FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání diplomové práce

Název:	Predikce výsledků tenisových utkání
Student:	Bc. Martin Zukal
Vedoucí:	Ing. Ondřej Hubáček
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

Cílem práce je navrhnout a implementovat model, který bude predikovat výsledky tenisových zápasů na základě historických dat a otestovat tento model na relevantních datech.

Konkrétní úkoly pro řešitele jsou následující:
1. Provést rešerši existujících modelů pro predikce sportovních výsledků.
2. Popsat a zpracovat dostupná data z tenisových utkání.
3. Navrhnout a implementovat model predikující výsledky tenisových utkání.
4. Otestovat model na relevantních datech a porovnat výsledky s jinými dostupnými modely.

[1] Kovalchik, Stephanie Ann. "Searching for the GOAT of tennis win prediction" Journal of Quantitative Analysis in Sports, vol. 12, no. 3, 2016, pp. 127-138.
[2] Lisi, Francesco. "Tennis betting: can statistics beat bookmakers?." Electronic Journal of Applied Statistical Analysis 10.3 (2017): 790-808.
[3] Wilkens, Sascha. "Sports prediction and betting models in the machine learning age: The case of tennis." Journal of Sports Analytics 7 (2021): 99-117.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 22. listopadu 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Predikce výsledků tenisových utkání

Bc. Martin Zukal

Katedra aplikované matematiky
Vedoucí práce: Ing. Ondřej Hubáček

3. května 2022

Poděkování

Chtěl bych poděkovat svému vedoucímu Ing. Ondřeji Hubáčkovi a firmě Ematiq a.s., že mi umožnili vypracovat tuto práci.

Prohlášení

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

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 3. května 2022 . . . . . .

Odkaz na tuto práci

Zukal, Martin. *Predikce výsledků tenisových utkání*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2022.

Abstrakt

Tato práce se zabývá návrhem a implementací modelů pro predikci pravděpodobnosti výsledků tenisových zápasů a jejich aplikací pro sázení. Vytvořil jsem dva různé modely (klasifikační a bodový) a zpracoval historická data, která sloužila pro napočítání statistik. Tyto statistiky byly použity pro natrénování modelů. Dále jsem provedl rozsáhlou rešerši, na jejímž základě jsem vybral state of the art model, který jsem naimplementoval a použil pro porovnání výsledků svých modelů. Oba mnou navržené modely dosahují vyšší přesnosti než výše zmíněný state of the art model. Nakonec jsem všechny tři modely použil pro analýzu ziskovosti, kde byly predikované pravděpodobnosti modelů použity pro sázení u sázkových kanceláří.

Klíčová slova prediktivní modelování, strojové učení, datová analýza, neuronové sítě, tenis

Abstract

This work deals with the design and implementation of models for predicting the probability of tennis match results and their applications for betting. I created two different models (classification and point-based) and processed historical data, which was used to calculate statistics. These statistics were used to train the models. I also conducted an extensive search, based on which I selected the state of the art model, which I implemented and used to compare the results of my models. Both models designed by me achieve higher accuracy than the above-mentioned state of the art model. Finally, I used all three models for profitability analysis, where the predicted probabilities were used for betting at bookmakers.

Keywords predictive modeling, machine learning, data analysis, neural networks, tennis

Obsah

Seznam obrázků

Seznam tabulek

Kapitola 1

Úvod

Tenisová klání nejsou v našich zemích žádnou novinkou. Již v době Rudolfa II. byla na Pražském hradě postavena budova, zvaná Míčovna, v níž se provozovaly aktivity ne nepodobné dnešnímu tenisu. Ten se k nám dostal pravděpodobně z Německa. Historicky pochází hra ze středověké Francie, kde je doložena k roku 1275 pod názvem jeu de la chasse (hra na lov), protože napodobovala chytání ptáků do sítí. Provozovali ji zejména řeholníci pocházející z urozených rodin jako náhradu rytířských zábav. Míčem se trefovalo do branky, a hrálo se na klášterních dvorech (francouzsky cour), odtud označení tenisového hřiště jako dvorec nebo kurt.[1]

Modelování tenisu jsem si vybral z jednoho prostého důvodu. Nikdy jsem se o tento sport nezajímal a chci svou prací dokázat, že i bez hlubších znalostí dané domény se dá vytvořit model, který bude stejně kvalitní, nebo lepší, než již existující modely. Stačí provést důkladnou analýzu a dát si práci s feature engineeringem. Položme si tedy otázku, může i amatérský model porážet světové sázkové kanceláře?

Pro sázkové firmy jsou šetření tohoto druhu za účelem stanovení sázkových kurzů součástí běžného každodenního života. Pokud bude predikci provádět velká firma, bude mít samozřejmě k dispozici rozsáhlejší zdroje dat než analytici, kteří se predikcemi zabývají ze zájmu. Sázkové firmy například průběžně shromažďují a analyzují data, aby mohly nastavit vhodné a zároveň ziskové kurzy pro nabízené sázky. K tomu používaná data se obvykle neomezují pouze na výsledky minulých her, ale také se zaměřují na probíhající změny okolností různé povahy. V případě potřeby jsou zohledněny i dodatečné vlivy jako zpráva o zranění hráče nebo velký objem sázek na jednoho z hráčů. Analýza je navíc prováděna tak, že se závěry neustále promítají do původně stanovených sázkových kurzů. To znamená, že aktuální kurzy poskytovatele sázek se mohou kdykoli v období před zahájením utkání

výrazně změnit, pokud budou k dispozici nové informace naznačující jinou předpověď.

Pokusím se tedy vytvořit model, který bude mít tak dobré predikce, že porazí i sázkovou kancelář a výsledek svého snažení si dovolím předložit v této práci.

1.1 Cíle

Cílem této diplomové práce je navrhnout a implementovat modely, které budou co nejpřesněji predikovat výsledky tenisových utkání. Abych toho dosáhl, musím nejprve provést rešerši již existujících modelů a pak implementovat nejlepší z nich, abych mohl provést porovnání se svými modely. Aby se mohly modely učit, bude potřeba sehnat co největší množství historických dat, předpřipravit je a napočítat na nich statistiky. Po natrénování a porovnání modelů bych rád získal historické kurzy sázkových kanceláří a na testovací množině vyzkoušel, jestli jsou tyto modely schopny sázkové kanceláře porážet.

Kapitola 2

Rešerše

V následující kapitole je popsáno několik přístupů, které byly v minulosti používány k predikcím výsledků tenisových utkání. Lze je rozdělit do několika kategorií:

- Klasifikační modely
- Bodové modely
- Modely užívající párové srovnání
- Modely založené na odhadech bookmakerů

2.1 Klasifikační modely

Klasifikace se zabývá zkoumáním vztahů v rámci množiny objektů s cílem zjistit, zda lze data shrnout do tříd, které jsou si blízké[2]. Jednou z klasifikačních metod je regresní analýza.
Regresní analýza je jednou z nejpoužívanějších statistických metod pro získávání znalostí o vztazích mezi proměnnými. Dříve jsme se s těmito technikami setkávali nejvíce při řešení problému v ekonometrii, ale později se rozvinuly i do dalších oblastí, včetně predikce sportovních utkání. [3]
Jedna z těchto technik je Probit regresní model, který využívá většina z mnou popisovaných klasifikačních modelů. Probit model popisuje vztah mezi diskrétní binární náhodnou událostí, v našem případě je to výhra nebo prohra tenisového utkání, a určitého souboru vysvětlujících proměnných. Poskytuje odhad pravděpodobnosti, že dojde k jednomu z dvou možných výsledků, pokud mají vysvětlující proměnné specifikované hodnoty[4]

Matematicky popisuje probitový model věta[5]:

Definice 2.1.1 (Probitový model)**.** Základní forma probitového modelu je

$$\pi_{\mathrm{ij}} = \Phi(x'_{\mathrm{ij}}\beta)$$

kde Φ označuje funkci kumulativní hustoty standardní normální proměnné a x_{ij} představuje vektor příznaků a β koeficienty regrese.

Příznaky mohou zahrnovat charakteristiky hráče, soupeře nebo zápasu. Modely se liší v sadě příznaků, které berou v úvahu.

2.1.1 Logistický model

Následující tenisový model[6] z roku 2003 popisuje, jak předpovědět vítěze tenisového zápasu na největších tenisových turnajích (Grand Slam) jak před zápasem, tak během zápasu. Tenisové turnaje probíhají formou takzvaného pavouka, to znamená, že do každého dalšího kola postupují hráči, kteří vyhráli svůj zápas. Do dalšího kola se tedy posune polovina hráčů. Tento model používá pouze jeden příznak, a to pořadí v žebříčku ATP, nebo WTA. Autoři ale tento ranking upravují, protože výkonnostní rozdíly mezi nejlepšími hráči v žebříčku jsou větší, než rozdíl mezi hráči na spodku žebříčku. Proto zavádějí proměnou R, která měří, do kterého kola v turnaji se daný hráč dostane.

$$R_{\mathrm{a}} = 8 - \log_2(RANK_a)$$

kde $(RANK_a)$ je pořadí hráče v žebříčku.

Pro výpočet pravděpodobnosti vítězství hráče je použit jednoduchý logit model

$$\pi_{\mathrm{j}} = \frac{exp(F_j)}{1 - exp(F_j)}$$

kde F_j je funkce, která transformuje hodnocení hráčů R_a a R_b. Buď $D_j \equiv R_a - R_b$. Když $D_j = 0$, pak $R_a = R_b$ a oba hráči jsou stejně silní a měli by mít stejnou šanci na výhru, takže $\pi = 0.5$ a tedy $F_j = 0$. Z toho tedy plyne, že $F_j = \lambda D_j$, kde λ je konstanta. Pak tedy

$$\pi_{\mathrm{j}} = \frac{exp(\lambda D_j)}{1 - exp(\lambda D_j)}$$

Buď $z_j = 1$, pokud hráč a vyhraje j-tý zápas a 0, pokud ho prohraje.

Potom pravěpodobnost vzorku je počítána jako

$$L = \prod_{j=1}^{N} \pi_z^{z_j}(1 - \pi_j)^{1-z_j}$$

2.1.2 Probit model s nasazením v turnajích

Tento model[4] z roku 1999 se zaměřuje na predikce basketbalu a tenisu. Na rozdíl od předchozího modelu používá Probit model a jako vysvětlující proměnnou místo žebříčku používá takzvaný seed. Seed je rozložení hráčů na tenisových turnajích takové, aby se nejlepších 16 hráčů mohlo potkat nejdříve ve 4. kole turnaje. Dva nejlepší hráči žebříčku se mohou potkat až ve finále. Autoři se tedy zaměřují pouze na 16 nejlepších hráčů turnaje. V tabulce 2.1 můžeme vidět vliv seedu na umístění nejlepších hráčů mezi nejlepšími šestnácti.
Jak již bylo zmíněno výše, model používá pouze jeden prediktor, a to seed ve formě $s_1 - s_2$, kde s_1 je seed lepšího z hráčů a s_2 horšího hráče. Pokud hraje tenista se seedem s někým bez seedu, je tomuto hráči přiřazen seed s hodnotou 17. Autoři zkoušeli přidat i další proměnnou a to pořadí v žebříčku, ale bezúspěšně.

Místo v žebříčku	Muži	Ženy
1	38	43
2	36	42
3	33	36
4	34	38
5	19	29
6	22	29
7	23	32
8	20	29
9	24	20
10	19	22
11	18	26
12	16	22
13	24	21
14	19	20
15	12	18
16	17	16

Tabulka 2.1: Počet umístění v TOP 16(z 44 možných) šestnácti nejlepších tenistů a tenistek

2.1.3 Probit model s žebříčkem a demografií

Tento model[7] je o proti předchozím modelům komplexnější a využívá 3 různé typy vysvětlujících proměnných:

- Charakteristiky zápasu
- Předchozí výsledky hráčů
- Fyzické vlastnosti hráčů

Všechny vysvětlující proměnné jsou popsané v tabulkách 2.2, 2.3 a 2.4. I v tomto případě používají autoři Probit model. Z jejich výsledků vyplývá, že nejrelevantnější vysvětlující proměnná je jak u mužů, tak i u žen rozdíl v žebříčku.

