2



Obrázek 5.12: Vývoj přesnosti hráčského modelu s vysokou granularitou v závislosti na posuvném okně.



Obrázek 5.13: Vývoj přesnosti týmového modelu s vysokou granularitou v závislosti na posuvném okně.

Parametr	Hodnoty
Granularita	Vysoká
Počet vrstev modelu	5
Velikost posuvného okna	16
Zvýhodnění novějších zápasů	0, 0.5, 1, 2
Konvoluční model	Ne
Resetování parametrů*	Ano

* Resetování parametrů modelu při posunu posuvného okna

Tabulka 5.8: Tabulka s hodnotami metaparametrů pro experiment zaměřený na zvýhodnění novějších zápasů.

5.6.2 Zvýhodnění novějších zápasů

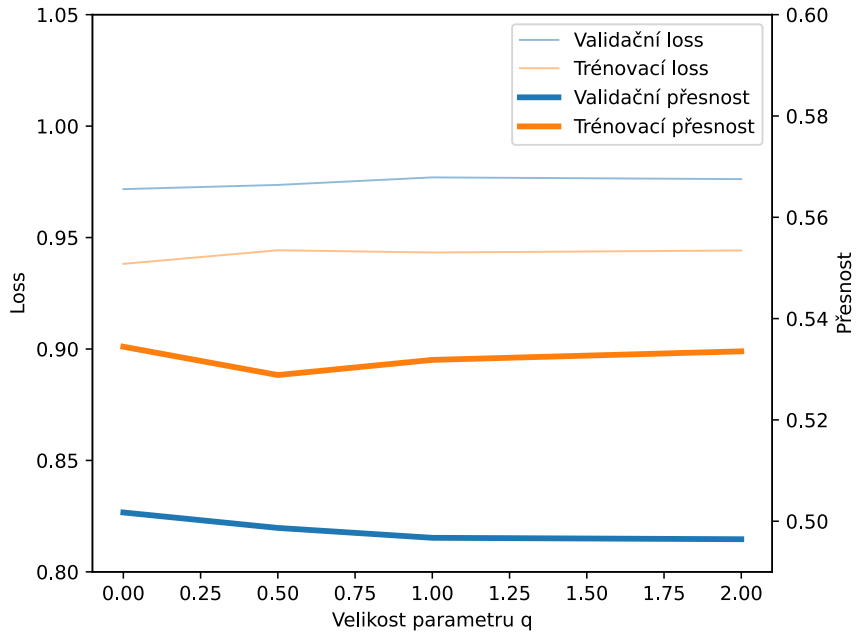
Tento experiment si pohrává s myšlenkou zvýhodnění novějších zápasů. Experiment se snaží využít kratšího časového rozdílu mezi zápasy z konce trénovací množiny a zápasy z množiny validační. Tímto se tak pokoušíme ověřit hypotézu, že zápasy hrané v kratším časovém intervalu spolu více korelují. U tohoto experimentu tak budou všechny ostatní metaparametry zafixovány na hodnotách uvedených v Tabulce 5.8.

Jelikož funkce `CrossEntropyLoss` umožňuje navrácení loss pro každý zápas zvlášť, byl vytvořen vektor vah $\mathbf{v}$, který závisí pouze na jednom parametru q, který ovlivňuje velikost zvýhodnění:

