Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

Modely logistické regrese v oblasti esportových dat

BAKALÁŘSKÁ PRÁCE

Studijní program: Aplikovaná informatika
Studijní obor: Aplikovaná informatika

Autor: Michal Lauer
Vedoucí práce: Ing. Zdeněk Šulc, Ph.D.

Praha, Duben 2022

Prohlášení

Prohlašuji, že jsem bakalářskou práci *Modely logistické regrese v oblasti esportových dat* vypracoval samostatně za použití v práci uvedených pramenů a literatury.

V Praze dne DD. Dubna 2021 .....

Podpis studenta

Poděkování

Rád bych poděkoval panu doktoru Zdeňku Šulcovi, který mou bakalářskou práci podpořil a vedl, i přes odlišný studijní obor. Dále děkuji autorům citovaných knih za poskytnutou příležitost se ve logistických modelech zlepšit. Bez nich by se práce psala velmi složitě.

Abstrakt

Esport je sport ve virtuálním světě, který se od počátku dvacátého prvního století rozrůstá mezi novou generací, která vyrůstala ve světě počítačů, mobilů a technologií. Esport nejsou jen klasické sporty jako fotbal, hokej či rugby, ale soutěžit se může i v různých počítačtích, mobilních či konzolových hrách. Tato práce je zaměřená na jednu z počítačových her, a to na hru Coutner-Strike: Global Offensive (CSGO). Téma je aktuální zejména díky tomu, že je esport ve světě relativně nová záležitost a neustále se vyvíjí. Na to se musí přizpůsobit například sázkové kanceláře, které využívají analýzu esportových dat pro nabízení mnoha různých kurzů.

Cílem bakalářské práce je kvantitativně zanalyzovat esportové zápasy ze hry CSGO, predikovat výhry hráčů a týmů a zjistit, jaké prediktory jsou pro výhru zápasu statisticky významné. Použitý datový soubor je z internetového portálu kaggle.com a obsahuje data od roku 2015 až do roku 2020.

Teoretická část práce se zabývá představením a historií esportu a esportové hry CSGO. Teoretická část je pak zaměřená na analýzu dat v programovacím jazyku R. Predikce výher je založena na logistickém vícerozměrném modelu a k jeho vyhodnocení je použita matice záměn a statistiky z ní vypočítané. Pro určení významnosti prediktorů je použit Waldův test.

Výsledek práce jsou logistické modely, které jsou schopné predikovat výhru hráče či týmu podle různých charakteristik. Zároveň jsou identifikované významné prediktory, které výhru zápasu ovlivňují. Toto zjištění by bylo možné použit např. pro stanovení kurzu sázkovou kanceláří na výhru či prohru daného hráče či týmu.

Klíčová slova

model, logistická regrese, predikce, esport

Abstract

DOPSAT DOPSAT DOPSAT DOPSAT DOPSAT DOPSAT DOPSAT DOPSAT

Keywords

model, logistic regression, prediction, esport

Obsah

1. Úvod

Esport je označení pro elektronický sport. Obsahuje všechny důležité oblasti jako klasický sport (např. turnaje, trénování, investice, stadiony, či sázení) s tím rozdílem, že se hraje na nějakém zařízení (počítač, konzole, mobil). Je to jedno z nejrychleji rostoucích odvětví v dnešní době. V roce 2021 se tržní hodnota esportu pohybovala kolem jedné miliardy dolarů — skoro 50% nárůst oproti roku 2020. Lze předpovídat, že v roce 2024 esport překročí hodnotu 1,5 miliardy dolarů. Dalo by se spekulovat, že za takový velký nárůst je zodpovědná pandemie koronaviru v letech 2019 — 2022. Kombinace rozvoje počítačových her a generace, která je na práci s počítačem zvyklá od mala, vzniklo výborné prostředí pro organický růst esportu. Většina populace byla nucena zůstat doma a to otevřelo dveře se s esportem přirozeně seznámit a nějakým způsobem se ho účastnit (online divák, soutěžící, organizátor, fanoušek...).

Práce se zaměřuje primárně na esportový titul Coutner-Strike: Global Offensive (CSGO). Je to jeden z nejdéle hraných esportových titulů, boří mnohé divácké rekordy a je aktuálně nejhranějším First-Person Shooter (FPS) esport titulem. CSGO vyniká nejen detailní herní mechanikou, ale i bohatou a zajímavou historií. Hra je unikátní i tím, že obsahuje mnoho různých módů[1] a hráč může strávit mnoho hodin pouze objevováním komunitních serverů, hraním klasických zápasů či trénováním na offline mapách.

Finálním cílem práce je vytvořit vícerozměrný logistický regresní model, který předpovídá výsledky zápasů. Pro tvorbu kvalitního modelu bude kritické zvolit vhodné charakteristiky hráčů a jednotlivých týmů. Pro predikci jsou použity charakteristiky, které se nacházejí ve dvou samostatných datových souborech, které podávají informace jak už o zápase (např. datum, výsledek zápasu, výsledek jednotlivých map, typ zápasu) a hráčích (např. charakteristiky za zápas).

Predikovat výhru hráče a týmu je důležité zejména v oblasti sázení, která je s esportem úzce spojená. Přesné predikce a kvalitní modely sázkovým kancelářím umožňují stanovovat výhodné a profitabilní kurzy. Kurzy pak nemusí být pouze na výhru či prohru, ale např. na charakteristiky hráčů či hranou mapu. Logistický model je pro predikci výhry či prohry preferován kvůli své snadné interpretaci.

Práce je rozdělená do tří částí. V první části je kladen důraz na esport, jeho vývoj, a na esportový titul CSGO. Jsou zde také představená pravidla, podle kterých se hra hraje. V druhé části jsou popsány popisné a statistické metody. Jsou zde definované grafické nástroje pro popis datového souboru, logistického regresního modelu, a evaluační nástroje pro model. Třetí část se zaměřuje na praktickou tvorbu modelů, jejich interpretaci, a vzájemné porovnání.

[1] rozšíření, jak hru hrát. Každý mód má svá vlastní pravidla, mapy, či herní fanoušky

2. Představení esportu

2.1 Historie esportu

I přes fakt, že esport není obecně známý pojem mezi širokou veřejností, má přes 70 let bohaté historie. Za jeho počátky by se daly považovat arkádové automaty, kde hráči z počátku soutěžili sami proti sobě. Největší rozvoj arkádových automatů se děl kolem 70 let minulého století. Nejen za tímto účelem byla 9. 2. 1982 založena Twin Galaxies National Scoreboard (TGNS). TGNS měla na starosti nejen udržování výsledkové tabulky (scoreboard), ale i tvorbu prvotních pravidel pro férovou hru. Za tímto účelem byla vydána kniha Twin Galaxies' Official Video Game & Pinball Book of World Records.

Na přelomu osmdesátých let minulého století se začal esport vyvíjet již více profesionálním směrem. V roce 1972 pořádala Stanfordská Universita historicky první esportový turnaj v arkádové hře Spacewar!. Výherce si mohl odnést předplatné magazínu Rolling Stones. Dále v roce 1983 byl založen první esportový profesionální tým, který se nacházel ve Spojených státech. Všechno toto se stalo díky podnikateli Walteru Day, který je zakladatel společnosti TGNS a založil již zmíněný prvních esportový tým. Ač se Walter považuje za jednoho z hlavních pionýrů esportu, v roce 2010 TGNS opustil kvůli své vášni pro hudbu.

Další důležitou kapitolou ve vývoji esportu je příchod internetu a výkonných počítačů. Hráči se dostali k rychlejším sestavám, stolní počítače se stali cenově dostupnějšími a díky tomu se zpřístupnili k více lidem. Klesala cena hardwaru, vývoj nové technologie a her se zrychloval. Díky rozvoji počítačových sítí se mohli hrát LAN[1] party či organizovat BYOC[2] turnaje. Dále už esport potřeboval jen čas na organický růst a dnes má tržní hodnotu přes jednu miliardu amerických dolarů (Gough, 2022), (Larch, 2022).

2.2 Zasazení do dnešní doby

V dnešní době je esport téměř miliardová záležitost. Díky pandemii, která trvá již od r. 2019, si esport ještě přilepšil. Dle průzkumu[3] z října roku 2020 si 73 % dotázaných myslelo, že se úroveň zájmu a obchodní činnost esportu v Q4 2020 a Q1 2021 zvětší. Respondenti, kteří se průzkumu zúčastnili, jsou považování za experty v oblasti esportu. Tento průzkum byl následné podpořen růstem že tržní hodnoty esportu a mezi lety 2019 a 2020 vzrostla o téměř 50 % (Gough, 2022).

[1] Hráči hrají v jedné místnosti na lokální počítačové síti.
[2] z ang. Bring Your Own Computer, kde si hráči si na akci donesou vlastní počítač
[3] https://www.statista.com/statistics/1247902/covid-impact-esports-investments

K takto prudkému růstu tržní hodnoty esportu z velké části přispěla právě pandemie. Mladá generace byla nucena zůstat doma, což dovolilo i esportem nedotčeným jedincům do tohoto světa proniknout. Větší zájem o esport přinesl i větší tržby herním studiím, která začala do esportových turnajů více investovat (Professeur, 2022), (liquipedia.net, 2021). S větším počtem diváku roste i marketingový potenciál, investiční příležitost a kariérní růst.

Jedním z dominantních žánrů je žánr FPS. V této kategorii jsou nejvýznamnější hry CSGO a Valorant. V tomto žánru proti sobě hrají dva týmy, většinou složené z pěti hráčů. Každý hráč pak má v týmu různou roli, jako např. velitel či odstřelovač. Jeden tým má obvykle za úkol něco zničit (položit bombu, unést rukojmí) a druhý tým jim v tom musí zabránit (ochránit oblast proti bombě, záchrana rukojmí).

Druhý významný a zajímavý žánr je žánr Battle Royale (BR). V těchto hrách hraje buď každý hráč sám za sebe, ve dvojici, nebo ve skupině po čtyřech. Zde hráči padají na začátku kola na velkou mapu. Jejich úkolem je získat vybavení, aby mohl porazit ostatní hráče a kolo sami, nebo s týmem vyhrát. Nacházejí se zde různé role, avšak trošku rozdílné oproti žánru FPS. Hlavním titulem této kategorie je hra Fortnite, která žánru dominuje. Stal se z ní jak esportový titul, tak perfektní marketingové místo pro teenagery. Hráči si zde mohou koupit oblečky různých filmových či komiksových postav. Pokud vychází nový film, ve hře se může objevit „event“ (událost), který daný film propaguje. Toto lze vidět například na propagaci Avengers: Endgame[4].

2.3 Představení titulu Counter-Strike: Global Offensive

CSGO, jak ho známe dnes, má bohatou a dlouho historii. Ne vždy se to ovšem jmenovalo stejně. Úplně první iterace hry se jmenovala čistě Counter-Strike a byl to pouze mód[5] do hry Half-Life. Half-Life byl vyvinutí společností Valve, tehdy primárně společností zaměřenou na vývoj her. Mód byl vytvořen studenty vysoké školy, panem Minh Le a Jess Cliffe. Toto rozšíření začali programovat v roce 1999. Jelikož mód byl neoficiálním rozšířením, Valve o něj neprojevovalo veliký zájem. Až po pěti betaverzích hry Counter-Strike si společnost Valve všimla rozšíření, její komunity, ale především jejich autorů. Minh a Jess se v roce 2000 stali oficiálními zaměstnanci Valve a duševní vlastnictví módu prodali. Autoři, nově jako zaměstnanci Valve, roku 2000 vydávají první oficiální verzi hry Counter-Strike. I přes toto „oficiální“ datum vydání je většina komunity přesvědčena, že výročí má CSGO v den svého úplně první vydání, a to 18. června 1999.

[4]Trailer pro propagaci události: https://www.youtube.com/watch?v=TanGK9o_d24

[5]upravení či rozšíření hry

Hra je z žánru FPS a hraje se primárně online proti skutečným hráčům. Counter-Strike se v herní komunitě rychle rozrostl díky své jednoduchosti. Hra se dá velmi dobře popsat pořekadlem „Lehké hrát, těžké vypilovat“. Hra má mechaniky[6], které jsou lehké na pochopení, ale velmi těžké na vypilování k dokonalosti. Spolu s touto vlastností je hra vlastně velmi jednoduchá a hráč hraje buď za policisty, nebo za teroristy. Hráči tak mohli, a stále mohou, hru velmi lehce a rychle začít hrát, jelikož se tento formát od roku 2000 nijak extrémně nezměnil.

Hra tedy rostla zejména díky své komunitě. Hráči hru různě upravovali, přidávali další módy, typy her, zbraně, mapy či audiovizuální obsah. Tento trend se přenášel přes mnoho různých verzí hry. První velký „průlom“ udělala verze 1.6, tedy Counter-Strike 1.6. Ta vynikala jak esportem, tak komunitním obsahem. Jen v České a Slovenské republice bylo několik herních serverů, na kterých se mohlo sejít sta tisíce hráčů. Např. na česko-slovenském herním portálu kotelna hrálo celkem přes 1,5 milionu unikátních hráčů (csko.cs, 2022). Hra byla populární nejen mezi obyčejnými hráči, ale i profesionály.