Název	Popis
TOURNAMENT	název grand slamu
SURFACE	typ povrchu

Tabulka 2.2: Charakteristiky zápasu

Název	Popis
DIFRANKING	Rozdíl logaritmů umístění hráčů v žebříčku
DIFROTOUR	Rozdíl mezi počty vyhraných zápasů na stejném turnaji v minulém roce
TOP10	1, pokud byl hráč v posledních 5 letech mezi 10 nejlepšími, jinak 0

Tabulka 2.3: Předchozí výsledky hráčů

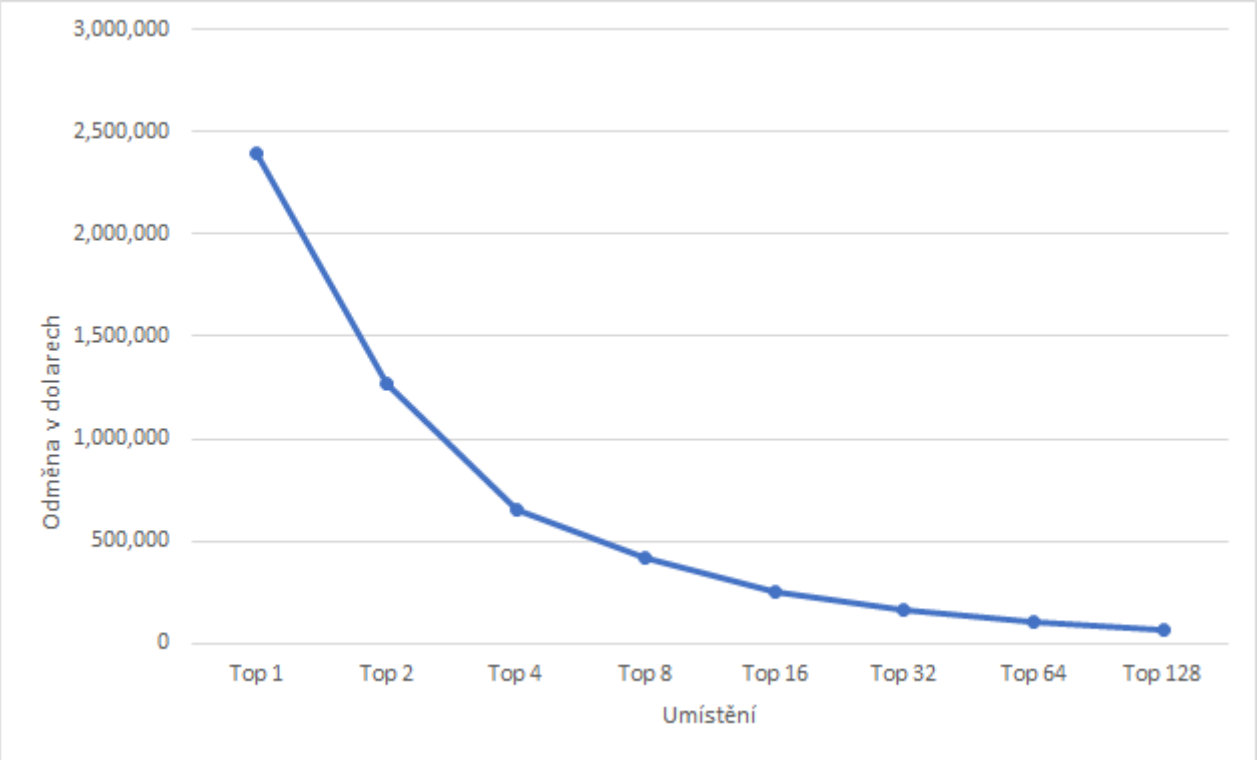
2.1.4 Probit model s žebříčkem, odměnami a demografií

U tohoto modelu[8] se autoři zaměřují mimo jiné také na odměny, které hráči dostanou za umístění v turnaji. Z grafu 2.1 můžeme vidět, že ceny nejsou rozložené lineárně, ale hráči na prvních příčkách získávají podstatně vyšší odměny než zbytek. Všechny vysvětlující proměnné použité v modelu jsou uvedeny v tabulce 2.5.
Z výsledků vyplývá, že větší rozdíly peněžních odměn mají pozitivní, statisticky významný vliv na pravděpodobnost, že favorizovaný hráč vyhraje

Název	Popis
DIFAGE	Rozdíl věku hráčů
DIFAGE2	Rozdíl čtverců věku hráčů
DIFHEIGHT	Rozdíl výšky hráčů
DIFHEIGHT2	Rozdíl čtverců výšky hráčů
BOTHRIGHT	Oba hráči jsou praváci
BOTHLEFT	Oba hráči jsou leváci
LEFTL	Hráč níže v žebříčku je levák
LEFTH	Hráč výše v žebříčku je levák

Tabulka 2.4: Fyzické vlastnosti hráče

zápas . Autoři to vysvětlují tak, že oba hráči vyvíjejí větší snahu zápas vyhrát a hráči výše postavení v žebříčku toho umí lépe využít.



Obrázek 2.1: Odměny za umístění na Wimbledonu 2021

Název	Popis
UNAGE	Věk hráče níže postaveného v žebříčku
FAVAGE	Věk hráče výše postaveného v žebříčku
FAVMATWIN	Počet výher favorita ve vzájemných zápasech
FAVMATLO	Počet proher favorita ve vzájemných zápasech
RDSLEFT	Kolik kol zbývá do konce turnaje po tomto zápase
SPECDIF	Počet vyhraných zápasů na daném povrchu
ATPPTDIF	Rozdíl postavení v žebříčku
MASTERS	1 pokud se jedná o Grand Slam nebo turnaj Masters
PDIFF	Rozdíl nejvyšší odměny a odměny, kterou dostane hráč, který prohraje daný zápas

Tabulka 2.5: Vysvětlující proměnné

2.1.5 Random forrest a GBM

Článek[9] analyzuje 39 000 tenisových zápasů mužů i žen v období od roku 2010 až 2019. Kombinují údaje o hráčích, zápasech i sázkových trzích. Autoři vyzkoušeli několik různých modelů včetně Random forrest a Gradient Boosting Machine (GBM). Oba tyto modely dosahovali na testovací množině přesnosti 69 %.

Random forrest

Random forrest je neparametrický model, který zobecňuje model náhodného stromu. Růstu každého stromu se dosahuje opakovaným rozdělováním dat v každém uzlu stromu na základě náhodně vybrané podmnožiny prvků. Proměnná, která maximalizuje informační zisk, je vybrána k dalšímu rozdělení. Konečný model je postaven z průměru mnoha stromů takto natrénovaných.

Gradient Boosting Machine

Zatímco náhodný les sestavuje soubor modelů a používá průměry jejich předpokládaných hodnot, Gradient Boosting je postaven na přidávání nových modelů do souboru sekvenčním způsobem. V každé iteraci je trénován nový model s ohledem na dosavadní chybu datasetu. Tyto nové modely jsou obvykle velmi mělké stromy. Právě maximální hloubka stromu je jendím z nejčasteji optimalizovaných parametrů.

2.2 Bodové modely

Tyto typy modelů používají velmi odlišný přístup k predikování výsledků tenisových zápasů. Pravděpodobnost výhry zápasu se počítá na základě pravděpodobnosti zisku bodu při podání obou hráčů. Předpokládá se, že se jedná o nezávislé realizace jedné náhodné veličiny. Poprvé tuto ideu popsali autoři v článku [10] z roku 2005. Hráč A může vyhrát gem, pokud dosáhne skóre 4:0, 4:1 nebo 4:2. Pokud je skóre 3:3, pak musí A vyhrát o 2 body(fifteeny). Když sečteme tyto pravděpodobnosti, tak dostáváme:

$$P_a = \sum_{j=0}^{2} P_A(4,j) + P_A(3,3) \sum_{n=0}^{\infty} P_A(n+2,n)$$

kde $P_A(i,j)$ je pravděpodobnost, že hráč A bude mít i bodů a hráč B j bodů.

Obdobně se dá vypočítat i pravděpodobnost, že hráč A vyhraje celý set. Set se skládá z gemů, kde se po každém z nich mění podávající hráč a vyhraje ten, kdo dříve dosáhne šesti vyhraných gemů. Pokud je ale skóre 5:5, pak musí jeden z hráčů získat 7 gemů, a pokud skóre dosáhne stavu 6:6 na gemy, pak se hraje takzvaný tiebreak.Typicky se v tiebreaku každý míč počítá jako jeden bod. Hráč, který jako první dosáhl sedmi bodů a má navíc nejméně dvoubodový náskok vůči soupeři (7:5), získává tuto zkrácenou hru a tím i celý set v poměru 7:6.
Pravděpodobnost výhry setu tedy můžeme zapsat jako:

$$P_a^S = \sum_{j=0}^{4} P_A^S(6,j) + P_A^S(7,5) + P_A^S(6,6) P_A^T$$

kde $P_A^S(i,j)$ je pravděpodobnost, že hráč A bude mít na konci setu i vyhraných gemů a hráč B j vyhraných gemů a P_A^T je pravděpodobnost, že hráč A vyhraje tiebrake. Autoři v této části také provádějí důkaz, že pravděpodobnost výhry setu, pokud hráč podává první, je stejná, jako když podává až jako druhý.
Obdobně můžeme spočítat i pravděpodobnosti pro výhru celého zápasu.

2.2.1 Bodový model na základě podání

Autoři[10] počítají pravděpodobnost získání bodu pouze z odhadu pravděpodobnosti úspěšného podání hráče. Vůbec zde nezáleží na tom, jak dobře je schopen podání přijmout protihráč. Pravděpodobnost, že hráč a vyhraje bod když podává, se počítá jako:

$$P_a = \frac{\alpha_a}{\beta_a}$$

kde α_a je počet vyhraných bodů hráčem a při podání za posledních dvanáct měsíců a β_a je celkový počet podání hráče a za posledních dvanáct měsíců.

2.2.2 Bodový model zohledňující sílu soupeře

Zde je potřeba zmínit další pravidlo tenisu. Pokud hráč pokazí své první podání, má možnost podávat ještě jednou. Tento model pracuje s pěti proměnnými a zahrnuje do nich výše zmíněné pravidlo. Nechť a_i = procento prvních podání ve hře pro hráče i, b_i = procento bodů vyhraných při prvním podání za předpokladu, že první podává hráč i, c_i = procento bodů získaných při druhém podání pro hráče i, d_i = procento bodů získaných při returnu prvního podání pro hráče i, e_i = procento získaných bodů při returnu druhého podání pro hráče i. Autoři[11] tyto statistiky sbírají z historických dat, a zároveň berou v úvahu, že pokud se zápas nehraje v prvním kole turnaje, tak tyto statistiky zastarávají. Z toho důvodu aktualizují statistiky i v průběhu turnaje a novější data mají větší váhu. Z těchto pěti statistik si vytvářejí dvě nové:

$$f_i = a_i b_i + (1 - a_i)c_i$$

kde f_i značí pravděpodobnost získání bodu, pokud má hráč i podání.

$$g_i = a_{av} d_i + (1 - a_{av})e_i$$

kde g_i značí pravděpodobnost výhry při returnu proti průměru 200 nejlepších hráčů z žebříčků. a_{av} je tedy průměrná pravděpodobnost prvního podání těchto 200 hráčů. Tyto dvě hodnoty ale stále nenesou informaci o tom, jak si proti sobě stojí dva určití hráči. Tuto informaci zavedeme následovně:

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av})$$

$$g_{ji} = g_t + (g_j - g_{av}) - (f_i - f_{av})$$

kde f_{ij} je procentuální zisk bodů při podání pro hráče i proti hráči j a g_{ji} je procentuální zisk bodů na returnu hráče j proti hráči i. f_t a g_t jsou průměry pro daný turnaj. Na těchto kombinovaných statistikách pak stojí celý model.

2.2.3 Markovský bodový model

Tento model[12] používá k výpočtu pravděpodobnosti vyhraného bodu Markovovské řetězce, jejichž schéma můžeme vidět na obrázku 2.2. Každý bod rozdělují na několik stavů

- 1st Serve - Hráč má první podání
- 2nd Serve - Hráč má druhé podání
- 1st Serve Rally - Hraje se míček po prvním podání
- 2nd Serve Rally - Hraje se míček po druhém podání
- Point Loss - Podávající prohrál
- Point Win - Podávající vyhrál

a přechodů mezi nimi

- P(1SN) - Při prvním podání byla síť
- P(1SF) - První podání bylo neúspěšné
- P(1SA) - První podání bylo eso
- P(1SR) - První podání bylo úspěšné a začala hra
- P(1SRW|1SR) - První podání bylo úspěšné a podávající vyhrál bod
- P(1SRL|1SR) - První podání bylo úspěšné a podávající prohrál bod
- P(2SN|1SF) - První podání bylo neúspěšné a druhé byla síť
- P(2SF|1SF) - První i druhé podání bylo neúspěšné (dvojchyba)
- P(2SA|1SF) - První podání bylo neúspěšné a druhé bylo eso
- P(2SR|1SF) - První podání bylo neúspěšné a druhé úspěšné a začala hra
- P(2SRW|2SR|1SF) - První podání bylo neúspěšné a druhé úspěšné a podávající vyhrál bod
- P(2SRL|2SR|1SF) - První podání bylo neúspěšné a druhé úspěšné a podávající prohrál bod

Z důvodu nedostatků dat a malého výskytu těchto situací, dávají autoři pravděpodobnosti sítí $P(1SN) = 0$ a $P(2SN|1SF) = 0$. Z Markovského řetězce už se dá pak velmi jednoduše zjistit pravděpodobnost vyhraného a prohraného bodu podávajícího:

$$\begin{aligned} P(\text{Podávající vyhrál}) =& P(1SA) + P(1SR)P(1SRW|1SR) + P(1SF)P(2SA|1SF) \\ &+ P(1SF)P(2SR|1SF)P(2SRW|2SR|1SF) \end{aligned}$$

$$\begin{aligned} P(\text{Podávající prohrál}) =& P(1SR)P(1SRL|1SR) + P(1SF)P(2SR|1SF)P(2SRL|2SR|1SF) \\ &+ P(1SF)P(2SF|1SF) \end{aligned}$$

Zároveň také platí, že $P(\text{Podávající vyhrál}) = 1 - P(\text{Podávající prohrál})$, protože jsou to vzájemně se vylučující události.
Aby bylo možné tyto pravděpodobnosti spočítat, je potřeba získat odhady pravděpodobností jednotlivých událostí. Tyto odhady se dají získat z historických dat a vypadají následovně:

