Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

****

Využití statistických metod pro predikci výsledků v oblasti bojových sportů

BAKALÁŘSKÁ PRÁCE

Studijní program: Aplikovaná informatika

Autor: Adam Zacpal

Vedoucí bakalářské práce: Ing. Jan Fojtík, Ph.D.

Praha, květen 2024

Poděkování

Děkuji…

Abstrakt

Lorem ipsum

Klíčová slova

amet, consectetuer, dolor, Lorem ipsum, sit.

Abstract

Lorem ipsum

Keywords

amet, consectetuer, dolor, Lorem ipsum, sit.

Obsah

[Úvod 10](#_Toc162483942)

[1 Smíšená bojová umění a predikce výsledků 11](#_Toc162483943)

[1.1 Historie MMA 11](#_Toc162483944)

[1.2 Pravidla MMA 12](#_Toc162483945)

[1.3 Predikce výsledků v MMA 13](#_Toc162483946)

[1.4 Statistické prediktivní modely 14](#_Toc162483947)

[1.4.1 Strojové učení 14](#_Toc162483948)

[1.4.2 Elo model 15](#_Toc162483949)

[1.5 Nástroje pro práci s daty 16](#_Toc162483950)

[1.5.1 Python 16](#_Toc162483951)

[1.5.2 Další nástroje 17](#_Toc162483952)

[2 Vstupní data 18](#_Toc162483953)

[2.1 Web scraping z tapology.com 18](#_Toc162483954)

[2.2 Web scraping z sherdog.com 20](#_Toc162483955)

[2.3 Propojení datasetů 22](#_Toc162483956)

[3 Analýza 23](#_Toc162483957)

[4 Modely 24](#_Toc162483958)

[5 Výsledky 25](#_Toc162483959)

[Závěr 26](#_Toc162483960)

[Použitá literatura 27](#_Toc162483961)

[Přílohy I](#_Toc162483962)

[Příloha A: Název první přílohy I](#_Toc162483963)

[Příloha B: Název druhé přílohy I](#_Toc162483964)

Seznam obrázků

[Obr. 2.1 Seznam turnajů UFC (www.tapology.com) 18](#_Toc162483528)

[Obr. 2.2 Profil zápasníka (www.sherdog.com) 22](#_Toc162483529)

Poznámka: Seznam obrázků je vhodný použít, pokud počet obrázků v textu práce je větší než 20. Seznam grafů je vhodný použít pouze v případě, že autor práce rozlišuje mezi obrázkem a grafem. Seznam grafů je utvářen, pokud je počet grafů větší než 20. V této šabloně závěrečné práce se grafy a obrázky nerozlišují.

Seznam tabulek

[Tab. 1.1 Záznam výsledků závodu robotů (vlastní zpracování) **Chyba! Záložka není definována.**](#_Toc529974579)

Poznámka: Seznam tabulek je vhodný použít, pokud počet tabulek v textu práce je větší než 20.

Seznam výpisů programového kódu

[Výpis 2.1 Procházení seznamu "page\_list" a uložení turnajů do DF (vlastní zpracování) 19](#_Toc162483495)

[Výpis 2.2 Získávání informací z profilu zápasníka (Obr. 2.2) (vlastní zpracování) 21](#_Toc162483496)

Poznámka: Seznam výpisů programového kódu je vhodný použít, pokud počet vložených objektů tohoto typu je větší než 20.

Seznam zkratek

MMA Mixed Martial Arts (smíšená bojová umění)

UFC Ultimate Fighting Championship

Poznámka: Seznam zkratek je vhodný použít, pokud počet zkratek v textu práce je větší než 20 a nejedná se o zkratky běžné.

Úvod

Bakalářská práce se zabývá využitím statistických metod pro predikci sportovních výsledků. Pro práci jsem si vybral sport smíšených bojových umění, konkrétně zápasy v nejprestižnější organizaci tohoto sportu UFC.

V práci budu muset sesbírat vhodná data z dostupných zdrojů, upravit je tak aby byla použitelná pro další části, analyzovat klíčové proměnné a následně vytvořit modely logistické regrese a Elo model. Logistická regrese identifikuje pravděpodobnosti výsledků na základě různých faktorů, zatímco Elo model ohodnocuje hráče či týmy na základě historie výkonů.

Výsledky těchto modelů poskytují hlubší vhled do sportovních událostí a v praxi by mohli pomáhat například při tvorbě kurzových nabídek sázkových kanceláří. Práce představuje komplexní přístup k analýze a predikci sportovních výsledků pomocí moderních statistických technik a knihoven programovacího jazyka python. Zajímavost práce spočívá i v tom, že je tento sport poměrně mladý a není zdaleka tak prozkoumaný jako například fotbal a hokej.

# Smíšená bojová umění a predikce výsledků

Smíšená bojová umění neboli MMA (z anglického Mixed Martial Arts), je plno kontaktní bojový sport kombinující techniky boxu, kickboxu, brazilského jiu-jitsu, zápasu, juda, thajského boxu a jiných disciplín (Britannica, 2024) zahrnující údery, kopy, strhy a další techniky boje na zemi. Zápasy se konají v kleci ze sloupků a pletiva, aby se zabránilo případnému vypadnutí a zranění zápasníků.

## Historie MMA

Smíšená bojová umění mají svůj původ už v roce 648 př. n. l., kdy se ve starověkých olympijských hrách objevil pankration (Britannica, 2024). Tohle bojové umění kombinovalo zápas, box a pouliční boj. Zakázáno bylo pouze dloubání očí a kousání. Zápas končil, když jeden z bojovníků uznal porážku nebo upadl do bezvědomí. V některých případech zápasníci během zápasu zemřeli, nejčastěji v důsledku udušení (Wikipedia, 2011). Pankration byl jednou z nejoblíbenějších disciplín starověkých olympijských her. V roce 393 n. l. císař Theodosius I. zakázal konání olympijských her, což vedlo k poklesu popularity pankrationu.