Counter-Strike 1.6 je pionýrem esportu pro FPS žánr. Za podpory Valve se hráli první major[7] turnaje, kde hráči mohli ukázat svůj um za tehdy relativně velkou sumu peněz. V dnešní době majory trhají světové rekordy a kouká se na ně miliony diváků[8] Hra se časem vyvíjela, hráči nalézali nové strategie či triky a Valve vydalo novou verzi — Counter-Strike: Source. Tato nová verze získala nepříliš pozitivní ohlas, jelikož velmi rozdělila herní komunitu. Představila nové mechaniky, staré mechaniky změnila a hráčům, zejména v esportu, se nechtělo učit něco úplně nového. Valve se rozhodlo sjednotit herní komunitu, a proto vydalo novou verzi hry s názvem CSGO

CSGO se snažilo sjednotit oba tábory z her Counter-Strike 1.6 a Counter-Strike: Source. Hra vyšla 21. srpna 2012 a z počátku nebyla tolik úspěšná, ale díky přidání různých skinů (Valve, 2013) na zbraně hra přilákala úplně nové publikum. Díky novému a velkému publiku se začali hrát menší esportové turnaje právě ve hře CSGO, ke kterým se později přidali i profesionálové z předchozích dvou verzí. Díky tomuto organickému růstu má Counter-Strike velmi silnou komunitu, která se o hru i nadále stará. I přes netradiční interakci mezi Valve a herní komunitou hra stále roste. CSGO se díky své dlouhé historii, bohaté komunitě a různým možnostem, jak hru hrát, dostala na špičku esportu. I přes několik titulů, které se s hrou snaží soutěžit, je hra stále největším a nejsledovanějším esport titulem v rámci FPS žánru (Henningson, 2020).

[6] herní prvky či unikátní vlastnosti

[7] turnaj pořádaný přímo Valve, který má největší prestiž

[8] https://www.invenglobal.com/articles/15619/csgo-major-breaks-viewership-records-overtakes-the-international

2.4 Propojení práce a titulu Counter-Strike: Global Offensive

Práce se zaměřuje na identifikování významných prediktorů a následně vytvoření regresního modelu. Před jakoukoliv prací s daty je ale nutné pochopit, jak se hra vlastně hraje a jaká jsou její pravidla. Ve hře CSGO hraje pět hráčů proti pěti (dále jen 5v5). Hra se většinou hraje online, avšak velké esportové turnaje se hrají offline, tedy v nějaké např. aréně. Hra má v základu 30 kol a po prvních patnácti se mění strany. Jedna strana jsou policisté (Counter-Terrorists či CT), kteří mají za úkol chránit „bomboviště“ - část mapy, která má vybouchnout. Naopak cíl Teroristů (T) je právě bombu položit a „bomboviště“ nechat vybouchnout. Vyhrává tým, který první vyhraje 16 kol. Pokud ovšem po první 30 kolech je stav nerozhodný, tedy 15:15, hraje se prodloužení. Tento formát není standardizovaný pro všechny turnaje, proto zmíním pouze pravidla, která se týkají turnajů od společnosti Valve (již zmíněné a nejvíc prestižní Majory). Zde se hraje prodloužení ve formát Bo6, tedy kdo první získá 4 body, vyhraje zápas. Takto může jít zápas teoreticky do nekonečna. Nejdelší semi-profesionální zápas, který se ovšem neodehrál na Majoru, se stal mezi týmem exceL a XENEX(Professeur, 2015). Zápas pokračoval do úctyhodných 88 kol.

V každém kole má tým určitý počet peněz. Každá hráč začíná polovinu (ted v první a šestnácté kolo) s $800. Finance každého hráče pak záleží na mnoha faktorech, jako kolik vyhrál jeho tým kol v řadě, kolik nakoupil zbraní, kolik zabil nepřátel, kolik peněz dostane hráč za zabití či jak kolo skončí. V profesionálním týmu je velmi obtížné pracovat s financemi, jelikož všichni musí být v tomto ohledu jednotní. V tuto chvíli přichází na řadu tzn. In-Game Leader (velitel týmu). Tuto roli má většinou jeden hráč v každém týmu. Je to ta nejdůležitější role ze všech Má na starosti např. finance týmu, rozhoduje kdy se koupí a kdy půjde tzn. eco (hráči nekoupí nic, aby ušetřili peníze), jaké se budou hrát mapy či jaká se půjde v daném kole strategie.V dnešní době k tomu In-Game Leader má i trenéra. Ten hru nehraje, ale pozoruje hráče a dává jim různé typy a triky.

Role trenéra není nijak silně definovaná a každý esportový tým má trošku jiného trenéra. V jednom případě může být trenér čistě jako podpora a pomáhá hráčům když se nedaří a řeší interní problémy. V jiném týmu může ovšem mít velký zásah do hry, pomáhat In-Game Leaderovi se strategiemi, obelstění soupeře či sledováním předchozích zápasů pro kontinuální zlepšování týmu. Další role v týmu jsou například Entry Fragger (má za úkol získat první zabití pro tým), support (podporuje svůj tým za pomoci různých granátů nebo se často pro svůj tým obětuje), AWP hráč (hráč je specifický tím, že hraje primárně s jednou zbraní) a Lurker (chodí po mapě sám a snaží se nepřítele odchytnout ze stran, které by nečekali)

Zápasy se pak hrají ve formátech „Best of“. Best of 3 například znamená, že se hrají tři mapy. Kdo první vyhraje dvě mapy, vyhrál celý zápas. Turnaje se pak odehrávají v tradičních formátech, jako je pavouk. Ten se charakterizuje tím, že vypadá jak pavučina, jde z leva doprava a každý tým může prohrát pouze jednou. Následně tu máme Upper/Lower bracket formát, který je v podstatě pavoučí formát, akorát jsou zde dvě „sítě“ a každý tým může prohrát maximálně jednou, jelikož druhá prohra znamená vyřazení z turnaje. Specifičtější formát pro CSGO je například swiss, který se počítá přes různé body a statistiky výsledných zápasů.

3. Teoretická část

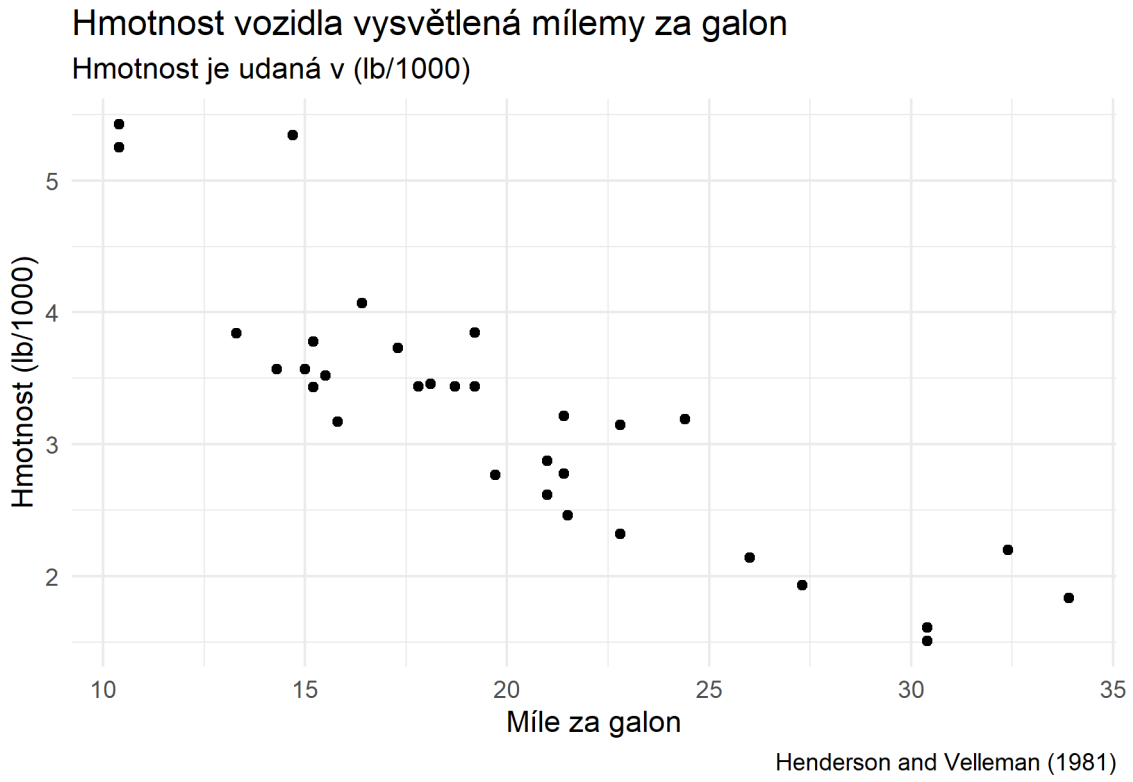
V následující části jsou popsány jak teoretické metody pro vizualizaci dat, tak i tvar, forma a vyhodnocení logistického regresního modelu. Ke každé části, která se věnuje popisu dat pomocí nějakého grafu, je přidána praktická ukázka s popisem a praktickým vysvětlením. vhodné.

3.1 Vizualizace dat

3.1.1 Bodový graf

Bodový graf slouží pro zobrazení vztahu dvou kvantitativních proměnných. Z pravidla se vysvětlovaná proměnná dává na osu Y, zatímco proměnná vysvětlující se nachází na ose X. Vysvětlovaná (nezávislá) proměnná je ta proměnná, která má být předvídaná. Vysvětlující proměnná se naopak snaží vysvětlovanou proměnnou předpovědět či popsat.

Zobrazením vysvětlované a vysvětlující proměnné na bodovém grafu lze vidět např. sílu korelace nebo vztah mezi proměnnými (např. lineární, kvadratický, logaritmický).



Obrázek 3.1: Bodový graf hmotnosti a míly za galon

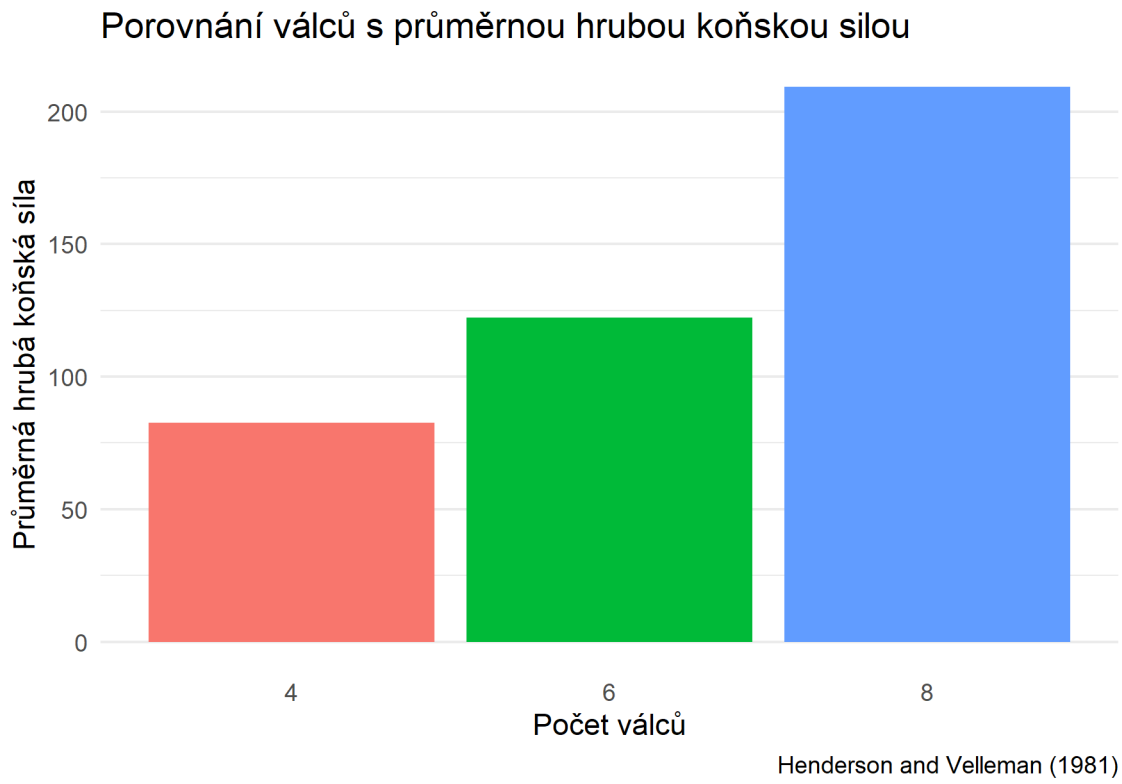
Obrázek 3.1 zobrazuje negativní korelaci mezi hmotností vozidla a mílemi ujetými za galon.

3.1.2 Sloupcový graf

Sloupcový graf slouží k zobrazení četností kategorií. Na jednu osu (z pravidla osu X) se položí možné kategorie. Na druhou osu se pak položí sledovaná statistika. Sledovat můžeme nejen četnost, ale i průměr či kteroukoli jinou statistiku, kterou bude možné na ose zobrazit.