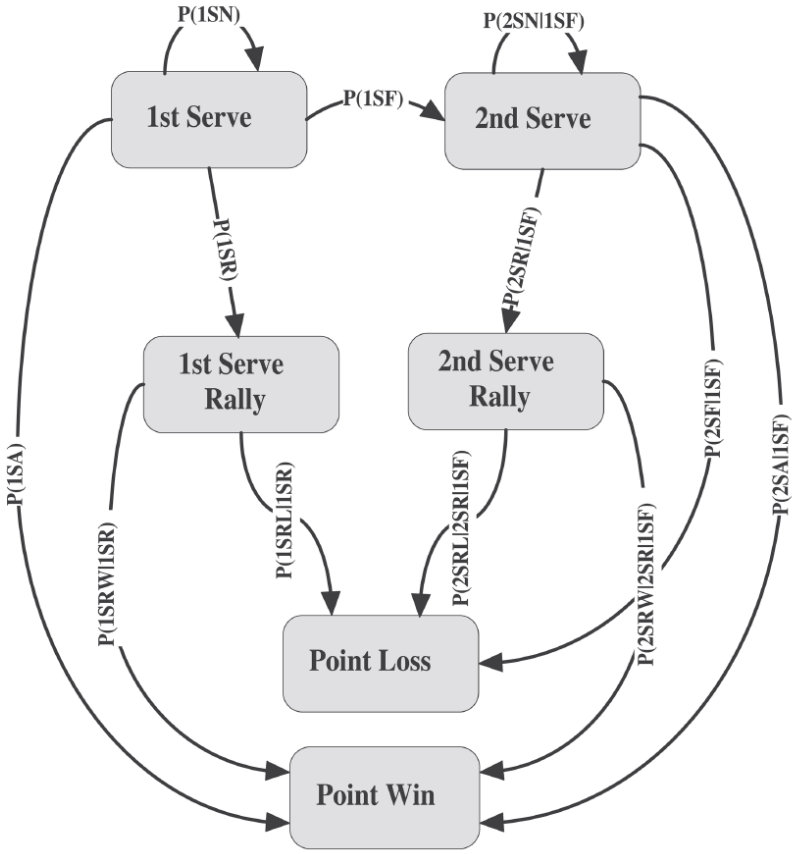
- $P(1SA) = a$
- $P(1SR) = b - a$
- $P(1SF) = 1 - b$
- $P(1SRW|1SR) = \frac{bc-a}{b-a}$
- $P(2SF|1SF) = d$
- $P(2SA|1SF) = 0$
- $P(2SR|1SF) = 1 - d$
- $P(2SR|2SR|1SF) = e$

kde

- p_{av} = Průměrný počet bodů za zápas
- z =Průměrný počet es za zápas
- $a = \frac{z}{p_{av}}$
- b =Procento prvních podání
- c =Počet vyhraných bodů po prvním podání
- y = Průměrný počet dvojchyb za zápas

- $d = \frac{y}{p_{av}}$
- e =Počet výher při druhém podání

V datech není rozděleno, kdy padlo eso při prvním, a kdy při druhém podání, a tak to autoři zjednodušují tvrzením, že všechna esa padají při prvním podání.



Obrázek 2.2: Markovský řetězec pro jeden bod

2.2.4 Bodový model na základě společných protivníků

V tomto modelu[13] je pravděpodobnost získaného bodu počítána na základě zápasů, které měli oba hráči se svými společnými protivníky. Buď A a B

hráči hrající proti sobě a C_i kde $1 <= i <= N$ jsou hráči, kterým v minulosti čelili oba soupeři. Pro každé C_i označíme $spw(A, C_i)$ jako procento vyhraných míčků při podání hráče A proti hráči C_i a $spw(B, C_i)$ jako procento vyhraných míčků hráče při podání hráče B proti hráči C_i. Obdobně $rpw(A, C_i)$ jako procento vyhraných míčků při returnu hráče A proti hráči C_i a $rpw(B, C_i)$ jako procento vyhraných míčků hráče při returnu hráče B proti hráči C_i. Pro každého společného protihráče spočítáme Δ_i^{AB}, která označuje míru výhody nebo nevýhody hráče A proti hráči B na základě C_i.

$$\Delta_i^{AB} = (spw(A, C_i) - (1 - rpw(A, C_i)) - (spw(B, C_i) - (1 - rpw(B, C_i)))$$

Tuto hodnotu lze použít k ovlivnění libovolné pravděpodobnosti zisku bodu na podání pro hráče A nebo hráče B v jakémkoliv hierarchickém modelu.

2.3 Párové srovnání

Tento typ modelů je alternativou k regresi či bodovému modelu a odvíjí se od modelu Bradley-Terry[14], který byl popsán už v roce 1952. Pokud tento model aplikujeme na tenis, s předpokladem, že pravděpodobnost vítězství hráče i nad hráčem j je $\pi_{ij} = \frac{\alpha_i}{\alpha_i + \alpha_j}$, kde α_i a α_j jsou kladné hodnoty, reprezentující, jak dobří hráči jsou.

2.3.1 Variace na Bradley-Terry model

Autoři tohoto modelu[15] vylepšují Bradley-Terry model, do kterého přidávají informace o vyhraných gemech, povrchu, a také mění váhy podle toho, jak daleko do minulosti se hry mezi hráči odehrály. Pravděpodobnost, že hráč i vyhraje g_i gemů a hráč j vyhraje g_j gemů je popsána jako

$$L(\alpha_i, \alpha_j) \propto \frac{\alpha_i^{g_i} \alpha_j^{g_j}}{(\alpha_i + \alpha_j)^{g_i + g_j}}$$

Aby zachytili i informace o povrchu a čase, tak zavádí dvě proměnné S_k a ϵ. S_k je rovna 1, pokud se zápas hrál na povrchu S a menší než 1, pokud se hrál na jiném povrchu. ϵ řídí poločas exponenciální funkce. Aby odhadli parametry pro predikce zápasů v čase t pro fixni ϵ a S_k, tak maximalizují

$$L(\alpha_i(t, S); i = 1, ..., n) = \prod_{k \in A_t} \left(\frac{\alpha_i(t, S)^{g_i} \alpha_j(t, S)^{g_j}}{(\alpha_i(t, S) + \alpha_j(t, S))^{g_i + g_j}}\right)^{exp(\epsilon(t - t_k))S_k}$$

kde k je index zápasů, t_k je čas, kdy byl zápas k odehrán, $A_t = k : t_k < t$ a n je počet hráčů v modelu. Výsledný žebříček vytvořený pomocí tohoto modelu můžeme vidět v tabulce 2.6.

Pořadí	Hráč	Ohodnocení modelem	žebříček ATP
1	R. Nadal	1.04	1
2	R. Federer	1.00	2
3	N. Djokovic	0.90	3
4	A. Murray	0.86	4
5	N. Davydenko	0.82	5
6	A. Roddick	0.81	8
7	J.M. Del Potro	0.81	9
8	R. Soderling	0.80	17
9	L. Hewitt	0.79	67
10	D. Nalbandian	0.78	11
11	R. Gasquet	0.76	25
12	D. Ferrer	0.76	12
13	J.W. Tsonga	0.76	6
14	N. Kiefer	0.74	38
15	T. Berdych	0.74	20

Tabulka 2.6: 15 nejlepších hráčů na konci roku 2008 na všech površích, spolu s odpovídajícím hodnocením ATP.

2.3.2 Variace Elo ratingu

Tento model[16] je založen na modelu popsaném v knize [17] z roku 1978, který se dodnes používá k evaluaci hráčů šachu. U původního modelu se pro odhad pravděpodobnosti výhry hráče A užívá vzorce

$$E_a = \frac{1}{1 + 10^{R_b - R_a/400}}$$

a obdobně pro hráče B

$$E_b = \frac{1}{1 + 10^{R_a - R_b/400}}$$

kde r_a je stávající rating hráče A a r_b je stávající rating hráče B. Vzorec se dá také vyjádřit jako $E_a = \frac{Q_a}{Q_a+Q_b}$ a $E_b = \frac{Q_b}{Q_a+Q_b}$, kde $Q_a = 10^{R_a/400}$ a $Q_b = 10^{R_b/400}$. Zde ale vyvstává otázka, jak zjistit R_a a R_b. Pokud hráč A ještě neměl žádný zápas, je mu přiděleno hodnoceni $R_a = 1000$. Po každé hře se provede přepočítání jeho hodnocení pomocí vzorce

$$R'_a = R_a + K(S - E_a)$$

kde R'_a je nové hodnocení, R_a je původní hodnocení, S je rovno 1, pokud hráč A vyhrál, 0,5 pro remízu a 0 pokud prohrál. A nakonec konstanta

K nazývaná K-faktor, která určuje rychlost růstu nebo poklesu hodnocení. Například mezinárodní šachová organizace FIDE používá K-faktor následujícím způsobem[18]:

- $K = 40$ pro nového hráče na ratingovém seznamu, dokud neodehraje alespoň 30 soutěžních her.
- $K = 20$, pokud hodnocení hráče s hodnocením do 2400
- $K = 16$ pro hodnocení vyšší než 2400

Autoři modelu [16] tento základní model jen lehce upravují. Přepočítání hodnocení hráče A je stejné jako v původním modelu, jen je zde volen jiný K-faktor.

$$K = 250/(m(t) + 5)^{0.4}$$

kde $m(t)$ je počet odehraných zápasů daného hráče do času t. Pokud se jedná o Grandslamové turnaje, pak násobí K-faktor ještě konstantou 1.1. Vývoj hodnoty K-faktoru můžeme pozorovat v grafu 2.3.

Pravděpodobnost výhry hráče A je pak také počítána stejně jako v originálním modelu. Jediný rozdíl je, že pokud hráč ještě neměl zápas, je mu přiděleno hodnocení $R_a = 1500$

2.4 Modely založené na odhadech bookmakerů

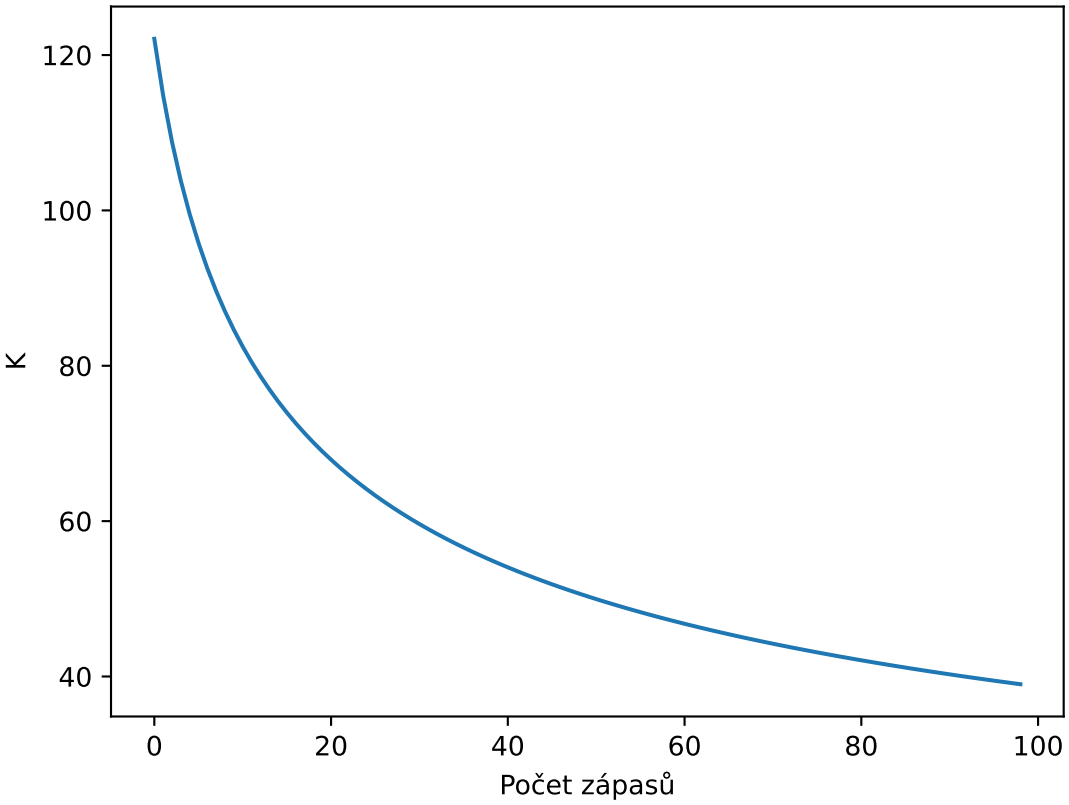
Tyto typy modelů nesouvisí čistě s tenisem, ale dají se aplikovat na jakýkoliv sport. V principu vezmeme pravděpodobnosti výsledku zápasu od bookmakerů a použijeme je jako jedinou vysvětlující proměnnou modelu. Kurzy od různých bookmakerů se mohou lišit a také se měnit podle toho, na kterou stranu více sázejí sázkaři, protože sázkové kanceláře tak mohou snižovat riziko.

2.4.1 Model konsenzu bookmakerů

Autoři tohoto modelu[19] využívají pro odhad pravděpodobnosti logit a to následujícím způsobem:

$$\frac{\pi_i}{1-\pi_i} = \frac{1}{B}\sum_{b=1}^{B}(\frac{\pi_{i,b}}{1-\pi_{i,b}})$$

kde B je počet sázkových kanceláří a $\pi_{i,b}$ je pravděpodobnost výsledku zápasu sázkové kanceláře b.



Obrázek 2.3: Vývoj K-faktoru

2.5 Porovnání

Většinu modelů popsaných v předchozích sekcích porovnává článek[5]. Autoři použily pro většinu modelů pouze data za rok 2014. Jen u modelu založeném na Elo použili kromě datasetu z roku 2014 i statistiky za celou kariéru hráčů.

Celkově použili dataset o velikosti 2395 zápasů, který je podrobněji popsán v tabulce 2.7. Kromě procentuálního počtu správně klasifikovaných zápasů je měřen i log-loss. Výsledky jsou v tabulce 2.8

Když nepočítám model založený na odhadu bookmakerů, tak nejlépe vycházejí modely párového srovnání využívající Elo. Jako baseline model ve zbytku práce tedy použiji model založený Elo. Jednak má nejlepší log-loss a také je výhodou, že jeho implementace by měla být vcelku jednoduchá.

