Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra statistiky

Diplomová práce

Využití prediktivního modelování pro predikci výsledku fotbalového zápasu

Martin Hemer

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Martin Hemer

Ekonomika a management
Provoz a ekonomika

Název práce

Využití prediktivního modelování pro predikci výsledku fotbalového zápasu

Název anglicky

Using predictive modelling for football match result prediction

---

Cíle práce

Cílem diplomové práce je vytvořit prediktivní model pro predikci výsledku fotbalového zápasu s využitím dat české nejvyšší soutěže s následným možným rozšířením i na jiné soutěže. První fází bude vytvoření jednoduchého prediktivního modelu na základě Poissonovy distribuce, který následně bude rozšiřován. Druhou fází bude složitější model, který bude pracovat i s individuálními vlastnostmi hráčů a specifickými proměnnými. Cílem práce bude dále zhodnocení možného ekonomického využití daného modelu jako investičního nástroje.

Metodika

Literární rešerše bude zaměřena na popis stávajících modelů pro predikci výsledků fotbalových zápasů. Pro tvorbu vlastního modelu budou použita data z internetových zdrojů oficiálních stránek vybraných soutěží, na základě kterých bude vytvořena potřebná databáze pro dané soutěže s daty o jednotlivých týmech (počet bodů, vstřelených branek, forma, atd.) a jednotlivých hráčích (zranění, počet branek, počet asistencí, atd.). Pro tvorbu modelů budou využity metody z oblasti matematického modelování, zejména pak modely založené na Poissonově distribuci a pravděpodobnosti vstřelení určitého množství branek danými týmy v daném zápase.

Doporučený rozsah práce

30-40 stran

Klíčová slova

fotbal, kvalita modelu, Poissonova distribuce, prediktivní modelování, výsledek zápasu

---

Doporučené zdroje informací

BIERMANN, C. Football Hackers: The Science and Art of a Data Revolution. Chichester, England: BLINK Publishing, 2019. ISBN 1788702050

FIELD, A. Discovering statistics using IBM SPSS Statistics. Thousand Oaks: SAGE Publications, 2013. ISBN 978-1-4462-4917-8

MEMMERT, D., RAABE, D. Data analytics in football: Positional data collection, modelling and analysis. Boca Raton, FL: CRC Press, 2018. ISBN 978-0815381549

MENDENHALL, W., SCHEAFFER, R. L., WACKERLY, D.D. Mathematical statistics with Applicatios. Boston: Duxbury Press, 1981. ISBN 0-87872-279-3

SUMPTER, D. Soccermatics. Bloomsbury Publishing, 2017. ISBN 978-1-4729-2414-8

Předběžný termín obhajoby

2021/22 LS – PEF

Vedoucí práce

Ing. Zuzana Pacáková, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 8. 9. 2021

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 19. 10. 2021

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 30. 03. 2023

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Využití prediktivního modelování pro predikci výsledku fotbalového zápasu" jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 31. 3. 2023 _____

Poděkování

Rád bych touto cestou poděkoval paní doktorce Zuzaně Pacákové za vstřícnost, podporu a trpělivost. Dále bych chtěl poděkovat rodině za podporu. A v poslední řadě mé přítelkyni a nejbližším přátelům, že mě dokázali přimět k dokončení této práce.

Využití prediktivního modelování pro predikci výsledku fotbalového zápasu

Abstrakt

Tato diplomová práce se zaměřuje na využití prediktivního modelování pro predikci výsledku fotbalového zápasu. Cílem této práce je vytvořit prediktivní model, který bude schopen odhadnout celkový počet branek obou týmů v zápase a celkového vítěze zápasu. Následně bude tento model aplikován na trhu kurzových sázek a bude zkoumána jeho profitabilita. Pro dosažení těchto cílů bude využita literární rešerše v teoretické části práce, která se věnuje problematice trhu kurzových sázek a historickým prediktivním modelům jako jsou například Poissonovo rozdělení, Random forest a Neuronové sítě. V praktické části následně budou vytvořeny vlastní modely v programovacím jazyce Python. Tato práce se tedy zabývá problémem, zda je možné pomocí prediktivních modelů vydělat na kurzových sázkách. Výsledkem této práce by měla být odpověď na otázku, zda je využití prediktivního modelování na trhu kurzových sázek vhodný investiční nástroj, nebo jen gambling.

Klíčová slova: fotbal, kvalita modelu, Poissonova distribuce, prediktivní modelování, výsledek zápasu

Using predictive modeling for football match result prediction

Abstract

This master's thesis focuses on the use of predictive modeling to predict the outcome of a football match. The aim of this work is to create a predictive model that will be able to estimate the total number of goals scored by both teams in a match and the overall winner of the match. Subsequently, this model will be applied to the sports betting market, and its profitability will be examined. To achieve these objectives, a literature review will be conducted in the theoretical part of the work, which will address the issue of the sports betting market and historical predictive models such as Poisson distribution, Random Forest, and Neural Networks. In the practical part, custom models will be created in the programming language Python. This work, therefore, deals with the issue of whether it is possible to profit from sports betting using predictive models. The result of this work should be an answer to the question of whether the use of predictive modeling in the sports betting market a suitable investment tool is or just gambling.

Keywords: football, quality of the model, Poisson distribution, predictive modeling, outcome of a football match

Obsah

Seznam obrázků

1 Úvod

Tato diplomová práce se zabývá tématem využití prediktivního modelování pro predikci výsledku fotbalových zápasů. Hlavním cílem práce bude vytvoření prediktivního modelu, který bude schopen odhadnout celkový počet branek obou týmů v zápase a celkového vítěze zápasu. Dále se bude věnovat otázce, zda je možné pomocí takového modelu vydělat peníze na trhu kurzových sázek.

V první, teoretické, části práce bude provedena literární rešerše, která bude zaměřená na objasnění problematiky prediktivního modelování a jeho využití v oblasti fotbalu. Dále bude vysvětlen trh kurzových sázek a jeho fungování. V této části budou popsány historické prediktivní modely, jako je například Poissonova distribuce, Random forest nebo neuronové sítě. Díky této části by mělo být snazší pochopit, jak prediktivní modelování může být využito v praxi v oblasti fotbalu.

Druhá, praktická, část práce se bude věnovat metodologii vytvoření prediktivního modelu. Pro vytvoření modelu bude použit programovací jazyk Python, který bude pracovat s vlastní databází, jež bude vytvořena z oficiálních webových stránek vybraných fotbalových soutěží. Na základě této databáze budou vytvořeny vlastní modely, které budou schopny předpovědět výsledek fotbalového zápasu.

Závěr práce se zaměří na aplikování výsledků vytvořeného prediktivního modelu na trhu kurzových sázek a jeho profitabilitu. Cílem bude zjistit, zda je možné pomocí tohoto modelu vydělat peníze na kurzových sázkách.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem diplomové práce je vytvořit prediktivní model pro predikci výsledku fotbalového zápasu s využitím dat české nejvyšší soutěže s následným možným rozšířením i na jiné soutěže. První fází bude vytvoření jednoduchého prediktivního modelu na základě Poissonovy distribuce, který následně bude rozšiřován. Druhou fází bude složitější model, který bude pracovat i s individuálními vlastnostmi hráčů a specifickými proměnnými. Cílem práce bude dále zhodnocení možného ekonomického využití daného modelu jako investičního nástroje.

2.2 Metodika

Literární rešerše bude zaměřena na popis stávajících modelů pro predikci výsledků fotbalových zápasů. Pro tvorbu vlastního modelu budou použita data z internetových zdrojů oficiálních stránek vybraných soutěží, na základě kterých bude vytvořena potřebná databáze pro dané soutěže s daty o jednotlivých týmech (počet bodů, vstřelených branek, forma atd.) a jednotlivých hráčích (zranění, počet branek, počet asistencí, atd.). Pro tvorbu modelů budou využity metody z oblasti matematického modelování, zejména pak modely založené na Poissonově distribuci a pravděpodobnosti vstřelení určitého množství branek danými týmy v daném zápase.

3 Teoretická východiska

3.1 Fotbal a statistika

Fotbal je jedním z nejoblíbenějších sportů na světě, který přitahuje velkou pozornost fanoušků a sázkařů. Aby týmy mohly uspět, je důležité pochopit a analyzovat jejich výkonnost. V dnešní době se výkonnost týmů může měřit pomocí různých statistických metod. Tyto metody umožňují analyzovat různé faktory, jako jsou například úspěšnost při přihrávce, střelba na branku, úspěšnost v obraně a mnoho dalších.

Jednou z nejčastěji používaných statistických metod pro analýzu výkonu týmů je regresní analýza. Regresní analýza umožňuje vyhodnotit vztah mezi dvěma nebo více proměnnými. Tato metoda může být využita pro analýzu mnoha faktorů, jako je například vliv počtu střel na branku na celkový počet vstřelených gólů. Další často využívanou metodou je analýza clusterů, která umožňuje týmy rozdělit do skupin podle společných vlastností, jako je například způsob hry. (Groll, A., & Ley, C., 2016, s. 257-279)

Další statistickou metodou, která se v poslední době stala velmi populární, je náhodný les (random forest). Jedná se o metodu strojového učení, která umožňuje vytvoření modelu pro predikci výsledků zápasů. Tento model umožňuje analyzovat mnoho různých faktorů, jako jsou například forma týmu, úspěšnost střelby na branku, síla obrany a podobné. Náhodný les umožňuje vytvořit velmi přesný model, a umožňuje tak lepší predikci výsledků zápasů.

Další statistické metody, které se využívají v analýze výkonu týmů, jsou například analýza faktorů, analýza variance či korelační analýza. (Lasek, J., Szczypinski, P., & Kowalczyk, R., 2013, s. 169-174)

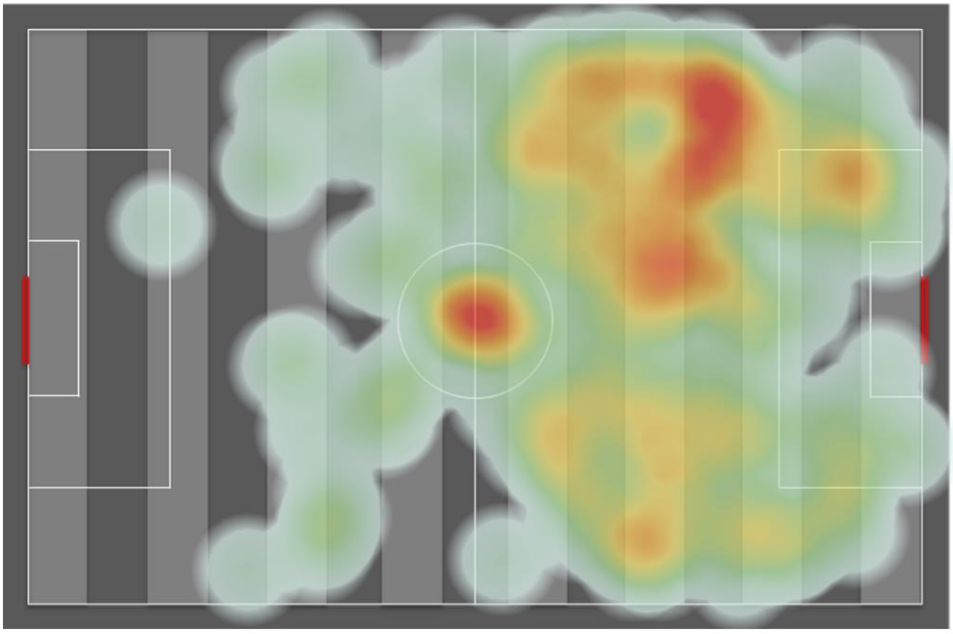
Je třeba podotknout, že statistika má v současnosti významné místo v analýze výkonu fotbalových týmů a predikci výsledků zápasů. Použití statistických metod jako je náhodný les, může být klíčové pro vytváření úspěšných prediktivních modelů. Tyto modely nejen pomáhají s výběrem týmu a sázek, ale také mohou přinést výhody při analýze a rozhodování týmů samotných. V budoucnosti se očekává, že se statistika bude stále více uplatňovat v

oblasti sportu, a že bude hrát stále důležitější roli při predikci výsledků a rozhodování týmů. (Shen, J., Zhang, H., & Cui, L., 2018)

3.1.1 Využití dat

S rozvojem datové analýzy samotných týmů ale využití statistiky nekončí. Fotbal, jako nejsledovanější sport, je i zároveň sportem, na který se nejvíce sází, proto se sportovní novináři a statistici snaží přinášet detailnější údaje o jednotlivých hráčích a týmech, jako jsou heat mapy (mapa pohybu při zápase včetně rychlosti), xG faktor – expected goals, vyhrané souboje apod.

Obrázek 1 Heat mapa



Zdroj: Kylian Mbappé Heat Map – Champions League 2021/22 – UEFA.com

Na obrázku výše je zmíněná heat mapa, která znázorňuje pohyb hráče Kyliana Mbappého z PSG v rámci všech zápasů Ligy Mistrů sezóny 2021/2022. Z obrázku je patrné, ve kterých prostorech se hráč nacházel nejčastěji (červenější místa) a v jakých částech hřiště se nenacházel téměř vůbec.

S popularitou sázení na fotbal, roste i zájem analytiků, kteří se snaží vytvářet statistické modely pro fotbalové zápasy. Ať už se jedná o samotné sázkové kanceláře a

bookmakery nebo naopak sázkaře, kteří se snaží prostřednictvím svých modelů najít příležitosti na trhu, tyto prediktivní modely s využitím znalosti risku se dají přirovnat například k finančním trhům s akciemi.

3.1.2 Prediktivní modely

Fotbal je sport, který po celém světě sleduje mnoho fanoušků. Zápas fotbalu může být plný překvapení, což ho dělá atraktivním pro diváky i sázkaře. Nicméně, s nástupem technologií a datových věd se fotbalový průmysl začíná spoléhat na prediktivní modely, které pomáhají předpovídat výsledky zápasů. Tento přístup využívá statistické analýzy, strojového učení a umělé inteligence k předpovídání výsledků, které mohou být použity při sázení na zápasy a vytváření strategií pro týmy.