Nejčastější se pomocí sloupcového grafu porovnává absolutní četnost přes kategorie. Řazení kategorií se dále odvíjí podle toho, zda je daná proměnná ordinální či nominální. V případě nominální proměnné se sloupce řadí podle absolutní četnosti, a to od nejvyšší po nejnižší. V případě ordinální proměnné se zachovává přirozené řazení. Příklad sloupcového grafu je zobrazen na obrázku

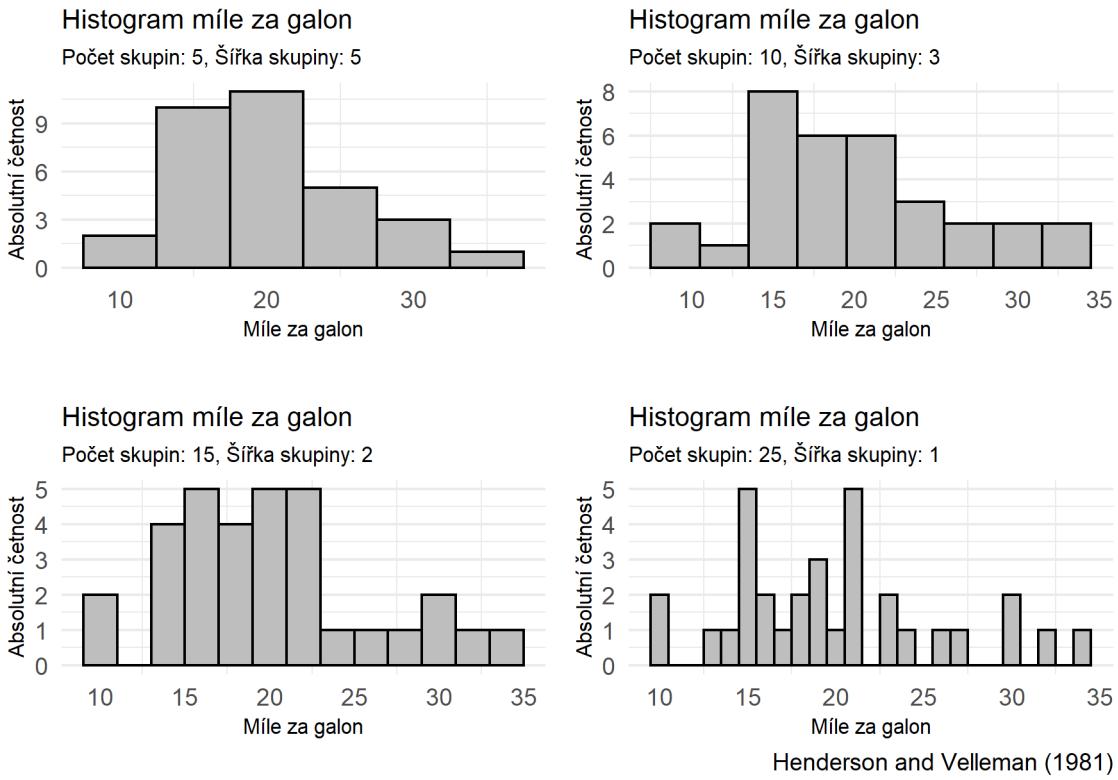
3.2 , který porovnává průměrnou hrubou koňskou sílu s počtem válců. Je na něm také vidět závislost hrubé koňské síly na počtu válců.



Obrázek 3.2: Sloupcový graf počtu válců a průměrné hrubé koňské síly

3.1.3 Histogram

Histogram je speciální typ sloupcového grafu. Hlavní rozdíl je v tom, že popisuje rozdělení spojité proměnné a mezi sloupci není žádná mezera. Pro histogram je třeba data sloučit do skupin o určité šířce. Správný výběr počtu skupin je kritický, jelikož může velmi silně ovlivnit interpretaci dat. Pokud se vybere příliš malý počet skupin, data se seskupí a může se ztratit důležitý vztah. Pokud se ovšem vybere moc velký počet skupin, v datech bude obtížné najít nějaký obecný vztah či trend, viz obrázek 3.3.



Obrázek 3.3: Porovnání histogramů s různým počtem skupin

Pro vhodný počet skupin existuje mnoho způsobů. Nejznámější je takzvané Sturgesovo pravidlo, které se spočítá následujícím vztahem:

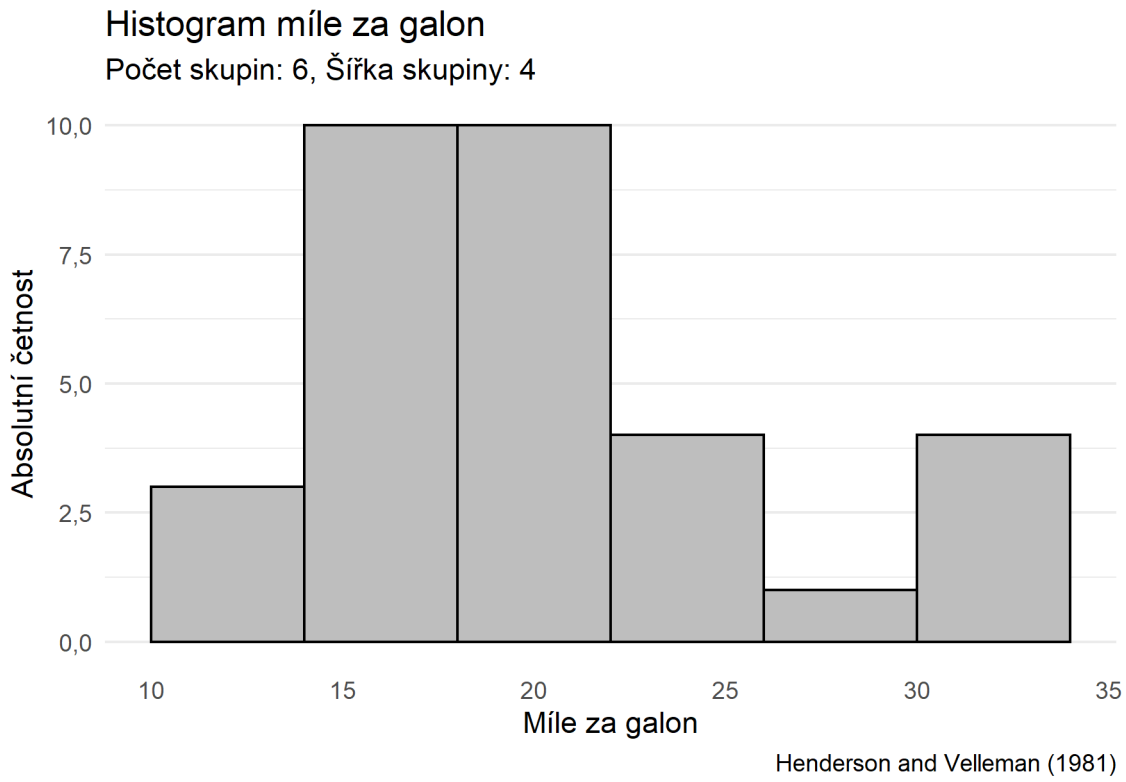
$$k \doteq 1 + 3{,}3 * log_{10}(n) \tag{3.1}$$

kde k je přibližný počet skupin a n je počet pozorování. Druhý parametr, který je pro tvorbu histogramu potřeba, je šířka skupiny. Ta by měla být ideálně stejná pro všechny skupiny. Pokud tomu tak není, histogram může být zavádějící a čtenář mu nemusí plně rozumět. Pro vypočtení počtu skupin má šířka skupiny následující tvar:

$$w = \frac{max(X) - min(X)}{k} \tag{3.2}$$

kde X je zobrazovaná proměnná, k je počet skupin a w je výsledná šířka intervalu. Nutné je však podotknout, že není pravidlem se danými výpočty řídit a výsledný sloupcový graf je nutné přizpůsobit

konkrétnímu datovému souboru.



Obrázek 3.4: Histogram s počtem skupin dle Sturgesova pravidla

Obrázek 3.4 ukazuje histogram proměnné míle za galon. Počet sloupců je vypočítán podle Sturgesova pravidla 3.1.

3.1.4 Krabičkový graf

Pětičíselné shrnutí

Pětičíselné shrnutí je číselná tabulka, která pomocí pěti různých čísel shrnuje seřazenou číselnou řadu. Základní statistický nástroj pro vytvoření takové tabulky jsou kvantily. Hodnota P-tého percentilu označuje číslo, které rozděluje seřazenou číselnou řadu na dva intervaly. První interval obsahuje $P * 100\%$ číselné řady a druhý analogicky $(1 - P) * 100\%$. Různé hodnoty percentilů mohou mít specifičtější pojmenování a značí se Q_P. Percentil $P = 0{,}5$ se označuje jako medián a rozděluje seřazenou číselnou řadu na polovinu. Percentily, kde $P = 0{,}25$ nebo $P = 0{,}75$, se označují jako kvartily a značí se Q_1 a Q_3. Oba tyto typy kvartilů jsou použité při tvorbě tabulky.

$Q_0(Q_0)$	$Q_{0,25}(Q_1)$	$Q_{0,50}$	$Q_{0,75}(Q_3)$	$Q_{1,00}$
1,513	2,58125	3,325	3,61	5,424

Tabulka 3.1: pětičíselné shrnutí hmotnosti vozidla (lb/1000)

Příkladem pětičíselné shrnutí je tabulka 3.1,

kde Q_0 a $Q_{1,00}$ označují minimum a maximum číselné řady. Kvartily Q_1, Q_2 (medián) a Q_3 jsou čísla, která rozděluji časovou řadu na na čtvrtiny. V prvním případě, tedy $Q_1 = Q_{0,25}$, je 25% čísel menší než 1,513 a 75% dat větší. Pro kvantil $Q_3 = Q_{0,75}$ je 75% čísel menších než 3,61 a 25% větších. $Q_{0,50}$ označuje medián.

Krabičkový graf

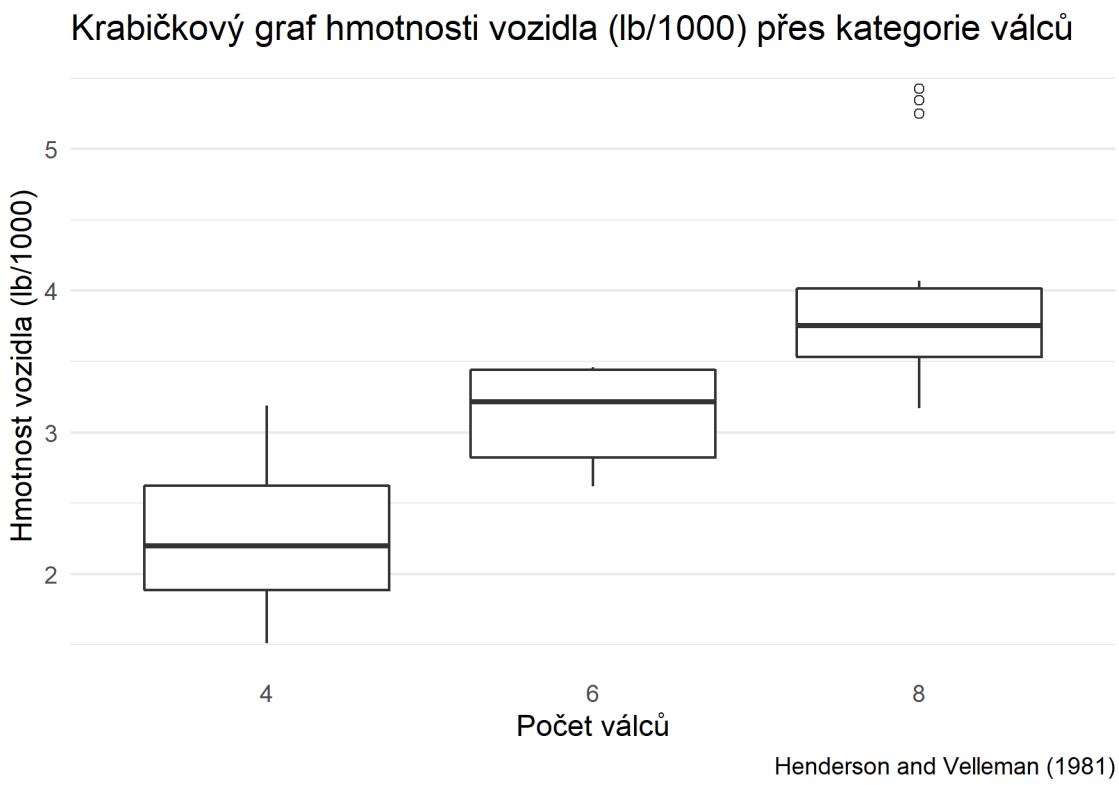
Krabičkový graf je grafické zobrazení a rozšíření pětičíselné shrnutí. Kromě grafického zobrazení pěti kvantilů ukazuje odlehlé a extrémní hodnoty. V Krabičkovém grafu se také nachází obdélník, který ukazuje mezikvartilové rozpětí (IQR), tedy prostředních 50 % dat. V obdélníku se také nachází černá čára, která značí medián. Z prostředního obdélníku vedou oběma směry čáry, jejichž konce značí hranici pro odlehlá pozorování. Pokud datový soubor neobsahuje žádná odlehlá pozorování, konec těchto čar značí minimum a maximum datového souboru. Pozorování, která jsou buď větší než horní hranice, nebo menší než spodní hranice, označujeme jako odlehlá nebo extrémní.

$$\begin{aligned} \textit{Spodní hranice} &= Q_1 - 1{,}5 * IQR \\ \textit{Horní hranice} &= Q_3 + 1{,}5 * IQR \end{aligned}$$

Hodnoty, které spadají do intervalu $\langle Q_1 - 1{,}5IQR; Q_1 - 3IQR\rangle$ a $\langle Q_3 + 1{,}5IQR, Q_3 + 3IQR\rangle$ se nazývají jako odlehlé. Hodnoty které leží mimo tento vztah, tedy hodnoty menší než $Q_1 - 3IQR$ nebo větší než $Q_3 + 3IQR$ se nazývají jako hodnoty extrémní.