Charakteristiky	Počet	Procenta
Celkový počet zápasů	2395	100
Typy turnajů		
Grand Slam	482	20.1
Masters	549	22.9
Ostatní	1364	57.0
Povrchy		
Antuka	790	33.0
Tráva	287	12.0
Asfalt	1318	55.0
Nejlépe hodnocení hráči		
Top 30	1235	51.6
Ostatní	1160	48.4

Tabulka 2.7: Validační dataset zápasů dvouhry ATP 2014

Model	Přesnost	Log-loss
Regresní		
Probit model s nasazením v turnajích	59 %	0.63
Probit s odměnami	68 %	0.61
Probit s žebříčkem a demografií	67 %	0.60
Logistický model	67 %	0.60
Bodové		
Na základě podání	63 %	0.67
Na základě soupeře	67 %	0.63
Markovský	64 %	0.68
Na základě společných protivníků	63 %	0.66
Párové srovnání		
Vylepšený Bradley-Terry	62 %	0.67
Elo s ročními daty	67 %	0.60
Elo s kompletními daty	70 %	0.59
Modely založené na odhadech bookmakerů		
Konsenzus bookmakerů	72 %	0.55

Tabulka 2.8: Shrnutí predikcí na validačních datech podle modelů.

Kapitola 3

Data

Historická data tenisových zápasů se mi podařilo získat ve spolupráci s firmou Ematiq a.s. Jedná se o data od roku 1993 až po únor roku 2022. Podrobnější statistiky jednotlivých zápasů, jako je počet úspěšných podání nebo počet es, je dostupný až od roku 2003.

3.1 Rozdělení a popis dat

Data původně nebyla logicky členěna a někde chyběly i identifikátory. Proto se nejprve musela rozdělit do tabulek, které spolu logicky souvisí.

3.1.1 Hráč

Tabulka 3.1 obsahuje identifikátor a základní údaje o hráčích a hráčkách.

Název	Popis
COMPETITOR_ORIGIN_ID	Identifikátor hráče
NAME	Jméno hráče
DATE_BIRTH	Datum narození hráče
PLAYER_COUNTRY	Země, kterou hráč reprezentuje
PRIZE_MONEY	Suma výher v turnajích

Tabulka 3.1: Popis tabulky Player

Jako příklad uvádím 10 nejvýdělečnější hráčů a hráček v tabulce 3.2

Jméno	Datum narození	Země	Suma výher v dolarech
Novak Djokovic	05/22/87	SRB	154,756,726
Roger Federer	08/08/81	SUI	130,594,339
Rafael Nadal	06/03/86	ESP	124,961,595
Serena Williams	09/26/81	USA	94,518,971
Andy Murray	05/15/87	GBR	62,314,306
Pete Sampras	08/12/71	USA	43,280,489
Venus Williams	06/17/80	USA	42,280,541
Maria Sharapova	04/19/87	RUS	38,777,962
Simona Halep	09/27/91	RUS	37,950,241
Caroline Wozniacki	07/11/90	DEN	35,233,415

Tabulka 3.2: 10 nejvýdělečnějších hráčů

3.1.2 Turnaj

Tabulka 3.3 turnaj už je podstatně obsáhlejší a jsou v ní důležité informace jako typ povrchu nebo typ turnaje.

Název	Popis
TOURNAMENT_ORIGIN_ID	Identifikátor turnaje
PREVIOUS_TOURNAMENT_ORIGIN_ID	Identifikátor předchozího turnaje
NAME	Název turnaje
SURFACE	Typ povrchu
TIME	Datum započetí turnaje
TOURNAMENT_RANK	Typ turnaje
COUNTRY	Země, ve které se turnaj odehrává
PRIZE_MONEY	Celková odměna rozdělená na turnaji

Tabulka 3.3: Popis tabulky Tour

Sloupec SURFACE je výčtový typ obsahující hodnoty Hard, Clay, Carpet, Grass a Acrylic s tím, že jednoznačně největší popularitu má trojice Hard(tvrdý), Clay(antuka) a Grass(trávník).

Typy turnajů ve sloupci TOURNAMENT_RANK jsou děleny následovně:

- 0 - ITF turnaje < $10K
- 1 - Challengery/ITF turnaje > $10K

- 2 - Hlavní turnaje ATP/WTA
- 3 - turnaje Masters
- 4 - Grand Slam
- 5 - Davis/Fed Cup
- 6 - Turnaje nespadající pod organizace ATP/WTA

3.1.3 Zápasy

Tato tabulka 3.4 obsahuje informace o výsledcích jednotlivých zápasů.

Název	Popis
GAME_ID	Identifikátor zápasu
COMPETITOR_ORIGIN_ID	Identifikátor hráče
TOURNAMENT_ORIGIN_ID	Identifikátor turnaje
ROUND_ORIGIN_ID	Identifikátor kola turnaje
RESULT_SET_X	výsledek setu X
RESULT_SET_X_TIEBREAKE	výsledek tiebreaku v setu X, pokud nastal
WINNER	1, pokud hráč vyhrál, 0, pokud prohrál

Tabulka 3.4: Popis tabulky Games

Každý zápas má dva záznamy pro každé GAME_ID, které se liší pouze v identifkátoru hráče a v tom, zda hráč zápas prohrál nebo vyhrál.

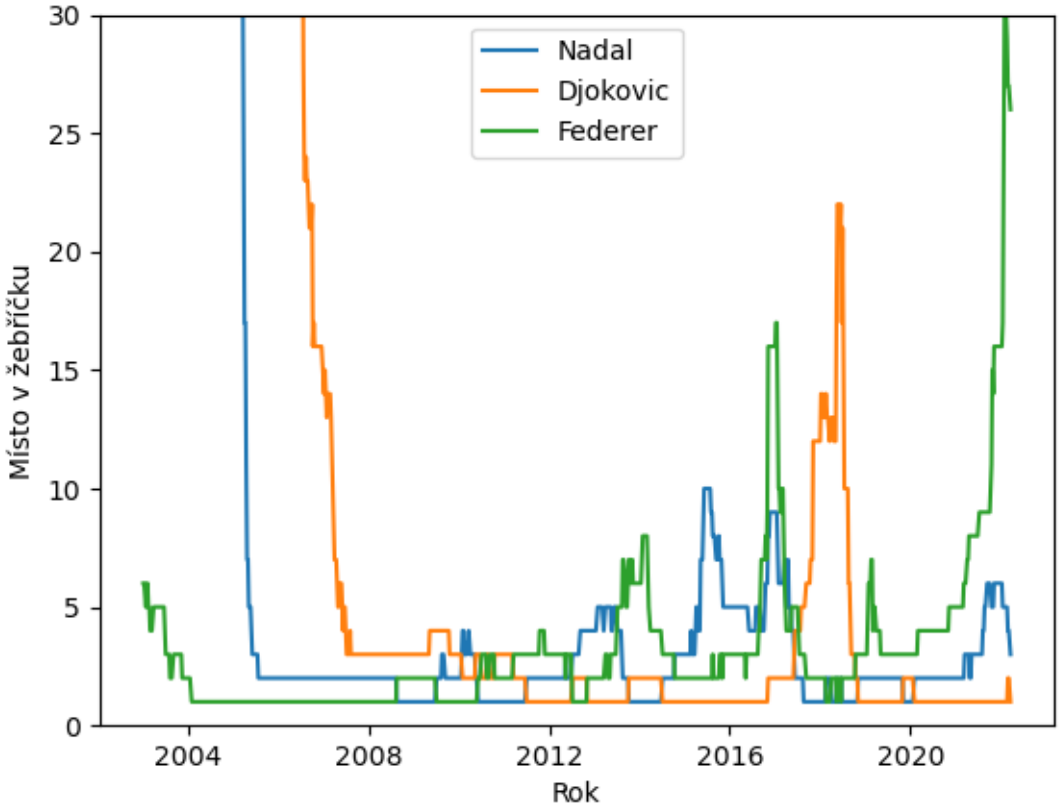
3.1.4 Hodnocení

Zde 3.5 jsou obsažené informace o hodnoceních hráčů v čase.

Název	Popis
COMPETITOR_ORIGIN_ID	Identifikátor hráče
TIME	Datum změny hodnocení X
NUM_POINTS	Hodnocení
POSITION_RANK	Pořadí v žebříčku ATP/WTA

Tabulka 3.5: Popis tabulky Rating

Jako příklad uvádím vývoj pozice tří nejznámnějších hráčů ATP, Novaka Djokoviče, Rafaela Nadala a Rodgera Federera - 3.1



Obrázek 3.1: Vývoj v žebříčku hráčů Djokovice,Federera a Nadala

3.1.5 Kurzy

Tato tabulka 3.6 obsahuje předzápasové kurzy jedné z největších světových sázkových kanceláří.

3.1.6 Statistiky

Tabulka 3.7 obsahuje statistky jednotlivých zápasů a je stěžejní pro moji další práci.

Název	Popis
COMPETITOR_ORIGIN_ID_1	Identifikátor hráče 1
COMPETITOR_ORIGIN_ID_2	Identifikátor hráče 2
TOURNAMENT_ORIGIN_ID	Identifikátor turnaje
ROUND_ORIGIN_ID	Identifikátor kola turnaje
K1	Kurz na výhru hráče 1
K2	Kurz na výhru hráče 2

Tabulka 3.6: Popis tabulky Odds

Název	Popis
COMPETITOR_ORIGIN_ID	Identifikátor hráče
TOURNAMENT_ORIGIN_ID	Identifikátor turnaje
ROUND_ORIGIN_ID	Identifikátor kola turnaje
GAME_ID	Identifikátor zápasu
WINNER	1, pokud hráč vyhrál, 0, pokud prohrál
FS_1	Počet úspěšně zahraných prvních podání hráče
FSOF_1	Celkový počet podání hráče
ACES_1	Počet es
DF_1	Počet dvojchyb
UE_1	Počet nevynucených chyb
W1S_1	Počet vyhraných bodů při prvním podání
W1SOF_1	Počet úspěšně zahraných prvních podání hráče
W2S_1	Počet vyhraných bodů při druhém podání
W2SOF_1	Počet druhých podání hráče
BP_1	Počet proměněných breakpointů
BPOF_1	Celkový počet možných breakpointů
TPW_1	Celkový počet vyhraných míčků
RPW_1	Počet vyhraných míčků na příjmu
RPWOF_1	Celkový počet míčků na příjmu

Tabulka 3.7: Popis tabulky Stats

V tabulce 3.8 můžeme vidět, jak které proměnné (vydělené počtem podání) korelují s vítězstvím v zápase.

Na tenise je specifické, že i přestože hráč vyhraje víc míčků než protihráč, může stále prohrát zápas. Z grafu 3.2 vidíme, že se jedná o 4.4 % zápasů.

Proměnná	Korelace
BP_1	0.605
RPW_1	0.467
TPW_1	0.438
W1S_1	0.272
W2S_1	0.271
ACES_1	0.222
UE_1	-0.133
DF_1	-0.181

Tabulka 3.8: Korelace proměnných s vítězstvím v zápase

3.2 Předzpracování

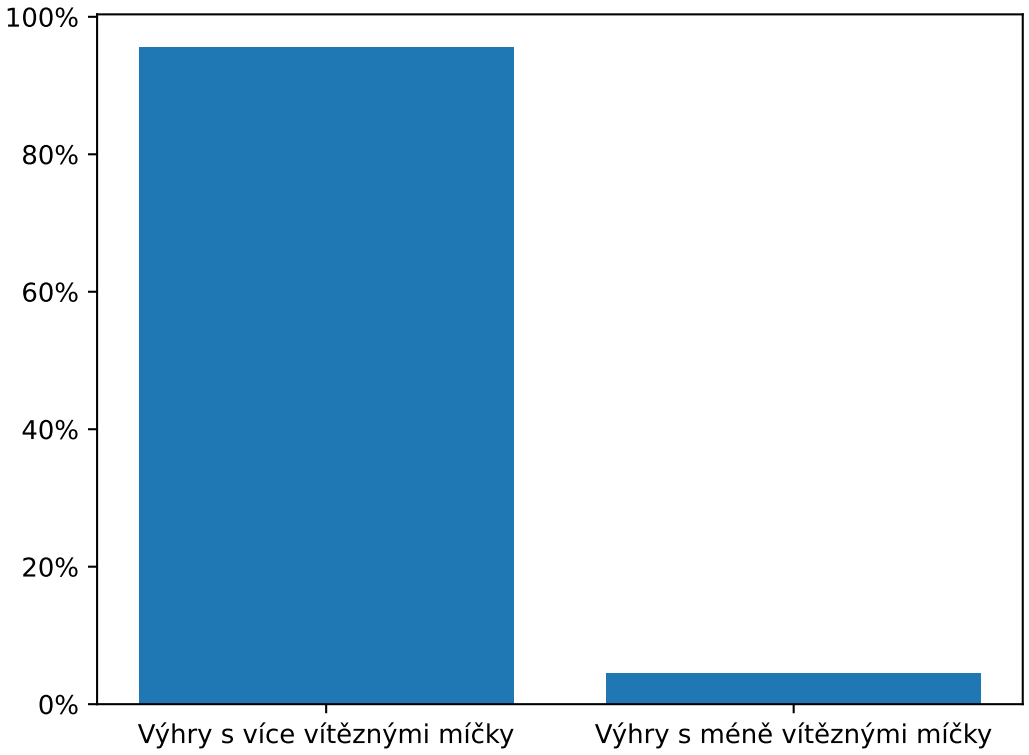
Pro napočítání vysvětlujících proměnných pro svůj model jsem použil data z tabulek obsahujících statistiky jednotlivých her, informace o hráčích a informace o turnajích. Tyto tabulky jsem spojil do jednoho datového souboru podle identifikátoru hráčů a turnajů.