Prediktivní modely jsou založeny na sběru a analýze velkého množství dat, jako jsou výsledky minulých zápasů, statistiky hráčů a týmů, povětrnostní podmínky, lokace stadiónu a mnoho dalších faktorů. Tyto modely pak používají algoritmy strojového učení k předpovídání výsledků zápasů s vysokou přesností. Tyto prediktivní modely jsou často používány ke zlepšení sázkových kurzů, což může být pro sázkaře velmi výhodné.

Využívání prediktivních modelů není nové. Už v roce 2003 vytvořili britští vědci z univerzity v Salfordu první složitější prediktivní model pro fotbalové zápasy. Tento model se zaměřil na to, jakým způsobem základní charakteristiky týmu jako jsou věk hráčů a úspěšnost minulých zápasů, ovlivňují jeho výkonnost v dalších zápasech. V současné době se prediktivní modely využívají v mnoha oblastech fotbalového průmyslu, včetně tvorby sázkových kurzů, výběru hráčů, sestavování týmů, a dokonce i v tréninku hráčů samotných. (Baio, G., 2019)

Jeden z nejúspěšnějších příkladů využití prediktivních modelů ve fotbale má na svědomí společnost StatsBomb, která se specializuje na sběr a analýzu dat pro fotbalové týmy. Tato společnost využívá strojové učení a umělou inteligenci k vytváření prediktivních modelů, které pomáhají fotbalovým týmům při výběru hráčů, taktiky a tréninku. Statistická analýza a modelování mohou pomoci trenérům identifikovat slabé stránky týmu a zlepšit výkon hráčů. (König, R., & Bühlmann, P., 2019)

Použití prediktivních modelů ve fotbale má potenciál změnit způsob, jakým fotbalový průmysl funguje. Tyto modely mohou pomoci týmům dosáhnout lepších výsledků a zlepšit výkon hráčů. Nicméně, stále existuje určitá míra kontroverze ohledně použití těchto modelů. Někteří lidé tvrdí, že tyto modely mohou být nepřesné a že spoléhání se na data může být nebezpečné. Další také argumentují, že použití prediktivních modelů může vést k ztrátě spontánnosti a překvapení, které dělají z fotbalu tak atraktivní sport.

Lze tedy říct, že prediktivní modely jsou významným prvkem moderního fotbalového průmyslu. Tyto modely mohou být využity k předpovídání výsledků zápasů, vytváření sázkových kurzů, výběru hráčů, taktiky a tréninku. Avšak, je důležité, aby byly použity s opatrností a s porozuměním jejich omezení a rizik. (Lasek, J., & Biecek, P., 2019, s. 39-53)

Fotbalové prediktivní modely se dají rozdělit na přímé a nepřímé:

3.1.2.1 Přímé modely

Přímé prediktivní modely se zabývají odhadem samotného výsledku, jejich výstupem je tedy pravděpodobnost výhry, remízy nebo prohry daného týmu v daném utkání. Vycházejí z historických dat zápasů (výsledků obou týmů), formy, postavení v tabulce a posledních vzájemných zápasů. Do těchto modelů se přidává i výhoda domácího hřiště, délka cestování pro hostující tým a např. i faktor rozhodčího.

3.1.2.2 Nepřímé modely

Nepřímé prediktivní modely naopak zkoumají jednotlivé proměnné v rámci utkání samostatně/odděleně. Nejtypičtějším modelem je model založení na lineární regresi, který zkoumá pravděpodobnost počtu branek domácího mužstva a pravděpodobnost branek hostujícího mužstva. Jsou i další modely na množství žlutých karet v zápase, kde je důležitým faktorem i rozhodčí. V poslední době velmi populárním modelem je model srovnávání síly jednotlivých mužstev. Vychází ze sehranosti mužstva, formace, které tým hraje, absence klíčových hráčů (zranění, disciplinární tresty) a samozřejmě porovnání důležitých atributů jednotlivých hráčů na daných postech.

3.2 Existující prediktivní modely

V kapitolách níže budou představeny důležité existující fotbalové prediktivní modely, jejich pořadí je chronologické. Jsou vybrány ty typy modelů, které svou strukturou inspirují autorův prediktivní model. U názvu modelu je uveden autor, rok a typ modelu.

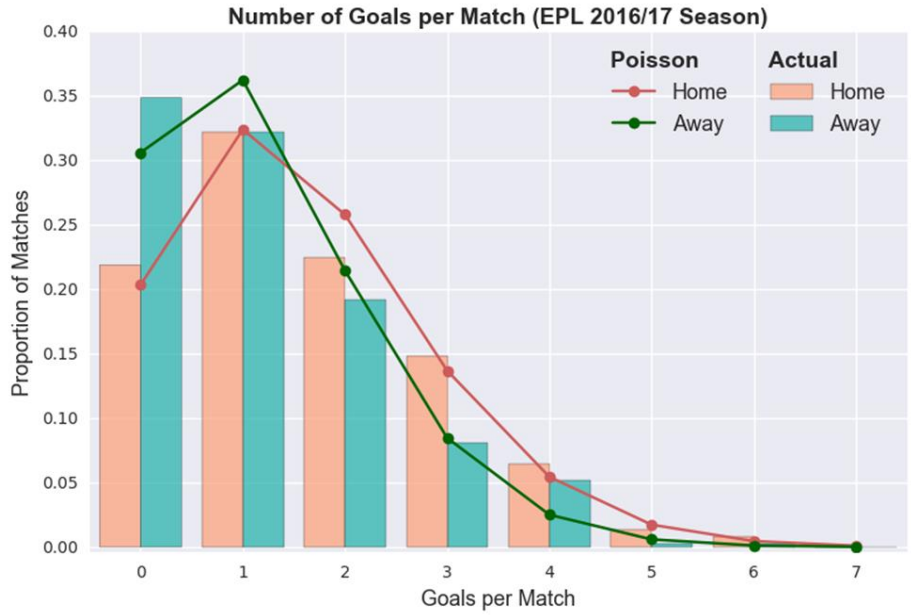
3.2.1 Maher (1982), Poissonovo rozdělení

Jeden z prvních veřejně publikovaných modelů. Poprvé publikován jako Maher-model v roce 1982. Tento model je v zásadě velmi jednoduchý, vychází z dat, která obsahují historické údaje o vstřelených a obdržených brankách a předpokládá, že tým, který v minulosti skóroval vícekrát v zápase má předpoklad k tomu vstřelit více branek i v budoucnu.

Zajímavostí je, že počítá nezávisle počet branek domácích a hostů, jedná se tedy o nepřímý model. Ve své práci zvolil čtyři nejvyšší anglické soutěže a data mezi lety 1973-1975. Pomocí Poissonova rozložení vypočítává jednotlivé pravděpodobnosti výsledku. Výsledkem je matice, kde hodnoty ve sloupcích označují pravděpodobnosti počtu branek hostí a v řádcích jsou pravděpodobnosti počtu branek domácích, průsečíky hodnot pak určuje pravděpodobnosti konkrétních výsledků. Tento model je pro svoji jednoduchost využívaný jako základní stavební kámen dalších modelů. Později byla k tomuto modelu přidána takzvaná výhoda domácího hřiště, která byla o několik let později statisticky prokázána. (Maher, M. J., 1982, s. 109-118)

Co se týká výhody domácího prostředí, zajímavostí je porovnání výsledků anglické nejvyšší soutěže v sezóně 2016/17 a v sezóně 2020/21. Pro sezónu 2016/17 platilo, že domácí týmy vstřelily průměrně za celou sezónu 1,591892 branky, zatímco hostující pouze 1,183784, což potvrzuje výhodu domácího prostředí. Sezóna 2020/21 se však kvůli pandemii koronaviru hrála bez diváků a průměrně vstřelili domácí 1,335484 branky a hosté 1,332258. Tento fakt dokládá, že hlavním faktorem, který způsobuje výhodu domácího prostředí jsou fanoušci. (Oficiální data Premier League)

Obrázek 2 Počet branek v zápase EPL



Zdroj: https://dashee87.github.io/images/home_away_goals.png (30.3.2023 15:21)

3.2.2 Dixon-Coles (1997), Poissonovo rozdělení

Dixon-Coles prediktivní model je jedním z nejznámějších a nejpoužívanějších prediktivních modelů v oblasti fotbalu. Tento model byl poprvé představen v roce 1997 Richardem Dixonem a Stuartem Coles. Cílem tohoto modelu je predikce výsledků fotbalových zápasů na základě statistických dat.

Model Dixon-Coles je založen na upravení tradičního Poissonova modelu. Hlavní myšlenkou tohoto modelu je zahrnout do výpočtu koeficient penalizace, který zohledňuje to, že některé týmy mají tendenci dávat více nebo méně gólů než průměr. Koeficient penalizace se vypočítává na základě historických dat a určuje, jaký vliv mají týmy na průměrný počet gólů v zápase. (Dixon, M. J., & Coles, S. G., 1997, s. 265-280)

V průběhu let bylo provedeno mnoho studií na ověření účinnosti modelu Dixon-Coles. V jedné z takových studií byl model porovnán s jinými prediktivními modely a ukázalo se, že model Dixon-Coles je účinnější v predikci výsledků fotbalových zápasů. Další studie ukázaly, že model Dixon-Coles je účinný i při predikci gólů v zápasech a může být úspěšně aplikován na trhu kurzových sázek (Dixon, M. J., & Coles, 1997).

Model Dixon-Coles je stále populární a používá se ve světě sportovních sázek velmi často. Mnoho webových stránek a aplikací nabízí možnost využít tento model pro sázení na fotbalové zápasy. Nicméně, jako u všech prediktivních modelů, existuje určitá míra nepřesnosti, a je důležité mít na paměti, že výsledek zápasu může být ovlivněn mnoha faktory, které nelze zahrnout do statistické analýzy.

Tento model tedy vychází z předchozího Maherova modelu. Získává historická data pro všechna mužstva a počítá na základě dvourozměrného Poissonova rozdělení pravděpodobností výstup pro budoucí zápas, na rozdíl od předchozího modelu, ale přidává faktor formy a upravuje četnost zápasů s nižším počtem branek.

Model může být formulován následující matematickou rovnicí:

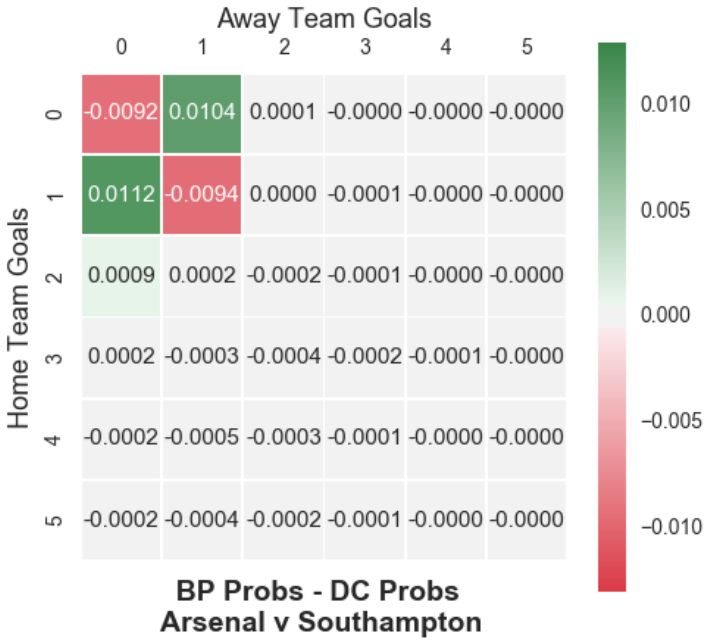
$$P\left(X_{i,j} = x, Y_{j,i} = \mathrm{y}\right) = \frac{e^{-\lambda}\lambda^{x}}{\mathrm{x!}}\frac{e^{-\mu}\mu^{y}}{y!}$$

$$kde \quad \lambda = \alpha_i\beta_j\gamma \qquad \mu = \alpha_j\beta_i$$

V této rovnici i a j odkazuje na domácí a hostující týmy. α a β značí sílu útoku a obrany každého z týmů a konečně γ znamená výhodu domácího prostředí pro domácí tým. α a β jsou jedinečné pro každý z týmů, zatímco γ je pro všechny týmy v dané soutěži stejná.

Dixon a Coles se ve své studii zabývali ziskovou strategií na sázkařském trhu v Anglii, svůj model postavili na datech anglických lig z let 1992-95. Zjistili však, že je v něm podhodnocená četnost zápasů s málo brankami (0:0, 1:0, 0:1 a 1:1), proto je do modelu přidán parametr, který modifikuje předpoklad nezávislosti pro tyto případy, a kromě toho se snažili přidat prvek aktuální formy. Toho docílili přidáním času, zápasy, které týmy hráli v poslední době mají vyšší váhu než staré zápasy. Výstupem modelu je stejně jako u předchozího matice pravděpodobností, jak lze vidět níže. (Dixon, M. J., & Coles, S. G., 1997, s. 265-280)

Obrázek 3 Matice Dixon Coles Model



Zdroj: https://dashee87.github.io/images/bp_dc_diff.png (29.3.2023 15:00)

3.2.3 Rue a Salvesen (2000), Bayesovský lineární model

Fotbalové sázky jsou oblíbenou formou zábavy, ale také obchodu. Vytvoření prediktivního modelu, který by dokázal správně předpovídat výsledky fotbalových zápasů, by mohl znamenat obrovský zisk. Jedním z takových modelů je Rue and Salvesen prediktivní fotbalový model z roku 2000.

Tento model se zaměřuje na predikci počtu gólů v zápase a používá několik faktorů jako jsou týmová síla, výkonnost na domácím hřišti a vzdálenost cesty na zápas. Jednou z klíčových vlastností modelu je jeho schopnost odhadnout nejistotu spojenou s predikcí, což umožňuje sázkařům vytvořit atraktivní sázky.