Na začátku 20. století se toto bojové umění objevilo znovu v Brazílii pod názvem „vale tudo“[[1]](#footnote-2). O jeho popularizaci se velkou mírou postarali bratři Carlos a Hélio Gracie (Britannica, 2024). Ti se učili judo od Japonského mistra Mitsuyo Maedy a sami rozvíjeli tento styl se zaměřením na boj na zemi. Později si založili vlastní školu, a aby dokázali, jak je jejich styl efektivní, vyzývali kohokoliv na zápasy bez pravidel, kde bylo vše povoleno. Jejich veřejné výzvy se staly populární a zápasy brzy vyprodávaly fotbalové stadiony (Britannica, 2024).

Do povědomí širší veřejnosti v Severní Americe se MMA dostalo v roce 1993, kdy se rodina Gracieových rozhodla rozšířit svou značku „brazilské jiu-jitsu“ a podílela se na organizaci turnaje „UFC 1“, kde se mělo rozhodnout, který styl zápasení je nejefektivnější (Britannica, 2024). Turnaje se zúčastnilo celkem osm zápasníků, každý reprezentující jeden z bojových stylů a to box, savate, sumo, kickbox, karate, taekwondo, shootfighting[[2]](#footnote-3) a brazilské jiu jitsu, které reprezentoval Héliův syn Royce Gracie. Ačkoliv byl Royce výrazně menší než všichni ostatní, zápasy poměrně jednoduše vyhrál a ihned získal přízeň fanoušků.

Organizace „Ultimate Fighting Championship“, která tento turnaj pořádala rychle získávala na popularitě. Turnaje sledovalo stále více lidí, s čímž přišlo i velké množství negativních reakcí zejména k brutalitě sportu a absenci dostatečných pravidel. V této době zápasy připomínaly pouliční rvačky a sport neměl dobrou pověst. Nové vedení „UFC“ v roce 2001 vytvořilo pravidla, která měla učinit sport bezpečnější. Byly přidány váhové kategorie, kola, časové limity a rozšířil se seznam zakázaných technik (Britannica, 2024).

V současnosti je MMA jedním z nejpopulárnějších sportů po celém světě a turnaje nejznámější organizace „UFC“ si kupují miliony lidí.

## Pravidla MMA

Regulační orgány ve Spojených státech v roce 2009 přijaly soubor pravidel známý jako „Unified rules of MMA“, který brzy přijaly organizace po celém světě a sjednotily tak pravidla sportu i mimo USA (Britannica, 2024). Tento standard vychází z pravidel vytvořených organizací „UFC“ v roce 2001.

Podle těchto pravidel bojují zápasníci ve speciálních rukavicích, s odhalenými prsty, bez bot a chráničem na zuby. Utkání je ve většině případů vypsáno na tři kola po pěti minutách s minutovými pauzami mezi koly. U titulových, či hlavních zápasů je to pět kol po pěti minutách. Každé kolo hodnotí 3 rozhodčí na základě pevně daných kritérií samostatně, na konci zápasu sečte každý zvlášť své body a po porovnání bodů s dalšími dvěma rozhodčími se určí vítěz. Mezi kritéria, která se hodnotí, patří v tomto pořadí od nejdůležitějšího po nejméně důležité: efektivní údery a práce na zemi, efektivní agresivita, kontrola zápasu (ASSOCIATION OF BOXING COMMISSIONS AND COMBATIVE SPORTS, 2022). Zápas končí remízou jen výjimečně za velmi specifických podmínek.

Souboj může skončit také předčasně, a to v případě, kdy jeden ze zápasníků není schopen dále pokračovat v boji. Takový stav může nastat v případě „knouckoutu“[[3]](#footnote-4), „submise“[[4]](#footnote-5) či například zranění.

Navzdory trvajícímu přesvědčení odpůrců sportu o jeho brutalitě, stále více výzkumů ukazuje, že MMA je například bezpečnější než box. Výzkum „Combative Sports Injuries: An Edmonton Retrospective“ říká, že ačkoliv je výskyt zranění u zápasníků MMA větší, často se jedná pouze o pohmožděniny či jiné drobné zranění, zatímco u boxerů je vyšší pravděpodobnost vážných zranění, jako je otřes mozku, ztráta vědomí nebo poranění očí (Kapman, Reid, Phillips, Qin, Gross, 2016).

## Predikce výsledků v MMA

V MMA, podobně jako v jakémkoliv jiném sportu je schopnost předpovědět výsledek zápasu velmi hodnotná. Fanoušci, sázkaři, trenéři i analytici by si jistě přáli křišťálovou kouli, která by jim řekla, jak bude zápas probíhat a jak dopadne. Umět správně předpovědět zápas však není vůbec snadné a je to složitý proces, který vyžaduje zohlednění mnoha faktorů a na který se dá nahlížet z mnoha úhlů pohledů.

Metody predikce výsledků:

1. Analýza statistik: Studium předchozích výsledků zápasníků, počtu vítězství a porážek, způsobů vítězství a porážek, počtu úderů, takedownů[[5]](#footnote-6) a podobně.
2. Fyzické a technické hodnocení: Hodnocení fyzické kondice, síly, rychlosti a technických schopností zápasníka.
3. Hodnocení taktické připravenosti: Sledování videí, podrobná analýza stylu zápasení v postoji i na zemi a schopnosti držet se stanovené taktiky boje.
4. Zohlednění tréninkových a psychologických faktorů: Hodnocení aktuální formy, motivace, tréninkových metod a dalších faktorů, jako jsou například zranění či rivalita zápasníků.
5. Využití predikcí expertů: Sledování prognóz a připomínek profesionálních analytiků, komentátorů a novinářů.
6. Využití statistických modelů: Statistické modely, jako je například logistická regrese, mohou být použity k analýze historických dat a identifikaci proměnných, které jsou spojeny s určitým vyústěním zápasu.

Nejlepšího výsledku lze dosáhnou pouze kombinací, všech těchto metod, ale ani tak nelze zaručit správnost predikce. Vzhledem k povaze sportu o úspěchu často rozhodují ty nejmenší detaily, které nemusí být zřejmé ani zkušenému analytikovi. Důležité je dlouhodobé a soustavné udržení pozitivní výkonnosti.

V MMA můžeme předpovídat například vítěze zápasu, způsob ukončení nebo kolo ukončení a různé kombinace těchto možností.

