FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Predikce výsledků zápasů v MMA
Student:	Kryštof Dostál
Vedoucí:	Ing. Magda Friedjungová, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

1) Proveďte rešerši dostupných zdrojů dat o zápasech a hráčích MMA.
2) Proveďte rešerši známých metod používaných pro predikce výsledků zápasů kontaktních sportů.
3) Ze získaných dat vytvořte vhodné příznaky a na nich experimentálně porovnejte vybrané metody predikce výsledků zápasů MMA.
4) Výsledky porovnejte také s predikcemi sázkových kanceláří.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 1. dubna 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Predikce výsledků zápasů v MMA

Kryštof Dostál

Katedra aplikované matematiky
Vedoucí práce: Ing. Magda Friedjungová, Ph.D.

11. května 2022

Poděkování

Velmi děkuji vedoucí mé práce Ing. Magdě Friedjungové, Ph.D. za její pomoc, ochotu, čas i cenné rady, které při zhotovování této práce nesmírně pomohly.

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 11. května 2022 . . . . . .

Odkaz na tuto práci

Dostál, Kryštof. *Predikce výsledků zápasů v MMA*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2022.

Abstrakt

Cílem této bakalářské práce je vytvoření modelu, který predikuje vítěze profesionálních MMA zápasů. Predikce jsou založeny na statistikách z předchozích zápasů. Provedená rešerše zkoumá dostupné datové zdroje. K statistikám z předchozích zápasů jsou připojeny predikce sázkových kanceláří v podobě historických kurzů. Také byl proveden průzkum již existujících prací s podobnou tématikou. Na základě rešerše bylo vybráno několik klasifikačních modelů, které se následně natrénovaly na předzpracovaných datech a jejich predikce byly vzájemně porovnány, a to i s predikcemi sázkových kanceláří. Výsledky nejlepších modelů se jeví jako příznivé.

Klíčová slova predikce, vytěžování znalostí z dat, strojové učení, smíšená bojová umění, UFC, klasifikace, Python

Abstract

The aim of this bachelor thesis is to create a model that predicts the winners of professional MMA matches. The predictions are based on statistics from previous matches. The research carried out examines the available data sources. To the statistics from previous matches are also added predictions by bookmakers in the form of historical odds. A research of already existing works

with a similar theme was also carried out. On the basis of the research, several classification models were selected, which were then trained on the preprocessed data and their predictions were compared with each other, including comparison with the predictions of bookmakers. The results of the best models appear to be favourable.

Keywords Prediction, Data mining, Machine learning, Mixed martial arts, UFC, Classification, Python

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Smíšená bojová umění (anglicky Mixed Martial Arts, zkráceně MMA) jsou komplexním bojovým sportem nejblíže simulující realný fyzický konflikt. MMA je jeden z nejrychleji rostoucích sportů současnosti, v roce 2021 dokonce hlavní tvář sportu Conor McGregor dokonce překonala etablované sportovní hvězdy jako je Lionel Messi a Cristiano Ronaldo (fotbalisté) v žebříčku nejlépe placených sportovců na světě. [1] Přesto vědecká literatura, která by zkoumala možnosti prediktivních modelů strojového učení, oproti jiným zavedeným sportům, zaostává. Data pro predikci nicméně k dispozici jsou, navíc relativně snadno dostupná a obsáhlá. Absence vědecké literatury na toto téma se dá tedy připsat novosti a možná i určité kontroverznosti tohoto sportu. Zatím bylo učiněno pouze několik jednotek pokusů a vytvoření modelu, který by predikoval výsledky zápasů na základě výkonnostních ukazatelů zápasníků.

Model, který popisuje takové souvislosti, by ale byl velmi užitečný. Zjištění, jaké bojové ukazatele a charakteristiky mají největší přínos pro vítězství, by mohlo pomoci při tréninku atletů a ukázat jim vhodnou strategii i slabé stránky na které se zaměřit, ať už své, nebo soupeřovy. Dále by tento model mohly využívat promotérské organizace pro domlouvání zápasů, ve kterých mají soupeři stejnou šanci na výhru, to by zajistilo atraktivnost utkání. V poslední řadě by přesnější predikce mohla pomoci i sázkovým kancelářím, pro další manipulaci se svými sázkovými kurzy, nebo samotným sázkařům. Tématem této práce je tedy predikce výsledků MMA zápasů ve smyslu výhry konkrétního zápasníka.

Motivací k výběru tématu je spolu s výše zmíněnými skutečnostmi i faktor výzvy, který vyplývá z práce s čerstvým tématem. Sport MMA sleduji, baví mne a myslím si, že tato práce by do sportovního prostoru mohla přinést nové poznatky.

Práce se věnuje hlavně získání relevantních dat o MMA zápasech, diskuzi o volbě vhodných příznaků, jejich extrakci do použitelného formátu pro klasifikační modely a vyhodnocení predikční přesnosti.

První část se věnuje popisu fungovaní MMA, existujících organizací a průběhu zápasení v nich. Dále se zabývá analýzou dostupných datových zdrojů. Ve třetí části se práce věnuje volbě nejvíce vypovídajících příznaků, ve smyslu majících největší vliv na predikci výsledku. Dále jsou v práci rozebrány vhodné klasifikační modely a v poslední části práce se poté porovnávají výsledky predikce s již existujícími modely a výsledky sázkových kanceláří.

Cíl práce

V teoretické části práce budou rozebrány existující organizace pořádající MMA zápasy, jejich rozdíly v rámci struktury žebříčků a samotných bojových pravidel. Dále rozeberu zdroje dat MMA zápasů a dat kurzů sázkových kanceláří. Poté dostupné zdroje zanalyzuji a vyberu nejvhodnější – s tím souvisí i volba surovosti a granulovanosti dat. Poslední cíl je popis struktury vybraných dat.

V praktické části provedu analýzu existujících metod pro predikci MMA zápasů s ohledem na identifikování nejdůležitějších výkonnostních znaků a dále provedu návrh nových příznaků. Obstarám data ze zdrojů zvolených na základě předchozí rešerše a data vhodně předzpracuji pro pozdější použití v prediktivních modelech. S finálními daty a modely zvolenými v teoretické části provedu experimenty za účelem dosažení nejlepší klasifikační přesnosti. Nad získanými výsledky se budu zamýšlet v kontextu možného použití pro přípravu MMA zápasníků na nadcházející utkání a porovnání přesnosti predikcí se sázkovými kancelářemi.

Kapitola 1

Smíšená bojová umění

1.1 MMA

Smíšená bojová umění (anglicky Mixed Martial Arts, zkráceně MMA) jsou bojovým sportem spojující různé manévry jiných bojových sportů a umění. Jedná se o plnokontaktní sport, umožňující velkou škálu úderů, pohybů a chvatů, vycházejících z tradičních bojových umění jako je Jiu Jistu, Taekwondo, Wrestling, Kickbox a Box. To umožňuje velkou taktickou rozmanitost, kde každý zápasník volí strategii podle svého zápasnického pozadí a fyzických předpokladů. Objektivně se dá tvrdit, že MMA má z bojových sportů nejblíže reálnému konfliktu a jedná se o jeden z nejvíce technicky i fyzicky náročných sportů vůbec. Techniky používané v MMA se mohou rozdělit do dvou hlavních kategorií, Postoje a Grapplingu. Manévry na zemi a u stěny klece spadají pod Grappling, používá se zde páčení (zámky), škrcení, lámání, tyto techniky se souhrnně označují anglickým výrazem *submission*. V postoji (anglicky *stand-up*) se naopak používají údery a kopy.

Vítězství lze dosáhnout čtyřmi způsoby. *Knouckoutem*, protivník upadne do bezvědomí důsledkem utržených úderů nebo nárazu o zem. *Technickým Knockoutem*, rozhodčí zastaví zápas kvůli velkému množství utržených úderů a rozhodnutí, že protivník není schopen dále pokračovat. *Submisí*, kdy se protivník dostane do pozice, například některého technického páčení, ve které se nutně musí vzdát. Pokud dojde k vypršení času, o vítězi rozhodnou tři přidělení rozhodčí. Tato takzvaná rozhodnutí jsou nejčastější způsob ukončení zápasu, ale jde i o způsob nejspornější, je velmi složité na základě výkonu a utržených zranění rozhodnout o nepochybném vítězi a metriky vyhodnocení nejsou jednoznačně uznávané ani mezi MMA špičkou.