Model byl testován na více než 2000 zápasech a dosáhl přesnosti předpovědí na úrovni 53,7 %. Přestože tento výsledek není dokonalý, byl model úspěšným příkladem využití statistických metod v predikci fotbalových zápasů. (Rue, H., & Salvesen, Ø., 2000, s. 399-418)

Model, který čerpá z modelu Dixon a Coles, avšak nevyužívá Poissonovo rozdělení, počítá s bayesovským dynamickým zobecněným lineárním modelem. Pomocí tohoto modelu vytvářeli autoři odhad útočné a obranné síly soupeřících týmů, které jsou v jejich modelu analyzovány ve spojitém čase. Protože je důležité v tomto modelu, aby se odhady dělaly pro všechny týmy ve vybrané soutěži současně, používají interaktivní simulační techniku Monte Carlo s využitím Markovova řetězce. (Rue, H., & Salvesen, Ø., 2000, s. 399-418)

3.2.4 Forrest, Simmons (2000), LOGIT model

Dalším z prediktivních modelů, který se stal již nedílnou součástí fotbalových prediktivních modelů je takzvaný logit model, který se stal oblíbeným především díky své schopnosti modelovat nelineární vztahy mezi nezávislými proměnnými.

Forrest a Simmons navrhli logit model pro predikci výsledků fotbalových zápasů. V tomto modelu se využívá statistického přístupu k predikci výsledků na základě dat z minulých zápasů. Jednou z klíčových vlastností modelu je schopnost zohlednit faktory, které mohou ovlivnit výsledek zápasu, jako jsou například zranění hráčů nebo povětrnostní podmínky.

Logit model se skládá z mnoha různých proměnných, které mají vliv na výsledek zápasu. Mezi tyto proměnné patří například úspěšnost týmu v minulosti, forma hráčů, výkonnost hráčů, domácí nebo hostující zápas, zranění hráčů, počet gólů v minulých zápasech a další. (Forrest, D., & Simmons, R., 2000, s.231-250).

Forrest a Simmons provedli analýzu dat z několika sezón a poté vytvořili model pro predikci výsledků zápasů. Své výsledky prezentovali v publikaci "Predicting football results and the efficiency of fixed-odds betting". Jejich model se ukázal jako úspěšný a byl široce přijat v oblasti kurzového sázení na fotbal.

V současné době existuje mnoho různých variant logit modelů pro predikci výsledků fotbalových zápasů, které se liší v zahrnutých proměnných a v použité metodologii. Nicméně

model Forresta a Simmonse z roku 2000 zůstává jedním z nejznámějších a nejúspěšnějších prediktivních modelů v oblasti fotbalu. (Koning, R. H., & van der Berg, G., 2007, s. 263-277).

3.2.5 Goddard (2005), Poissonovská a logistická regrese

Jeden z významných modelů, který byl používán pro predikci výsledků fotbalových zápasů, je model navržený Markem Goddardem v roce 2005. Jeho model spočívá v kombinaci tří různých typů modelů: lineární regrese, modelu založeném na random forest a modelu založeném na neuronových sítích. Kombinace těchto modelů poskytuje vysoce přesné predikce výsledků fotbalových zápasů.

Goddardův model využívá celkovou formu týmů, výkonnost hráčů a další faktory, jako jsou zranění, aby určil pravděpodobnost výhry pro každý tým. Model také zohledňuje faktory jako hřiště, počasí a další okolnosti, které mohou ovlivnit výsledek zápasu. Vyhodnocení modelu bylo provedeno na základě databáze s výsledky anglické Premier League. Goddardův model byl úspěšný při predikci výsledků a mnoho sázkových společností jej využívá pro určení kurzů pro fotbalové zápasy. (Goddard, M., 2005, s.331-340)

3.2.6 Koopman a Lit (2015), Poissonovský model

Jedním z nejnovějších a nejvíce citovaných modelů je model Koopman a Lit, který kombinuje několik metod strojového učení, včetně Bayesian Additive Regression Trees (BART) a Gaussian Process Regression (GPR), aby vytvořil přesný model pro predikci výsledků zápasů.

V roce 2015 přišli autoři Koopman a Lit s velmi efektivním sázkařským algoritmem, využívají dvourozměrný Poissonovský model na analýzu týdenních výsledků. Do svého modelu zahrnuli koeficienty, které představují sílu útoků a obran jednotlivých týmů. Tyto koeficienty se stochasticky mění v čase. Autoři model aplikovali na anglickou nejvyšší soutěž a vypočítávali ziskovost modelu na dvou sezónách této soutěže, kdy se nejlepší model zhodnocoval o 50 % vůči investici.

Stejní autoři o dva roky později vydali publikaci, kde vytváří a porovnávají tři různé složitější modely na základě dat ze šesti nejsledovanějších evropských soutěží. Z jejich zkoumání vyplývá, že nejlépe vychází model, kde je výsledek chápán jako párové pozorování vycházející z dvourozměrného Poissonova rozdělení.

Koopman a Lit modelují několik různých aspektů fotbalu, kterými jsou útok, obrana, forma týmu a další faktory. Tyto faktory jsou poté využity pro predikci výsledků zápasů. Model byl testován na několika velkých evropských ligách s velmi přesnými výsledky. (Koopman, B., & Lit, R., 2015, s.481-502)

Jedním z hlavních důvodů, proč se tento model stal populárním, je jeho schopnost zohlednit náhodné vlivy jako jsou zranění nebo neočekávané výkony týmů. To umožňuje přesnější predikce a lepší rozhodování při sázení.

Nicméně, stejně jako u každého modelu, jsou zde některé omezení. Koopman a Litův model, stejně jako ostatní modely, se spoléhají na kvalitu dat, která jsou k dispozici. Pokud jsou data neúplná nebo nesprávná, může to ovlivnit přesnost predikce. Navíc může být ovlivněna i výkonnost modelu v závislosti na zvolených parametrech a předpokladech.

Celkově lze říci, že model koopman a Lit je velmi účinný nástroj pro predikci výsledků fotbalových zápasů a jeho použití se stává stále běžnějším v rámci fotbalového průmyslu a sázkového trhu.

3.3 Statistické metody

V této kapitole jsou představeny statistické metody, z nichž se v praktické části vycházelo pro sestavení modelu. Kromě použitých statistických a analytických metod, jsou zde také nejčastější statistické metody použité pro modely, ne všechny budou v autorově modelu využity v plné míře.

3.3.1 Příprava dat

Přípravě dat pro autorův prediktivní model bude věnován detailnější prostor v kapitole 4, zde je jen základní představení přípravy dat. Je důležité, aby použitá data byla standardizována. Pro standardizaci dat byla použita standardizace variačním rozpětím, důvodem bylo různé rozpětí a formát proměnných v datovém souboru a odlehlé hodnoty, proto je vhodné proměnné upravit na hodnoty mezi 0 a 1.

Výběr proměnných, představení datového souboru, zdroje dat a samotný popis modelu, včetně konkrétních dat bude popsán v kapitole 4.

3.3.2 Binomické rozložení

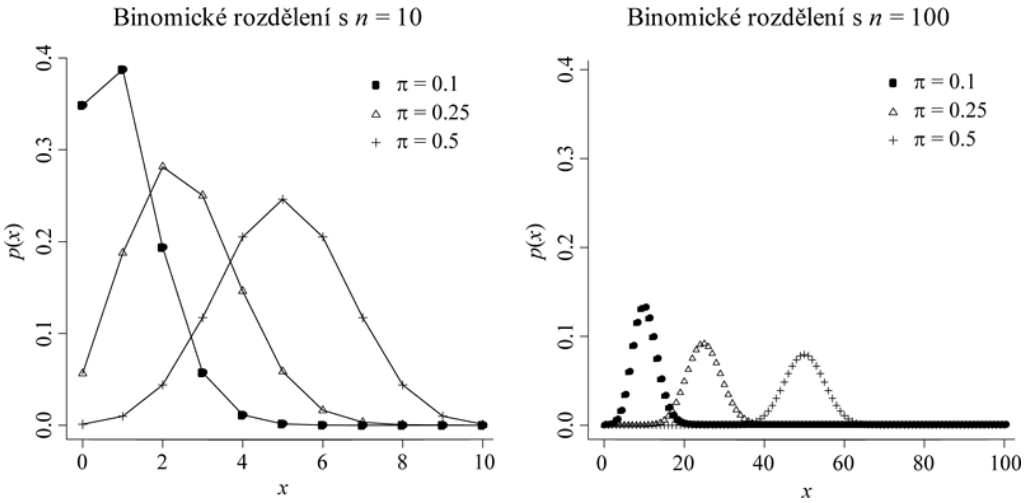
Fotbalový zápas se ze statistického pohledu dá pochopit jako náhodný jev s neznámým výsledkem. Tento zápas na základě předpokladů výše zmíněných autorů můžeme rozdělit na více nezávislých jevů, počet branek domácích je nezávislý vzhledem k počtu branek hostí. Jednotlivé šance jsou samostatné statistické jevy s určitou pravděpodobností na vstřelení branky, v tomto případě můžeme předpokládat, že pro každý samostatný jev je pravděpodobnost stejná. Z tohoto vyplývá, že počet vstřelených branek má binomické rozložení.

Definice binomického rozdělení zní: v posloupnosti n vzájemně nezávislých pokusů pozorujeme počet nastoupení sledovaného jevu, který v každém pokusu nastává s pravděpodobností p. Náhodná veličina X udávající, kolikrát námi sledovaný jev nastal má binomické rozdělení pravděpodobnosti s parametry n a p, kde p∈ (0,1), píšeme: X∼Bi(n,p).

Pravděpodobnostní funkce je tvaru:

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{pro } x = 0, 1, \ldots, n \\ 0 & \text{jinak} \end{cases}$$

Obrázek 4 Příklad binomického rozdělení



Zdroj:https://portal.matematickabiologie.cz/res/image/biostatistika/02-06.png (5.3.2023 10:00)

Proti využití binomického rozdělení může být práce s xG faktorem. Tato proměnná počítá a přiřazuje na základě různých faktorů každé příležitosti určitou pravděpodobnostní hodnotu, jaká byla v konkrétní šanci pravděpodobnost, že padne branka. Vychází z postavení útočícího hráče v době zakončení, vzdálenosti, bránících hráčů apod. Díky této pokročilé fotbalové statistice však nelze tvrdit základní předpoklad binomického rozdělení, a to že pro každou příležitost máme stejnou pravděpodobnost.

3.3.3 Logistická regrese

Logistická regrese je statistická metoda, která slouží k predikci pravděpodobnosti výskytu kategorie nezávislé proměnné na základě hodnot několika vysvětlujících proměnných. Tento typ regrese je často používán v oblasti datové analýzy a strojového učení. (Agresti, A., 2018)

Využití logistické regrese může být například při predikci pravděpodobnosti vzniku určité nemoci na základě věku pacienta, jeho BMI nebo kouření. Logistická regrese také může být využita v oblasti marketingového výzkumu k predikci pravděpodobnosti nákupu určitého produktu na základě předchozích nákupů nebo demografických údajů zákazníka. (King, G., & Zeng, L., 2001, s.137-163)

Pro vytvoření logistické regrese jsou potřebné statistické znalosti a znalosti programování. Analytické nástroje, jako je například Python, mohou být použity k vytvoření a analýze modelů logistické regrese. Výsledky logistické regrese mohou být prezentovány v grafické formě, což usnadňuje interpretaci výsledků.

Mezi výhody použití logistické regrese patří snadná interpretace výsledků, schopnost pracovat s binárními proměnnými, a schopnost pracovat s velkými soubory dat. Na druhé straně, mezi nevýhody patří citlivost na nesprávné specifikace modelu, potřeba správně specifikovat vztahy mezi nezávislými proměnnými a závislou proměnnou a nutnost interpretace a vysvětlení výsledků. (Kuhn, M., & Johnson, K., 2013)

Obecný tvar logistické regrese vypadá takto:

$$\ln \frac{\pi}{1-\pi} = x^T \beta$$

Kde $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ značí vektor regresních parametrů a π je pravděpodobnost nastolení náhodného jevu. V rovnici je vidět, že hodnota lineárního prediktoru odpovídá hodnotě transformované pravděpodobnosti, která se označuje jako logit.

Pokud by logit nebyl do modelu zařazen a byla použita identita, pak by pravděpodobnost mohla nabývat i hodnoty mimo rozmezí 0-1.

3.3.4 Naivní bayesovský klasifikátor

Naivní Bayesovský klasifikátor (NBC) je statistický klasifikační algoritmus, který se často používá pro kategorizaci textových dat, jako je například klasifikace spamu v e-mailech nebo kategorizace zpráv do tematických kategorií. Jeho základem je Bayesova teorie pravděpodobnosti a předpoklad nezávislosti mezi rysy, což znamená, že přítomnost jednoho rysu v textu neovlivňuje pravděpodobnost výskytu jiného rysu. (Manning, C. D., Raghavan, P., & Schütze, H., 2008)

NBC se skládá z trénovací fáze a testovací fáze. V trénovací fázi se na základě poskytnutých dat vypočítají pravděpodobnosti jednotlivých rysů v každé třídě. V testovací

fázi se pak pro nový vstupní text spočítají pravděpodobnosti jeho příslušnosti k jednotlivým třídám, na základě kterých se určí jeho konečné zařazení.

NBC má několik výhod, jako je rychlost a efektivita pro klasifikaci velkého množství dat, a to i při použití většího počtu rysů. Mezi jeho nevýhody patří předpoklad nezávislosti rysů, který se v praxi často neplní, a náchylnost k přetrénování. (Domingos, P., & Pazzani, M., 1997, s.103-130)

Použití NBC se však neomezuje pouze na klasifikaci textových dat. Lze jej uplatnit i v dalších oblastech, jako je biomedicína, obrazové zpracování a rozpoznávání řeči.

Metoda v prediktivních modelech velmi často využívána vychází z pravděpodobnostního modelu a Bayesova teorému. Základem naivního bayesovského klasifikátoru je, že každému pozorování x jsou přiřazeny posteriorní pravděpodobnosti P (y|x). Pozorování x je pak zařazeno do kategorie y s nejvyšší posteriorní pravděpodobností. Níže následuje podmínka zařazení pozorování x do kategorie y dle Bayesova teorému:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

P (y) je rovno relativní četnosti kategorie y na všech pozorováních.