3.2.1 Základní přiznaky

Bylo potřeba přepočítat některé proměnné na procenta, protože zápasy jsou různě dlouhé. Procenta byla napočítána pro sloupce s esy, breakbally, vyhranými podáními na příjmu, vyhranými podáními při prvním a při druhém podání. Také bylo potřeba vytvořit sloupec s celkovým počtem vítězných podání za pomocí vzorce:

$$p = \frac{w1s + w2s}{w1sof + w2sof}$$

kde $w1s$ je počet vyhraných prvních podání, $w2s$ je počet vyhraných druhých podání, $w1sof$ je celkový počet prvních podání a $w2sof$ je celkový počet druhých podání. Toto je extrémně důležitá statistika, kterou využívají bodové modely pro predikce pravděpodobností vítěze zápasu. Na závěr předzpracování bylo potřeba setřídit data podle identifikátoru hráče, času a identifikátoru kola, aby šla data korektně za sebou podle toho, jak každý hráč hrál své zápasy. Tyto statistiky budou dále nazývány **realized**.



Obrázek 3.2: Statistika zápasů, kde hráč vyhrál méně míčků než oponent a přesto vyhrál zápas

3.2.2 Příznaky soupeře

Velkou přidanou hodnotu mají také informace, jak dobře nebo špatně si vedl v zápase protivník. Stačilo spojit již vypočítané realized statistiky přes game_id. Tyto statistky jsou dále značené jako **allowed**.

3.2.3 Přizpůsobené příznaky

Další statistika, které se mi již v minulosti osvědčila, je odečítání rozšiřujícího průměru allowed příznaků soupeře od realized příznaků hráče a naopak rozšiřujícího průměru realized příznaků soupeře od allowed přízaků hráče. Zdvojnásobí se tak sice počet přízna, ale získávám tak informaci, jak dobře nebo špatně si hráč vedl oproti průměru předchozích zápasů soupeře. Tedy

například rozdíl počtu es, které hráč soupeři dal a počtu es, které v průměru historicky soupeř za zápas dostal. Tyto statistky nazývám **adjusted**.

3.3 Výpočet předzápasových statistik

V tuto chvíli bylo připraveno 52 vysvětlujících proměnných pro každý zápas. 13 realized, 13 allowed, 13 realized_adjusted a 13 allowed_adjusted.

Pro učení modelů a pro predikce jsem musel napočítat statistiky, se kterými hráči do každého zápasu vstupují. Vyzkoušel jsem 3 různé přístupy a všechny pak použil při učení modelů.

3.3.1 Klouzavý průměr

Použil jsem klouzavý průměr, protože odráží aktuální formu hráčů. Využívám vzorce:

$$MA_k = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i$$

kde n je celkový počet zápasů, k je takzvané okno klouzavého průměru a p_i je datový bod. Čím užší je okno klouzavého průměru, tím více statistika odráží aktuální formu hráče. Zkusil jsem použít $k \in \{3, 5, 10, 25\}$.
Dále jsem bral v potaz, že různí hráči hrají rozdílně podle typu povrchu, a proto jsem napočítal klouzavé průměry i přes jednotlivé povrchy. Zde jsem volil pouze $k \in \{3, 5\}$.

3.3.2 Rozšiřující průměr

Jako další jsem zkusil využít rozšiřující průměr, který na rozdíl od klouzavého průměru nepopisuje aktuální formu, ale naopak formu za celou kariéru. Pro každý zápas se tedy statistiky počítají jako průměr všech doposud známých dat. Jedná se tedy o klouzavý průměr, kde k je vždy rovno počtu všech dosavadních zápasů. Stejně tak jsem tento průměr použil pro napočítání statistik přes jednotlivé povrchy.

3.3.3 Vážený klouzavý průměr

Jako poslední jsem vyzkoušel vážený klouzavý průměr. Ten pro výpočet aktuální statistiky využívá vzorce:

$$y_t = \frac{x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + ... + (1-\alpha)^t x_{t_0}}{1 + (1-\alpha) + (1-\alpha)^2 + ... + (1-\alpha)^t}$$

Pro výpočet parametru α jsem vyzkoušel dva různé přístupy. Jednou jsem volil α jako konstantu, kdy $\alpha = 0.95$ a podruhé použil:

$$\alpha = 1 - exp(-ln(2)/halflife)$$

kde *halflife* určuje počet zápasů, kdy se váha dostane na polovinu. Pro počítání statistik přes jednotlivé povrchy jsme použili $halflife = 10$ a pro celkové statistiky $halflife = 30$.

Kapitola 4

Návrh a implementace modelu

V této kapitole budu popisovat modely, které byly použity pro experimenty ve zbytku práce. Rozhodli jsme se vyzkoušet dva různé přístupy. Základem bylo samozřejmě rovnou predikovat pravděpodobnost výhry, ale také jsme využili toho, že v datech máme pravděpodobnost výhry při podání, a tak se naskytla možnost predikovat tyto pravděpodobnosti pro oba hráče a samotnou pravděpodobnost výhry pak počítat pomocí bodového modelu. Vydali jsme se cestou neuronových sítí.

4.1 Umělé neuronové sítě

Neuronová síť se skládá ze vstupní vrstvy neuronů, několika skrytých vrstev neuronů a výstupní vrstvy neuronů. Neurony jsou propojeny a každému z těchto propojení je dána číselná hodnota nazývaná *váha*. Schéma neuronové sítě můžeme vidět na obrázku 4.1.

Výstup h_i neuronu i ve skryté vrstvě počítáme jako:

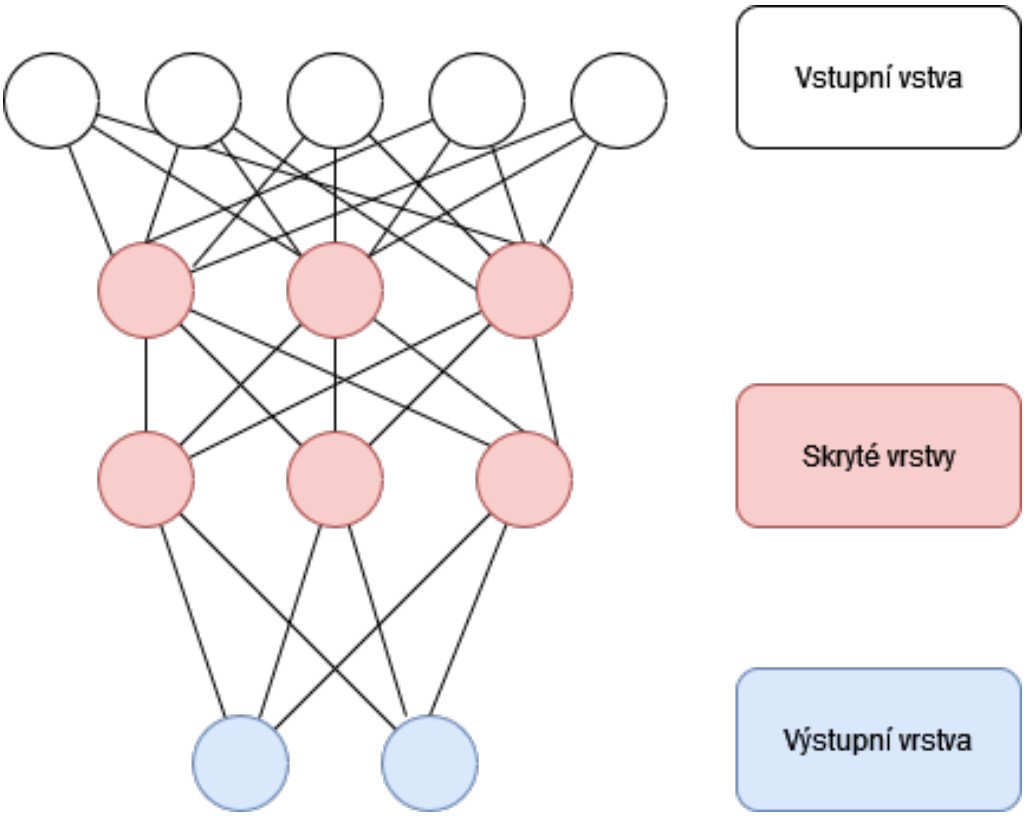
$$h_i = \sigma(\sum_{j=1}^{N} V_{ij} x_j + T_i^{hid})$$

kde σ je takzvaná aktivační funkce, N je počet vstupních neuronů, V_{ij} váhy, x_j vstupy do vstupních neuronů a nakonec T_i^{hid} je práh, který určuje, jestli je neuron aktivní nebo ne[20].

Účelem aktivační funkce je , kromě zavedení nelinearity do sítě, ohraničit hodnotu neuronu. V této práci byly používané tři různé a to:

Sigmoida

$$\sigma(u) = \frac{1}{1 + exp(-u)}$$



Obrázek 4.1: Schéma neuronové sítě

Rectified Linear Unit (ReLU)

$$\sigma(u) = max(0, u)$$

Hyperbolický tangens

$$\sigma(u) = \tanh(u) = \frac{\exp(u) - \exp(-u)}{\exp(u) + \exp(-u)}$$

4.2 Vstupní a výstupní data

Neuronové sítě jsem zkoušel učit na různých statistikách, popsaných v předchozí kapitole.

Vybral jsem 8 datasetů jak pro muže, tak pro ženy:

- ewm - Vážený klouzavý průměr bez ohledu na povrch
- ewm_surface - Vážený klouzavý průměr přes jednotlivé povrchy

- ewm_all - Vážený klouzavý průměr přes jednotlivé povrchy i bez nich
- expanding - Rozšiřující průměr bez ohledu na povrch
- expanding_surface - Rozšiřující průměr přes jednotlivé povrchy
- expanding_all - Rozšiřující průměr přes jednotlivé povrchy i bez nich
- rolling_10 - Klouzavý průměr bez ohledu na povrch s oknem velikosti 10
- rolling_5 - Klouzavý průměr přes jednotlivé povrchy s oknem velikosti 5

Pro každý z těchto datasetů libovolně kombinujeme jednotlivé typy příznaků:

- realized - Základní přiznaky
- allowed - Přiznaky soupeře
- realized_adjusted - Přizpůsobené základní přiznaky
- allowed_adjusted - Přizpůsobené přiznaky soupeře

Jak už bylo zmíněno v úvodu této kapitoly, tak jsem učil neuronové sítě buďto s jedním neuronem ve výstupní vrstvě a predikovali přímo pravděpodobnost výhry, a nebo byly ve výstupní vrstvě dva neurony a pak jsme síť trénovali tak, aby predikovala pravděpodobnosti výhry míčku při podání obou hráčů. Pro oba tyto přístupy se musely použít různé ztrátové funkce. Při predikování výhry jsme využili binární křížovou entropii (log-loss):

$$L_{bce}(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

kde N je počet datových bodů, $y = 1$, pokud hráč vyhrál a $y = 0$, pokud prohrál a p(y) pravděpodobnost toho, že hráč vyhrál.

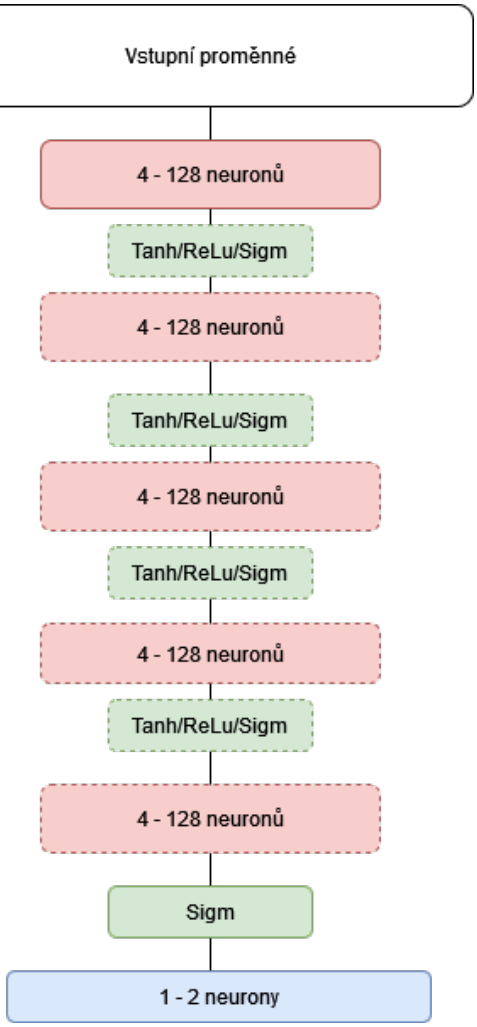
Pro druhý případ jsme využili průměrnou absolutní chybu:

$$L_1(q) = \sum_{i=1}^{N} |y_{true} - y_{predicted}|$$

kde N je počet datových bodů, $y_{predicted}$ je predikovaná hodnota a y_{true} je reálná hodnota.

4.3 Architektura

Vstupní vrstva se sestává z výše popsaných vstupních proměnných. Dále následuje 1-5 lineárních vrstev, kdy každá obsahuje 4-128 neuronů. Mezi každou z těchto vrstev je aplikována aktivační funkce. Testovali jsme i modely bez aktivačních funkcí. Před výstupní vrstvou, která je tvořena jedním nebo dvěma neurony podle typu úlohy, je aplikována Sigmoida, aby výstupy byly vždy v rozmezí (0, 1). Mezi skrytými vrstvami je vždy použita stejná aktivační funkce. Návrh architektury můžeme vidět v diagramu 4.2.



Obrázek 4.2: Architektura neuronové sítě

4.4 Optimalizační algoritmy

Optimalizační algoritmy slouží k postupnému snižování ztrátové funkce. Upravují atributy neuronové sítě, jako jsou váhy a rychlost učení. Pomáhají tedy zlepšit přesnost predikce. Dva nejpopulárnější, které jsme taky použili v této práci, jsou Stochastický gradientní sestup (SGD) a Adam.