V novodobém MMA se zápasí v předem ustanovených váhových kategoriích. Rozpětí jedné váhové kategorie zpravidla bývá v rozmezí 4,5–9 kilogramů. Jestli zápasník splnil limit a váhově spadá do určené kategorie, se většinou kontroluje den před zápasem na oficiálním vážení. Velmi často ale dochází k extrémně drastickému procesu shazování přibližně 10 dnů před vážením,

zápasníci se snaží zhubnout co nejvíce kilo (dochází i k dehydrataci) tak, aby při vážení splnili váhový limit. Po vážení pak mají den na to, aby se zotavili a nabrali co nejvíce hmoty, to by jim pak v samotném zápase mělo poskytnout výhodu. Toto shazování je zvyklost, která souvisí s nároky soutěže – pokud by se zápasník rozhodl neshazovat a zůstat ve své normální váhové kategorii, riskuje že by musel čelit fyzicky větším a silnějším soupeřům, kteří do dané kategorie shazovali zase ze své vyšší váhy. V praxi ale shazování podstupují všichni a proto jsou zápasy stejně váhově vyrovnané, byť lehce posunuté o maximální limit dané váhové kategorie.

MMA sport na nejvyšší úrovni zprostředkovává několik hlavních promotérských společností, mezi ty nejvýznamnější z nich patří Ultimate Fighting Championship (UFC)[1], ONE Championship (ONE)[2] , Bellator MMA[3] a Rizin[4]. Organizace zajišťují své sportovce a snaží se vyjednat ty nejatraktivnější zápasy, v naprosté většině mezi zápasníky spadajícími pod jednu organizaci. Organizaci domluvených zápasů má na starost také daná společnost, v jeden den se tak odehrává více zápasů na takzvaných galavečerech nebo *fight nights*. Hlavní rozdíl mezi zápasy jednotlivých organizací spočívá v dbání na různé sady pravidel. Díky svým širokým kořenům se MMA pravidla vyvíjela více způsoby v různých zemích, v každé odlišně zohledňující zdravotní, legální i morální otázky dle sportovního pozadí a požadavků dané promotérské firmy. Výrazné rozdíly v pravidlech se týkají počtu zápasových kol, jejich doby trvání a výčtu zakázaných úderů, ale pravidla se liší i mimo samotný zápas, to se týká například regulace shazování pro splnění váhového limitu. Hlavně odlišné trvání jednotlivých kol výrazně mění strategii boje. Nejvíce převládá sada zvaná *Unified Rules of Mixed Martial Arts*, která je přijata komisemi regulujícími MMA ve Spojených státech a zejména ji používá organizace UFC.[2] [3] [4]

1.1.1 UFC

Zde stručně popisuji základní informace o UFC. Jelikož se jedná o tu největší a nejsledovanější organizaci, která většinou i určuje směřování novodobého MMA, popisuji pouze ji. V sekci Diskuze výběru dat pak rozepisuji relevantní rozdíly ostatních předních organizací.

UFC je právem označována za nejprominentnější MMA organizaci. Skoro každý profesionální MMA zápasník sní o zápase na hlavní kartě UFC turnaje, jelikož právě UFC přitahuje ten největší talent a také kontroluje největší podíl MMA trhu. UFC hostuje *fight night* každý týden, a minimálně jednou do měsíce takzvaný *Pay per view* zápasový večer, na nich probíhají ty nejdůležitější a divácky nejsledovanější zápasy jako jsou utkání o titul divize, na kterých se rozhodne o divizním šampionovi.

[1] `https://www.ufc.com/`
[2] `https://www.onefc.com`
[3] `https://www.bellator.com/`
[4] `http://www.rizinff.com/en/`

V UFC trvají zápasy tři kola po pěti minutách, hlavní zápasy večera mají kol pět. UFC pro zápasy používá osmistrannou oplocenou klec oktagon, mající průměr 9,14 metru, jiné organizace často využívají jinak velký druh klece či ringu. Všichni UFC zápasníci se musí podrobit tvrdé a prakticky neprůstřelné antidopingové kontrole po celou dobu svého aktivního působení v organizaci.

Obrázek 1.1: Pokus o submisi v oktagonu

Zdroj: `https://www.shutterstock.com/cs/image-photo/paulo-brazil-september-22-2018-ufc-2042078177`

Kapitola 2

Analýza zdrojů

V této kapitole popisuji jednotlivé zdroje dat o proběhlých zápasech a detailech, které obsahují.

2.1 UFC

Je vhodné abych měl pro práci přístup k otevřeným, tedy volně dostupným datům, buď skrze API, nebo veřejně dostupný hotový dataset, případně si data „ručně“ postahovat přes vlastní parser a napodobení GET požadavku. Dříve existovala oficiální UFC API[5], poté byla zrušena a její obdoba přesunuta pod ESPN, která je ale od roku 2014 draze zpoplatněna[6] a obsahuje pouze informace o zápasnících, ale ne o žebříčcích a zápasech. V současnosti oficiální veřejně dostupná API neexistuje. V minulosti fungovalo několik amatérských API vytvořených fanoušky, ale ta už jsou stará minimálně 5 let a kvůli tomu, nebo kvůli závislosti na oficiální UFC API také nejsou funkční. Jediná amatérsky vytvořená API[7], má dva problémy – vyžaduje přístupové údaje a data stahuje z domény Sherdog, která nabízí omezené údaje k proběhlým zápasům – jde pouze o kolo a čas ukončení zápasu a metoda použitá vítězem k poražení protivníka. Dále existuje několik placených API od komerčních data poskytovatelů. Od první UFC události v roce 1993 se zápasilo přibližně 6000×. Bohužel přesný počet zápasů není snadno dohledatelný, v článku z roku 2019[5] se pracuje s 5144 zápasy od roku 1993, dataset z Kaggle[8] z roku 2021 pak s 6012, obě práce mechanicky stahují data od prvního zápasu z Ufcstats. Pro predikci budoucím modelem by bylo ideální mít alespoň 2000 vyčištěných záznamů. Důležité ale je, že prvních 30 UFC večerů se nepoužívala oficiální *Unified Rules*

[5] `http://ufc-data-api.ufc.com/api/v3/us/`
[6] `http://www.espn.com/apis/devcenter/blog/read/publicretirement.html`
[7] `http://www.mmaapi.com/`
[8] `https://www.kaggle.com/rajeevw/ufcdata?select=data.csv`

of Mixed Martial Arts, která se začala používat až od UFC 28 v roce 2000[9] [10]. Dále pak byla tato pravidla začátkem roku 2017 lehce pozměněna[11].

2.1.1 UFCStats

Nyní na portálu UFCStats[12] existuje k datu 22. 12. 2021 záznam k přibližně 5900 UFC zápasům, datujícím se od roku 1994. Ke každému zápasu patří příznaky, jako je metoda ukončení, celkový čas a počet kol, dále se zaznamenává počet a úspěšnost signifikantních úderů plus místo kam údery padaly a z jaké pozice byly poslány, poslední sledované ukazatele jsou pokusy o takedown nebo submisi – techniky souboje používané na zemi. Výše zmíněné statistiky se sledují pro každé zápasové kolo zvlášť – celkově jde o přibližně 130 unikátních příznaků za zápas. U jednotlivých zápasníků se sledují další příznaky jako celková zápasová bilance, výška a váha zápasníků, rozpětí rukou a zápasový postoj, v součtu 16 příznaků. Existuje několik neodborných prací, zaměřených na predikci UFC zápasů. Čtyři z těchto prací[13,14,15,16] využívají data scrapping (automatizované získávání dat přímo z webových stránek) oficiálních UFC stránek. V případě zvolení tohoto způsobu stahování dat, se lze jejich postupem inspirovat.

2.1.2 Kaggle

„Kaggle je AirBnB pro datové vědce – zde tráví noci a víkendy. Je to davem poháněná platforma, která přitahuje, vychovává, školí a vyzývá datové vědce z celého světa, aby řešili problémy datové vědy, strojového učení a prediktivní analýzy.“ [6] Jednou ze služeb poskytovanou Kaggle je sdílení datasetů vytvořených uživateli, nad nimi lze pak samostatně pracovat, nebo se snažit překonávat a vylepšovat zadané úkoly k daným datasetům. Lze zde najít několik datasetů věnujících se MMA problematice, většinou jejich zdrojem jsou také stránky UFCStats, ale kromě syrových dat obsahuje i data předzpracovaná někdy i spojená s jinými zdroji. Ty nejzajímavější jsou vypsány.

První, nejspíše nejpopulárnější MMA Kaggle dataset, je od uživatele Rajeevw[17]. Obsahuje zpracovaná i nepředzpracovaná data také ze stránky Ufcstats. Dataset byl naposled autorem aktualizován před 2 měsíci (psané v květnu 2022), navíc autor poskytuje návod jak si data vlastnoručně aktualizovat pomocí jednoduchého příkazu. Za zmínku stojí příznaky, které popisují zprůměrovaný

[9] http://statleaders.ufc.com/en/event
[10] http://www.ufcstats.com/statistics/events/completed?page=all
[11] https://www.dca.ca.gov/csac/forms_pubs/publications/unified_rules_2017.pdf
[12] http://ufcstats.com/
[13] https://github.com/shortlikeafox/tiger-millionaire/wiki
[14] https://github.com/vinaykanigicherla/mmapredict
[15] https://github.com/WarrierRajeev/UFC-Predictions
[16] https://github.com/jasonchanhku/UFC-MMA-Predictor
[17] https://www.kaggle.com/rajeevw/ufcdata

počet úderů, kopů atd. všech dosavadních oponentů daného zápasníka. Dataset v současnosti popisuje 6012 zápasů před jeho aktualizací. Jedná se o všechny dostupné zápasy na ufcstats od roku 1993, v případě použití tohoto datasetu bych vynechal starší zápasy, při kterých se doposud nepoužívala pravidla *Unified Rules of Mixed Martial Arts*.