P (x) je pro každou kategorii y stejná, takže není pro kalkulaci důležitá.

P (x|y) je jedinou hodnotou, která musí být odhadnuta a je neznámá. Pokud ale datový soubor čítá více proměnných, musí se při odhadu najít sdružené pravděpodobnosti jednotlivých proměnných, tudíž při vyšším počtu proměnných počet sdružených pravděpodobností exponenciálně roste.

3.3.5 Podpůrné vektory

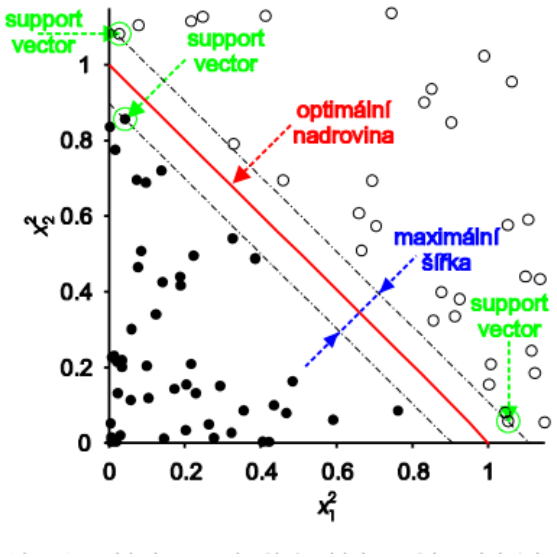
Metoda podpůrných vektorů tvoří základ jádrových algoritmů. Základním principem je převod daného vstupního prostoru do jiného, který má více dimenzí, a ve kterém je již

možné oddělit třídy lineárně. Tato metoda se tedy snaží nalézt lineární klasifikátor, který rozdělí data do dvou kategorií.

Metodu podpůrných vektorů lze rozdělit do tří kategorií:

3.3.5.1 Lineárně rozdělitelné kategorie

Obrázek 5 Lineárně rozdělitelné vektory

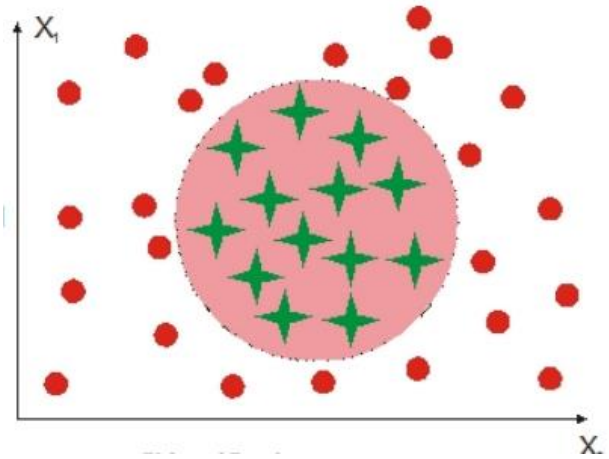


Zdroj: Žižka, J.: Support vector machines (SVM). 2005-10-21 [cit. 2011-01-09]. URL http://is.muni.cz/el/1433/podzim2006/PA034/09_SVM.pdf

V této kategorii je předpokladem, že existuje dělící hranice kategorií, musí se stanovit podmínka pro nejrobustnější řešení, tzn. určit vzdálenost nejbližších objektů a následně najít takové objekty, jejichž součet vzdáleností od hranice je největší.

3.3.5.2 Lineárně nerozdělitelné kategorie

Obrázek 6 Lineárně nerozdělitelné kategorie

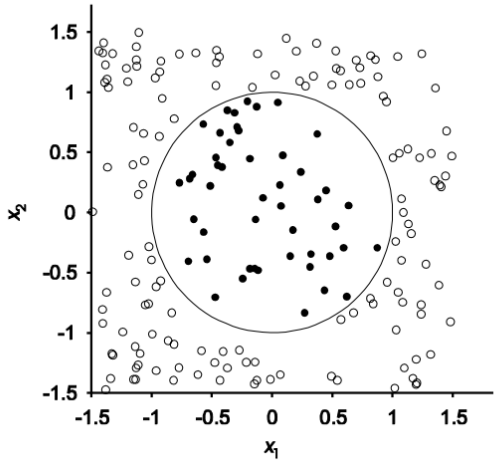


Zdroj: stejný jako u obrázku 5.

Pro tuto kategorii, která je stejná jako první lineární verze, nelze uplatnit podmínky pro první kategorii. Proto se zde vytváří tzv. relaxační proměnné, které vyjadřují porušení daných podmínek konkrétním objektem.

3.3.5.3 Nelineární verze

Obrázek 7 Nelineární verze



Zdroj: stejný jako u obrázku 5.

U nelineární verze musí dojít k zvýšení počtu dimenzí, aby bylo možné objekty separovat. V tomto případě, na obrázku nahoře, stačí přidat třetí rozměr a přítomné objekty do něj transformovat, díky čemuž jsou následně kategorie separovatelné. Transformace se provádí pomocí jader (kernel trick).

3.3.6 Náhodné lesy

Random Forest (náhodný les) je algoritmus strojového učení používaný pro klasifikaci, regresi a další prediktivní úkoly. Tento algoritmus kombinuje několik stromů rozhodování, které vytvářejí prediktivní modely. Random Forest se stal velmi populárním díky své schopnosti pracovat s velkými datovými sady a vykazovat dobré výsledky bez přílišného úsilí při ladění parametrů.

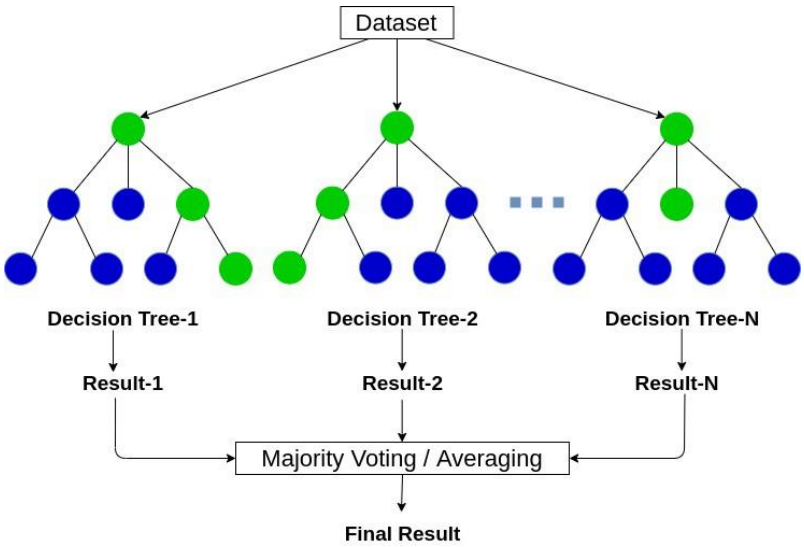
Tento algoritmus je založen na rozhodovacích stromech, které mají výhodu v jednoduchosti interpretace. Random Forest vytváří větší množství těchto stromů a následně hlasováním určuje výsledek. Každý strom je vytvořen náhodně vybranou podmnožinou trénovacích dat a náhodně vybranými příznaky. To umožňuje Random Forestu být robustnější vůči šumu v datech a přenést větší množství informací z trénovacích dat na testovací data. (Breiman, L., 2001, s.5-32).

Random Forest má širokou škálu aplikací, například v medicíně, biologii, financích, marketingu a dalších oblastech. Jedním z příkladů použití Random Forest je predikce onemocnění srdce na základě různých klinických faktorů, jako jsou věk, pohlaví, hladina cholesterolu a krevní tlak. Dalším příkladem může být klasifikace spamových e-mailů na základě různých příznaků, jako jsou klíčová slova, délka e-mailu a jeho obsahu. (Liu, J., & Kumar, S., 2011, s. 573-578).

Existuje mnoho nástrojů, které umožňují implementaci Random Forest v různých programovacích jazycích, jako jsou R, Python nebo MATLAB. Tyto nástroje poskytují uživatelům možnost vytvářet a optimalizovat své prediktivní modely, a tak získávat tak větší přesnost v predikcích.

Níže je příklad schéma náhodných lesů:

Obrázek 8 Schéma Random Forest



Zdroj:https://cdn.analyticsvidhya.com/wp-content/uploads/2020/02/rfc_vs_dt1.png (31.3.2023 17:30)

3.3.7 Neuronová síť

Neuronové sítě se staly v posledních letech velmi populárním nástrojem pro predikci v různých oblastech, včetně sportovních her, jako je fotbal. Tyto sítě jsou schopné

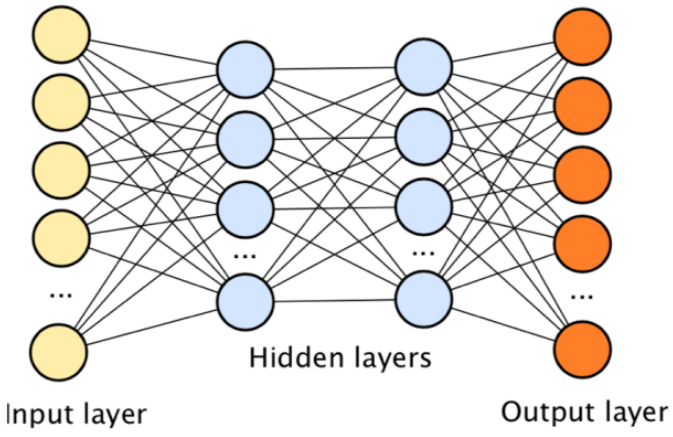
analyzovat obrovské množství dat a naučit se komplexní vztahy mezi různými faktory, což umožňuje vytvoření přesných prediktivních modelů.

V kontextu fotbalu lze neuronové sítě využít k predikci výsledků zápasů, pravděpodobnosti vstřelení gólu, průměrného počtu gólů v zápase a mnoho dalších ukazatelů. Tyto predikce mohou být užitečné pro sázení na zápasy, předpovědi výsledků sezóny nebo pro trénink týmu.

Při použití neuronových sítí v prediktivních modelech pro fotbal je klíčové mít k dispozici kvalitní data, včetně historických výsledků zápasů, statistik hráčů a dalších faktorů ovlivňujících výsledek zápasu. Dále je třeba provádět důkladné trénování sítě, aby byla schopna odhalit komplexní vztahy mezi těmito faktory a výsledky zápasů. (Tsang, S. H. K., et al., 2019, s. 1812-1819)

Použití neuronových sítí v predikci fotbalových výsledků se v posledních letech stalo velmi populárním a mnoho výzkumných prací se zaměřuje na vylepšení těchto modelů. Například výzkum publikovaný v roce 2019 ukázal, že použití hlubokých neuronových sítí může vést k vylepšení přesnosti predikce oproti tradičním metodám.

Obrázek 9 Schéma neuronové sítě



Zdroj: https://labels.ff.cuni.cz/neuronovesite (31.3.2023 17:41)

Výše je uveden příklad vícevrstvé neuronové sítě. Ta obsahuje vstupní vrstvu, skrytou vrstvu a výstupní vrstvu. Jak již bylo zmíněno, do vstupní neuronové vrstvy vstupují data z datového souboru, ty se přes tuto vrstvu rozhodí do skryté vrstvy, kdy je každému vstupu přiřazena určitá váha, ve skryté vrstvě z každého neuronu vyjde jen jeden výstup, který je zpracován ve výsledné vrstvě do jednoho výstupu. Při trénování tohoto modelu se upravují váhy jednotlivých vstupů a podmínky v jednotlivých neuronech. Tento typ modelu je velmi často využíván například pro předpověď burzovních indexů.

3.4 Teorie kurzového sázení

Posledním bodem teoretické části je seznámení s problematikou kurzového sázení, představení trhu kurzového sázení v České republice, vysvětlení vybraných typů sázek a v poslední řadě představení roli bookmakera. Zajímavá je i otázka arbitráže na trhu kurzového sázení.

Kurzové sázení, často nazývané také sázení na sport, je forma hazardní hry, která spočívá v předpovídání výsledků sportovních událostí a následném vsazení peněz na tyto výsledky. Sázkoví operátoři, kterými jsou sázkové kanceláře, nabízejí různé kurzy na různé výsledky sportovních událostí, a hráči si mohou vsadit na jednotlivé výsledky podle svých předpokladů a odhadů. Výhra nebo prohra závisí na správnosti hráčových předpovědí.

Jedním z klíčových principů kurzového sázení je práce s kurzy. Když sázková kancelář nabízí kurz na konkrétní výsledek, tento kurz vychází z pravděpodobnosti, že se tento výsledek skutečně stane. Pokud je vysoká pravděpodobnost, že se tento výsledek stane, pak bude kurz nižší. Naopak, pokud je nízká pravděpodobnost, že se tento výsledek stane, pak bude kurz vyšší. Tímto způsobem sázkové kanceláře pracují s pravděpodobnostmi a kurzy, aby maximalizovaly svůj zisk. (Hing, N., Gainsbury, S., & Blaszczynski, A., 2014)

Dalším klíčovým principem je analýza dat. Sázkové kanceláře a hráči se snaží využít různé statistické metody a analýzy dat k předpovídání výsledků sportovních událostí a určení hodnotných sázek. Toto využití datových analýz se stává čím dál tím důležitějším pro sázkové operátory a hráče.

Kurzové sázení je formou hazardní hry, která je založena na předpovídání výsledků sportovních událostí a vsazení peněz na tyto výsledky. Klíčovými principy tohoto sázení jsou práce s kurzy a analýza dat. Kromě toho je také důležité věnovat pozornost rizikům, která jsou spojena s touto formou hry, a snažit se minimalizovat negativní dopady.

3.4.1 Trh kurzového sázení v ČR

Trh kurzových sázek v ČR se v posledních letech rozvíjí neuvěřitelně rychle. Sázkařské kanceláře se objevují na každém rohu a sázkařské weby se staly běžnou součástí internetového prostředí. Přestože trh kurzových sázek má své kořeny v minulosti, byl v ČR zákaz sázení až do roku 1990, kdy byl zrušen. Od té doby se trh sázkových kanceláří rozrůstal až do současné podoby.