SGD

Gradientní sestup je iterativní optimalizační algoritmus pro nalezení lokálního minima diferencovatelné funkce. Myšlenkou metody je posouvat se z výchozího bodu po krocích v opačném směru gradientu funkce v daném bodě, protože to je směr nejstrmějšího klesání její hodnoty. Algoritmus stochastického gradientního sestup (SGD) je zjednodušením. Namísto přesného výpočtu gradientu každá iterace odhaduje tento gradient na základě gradientu, který vypočítá jen z náhodně vybrané podmnožiny[21].

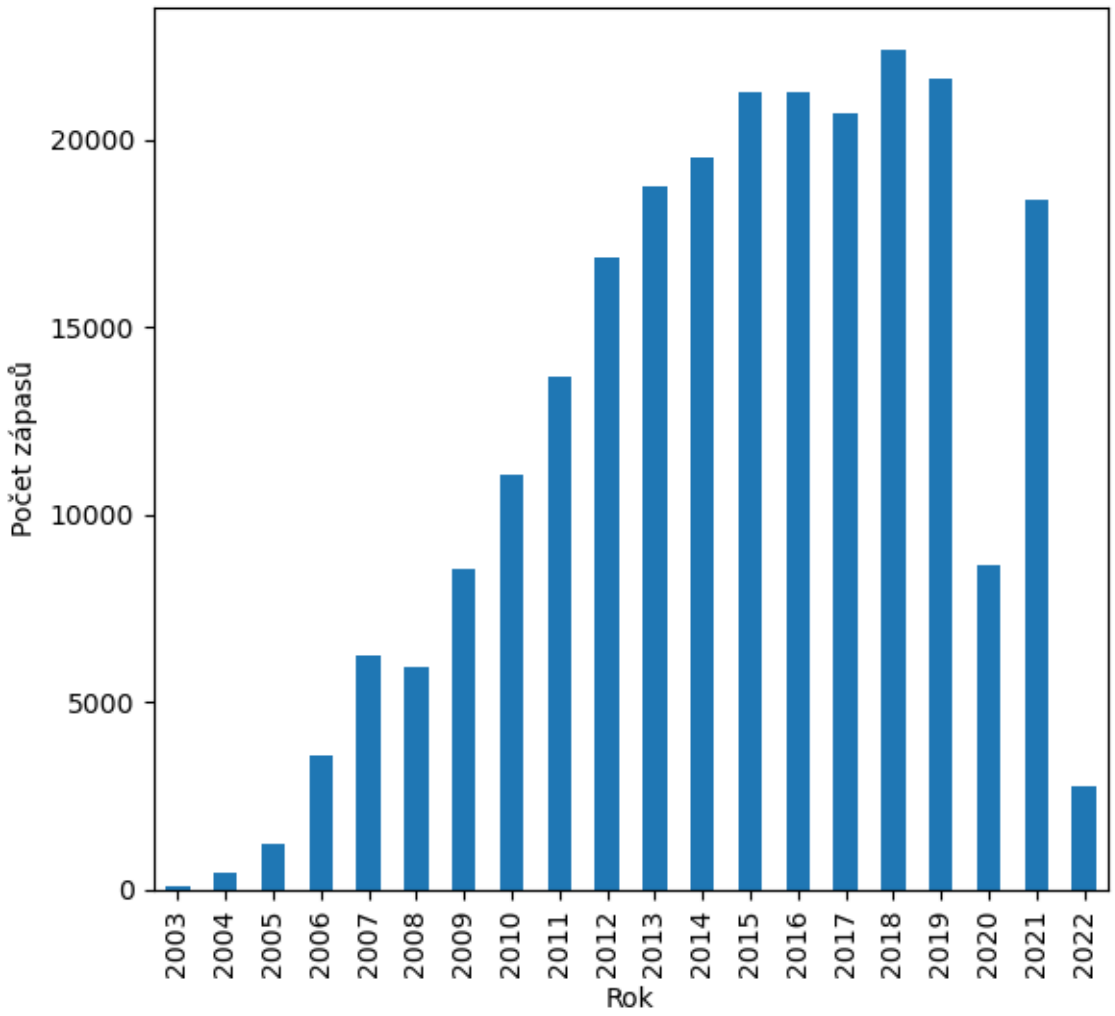
Adam

Adam je metoda pro efektivní stochastickou optimalizaci, která vyžaduje pouze gradienty prvního řádu. Metoda počítá individuální adaptivní rychlosti učení pro různé parametry z odhadů prvního a druhého momentu gradientu[22].

4.5 Trénování modelu

Typicky se trénování modelu ve strojovém učení provádí pomocí rozdělení datové sady na dvě nezávislé podmnožiny, které nazýváme trénovací a validační. Poté je model trénován na trénovací sadě a vyhodnocen na sadě validační. Datasety obsahují necelých 250 000 datových bodů, ale 70 000 z nich jsme vyhradili do testovací sady. Trénovací sada tedy končí na konci roku 2015 a obsahuje 127 000 datových bodů a validační sada obsahuje roky 2016 a 2017 a obsahuje 42 000 datových bodů. Zbytek spadá do testovací sady. V grafu 4.3 můžeme vidět počty zápasů v jednotlivých letech.

Samotné trénování probíhá v cyklech. Každé iteraci říkáme epocha. Učení se zastaví ve chvíli, kdy uběhne 200 epoch a nebo pokud se chyba na validační množině po 10 epoch nesnížila.



Obrázek 4.3: Počty zápasů v jednotlivých letech

Kapitola 5

Experimenty a vyhodnocení

Rozhraní pro trénink neuronové sítě bylo připraveno, a tak jsem mohl začít s učením. Prostor, z kterého se vybíraly parametry byl extrémně velký, ale i přesto jsem zvažoval použít *grid search*, tedy všechny možné kombinace parametrů. Celý stavový prostor je popsán v tabulce 5.1.

Jak lze z tabulky vyčíst, je celkový počet kombinací:

$$8 \cdot (2^4 - 1) \cdot 2 \cdot \sum_{i=1}^{5} 124^i \cdot 4 = 6.65374878 \cdot 10^{21}$$

Po tomto zjištění jsem se rozhodl cestou grid searche nevydávat a nesnažit se omezovat prostor hyperparametrů a místo toho jsem vyzkoušel optimalizační nástroj Optuna.

Optuna

Optuna je open-source black-box optimalizační software, který se snaží najít nejlepší hyperparametry pro daný problém[23]. Samotná práce s tímto softwarem je vcelku jednoduchá. Nejdříve je potřeba definovat *cíl*, což je prostor hyperparametrů, na kterých se optimalizace provádí, a funkce, která na základě vstupních parametrů vrací výsledek. V mém případě je výsledkem chyba na validační množině a cílem je tuto chybu minimalizovat, nebo přesnost a pak je cílem tuto hodnotu maximalizovat. Pak už je potřeba jen nadefinovat počet experimentů, tedy kolikrát se má Optuna snažit najít nejlepší hyperparametry pro daný problém.

Název	Hodnoty
Dataset	ewm ewm_surface ewm_all expanding expanding_surface expanding_all rolling_10 rolling_surface_5
Statistiky	realized allowed realized_adjusted allowed_adjusted
Optimalizátor	Adam SGD
Počet vrstev	1-5
Počet neuronů v každé vrstvě	4-128
Aktivační funkce	ReLu Tanh sigmoida bez aktivační funkce

Tabulka 5.1: Hyperparametry

5.1 Experimenty

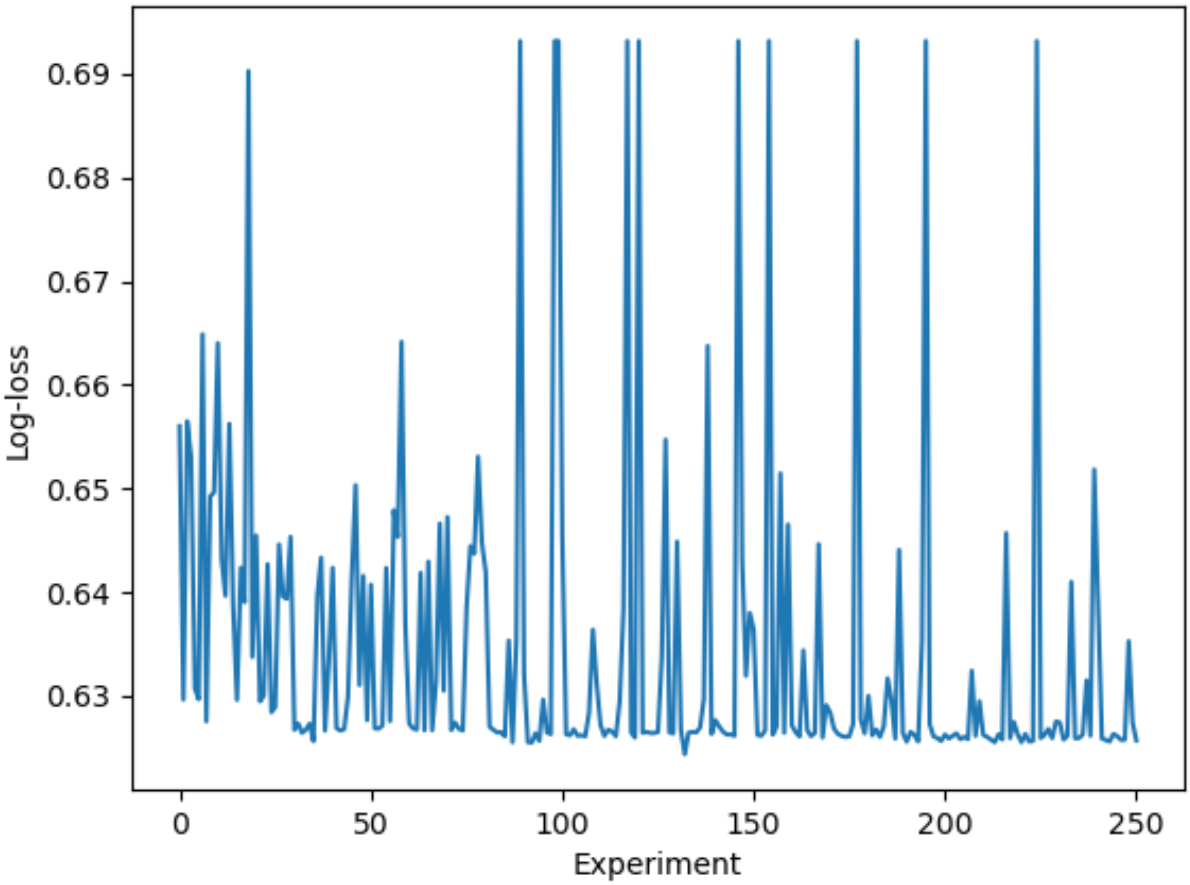
Experimenty jsem rozdělil na dvě části. Jednu neuronovou síť jsem trénoval čistě na predikce pravděpodobnosti výsledků utkání a druhou na predikce pravděpodobností zisku bodu při podání.

5.1.1 Predikce pravděpodobnosti výsledků utkání

Při učení tohoto modelu používá neuronová síť vždy jako ztrátovou funkci binární křížovou entropii, ale pro optimalizaci hyperparametrů jsem vyzkoušel kromě této ztrátové funkce také *přesnost*, tedy procento správně predikovaných zápasů. Tato metrika nebývá vždy vypovídající kvůli procentuálnímu rozložení výsledků v datasetu, ale jelikož v mém datasetu je přesně 50 % výher a 50 % proher, mohl jsem tuto metriku použít.

Užití binární křížové entropie pro hyperparametrovou optimalizaci

Za pomoci binární křížové entropie bylo provedeno 250 experimentů. Model se na validační množině už od začátku dostával pod chybu 0.63 a už ve 133. experimentu se dostal do minima na 0.6243. Poté se ještě několikrát přiblížil, ale tuto hranici už nepřekonal. Vývoj chyby na validační množině je zanesen do grafu 5.1.



Obrázek 5.1: Vývoj binární křížové entropie na validační množině při klasifikaci

Z tabulky 5.2 můžeme vidět, že ve velké většině případů se nepoužívala aktivační funkce a navíc často jen s jednou skrytou neuronovou vrstvou.

I přesto sítě tohoto typu byly schopny dosáhnout na validační množině relativně dobrých výsledků.

Název	Hodnoty	Počet
Dataset	ewm	15
	ewm_surface	9
	ewm_all	181
	expanding	10
	expanding_surface	9
	expanding_all	9
	rolling_10	9
	rolling_surface_5	9
Počet vrstev	1	206
	2	21
	3	11
	4	8
	5	5
Aktivační funkce	ReLu	13
	Tanh	13
	sigmoida	13
	bez aktivační funkce	212

Tabulka 5.2: Distribuce parametrů v experimentech pro klasifikaci při použití binární křížové entropie

Hyperparametry nejlepšího experimentu jsou zaneseny v tabulce 5.3.

Užití přesnosti pro hyperparametrovou optimalizaci

U těchto experimentů nebylo cílem Optuny minimalizovat chybu, ale naopak maximalizovat přesnost. Tímto způsobem bylo provedeno dalších 250 experimentů. Vývoj přesnosti na validační množině je zanesen do grafu 5.2. Z grafu bylo odfiltrováno 21 experimentů, které měly přesnost nižší než 62% a znepřehledňovaly tak graf.

Z tabulky5.4 můžeme vidět, že se optimalizační software vydal úplně jiným směrem než v předchozím případě. Namísto jednovrstvých sítí používá nejhlubší možné, tedy pětivrstvé sítě. Velmi podobně naopak vybírá z datasetů i aktivačních funkcí.