Příklady různých úrovní predikce:

1. Zápasník „A“ vyhraje.
2. Zápasník „B“ vyhraje na body.
3. Zápasník „A“ vyhraje na K.O. ve třetím kole.

## Statistické prediktivní modely

Prediktivní modelování je statistický proces používaný k předvídání budoucích událostí nebo výsledků analýzou vzorců v daném souboru vstupních dat. Jedná se o klíčovou součást prediktivní analýzy, což je typ datové analytiky, který využívá aktuální a historická data k předpovídání aktivit, chování a trendů (Lawton, Carew, Burns, 2022).

### Strojové učení

V rámci strojového učení existují dva základní přístupy: učení s učitelem a učení bez učitele.

Učení s učitelem je přístup, který je definován použitím označených datasetů, to znamená, že pro vstupní data je známý správný výstup. Datasety jsou určeny k trénování algoritmů, které poté přesně klasifikují data nebo správně předpovídají výsledky (Delua, 2021). Učení s učitelem lze dále rozdělit na dva typy problémů:

1. Klasifikační modely: Využívají algoritmus k přesnému zařazení testovacích dat do určitých kategorií, například k oddělení housek od rohlíků nebo k oddělení spamu od doručené pošty (Delua, 2021). Mezi běžné typy klasifikačních algoritmů patří metoda podpůrných vektorů nebo rozhodovací stromy.
2. Regrese: Využívá algoritmy k pochopení vazem mezi závislými a nezávislými proměnnými. Regresní modely jsou užitečné pro předpovídání číselných hodnot na základě různých datových bodů, například pro prognózy příjmů z prodeje pro danou firmu (Delua, 2021). Mezi oblíbené regresní algoritmy patří lineární regrese, logistická regrese a polynomiální regrese.

Učení bez učitele využívá algoritmy strojového učení k analýze a shlukování souborů neoznačených dat. Tyto algoritmy objevují skryté vzory v datech bez nutnosti lidského zásahu. Modely učení bez učitele se používají pro tři hlavní úlohy (Delua, 2021):

1. Shlukování (Clustering): Shlukování má za úkol rozdělit neoznačená data do různých skupin, tak aby podobné datové body spadaly do stejného shluku, a odlišné datové body do jiného (Kaushik, 2019). Technika je užitečná například pro kompresi obrázků či segmentaci trhu.
2. Asociace: Využívá různá pravidla k nalezení vztahů mezi proměnnými v dané množině dat. Tyto metody se často používají pro analýzu tržního koše a algoritmů, které například doporučují, který film by se vám mohl líbit na základě pozorovaných pravidel.
3. Redukce dimenzionality: Je to technika učení, která se používá v případě, že počet proměnných v daném souboru dat je příliš vysoký. Snižuje počet datových vstupů na zvládnutelnou velikost a zároveň zachovává integritu dat (Delua, 2021).

### Elo model

Elo rating je statistický model používaný k hodnocení relativní síly hráčů v hrách s nulovým součtem („zero-sum games“). To jsou hry, při kterých zisk jednoho hráče ekvivalentní znamená ztrátu druhého hráče. Elo rating byl vyvinut maďarským profesorem Arpadem Elo pro hodnocení hráčů šachu, ale později byl adaptován i pro jiné hry a sporty. Hlavním cílem ratingu je poskytnout objektivní měřítko pro porovnání síly hráčů a předpovídání výsledků zápasů.

Princip modelu spočívá v relativním hodnocení hráče na základě jeho vlastního ratingu a ratingu jeho soupeře. Každý hráč začíná s určitým počátečním hodnocením, který se po každém zápase aktualizuje na základě jeho očekávaného a skutečného výsledku. Když hráč s horším ratingem porazí ratingově lepšího soupeře sebere mu hodně bodů, když hráč s lepším ratingem porazí horšího soupeře sebere mu málo bodů a když zápas skončí remízou, hráč s nižším ratingem získá malou část bodů od toho s vyšším (Mittal, 2020). Hodnocení hráčů se mění postupně v čase, výrazná změna v hodnocení vyžaduje dlouhodobě stabilní výkony.

Pravděpodobnost výhry neboli očekávané skóre hráče Alice se vypočítá pomocí vzorce:

RB značí rating hráče Bob a RA rating hráče Alice.

Vzorec pro úpravu počátečních skóre, kde K je K-faktor úpravy skóre a SA počet bodů za výsledek zápasu (1 za výhru, 0 za prohru, 0,5 za remízu):

Řekněme tedy, že se utká Alice s ratingem 1600 a Bob s ratingem 1400.

Očekávané skóre Alice = 0,759746926:

Očekávané skóre Boba = 0,240253074:

Vidíme, že součet očekávaného skóre Alice a Boba je roven 1.

Řekněme dále, že K = 32 a vypočítejme nové ratingy obou hráčů.

Nový rating hráče Alice je nyní 1608 a Boba 1392.

V současnosti se Elo model využívá kromě šachu také ve sportu, deskových hrách, videohrách, ale také v různých seznamovacích aplikacích, které na základě Elo modelu hodnotí profily uživatelů a podle nich pak profily nabízejí méně či více.

## Nástroje pro práci s daty

Práce s daty je nedílnou součástí statistické analýzy a má několik částí, které zahrnují sběr dat, porozumění datům, přípravu dat, prediktivní modelování, vyhodnocení výsledků a nasazení výsledků do praxe. Každá z těchto částí má svá specifika a existuje pro ně široká škála nástrojů a technik. Tato kapitola se zaměřuje na představení a popis některých klíčových nástrojů, které jsou k dispozici pro práci s daty. Dva nejpopulárnější jazyky v této oblasti jsou Python a R (Datacamp, 2022).

### Python

Python je univerzální programovací jazyk, který se stal nedílnou součástí datové analýzy a datové vědy. Jeho jednoduchost, výkonnost, ale hlavně velké množství knihoven, za kterými stojí rozsáhlá komunita uživatelů, je příčinou jeho stále rostoucí popularity (Datacamp, 2022). Python nabízí téměř neomezené možnosti, které zprostředkovávají právě zmíněné knihovny v oblasti získávání, čištění a vizualizace dat nebo strojového učení.