Odlehlá pozorování se v krabičkovém grafu většinou značí kolečkem, zatím co pozorování extrémní hvězdičkou.

Díky grafickému zobrazení lze lehce porovnávat rozdělení jedné vysvětlované kvantitativní proměnné tříděné přes několik kategorií.



Obrázek 3.5: Krabičkový graf hmotnosti auta pro různý počet válců

Průhledná kolečka v obrázku 3.5 v kategorii osmi válců značí odlehlé hodnoty, t.j. hodnoty v intervalu $\langle Q_3 + 1{,}5IQR, Q_3 + 3IQR\rangle$.

3.1.5 Korelační matice

Korelace

Korelace popisuje směr a sílu vztahu mezi dvěma proměnnými X a Y. Značí se r a nabývá hodnot $\langle -1; 1\rangle$.

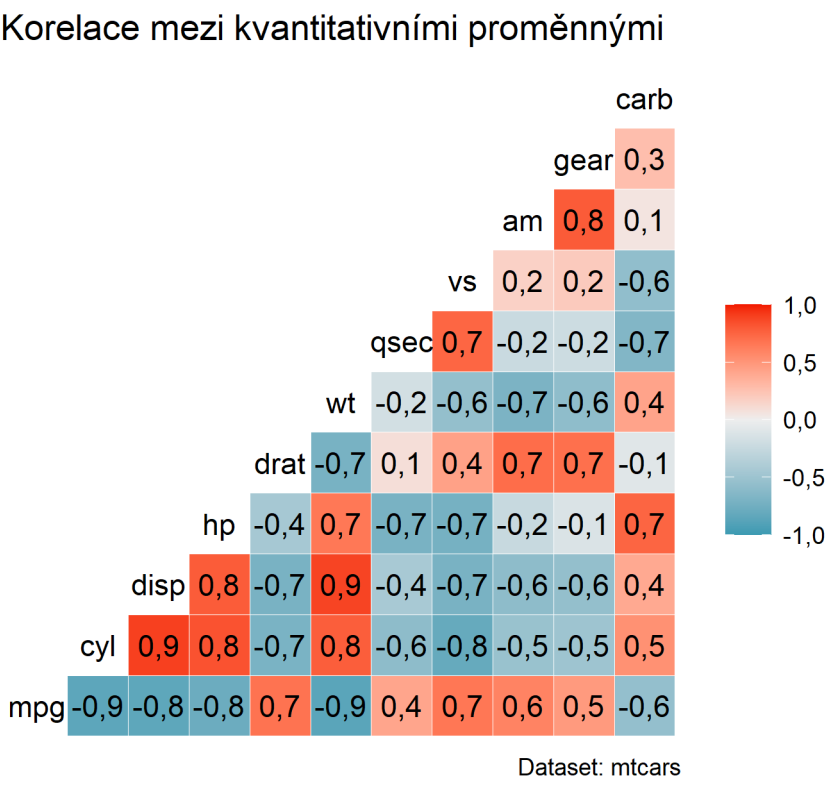
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}} \tag{3.3}$$

Čím větší je absolutní hodnota korelace mezi proměnnou X a Y, tím lépe lze pomocí jedné proměnné vysvětlit proměnnou druhou. Kladnost, případně zápornost korelace pak značí směr vztahu, tedy zda proměnná X s růstem proměnné Y klesá či stoupá. Pokud je korelace záporná, tedy $r < 0$, s růstem jedné proměnné klesá proměnná druhá. Naopak při kladné korelace, tedy $r > 0$, s růstem jedné proměnné roste i druhá.

Pokud se korelace r vychází kolem nuly, neexistuje lineární závislost mezi proměnnou X a Y. Důležité je také podotknout, že korelace neznamená kauzalitu. Pokud existuje kladná korelace mezi proměnnou X a Y, neznamená to, že růst jedné proměnné způsobil růst druhé proměnné.

Korelační matice

Korelační matice je nástroj, díky kterému lze zobrazit korelaci mezi více jak dvěma páry proměnných. Matice může být zobrazená pomocí grafu a je velmi užitečná v regresní analýze kvůli předpokladu nezávislosti vysvětlujících proměnných. Pokud jsou při tvorbě modelu prediktory korelované, vzniká problém tzn. multikolinearity. Při multikolinearitě se zhoršuje přesnost a vypovídací hodnota koeficientů (Kleinbaum et al., 2010). V takovém případě je potřeba zvýšit počet pozorování nebo z modelu jeden z vysoce korelovaných prediktorů odebrat.



Obrázek 3.6: Korelační matice

Graf korelační matice může mít mnoho podob. V příkladu obrázku 3.6 je zobrazená korelační matice jako teplotní mapa. Z obrázku je možné pozorovat vysokou pozitivní korelaci mezi páry proměnných *cyl*, *disp* a *hp*. Naopak skoro žádná korelace není mezi proměnnou *qsec* a proměnnou *drat*. Korelační matice je zároveň symetrická, jelikož korelace mezi X a Y je stejná jako korelace mezi Y a X. Díky této vlastnosti je možné zobrazit pouze část korelační matice pod úhlopříčkou bez ztráty jakékoliv informace.

3.2 Logistická regrese

Logistická regrese je způsob, jak popsat vztah mezi jedním či několika prediktory a jednou binární vysvětlovanou proměnnou. K tomu slouží spojovací funkce, která transformuje lineární kombinaci prediktorů na index z. V případě logistické regrese se tato funkce nazývá logistická a je definovaná jako

$$f(z) = \frac{1}{1 + e^{-z}}. \tag{3.4}$$

Obor hodnot funkce je interval $\langle 0, 1 \rangle$. Proměnná z je lineární kombinace prediktorů $X_1, X_2, ..., X_k$, jejich koeficientů $\beta_1, \beta_2, ..., \beta_k$ a parametru α.

$$\begin{aligned} z &= \alpha + \beta_1 X_1 + ... + \beta_2 X_2 + \beta_k X_k \\ &= \alpha + \sum_{i=1}^{k} \beta_i X_i \end{aligned} \tag{3.5}$$

Mějme tedy binární vysvětlovanou proměnnou Y, u které hodnota 1 značí výskyt jevu. Pravděpodobnost, že jev nastane vzhledem k definovaným prediktorům lze zapsat jako

$$P(Y = 1 \mid X_1, X_2, ..., X_k) = \frac{1}{1 + e^{-\left(\alpha + \sum_{i=1}^{k} \beta_i X_i\right)}}, \tag{3.6}$$

kde α a β_i jsou parametry odhadnuté z datového souboru (Kleinbaum et al., 2010).

3.2.1 Interakce mezi prediktory

Prediktory mohou na výslednou pravděpodobnost působit i dohromady, tzn. interakcí. Jako příklad lze uvést model, který predikuje *nehodu* podle číselného prediktoru *rychlost* a dichotomického prediktoru *dálnice*. Šance na nehodu může být ovlivněná jak rychlostí, tak tím, zda je řidič na dálnici.

3.2.2 Interpretace parametrů

Parametry α a β_i značí logaritmus šance. α je logaritmus šance v případě, že všechny prediktory jsou teoreticky rovné 0. Parametr β_i značí logaritmus šance pro prediktor X_i V případě, že všechny prediktory jsou konstantní a prediktor X_i se změní o jednotku, přirozený logaritmus šance se změní o B_i. Toto lze pozorovat například u binárních prediktorů, kdy typicky přítomnost daného prediktoru, značená jedničkou, změní výslednou šanci právě o odhadnutý parametr β. Pro přechod z přirozeného logaritmu šance na šanci lze využít vztahu

$$šance \ = e^{\beta_i}. \tag{3.7}$$

Šance je podíl dvou pravděpodobností. Pokud bychom měli šanci jevu A oproti jevu B $2:1$, značí to, že výskyt jevu A je dvakrát pravděpodobnější než výskyt jevu B a jev A se vyskytuje ve $\frac{2}{3}$ případů. Šance e^{β_i} tedy značí vztah mezi prediktorem X_i a vysvětlovanou proměnnou Y. Pokud je šance kladná, značí to, že s vyšší hodnotou prediktoru X_i se zvyšuje šance že $Y = 1$. Pokud je naopak nižší, pravděpodobnost se zmenšuje. Pokud je potřeba interpretovat pravděpodobnost jako šanci, použije se logitová funkce

$$šance \ jevu \ A = \frac{p}{1-p}, \tag{3.8}$$

kde p je pravděpodobnost výskytu jevu A.

3.2.3 Maximální pravděpodobnost

Parametry logistického modelu v rovnici 3.6 jsou pouze teoretické a je třeba je odhadnout. Již vypočtené odhady se proto neznačí pouze β_i, ale $\hat{\beta}_i$. Pro odhad parametrů se při logistické regresi používá metoda maximální věrohodnosti. Pro výpočet maximální věrohodnosti se používá věrohodnostní funkce $L(\theta)$ kde θ jsou parametry logistického modelu $\alpha, \beta_1, ..., \beta_k$. Pro logistickou regresi má věrohodnostní funkce tvar

$$L(\theta) = \Pi_{l=1}^{m_1} P(X_i) \Pi_{l=m_1+1}^{n} (1 - P(X_i)), \tag{3.9}$$

kde n je počet pozorování a m_1 je počet příznivých ($Y = 1$) jevů. Funkce předpokládá, že datový soubor je seřazen tak, že prvních m_1 výskytů jsou jevy příznivé. $P(X_i)$ poté značí logistickou funkci 3.4. Pro vypočtení optimálního parametru β_i je nutné vypočítat maximum funkce $L(\theta)$ vzhledem k parametru β_i. Parametr β_i lze tedy získat derivací funkce $L(\theta)$ vzhledem k parametru β_i (Kleinbaum et al., 2010).

$$\frac{\partial L(\theta)}{\partial \beta_i} = 0 \tag{3.10}$$

3.2.4 Křížová validace

Při tvorbě logistického modelu může dojít k takzvanému přeučení modelu. To znamená, že výsledný model je velmi přizpůsobený na data, ze kterých byl vytvořen, a nebude připravený na náhodnou variaci, která může v nových datech nastat. Z tohoto důvodu se datový soubor rozděluje na dvě podmnožiny. Jedna podmnožina, většinou zvaná *trénovací*, slouží k sestavení a natrénování modelu. Model se pak otestuje na druhé množině dat, na které nebyl natrénovaný. Druhá množina dat se většinou nazývá *validační* nebo *testovací*. Pokud je následně model použit na *testovací* množinu dat a výsledky jsou vyhodnocené např. pomocí matice záměn, jsou zachycené variace, na které model není připraven a lze tak objektivněji určit kvalitu modelu. Způsobů, jak datový soubor rozdělit, je mnoho. Existuje například k-fold validace, kdy se trénovací množina dat rozdělí na k náhodných podmnožin. Jedna podmnožina dat se použije pro validaci a zbylých $k-1$ podmnožin se použije pro trénování. Celý proces se opakuje k krát, tedy každá podmnožina bude právě jednou použita pro testování. Výsledné statistiky lze zprůměrovat a použít jako hodnocení daného modelu.

3.2.5 Matice záměn

Matice záměn je nástroj pro vyhodnocení predikcí modelu. Matice je o velikosti 2×2. Pro potřeby logistické regrese se matice skládá ze dvou řádku a dvou sloupců. Ve sloupcích se nachází původní hodnoty, tedy hodnoty, které chceme předpovídat. Ve řádcích se pak nachází předpovědi z modelu.

		Původní pozitivní	Původní negativní
		1	0
Pozitivní predikce	1	Skutečně pozitivní	Falešně pozitivní
Negativní predikce	0	Falešně negativní	Skutečně negativní

Tabulka 3.2: Matice záměn

Pro sestavení matice je potřeba množina dat, u kterých známe vysvětlovanou proměnnou. Na datech pak provedeme predikci, díky čemuž získáme predikované hodnoty. Porovnáním původních a predikovaných hodnot vznikne matice 3.2. Každá ze čtyř vnitřních buněk má vlastní označení a interpretaci:

- **Skutečně pozitivní** — počet správných predikcí, které byly rovné jedné.
- **Falešně pozitivní** — počet predikcí rovných jedné, kde byla původní hodnota rovná nule.
- **Skutečně negativní** — počet správných predikcí, které byly rovné nule.
- **Falešně negativní** — počet predikcí rovných nule, kde byla původní hodnota rovná jedné.