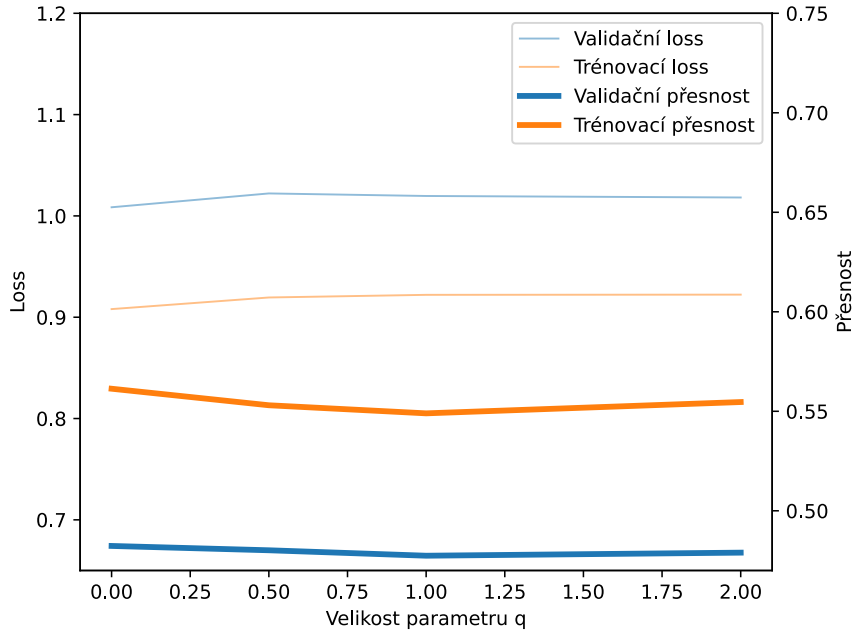
$$\mathbf{v} = \frac{1 + q \cdot \mathbf{z}}{\sum_{i=1}^{Z} i} \qquad \mathbf{z} = [1, 2, 3, \ldots, Z] \tag{5.4}$$

kde Z je počet zápasů v dané množině dat. Výsledný loss se spočítá jako skalární součin vektoru obsahujícího loss pro jednotlivé zápasy s vektorem $\mathbf{v}$. S rostoucím q tak roste i zvýhodnění nových zápasů na úkor starších.

Na Obrázcích 5.14 a 5.15 jsou grafy zobrazující výsledky experimentu. Ani v jednom případě se zvýhodnění zápasů neprojevilo zlepšením vlastností modelu. Hodnota 0 parametru q značí nulové zvýhodnění, což je stejné jako aritmetický průměr loss jednotlivých zápasů, který dělá funkce `CrossEntropyLoss`.



Obrázek 5.14: Vývoj přesnosti hráčského modelu s vysokou granularitou v závislosti na zvýhodnění nových zápasů.



Obrázek 5.15: Vývoj přesnosti týmového modelu s vysokou granularitou v závislosti na zvýhodnění nových zápasů.

Parametr	Hodnoty
Typ statistik	Hráčské, Týmové
Granularita	Vysoká
Počet vrstev modelu	5
Velikost posuvného okna	4
Konvoluční model	Ne
Resetování parametrů	Ano, Ne

Tabulka 5.9: Tabulka s hodnotami metaparametrů pro experiment zaměřený na resetování parametrů modelu při posunu trénovacího okna.

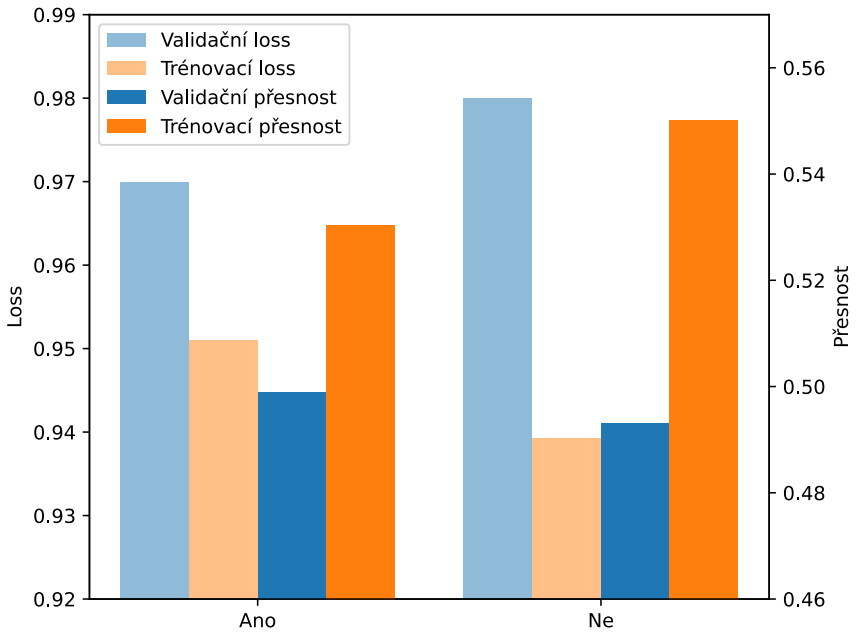
5.6.3 Resetování parametrů

Za tímto experimentem stojí myšlenka přetrénování naučeného modelu na novou sadu dat. Ve všech ostatních experimentech byly parametry modelu, vždy při posunu okna s daty, resetovány. Informace získaná tréninkem na předchozím okně, tak byla ztracena. Při tomto experimentu byly zbylé metaparametry zafixovány na hodnotách v Tabulce 5.10.