1. Pandas: Open-source knihovna pro manipulaci s daty, která poskytuje nástroj „DataFrame“, což je snadno použitelná a výkonná tabulková struktura, která umožňuje snadné ukládání, načítání a indexování dat. Pandas poskytuje také jednorozměrnou datovou strukturu „Series“ a další funkce pro operace s nimi.
2. NumPy: NumPy je knihovna pro vědecké výpočty v Pythonu. Poskytuje vysokoúčinné datové struktury, jako jsou vícedimenzionální pole „ndarray“, a širokou škálu matematických funkcí pro práci s daty.
3. Matplotlib: Je to knihovna pro vizualizaci dat v Pythonu. Poskytuje širokou škálu funkcí pro tvorbu statických, interaktivních a animovaných grafů a vizualizací dat.
4. Seaborn: Seaborn je knihovna pro vizualizaci dat založená na Matplotlib, která usnadňuje tvorbu esteticky příjemných a informativních grafů. Často se používá pro vizualizaci výsledků analýzy dat a pro zkoumání vztahů mezi proměnnými.
5. Scikit-learn: Scikit-learn je knihovna pro strojové učení v Pythonu. Obsahuje širokou škálu algoritmů pro klasifikaci, regresi, shlukování, redukci dimenzionality a další techniky strojového učení, stejně jako nástroje pro evaluaci a předzpracování dat.
6. Requests: Requests je knihovna pro práci s HTTP požadavky v Pythonu. Je často používána pro komunikaci s webovými API a stahování dat z internetu.
7. BeautifulSoup: Populární knihovna, která slouží k extrakci dat z HTML a XML dokumentů. Je často používána pro web scraping, což je proces získávání strukturovaných dat z webových stránek.

### Další nástroje

R je další programovací jazyk, který se často používá v statistice a analýze dat. Podobně jako Python obsahuje mnoho balíčků pro manipulaci s daty, vizualizaci a statistické analýzy, jako je například dplyr, ggplot2 a tidyr. Stejně jako Python i R dokáže pokrýt všechny oblasti datové vědy. Oproti Pythonu je však v R náročnější řešení složitějších úloh (Datacamp, 2022).

Excel je běžně používaný nástroj pro jednoduchou analýzu a zpracování dat pomocí vestavěných funkcí. Není však zdaleka tak robustní a pro velké datasety může být pomalý a neefektivní. Python oproti Excelu umožňuje reprodukci a sdílení kódu pomocí skriptů a notebooků nebo pokročilejší metody strojového učení.

Jupyter Notebook je interaktivní vývojové prostředí, které umožňuje kombinovat kód, text a vizualizace do jednoho dokumentu. Často se používá pro explorativní analýzu dat a podporuje různé programovací jazyky včetně Pythonu a R.

# Vstupní data

Vstupní data pro bakalářskou práci byla získána pomocí technik web scrapingu veřejně dostupných dat z nejpopulárnějších webových stránek pro zaznamenávání výsledků MMA zápasů tapology.com a sherdog.com, které byly následně spojeny k vytvoření jednoho datasetu. Obě stránky poskytují informace o zápasnících, jejich zápasech a organizacích. Tento projekt byl omezen na data z nejprestižnější MMA organizace UFC, a to kvůli konzistenci úrovně zápasů, která se napříč organizacemi velmi liší. K vytěžení těchto webů byly použity knihovny Pythonu requests a beatifulsoup, pro práci a ukládání dat knihovna pandas. Použité vývojové prostředí bylo Jupyter Notebook.

## Web scraping z tapology.com

Tato kapitola popisuje proces extrakce dat v příloze č. 1 – tapology\_events.ipynb.

Tento notebook má za cíl získat data o všech zápasech v UFC s dostupnými podrobnostmi o zápasnících, kteří se jich zúčastnili. Stránka tapology.com byla vybrána z důvodu přehledného členění turnajů podle organizací a bylo tak jednoduché získat potřebná data.

Obsah obrázku text, snímek obrazovky, Písmo, číslo

Popis byl vytvořen automatickyNejprve se pošle request na url, která odkazuje na první stránku seznamu turnajů organizace UFC. Navrácený obsah se přečte pomocí html parseru a uloží do BeautifulSoup objektu. Pomocí vyhledávací metody a cyklu se z BeautifulSoup objektu vždy vytěží odkaz na další stránku se seznamem turnajů (Obr. 2.1), uloží se do seznamu stránek a přejde na tento odkaz. Tohle se opakuje, dokud nedojdeme na poslední stránku. Výsledkem je seznam „page\_list“ plný odkazů, které ukazují na stránky se seznamem turnajů.

Obr. 2.1 Seznam turnajů UFC (www.tapology.com)

Výpis 2.1 Procházení seznamu "page\_list" a uložení turnajů do DF (vlastní zpracování)

events = pd.DataFrame(columns=["Event Link", "Event Name", "Date", "Venue", "Location"])

# Iterates through page\_list

for page in tqdm(page\_list):

data = req.get(page, headers = headers).text

soup = BeautifulSoup(data,"html.parser")

# Finds all listings on a page

listings = soup.find\_all(class\_="fcListing")

# Iterates listings and saves each one in DataFrame

for listing in listings:

link = "https://www.tapology.com" + listing.find("a")["href"]

name = listing.find("a").text

date = listing.find(class\_="datetime").text

venue = ""

if listing.find(class\_="venue") != None:

venue = listing.find("span", class\_="venue").text

location = ""

if listing.find(class\_="venue-location") != None:

location = listing.find(class\_="venue-location").text

region = ""

if listing.find(class\_="region") != None:

region = listing.find(class\_="region").text

events = events.\_append({"Event Link":link, "Event Name":name, "Date":date, "Venue":venue, "Location":location}, ignore\_index=True)

events = events.replace(r'\n','', regex=True)

V druhém kroku se „page\_list“ prochází cyklem, který podobným procesem, jako v předchozím odstavci, najde všechny turnaje na stránce (Obr. 2.1) a uloží je do DataFramu „events“ (Výpis 2.1).

DataFrame „events“ obsahuje sloupce: „Event Link“, „Event Name“, „Date“, „Venue“, „Location“.