Dataset uživatele Mdabbert[18] obsahuje částečně zpracovaná data. Je k němu dostupný i repozitář s kompletní dokumentací[19], autor zde pracuje na predikci zápasů, se zaměřením na určení ziskových vítězů v případě, že by na ně vsadil, k čemuž používá i sázkové kurzy. Dataset nebyl v posledních několika měsících aktualizován, obsahuje záznamy zápasů od poloviny roku 2010 do února 2021. Celkově jde o 4896 záznamů. Dataset obsahuje navíc pár zajímavých příznaků – sázkový kurz, prázdná aréna a rozdíly některých metrik například `avg_sub_att_dif`: rozdíl průměrných pokusů o *submise* obou zápasníků. Neobsahuje ale informace o průměrných statistikách oponentů. Oba zmíněné datasety zaznamenávají lokaci proběhlého UFC večera, bylo by zajímavé přidat příznak domácí výhoda (zápas se odehrává v rodném městě zápasníka) a sledovat jak tato psychologická výhoda ovlivňuje výsledky.

Poslední zajímavý dataset[20] obsahuje už zpracovaná a nejvíce granulovaná data, ke každému zápasu popisuje 894 příznaků, jde o podrobně zpracované statistiky jako jsou procentuální úspěšnosti úderů a časy strávené v jednotlivých postojích rozdělené pro každé zápasové kolo, dále pak informace o události jako je datum a místo konání. U některých příznaků, ale chybí značné množství dat, případné vyčištění datasetu je na místě. Obsahuje 2318 záznamů od roku 2013 do 2019, je tedy bohužel zastaralý.

2.2 Ostatní organizace

Další dvě největší MMA organizace Bellator a ONE neposkytují tak detailní statistiky jako UFC, k dispozici jsou velmi základní informace o předchozích zápasech, které má v režii několik fanouškovských stránek. Z nich je doména Sherdog[21] uživateli nejčastěji používaná, například pro extrakci[22] [23] dat.

Jako zdroj dat bych volil zmíněnou stránku Sherdog nebo dále MMA-ORACLE[24], obě nabízejí totožné statistiky pro všechny další organizace. Mezi dostupné příznaky patří čas a metoda ukončení zápasu a dále dosavadní statistiky o jednotlivých zápasnících – věk, váha, váhová kategorie a zápasová bilance rozdělená podle způsobu ukončení. MMA-ORACLE disponuje vlastním hodnotícím algoritmem, poskytuje proto navíc informace o zápasníkovo umístění

[18] `https://www.kaggle.com/mdabbert/ultimate-ufc-dataset`
[19] `https://github.com/shortlikeafox/tiger-millionaire/wiki`
[20] `https://www.kaggle.com/calmdownkarm/ufcdataset`
[21] `https://www.sherdog.com/events`
[22] `https://github.com/seanpquig/betting-odds-analyzer`
[23] `https://github.com/fight-api/ufc-fight-api`
[24] `https://mma-oracle.com/en/event/future/list/page-0`

v jeho váhové kategorii, a v zemi za kterou zápasí, plus vede záznamy o typu submisí, pomocí kterých vyhrál zápas.

2.3 Data sázkových kanceláří

Pro pozdější srovnání a možné použití v modelech budou dále potřeba historické sázkové kurzy. Pro tuto práci je nejvhodnější použít desetinný způsob zápisu kurzů. [7] I proto je jedna z nejlepších variant využití MMA záznamů z webové stránky BetMMA[25]. Konkrétní odkaz obsahuje přehledně seřazené proběhlé zápasové události od dubna roku 2013 spolu s desetinnými kurzy. Portál neobsahuje podmínky použití, a tak plánuji data stahovat napodobením GET požadavku ve vlastním skriptu.

2.4 Diskuze výběru dat

V práci jsou použity pouze záznamy ze zápasů organizace UFC, k tomuto rozhodnutí mě vede kombinace více faktorů. Nízká granularita a značně omezená dostupnost statistik o minulých zápasech je problém, který mají všechny organizace kromě UFC, bude to pravděpodobně také důvod, proč ve všech odborných článcích, snažících se predikovat výsledky MMA zápasů, byly doposud použita data pouze z UFC.

V některých případech by bylo velmi obtížné porovnat zápasy rozdílných organizací, i když jsou velmi kvalitní – například pravidla společnosti Rizin jsou diametrálně odlišná – v Rizinu trvá první kolo zápasu 10 minut, případně někdy je zápas pouze na dvě 10minutová kola, to je skoro nesrovnatelné s normálním zápasem z hlediska vyčerpání zápasníků. Rizin také povoluje údery, které jsou zakázané v ostatních organizacích. Dalším faktorem je kvalita zápasníků pod danou organizací, i když PFL patří mezi ty známější organizace s dobrým fungování, z pohledu fanouška MMA se dá říci, že jde zatím o znatelně horší soutěž, než u konkurence, nepůsobí v ní žádný světový zápasník a ani vysloužilí zápasníci z jiných organizací nemají zájem do téhle organizace přejít ke konci kariéry. V užším výběru byly zvažovány organizace UFC, Bellator a ONE FC, patří mezi tři nejvrcholnější MMA soutěže. UFC ale stále velmi vyčnívá.

Porovnatelnost odlišných zápasů pod UFC, Bellator a ONE FC je také faktorem, rozdíly v pořádání zápasů a v pravidlech organizací jsou značné. UFC a Bellator se liší hlavně v rozdílných váhových kategoriích a pořádání turnajů. UFC volí nadcházející zápasy podle rozhodnutí z vyšších pozic, tak aby proti sobě zápasili co nejatraktivnější soupeři, kromě vyrovnaných šancí je pro UFC důležité udržovat i dějovou linku, a tak zápasník, který se dobře mediálně prezentuje, má často větší šance na zápas než jiný zápasník, byť tabulkově výše postavený. Bellator naopak organizuje zápasy dle turnajového systému,

[25] `https://www.betmma.tips/mma_betting_favorites_vs_underdogs.php?Org=1`

kdy v jedné váhové kategorii pořádá související zápasy v několika navazujících večerech, zápasníci postupují v pyramidě nahoru a vítěz turnaje pak dostane šanci zápasit o titul šampiona – tento systém je tabulkově spravedlivější, ale fanoušci mohou přijít o vysněné zápasy např. dvou rivalů. ONE FC používá oproti UFC rozdílnou sadu pravidel a také jiná kritéria pro rozhodování rozhodčích v případě neukončení zápasu před limitem, jelikož většina zápasů končí rozhodnutím, jedná se o další faktor ovlivňující rozdíly mezi daty. Dále má ONE FC jiná pravidla pro shazování váhy před zápasem a maximální povolenou dehydrataci zápasníků – nedovoluje se tak drastické shazování jako v UFC a tím pádem jsou efektivně všechny váhové kategorie posunuté o jeden stupeň (zápasník lehké váhy v UFC by patřil do velterové váhy v ONE FC). Samozřejmě se jedná i o velké rozdíly v kvalitě MMA zápasníků, kdy UFC reprezentuje naprostou světovou špičku, která je cílem i zápasníků z ostatních organizací.

2.5 Analýza současného stavu řešení problému

V první práci[5] jsou použitá data strojově stažena ze stránky UFCStats[26] a předzpracována pro výsledných 3 355 záznamů o 134 příznacích. Dvě nejvýznamější metody pro predikci v této práci jsou *rozhodovací stromy* a *Gradient boosting*, po použití těchto dvou technik bylo dosaženo predikční úspěšnosti 60,25 % a 61,22 % respektive. Autoři uznávají, že to není velmi úspěšné skóre – v celém datasetu vyhrává zápasník v červeném rohu v 62,6 %, tedy jediný příznak má větší vypovídající hodnotu než aplikace predikčních modelů. Každopádně je způsob stahovaní a předzpracovávání dat i samotného používání modelů korektní a systematický, a proto se jím budu inspirovat v experimentální části práce.