Z matice záměn lze následně vypočítat mnoho statistik. Pro vyhodnocení regresního modelu lze použít např. přesnost, která se vypočítá jako počet všech správných predikcí nad počtem všech provedených predikcí n.

$$\textit{Přesnost} = \frac{\text{Skutečně pozitivní} + \text{Skutečně negativní}}{n} \tag{3.11}$$

Přesnost říká, jaké procento objektů bylo klasifikováno správně. Pokud je ovšem poměr původních pozitivních a negativních hodnot velmi nevyrovnaný, tato statistika není vhodná. V případě velké nevyrovnanosti predikovaných hodnot přesnost zkresluje schopnost modelu predikovat méně zastoupenou predikovanou hodnotu. Toto se může stát například v lékařství při identifikaci nemocného pacienta. Zde většinou dochází k velkému nepoměru mezi počtem nemocných a počtem zdravých. V takovém případě lze použít statistiku zvanou senzitivita. Ta se rovná poměru správných pozitivních predikcí a úhrnu všech pozitivních predikcí, neboli

$$\textit{Senzitivita} = \frac{\text{Skutečně pozitivní}}{\text{Skutečně pozitivní} + \text{Falešně pozitivní}}. \tag{3.12}$$

Senzitivita tedy určuje poměr správně klasifikovaných pozitivních případů a všech pozitivně klasifikovaných případů. Pokud by bylo vhodné preferovat spíše negativní klasifikace, tedy zdravé pacienty, lze použit statistiku zvanou specificita a je definovaná jako

$$\textit{Specificita} = \frac{\text{Skutečně negativní}}{\text{Skutečně negativní} + \text{Falešně negativní}}. \tag{3.13}$$

3.2.6 Testování hypotéz

Cílem testovací hypotézy je zjistit, zda neznámý parametr θ patří do nějaké prostoru Ω_0. V případě testování parametrů logistického modelu bude neznámý parametr θ roven parametru β_i. Prostor Ω_0 je pak populace. Testuje se nulová hypotéza, značená H_0, která říká, že parametr θ do prostoru Ω_0 patří. Proti ní je postavená hypotéza alternativní, značená H_1, která tvrdí, že parametr θ do prostoru Ω_0 nepatří. Pro testování je nutné zvolit parametr α, který značí maximální hodnotu chyby prvního druhu. Chyba prvního druhu stanovuje, jaká je pravděpodobnost, že se zamítne testovaná nulová hypotéza H_0 za předpokladu, že je H_0 pravdivá (Härdle et al., 2015).

$$P(\text{Zamítnutní } H_0 | H_0 \text{ je platná}) = \alpha \tag{3.14}$$

Existuje také chyba druhého typu β, která značí pravděpodobnost nezamítnutí neplatné nulové hypotézy.

$$P(\text{Nezamítnutí } H_0 | H_0 \text{ je neplatná}) = \beta \tag{3.15}$$

Chyby nelze eliminovat a je nutné je s nimi při testování hypotézy počítat. Pro parametr α se ustálily hodnoty $0.1, 0.05$ a 0.01.

3.2.7 Waldův test

Koeficienty v logistickém regresním modelu nemusí být statisticky významné. Pokud je prediktor nevýznamný, znamená to, že není významný při predikci prediktoru. K otestování významnosti prediktoru lze použít Waldův test.

Waldův test ověřuje, zda je parametr β_i v populaci významný či nikoliv. Definice testu hypotézy je tedy:

H_0 : Koeficient β_i je rovný nule.
H_A : Koeficient β_i je různý od nuly.

Pro vyhodnocení hypotézy se používá

testové kritérium

Z, který se vypočítá jako poměr testovaného parametru β_i a směrodatné chyby koeficientu $S_{\hat{\beta}_i}$

$$Z = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}. \tag{3.16}$$

Testové kritérium Z má za platnosti nulové hypotézy normální $N(0, 1)$ a její druhá mocnina, Z^2, má chí-kvadrát rozdělení s jedním stupněm volnosti (Kleinbaum et al., 2010).

4. Teoretická část

Cílem praktické části je prozkoumat dva typy modelů. První typ modelu bude predikovat výhru zápasu pro jednotlivé hráče a druhý typ pro jednotlivé mapy. Hlavním cílem modelu pro hráče bude identifikovat významné charakteristiky hráčů na vybraných mapách, a následně rozdíly interpretovat. V druhém modelu bude zkoumaná výhra týmu v zápase. V tomto typu modelu se jako prediktory použijí agregované charakteristiky hráčů jednotlivých týmu. Pro agregaci bude použit buď aritmetický nebo geometrický průměr. Všechny modely jsou vyhodnoceny pomocí testovacího datového souboru a matice záměn. Pro tvorbu modelů a veškerou práci s daty je použit programovací jazyk R (Team, 2022) a vývojové prostředí RStudio (RStudio, 2022).

4.1 Cíle analýzy

Cílem analýzy je vytvořit logistické modely pro předpověď výhry jak pro individuální hráče, tak pro tým. Jelikož se charakteristiky hráčů i zápasů mohou měnit dle mapy, na které se zápas odehrává, modely pro individuální hráče jsou tříděné přes kategorie map. Modely pro předpověď výhry týmu používají interakci mezi mapou a začínající stranou.

Pro hráče budou sestaveny logistické modely z celého datového souboru a pro každou mapu zvlášť. Modely budou následně vyhodnoceny pomocí statistik vycházejících z matice záměn a bude porovnána významnost jejich parametrů. Modely budou porovnané na vybraných dvou mapách, a to na mapě Mirage a Vertigo. Mirage je klasický mapa, která je ve hře od jeho vydání. Vertigo je naopak nejnovější přírůstek do profesionální scény a hráči, v době extrakce datového souboru, mapu ještě plně strategicky neznali. Modely budou porovnány podle přesnosti predikce a podle toho, jaké prediktory jsou pro model významné.

Modely pro týmy budou sestavené nejprve z agregovaného datového souboru a následně pro dva vybrané referenční týmy. První tým, pro který se vytvoří model, bude tým Astralis. Ten je dlouhodobě považován za jeden z nejlepších týmu na světě a v grafu 4.3 je vidět, že má v datovém souboru největší míru výhry. Druhým týmem bude německý celek Sprout. Tým Sprout se řadí v době extrakce dat k profesionálnímu týmu s průměrným třicátým místem na světovém žebříčku. Modely jsou vytvořené pouze pro dva týmy z toho důvodu, že je celkový počet týmů velmi vysoký a bylo by komplikované porovnávat všechny týmy naráz. Z toho důvodu se budou hledat významné rozdíly mezi dvěma referenčními modely a jedním celkovým modelem, který je sestaven na celém datovém souboru. Mezi modely bude porovnaná jak přesnost predikce, tak významnost parametrů. Jelikož je původní spojený datový soubor na úrovni hráčů, charakteristiky hráčů se musí agregovat na úroveň týmů a zápasů.

4.2 Příprava dat

Dataset[1] obsahuje čtyři datové soubory, které popisují zápasy ve hře CSGO. K potřebám této bakalářské práce budou použity pouze soubory *players.csv* a *results.csv*. Datové soubory jsou následně spojeny do jednoho datového souboru, který obsahuje charakteristiky všech hráčů v právě jednom zápase, potřebné informace o zápase a výsledek (zda hráč zápas vyhrál či nikoliv). Zbylé dva soubory obsahují informace, které jsou již z probíhajících zápasů a z volby map. Tyto informace pro predikci výhry ještě před začátkem zápasu nelze využít. Žádný z těchto zbylých dvou souborů (*picks.csv*, *economy.csv*) proto v bakalářské práci použit není.

4.2.1 Soubor players.csv

Datový soubor *players.csv* obsahuje charakteristiky jednotlivých hráčů v daném zápase. Původní datový soubor obsahuje 101 proměnných a 383 317 pozorování. V původním datovém souboru se jeden řádek (pozorování) rovná charakteristikám jednoho hráče za celý zápas, který se může odehrávat až na třech mapách. Pro potřeby bakalářské práce je tak nutné získat charakteristiky hráčů za jednotlivé mapy. Proto je původní datový soubor transformován do podoby, kde se jedno pozorování rovná charakteristikám právě jednoho hráče na právě jedné mapě, a to bez ohledu na to, kolik map se v daném zápase hrálo. Jinak řečeno, transformovaný datový soubor nebere v potaz, zda se daná mapa hrála jako první, druhá, či třetí.

Z datového souboru jsou odstraněné záznamy o mapě Default. Ta značí automatickou výhru pro jeden team, například díky formátu turnaje. V souboru je také možné narazit na tým, pro který za zápas na jedné mapě nehrálo právě 5 hráčů. Více než pět hráčů mohlo na mapě hrát v případě použitého náhradníka. Může se také stát, že hraje méně hráčů, pokud se jeden například nestihne dostavit. V obou případech jsou data pro danou mapu v zápase odstraněná.

[1] https://www.kaggle.com/datasets/mateusdmachado/csgo-professional-matches

Transformovaný a očištěný datový soubor má 10 proměnných a 640 225 pozorování. Příklad jednotlivých pozorování v transformovaném datovém souboru je v přiložené tabulce A.1. Interpretace charakteristik je následující:

- **match_id** — identifikátor zápasu
- **player_id** — identifikátor hráče
- **team** — jméno týmu
- **map** — název hrané mapy
- **kills** — počet zabití hráče v zápase na dané mapě
- **assists** — počet asistencí hráče v zápase na dané mapě
- **deaths** — počet smrtí hráče v zápase na dané mapě
- **hs** — procento zabití, které lze označit jako headshot[2]
- **fkdiff** — rozdíl, kolikrát hráč zabil jako první nepřítele versus kolikrát byl zabit jako první
- **rating** — shrnutí více charakteristik za jeden zápas do jednoho ukazatele výkonu[3]

4.2.2 Soubor results.csv

Druhý datový soubor, který je pro analýzu použit, obsahuje výsledky daných zápasů. Soubor se původně skládá z 19 proměnných a 45 773 pozorování. Datový soubor *results.csv* obsahuje na rozdíl od datového souboru *players.csv* jedno chybné pozorování. Dle něho hrál tým sám proti sobě, což nedává věcný a logický smysl. Jelikož je zápas na webovém portálu zadán správně, nejspíše se jedná o neznámou chybu, která nastala při exportu dat z webového portálu. Zároveň jsou vybrané pouze proměnné, které pro predikci lze využít. Dále je zde rozdělený tým Astralis do dvou týmů: týmu Astralis a tým „?“. Důvod rozdělení je ten, že historicky hrál tým bez organizace a označoval se jako „?“. Po vytvoření organizace Astralis jsou hráči řazeni pod tuto organizaci. Jelikož se tímto týmem zabývá specializovaný logistický model, je označení týmu sjednoceno pod název Astralis.

Po transformacích vznikne tabulka o 7 proměnných a 91 502 pozorování. Každé pozorování identifikuje výsledek právě jednoho týmu v jednom zápase na jedné mapě. Příklad je zobrazen v přiložené tabulce A.2. Jednotlivé proměnné lze interpretovat následovně:

- **date** — datum, kdy se hrál zápas
- **match_id** — identifikátor zápasu
- **team** — jméno týmu
- **map** — název hrané mapy
- **map_winner** — binární značení, zda tým vyhrál (1) či prohrál (0)
- **starting_ct** — binární značení, zda tým začal zápas na straně Counter-Teroristů (1) či Teroristů (0)
- **team_rank** — rank týmu v okamžik, kdy se zápas hrál[4]

[2] hráč zabil nepřítele střelou do hlavy
[3] https://www.hltv.org/news/20695/introducing-rating-20
[4] https://www.hltv.org/news/16061/introducing-csgo-team-ranking

4.2.3 Datový soubor pro modelování

Pro modelování je nutné vytvořit jeden datový soubor, na kterém bude model sestaven. Datový model je získán spojením představených souborů *players.csv* a *results.csv*. Datové soubory jsou spojené pomocí proměnných *match_id*, *team* a *map*. Při propojování souborů jsou smazané záznamy, které nejsou reprezentovány v obou souborech. To znamená, že charakteristiky hráčů v zápase, pro který není výsledek, jsou smazané. Stejně jsou smazané zápasy, pro které nejsou zadané charakteristiky hráčů. Spojený datový soubor je pak na úrovni jednotlivých hráčů a ukazuje jejich charakteristiky v právě jednom zápase na právě jedné mapě. Příklad spojeného datového souboru je v přiložené tabulce A.3.

4.2.4 Agregovaný datový soubor

Spojený datový soubor obsahuje charakteristiky hráčů na úrovni jednotlivých zápasů. To je vhodné pro první logistický model, který se zabývá predikcí a identifikací významných prediktorů pro individuální hráče. Pro predikci výhry týmu a identifikaci významných prediktorů pro tým je nutné data agregovat. Agregace charakteristik, u kterých to dává věcný smysl, je použit aritmetický průměr. Pro statistiky, u kterých dává věcný smysl násobení, je použit průměr geometrický. Jako příklad lze uvést charakteristika počtu zabití (prediktor *kills*), která je agregovaná pro daný tým jako průměrný počet zabití (prediktor *mean_kills*) hráčů týmu na dané mapě. Pro charakteristiku procent zabití do hlavy (prediktor *hs*) je použit geometrický průměr a v modelu je použitý průměr procenta zabití do hlavy (prediktor *mean_hs*). Příklad agregovaných dat je v přiložené tabulce A.4. Popis agregovaných prediktorů je pak v následující tabulce.