Ve třetím kroku se konečně získávají cílová data o zápasech procházením odkazů DataFramu „events“. Každá stránka má informace o turnaji a zápasech na nich. Algoritmus nejdříve zjistí, zda se turnaj již uskutečnil, vyhledá obecné informace o turnaji a poté iteruje po jeho zápasech a hledá informace o zúčastněných zápasnících, výsledku zápasu, způsobu ukončení zápasu, váhové kategorii a dalších dostupných informacích. Dále najde tabulku s detaily zápasu, kde vyhledá informace například o vypsaných kurzech, výsledku vážení či věku zápasníků. Na konci iterace se všechny získané informace uloží do DataFramu „bouts“, kde jeden řádek znamená jeden zápas.

DataFrame „bouts“ má 7577 řádků a obsahuje sloupce: "Bout Link", "Fighter A", "Link A", "Fighter B", "Link B", "Nickname A", "Nickname B", "Record A", "Record B", "Odds A", "Odds B", "Title A", "Title B", "Weight A", "Weight B", "Age A", "Age B", "Height A", "Height B", "Reach A", "Reach B", "Result", "Time", "Weightclass", "Rounds", "Event Link", "Venue", "Date", "Location", "Billing", "Event Name".

Pro další vyhledávání bylo nezbytné zjistit, zda v DataFramu nemáme zápasníky se stejným jménem. Víme, že zápasníci sice můžou mít stejná jména, ale odkaz na jejich profil musí být vždy unikátní. Pomocí sloupců „Fighter A“, „Fighter B“, „Link A“ a „Link B“ a metody „groupby“ jsme tedy prozkoumali, zda není jedno jméno propojeno s více odkazy a bylo zjištěno, že ve dvou případech se tak stalo. V jednom případě se jednalo o bratry se stejným prvním jménem, proto bylo přidáno jejich prostřední jméno pro jednoznačnou identifikaci. V druhém případě byla pro jednoznačnou identifikaci přidána přezdívka zápasníků. Upravený DataFrame s 7577 zápasy a 2479 unikátními zápasníky byl uložen do souboru „bouts\_new.csv“ (příloha).

## Web scraping z sherdog.com

Tato kapitola popisuje proces extrakce dat v příloze č. 2 – sherdog\_records.ipynb.

Tato část má za cíl získat data primárně o předchozích zápasech všech zápasníků z datasetu „bouts\_new.csv“, ale také doplňující data o samotných zápasnících, která budou později připojena k prvnímu datasetu. Stránka sherdog.com byla vybrána z důvodů konzistentního formátu zaznamenávání metod ukončení zápasu, jako tomu není u tapology.com.

Nejdříve byl vytvořen seznam unikátních jmen zápasníků z datasetu získaného v předchozí kapitole. Pro každé jméno se spustil algoritmus, který pomocí google vyhledávače našel odkaz na Sherdog profil zápasníka (Obr. 2.2). Pomocí html parseru byla stránka z odkazu opět přečtena a uložena do BeautifulSoup objektu. Z profilu byl opět pomocí vyhledávacích metod objektu vytěženy obecné informace o zápasnících a poté požadovaná tabulka se záznamy o jednotlivých zápasech zápasníka (Výpis 2.2). Tyto informace byly uloženy do DataFramu „records“, kde jeden řádek znamená opět jeden zápas.

DataFrame „records“ má 54860 záznamů a obsahuje sloupce: "Fighter", "URL", "Nationality", "Nickname", "Birthday", "Height ft", "Height cm", "Opponent", "Result", "Method", "Round", "Time", "Event", "Date", "Referee" byl uložen do souboru „records.csv“.

Testováním algoritmu bylo zjištěno, že pro některá jména vyhledávání profilu vykazovalo chybné chování, proto byla jména ještě před spuštěním algoritmu změněna. Jednalo se o jména „Ian Machado Garry“, „Sako Chivichitan“ a „Razak Al-Hassan“. Po dokončení byla jména nahrazena původními hodnotami.

Výpis 2.2 Získávání informací z profilu zápasníka (Obr. 2.2) (vlastní zpracování)

def get\_record(fighter):

fighter\_data, url = get\_website(fighter)

fighter\_soup = BeautifulSoup(fighter\_data, "html.parser")

# Finds fighter information

nationality = ""

if fighter\_soup.find("strong", itemprop="nationality") != None:

nationality = fighter\_soup.find("strong", itemprop="nationality").text

nickname = ""

if fighter\_soup.find("span", class\_="nickname") != None:

nickname = fighter\_soup.find("span", class\_="nickname").find("em").text

birthday = ""

if fighter\_soup.find("span", itemprop="birthDate") != None:

birthday = fighter\_soup.find("span", itemprop="birthDate").text

height\_ft = ""

height\_cm = ""

if fighter\_soup.find("b", itemprop="height")!= None:

height\_ft = fighter\_soup.find("b", itemprop="height").text

if fighter\_soup.find("b", itemprop="height").nextSibling.nextSibling.nextSibling != None:

height\_cm = fighter\_soup.find("b", itemprop="height").nextSibling.nextSibling.nextSibling

# Finds table with fighters record

table = fighter\_soup.find("table", class\_="new\_table fighter")

fighter\_record = pd.DataFrame(columns=["Fighter", "URL", "Nationality", "Nickname", "Birthday", "Height ft", "Height cm", "Opponent", "Result", "Method", "Round", "Time", "Event", "Date", "Referee"])

# Fills the DataFrame with fighter record

for row in table.find\_all("tr", class\_=""):

col = row.find\_all("td")

if (col != []):

result = col[0].text

opponent = col[1].text

event = col[2].a.text

date = col[2].find("span", class\_="sub\_line").text

method = col[3].b.text

referee = col[3].span.text

rnd = col[4].text

time = col[5].text

fighter\_record = fighter\_record.\_append({"Fighter":fighter, "URL":url, "Nationality":nationality, "Nickname":nickname, "Birthday":birthday, "Height ft":height\_ft, "Height cm":height\_cm, "Opponent":opponent, "Result":result, "Method":method, "Round":rnd, "Time":time, "Event":event, "Date":date, "Referee":referee}, ignore\_index=True)

return fighter\_record

Obsah obrázku text, muž, snímek obrazovky, Webové stránky

Popis byl vytvořen automaticky

Obr. 2.2 Profil zápasníka (www.sherdog.com)

## Spojení datasetů

Tato kapitola popisuje propojení datasetů v příloze č. 3 – creating\_dataset.ipynb.