V další práci[8] jsou data obdržena od portálu FightMetric LLS[27], v současnosti je dostupná pouze omezená verze, tento portál fungoval jako oficiální poskytovatel UFC statistik, a jeho funkci později převzal již zmíněný UFCStats. Pracuje se s přibližně 1477 záznamy (není zmíněno, zda některé byly po předzpracování vyhozeny) o 895 příznacích, jde tedy o velmi granulární data. Nejlepší prediktivní model je zde *Support Vector Machines* (také *SVM*), dosahujících konstantní úspěšnosti okolo 61 %, s maximem 62,8 %. Druhý nejlepší model je *náhodný les*. Dále zápasník reprezentující červený roh vítězí dokonce v 58,7 % zápasů – bylo by tedy vhodné data předzpracovat tak aby zastoupení vítězů bylo rovnoměrné. Užitečný postřeh ohledně předzpracování: *„(. . .)the highest correlations are with Round 4 and Round 5 features, since msot fights do not have Round 4 and Round 5. To deal with this sparsity, we summed the respective features of each round. Finally, we then attempt to half the numbers of features again, by taking the ratio of features from red and blue side fighters.“*

[26] `http://ufcstats.com/statistics/events/completed`
[27] `https://fightmetric.rds.ca/events/completed`

Práce Lachlan P. James[9] se nevěnuje samotné predikci výsledků, ale identifikaci nejvíce vypovídajících příznaků. V práci se ale stejně musely použít zápasové statistiky, zde 236 záznamů o 11 příznacích, stažené z portálu FightMetric LLS. Podobně jako v níže zmíněné práci se z 11 součtových příznaků vytvořily příznaky popisující míru závislosti daných technik na čase nebo úspěšnosti ku počtu pokusů. V určení nejdůležitějších příznaků byly úspěšnější modely používající *rozhodovací stromy* oproti metodě používající konečný automat. Ve výsledcích se nejčastěji objevuje aktivita grapplingu (tedy úspěšné provedení technik na zemi, jako je posunutí ze do výhodnější pozice), spolu s přesností technik (poměr úspěšných manévrů ku pokusům o ně). Oproti tomu hrubé množství uštědřených úderů apod. se v *rozhodovacích stromech* vysoko nevyskytuje. *„The decision tree partitioning also reveals the highly technical nature of MMA activity across both modes of combat. In particular, the accuracy of strikes and takedowns, in addition to strikes landed per minute, which are representative of successfully executed techniques, were featured in the model.“* Nicméně stojí za zmínku, že vyšší váha grapplingových technik může být dána popularitou tohoto přístupu v době zápasů, ze kterých data pochází. V současnosti se ale nejvýhodnější soubor technik mohl změnit na postoj, možná by bylo vhodné záznamy ováhovat podle jejich data konání tak, aby čerstvější zápasy byly důležitější, a tím více reprezentovaly současnou strategii.

Poslední zmíněný článek Jeremiah Johnson[10] se věnuje predikci výsledků podle metod matematické statistiky, hlavně *logistické regrese*. Výsledné poznatky mimo jiné objasňují užitečnost různých typů příznaků. Do první kategorie spadají podle autora tzv. součtové příznaky, ty označují jednoduše statistiky, u kterých se dá v jednotkách spočítat kolikrát za zápas případ nastal. Jde např. o počet uštědřených úderů, počet úspěšných takedownů a podobně. Z těchto příznaků a několika dalších informací (celková délka zápasu atd.) lze pak odvodit druhou kategorii příznaků – ty popisující určitou úspěšnost zápasníkových technik. Jde tedy o procentuální úspěšnost, například poměr úspěšných úderů. Dále statistiky úspěšnosti za minutu, například rozdíl úderů za minutu obou zápasníků. A také statistiky úspěšnosti určitých technik v poměru ku související (hlavně předcházející) technice, tam patří kupříkladu poměr zápasníkovo aktivity na zemi ku všem úspěšným *takedownům* v zápase. Z výsledků vyplývá, že logistická regrese dosahuje nejlepších výsledků, pokud jsou do celkového modelu zařazeny obě dvě kategorie příznaků. Tento závěr podporují i výsledky předchozí studie, proto by bylo vhodné v datasetu použít určité příznaky z druhé kategorie.

Na základě zmíněných článků a svých znalostí o sportu jsem se rozhodl pro následující postupy při tvorbě výsledného modelu. Statistiky budu stahovat ručně pomocí vlastního skriptu z portálu UFCStats[28], jde o nejčastější způsob, který umožňuje individuální manipulaci. Data budu stahovat od dubna roku

[28] `http://ufcstats.com/`

2013, konkrétně večera UFC 159, to kvůli synchronizaci s dostupnými sázkovými kurzy, jak je zmíněno v sekci 2.3. Navíc, jak zmiňuje článek Hitkul a spol.[8], více granulární data jsou dostupná až po roce 2012, což jsem si ověřil vlastní inspekcí dat, jde hlavně o časté chybějící hodnoty u detailních statistik zápasníků. Předzpracování stažených dat a transformace příznaků je popsána v kapitole 4, zde zmíním nutnost rozhodnutí se, zda vybalancovat výskyt počtu výher u zápasníků v červeném a modrém rohu. Jelikož je zpravidla favorit zápasu v červeném rohu, v závislosti na volbě datasetu vyhrávají tito zápasnící v 56,7 % až 62,6 % případech. Po konzultaci jsem se ale rozhodl rozložení v datasetu ponechat, místo vybalancování tak, aby obě strany měly v průměru 50 % zastoupení vítězství. Místo toho se soustředím na nalezení takového modelu, který bude mít vysokou senzitivitu (viz. sekce 5.1) pro zápasníky v modrém rohu – takový model bude častěji správně predikovat podceňované vítěze a tím pádem bude i výhodnější pro případné budoucí sázení. Predikční modely, se kterými budu pracovat, jsou *náhodné lesy, logistická regrese, SVM* (*Support Vector Machines*, metoda podpůrných vektorů), *XGB klasifikátor a MLP* (Multi-layer Perceptron, vícevrstvý perceptron, umělá neuronová síť). Důvodem zvolení těchto modelů je vyšší predikční úspěšnost ve zmíněných pracích, a také celková úspěšnost a robustnost v praxi. U modelů *náhodných lesů, logistické regrese, XGB klasifikátoru a SVM* (pokud použiji lineární jádro), je také významná možnost zobrazení důležitosti jednotlivých příznaků. Práce s modely je popsána v sekci 5.2.

Kapitola 3

Popis dat

V této kapitole popíši zvolená data, se kterými budu dále v práci manipulovat a vycházet z nich. Data byla stažena vlastními skripty.

3.1 Záznam o zápase

Nejdříve popíši záznamy o jednotlivých zápasech stažených z portálu UFCStats. Data jsou ve formátu jednoduché tabulky s více sloupci – obsahuje detailní zápasové statistiky pro zápasníky v červeném i modrém rohu, jména obou zápasníků (slouží také jako identifikátor), způsob, čas ukončení a formát zápasu, datum a vítěze. Stažený dataset obsahuje 4226 záznamů.

Kromě zmíněných statistik se v datech nachází ještě sloupce popisující sloupec `Sig. Str.`, jde o `Head, Body, Leg, Distance, Clinch, Ground`. Ty jsou ve stejném formátu jako `Sig. Str.` a jen upřesňují, kam údery padaly a zda se zápasníci nacházeli v postoji nebo na zemi.

3.2 Záznam o zápasníkovi

Data jsou opět stažena z portálu UFCStats ve formátu tabulky. Sloupce nyní popisují pouze jméno bojovníka, jeho výšku a rozpětí paží měřených v palcích, váhu v librách a datum narození. Portál dále poskytuje numerické statistiky zápasníkovo průměrných signifikantních úderů apod. ze všech jeho proběhlých zápasů, pro predikční účely jsou tyto statistiky vhodné pouze pro odhad budoucího zápasu. Stejné statistiky, ale relevantní pro každý záznam zvlášť (popisující průměry ze všech předcházejících zápasů), si vytvářím sám během předzpracování datasetu, proto zde nejsou ukládány. Oproti zápasovým záznamům se zde vyskytuje značně více chybějících hodnot. Dataset obsahuje záznamy o 3742 zápasnících.

Název	Ukázka formátu	Popis
KD	2	Počet *Knockdownů* (shození na zem pomocí úderů)
Sig. Str.	7 of 17	Počet pokusů a úspěšně zasažených signifikantních úderů
Sig. Str.%	41%	Procentuální úspěšnost signifikantních úderů
Total Str.	10 of 14	Počet pokusů a úspěšně zasažených všech úderů
TD	0 of 1	Počet pokusů a úspěšně provedených *Takedownů* (shození na zem pomocí vlastní váhy)
TD%	0%	Procentuální úspěšnost *Takedownů*
Sub Att.	2	Počet pokusů o *Submisi* (poražení soupeře pomocí pák a škrtících chvatů)
Rev.	1	Počet úspěšných přechodů z nevýhodné pozice na zemi do výhodné
Ctrl	0:45	Jak dlouho zápasník drží protivníka na zemi v dominantní pozici

Tabulka 3.1: Popis zápasových statistik

3.3 Sázkové kurzy

Jednotlivé sázkové kurzy na zápasy, stažené z portálu BetMMA[29]. Kurzy jsou v desetinném formátu připojována k odpovídajícím zápasům během samotného stahování pomocí skriptu. Oproti zápasovému datasetu zde chybí lehce přes 1100 záznamů.