Tabulka 4.1: Přehled agregací charakteristik pro daný tým na dané mapě v daném zápase

Prediktor	Agregace	Popis
mean_kills	Aritmetický průměr	Průměrný počet zabití
mean_assists	Aritmetický průměr	Průměrný počet asistencí
mean_deaths	Aritmetický průměr	Průměrný počet smrtí
mean_hs	Geometrický průměr	Průměrné procento zabití do hlavy
mean_fkdiff	Aritmetický průměr	Průměrný rozdíl mezi prvním zabitím a první smrtí

4.2.5 Trénování a validace modelů

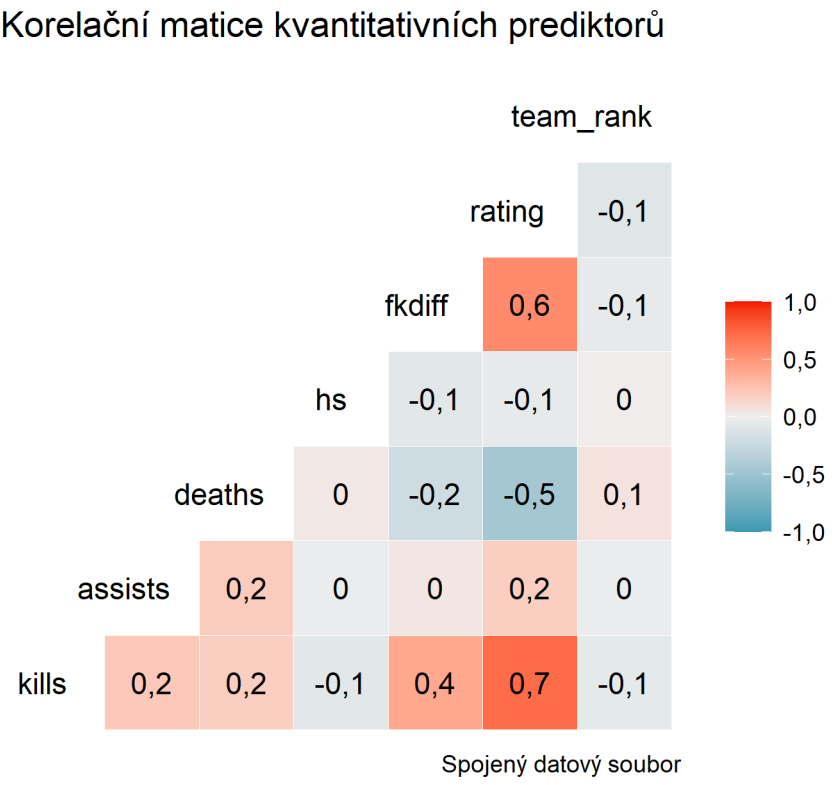
Pro trénování a validaci modelů se vytvoří náhodné rozdělení souboru v poměru 8:2. Pro trénování modelu bude použito 80 % z datového souboru, na validaci je použito zbylých 20 %. Natrénované modely jsou vyhodnoceny pomocí Waldova testu, který je v každé tabulce v posledním sloupci $Pr(>|z|)$. Případné nevýznamné prediktory jsou pak z modelu odstraněny a model je natrénován znovu na stejné trénovací množině. Pro validaci modelu je vytvořena matice záměn, která popisuje přesnost predikcí modelu na validačních datech. K následnému vyhodnocení modelů jsou použity statistiky Přesnost, Senzitivita a Specificita.

4.3 Průzkumová analýza dat

Průzkumová analýza vizualizuje prediktory, hledá různé vztahy a rozdělení proměnných. Díky průzkumu lze určit, které proměnné není vhodné použít pro tvorbu logistického regresního modelu, např. kvůli problému multikolinearity.

4.3.1 Korelační matice

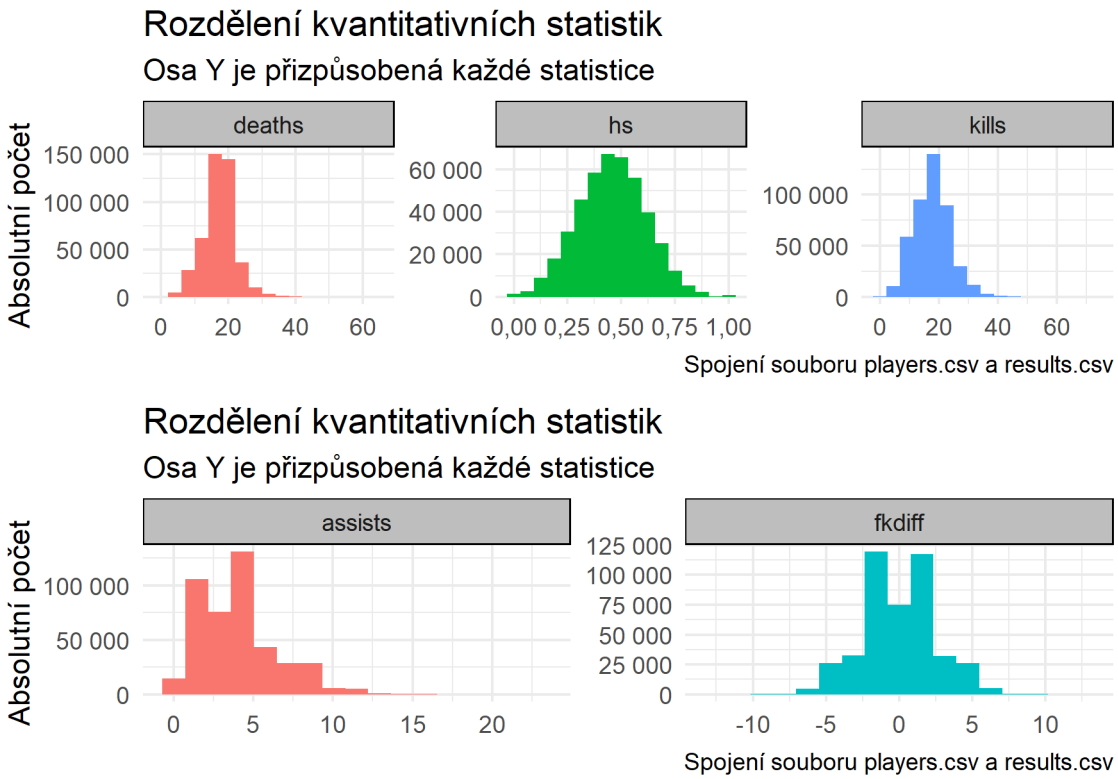
Pro logistickou regresi je důležité, aby prediktory nebyly lineárně závislé. Přehled korelací mezi kvantitativními prediktory lze zjistit z korelační matice.



Obrázek 4.1: Korelační matice

Z korelační matice 4.1 lze vyčíst, že korelace mezi rankem týmu a charakteristikami hráčů se blíží nule. Z toho plyne, že neexistuje lineární závislost mezi výkonem hráče a rankem týmu. Zároveň je vidět silná korelace mezi prediktorem *rating* a prediktory *fkdiff*, *deaths* a *kills*. Jelikož by díky vysoké korelaci prediktorů vznikl problém multikolinearity, prediktor *rating* ve finálních modelech není použit.

4.3.2 Histogramy kvantitativních prediktorů

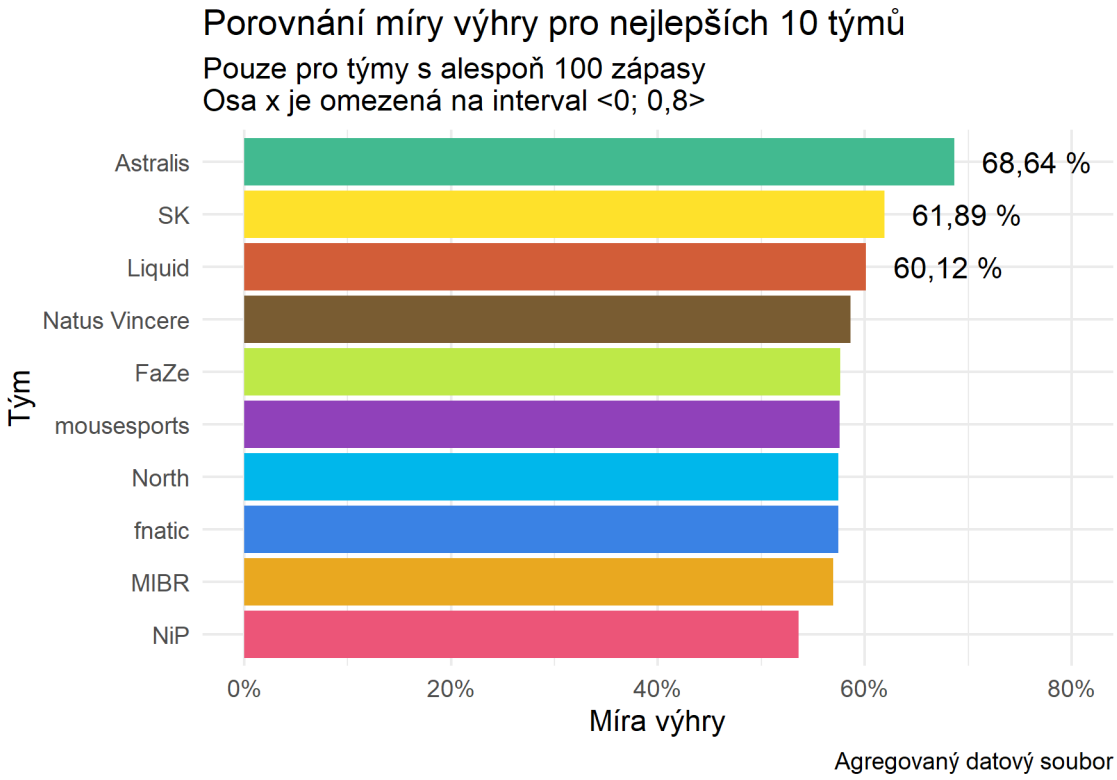


Obrázek 4.2: Histogram prediktorů

Histogramy prediktorů z obrázku 4.2 ukazují, že prediktory *kills*, *deaths* a *hs* mají normální rozdělení a v prediktorech se nenachází mnoho extrémních hodnot. Prediktor *fkdiff* má bimodální rozdělení. Prediktor *assists* je zešikmení doprava, což značí velké množství odlehlých či extrémních hodnot. Pro logistickou regresi není předpoklad normálního rozdělení prediktorů, a proto analýza slouží čistě k získání povědomí o tom, jakých hodnot každý prediktor nabývá a jaké je jejich rozdělení.

4.3.3 Míra výhry pro nejlepší týmy

Pro výběr do logistického modelu je nutné vybrat referenční tým, který je považován za jeden z nejlepších.



Obrázek 4.3: Procento vyhraných zápasů na dané mapě za stranu Counter-Teroristů

Obrázek 4.3 ukazuje, že tým Astralis má v datovém souboru nejlepší míru výhry. Rozdíl mezi mírou výhry Astralis a týmy SK je necelých 7 procent. Z tohoto důvodu je tým Astralis použit při analýze dat jako jeden z referenčních týmů.

4.4 Predikce výhry hráčů na různých mapách

Cílem modelu je predikovat výhru zápasu pro jednotlivé hráče na různých mapách a identifikovat významné prediktory s nimi spojené. Model lze využít na profesionální úrovni k identifikaci toho, jaká charakteristika nejvíce přispívá k výhře či prohře hráče. Prediktory se týkají pouze výkonu jednotlivých hráčů, model tedy pro předpověď výhry hráče nepoužívá charakteristiky spoluhráčů. K modelování je použit spojený datový soubor, který je detailně popsaný v sekci **??**. Obecně by šlo model zapsat následovně:

$$
\begin{aligned}
&P(1|X_{kills}, X_{assists}, X_{deaths}, X_{fkdiff}, X_{starting_{ct}}) = \frac{1}{1+e^{-z}} \\
&z = \beta_0 + \beta_1 * X_{kills} + \beta_2 * X_{assists} + \beta_3 * X_{deaths} + \\
&+ \beta_4 * X_{fkdiff} + \beta_5 * X_{starting_{ct}}.
\end{aligned}
$$

Pro porovnání jsou vybrány mapy Mirage a Vertigo. Mapa Mirage je jednou z nejvíce tradičních map a mapa Vertigo je naopak nejnovější přídavek do hry. Díky rozdílným modelům bude možné zkoumat, na čem pravděpodobnost výhry na mapách záleží a jaké strategie je na dané mapě výhodnější použít.