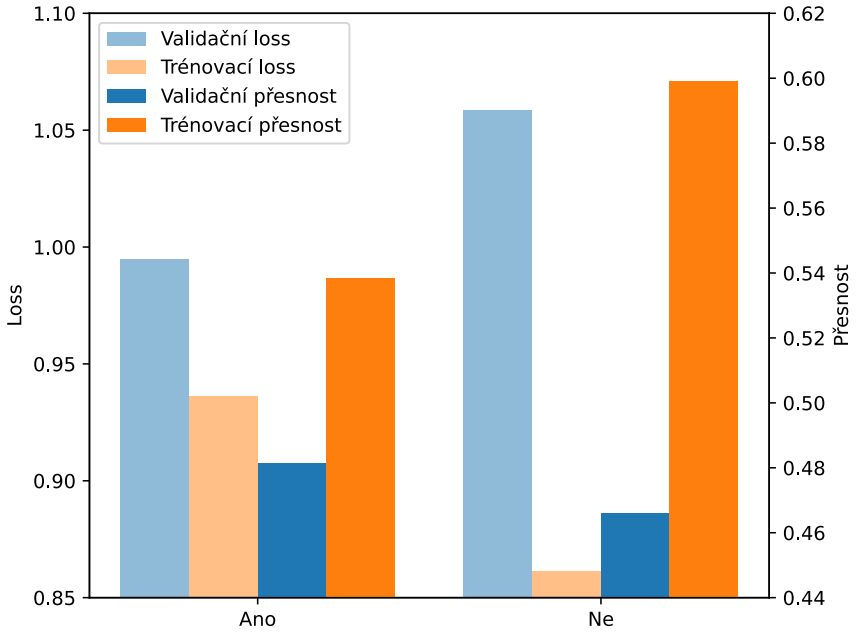
Z grafů na Obrázcích 5.16 a 5.17 je vidět, že resetování parametrů modelu, mezi jednotlivými iteracemi posuvného okna, má regularizační účinky jak na model s hráčskými, tak na model s týmovými statistikami.

5.6.4 Diskuze

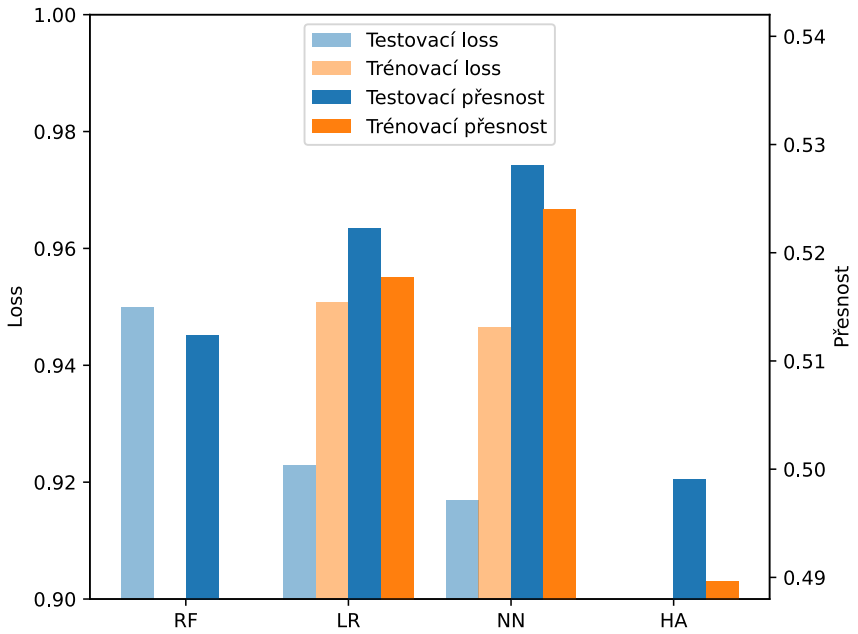
Z výsledků těchto experimentů se ukázalo, že jednotlivé hokejové zápasy mezi sebou v tomto časovém měřítku nemají časovou závislost. Lepší dosažené výsledky pro velká tréninková okna byly vysvětleny skokovou změnou v počtu remíz tedy nejhůře predikovatelné třídy.