[29] https://www.betmma.tips/mma_betting_favorites_vs_underdogs.php?Org=1

Kapitola 4

Práce s daty

Tato kapitola popisuje použité nástroje a postup pro stažení a zpracování zvolených dat do výsledné formy, která bude použita v predikčních modelech.

4.1 Použité nástroje

Všechny operace stažení a zpracování dat byly provedeny v programovacím jazyce Python, verze 3.7.11, pomocí vlastních skriptů.

Pro automatizované stahovaní dat pomocí skriptu bylo použito vývojové prostředí (IDE) PyCharm od společnosti JetBrains. Důvodem jsou možnosti debuggování a kontroly kódu v reálném čase, tyto funkcionality jsou užitečné při parsování statistik z mechanicky stažených HTML reprezentací stránek UFCStats.

Pro následné zpracování dat do finální podoby, a i provádění experimentů nad nimi, byl použit nástroj Jupyter notebook. Jedná se o aplikaci spouštěnou ve webovém prohlížeči, která poskytuje základní funkcionality vývojového prostředí (chybí například kontrola kódu v reálném čase), ale navíc umožňuje spustit kód a zobrazit výstupy po volitelných blocích, a přehledně celý postup komentovat v Markdown jazyce.

4.1.1 Knihovny

Seznam nejdůležitějších Python knihoven použitých v práci:

- *Requests* – Standardní knihovna pro HTTP požadavky. Zde použita pro stažení HTML souborů zápasových a sázkových statistik pomocí GET požadavku.

- *BeautifulSoup* – Knihovna umožňující procházení a získávání konkrétních struktur ze stažených HTML souborů.

- *Pandas* – Knihovna pro snadnou a výkonnou práci s daty, souvisejícími datovými strukturami a následnou datovou analýzu.
- *NumPy* – Knihovna poskytující snadno použitelné matematické operace, zejména nad seznamy, vektory a maticemi.
- *Scikit-learn* – Knihovna zahrnuje většinu technik strojového učení, klasifikační i regresní modely, dále funkce pro předzpracování, transformaci a výběr dat i modelů.

4.1.2 Hardware

Praktická část práce byla provedena na notebooku s těmito parametry:

```
Výrobce a název notebooku: HP Pavilion 15
Operační systém: Windows 10 Home
Procesor: Intel Core i5-7300HQ
Grafické karty: NVIDIA GeForce GTX 150, Intel HD Graphics 630
RAM: 8 GB
Pevný disk: 128 GB SSD M.2 + 1 TB HDD
```

4.2 Stažení dat

Data byla stahována ze dvou zdrojů, za prvé z hojně zmiňovaného portálu UFCStats. Zde se stahovaly nejdříve statistiky o průběhu jednotlivých zápasů a poté informace o jednotlivých zápasnících. Během stahování zápasových dat k nim byly souběžně připojovány odpovídající sázkové kurzy stažené z portálu BetMMA. Konkrétní popis jednotlivých statistik a počty záznamů jsou popsány v kapitole 3.

Nejdříve byl pomocí HTTP požadavku z knihovny *Requests* stažen HTML soubor reprezentující strukturu webové stránky se statistikami. Poté, za použití nástrojů z *BeautifulSoup*, byly vyparsovány cílené informace, které byly nejdříve ukládány v paměti do datových struktur seznamů nebo slovníků, a ty se nakonec uložily na disk v CSV (Comma-separated values, hodnoty oddělené čárkami) formátu. Výstupem stahování jsou dva datasety, jeden popisující zápasy, a druhý popisující samostatné zápasníky. Jelikož skript nebyl nijak optimalizován (běh na více vláknech v jednom procesoru) a obsahuje vnořené stahování sázkových kurzů během stahování zápasových detailů, jeho běh trval necelé dvě hodiny.

4.3 Tvorba finálního datasetu

Tvorba finálního datasetu se nachází na přiloženém CD v Jupyter notebooku *Preprocess.ipynb*, ve kterém jsou i podrobně popsány všechny kroky. V této části vytyčím to nejdůležitější z předzpracování dat do finální podoby. Výsledný

dataset obsahuje 2944 řádků, zápasů, a 81 sloupců, tedy příznaků včetně predikované proměnné. Dataset je na CD uložen jako *final.csv*.

Prvním krokem je inspekce staženého datasetu a provedení základních kroků pro použití dat v predikčních modelech jako je sjednocení chybějících hodnot do stejného formátu (zde reprezentovány pomlčkami) a odstranění nevhodných záznamů, například zápasů, které skončili diskvalifikací jednoho ze zápasníků. Dále pak je nutné převést kategorické příznaky na jejich binární reprezentaci pomocí takzvaných dummy příznaků a převod některých dalších příznaků např. časů v minutách a vteřinách na vteřiny, kurzů v desetinném formátu na procentuální hodnoty a podobně. Z příznaků ve formátu `10 of 15` (10 úspěšných pokusů z celkových 15, například u počtů úderů) jsem vytvořil dva nové příznaky, jeden popisuje celkový počet pokusů a druhý procentuální úspěšnost z celku.

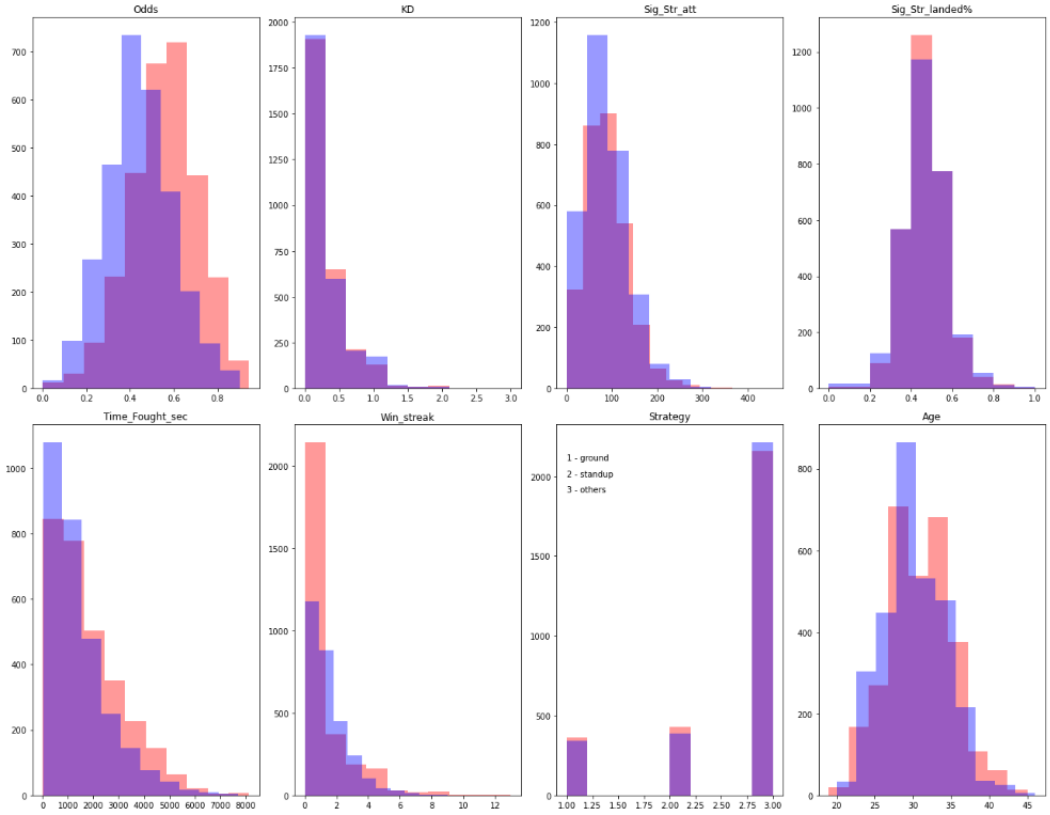
Jedním z hlavních kroků předzpracování je transformace zápasových příznaků tak, aby každý záznam zápasu neobsahoval statistiky popisující to, co se v zápase událo, ale místo nich byly průměrné hodnoty ze všech předchozích zápasů pro oba odpovídající zápasníky. Díky této transformaci se z datasetu musí odstranit všechny záznamy ve kterých zápasí nováčci, jelikož ti nemají v datech žádné předchozí statistiky, které by šly zprůměrovat.