Podle statistiky, kterou zveřejnila Asociace sázkových společností, se v roce 2020 celková částka sázek v ČR vyšplhala na více než 228 miliard korun. Z toho téměř 70 % sázek bylo podle Asociace realizováno prostřednictvím internetu. V současné době existuje na českém trhu několik velkých hráčů, mezi které patří Tipsport, Fortuna a Betano. ("Sázkové kanceláře v roce 2020," Tip na kurz, 30. prosince 2020. [Online]. Dostupné z: https://tipnakurz.cz/sazkove-kancelare-v-roce-2020.)

Sázkaři mají možnost vsadit si na různé sportovní události, jako jsou fotbalové, hokejové, tenisové nebo basketbalové zápasy. Kromě toho je možné vsadit si i na jiné druhy událostí, jako jsou politické volby nebo výsledky reality show. Sázkové kanceláře využívají různých modelů a strategií pro stanovení kurzů na jednotlivé sázky. Jedním z nejpoužívanějších modelů je tzv. Kellyho kritérium, které určuje optimální velikost sázky na základě pravděpodobnosti výhry.

Sázkařský trh je v ČR regulován Ministerstvem financí, které uděluje licenci sázkovým společnostem. Sázkařské společnosti musí splňovat řadu požadavků a podléhají pravidelnému dohledu. Kromě toho existuje i Národní protidrogová centrála, která má na starosti kontrolu sázkových kanceláří a boj proti nelegálnímu sázení.

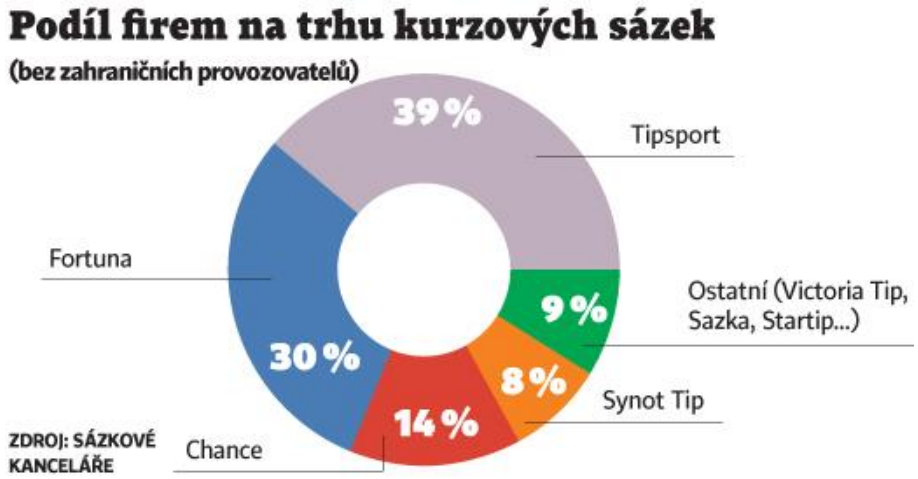
Někteří kritici se obávají, že trh sázkových kanceláří může mít negativní dopad na společnost, zejména na mladé lidi. Existují také obavy z praní špinavých peněz a podvodu. Sázkařské společnosti však argumentují tím, že regulace a dohled jsou dostatečné

V závěru lze shrnout, že trh kurzových sázek v ČR je v současné době velmi dynamický a rychle se rozvíjí. Sázkové společnosti nabízejí širokou škálu možností sázek na různé sporty a události, a to jak prostřednictvím kamenných poboček, tak i online prostřednictvím webových stránek a mobilních aplikací. Zákaznický servis je vysoko profesionalizovaný a sázkové společnosti se snaží své zákazníky získat a udržet pomocí různých bonusů a promoakcí. ("Sázkové kanceláře: Jak fungují kurzové sázky?" Kurzy.cz, 8. srpna 2018. [Online]. Dostupné z: https://www.kurzy.cz/sazky-a-sazeni)

Je však důležité si uvědomit, že sázkařství přináší riziko ztráty peněz a může být návykové. Proto je vhodné pečlivě zvažovat každou sázku a používat různé strategie, jako například kontrolu bankrollu, sázení s rozumem a sázky pouze na sporty, ve kterých má sázející dostatečné znalosti a zkušenosti.

Nejstarší sázkovou kanceláří v České republice ze stále aktivních je Fortuna, která byla založena v roce 1990, jejich podíl na trhu z roku 2016 můžeme vidět níže:

Obrázek 10 Podíl SK v ČR



Zdroj: Tipsport

3.4.2 Typy sázek

Jak již bylo zmíněno výše, v dnešní době existují sázky téměř na cokoli. Níže jsou představeny základní typy sázek, které budou stěžejní v budoucích kapitolách pro valuaci autorova modelu.

Nejprve ale, jak samotné kurzové sázení funguje. Zjednodušeně lze říci, že kurzové sázení je hra, kterou pořádá sázková kancelář, ta za pomocí bookmakera vypisuje kurzy na sázkové příležitosti. Sázející si vybere událost, na kterou chce vsadit a tipne si její výsledek, který podloží vkladem (minimální a maximální vklad se liší dle sázkové kanceláře). V případě, že se mu podaří správně uhodnout výsledek dané události, sázející vyhraje částku, která je vytvořena součinem jeho vkladu a kurzu, který je vypsaný na vsazený výsledek. V případě, že se sázkaři výsledek nepodaří uhodnout, celou část svého vkladu ztrácí. Pokud sázející sází na více sázkových příležitostí, tak se jejich kurzy násobí.

3.4.2.1 Sázka na vítěze

Nejzákladnější typ sázky. Jedná se o sázku na vítěze daného zápasu. Dle druhu sportu sázková kancelář vypisuje kurzy na výhru domácích, remízu (pokud je možná) a výhru hostí. Zároveň se vypisuje i takzvaný dvojtip – neprohra domácích/ neprohra hostí. Typické označení těchto sázek je formát 1 – 0 – 2 – 10 – 02, kde 1 znamená výhru domácích, 0 remízu a 2 vítězství hostujícího družstva.

3.4.2.2 Handicap

Sázka s handicapem, znamená, že některé mužstva vyhraje vyšším rozdílem, tento druh sázky se často uplatňuje u velkých favoritů, u kterých se očekává vyšší výhra a pro sázející tedy není kurz na obyčejnou výhru příliš lákavý, protože bývá velmi nízký. Při sázce s handicapem, např. že favorit vyhraje alespoň o dvě branky, bývá kurz vyšší, ale zároveň samozřejmě roste i míra rizika.

3.4.2.3 Počet branek

Sázky na počet branek neboli Over/Under, znamenají, že sázející tipuje počet branek v daném utkání, nebo daného týmu. Sázkové příležitosti bývají vypisovány jako: Počet branek v utkání 1,5 +/-. To znamená, že pokud sázející vsadí na (Over) 1,5+, musí v utkání padnout alespoň dvě branky, aby byl jeho tiket výherní.

3.4.2.4 Ostatní (události v zápase)

Tato kategorie čítá nepřeberné množství sázek na události v zápase, asi nejpopulárnější jsou sázky na počet rohů, žlutých karet, střelce branek. Avšak v nabídkách sázkových kanceláří jsou i počty faulů, kdo bude mít zahajovací výkop a podobné sázky.

3.4.2.5 Live sázky

Samostatnou a velmi populární kategorií jsou pak Live sázky neboli sázky na zápas v reálném čase. Zde je výhoda, že sázející může přímo reagovat na vývoj daného zápasu, na který ale reaguje i bookmaker a kurzy průběžně mění.

3.4.3 Role bookmakera

Role bookmakera při kurzovém sázení je zásadní pro provozování sázkového trhu. Bookmaker je subjekt, který stanovuje kurzy a nabízí sázky na různé události, jako jsou sportovní zápasy, politické volby nebo události z oblasti zábavy.

Bookmaker zpravidla vytváří nabídku sázek na základě svého odhadu pravděpodobnosti, s jakou se daná událost stane. Tento odhad pak převede na kurz, který představuje poměr mezi výší sázky a výší výhry. Například kurz 2.0 znamená, že pokud vsadíte 100 Kč na danou událost a ta se stane, vyhrajete 200 Kč.

Bookmaker si musí být vědom rizika, které s sebou kurzové sázení nese. Pokud totiž vsadí příliš mnoho lidí na jednu stranu sázky, může se stát, že bude muset vyplatit výhry větší než výše přijatých sázek. Proto se bookmaker snaží vytvořit takovou nabídku sázek, která rozloží sázky na obě strany rovnoměrně.

Bookmaker může také využívat různé strategie k minimalizaci rizik. Například může vytvářet nabídku sázek tak, aby se většina sázek konala na událostech s nízkým výskytem, což snižuje riziko vysokých výher. Dále může využívat techniky jako je hedging, tedy uzavírání protisměrných sázek na stejnou událost, čímž minimalizuje riziko výrazných ztrát. (DICKINSON, Peter, 2012, s. 112-139)

Bookmaker rovněž podléhá regulaci ze strany státních orgánů. Například v České republice je provozování sázkových kanceláří regulováno zákonem o hazardních hrách. Bookmaker musí splnit přísná kritéria, aby mohl získat licenci pro provozování sázkové kanceláře a je povinen dodržovat stanovená pravidla, například ohledně maximální výše sázek a minimální výše výher.

3.4.4 Arbitráž

Zajímavá je otázka arbitráže, i k té může na trhu kurzového sázení dojít, v případě, že různé sázkové kanceláře nabízejí jinou výši kurzu na stejnou událost. V tom případě si sázející může vybrat výhodnější kurz u každé z nich a u jedné vsadit na výhru a u druhé na prohru. Bohužel pro sázející je teorie jednodušší než samotná praxe, sázkové kanceláře dokáží na výkyvy kurzů na trhu reagovat, takže časové okno, než se kurzy znovu přiblíží, je relativně malé, druhým faktorem jsou marže sázkových kanceláří, kvůli kterým je profit velmi nízký.

Marže sázkové kanceláře je marže, díky níž získává sázková kancelář další profit. Jedná se o rozdíl mezi pravděpodobností a výší kurzu od sázkové kanceláře. Marže se napříč sázkovými kancelářemi liší, záleží na politice každé z nich, jestli chtějí na úkor vyšší marže riskovat ztrátu klientů. Obecně platí, že větší sázkové kanceláře s vyšším objemem sázek si mohou dovolit mít marži nižší.

Nejjednodušším příkladem marže je sázková příležitost se dvěma možnostmi, které mají stejnou pravděpodobnost, např. který tým bude mít úvodní výkop (ten se určuje hodem mincí) v tomto případě by měli být spravedlivé kurzy 2,0 – 2,0, většinou ale bývají v sázkových kancelářích ve výši 1,85 – 1,85 a tento rozdíl vytváří onu marži. V tomto případě je dokonce 8,1 %, standardně bývá mezi 3-6 %.

4 Vlastní práce

4.1 Data – příprava dat

V rámci praktické části je důležitý první krok, najít vhodný zdroj dat a následně data zpracovat a standardizovat do odpovídajícího datasetu. V této kapitole je vysvětleno právě zpracování a standardizace dat. Data jsou k dispozici ve formátu CSV a budou dále analyzována pomocí programovacího jazyka Python 3.0 (dále jen Python).

Prvním krokem bude načtení dat z CSV souboru pomocí knihovny Pandas. Poté bude prováděno čištění dat, například odstranění duplicitních záznamů a chybějících hodnot. Dalším krokem je příprava dat pro modelování. To zahrnuje kódování kategoriálních proměnných, normalizaci numerických dat a přidání nových proměnných, například z historických dat o výkonnosti týmů.

Poslední fází práce s daty je standardizování, aby byla všechna ve stejném formátu. Toto zahrnuje například úpravu jednotek měření a normalizaci hodnot, aby byla všechna data v měřítku 0 až 1. Tento krok je důležitý pro správnou funkci prediktivních modelů, které budou následně vytvořeny. Výsledkem této kapitoly bude připravená a standardizovaná datová sada, kterou lze použít pro prediktivní modelování.

4.1.1 Zdroj dat

V dnešní době existuje nespočet fotbalových databází. Samotná data se dají vytěžit i z oficiálních stránek daných evropských soutěží, kde ovšem nebývají ve vhodných formátech. Ideální je tedy využít jednu z níže zmíněných ověřených fotbalových databází, které čítají vybrané fotbalové soutěže a ročníky. Všechna data se následně dají legálně stáhnout ve formátu CSV, se kterým se pak velmi dobře pracuje a jsou tak vhodná pro prediktivní modelování. Mezi nejznámější tyto databáze patří:

- Open Football: Jedná se o otevřenou databázi s fotbalovými výsledky z celého světa. Obsahuje výsledky od roku 1872 a je neustále aktualizována. Dostupná je zdarma pro osobní nekomerční použití.

- Football-Data.co.uk: Databáze s fotbalovými výsledky z Anglie, Skotska, Německa, Španělska, Itálie a Francie,... Obsahuje detailní statistiky jako jsou střely na branku, vyloučení, karty atd.

- Kaggle: Kaggle je platforma pro datovou vědu a strojové učení a obsahuje mnoho datasetů s fotbalovými výsledky a statistikami.

- FIFA: FIFA má svou vlastní databázi s fotbalovými výsledky a statistikami z celého světa.

- Opta: Opta je profesionální společnost zabývající se sběrem dat v reálném čase během fotbalových zápasů. Jejich databáze obsahuje velké množství statistik jako jsou například driblingy, přihrávky, nahrávky a mnoho dalších.

Nakonec byla vybrána databáze z https://football-data.co.uk/ , protože představovala nejvíce uživatelsky přívětivou databázi, jejíž data jsou z oficiálních databází evropských soutěží. Následné databáze ve formátu CSV byly rozděleny na trénovací a testovací sadu.