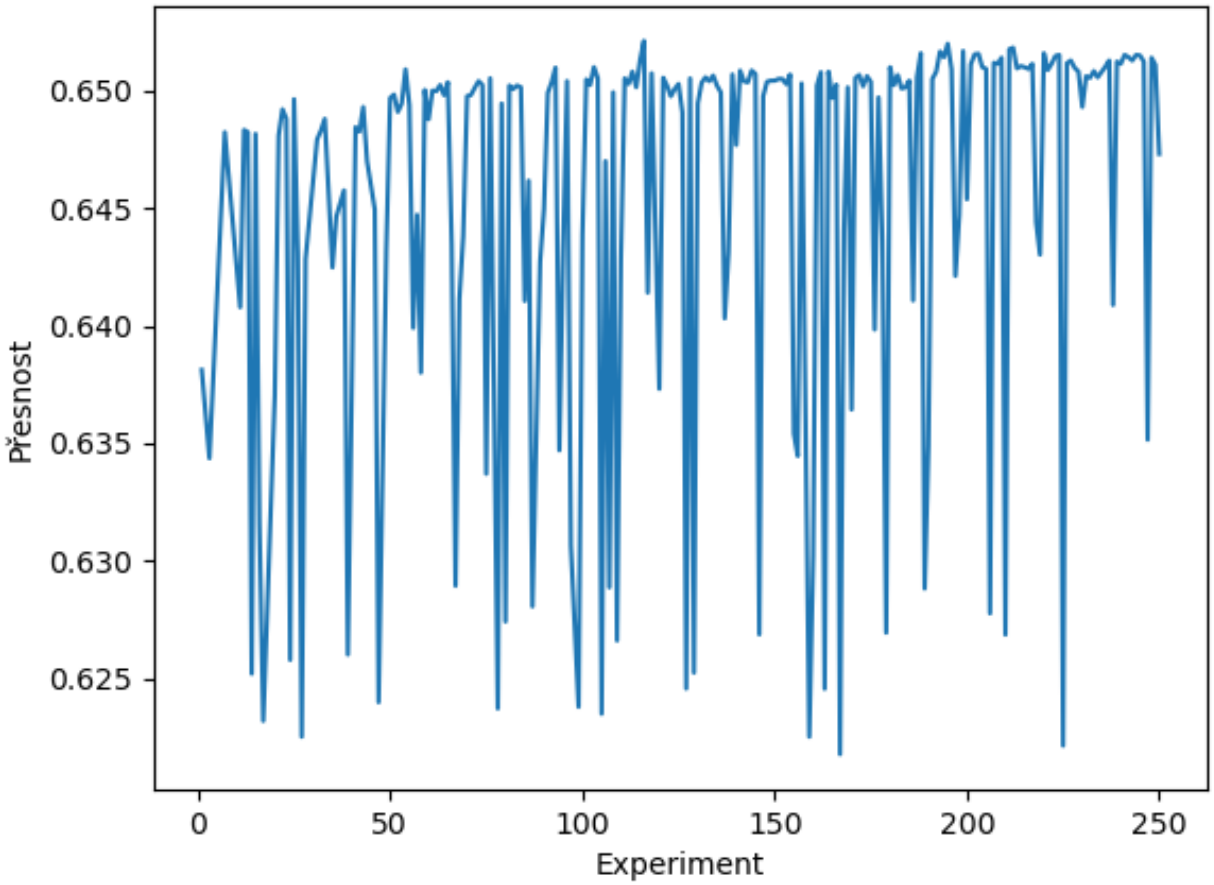
Hyperparametry nejlepšího experimentu jsou zaneseny v tabulce 5.5. Nejlepší výsledek na validační množině byl nakonec 65.21%.

Název	Hodnoty
Dataset	ewm_all
Statistiky	realized allowed realized_adjusted allowed_adjusted
Optimalizátor	Adam
Počet vrstev	1
Počet neuronů v jednotlivých vrstvách	7
Aktivační funkce	bez aktivační funkce

Tabulka 5.3: Hyperparametry nejlepšího experimentu pro klasifikaci při použití binární křížové entropie

Název	Hodnoty	Počet
Dataset	ewm	26
	ewm_surface	9
	ewm_all	173
	expanding	8
	expanding_surface	9
	expanding_all	10
	rolling_10	8
	rolling_surface_5	8
Počet vrstev	1	11
	2	27
	3	11
	4	10
	5	192
Aktivační funkce	ReLu	13
	Tanh	12
	sigmoida	13
	bez aktivační funkce	213

Tabulka 5.4: Distribuce parametrů v experimentech pro klasifikaci při použití přesnosti



Obrázek 5.2: Vývoj přesnosti na validační množině při klasifikaci

5.1.2 Predikce pravděpodobností zisku bodu při podání

U modelování pravděpodobností zisku bodu při podání jsem udělal 500 různých experimentů. Ztrátová funkce pro optimalizaci byla použita průměrná absolutní chyba, stejně jako pro učení modelu samotného. Z grafu 5.3 je patrné, že se optimalizační software neustále zlepšoval a nejlepší řešení našel až v 496. experimentu. Z tabulky 5.6 vidíme, že se při optimalizaci hyperparametrů optimalizační software opět vydal jinou cestou než v předchozích měřeních. Tentokrát se nejvíce snažil využívat dvouvrstvých neuronových sítí s aktivační funkcí ReLu.

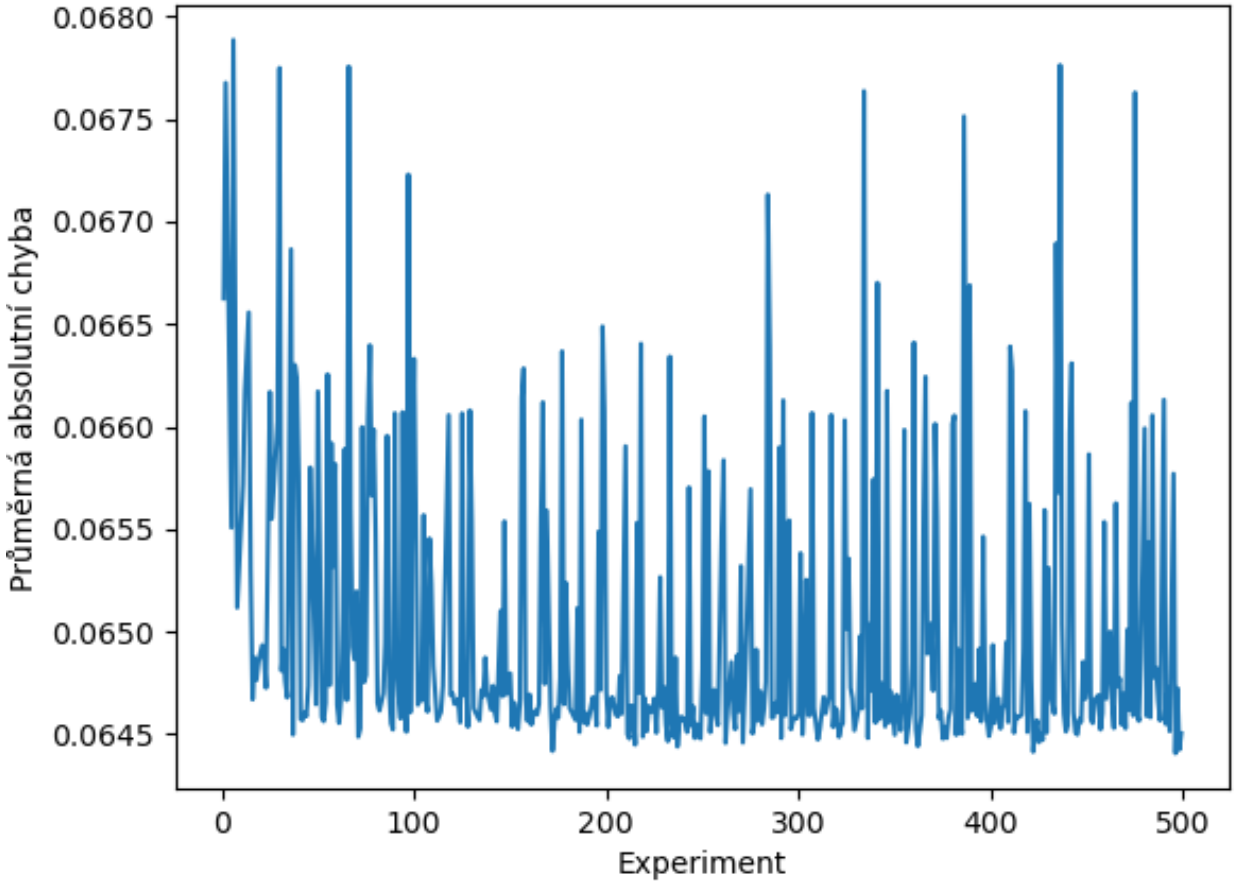
Název	Hodnoty
Dataset	ewm_all
Statistiky	realized allowed allowed_adjusted
Optimalizátor	Adam
Počet vrstev	5
Počet neuronů v jednotlivých vrstvách	128 117 117 90 106
Aktivační funkce	bez aktivační funkce

Tabulka 5.5: Hyperparametry nejlepšího experimentu pro klasifikaci při použití přesnosti

Hyperparametry nejlepšího experimentu jsou zaneseny v tabulce 5.7. Ačkoliv se nejvíce používaly neuronové sítě s dvěma skrytými vrstvami, tak nejlepší výsledek byl se sítí, které měla pouze jednu vrstvu. Tento výsledek byl 0.0644.

5.1.3 Shrnutí experimentů

Celkově bylo provedeno 1000 experimentů rozdělených do tří různých částí. Ve všech třech částech docházelo k postupné konvergenci do lokálních extrémů, ačkoliv se optimalizační software vydával velmi odlišnými směry, co se týče architektury neuronových sítí. Zajímavé je, že ve všech třech případech byly nejpoužívanější statistiky napočítané pomocí váženého klouzavého průměru. Stejně tak se ve většině případů použil optimalizátor ADAM. Hodnoty všech experimentů jsou uloženy ve složce Experiments na přiloženém datovém médiu.



Obrázek 5.3: Vývoj průměrné absolutní chyby na validační množině při predikování pravděpodobnosti zisku bodu při podání

Název	Hodnoty	Počet
Dataset	ewm	17
	ewm_surface	19
	ewm_all	377
	expanding	19
	expanding_surface	17
	expanding_all	19
	rolling_10	17
	rolling_surface_5	17
Počet vrstev	1	15
	2	353
	3	82
	4	33
	5	17
Aktivační funkce	ReLu	420
	Tanh	24
	sigmoida	22
	bez aktivační funkce	34

Tabulka 5.6: Distribuce parametrů v experimentech pro predikce pravděpodobnosti zisku bodu při podání

Název	Hodnoty
Dataset	ewm_all
Statistiky	realized allowed allowed_adjusted
Optimalizátor	Adam
Počet vrstev	1
Počet neuronů v jednotlivých vrstvách	78
Aktivační funkce	ReLu

Tabulka 5.7: Hyperparametry nejlepšího experimentu pro predikce pravděpodobnosti zisku bodu při podání

Kapitola 6

Výsledky

V této kapitole popíši, jak byly vybrány nejlepší modely podle experimentů a jejich výsledky na testovací množině dat. Také popíši výsledky dalšího modelu, který jsem vybral na základě rešerše v druhé kapitole, který má sloužit jako ukazatel toho, jak dobře si vedou mnou vymyšlené a implementované modely. Testovací množinu tvoří dataset od roku 2018 až do současnosti.

6.1 Výsledky srovnávacího modelu

Jako srovnávací model jsem využil nejlépe performující model z rešerše, tedy modelu Elo - ve zdrojových kódech *baseline_model.py*. Tento ratingový model aktualizuje hodnocení hráčů po každém zápase. Na testovací množině měl model

- **log-loss** = 0.6474
- **přesnost** = 62.26 %

6.2 Výsledky klasifikačního modelu

V předchozí kapitole je popsáno, jak jsem při optimalizaci hyperparametrů využil dva různé přístupy a získal tak dvě různé množiny hyperparametrů. Musel jsem tedy model, u kterého jsem počítal pouze s přesností, naučit znovu a tentokrát zjistit ještě log-loss, abych mohl oba modely porovnat. U modelu, jehož hyperparametry Optuna původě optimalizovala na přesnost, dosáhl log-loss = 0.6260. Model, jehož hyperparametry Optuna optimalizovala přímo na log-loss, dosáhl výsledku 0.6243, byl tedy o 1.7 tisíciny

přesnější, a proto jsem tyto hyperparametry použil pro trénování finálního modelu. Na testovací množině dosahoval model výsledků:

- **log-loss** = 0.6408
- **přesnost** = 63.24 %

6.3 Výsledky modelu predikujícího pravděpodobnost zisku bodu při podání

Výsledkem tohoto modelu jsou pravděpodobnosti zisku bodu při podání pro oba hráče a jako takové nám neříkají, jakou pravděpodobnost mají hráči na výhru celého zápasu. Bylo tedy potřeba na tyto pravděpodobnosti aplikovat bodový model, jehož výstupem byly právě pravděpodobnosti výhry. Využil jsem již existující implementaci tohoto modelu, který pro výpočet potřeboval pouze pravděpodobnosti pro oba hráče a počet setů. Po aplikaci dosáhl model těchto výsledků:

- **log-loss** = 0.6537
- **přesnost** = 63.90 %

Model tedy dosahoval horší log-loss, ale překvapivě byl nejpřesnější.