Dále je potřeba doplnit chybějící hodnoty, zde se jedná o výšku, váhu a rozpětí paží zápasníků a pak chybějící sázkové kurzy u přibližně jedné čtvrtiny zápasů. Pro oba druhy hodnot jsem zvolil doplnění pomocí metody *KNN Imputer*[30], ta doplňuje hodnoty na základě podobných hodnot u nejbližších sousedů v podprostoru příznaků. Míry zápasníků doplňuji tímto způsobem na základě předpokladu, že v podprostoru příznaků reprezentujících rozměry zápasníků, by se měli podobně stavění (podobný typ postavy) zápasníci nacházet blízko u sebe. Proto je i doplnění provedeno nad podmožinou datasetu, který obsahuje pouze příznaky výška, váha a rozpětí paží, a poté jsou doplněné hodnoty vloženy zpět do původního datasetu detailů samostatných zápasníků. Po konzultaci je toto doplnění provedeno už v samotném notebooku *Preprocess.ipynb* před rozdělením na trénovací a testovací množiny, je to z důvodu povahy dat, kdy by při pozdějším doplnění *KNN Imputer* doplnil hodnoty podle jiných vazeb, než je pouze podobný typ postavy. Doplnění sázkových kurzů popisuji dále v sekci 5.1.

Zmiňuji také vytvoření nových příznaků `Win Streak` (počet aktuálních vítězství v řadě) a `Age` (věk zápasníka v době konání zápasu). Dále jsem také použil diskretizaci dat pro vytvoření nového příznaku, který by měl popisovat, zda se zápasník specializuje na boj v postoji nebo na zemi, toho jsem dosáhl analýzou percentilů u příznaků popisujících počet zasažených úderů v postoji a doby držení protivníka v dominantní pozici na zemi. V průběhu práce nad datasetem jsem sem také musel vypořádávat s odlehlými hodnotami

[30] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html`

a nezvyklostmi, zde například šlo o 5 párů zápasníků se stejnými jmény, ty jsou použity jako unikátní identifikátor, a proto mohlo dojít k promíchání hodnot během transformací příznaků. Příznak popisující váhu zápasníků jsem se rozhodl do datasetu nezahrnout, jelikož zápasy se odehrávají ve stejné váhové kategorii, a tedy váha obou zápasníků je pro každý záznam totožná. Ve finálním datasetu je zastoupení predikované proměnné 56,76 % pro vítězství zápasníka v červeném rohu.



Obrázek 4.1: Rozdělení distribucí u vybraných příznaků pro zápasníky v červeném a modrém rohu

Název	Popis
Odds	Procentuální reprezentace zápasových kurzů
KD	Průměr *Knockdownů* v předchozích zápasech daného zápasníka
Sig. Str. att.	Průměr pokusů o signifikantní údery v předchozích zápasech daného zápasníka
Sig. Str.%	Průměrná procentuální úspěšnost signifikantních úderů v předchozích zápasech daného zápasníka
Time Fought	Součet doby všech předchozích zápasů ve vteřinách
Win	Počet všech dosavadních vítězství
Lose	Počet všech dosavadních proher
Method	Počet jednotlivých metod ukončení dosavadních zápasů, v datasetu příznaky dumifikovány
Win Streak	Současný počet vítězství v řadě
Strategy	Rozdělení zápasníků do tříd podle specializace na boj na zemi nebo v postoji
Age	Věk zápasníka k datu konání zápasu

Tabulka 4.1: Popis vybraných typových příznaků

Kapitola 5

Experimenty

V této kapitole popisuji práci nad již zpracovaným finálním datasetem, vyhodnocení výsledků jednotlivých prediktivních modelů a porovnání s úspěšností získaných sázkových kurzů.

5.1 Evaluace modelů a rozdělení dat

Pro určení kvality jednotlivých modelů zjišťuji predikční přesnost a senzitivitu. Terminologie typů predikcí pro výpočet těchto veličin je následující (v závorkách uvádím anglické značení). Mezi *skutečně pozitivní (TP)* patří záznamy kdy jsou správně predikovaní vítězové z modrého rohu, analogicky *skutečně negativní (TN)* jsou správně určení vítězové z rohu červeného. Jako *falešně pozitivní (FP)* výsledek, je zde označována predikce vítězství zápasníka v modrém rohu, kdy ale ve skutečnosti vyhrál zápasník z červeného rohu, opak se nazývá jako *falešně negativní (FN)* výsledek. Níže jsou uvedeny vzorce pro sledované hodnoty, podrobněji viz. článek Baratloo a spol.[11]

$$Přesnost = \frac{skutečně\,pozitivní + skutečně\,negativní}{všechny\,predikce}$$

$$Senzitivita = \frac{skutečně\,pozitivní}{skutečně\,pozitivní + falešně\,negativní}$$

Korektní porovnání úspěšnosti jednotlivých modelů je důležitou součástí získání kvalitní finální predikce. Finální dataset obsahuje 2944 záznamů, nicméně pouze u příznaků reprezentujících sázkové kurzy v daném zápase chybí přes 600 kurzů, tedy dataset, který má kompletní nedoplněné kurzy (predikce sázkových kanceláří) obsahuje 2308 záznamů. Jelikož chci z dat vytěžit co nejvíce informací, na trénování zvolených modelů použiji větší dataset. Po vybrání a doladění finálního modelu, porovnám jeho predikční úspěšnost i na menším datasetu bez imputovaných hodnot. To zároveň umožní porovnat

finální predikce s predikcemi sázkových kanceláří. Jako předpokládaného vítěze zápasu podle predikcí sázkových kanceláří označuji toho zápasníka, na kterého je vypsán příznivější sázkový kurz, proto zde nelze použít imputované hodnoty, ale pouze ty vypsané kancelářemi.

Z většího předzpracovaného datasetu, který je použit k trénování modelů, byly vytvořeny dvě množiny záznamů, trénovací a testovací. Místo vytvoření i třetí, validační množiny je zde k ladění hyperparametrů a získání úspěšnosti modelů při trénování použita *Scikit-learn* funkce *cross_val_score*, která implementuje rozdělení křižovou validací[31]. Jelikož počet záznamů v datasetu se pohybuje alespoň v řádech tisíců, a kvůli menší výpočetní náročnosti, je při křížové validaci zvoleno výchozí rozdělení na 5 podmnožin, viz. článek Marcot[12]. Rozdělení bylo náhodně provedeno v poměru 80 : 20 pro trénovací a testovací množiny respektive. Při práci s jednotlivými podmnožinami, je důležité dbát na správné rozdělení dat, tak aby do testovací sady nebyly zaneseny žádné závislosti z trénovací sady. Po rozdělení jsem samostatně na trénovací a testovací množinu použil metodu *KNN Imputer*[32], která doplnila chybějící sázkové kurzy. Dále byly také z obou množin pomocí třídy *StandardScaler* vytvořeny další 2 datové sady, které obsahují strandardizovaná data[33]. Testovací statická množina dat je použita ke konečnému ohodnocení každého jednotlivého modelu po vybrání nejlepších hyperparametrů během křížové validace, to simuluje reálné použití na budoucích datech. Následný výběr nejlepšího finálního modelu je proveden porovnáním výsledků právě nad testovací množinou.

5.2 Klasifikační modely

Klasifikační modely, se kterými zde pracuji jsou *náhodné lesy, logistická regrese, SVM* (*Support Vector Machines*, metoda podpůrných vektorů), *XGB klasifikátor a MLP* (Multi-layer Perceptron, vícevrstvý perceptron, umělá neuronová síť). Zde nastíním jejich podstatu, důležité hyperparametry, které v práci ladím a poznatky při trénování modelů.

5.2.1 Náhodný les

První použitý model je *náhodný les*, ten patří mezi takzvané *ensemble* metody, kdy výsledný model je tvořen z více slabších modelů, zde z *rozhodovacích stromů*. Konečná predikce je tvořená z průměru hlasování jednotlivých stromů.

Zvolené hyperparametry k ladění:

[31] `https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation`

[32] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html`

[33] `https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler`

- *n_estimators* – Počet rozhodovacích stromů, ze kterých je model sestaven.
- *criterion* – Funkce posuzující kvalitu rozdělení dat na dvě podmnožiny, podle určité hodnoty vybraného příznaku. Tato funkce je součástí rozhodovacích stromů.
- *max_depth* – Nejvyšší možná hloubka jednotlivých rozhodovacích stromů.
- *max_features* – Množství příznaků, které zohlednit při hledání nejlepšího možného rozdělení dat na dvě podmnožiny.

Strandardizace dat nemá u tohoto modelu značný vliv, výsledky predikcí jsou skoro totožné.

5.2.2 Logistická regrese

Důležitý model, který se používá k predikcím kategorických proměnných. Využívá logistickou funkci neboli *sigmoidu*, tak aby se její výstup pohyboval v rozmezí od 0 do 1.

Zvolené hyperparametry k ladění:

- *C* – Takzvaná míra regularizace. Čím je tato hodnota nižší, tím méně příznaků je zvoleno k dalším operacím nad nimi, to snižuje riziko přeučení modelu.
- *solver* – Algoritmus použitý při optimalizaci modelu, tedy snaze konvergovat k řešení.
- *fit_intercept* – Zda se má v rozhodovací funkci použít určitá konstantní hodnota (neboli *intercept* či *bias*).
- *max_iter* – Maximální možný počet iterací, než by měl *solver* zkonvergovat.