4.2 Model 1 – Poissonovo rozdělení – MS Excel

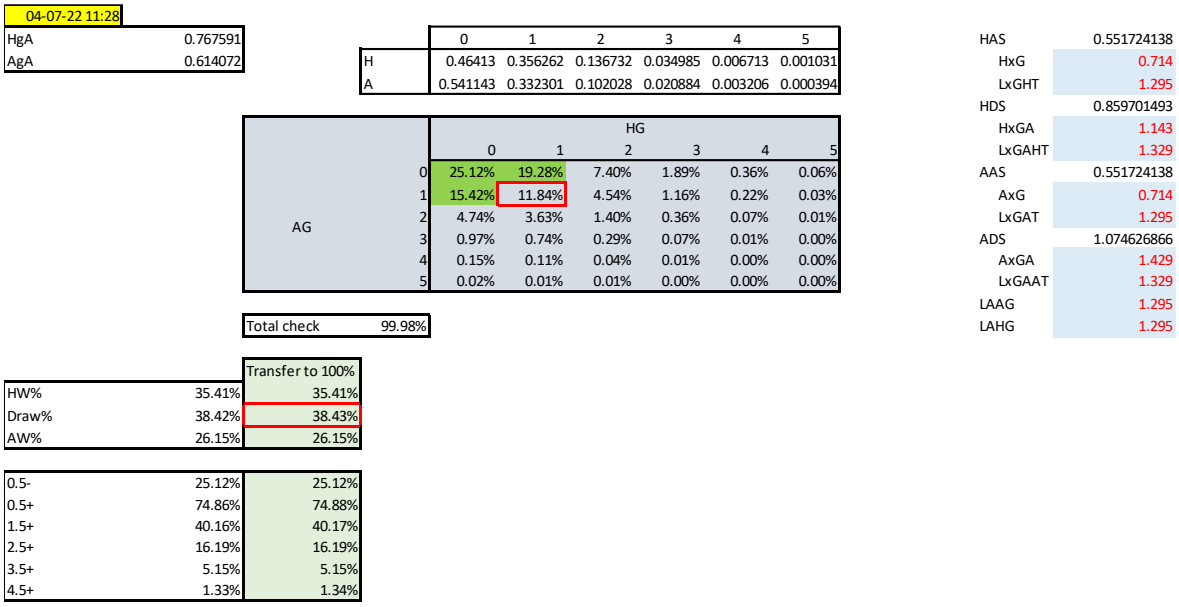
Poissonovo rozdělení je jedním ze základních matematických modelů, který se nejčastěji používá při predikci výsledků fotbalových zápasů. Model je založen na statistickém rozdělení, které určuje pravděpodobnost výskytu určitého počtu událostí v určitém časovém intervalu. V případě fotbalových zápasů se tento model používá k predikci počtu gólů, které budou vstřeleny v daném zápase.

Konkrétní prediktivní modely založené na Poissonově rozdělení byly zmíněny v teoretické části. V praktické části byly sestaveny takovéto modely dva. Nejprve bude

vysvětlen model poissonova rozdělení v programu MS Excel, ve kterém byly využity jednoduché MS Excel vzorce a následné výsledky byly interpretovány v matici pravděpodobností. Pro každý tým byl vypočítán průměrný počet vstřelených a inkasovaných gólů na základě historických dat z minulých zápasů. Poté byly použity tyto průměrné hodnoty jako parametry pro poissonovo rozdělení a z nich vypočítána pravděpodobnost výskytu určitého počtu gólů v zápase.

Tento jednoduchý model byl dále vylepšen faktorem domácího prostředí, respektive faktorem síly týmů doma/venku.

Obrázek 11 Poissonovo rozdělení Excel



04-07-22 11:28	
HgA	0.767591
AgA	0.614072

	0	1	2	3	4	5
H	0.46413	0.356262	0.136732	0.034985	0.006713	0.001031
A	0.541143	0.332301	0.102028	0.020884	0.003206	0.000394

AG \ HG	0	1	2	3	4	5
0	25.12%	19.28%	7.40%	1.89%	0.36%	0.06%
1	15.42%	11.84%	4.54%	1.16%	0.22%	0.03%
2	4.74%	3.63%	1.40%	0.36%	0.07%	0.01%
3	0.97%	0.74%	0.29%	0.07%	0.01%	0.00%
4	0.15%	0.11%	0.04%	0.01%	0.00%	0.00%
5	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%

Total check	99.98%

		Transfer to 100%
HW%	35.41%	35.41%
Draw%	38.42%	38.43%
AW%	26.15%	26.15%

0.5-	25.12%	25.12%
0.5+	74.86%	74.88%
1.5+	40.16%	40.17%
2.5+	16.19%	16.19%
3.5+	5.15%	5.15%
4.5+	1.33%	1.34%

HAS	0.551724138
HxG	0.714
LxGHT	1.295
HDS	0.859701493
HxGA	1.143
LxGAHT	1.329
AAS	0.551724138
AxG	0.714
LxGAT	1.295
ADS	1.074626866
AxGA	1.429
LxGAAT	1.329
LAAG	1.295
LAHG	1.295

Zdroj: vlastní zpracování

Na obrázku vidíme samotný model v aplikaci MS Excel. Vysvětlení parametrů je následující:

Tabulka vpravo značí průměrné počty branek, zkratky znamenají toto:

HAS – síla útoku domácího družstva (HxG / LxGHT)

HxG – průměrný počet vstřelených branek domácího týmu na zápas

LxGHT – průměrný počet vstřelených branek na zápas od všech týmů (ligový průměr domácích týmů)

HDS – síla obrany domácího družstva (HxGA / LxGAHT)

HxGA – průměrný počet obdržených branek domácího týmu na zápas

LxGAHT – průměrný počet obdržených branek na zápas od všech týmů (ligový průměr domácích týmů)

AAS – síla útoku venkovního družstva (AxG / LxGAT)

AxG – průměrný počet vstřelených branek hostujícího týmu na zápas

LxGAT – průměrný počet vstřelených branek na zápas od všech týmů (ligový průměr hostujících týmů)

ADS – síla obrany venkovního družstva (AxGA / LxGAAT)

AxGA – průměrný počet obdržených branek hostujícího týmu na zápas

LxGAAT – průměrný počet obdržených branek na zápas od všech týmů (ligový průměr hostujících týmů)

LAAG – ligový průměr vstřelených branek hostujících týmů na zápas

LAHG – ligový průměr vstřelených branek domácích týmů na zápas

HgA – koeficient pro domácí tým (HAS*ADS*LAHG)

*Síla útoku domácího týmu*síla obrany hostujícího týmu*ligový průměr vstřelených branek domácích týmů*

AgA – koeficient pro hostující tým (HDS*AAS*LAAG)

*Síla útoku hostujícího týmu*síla obrany domácího týmu*ligový průměr vstřelených branek hostujících týmů*

Z následných koeficientů HgA a AgA je vytvořeno Poissonovo rozdělení a matice pravděpodobností, kterou můžeme vidět uprostřed obrázku 11. V matici máme na horní ose počty branek domácího týmu 1-5 a na boční levé ose máme branky hostujícího týmu 1-5. Průsečík hodnot značí pravděpodobnost výsledku. Například výhra domácích 2:1 má pravděpodobnost 4,54 %. V tabulce jsou zeleně znázorněny nejvíc pravděpodobné výsledky, v tomto konkrétním případě se jedná o výsledky 0:0, 1:0 a 0:1. Následně je červeným ohraničením označen konečný výsledek zápasu – v tomto případě 1:1, výsledek, který měl dle modelu pravděpodobnost 11,84%.

Je důležité zmínit, že celkový součet procent v matici pravděpodobností není 100 %, ale je mu velmi blízko (hodnotu součtu pravděpodobností vidíme v řádku TOTAL Check). V tomto konkrétním případě je součet pravděpodobností 98,43 %. Je to z toho důvodu, že v matici vidíme jen výsledky v rozmezí 0-5 branek obou týmů a vždy existuje velmi malá pravděpodobnost, že padne branek více. Vše je velmi dobře vidět na obrázku 11.

Poslední důležitá tabulka na obrázku 11 je vlevo dole a ukazuje HW% (pravděpodobnost výhry domácího družstva), Draw% (pravděpodobnost remízy) a AW% (pravděpodobnost výhry hostujícího družstva).

Tento MS Excel model byl testován na množině pouze 30 zápasů, čínské nejvyšší fotbalové soutěže a švédské nejvyšší fotbalové soutěže, jejíž data vycházela ze sportovního statistického serveru https://www.livesport.cz/ . Pro svou manuální náročnost nemohlo být zápasů více, protože si každý žádal určité manuální úpravy. Cílem bylo zjistit / odhadnout vítěze zápasu a počet branek v zápasu.

Model uvedl nejvyšší pravděpodobnost u týmu (remízy), který nakonec vyhrál v 73,33 % případů, neboli 22 zápasů z 30 odhadl model správně. Co se týká odhadu počtu branek na zápas, byl to původní cíl využití tohoto modelu, ale vzhledem k již zmíněné náročnosti a specifičnosti daného modelu v rozhraní MS Excel, se touto problematikou bude zabývat až model v programu Python.

Ačkoli je procento úspěšnosti velmi vysoké, hlavním důvodem bude pravděpodobně malá množina zápasů, který model otestoval, s více zápasy by pravděpodobně klesala i úspěšnost na hranici okolo 60 %.

4.3 Model 1 – Poissonovo rozdělení - Python

Stejný druh modelu jako byl použit v MS Excel, akorát z důvodu automatizace, úspory času a efektivity byl napsán v programovacím jazyce Python. Model je tedy založen na Poissonově distribuci a snaží se predikovat výsledek fotbalového zápasu a počet branek

v daném zápase. Model počítá pravděpodobnost, že každý tým dá v zápase určitý počet gólů. Tato pravděpodobnost je založena na historických datech o tom, kolik gólů daný tým dává a dostává v průměru v jednom zápase.

Vstupem modelu jsou historická data o výsledcích zápasů a počtu gólů, které daly jednotlivé týmy. Tyto informace jsou uloženy v CSV souboru a jsou načteny do programu pomocí knihovny pandas.

Poté je provedena příprava dat, která zahrnuje vytvoření nových sloupců, jako jsou počet zápasů, počet vítězství, remíz a proher, průměrný počet gólů na zápas a další.

Model využívá algoritmus Maximum Likelihood Estimate (MLE) pro určení parametrů Poissonovy distribuce pro každý tým. Tyto parametry jsou následně použity pro výpočet pravděpodobnosti, že každý tým dá v zápase určitý počet gólů. Poté jsou vypočteny pravděpodobnosti výsledků zápasů (výhra, remíza, prohra) pro každý tým a pro oba týmy dohromady.

Nakonec jsou tyto pravděpodobnosti použity pro vytvoření předpovědi výsledku zápasu a porovnány s reálným výsledkem.

Celkově je tento model relativně jednoduchý, ale může být přesný, pokud jsou k dispozici dostatečná a přesná historická data.

Níže je popsán kód v programovacím jazyce Python:

Nejprve naimportovat potřebné knihovny:

Obrázek 12 Python kód

```
import numpy as np
import pandas as pd
from scipy.optimize import minimize
```

Zdroj: vlastní

Dále je třeba načíst dataset a připravit data pro výpočet parametrů Poissonovy distribuce:

Obrázek 13 Python kód 2

```
# Načtení datasetu
data = pd.read_csv('football_data.csv')

# Příprava dat pro Poissonovu distribuci
home_goals = data['HomeGoals'].values
away_goals = data['AwayGoals'].values
```

Zdroj: vlastní

Následuje definice funkce, která spočítá log-likelihood hodnotu pro Poissonovu distribuci:

Obrázek 14 Python kód 3

```
# Definice funkce log-likelihood
def poisson_log_likelihood(params, observed):
    lambdas = np.exp(np.dot(params, observed))
    log_likelihood = np.sum(np.log(lambdas) - observed * np.log(lambdas))
    return -log_likelihood
```

Zdroj: vlastní

Poté se spočte počet gólů v domácích a hostujících týmech pro všechny zápasy a zkonstruuje se matice s počtem gólů:

Obrázek 15 Python kód 4

```python
# Spočtení počtu gólů v domácích a hostujících týmech pro všechny zápasy
num_games = len(home_goals)
max_goals = np.max(np.concatenate((home_goals, away_goals)))
num_goals = max_goals + 1

# Konstrukce matic s počtem gólů
home_matrix = np.zeros((num_goals, num_games))
away_matrix = np.zeros((num_goals, num_games))

for i in range(num_games):
    home_matrix[home_goals[i], i] = 1
    away_matrix[away_goals[i], i] = 1
```

Zdroj: vlastní

Nyní je třeba použít funkci minimize z knihovny Scipy pro optimalizaci parametrů Poissonovy distribuce:

Obrázek 16 Python kód 5

```python
# Výpočet parametrů Poissonovy distribuce pomocí maximum likelihood
initial_params = np.zeros(num_goals)
result = minimize(poisson_log_likelihood, initial_params, args=(home_matrix - away_matrix))
params = result.x
```

Zdroj: vlastní

Nakonec se získané parametry můžou použít k predikci výsledku zápasu mezi domácím a hostujícím družstvem:

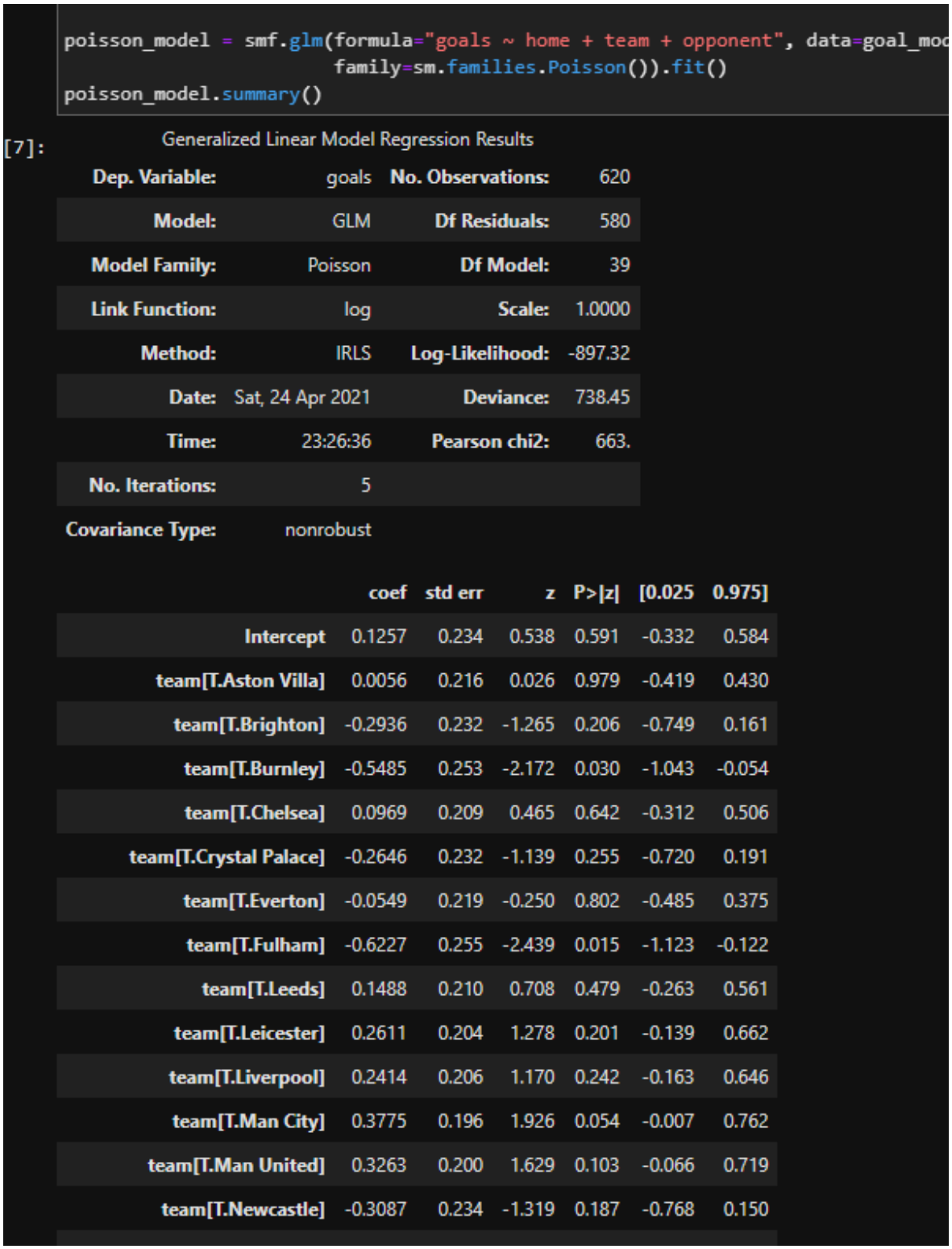
Obrázek 17 Python kód 6

```
# Predikce výsledku zápasu
home_team = 'FC Barcelona'
away_team = 'Real Madrid'
home_index = data['HomeTeam'] == home_team
away_index = data['AwayTeam'] == away_team
home_goals_avg = np.mean(data.loc[home_index, 'HomeGoals'])
away_goals_avg = np.mean(data.loc[away_index, 'AwayGoals'])
```

Zdroj: vlastní

Po úpravě kódu dostáváme následný výstup z programu Python, který se týká zápasů anglické Premier League:

Obrázek 18 Výstup Python EPL



```
poisson_model = smf.glm(formula="goals ~ home + team + opponent", data=goal_mod
                        family=sm.families.Poisson()).fit()
poisson_model.summary()
```

[7]:

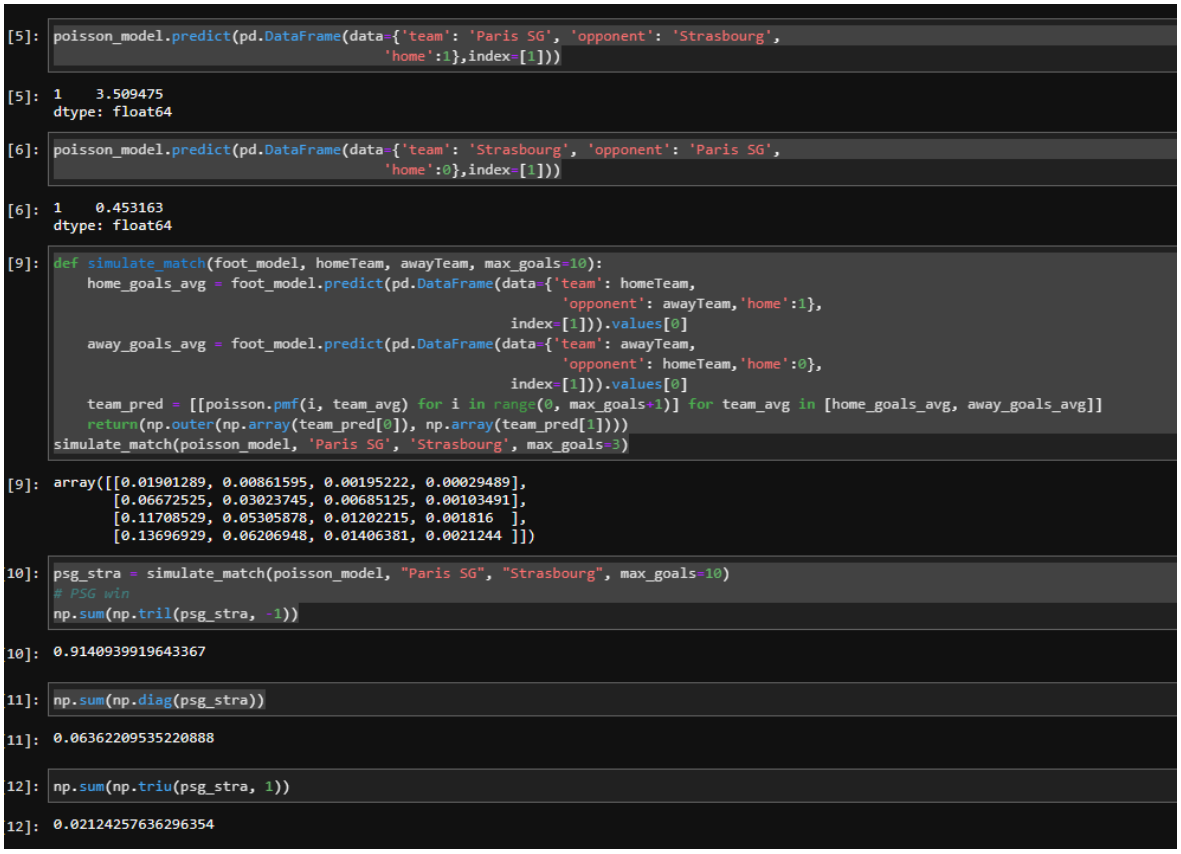
Generalized Linear Model Regression Results

Dep. Variable:	goals	**No. Observations:**	620
Model:	GLM	**Df Residuals:**	580
Model Family:	Poisson	**Df Model:**	39
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-897.32
Date:	Sat, 24 Apr 2021	**Deviance:**	738.45
Time:	23:26:36	**Pearson chi2:**	663.
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.1257	0.234	0.538	0.591	-0.332	0.584
team[T.Aston Villa]	0.0056	0.216	0.026	0.979	-0.419	0.430
team[T.Brighton]	-0.2936	0.232	-1.265	0.206	-0.749	0.161
team[T.Burnley]	-0.5485	0.253	-2.172	0.030	-1.043	-0.054
team[T.Chelsea]	0.0969	0.209	0.465	0.642	-0.312	0.506
team[T.Crystal Palace]	-0.2646	0.232	-1.139	0.255	-0.720	0.191
team[T.Everton]	-0.0549	0.219	-0.250	0.802	-0.485	0.375
team[T.Fulham]	-0.6227	0.255	-2.439	0.015	-1.123	-0.122
team[T.Leeds]	0.1488	0.210	0.708	0.479	-0.263	0.561
team[T.Leicester]	0.2611	0.204	1.278	0.201	-0.139	0.662
team[T.Liverpool]	0.2414	0.206	1.170	0.242	-0.163	0.646
team[T.Man City]	0.3775	0.196	1.926	0.054	-0.007	0.762
team[T.Man United]	0.3263	0.200	1.629	0.103	-0.066	0.719
team[T.Newcastle]	-0.3087	0.234	-1.319	0.187	-0.768	0.150

Zdroj: Vlastní

Obrázek 19 Python Poisson distribuce výstup



```
[5]: poisson_model.predict(pd.DataFrame(data={'team': 'Paris SG', 'opponent': 'Strasbourg',
                                              'home':1},index=[1]))

[5]: 1    3.509475
     dtype: float64

[6]: poisson_model.predict(pd.DataFrame(data={'team': 'Strasbourg', 'opponent': 'Paris SG',
                                              'home':0},index=[1]))

[6]: 1    0.453163
     dtype: float64

[9]: def simulate_match(foot_model, homeTeam, awayTeam, max_goals=10):
         home_goals_avg = foot_model.predict(pd.DataFrame(data={'team': homeTeam,
                                                                'opponent': awayTeam,'home':1},
                                                          index=[1])).values[0]
         away_goals_avg = foot_model.predict(pd.DataFrame(data={'team': awayTeam,
                                                                'opponent': homeTeam,'home':0},
                                                          index=[1])).values[0]
         team_pred = [[poisson.pmf(i, team_avg) for i in range(0, max_goals+1)] for team_avg in [home_goals_avg, away_goals_avg]]
         return(np.outer(np.array(team_pred[0]), np.array(team_pred[1])))
     simulate_match(poisson_model, 'Paris SG', 'Strasbourg', max_goals=3)

[9]: array([[0.01901289, 0.00861595, 0.00195222, 0.00029489],
            [0.06672525, 0.03023745, 0.00685125, 0.00103491],
            [0.11708529, 0.05305878, 0.01202215, 0.001816  ],
            [0.13696929, 0.06206948, 0.01406381, 0.0021244 ]])

10]: psg_stra = simulate_match(poisson_model, "Paris SG", "Strasbourg", max_goals=10)
     # PSG win
     np.sum(np.tril(psg_stra, -1))

10]: 0.9140939919643367

11]: np.sum(np.diag(psg_stra))

11]: 0.06362209535220888

12]: np.sum(np.triu(psg_stra, 1))

12]: 0.02124257636296354
```

Zdroj: vlastní

Na obrázku výše můžeme vidět výstup z programu Python pro francouzskou League 1, konkrétně zápas Paris Saint-Germain vs. Strasbourg. Výstup obsahuje i matici pravděpodobností.

Tento model byl testován na datech anglické Premier League a francouzské League 1, v obou případech na sezónách 2020/2021 a 2021/2022. Jednalo se tedy o několik stovek zápasů, konkrétně každá sezóna čítá 380 zápasů. Bylo tedy analyzováno 1520 zápasů, přičemž úvodní zápasy sezóny 2020/2021 se braly jako první historická data. Model tedy predikoval celkem 1420 zápasů a dosáhl úspěšnosti 820 zápasů – tedy 57,74648% . Což je očekávaná hodnota, většina modelů tohoto typu se pohybuje mezi 50-60% úspěšnosti.

4.4 Model 2 – Náhodné lesy

Nejkomplexnějším modelem v této práci je model založený na náhodných lesích, pro nějž je potřeba zajistit dva datasety. Obě vstupní tabulky ve formátu CSV. První tabulka obsahuje historická data o výsledcích fotbalových zápasů a druhá tabulka obsahuje informace o týmech a jejich vlastnostech.

Pro jednoduchost je uvažováno, že celkový počet branek v zápase bude následovat Poissonovo rozdělení a výsledek zápasu bude záviset na rozdílu počtu vstřelených branek obou týmů. Tento model bude založen na lineární regresi.

Následující kroky byly potřebné k vytvoření modelu:

- Nainstalovat knihovny potřebné pro modelování. Konkrétně knihovny pro práci s daty (pandas), strojové učení (scikit-learn) a pro vizualizaci (matplotlib).

- Načíst historická data o výsledcích fotbalových zápasů z prvního CSV souboru do pandas DataFrame. Tento DataFrame by měl obsahovat informace o datu, domácím týmu, hostujícím týmu, počtu branek vstřelených domácím týmem a počtu branek vstřelených hostujícím týmem.

- Vypočítat rozdíl počtu branek vstřelených domácím a hostujícím týmem pro každý zápas v historických datech a uložit tento rozdíl jako nový sloupec v DataFrame.

- Vytvořit nový DataFrame s informacemi o týmech z druhého CSV souboru. Tento DataFrame by měl obsahovat jména týmů a jejich vlastnosti, jako je například průměrný počet vstřelených branek za zápas, průměrný počet inkasovaných branek za zápas atd.

- Slepovat informace o týmech s historickými daty pomocí názvů týmů jako klíče. Výsledkem bude DataFrame, který bude obsahovat informace o každém

zápase, jako jsou jména domácího a hostujícího týmu, jejich vlastnosti a rozdíl počtu vstřelených branek.

- Rozdělit data na trénovací a testovací sady.

- Vytvořit model založený na lineární regresi, který bude predikovat rozdíl počtu vstřelených branek v zápase. Tento model bude trénován na trénovací sadě dat.

- Použít predikce z modelu a vypočítat předpovědi výsledků zápasů na testovací sadě.