V této části propojíme dva datasety získané v předchozích kapitolách do jednoho, s kterým budeme dále pracovat. Cíl, kterého chceme dosáhnout, je vypočítat pro každý záznam v souboru „bouts\_new.csv“ (tedy pro každý zápas v UFC), počet vítězství a proher obou zápasníků na body, K.O. a submisi pro oba zápasníky a připojit k nim doplňující informace. Tyto hodnoty se budou počítat vždy pouze do posledního zápasu, který předcházel zápasu, pro který hodnoty počítáme. Abychom mohli tyto hodnoty vypočítat a datasety propojit, musíme nejdříve upravit některé sloupce v tabulce ze souboru „records.csv“.

### Úprava souboru „records.csv“

Sloupec „Method“ reprezentuje způsob ukončení zápasu. V MMA existují pouze čtyři hlavní způsoby ukončení: na body, K.O., submise a „no contest“[[6]](#footnote-7). Po prozkoumání sloupce bylo zjištěno, že obsahuje až 733 různých hodnot, a to z důvodu, že každý záznam, kromě metody ukončení, obsahuje také popis v závorce, který už může mít mnoho podob. Tento popis byl ze sloupce „Method“ odstraněn a uložen do nového sloupce „Description“. Sloupec „Method“ nyní z důvodu nekonzistence obsahoval už jen 20 různých hodnot, které byly dále sloučeny do požadovaných čtyř.

1. „KO“, „TKO a „K.O“ bylo sloučeno pod „KO/TKO“.
2. „Submission“, „Technical Sumbission“, „Technial Submission“ a „"Submission (Rear-Naked Choke" bylo sloučeno do „Submission“.
3. „Draw“ and „Technical Draw“ není metoda ukončení, ale výsledek zápasu a každá remíza končí na body, proto byly tyto hodnoty sloučeny do „Decision“.
4. "No Contest", "NC", "No Decision - Overturned by CSAC", "No Contest - Collard Failed Drug Test", "No Decision" a "ND" znamenají, že zápas nemá výsledek, tudíž je můžeme sloučit do „No Contest“.
5. „Technical Decision“ můžeme sloučit do „Decision“, jelikož je to pouze jiný druh toho samého.
6. Všechny diskvalifikace se v MMA počítají jako „KO/TKO“, proto můžeme záznamy "Disqualification", "DQ" a „DG“ přidat k této hodnotě.

Po provedení těchto operací zůstaly ve sloupci „Method“ pouze požadované čtyři různé hodnoty (Tab. 2.1). DataFrame byl uložen do souboru „records\_clean.csv“.

Tab. 2.1 Četnost metod ukončení v datasetu (vlastní zpracování)

|  |  |
| --- | --- |
| Metoda ukončení | Počet |
| KO/TKO | 20223 |
| Decision | 17871 |
| Submission | 16213 |
| No Contest | 457 |

### Připojení dat o počtu a typu ukončení

Nyní použijeme DataFrame ze souboru „records\_clean.csv“, který jsme upravili v minulé kapitole a přidáme ho k datasetu o zápasech v UFC „bouts\_new.csv“.

Tato operace byla provedena na základě jmen zápasníků a dat zápasů. Jména zápasníků byla v obou datasetech totožná, data zápasů však měla v obou datasetech jiný formát, a musela tak být pomocí jednoduché funkce a knihovny „datetime“ převedena.

Výpis 2.3 Funkce, která získá "record" a informace zápasníka a uloží ho do seznamu (vlastní zpracování)

def calculate\_records(fighter\_name, bout\_date):

df = records.loc[records["Fighter"] == fighter\_name]

# Finds fighter information

nationality, nickname, birthday, referee = "", "", "", ""

nationality = df.iloc[0]["Nationality"]

nickname = df.iloc[0]["Nickname"]

birthday = df.iloc[0]["Birthday"]

if not df.loc[df["Date"] == bout\_date, "Referee"].empty:

referee = df.loc[df["Date"] == bout\_date, "Referee"].iloc[0]

# Filters DataFrame for the date of a given fight

df = df.loc[df["Date"] < bout\_date]

# Calculates fighters record

counts = df.groupby(["Result","Method"])["Fighter"].count()

ko\_win, ko\_loss, sub\_win, sub\_loss, dec\_win, dec\_loss, draws = 0, 0, 0, 0, 0, 0, 0

for index, value in counts.items():

if index[0] + index[1] == "winKO/TKO":

ko\_win = value

elif index[0] + index[1] == "lossKO/TKO":

ko\_loss = value

elif index[0] + index[1] == "winSubmission":

sub\_win = value

elif index[0] + index[1] == "lossSubmission":

sub\_loss = value

elif index[0] + index[1] == "winDecision":

dec\_win = value

elif index[0] + index[1] == "lossDecision":

dec\_loss = value

elif index[0] == "draw":

draws = value

info\_list=[]

info\_list = [ko\_win, ko\_loss, sub\_win, sub\_loss, dec\_win, dec\_loss, draws, nationality, birthday, nickname, referee]

return info\_list

Poté byla na každý záznam v tabulce „bouts\_new“ zavolána funkce „calculate\_records“ (Výpis 2.3), která pomocí jména zápasníka a data zápasu vypočítá informace o jeho předchozích zápasech a vrátí je spolu s doplňujícími informacemi o něm. Informace vrácené z funkce „calculate\_records“ se následně nalepí k záznamu v tabulce „bouts\_new“.

Výsledný dataset má tyto sloupce: „Bout Link“, „Fighter A“, „Link A“, „Fighter B“, „Link B“, „Nickname A“, „Nickname B“, „Record A“, „Record B“, „Odds A“, „Odds B“, „Title A“, „Title B“, „Weight A“, „Weight B“, „Age A“, „Age B“, „Height A“, „Height B“, „Reach A“, „Reach B“, „Result“, „Time“, „Weightclass“, „Rounds“, „Event Link“, „Venue“, „Date“, „Location“, „Billing“, „Event Name“, „KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „Nationality A“, „Birthday A“, „Nick A“, „Referee“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“, „Nationality B“, „Birthday B“, „Nick B“, „Referee B“ a byl uložen do souboru „bouts\_detailed.csv“.