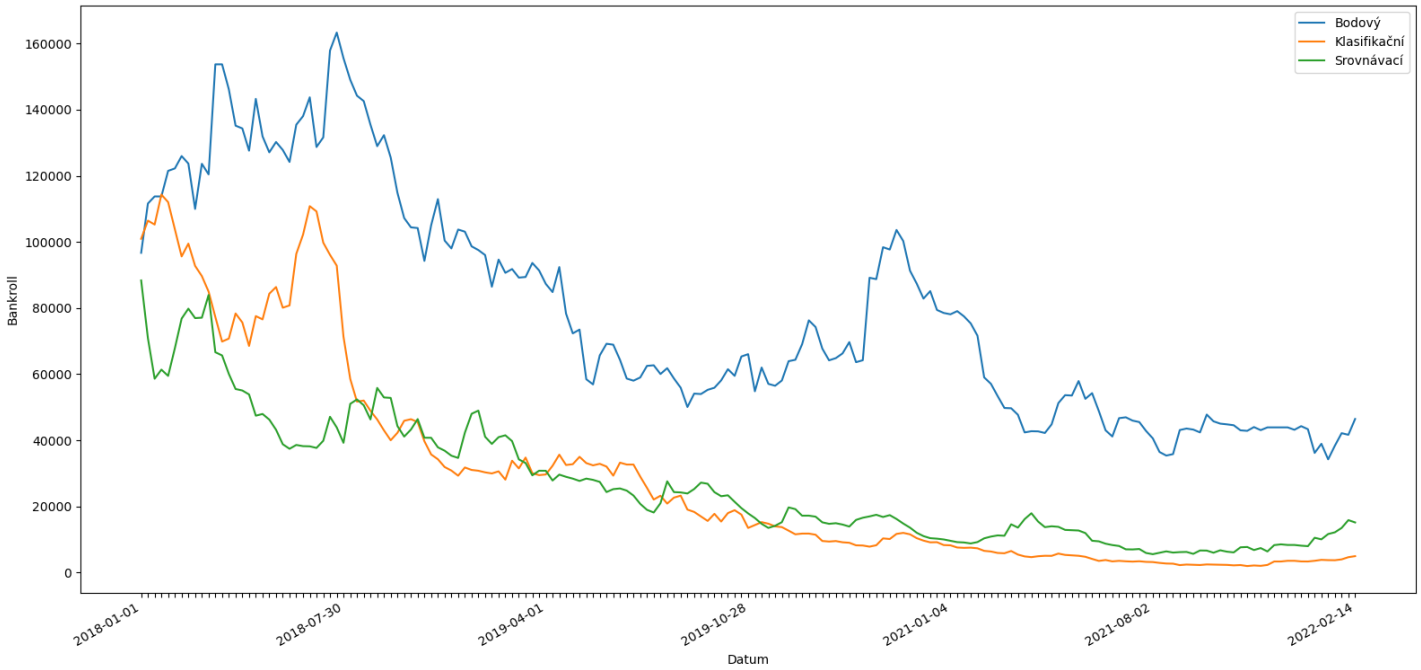
6.4 Ziskovost modelů

Pro každý z modelů jsem vyzkoušel dvě různé investiční strategie u dvou různých zahraničních sázkových kanceláří. Strategií ohledně sázení existuje nepřeberné množství, ale já jsem vyzkoušel dvě vlastní. Pro obě bylo nejprve nutné spočítat, jestli má vůbec smysl na danou příležitost vsadit, a jak moc je pro mě vsazení výhodné. Této hodnotě říkám *value* a počítám ji následujícím způsobem:

$$\text{prediction} - 1/\text{rate}$$

Jedná se tedy o rozdíl pravděpodobností predikovaných mým modelem a těch, které určila sázková kancelář. Platí, že čím vyšší value, tím bude ve střední hodnotě větší zisk, ale pouze za předpokladu, že mé modely predikují pravděpodobnost vítězství lépe, než modely sázkové kanceláře. Obě strategie měly počáteční bankroll(finance) 100 000 a obchodovaly na celém testovacím datasetu, tedy od začátku roku 2018, až do února roku 2022.

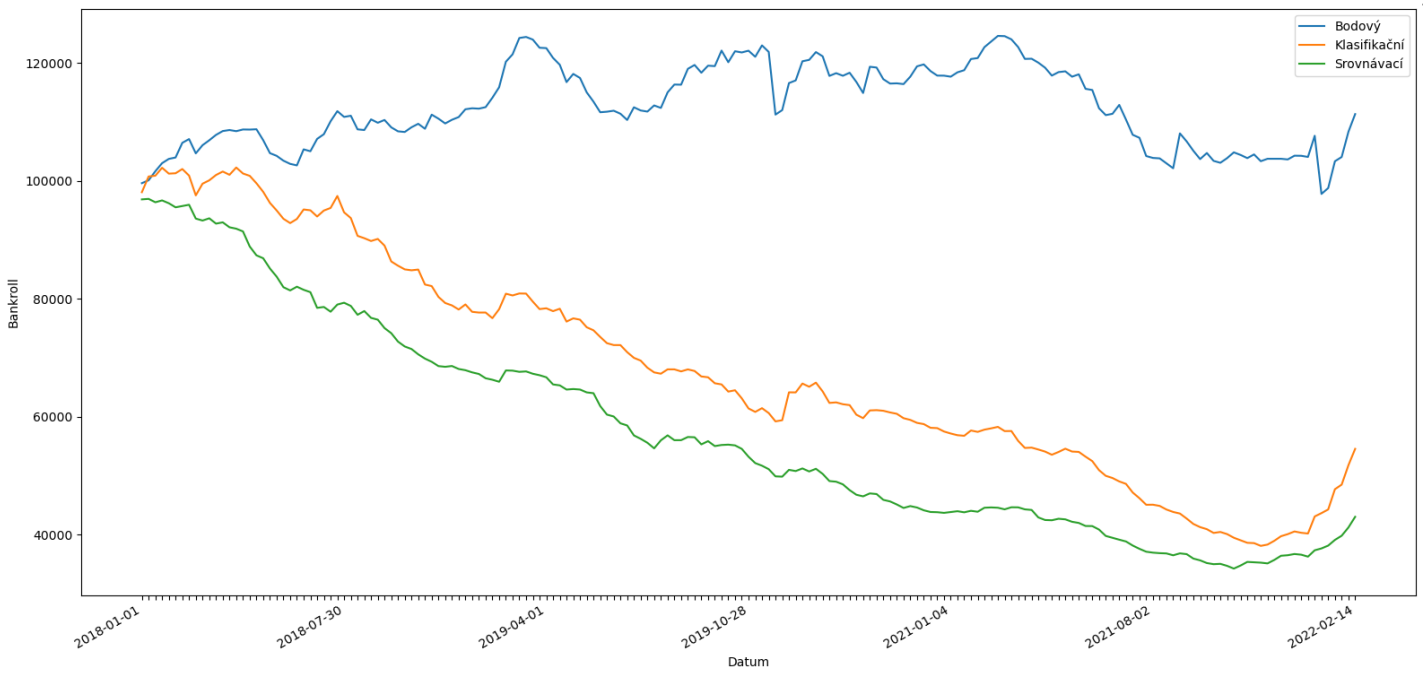
První strategie spočívala v tom, že se vždy vsadilo procento bankrollu rovné value na danou příležitost. Vyzkoušel jsem dva přístupy. První, konzervativnější, spočíval v tom, že se sázelo pouze pokud na některého z hráčů byla value 0–3 %. Jak můžeme vidět z grafů 6.1 a 6.2, ani jeden model nebyl schopen porazit sázkovou kancelář, ačkoliv některé byly ziskové i více než rok.



Obrázek 6.1: Vývoj bankrollu při strategii 0–3 u první sázkové kanceláře

Druhá strategie spočívala v tom, že se sázelo na každou příležitost, na které byla value vždy konstantně 3 % z bankrollu. Jak můžeme pozorovat z grafů 6.3 a 6.4, při této strategii je alespoň bodový model schopen sázkovým kancelářím do nějaké míry konkurovat, ale stejně je není schopen dlouhodobě překonávat.

Obrázek 6.2: Vývoj bankrollu při strategii 0–3 u druhé sázkové kanceláře



Obrázek 6.3: Vývoj bankrollu při strategii fix 3 u první sázkové kanceláře

Obrázek 6.4: Vývoj bankrollu při strategii fix 3 u druhé sázkové kanceláře

Kapitola 7

Závěr

Provedl jsem obsáhlou rešerši již existujících modelů a na jejím základě jsem vybral a naimplemetoval nejlepší z nich. Poté jsem vytvořil dva vlastní modely, jejichž základem byly neuronové sítě a provedl jsem rozsáhlou analýzu a hledání hyperparametrů za pomoci optimalizačního softwaru Optuna. Všechny tři modely jsem natrénoval na totožných datech a posléze porovnal. Pozitivním zjištěním bylo, že oba mé modely byly přesnější a jeden z nich měl i nižší log-loss. Poté jsem vyzkoušel, jestli by byly modely ziskové, kdyby se podle jejich predikcí sázelo u světových sázkových kanceláří. Bohužel ani jeden z mých modelů nebyl schopen modely sázkových kanceláří dlouhodobě překonávat. I přesto ale považuji tuto práci za úspěšnou, protože mé modely byly schopny překonat Elo model, který jsem vybral pro srovnání, protože byl označen jako state of the art. Také jsem vytvořil mnoho statistik, které mohu použít pro další zlepšování stávajících modelů.

7.1 Budoucí práce

Myslím si, že je stále prostor pro nalezení lepších hyperparametrů pro neuronové sítě. Mohu vyzkoušet lepší kombinace statistik a nebo využít regularizaci. Další z možností je zkusit využít ještě hlubší neuronové sítě a nechat optimalizační software, aby provedl více experimentů a našel tak ještě lepší hyperparametry.

Není to ale jediná cesta, kudy bych se v budoucnu rád vydal. Byl jsem velmi překvapen, jak dobře si vedly ratingové modely z rešerše, a proto bych rád vyzkoušel podobný přístup za pomocí neuronových sítí. Vstupní data by byla pouze identifikátory jednotlivých hráčů a typ povrchu, na kterém se zápas hraje. Vytvořil bych pak embedding, na jehož základě bych stavěl zbytek neuronové sítě. Ostatní ratingové modely ale upravují ratingy po

každém zápasu, ale to u neronových sítí nejde dost dobře udělat. Musel bych vymyslet nějaký způsob dávkování, který by ale zároveň netrval moc dlouho na naučení.

Věřím, že oba zmíněné postupy by mohly vést k ještě lepším výsledkům, než jakých jsem dosáhl doposud, a že se mi jednoho dne podaří sázkové kanceláře v predikcích výsledků tenisu překonat.

Literatura

[1] Olivová, V.: *Odvěké kouzlo sportu*. Olympia, 1989.

[2] Gordon, A. D.: *Classification*. CRC Press, 1999.

[3] Sykes, A. O.: An introduction to regression analysis. 1993.

[4] Boulier, B. L.; Stekler, H.: Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, ročník 15, č. 1, 1999: s. 83–91, ISSN 0169-2070.

[5] Kovalchik, S. A.: Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, ročník 12, č. 3, 2016: s. 127–138, doi:doi:10.1515/jqas-2015-0059.

[6] Klaassen, F. J.; Magnus, J. R.: Forecasting the winner of a tennis match. *European Journal of Operational Research*, ročník 148, č. 2, 2003: s. 257–267.

[7] Del Corral, J.; Prieto-Rodríguez, J.: Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, ročník 26, č. 3, 2010: s. 551–563.

[8] Gilsdorf, K. F.; Sukhatme, V. A.: Testing Rosen's sequential elimination tournament model: Incentives and player performance in professional tennis. *Journal of Sports Economics*, ročník 9, č. 3, 2008: s. 287–303.

[9] Wilkens, S.: Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, , č. Preprint, 2021: s. 1–19.

[10] Newton, P. K.; Keller, J. B.: Probability of winning at tennis I. Theory and data. *Studies in applied Mathematics*, ročník 114, č. 3, 2005: s. 241–269.

[11] Barnett, T.; Clarke, S. R.: Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, ročník 16, č. 2, 2005: s. 113–120.

[12] Spanias, D.; Knottenbelt, W. J.: Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, ročník 24, č. 3, 2013: s. 311–320.

[13] Knottenbelt, W. J.; Spanias, D.; Madurska, A. M.: A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, ročník 64, č. 12, 2012: s. 3820–3827.

[14] Bradley, R. A.; Terry, M. E.: Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, ročník 39, č. 3/4, 1952: s. 324–345.

[15] McHale, I.; Morton, A.: A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, ročník 27, č. 2, 2011: s. 619–630.

[16] Bialik Carl, M. B.: Serena Williams And The Difference Between All-Time Great And Greatest Of All Time. 2015, navštíveno dne: 06.03.2022. Dostupné z: `https://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/`

[17] Elo, A. E.: *The rating of chessplayers, past and present*. BT Batsford Limited, 1978.

[18] FIDE, C.: FIDE RATING REGULATIONS. 2022, navštíveno dne: 06.03.2022. Dostupné z: `https://handbook.fide.com/chapter/B022022`

[19] Leitner, C.; Zeileis, A.; Hornik, K.: Is Federer Stronger in a TournamentWithout Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009. *Austrian Journal of Statistics*, ročník 38, č. 4, 2009: s. 277–286.

[20] Wang, S.-C.: Artificial neural network. In *Interdisciplinary computing in java programming*, Springer, 2003, s. 81–100.

[21] Bottou, L.: Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, Springer, 2012, s. 421–436.

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

[23] Akiba, T.; Sano, S.; Yanase, T.; aj.: Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, s. 2623–2631.

Příloha **A**

Seznam použitých zkratek

ATP Association of Tennis Professionals

WTA Women's Tennis Association

Příloha **B**

Obsah přiloženého CD

- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
- `experiments` ..... Záznamy jednotlivých experimentů