Obrázek 5.16: Graf přesnosti hráčského modelu s vysokou granularitou pro experiment s resetováním parametrů.



Obrázek 5.17: Graf přesnosti týmového modelu s vysokou granularitou pro experiment s resetováním parametrů.



Obrázek 5.18: Porovnání výsledného modelu s modely základními. RF - random forest. LR - logistická regrese. NN - neuronová síť. HA - procentuální výhra domácího týmu.

5.7 Model vytvořený s poznatky z experimentů

Na základě výsledků jednotlivých experimentů, byl vytvořen model pro predikci hokejových výsledků s metparemetry vypsanými v Tabulce 5.6. Důvodem proč bylo vybráno okno s velikostí 16 je to, že nastolený trend nízkého počtu remíz je očekáván i v následujících sezónách. Model byl nejdříve trénován na sezónách 2002/03 až 2017/18 a sezóna 2018/19 byl brána jako validační, aby funkce early stopping měla z čeho určovat, kdy je vhodné zastavit trénování. Tato epocha e byla zaznamenána a validační rok se připojil za trénovací. Poté byl proces učení opět spuštěn, tentokrát již na sezónách 2002/03 až 2018/19, učení bylo zastaveno po epoše e. Model byl následně otestován na sezóně 2019/20. Stejným způsobem byla trénována a testována i logistická regrese, u metody random forest nebylo potřeba dvojího tréninku pro zjištění epochy e a proto byl model rovnou trénován a následně otestován.

Výsledky modelů jsou vyneseny do grafu na Obrázku 5.18. Trénovací loss a přesnost pro model random forest nebyly do grafu vyneseny. Do porovnání je ještě zařazena procentuální výhra domácího týmu na trénovacím a testovacím období. Nejlépe si v porovnání vedla neuronová síť, která na testovací množině dosahovala přesnosti 52.8%, Logistická regrese měla přesnost 52.2% a metoda random forest dosahovala přesnosti 51.1%. Zajímavým faktem je, že si modely vedli lépe na testovací množině lépe než na množině trénovací. Toto je způsobeno změnou zastoupení jednotlivých tříd v měřených množinách.

Parametr	Hodnoty
Granularita	Vysoká
Typ statistik	Hráčské
Počet vrstev modelu	4
Velikost posuvného okna	16
Zvýhodnění novějších zápasů	0
Konvoluční model	Ne
Resetování parametrů*	Ne
* Při této velikosti se nemá okno kam posunou	

Tabulka 5.10: Tabulka s metaparametry výsledného modelu.

5.8 Komplikace s experimenty

První komplikace nastaly při přesunu databáze z notebooku na databázový server, tentokrát kvůli rychlosti a stabilitě internetového připojení. Notebook na kterém se nacházela původní databáze neměl možnost připojení internetu pomocí kabelu a připojení WiFi nebylo příliš rychlé ani stabilní. Spojení mezi databázovým serverem a notebookem tak nevydrželo otevřené po celou dobu přenosu dat. Řešením tohoto problému bylo přesunutí dat po menších částech ve velmi pozdních večerních/brzkých ranních hodinách, kdy bylo připojení rychlejší a stabilnější.

Při spuštění experimentů na serveru nastaly další komplikace s databázovým serverem, jelikož výpočetní server jednotlivé úlohy spouští podle obsazenosti hardwaru, nastala situace, kdy si o data na databázový server žádalo příliš mnoho experimentů. To mělo za následek ukončení všech otevřených připojení k databázovému serveru a po určitou dobu nebyla možnost navázání nových spojení. Z chybových hlášení nebylo jasné, zda se jedná o chybu knihovny *sshtunel*, která se starala o připojení k databázovému serveru, nebo o přetížení databázového serveru. Tento problém byl vyřešen vlastním spouštěcím programem, který zajišťoval aby na serverech neběželo příliš úloh ve fázi načítání dat. Program měl přehled, ve které fázi se jednotlivé spuštěné úlohy nacházejí a další úlohu spustil teprve až ve chvíli, kdy se v načítcí fázi nacházelo méně než 10 úloh.