## Čištění datasetu

Konečně jsme získali dataset se všemi potřebnými informacemi k další analýze a modelování. Nyní musíme tato data vyčistit, upravit a doplnit chybějící hodnoty. Dataset „bouts\_detailed“ nyní obsahuje padesát tři sloupců z nichž je velká část zdvojená pro zápasníka „A“ a zápasníka „B“.

Všechny sloupce byly podrobeny základní analýze, kde bylo zjištěno, že sloupce „Nickname A“ a „Nickname B“ se nepodařilo naplnit daty. Sloupce můžeme odstranit a použít místo nich „Nick A“ a „Nick B“, o které jsme data doplnili v minulém kroku.

Podobně jako v kapitole 2.3.1 musíme v tomto datasetu upravit sloupec „Result“, který kromě metody ukončení zápasu obsahuje také její popis. Vznikl tak nový sloupec „Description“ a sloupec „Result“ nyní obsahuje pouze metody ukončení, proto byl jeho název změněn na „Method“. Sloupec „Result“ naplníme skutečným výsledkem zápasu. Protože jsou data strukturovaná tak, že zápasník „A“ je vždy, mimo případů remízy či zápasu bez výsledku, vítěz, sloupec bude obsahovat pouze hodnoty „win“, „NC“ a „draw“. Hodnotu „Draw“ tak nyní můžeme v sloupci „Method“ změnit na „Decision“, jelikož ji už máme ve sloupci „Result“.

Dále bylo ze sloupce „Time“ vytaženo kolo ukončení zápasu a uloženo do sloupce „Finish“. V případě, že zápas skončil na body, byla do sloupce přidána hodnota 0, v opačných případech hodnoty 1 – 5 pro ukončení v prvním až pátém kole.

Pro každý zápas máme dva záznamy o ringovém rozhodčím, každý zápas však má pouze jednoho, proto byly doplněny chybějící hodnoty v sloupci „Referee“ hodnotami ve sloupci „Referee B“ a sloupec „Referee B“ byl smazán.

V dalším kroku byly upraveny sloupce „Height A“, „Height B“, „Reach A“ a „Reach B“, které obsahují hodnoty v palcích a centimetrech. Ponechány byly pouze číselné hodnoty v centimetrech pro snadnější manipulaci.

Sloupce „Weight A“ a „Weight B“, obsahují hmotnost v librách i v kilogramech, ponechány byly pouze číselné hodnoty v librách. Tohle rozhodnutí bylo učiněno z důvodu, že sloupec „Weightclass“ obsahuje také hodnoty v librách a při přepočítávání by mohlo dojít k nepřesnostem.

Ze sloupců „Age A“ a „Age B“ ponecháme pouze roky. Měsíce, týdny a dny jsou u věku zápasníků irelevantní.

Sloupce „Odds A“ a „Odds B“ obsahují americké kurzy zápasníků a jejich kategorii. Kategorii extrahujeme a uložíme do sloupců „Odds cat A“ a „Odds cat B“. V původních sloupcích zůstanou pouze číselné hodnoty.

### Odstranění nepotřebných hodnot

Pro další analýzu a modelování už nebudeme potřebovat tyto sloupce, proto byly odstraněny:

1. „Bout Link“,
2. „Time“ – důležitou informaci jsme umístili do sloupce „Finish“,
3. „Event Link“ – ponecháme „Event Name“, který má podobný význam,
4. „Link A“ a „Link B“,
5. „Record A“, „Record B“ – detalní „record“ máme rozložen v jiných sloupcích,
6. „Birthday A“ a „Birthday B“ – ve sloupci „Age A“ a „Age B“ máme věk zápasníků, se kterým se bude lépe pracovat.

Dále bylo rozhodnuto smazat všechny záznamy zápasů, které se odehrály před adopcí „NJSACB“ pravidel, vytvořených v roce 2001 (1.1), protože obsahovali velké množství nekompatibilních dat se zbytkem datasetu. Smazáno bylo 283 řádků.

### Doplnění chybějících hodnot

V sloupci „Rounds“ byly nalezeny dva záznamy, které nebyli v souladu s pravidly MMA od zmíněné změny. Tyto záznamy byly manuálně prověřeny a došlo se k závěru, že jsou pouze špatně zapsané a byly změněny na odpovídající hodnotu. 69 chybějících hodnot bylo nahrazeno hodnotou „3 x 5“ pro běžné zápasy a hodnotou „5 x 5“ pro titulové a hlavní zápasy.

Chybějící hodnoty ve sloupci „Odds A“ a „Odds B“ byly nahrazeny hodnotou „+100“, která v amerických kurzech znamená vyrovnané šance. Pro sloupce „Odds cat A“ a „Odds cat B“ to byla hodnota „Even“.

Ve sloupci „Weight A“ a „Weight B“ nahradíme chybějící hodnoty, hodnotou sloupce „Weightclass“. To znamená, že zápasník navážil přesně domluvenou hmotnost.

„Age A“ a „Age B“, tedy věk zápasníků, byl nahrazen průměrem datasetu „30“.

Výšku a rozpětí zápasníků „Height A“, „Height B“, „Reach A“ a „Reach B“ nahradíme průměrem v jejich váhové kategorii. U jednoho zápasu však stále chyběla hodnota „Reach B“, protože pro tuto váhovou kategorii nebyl v datasetu žádný záznam. Informace se nepodařila dohledat, proto byla hodnota v záznamu nahrazena průměrem nejbližší nižší váhové kategorie.

Po doplnění hodnot byly změněny datové typy sloupců:

1. Na datový typ „int“: „Rounds“, „Odds A“, „Odds B“, „Age A“, „Age B“ a „Finish“.
2. Na datový typ „float“: „Weight A“, „Weight B“, „Height A“.

Dataset byl uložen do souboru „ufc\_bouts.csv“.

## Finální podoba datasetu

Dataset „ufc\_bouts“ obsahuje 7294 záznamů a má 46 sloupců popsaných v tabulce (Tab. 2.2), z nichž je 19 kategoriálních a 27 numerických.