4.4.1 Model pro mapu Mirage

Tabulka 4.2: Výstup z programu R pro logistický model na mapě Mirage

Estimate	Std. Error	z value	Pr(>|z|)
2,3954	0,0599	39,9740	0,0000
0,1807	0,0021	87,4819	0,0000
0,3031	0,0047	64,8606	0,0000
-0,3661	0,0032	-115,9604	0,0000
-0,1957	0,0634	-3,0854	0,0020
0,0171	0,0046	3,6881	0,0002
-0,1995	0,0199	-10,0442	0,0000

Tabulka 4.2 představuje model se všemi významnými prediktory. Prediktory *kills*, *assists* a *fkdiff* šanci na výhru hráče zvyšují. Naopak prediktory *deaths* a *starting_ct* šanci snižují. Model lze také zapsat jako přepis funkce.

$$
\begin{aligned}
&P(1|X_{kills}, X_{assists}, X_{deaths}, X_{fkdiff}, X_{starting_{ct}}) = \frac{1}{1+e^{-z}} \\
&z = 2{,}3954 + 0{,}1807 * X_{kills} + 0{,}3031 * X_{assists} - 0{,}3661 * X_{deaths} - \\
&- 0{,}1957 * X_{hs} + 0{,}0171 * X_{fkdiff} - 0{,}1995 * X_{starting_{ct}}
\end{aligned}
\tag{4.1}
$$

Ze získaných koeficientů lze získat změnu šance. Změnu šance s každým dalším zabitím lze získat umocnění parametru $\beta_1 = 0{,}1807$ na konstantu e.

$$\begin{aligned} \textit{Změna šance} &\sim e^{\beta_1} \\ &\sim e^{0{,}1807} \\ &\sim 1.1980 \end{aligned} \tag{4.2}$$

S každým dalším zabitým hráčem se zvyšuje šance na výhru hráče zhruba 1,2 krát. U prediktoru *deaths* je parametr β_3 záporný. To znamená, že se šance na výhru snižuje.

$$\begin{aligned} \textit{Změna šance} &\sim e^{\beta_3} \\ &\sim e^{-0{,}3661} \\ &\sim 0.6934 \end{aligned} \tag{4.3}$$

Parametr lze interpretovat tak, že s každou další hráčovou smrtí se jeho šance na výhru sníží o zhruba 31 %.

Matice záměn pro mapu Mirage

Na predikci je použitá validační podmnožina a model z tabulky 4.2.

Tabulka 4.3: Vybrané statistiky z matice záměn pro mapu Mirage

Původní pozitivní	Původní negativní
7037	1938
1618	6804

Tabulka 4.4: Statistiky z matice záměn pro mapu Mirage

statistika	hodnota
Přesnost	0,7956
Senzitivita	0,7783
Specificita	0,8131

Z tabulky 4.3 je vidět, že model predikoval správně 7 037 výher ($\sim$ 81,31%) a 6 804 proher ($\sim$ 77,83%). Celkově model určil správně 13 841 objektů ($\sim$ 79,56%).

4.4.2 Model pro mapu Vertigo

Model pro mapu Vertigo má identické vstupní prediktory, podle kterých se predikuje proměnná *map_winner*, jako model pro mapu Mirage.

Tabulka 4.5: Výstup z programu R pro logistický model na mapě Vertigo

Estimate	Std. Error	z value	Pr(>|z|)
1,8170	0,2199	8,2645	0,0000
0,1826	0,0081	22,6135	0,0000
0,2983	0,0174	17,1061	0,0000
-0,3508	0,0120	-29,2434	0,0000
0,2077	0,2500	0,8306	0,4062
0,0092	0,0180	0,5099	0,6101
-0,0021	0,0773	-0,0265	0,9788

Na rozdíl od mapy Mirage jsou v modelu pro mapu Vertigo nevýznamné parametry. Statisticky nevýznamné prediktory *hs*, *fkdiff* a *starting_ct* jsou z modelu odebrány a model je znovu natrénován.

Tabulka 4.6: Výstup z programu R pro optimalizovaný logistický model na mapě Vertigo

Estimate	Std. Error	z value	Pr(>|z|)
1,9090	0,1803	10,5872	0,0000
0,1841	0,0074	24,8957	0,0000
0,2981	0,0174	17,1172	0,0000
-0,3521	0,0116	-30,2299	0,0000

Pro hráče jsou na mapě Vertigo významné pouze prediktory *kills*, *assists* a *deaths*. Ostatní prediktory *hs*, *fkdiff* a *starting_ct* jsou pro model nevýznamné. Optimalizovaný model lze zapsat do následující rovnice:

$$\begin{aligned} P(1|X_{kills}, X_{assists}, X_{deaths}) &= \frac{1}{1+e^{-z}} \\ z &= 1{,}9090 + 0{,}1841 * X_{kills} + 0{,}2981 * X_{assists} - 0{,}3521 * X_{deaths} \end{aligned} \tag{4.4}$$

Každé hráčovo zabití vede ke zvýšení šance na výhru o zhruba 20 %. Každá další hráčova smrt vede ke snížení šance na výhru o zhruba 30 %. Matice záměn i statistiky modelu jsou potom vyhodnocené na optimalizovaném modelu pomocí validačních dat.

Tabulka 4.7: Vybrané statistiky z matice záměn pro mapu Vertigo

Původní pozitivní	Původní negativní
455	113
131	411

Tabulka 4.8: Statistiky z matice záměn pro mapu Vertigo

statistika	hodnota
Přesnost	0,7802
Senzitivita	0,7844
Specificita	0,7765

Z tabulky 4.8 je vidět, že model predikoval správně 455 výher ($\sim$ 77,65%) a 411 proher ($\sim$ 78,44%). Celkově model určil správně 866 objektů ($\sim$ 78,02%).

4.4.3 Interpretace a porovnání modelů

Na mapě Mirage jsou pro hráče významné všechny prediktory. Nejvíc pozitivní vliv má na výhru hráče počet asistencí a největší negativní vliv na výhru hráče má pak počet smrtí. Významnost prediktorů *fkdiff* a *assists* naznačuje, že je důležitá souhra hráčů a zkušenější týmy mají na mapě výhodu. Pokud hráči hrají spolu, mohou se při zabíjení nepřátel doplňovat (prediktor *assists*). Zároveň je důležité, aby se v zápase hráči navzájem podporovali a mohli získat první zabití v daném kole (prediktor *fkdiff*).

Pro mapu Vertigo jsou naopak významné pouze prediktory *kills*, *assists* a *deaths*. Pro hráče tedy není důležité, jak přesně střílí (prediktor *hs*), jak dobře se tým podporuje na začátku kola (prediktor *fkdiff*) ani na jaké straně hráč začíná (prediktor *starting_ct*). Největší vliv na výhru zde má prediktor *assists*, největší vliv na výhru má pak prediktor *deaths*. Porovnání obou statistik obou modelů je pak v následující tabulce.

Tabulka 4.9: Porovnání statistik pro mapu Mirage a Vertigo

statistika	mapa_mirage	mapa_vertigo
Přesnost	0,7956	0,7802
Senzitivita	0,7783	0,7844
Specificita	0,8131	0,7765

Mapa Mirage se hodí nejvíce na predikci proher díky své vysoké specificitě. Rozdíly mezi statistikami pro mapu Vertigo jsou pak nepatrné a menší než 1 %. Největší rozdíl mezi modely je ve Specificitě, která je pro model na mapě Mirage větší o zhruba 3,5 procentních bodů.

Modely, matice záměn i vybrané statistiky z matic záměn pro ostatní mapy jsou v příloze B. Zajímavé je, že pro všechny ostatní modely jsou významné všechny prediktory. To z mapy Vertigo, která má významné pouze 3 prediktory, dělá poněkud unikátní mapu. Z tohoto závěru lze usoudit, že pro všechny mapy, které se hrají už nějakou dobu, je důležitý jak výkon hráče, tak sehranost týmu. Jelikož je mapa Vertigo nová, je na začátku svého vývoje a pro hráče je stále relativně neznámá. Jelikož je na mapě Vertigo méně významná sehranost, zkušenosti a spolupráce v týmu, měli by jí preferovat nové či semi-profesionální týmy. Mapu také mohou využít profesionální týmy, které spoléhají spíše na individuální výkon hráčů, než na týmové strategie.

4.5 Predikce výhry týmu

Cílem modelů je predikovat výhru na základě agregovaných charakteristik hráčů za tým na mapě. Tento typ modelu se hodí zejména v sázkových kancelářích. Díky referenčním modelům lze určit šanci, že určitý tým vyhraje. To umožní kancelářím stanovit výdělečný kurz pro zápas ještě před tím, než vůbec začal. Matematický přepis modelu je následující:

$$
\begin{aligned}
P(1|&X_{mean\_kills}, X_{mean\_assists}, X_{mean\_deaths}, X_{mean\_hs}, X_{mean\_fkdiff}, X_{team\_rank},\\
&X_{mapCache*starting\_ct}, X_{mapCobblestone*starting\_ct}, X_{mapDust2*starting\_ct}, X_{mapInferno*starting\_ct},\\
&X_{mapMirage*starting\_ct}, X_{mapNuke*starting\_ct}, X_{mapOverpass*starting\_ct}, X_{mapTrain*starting\_ct},\\
&X_{mapVertigo*starting\_ct}) = \frac{1}{1+e^{-z}}\\
z =&\beta_0 + \beta_1 * X_{mean\_kills} + \beta_2 * X_{mean\_assists} + \beta_3 * X_{mean\_deaths} +\\
&+ \beta_4 * X_{mean\_hs} + \beta_5 * X_{mean\_fkdiff} + \beta_6 * X_{team\_rank} +\\
&+ \beta_7 * X_{mapCache*starting\_ct} + \beta_8 * X_{mapCobblestone*starting\_ct} +\\
&+ \beta_9 * X_{mapDust2*starting\_ct} + \beta_{10} * X_{mapInferno*starting\_ct} +\\
&+ \beta_{11} * X_{mapMirage*starting\_ct} + \beta_{12} * X_{mapNuke*starting\_ct} +\\
&+ \beta_{13} * X_{mapOverpass*starting\_ct} + \beta_{14} * X_{mapTrain*starting\_ct} +\\
&+ \beta_{15} * X_{mapVertigo*starting\_ct}
\end{aligned}
\tag{4.5}
$$

Prediktory *mean_kills*, *mean_assists*, *mean_deaths*, *mean_hs*, *mean_fkdiff* jsou průměrné charakteristiky hráčů v týmu na dané mapě v daném zápase a lze očekávat, že šanci na výhru ovlivňují. Dále do modelu vstupuje prediktor *team_rank*, který značí rank týmu v daném zápase. Model následně obsahuje interakci mezi proměnnou *map* a *starting_ct*. Interakce je z toho důvodu, že vliv počáteční strany je na každé mapě jiný. Přesněji jsou prediktory definované v sekci 4.2.4

4.5.1 Celkový model

Tabulka 4.10: Výstup z programu R pro logistický model pro všechny týmy

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0,4886	0,1651	2,9598	0,0031
mean_kills	1,3729	0,0145	94,7439	0,0000
mean_assists	0,1507	0,0161	9,3446	0,0000
mean_deaths	-1,4106	0,0143	-98,4641	0,0000
mean_hs	-0,4379	0,2513	-1,7424	0,0814
mean_fkdiff	-0,0807	0,0197	-4,1010	0,0000
team_rank	-0,0011	0,0003	-3,6420	0,0003
starting_ct:mapCache	-0,0032	0,0862	-0,0370	0,9705
starting_ct:mapCobblestone	-0,2966	0,0979	-3,0291	0,0025
starting_ct:mapDust2	-0,3174	0,0903	-3,5166	0,0004
starting_ct:mapInferno	-0,3189	0,0706	-4,5185	0,0000
starting_ct:mapMirage	-0,2587	0,0637	-4,0615	0,0000
starting_ct:mapNuke	-0,1961	0,0905	-2,1662	0,0303
starting_ct:mapOverpass	-0,5027	0,0800	-6,2806	0,0000
starting_ct:mapTrain	-0,2420	0,0746	-3,2456	0,0012
starting_ct:mapVertigo	0,1061	0,2213	0,4794	0,6317

Parametry celkového modelu se nachází v tabulce 4.10. První vytvořený model je sestavený na celém trénovacím datovém souboru. Pro model jsou významné všechny prediktory bez interakce, jmenovitě *mean_kills*, *mean_assists*, *mean_deaths*, *mean_hs*,*mean_fkdiff* a *team_rank*. Interakce mezi prediktory *map* a *starting_ct* není významná u map Vertigo, Cache a Nuke.