Kapitola 6

Závěr

V této práci byl podrobně popsán postup výběru vhodné hokejové soutěže pro následnou predikci výsledku zápasů. Popsány jsou pak také jednotlivé zdroje nabízející statistiky pro vybranou soutěž NHL. Dále byla jako součást práce vytvořena databáze obsahující statistiky NHL sahající až do roku 1917. V práci je vysvětlen způsob, jakým byly statistiky získávány a následně transformovány do tvaru, který je možné použít jako vstup pro predikční modely. Před samotnými experimenty jsou ještě vysvětleny principy použitých modelů a metoda posuvného okna, která je využita při načítání dat.

Při experimentu s granularitou dat byla potvrzena hlavní hypotéza a tedy že u modelů s hráčskými statistikami přináší vyšší granularita dat lepší predikční schopnosti. Dále tento experiment ukázal, že modely založené na hráčských statistikách dosahují při střední a vysoké granularitě dat lepších výsledků než modely týmové. Zajímavostí je, že při použití embeddingu není mezi týmovými a hráčskými modely žádný rozdíl. Experiment s velikostí sítě zase prokázal, že velké sítě nemají dobrý vliv na vlastnosti modelu a spíše modely s méně vrstvami dosahují lepších výsledků. Jako velmi dobrá náhrada za model neuronové sítě se pro predikci ukázala logistická regrese, jež ve spoustě případů dosahovala stejných nebo velmi podobných výsledků. Při experimentech zabývajících se časovým vývojem bylo zjištěno, že skoková změna rozdělení klasifikovaných tříd v nových sezónách má na výsledky modelu velký vliv. Výsledkem experimentálního snažení byl model, který využíval získaných poznatků a při otestování dosahoval přesnosti 52.8%. Důvodem proč není výsledný model výrazněji lepší, než modely figurující v dílčích experimentech je velmi dobré nastavení základních hodnot fixovaných metaparametrů. Toto nastavení bylo pouze kvalifikovaným odhadem a velmi se podobalo nastavení výsledného modelu.

Při porovnání výsledného modelu[1] (52.8%) s modely z ostatních prací je nutno zmínit, že hokejové modely vytvořené Wei Gu a další (77.5%) [9], Gianni Pischedda (61.54%)

[1] Programy pro vytvoření modelu jsou dostupné v repozitáři GitHub

[18] nebo Joshua Weissbock (59.38%) [22], provádí pouze binární klasifikaci, u které je průměrná přesnost 50%. Zatímco u problému klasifikace 3 tříd, kterou řeší model vytvořený v této práci, činí průměrná přesnost pouhých 33%. Dále je ve zmíněných pracích používán jako testovací vzorek mnohem menší počet zápasů, jedná se tak často o modely trénované a testované na jedné sezóně rozdělené v poměru 10:1, což ze statistického hlediska dělá model vytvořený v této práci signifikantnějším.

Na základě výsledků této práce je možné provádět další experimenty, které by objasnily zda například kombinace týmových a hráčských statistik pomůže zlepšit predikční vlastnosti modelu. Díky vytvořené databázi celé historie NHL je možné vytvářet modely pro predikci nejen výsledku zápasu, ale i pro predikci různých hráčských statistik. Například modely predikující vstřelené góly jednotlivými hráči lze využít při sázení u sázkových kanceláří.

Příloha A

Literatura

[1] P. Andersson, J. Edman, and M. Ekman. Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting*, 21(3):565–576, 2005.

[2] B. G. Aslan and M. M. Inceoglu. A comparative study on neural network based soccer result prediction. In *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, pages 545–550, 2007.

[3] G. Bonaccorso. *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[4] R. P. Bunker and F. Thabtah. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33, 2019.

[5] S. Buttrey, A. Washburn, and W. Price. Estimating nhl scoring rates. *Journal of Quantitative Analysis in Sports*, 7:24–24, 01 2011.

[6] E. Experts. Processing power compared. `https://www.slideshare.net/Experts-Exchange/processing-power-compared`, 2015.