- Vyhodnotit výstup

Obrázek 20 Python random forest

```python
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split

# načtení dat ze souboru csv
data = pd.read_csv(r'C:\Users\mross\Desktop\Mantan\F1.csv')

# výběr sloupců se vstupními daty a cílovými daty
X = data[['home_team_goals', 'away_team_goals', 'home_team_ranking', 'away_team_ran
y = data['result']

# rozdělení dat na trénovací a testovací sady
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_sta

# vytvoření a trénování modelu
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

# výpočet a výpis přesnosti modelu na trénovací a testovací sadě
print("Přesnost na trénovací sadě: {:.2f}".format(rf.score(X_train, y_train)))
print("Přesnost na testovací sadě: {:.2f}".format(rf.score(X_test, y_test)))
```

Zdroj: vlastní

Dle odborné literatury jsou fotbalové prediktivní modely založené na Random Forest obvykle jedny z nejpřesnějších, kde se procento úspěšnosti modelu pohybuje nad 70 %. Z tohoto důvodu byl tento druh modelu vybrán. Bohužel se ale na základě vstupních historických dat a vybraných parametrů nepovedlo vytrénovat vhodný model ani v jednom z pokusů a pokaždé byl model úspěšný pod 50 % z odhadovaných zápasů, což znamená, že nebyl sestaven správně, i při změně proměnných.

4.5 Model 3 - Marže sázkových kanceláří

Posledním modelem, který nakonec byl jen zvažovaným, je model vyhodnocení marže sázkové kanceláře. Tento model vychází z předpokladu, že sázkové kanceláře stanovují koeficienty na základě pravděpodobnosti výsledku, přičemž tyto koeficienty zahrnují marži, kterou si sázková kancelář ponechává.

Pro ukázkový výpočet tohoto modelu se nejprve vezme v úvahu marže sázkové kanceláře, která se obvykle pohybuje okolo 5 procent. Poté se vypočte pravděpodobnost výhry každého týmu na základě koeficientů, které sázková kancelář stanovila.

Například, pokud má tým A koeficient 2,5 a tým B koeficient 3,0, pak pravděpodobnost výhry týmu A se vypočte jako:

pravděpodobnost výhry týmu A = 1/2,5 = 0,4 = 40 %

Pravděpodobnost výhry týmu B se vypočte jako:

pravděpodobnost výhry týmu B = 1/3,0 = 0,33 = 33 %

Celková pravděpodobnost remízy se pak vypočte jako:

pravděpodobnost remízy = 1 - pravděpodobnost výhry týmu A – pravděpodobnost výhry týmu B

Například, pokud jsou koeficienty na remízu 3,5, pak se pravděpodobnost remízy vypočte jako:

pravděpodobnost remízy = 1 - 0,4 - 0,33 = 0,27 = 27 %

Poté se tyto pravděpodobnosti použijí k výpočtu očekávaného počtu gólů pro každý tým pomocí Poissonova rozdělení. Nakonec se očekávané počty gólů sečtou a porovnají se s marží sázkové kanceláře. Pokud jsou očekávané počty gólů vyšší než marže, pak je to signál k sázení na tento výsledek, protože to naznačuje, že sázková kancelář má špatné koeficienty.

Celkově lze říci, že tento model na základě hodnocení marže sázkové kanceláře je poměrně jednoduchý a snadno pochopitelný. Nicméně, přesnost predikce může velmi kolísat a po delším studování těchto modelů nakonec nebyl aplikován pro tuto diplomovou práci.

5 Výsledky a diskuse

5.1 Diskuze výsledků

V této diplomové práci se autor zaměřil na využití prediktivního modelování pro predikci výsledku fotbalového zápasu. Použil dva různé typy modelů - Random Forest a Poisson distribution.

Předpokládaným výsledkem bylo, že model založený na Random Forest dosáhne úspěšnosti 70 %, což je poměrně vysoké číslo a indikuje, že tento model může být velmi užitečný při predikci výsledků fotbalových zápasů. Bohužel se ale autorovi nepovedlo naprogramovat tak kvalitní prediktivní model a jeho úspěšnost byla pod 50 %.

Na druhé straně, Poisson distribution model, který byl modelován v programovacím jazyce Python, dosáhl úspěšnosti 57,74648 %, což je nižší než potenciální úspěšnost Random Forest modelu, ale stále představuje určitou přesnost, která splnila autorova očekávání, vzhledem k tomu, že tyto modely mívají přesnost okolo 60 % v lepších případech.

Nízká úspěšnost Random Forest modelu může být vysvětlena špatnou interpretací proměnných a špatným výběrem proměnných. Tento model by měl být dále upraven, předělán a vylepšen, aby zahrnoval další faktory, jako je forma týmu nebo výkonnost hráčů.

Na druhé straně, Poisson distribution model je jednodušší a založený na základním matematickém modelu. Jeho úspěšnost okolo 58 % může být způsobena tím, že nezahrnuje tolik faktorů, ale přesto se jedná o relativně kvalitní prediktivní model.

6 Závěr

Hlavním cílem této práce bylo zjistit, zda-li se pomocí prediktivního modelování dá odhadnout výsledek fotbalového zápasu a následně toto využít na trhu kurzových sázek a přijít na to, zda je nějaký model profitabilní.

Dle odborné literatury existuje mnoho ověřených modelů, které dokáží v dlouhodobém horizontu dosáhnout určitého profitu. Tyto modely jsou zmíněné v teoretické části a bylo o nich napsáno mnoho vědeckých knih a článků.

Autor se na základě znalostí o jednotlivých modelech rozhodl vybrat dva. První model založený na Poisson distribuci, dosáhl úspěšnosti téměř 58 %. I přestože tato metoda není tak přesná jako by měla být metoda Random Forest, má své výhody, zejména v případech, kdy jsou k dispozici pouze omezená data. Navíc, tato metoda je jednoduše implementovatelná a relativně snadno pochopitelná.

Autor tento model aplikoval nejprve v programu MS Excel, kde se model odhadující zápasy švédské a čínské nejvyšší fotbalové soutěže dostal dokonce přes 70 %, interpretací tohoto vysokého procenta ale bude pravděpodobně malý soubor testovaných zápasů. Vzhledem k časové a manuální náročnosti Excel Modelu vytvořil autor totožný model v programu Python, který na datech anglické a francouzské nejvyšší soutěže dosáhl již zmíněnou úspěšnost téměř 58 %.

Druhý model byl model založený na Random Forest, který ale autor nezvládl správně namodelovat a tento model i přes vysoká očekávání selhal a nedosáhl úspěšnosti ani 50 %.

Při pohledu na úspěšnost modelů a aplikování jich na trh kurzových sázek, tak ani jeden nedosáhl výsledku, který by zajistil na trhu kurzových sázek jakýkoli profit. První model založený na Poissonově distribuci sice udržel solidní úspěšnost odhadů, ale vzhledem k charakteristice kurzových sázek, že velký favorit má velmi malý kurz na výhru a tím pádem to znamená pro sázejícího i malý profit. Byl i tento model nerentabilní, protože měl vyšší úspěšnost právě u těch zápasů, kde byl jasný favorit a tím pádem nedokázal ziskem zacelit ztráty.

Co se týká modelu náhodných lesů ten nebyl autorem správně namodelován a tím pádem ani nemohl být testován na trhu kurzových sázek.

Je však třeba poznamenat, že úspěšnost těchto modelů se může lišit v závislosti na konkrétních podmínkách a kvalitě dat, které jsou k dispozici. Nicméně výsledky, kterých bylo dosaženo v rámci této práce, ukazují, že prediktivní modelování může být velmi užitečnou metodou pro predikci výsledku fotbalových zápasů.

Výsledky této práce mohou být užitečné pro různé subjekty, jako jsou sázkové kanceláře, fanoušci, sportovní analytici a trenéři. Sázkové kanceláře mohou využít predikce výsledků pro stanovení kurzů a odhadnutí rizika, zatímco fanoušci, sportovní analytici a trenéři mohou využít predikce výsledků k lepšímu pochopení hry a k lepšímu plánování strategie.

Vzhledem k úspěšnosti prediktivních modelů použitých v této práci, může být další výzkum v této oblasti velmi užitečný a může poskytnout nové nástroje pro lepší predikci výsledků fotbalových zápasů.

Závěrem je třeba vyhodnotit prediktivní modelování jako příjemné zpestření času a obor statistiky, který bude mít ve sportu čím dál tím větší místo a bude se v budoucnosti rapidně rozrůstat, ale autorova původní premise, že se dá považovat za investiční nástroj je v tuto chvíli spíše lichá.

7 Seznam použitých zdrojů

1. AGRESTI, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.
2. BAIO, G. (2019). Introduction to Bayesian statistics in football. Boca Raton: Chapman and Hall/CRC.
3. BIERMANN, C. Football Hackers: The Science and Art of a Data Revolution. Chichester, England: BLINK Publishing, 2019. ISBN 1788702050
4. BREIMAN, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
5. BREIMAN, L. Random forests. 2001
6. DICKINSON, Peter. Betting on football. In: Handbook of sports and lottery markets. 2012. S. 112-139.
7. DIXON, M. J., & COLES, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265-280.
8. DOMINGOS, P., & PAZZANI, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29(2-3), 103-130.
9. FIELD, A. Discovering statistics using IBM SPSS Statistics. Thousand Oaks: SAGE Publications, 2013. ISBN 978-1-4462-4917-8
10. FORREST, D., & SIMMONS, R. (2000). Predicting football results and the efficiency of fixed-odds betting. Journal of Forecasting, 19(3), 231-250.
11. GODDARD, J. (2005). Regression models for forecasting goals and match
12. GODDARD, M. (2005). Regression models for forecasting goals and match results in association football. International Journal of Forecasting, 21(2), 331-340.
13. GROLL, A., & LEY, C. (2016). A probabilistic model for predicting the outcome of a football match. Journal of the Royal Statistical Society: Series C (Applied Statistics), 65(2), 257-279.
14. HING, N., GAINSBURY, S., & BLASZCZYNSKI, A. (2014). "Bet anywhere, anytime": An analysis of Internet sports bettors' responses to gambling promotions during sports broadcasts by problem gambling severity. Journal of Gambling Studies, 30(2), 321-334. doi: 10.1007/s10899-013-9361-5
15. CHAPMAN, HALL/CRC. Handbook of Statistical Methods and Analyses in Sports. CRC Press, 2019. ISBN 978-1-4987-3736-4

16. KING, G., & ZENG, L. (2001). Logistic regression in rare events data. Political analysis, 9(2), 137-163.
17. KOMPRDOVÁ, K. (2012). Rozhodovací stromy a lesy. Akademické nakladatelství Cerm, s.r.o., Brno
18. KÖNIG, R., & BÜHLMANN, P. (2019). Statistical learning and data science in football. Boca Raton: Chapman and Hall/CRC.
19. KONING, R. H., & VAN DER BERG, G. (2007). Statistical methods for the prediction of football matches. Journal of the Royal Statistical Society: Series D (The Statistician), 56(2), 263-277.
20. KOOPMAN, B., & LIT, R. (2015). A probabilistic approach to modelling football scores. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(2), 481-502.
21. KUHN, M., & JOHNSON, K. (2013). Applied predictive modeling. Springer.
22. LASEK, J., & BIECEK, P. (2019). Machine Learning for Soccer Analytics: An Overview. Journal of Quantitative Analysis in Sports, 15(2), 39-53.
23. LASEK, J., SZCZYPINSKI, P., & KOWALCZYK, R. (2013). Soccer outcome prediction with high accuracy. In International Conference on Computer Science and Information Systems (pp. 169-174). Springer, Berlin, Heidelberg.
24. LIU, J., & KUMAR, S. (2011). A study on feature selection using random forest for real-time heart disease diagnosis. Procedia Computer Science, 6, 573-578
25. MAHER, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109-118.
26. MANNING, C. D., RAGHAVAN, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge: Cambridge University Press.
27. MEMMERT, D., RAABE, D. Data analytics in football: Positional data collection, modelling and analysis. Boca Raton, FL: CRC Press, 2018. ISBN 978-0815381549
28. MENDENHALL, W., SCHEAFFER, R. L., WACKERLY, D.D. Mathematical statistics with Applicatios. Boston: Duxbury Press, 1981. ISBN 0-87872-279-3 results in association football. International Journal of forecasting.
29. RUE, H., & SALVESEN, Ø. (2000). Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 399-418.

30. SHEN, J., ZHANG, H., & CUI, L. (2018). A comparative study of soccer match prediction methods. PloS one, 13(9), e0202793.
31. SUMPTER, D. Soccermatics. Bloomsbury Publishing, 2017. ISBN 978-1-4729-2414-8
32. TSANG, S. H. K., et al. (2019). Neural Networks for Soccer Match Prediction. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1812-1819). IEEE.

33. https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/

34. https://medium.com/geekculture/building-a-simple-football-prediction-model-using-machine-learning-f061e607bec5

35. https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8

36. https://www.academia.edu/download/36217389/B04124012020.pdf

37. "Lotynka vydělala miliardy, kurzové sázení se stalo pro spoustu lidí zdrojem příjmů," Novinky.cz, 10. února 2019. [Online]. Dostupné z: https://www.novinky.cz/finance/clanek/lotynka-vydelala-miliardy-kurzove-sazeni-se-stalo-pro-spoustu-lidi-zdrojem-prijmu-40121027.

38. "Sázkové kanceláře v roce 2020," Tip na kurz, 30. prosince 2020. [Online]. Dostupné z: https://tipnakurz.cz/sazkove-kancelare-v-roce-2020.

39. "Začátek kurzového sázení v ČR," iDNES.cz, 6. ledna 2014. [Online]. Dostupné z: https://www.idnes.cz/sport/fotbal/zacatek-kurzoveho-sazeni-v-cr.A140105_111437_fotbal_min.

40. "Výsledky průzkumu: Jaké jsou návyky sázkařů v ČR?" Sazka.cz, 21. června 2019. [Online]. Dostupné z: https://www.sazka.cz/press/aktuality/vysledky-pruzkumu-jake-jsou-navyky-sazkaru-v-cr.

41. "Sázkové kanceláře: Jak fungují kurzové sázky?" Kurzy.cz, 8. srpna 2018. [Online]. Dostupné z: https://www.kurzy.cz/sazky-a-sazeni

42. https://football-data.co.uk/

8 Přílohy

1. EPL 20_21.csv
2. EPL 21_22.csv
3. F1_20_21.csv
4. F1_21_22.csv
5. Poisson Excel_v2.xlsx
6. Poisson Python 1.txt
7. Poisson Python 2 Dixon Cole.txt
8. Random forest Python.txt