Kategoriální proměnné jsou: „Fighter A“, „Fighter B“, „Result“, „Venue“, „Date“, „Location“, „Billing“, „Event Name“, „Referee“, „Description“, „Method“, „Title A“, „Nationality A“, „Nick A“, „Odds cat A“, „Title B“, „Nationality B“, „Nick B“, „Odds cat B“.

Numerické proměnné jsou: „Weightclass“, „Rounds“, „Finish“, „Odds A“, „Weight A“, „Age A“, „Height A“, „Reach A“, „KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „Odds B“, „Weight B“, „Age B“, „Height B“, „Reach B“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“.

Tab. 2.2 Popis sloupců tabulky "ufc\_bouts" (vlastní zpracování

|  |  |  |  |
| --- | --- | --- | --- |
| Název sloupce | Popis | Název sloupce | Popis |
| Fighter A | Jméno zápasníka A | Location | Lokace zápasu |
| Fighter B | Jméno zápasníka B | Billing | Označení zápasu v rámci turnaje |
| Result | Výsledek zápasu | Event Name | Jméno turnaje |
| Weightclass | Váhová kategorie, ve které byl zápas vypsán | Referee | Rozhodčí v kleci |
| Rounds | Počet vypsaných kol zápasu | Description | Popis ukončení zápasu |
| Venue | Aréna, ve které se zápasilo | Method | Metoda ukončení zápasu |
| Date | Datum zápasu | Finish | Kolo ukončení zápasu |
| Odds A | Kurz na vítězství zápasníka A | Odds B | Kurz na vítězství zápasníka B |
| Title A | Titulový status zápasníka A | Title B | Titulový status zápasníka B |
| Weight A | Výsledek vážení záp. A | Weight B | Výsledek vážení záp. B |
| Age A | Věk v době zápasu záp. A | Age B | Věk v době zápasu záp. B |
| Height A | Výška záp. A | Height B | Výška záp. B |
| Reach A | Rozpětí paží záp. A | Reach B | Rozpětí paží záp. B |
| KO win A | Počet výher KO/TKO záp. A | KO win B | Počet výher KO/TKO záp. B |
| KO loss A | Počet proher KO/TKO záp. A | KO loss B | Počet proher KO/TKO záp. B |
| Sub win A | Počet výher submisí záp. A | Sub win B | Počet výher submisí záp. B |
| Sub loss A | Počet proher submisí záp. A | Sub loss B | Počet proher submisí záp. B |
| Dec win A | Počet výher na body záp. A | Dec win B | Počet výher na body záp. B |
| Dec loss A | Počet proher na body záp. A | Dec loss B | Počet proher na body záp. B |
| Draws A | Počet remíz záp. A | Draws B | Počet remíz záp. B |
| Nationality A | Národnost záp. A | Nationality B | Národnost záp. B |
| Nick A | Přezdívka záp. A | Nick B | Přezdívka záp. B |
| Odds cat A | Kurzová kategorie na vítězství záp. A | Odds cat B | Kurzová kategorie na vítězství záp. B |

# Analýza

# Modely

# Výsledky

Závěr

Lorem ipsum

Použitá literatura

ASSOCIATION OF BOXING COMMISSIONS AND COMBATIVE SPORTS. (2022). Official Unified Rules of MMA.

Britannica, T. Editors of Encyclopaedia (2024, March 21). mixed martial arts. Encyclopedia Britannica. <https://www.britannica.com/sports/mixed-martial-arts>

Přispěvatelé projektů Wikimedia. (2015, May 8). *starověké bojové umění z Řecka*. Wikipedia.org; nadace Wikimedia. <https://cs.wikipedia.org/wiki/Pankr%C3%A1tion>

Karpman, S., Reid, P., Phillips, L., Qin, Z., & Gross, D. P. (2016). Combative Sports Injuries: An Edmonton Retrospective. *Clinical journal of sport medicine : official journal of the Canadian Academy of Sport Medicine*, *26*(4), 332–334. <https://doi.org/10.1097/JSM.0000000000000235>

Lawton, G. (2022, January). *What is Predictive Modeling?* SearchEnterpriseAI. <https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling>

Delua, J. (2021, March 12). *Supervised vs. Unsupervised Learning: What’s the Difference?* IBM Blog. <https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>

Kaushik, S. (2019, March 11). *An Introduction to Clustering & different methods of clustering*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>

Mittal, R. (2020, November 6). *What is an ELO Rating?* Medium. <https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0>

Datacamp. (2022, December). *Python vs R for Data Science: Which Should You Learn?* Www.datacamp.com. <https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference>

Poznámka: Za koncem Použité literatury nutné nesmazat *konec oddílu*, aby fungovalo různé číslování hlavního textu práce a příloh

Přílohy

Lorem ipsum

1. Název první přílohy

Lorem

1. Mauris elementum mauris vitae tortor.
2. Aliquam id dolor.

Maecenas fermentum, sem in pharetra pellentesque, velit turpis volutpat ante, in pharetra metus odio a lectus.

1. Název druhé přílohy

Lorem ipsum

1. „Vale tudo“ v překladu „vše dovoleno“. [↑](#footnote-ref-2)
2. Shootfighting je termín, který předcházel MMA. Pochází z Japonských soutěží obdobného typu a dnes se již nepoužívá. [↑](#footnote-ref-3)
3. Knockout, zkráceně K.O., je situace, kdy zápasník po inkasovaném úderu či kopu buď plně ztratí vědomí nebo rozhodčí usoudí, že už se není schopen dále efektivně bránit. V takovém případě se jedná o takzvaný „technický knockout“. [↑](#footnote-ref-4)
4. Submise je způsob ukončení při, kterém jeden ze zápasníku nějakým chvatem donutí protivníka vzdát se. V případě, kdy se zápasník nechce vzdát a například při škrcení ztratí vědomí nebo utrží zranění, se kterým nelze pokračovat v zápase, rozhodčí ukončí zápas takzvanou „technickou submisí“. [↑](#footnote-ref-5)
5. Takedown je v terminologii MMA označení pro strh/poraz bojovníka k zemi a následnou kontrolu. [↑](#footnote-ref-6)
6. „No contest“ znamená, že zápas je bez výsledku. Nikam se nepočítá a je na něj nahlíženo, jako by se nestal. [↑](#footnote-ref-7)