[7] D. Feltz and C. Lirgg. Perceived team & player efficacy in hockey. *The Journal of applied psychology*, 83:557–64, 09 1998.

[8] W. Gu, K. Foster, J. Shang, and L. Wei. A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, 130:293–305, 2019.

[9] W. Gu, T. Saaty, and R. Whitaker. Expert system for ice hockey game prediction: Data mining with human judgment. *International Journal of Information Technology & Decision Making*, 15, 06 2016.

[10] D. Hynes. nhlapi. `https://gitlab.com/dword4/nhlapi`, 2021.

[11] J. Kahn. Neural network prediction of nfl football games. *World Wide Web electronic publication*, pages 9–15, 2003.

[12] O. Kouropteva, O. Okun, and M. Pietikäinen. Selection of the optimal parameter value for the locally linear embedding algorithm. *FSKD*, 2:359–363, 2002.

[13] M. Leung. $4,718 - using machine learning to bet on the nhl, Sep 2020.

[14] M. J. Mauboussin. *The success equation: Untangling skill and luck in business, sports, and investing.* Harvard Business Review Press, 2012.

[15] A. McCabe and J. Trevathan. Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*, pages 1194–1197, 2008.

[16] J. McCullagh et al. Data mining in sport: A neural network approach. *International Journal of Sports Science and Engineering*, 4(3):131–138, 2010.

[17] I. B. Mohamad and D. Usman. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17):3299–3303, 2013.

[18] G. Pischedda. Predicting nhl match outcomes with ml models. *International Journal of Computer Applications*, 101:15–22, 09 2014.

[19] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[20] P. Somervuo and T. Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10(2):151–159, 1999.

[21] A. Thomas, S. L. Ventura, S. T. Jensen, and S. Ma. Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics*, pages 1497–1524, 2013.

[22] J. Weissbock. *Forecasting success in the National Hockey League using in-game statistics and textual data.* PhD thesis, Université d'Ottawa/University of Ottawa, 2014.

[23] F. Šimsa. Analysis and prediction of league games results, 2015.

Příloha B

Zkratky statistik

Zde se nacházejí vysvětlivky pro zkratky statistik.

GF - Vstřelené góly
GA - Obdržené góly
S - Střely
SA - Střely soupeře
S/G - Počet střel na gól
SA/GA - Počet soupeřových střel na obdržený gól
PIM - Trestné minuty
PPGF - vstřelené góly v přesilovkách
PPO - Počet přesilovek
PP% - Úspěšně využité přesilovky
PKGA - Obdržené góly v oslabení
PKO - Počet oslabení
PK% - Úspěšně ubráněná oslabení
FOW% - Procento vyhraných buly
FOW - Vyhraná buly
FO - Hraná buly
BS - Zblokované střely
BS% - Procento zblokovaných střel
Hits - Bodyčky
GvA - Ztracené puky
TkA - Získané puky
TOI - Čas na ledě
SHSV - Zákroky v oslabení
SHSA - Střely na brankáře v oslabení
SHA - Asistence v oslabení
SHG - Góly v oslabení

SHTOI - Čas na ledě v oslabení
PPSV - Zákroky v přesilovce
PPSA - Střely na brankáře v přesilovce
PPA - Asistence v přesilovce
PPG - Góly v přesilovce
PPTOI - Čas na ledě v přesilovce
EVSV - Zákroky při hře ve stejném počtu
EVSA - Střely na brankáře při hře ve stejném počtu
EVTOI - Čas na ledě při hře ve stejném počtu
A - Asistence
G - Góly
SV - Zákroky
PM - Plusy mínusy

Příloha C

Tabulky

V této části se nacházejí tabulky s informacemi o poskytovaných statistikách jednotlivými zdroji popisované v Kapitole 3.