V průběhu ladění hyperparametrů jsem zkusil zvýšit příznak *max_iter*, jelikož často testované modely nedosáhly konvergence. Zvýšení pomohlo a nejlépe predikující modely zde mají *max_iter* nastavený na vyšší hodnotu. Standardizace data zde pomohla a díky ní model dosáhl přibližně o 2 % vyšší predikční přesnosti.

5.2.3 Support Vector Machines

Zkráceně *SVM*, česky se tento model dá také nazvat jako metoda podpůrných vektorů. Model přímo hledá nadrovinu v prostoru příznaků, která odděluje predikované třídy a pracuje s body ležícími blízko u této nadroviny (body se nazývají podpůrné vektory). Zde jsem po více běžích experimentů použil dvě

implementace, *LinearSVC* a *SVC*, obě z knihovny *scikit-learn*. *SVC* implementace umožňuje v algoritmu použít jiná jádra, než pouze lineární. Po mých pokusech, však i zde lineární jádro dosahovalo značně lepších výsledků než *rbf* a *sigmoid* jádra, proto jsou ve finálním notebooku vynechána.

Zvolené hyperparametry k ladění:

- *C* – Hodnota míry regularizace, totožně jako u *logistické regrese* 5.2.2.
- *penalty* – Norma sloužící k výpočtu penalizace.
- *dual* – Určuje jakou formulaci řešeného problému má algoritmus počítat[34].

Nestandardizovaná data jsem u implementace *SVC* ve finálním notebooku vynechal, kvůli značně delší době výpočtů a konstantně horším výsledkům. Výsledky predikcí jsou ale jinak příznivé, nejlepších dosahuje *LinearSVC* model nad standardizovanými daty.

5.2.4 XGBoost klasifikátor

XGBoost patří také mezi *ensemble* metody, a stejně jako *náhodný les* je tvořen z *rozhodovacích stromů*. Jako jediný z použitých modelů pro tento model neexistuje implementace v knihovně *scikit-learn*, místo toho je použita knihovna *xgboost*. Tento model, dosahuje všeobecně velmi dobrých výsledků, efektivně pracuje s pamětí a vnitřními výpočty a dokáže pracovat i s chybějícími daty. Standardizace dat nemá na výsledky modelu vliv, v notebooku jsou ponechány predikce nad oběma variantami dat, a lze vidět, že výsledky jsou naprosto totožné. Kvůli počtu zvolených hyperparametrů k ladění zabralo trénování modelu nejdelší dobu, přibližně 45 minut.

Zvolené hyperparametry k ladění:

- *gamma* – Minimální nutné snížení chyby nutné k dalšímu rozdělení listu ve stromu. Čím vyšší hodnota, tím konzervativnější algoritmus bude.
- *learning_rate* – Po jak velkých krocích se mají aktualizovat váhy příznaků z výstupů jednotlivých stromů. Slouží k větší opatrnosti algoritmu, tedy zabránění přeučení.
- *max_depth* – Nejvyšší možná hloubka jednotlivých rozhodovacích stromů.
- *n_estimators* – Počet rozhodovacích stromů, ze kterých je model sestaven.
- *reg_alpha* – Regularizační hodnota, snižuje složitost modelu a tím pomáhá zabránit jeho přeučení.

[34] `http://www.adeveloperdiary.com/data-science/machine-learning/support-vector-machines-for-beginners-duality-problem/`

5.2.5 Vícevrstevný perceptron

Anglicky *Multi-layer Perceptron, MLP*, jedná se o jeden z nejvíce používaných typů umělé neuronové sítě. Skládá se z více výpočetních jednotek zvaných neuronů, které jsou propojeny tak, že přijímají výstupy z předchozích vrstev a posílají své výsledků do vrstev následujících. Dokáže rozpoznat velmi složité vztahy i v mnohorozměrných prostorech dat. Sestavení perceptonu do konečné podoby, tedy rozpoznání skrytých stavů, se dělá během takzvaného učení sítě pomocí zpětného šíření chyby. Nevýhodou je absence možnosti zobrazení nejvíce vypovídajících příznaků po natrénovaní modelu, je to kvůli složité vnitřní reprezentaci vztahů.

Zvolené hyperparametry k ladění:

- *hidden_layer_sizes* – Počet neuronů v jednotlivých skrytých vrstvách.
- *activation* – Aktivační funkce pro skryté vrstvy.
- *alpha* – Regularizační hodnota, snižuje složitost modelu a tím pomáhá zabránit jeho přeučení.
- *learning_rate_init* – Hodnota použitá při aktualizaci vah jednotlivých výstupů během učení.
- *learning_rate* – Zda zůstane hodnota aktualizace vah konstatní, nebo se bude měnit v závislosti na průběžné chybě modelu.
- *max_iter* – Maximální počet běhů učení, takzvaných epoch. Kolikrát může být každý datový bod použit pro učení modelu.

U tohoto modelu jsem nejčastěji měnil hodnoty možných kombinací hyperparametrů. V prvních bězích učení jsem používal hlubší a širší vrstvy u parametru *hidden_layer_sizes*, který reprezentuje počet neuronů ve skrytých vrstvách sítě. Poté jsem zjistil, že například síť s jednou skrytou vrstvou o 20 neuronech dosahuje příznivějších výsledků. Také jsem zkusil zvýšit *learning_rate_init*, jelikož občas model nekonvergoval po dosažení *max_iter* (také zvětšeno). A naposledy jsem změnil výchozí *solver* na *sgd*, tedy optimalizaci vah používaných k učení modelu pomocí gradientního sestupu[35]. Síť natrénovaná na nestandardizovaných datech ve všech případech predikuje stejný výsledek, tedy je nepoužitelná. Naopak po standardizaci dosahuje tento model velmi dobrých výsledků.

[35] `https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31`

5.3 Porovnání výsledků modelů

Obě sledované statistiky vypočítávám nad daty z testovacího datasetu. Přesnost modelů se pohybuje od 57,3 % do 64,8 %. Výsledky natrénovaných modelů jsou zapsané v tabulce 5.1.

Lze vidět, že nejúspěšnější model je *vícevrstevný percepton* a hned za ním *lineární SVM*, oba modely predikují nad standardizovanými daty. Nejlepší nalezené hyperparametry jsou následující pro percepton: *solver='sgd', max_iter=200, learning_rate_init=0.003, learning_rate='constant', hidden_layer_sizes=(20,), alpha=0.0001, activation='tanh'*. Pro *SVM* pak: *penalty='l2', dual=False, C=0.001*. Standardizace dat pomohla k lepší přesnosti o jednotky procent ve všech případech. Je zajímavé, že *vícevrstevný percepton* natrénovaný na nestandardizovaných datech se umístil jako nejhorší model, jelikož v 99 % predikuje stejnou třídu. *XGBoost klasifikátor* se umístil předposlední i přes to, že ladění jeho hyperparametrů jsem věnoval zvýšenou pozornost.

Model	Standardizovaná data	Přesnost	Senzitivita
Vícevrstevný percepton	Ano	64,85 %	49 %
Lineární SVM	Ano	64,51 %	48 %
Lineární SVM	Ne	63,83 %	45 %
Logistická regrese	Ano	63,66 %	45 %
SVM (jiná implementace)	Ano	63,66 %	44 %
Náhodný les	Ano	63,15 %	37 %
Náhodný les	Ne	62,13 %	33 %
Logistická regrese	Ne	61,29 %	44 %
XGBoost klasifikátor	Bez rozdílu	59,59 %	43 %
Vícevrstevný percepton	Ne	57,38 %	2 %

Tabulka 5.1: Seřazené výsledky všech modelů

5.4 Zkoumání predikcí nejlepších modelů

Na základě výsledků jsem vybral dva nejúspěšnější modely a srovnal jejich úspěšnost na menším datasetu, který neobsahuje uměle doplněné hodnoty. Použil jsem jak *vícevrstevný percepton*, tak *lineární SVM* kvůli velmi podobným výkonům, a také díky možnosti *SVM* modelu zobrazit nejvíce vypovídající příznaky.

Postup rozdělení dat, trénování a otestování úspěšnosti byl totožný jako při práci s ostatními modely a je také popsán v notebooku *Experiments.ipynb*. Dále jsem také porovnal tato testovací data s predikcemi na základě sázkových kurzů. Výsledky jsou popsány v tabulce 5.2. Ukázalo se, že oba modely pracují nad

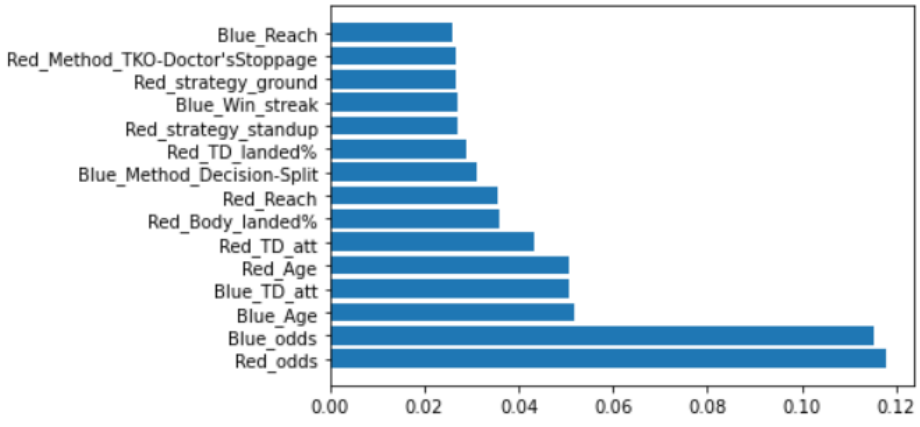
neimputovanými daty ještě lépe, než nad původním datasetem. Oba dosahují přesnosti okolo 66 % a senzitivita perceptonu činí celých 58 %.

Pozoruhodnější je však úspěšnost predikce čistě na základě sázkových kurzů, ta dosahuje 67 %. Dá se tedy říci, že finální model, který se skládá z dat sázkových kurzů i dalších předzpracovaných dat, není přesnější než predikce ze samotných kurzů. V potaz se nicméně musí vzít i hodnota senzitivity pro vítězství zápasníka v modrém rohu, ta je stále vyšší u predikcí obou zvolených modelů, tedy modely lépe odhadují vítězství u zpravidla podceňovaných zápasníků. Tato schopnost je užitečná například pro reálné sázení na vítěze.

Z grafu 5.1, který zobrazuje nejvíce vypovídající příznaky podle *SVM*, lze vidět, že zápasové kurzy jsou zdaleka nejdůležitější příznak. To vysvětluje zlepšení modelu při práci s daty bez uměle doplněných kurzů. Další v pořadí podle důležitosti jsou příznaky popisující věk bojovníka v den zápasu a také počet pokusů o *takedown*, tedy snaha dostat protivníka na zem. Z dalších typů příznaků objevujících se v grafu nelze jednoznačně usoudit, jaký styl boje má největší dopad na vítězství. Mohu ale jednoznačně doporučit vytvoření příznaku `Age`.

Model	Standardizovaná data	Přesnost	Senzitivita
Sázkové kurzy		67,09 %	52 %
Vícevrstevný percepton	Ano	66,01 %	58 %
Lineární SVM	Ano	65,80 %	54 %

Tabulka 5.2: Seřazené výsledky nejúspěšnějších modelů



Obrázek 5.1: 15 nejvíce vypovídajících příznaků pro predikci

Závěr

V práci se zabývám predikcí vítězů profesionálních MMA zápasů. Hlavní částí rešerše bylo popsat nejvýznamnější MMA organizace a najít vhodné zdroje o proběhlých zápasech v jednotlivých organizacích. Některé zdroje se povedlo najít, ale také jsem zjistil, že i významné organizace neposkytují snadno dostupné statistiky a často existují jen velmi omezené záznamy nebo pouze amatérsky vytvořené datasety, například na doméně Kaggle. Proto jsem se rozhodl provést predikce pouze nad daty z organizace UFC, a k tomu využít strojové stažení dat z oficiálních stránek UFC statistik. Také jsem vybral vhodný zdroj historických sázkových kurzů na proběhlé UFC zápasy.

Dále jsem provedl rešerši existujících prací zabývajících se také predikcí MMA zápasů. Jelikož je MMA relativně čerstvý sport, neexistuje velké množství odborných publikací na toto téma. Několik existujících je nicméně představeno a popsáno, důraz je kladen na použité klasifikační modely a úspěšnost provedených predikcí. Při rešerši jsem zjistil, že zápasník nastupující do zápasu v červeném rohu je zpravidla očekávaný vítěz a v datasetech vyhrávají tito zápasníci znatelně častěji. Proto byl při výběru vhodného finálního modelu kladen důraz na vysokou úspěšnost při predikování vítěze z modrého rohu.

První cíl praktické části byl získání všech nutných dat a jejich předzpracování do vhodné podoby. Zápasová data i sázkové kurzy byly staženy bez problémů napodobením GET požadavků ve vlastních skriptech. Předzpracování také proběhlo příznivě, důležité bylo takzvané doménové porozumění. Hlavní bylo transformovat data tak, aby statistiky k danému zápasu nepopisovaly to, co se odehrálo během zápasu, ale místo toho obsahovaly průměrné statistiky ze všech doposud proběhlých zápasů pro oba zúčastněné zápasníky.

Poslední cíl byl už samotné vytvoření finálního prediktivního modelu a otestování jeho kvality. Na základě rešerše existujících prací jsem vybral 5 vhodných modelů a ty následně trénoval a ladil na získaných datech. Model se vytvořit podařilo, a jeho výsledky považuji za obstojné. klasifikační přesnost na testovacích datech je vyšší než v předchozích pracích, výsledný nejúspěšnější model je *vícevrstevný percepton* s přesností 66,01 %. Nicméně i zde je prostor

pro zlepšení. Umělé doplnění chybějících hodnot sázkových kurzů se ukázalo jako škodlivé pro predikční schopnost modelu, a jednoduchá predikce pouze na základě sázkových kurzů má vyšší přesnost než nejúspěšnější model. V budoucnu bych se zaměřil na získání většího množství dat, případně na robustnější způsob doplnění těch chybějících (například aby se součet doplněných sázkových kurzů obou zápasníků rovnal 1). Dále bych zkusil ováhovat záznamy, aby novější zápasy měly větší vliv na výslednou predikci a také experimentovat s laděním dalších hyperparametrů u výsledných modelů.

Literatura

[1] Muthanna, P.: Conor McGregor surpasses Messi and Ronaldo to become highest paid athlete. *International Business Times*, May 2021, [cit. 2021-12-12]. Dostupné z: https://www.ibtimes.co.uk/conor-mcgregor-surpasses-messi-ronaldo-become-highest-paid-athlete-1690301

[2] of Encyclopaedia, T. E.: mixed martial arts. *Encyclopedia Britannica*, 09 2019, [cit. 2021-12-12]. Dostupné z: https://www.britannica.com/sports/mixed-martial-arts

[3] The Unified Rules of Mixed Martial Arts. [cit. 2021-12-12]. Dostupné z: https://www.ufc.com/unified-rules-mixed-martial-arts

[4] Taylor, T.: Only Watching the UFC? You're Missing Out on These World-Class MMA Promotions. *Men's Journal*, Oct 2021, [cit. 2021-12-12]. Dostupné z: https://www.mensjournal.com/sports/mma-promotions/

[5] McQuaide, M.: Applying Machine Learning Algorithms to Predict UFC Fight Outcomes. Aut 2019, [cit. 2021-13-12]. Dostupné z: http://cs229.stanford.edu/proj2019aut/data/assignment_308875_raw/26426025.pdf

[6] ul-hassan Usmani, Z.: What is Kaggle, Why I Participate, What is the Impact? *Kaggle*, 2017, [cit. 2021-13-12]. Dostupné z: https://www.kaggle.com/getting-started/44916#

[7] Sohail, S.: How Do Odds Work in Betting? *Investopedia*, Feb 2021, [cit. 2022-01-10]. Dostupné z: https://www.investopedia.com/articles/investing/042115/betting-basics-fractional-decimal-american-moneyline-odds.asp

[8] Hitkul, K. A.; Yadav, N.; Dwivedy, M.: A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights. *Harmony Search*

and Nature Inspired Optimization Algorithms Advances in Intelligent Systems and Computing, 2018: s. 67–76.

[9] James, L. P.; Robertson, S.; Haff, G. G.; aj.: Identifying the performance characteristics of a winning outcome in elite mixed martial arts competition. *Journal of science and medicine in sport*, ročník 20, č. 3, 2017: s. 296–301.

[10] Johnson, J. D.: *Predicting outcomes of mixed martial arts fights with novel fight variables*. Dizertační práce, University of Georgia, 2012.

[11] Baratloo A, N. A. E. A. G. P. . S. D., Hosseini M; Calculation of Accuracy, S.; PMC4614595., S. E. T. . S.-. P. . P.: Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency (Tehran, Iran)*, ročník 3, č. 2, 2015: s. 48–49, [cit. 2022-22-04].

[12] Marcot, B.; Hanea, A.: What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, ročník 36, 09 2021, doi:10.1007/s00180-020-00999-9, [cit. 2022-15-04].

Příloha **A**

Seznam použitých zkratek

GUI Graphical user interface

XML Extensible markup language

API Application Programming Interface

Příloha B

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src` ..... kódy implementace a použité datové soubory
  - `scraping`
    - `data`
  - `notebooks`
- `text` ..... text práce
  - `thesis.pdf`
  - `thesis.tex`
  - `thesis.zip`