Statistika	Hráči	Týmy
Čas na ledě	✓[1]	
Góly	✓	✓
Asistence	✓	
Střely	✓[1]	✓
Zákroky	✓	✓
Trestné minuty	✓	✓
+/-	✓	
Vhazování		
Vyhraná vhazování	✓[2]	
Zblokované střely	✓[1]	✓[1]
Získané puky		
Ztracené puky		
Hity	✓[1]	✓[1]
Střídání		
Počet přesilovek		✓
Čas na ledě v přesilovkách		
Góly v přesilovkách	✓	✓
Asistence v přesilovkách		
Střely v přesilovkách		
Zákroky v přesilovkách		
Počet oslabení		✓
Čas na ledě v oslabení		
Góly v oslabení	✓	✓
Asistence v oslabení		
Střely v oslabení		
Zákroky v oslabení		

[1] od play-off v sezóně 2014/15
[2] od play-off v sezóně 2015/16 a pouze v %

Tabulka C.1: Tabulka s přehledem statistik nabízených webem nhl.cz

Statistika	Hráči	Týmy
Čas na ledě	✓	
Góly	✓	✓
Asistence	✓	
Střely	✓	✓
Zákroky	✓	
Trestné minuty	✓	✓
+/-	✓	
Vhazování		
Vyhraná vhazování	✓[1]	✓
Zblokované střely	✓	✓
Získané puky	✓	✓
Ztracené puky	✓	✓
Hity	✓	✓
Střídání		
Počet přesilovek		✓
Čas na ledě v přesilovkách		
Góly v přesilovkách		✓
Asistence v přesilovkách		
Střely v přesilovkách		✓
Zákroky v přesilovkách	✓	✓
Počet oslabení		✓
Čas na ledě v oslabení		
Góly v oslabení		✓
Asistence v oslabení		
Střely v oslabení		✓
Zákroky v oslabení	✓	✓

[1] pouze v %

Tabulka C.2: Tabulka s přehledem statistik nabízených webem nhlportal.cz

Statistika	Hráči	Týmy
Čas na ledě	✓[1]	
Góly	✓	✓
Asistence	✓	
Střely	✓[2]	
Zákroky	✓[3]	
Trestné minuty	✓	
+/-	✓[2]	
Vhazování		
Vyhraná vhazování		
Zblokované střely	✓[4]	
Získané puky		
Ztracené puky		
Hity	✓[4]	
Střídání	✓[1]	
Počet přesilovek		
Čas na ledě v přesilovkách		
Góly v přesilovkách	✓[5]	✓[7]
Asistence v přesilovkách	✓[6]	
Střely v přesilovkách		
Zákroky v přesilovkách		
Počet oslabení		
Čas na ledě v oslabení		
Góly v oslabení	✓[5]	✓[7]
Asistence v oslabení	✓[6]	
Střely v oslabení		
Zákroky v oslabení		

[1] od sezóny 1997/98 [2] od sezóny 1959/60
[3] od sezóny 1955/56 [4] od sezóny 2007/08
[5] od sezóny 1933/34 [6] od sezóny 2014/15
[7] od sezóny 1935/36

Tabulka C.3: Tabulka s přehledem statistik nabízených webem hockey-reference.com

Statistika	Hráči	Týmy
Čas na ledě	✓[1]	
Góly	✓	✓
Asistence	✓	
Střely	✓[2]	✓[2]
Zákroky	✓[3]	
Trestné minuty	✓	✓
+/-	✓[2]	
Vhazování	✓[1]	✓[1]
Vyhraná vhazování	✓[1]	✓[1]
Zblokované střely	✓[4]	✓[4]
Získané puky	✓[4]	✓[4]
Ztracené puky	✓[4]	✓[4]
Hity	✓[4]	✓[4]
Střídání	✓[1]	
Počet přesilovek		✓[7]
Čas na ledě v přesilovkách	✓[1]	
Góly v přesilovkách	✓[5]	✓[5]
Asistence v přesilovkách	✓[5]	
Střely v přesilovkách	✓[1]	
Zákroky v přesilovkách	✓[1]	
Počet oslabení		✓[7]
Čas na ledě v oslabení	✓[1]	
Góly v oslabení	✓[5]	✓[5]
Asistence v oslabení	✓[5]	
Střely v oslabení	✓[6]	
Zákroky v oslabení	✓[1]	

[1] od sezóny 1997/98 [2] od sezóny 1959/60
[3] od sezóny 1955/56 [4] od sezóny 2005/06
[5] od sezóny 1933/34 [6] od sezóny 2009/10
[7] od sezóny 1977/78

Tabulka C.4: Tabulka s přehledem některých statistik nabízených webem nhl.com