Agregované charakteristiky hráče *mean_kills* a *mean_assists* šanci na výhru týmu zvyšují. Pokud se průměr zabitých nepřátel za tým zvýší o jednotku, šance na výhru týmu se zvýší zhruba 3,96 krát. Pokud se průměr smrtí hráčů za tým zvýší o jednotku, šance na výhru se sníží zhruba 0,25 krát. Všechny statisticky významné interakce mezi prediktory *map* a *starting_ct* naznačují, že je pro tým nevýhodné začínat na straně Counter-Teroristů. Jejich šance na výhru se vždy sníží, a to nejvíce na mapě Overpass, kde se šance sníží zhruba 0,62 krát. Zajímavý je koeficient u prediktoru *team_rank*, který říká, že s růstem ranku týmu se šance na výhru sníží zhruba 0,99 krát. To lze vysvětlit tím, že lepší týmy hrají proti lepším týmům a jejich šance na výhru je nižší. Model lze zapsat jako přepis funkce následovně:

$$
\begin{aligned}
P(1|&X_{mean\_kills}, X_{mean\_assists}, X_{mean\_deaths}, X_{mean\_hs}, X_{mean\_fkdiff}, X_{team\_rank},\\
&X_{mapCache*starting\_ct}, X_{mapCobblestone*starting\_ct}, X_{mapDust2*starting\_ct}, X_{mapInferno*starting\_ct},\\
&X_{mapMirage*starting\_ct}, X_{mapNuke*starting\_ct}, X_{mapOverpass*starting\_ct}, X_{mapTrain*starting\_ct},\\
&X_{mapVertigo*starting\_ct}) = \frac{1}{1+e^{-z}}\\
z =&0{,}4135 + 1{,}376 * X_{mean\_kills} + 0{,}143 * X_{mean\_assists} - 1{,}402 * X_{mean\_deaths} -\\
&- 0{,}615 * X_{mean\_hs} - 0{,}064 * X_{mean\_fkdiff} - 0{,}001 * X_{team\_rank} -\\
&- 0{,}040 * X_{mapCache*starting\_ct} - 0{,}282 * X_{mapCobblestone*starting\_ct} -\\
&- 0{,}318 * X_{mapDust2*starting\_ct} - 0{,}272 * X_{mapInferno*starting\_ct} -\\
&- 0{,}239 * X_{mapMirage*starting\_ct} - 0{,}113 * X_{mapNuke*starting\_ct} -\\
&- 0{,}418 * X_{mapOverpass*starting\_ct} - 0{,}177 * X_{mapTrain*starting\_ct} +\\
&+ 0{,}259 * X_{mapVertigo*starting\_ct}
\end{aligned}
\tag{4.6}
$$

5. Závěr

Bakalářská práce se zabírala predikcí výher zápasů ve hře CSGO pro hráče i pro týmy. Významnost modelů byla zjištěna pomocí Waldova testu a hladiny významnosti $\alpha = 0.05$. Predikce modelů byla vyhodnocena pomocí matice záměn a statistik Přesnost, Senzitivita a Specificita.

Pro práci s modely bylo prve nutné data spojit do jednoho datového souboru. Modely byly vytvořené pomocí trénovací množiny dat, která činila 80% ze spojeného datového souboru. Validační množina pak tvořila zbylých 20% dat. Ta byla použita k tvorbě matice záměn.

Modely pro hráče byly vytvořené přes všechny kategorie map a modely pro mapu Mirage a Vertigo mezi sebou byli porovnané. Modely se lišily hlavně tím, zda se hodí pro predikci výher či proher. Model pro mapu Mirage je díky své vyšší specificitě vhodný pro identifikaci proher. Model pro mapu Vertigo má naopak vyšší senzitivitu a hodí se spíše pro predikci výhry hráče. Každý model má také jiné významné prediktory, což značí, že každá se specifikuje jiným stylem hraním.

Modely pro celé týmy byly vytvořeny tři. Celkový model měl významnou většinu prediktorů. Jeho přesnost, senzitivita i specificita jsou velmi podobné a model díky tomu predikuje stejně úspěšně jak výhry, tak prohry. Model pro referenční tým Astralis má významné pouze dva prediktory *mean_kills* a *mean_deaths*. Statistiky z matice záměn naznačují, že se model hodí spíše k predikci výher. Model referenčního týmu Sprout má významné statistiky také pouze *mean_kills* a *mean_deaths*. Z matice záměn lze usoudit, že se model hodí spíše k predikci výher týmu.

Seznam použitého softwaru

RSTUDIO, 2022. *RStudio | Open source & professional software for data science teams* [online]. Osobní počítač: RStudio. 2022.02.2+485 [cit. 2022-04-30]. Dostupné z: https://www.rstudio.com/.

TEAM, R Core, 2022. *R: The R Project for Statistical Computing* [online]. Osobní počítač: R Project. Ver. 4.2.0 [cit. 2022-04-30]. Dostupné z: https://www.r-project.org/.

Seznam použité literatury

CSKO.CS, 2022. *Kotelna - Contents* [online] [cit. 2022-04-24]. Dostupné z: https://stats.csko.cz/statsx/hlstats.php.

GOUGH, Christina, 2022. *Global eSports market revenue 2024* [Statista] [online] [cit. 2022-04-24]. Dostupné z: https://www.statista.com/statistics/490522/global-esports-market-revenue/.

HÄRDLE, Wolfgang; SIMAR, Léopold, 2015. *Applied multivariate statistical analysis*. 4. ed. Heidelberg Berlin: Springer. ISBN 978-3-662-45170-0.

HENNINGSON, Joakim, 2020. *The history of Counter-Strike* [Red Bull] [online] [cit. 2022-04-24]. Dostupné z: https://www.redbull.com/se-en/history-of-counterstrike.

KLEINBAUM, David G.; KLEIN, Mitchel, 2010. *Logistic regression: a self-learning text*. Third Edition. Ve spol. s RIHL PRYOR, Erica. New York Dordrecht Heidelberg London: Springer. Statistics for Biology and Health. ISBN 978-1-4939-3697-7.

LARCH, Florian, 2022. *History of eSports: How it all began* [online] [cit. 2022-04-24]. Dostupné z: https://www.ispo.com/en/markets/history-origin-esports.

LIQUIPEDIA.NET, 2021. *PGL Major Stockholm 2021* [Liquipedia Counter-Strike Wiki] [online] [cit. 2022-04-24]. Dostupné z: https://liquipedia.net/counterstrike/PGL/2021/Stockholm.

PROFESSEUR, 2015. *HLTV.org - The home of competitive Counter-Strike* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1.

PROFESSEUR, 2022. *ESEA increase prize pool and number of seasons for 2021; simplify path to Pro League* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: https://www.hltv.org/news/30926/esea-increase-prize-pool-and-number-of-seasons-for-2021-simplify-path-to-pro-league.

RSTUDIO, 2022. *RStudio | Open source & professional software for data science teams* [online]. Osobní počítač: RStudio. 2022.02.2+485 [cit. 2022-04-30]. Dostupné z: https://www.rstudio.com/.

TEAM, R Core, 2022. *R: The R Project for Statistical Computing* [online]. Osobní počítač: R Project. Ver. 4.2.0 [cit. 2022-04-30]. Dostupné z: https://www.r-project.org/.

VALVE, 2013. *Counterstrike: Global Offensive - Arms Deal* [online] [cit. 2022-04-24]. Dostupné z: http://counter-strike.net/armsdeal.

Seznam elektronických zdrojů

CSKO.CS, 2022. *Kotelna - Contents* [online] [cit. 2022-04-24]. Dostupné z: https://stats.csko.cz/statsx/hlstats.php.

GOUGH, Christina, 2022. *Global eSports market revenue 2024* [Statista] [online] [cit. 2022-04-24]. Dostupné z: https://www.statista.com/statistics/490522/global-esports-market-revenue/.

HENNINGSON, Joakim, 2020. *The history of Counter-Strike* [Red Bull] [online] [cit. 2022-04-24]. Dostupné z: https://www.redbull.com/se-en/history-of-counterstrike.

LARCH, Florian, 2022. *History of eSports: How it all began* [online] [cit. 2022-04-24]. Dostupné z: https://www.ispo.com/en/markets/history-origin-esports.

LIQUIPEDIA.NET, 2021. *PGL Major Stockholm 2021* [Liquipedia Counter-Strike Wiki] [online] [cit. 2022-04-24]. Dostupné z: https://liquipedia.net/counterstrike/PGL/2021/Stockholm.

PROFESSEUR, 2015. *HLTV.org - The home of competitive Counter-Strike* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1.

PROFESSEUR, 2022. *ESEA increase prize pool and number of seasons for 2021; simplify path to Pro League* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: https://www.hltv.org/news/30926/esea-increase-prize-pool-and-number-of-seasons-for-2021-simplify-path-to-pro-league.

VALVE, 2013. *Counterstrike: Global Offensive - Arms Deal* [online] [cit. 2022-04-24]. Dostupné z: http://counter-strike.net/armsdeal.

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

CSGO Coutner-Strike: Global Offensive
BR Battle Royale
MOBA Multiplayer Online Battle Arena
FPS First-Person Shooter
TGNS Twin Galaxies National Scoreboard

Část I

Přílohy

A. Datové soubory

A.1 Transformovaný datový soubor players.csv

Tabulka A.1: Záznam z transformovaného datového souboru players.csv

match_id	player_id	team	map	kills	assists	deaths	hs	fkdiff	rating
2309869	9859	EURONICS	Mirage	13	4	19	0,5385	-2	0,7800
2322900	15631	FURIA	Nuke	18	1	14	0,5000	2	1,1400
2304505	7205	DarkPassage	Cache	15	3	21	0,2667	-1	0,8400
2301029	3997	PENTA	Cache	21	3	17	0,4762	1	1,2800
2333140	10264	OpTic	Dust2	5	1	16	0,4000	0	0,3800
2320084	9032	Astralis	Overpass	16	9	21	0,5625	-1	1,0600

A.2 Transformovaný datový soubor results.csv

Tabulka A.2: Příklad záznamu z transformovaného datového souboru results.csv

date	match_id	team	map	map_winner	starting_ct	team_rank
2017-03-29	2309175	New4	Inferno	0	0	123
2017-09-09	2314493	FaZe	Inferno	1	1	8
2019-04-22	2332676	Epsilon	Mirage	1	0	22
2019-07-29	2335195	North	Nuke	1	1	12
2019-10-30	2337316	Windigo	Vertigo	0	1	40
2019-10-24	2337128	Evil Geniuses	Dust2	1	1	3

A.3 Spojený datový soubor

Tabulka A.3: Příklad záznamu z transformovaného datového souboru results.csv

date	match_id	team	map	map_winner	starting_ct	team_rank	player_id	k
2017-05-17	2310786	ENCE	Cache	0	1	214	684	
2016-07-24	2303591	Chiefs	Dust2	0	0	88	9636	
2019-03-23	2331595	Liquid	Overpass	1	0	2	8738	
2019-12-22	2338373	Vitality	Mirage	1	1	8	8184	
2019-12-09	2338387	OFFSET	Inferno	1	0	73	11205	
2020-01-24	2339089	AUGUST	Dust2	0	0	114	10870	

A.4 Agregovaný datový soubor

Tabulka A.4: Agregovaná data pro týmy za zápas a mapu

match_id	map	team	mean_kills	mean_assists	mean_deaths	mean_hs	mean_fkdiff	map_winner	starting_ct	team_rank
2302745	Train	Virtus.pro	10,8000	2,0000	19,8000	0,3860	-2,0000	0,0000	0,0000	6,0000
2306295	Train	Natus Vincere	14,6000	2,2000	18,4000	0,3823	-0,8000	0,0000	1,0000	4,0000
2338225	Dust2	PC419	13,0000	2,4000	19,8000	0,4057	0,0000	0,0000	0,0000	172,0000
2334630	Dust2	Nordavind	16,4000	5,0000	20,4000	0,5492	0,6000	0,0000	1,0000	54,0000
2313618	Train	Adaptation	14,8000	4,0000	19,4000	0,4322	-1,4000	0,0000	0,0000	121,0000
2333754	Vertigo	Singularity	17,4000	4,8000	7,0000	0,4448	2,0000	1,0000	1,0000	52,0000

B. Modely, matice záměn a statistiky pro individuální hráče

B.1 Mapa Cache

Tabulka B.1: Výstup z programu R pro logistický model na mapě Cache

Estimate	Std. Error	z value	Pr(>|z|)
2,4062	0,0662	36,3650	0,0000
0,1852	0,0023	78,9265	0,0000
0,2821	0,0047	60,1204	0,0000
-0,3798	0,0036	-106,1746	0,0000
-0,2166	0,0703	-3,0802	0,0021
0,0239	0,0052	4,6146	0,0000
-0,2021	0,0220	-9,1647	0,0000

Tabulka B.2: Vybrané statistiky z matice záměn pro mapu Cache

Původní pozitivní	Původní negativní
5827	1659
1270	5663

Tabulka B.3: Statistiky z matice záměn pro mapu Cache

statistika	hodnota
Přesnost	0,7969
Senzitivita	0,7734
Specificita	0,8211

B.2 Mapa Cobblestone

Tabulka B.4: Výstup z programu R pro logistický model na mapě Cobblestone

Estimate	Std. Error	z value	Pr(>|z|)
2,3603	0,0657	35,9049	0,0000
0,1847	0,0023	79,0282	0,0000
0,2833	0,0047	60,5620	0,0000
-0,3753	0,0035	-105,7767	0,0000
-0,2457	0,0702	-3,5028	0,0005
0,0201	0,0051	3,9087	0,0001
-0,2108	0,0220	-9,5825	0,0000

Tabulka B.5: Vybrané statistiky z matice záměn pro mapu Cobblestone

Původní pozitivní	Původní negativní
5876	1530
1330	5683

Tabulka B.6: Statistiky z matice záměn pro mapu Cobblestone

statistika	hodnota
Přesnost	0,8017
Senzitivita	0,7879
Specificita	0,8154

B.3 Mapa Dust2

Tabulka B.7: Výstup z programu R pro logistický model na mapě Dust2

Estimate	Std. Error	z value	Pr(>|z|)
2,3909	0,0661	36,1522	0,0000
0,1863	0,0024	79,1851	0,0000
0,2872	0,0047	60,9470	0,0000
-0,3790	0,0036	-106,1633	0,0000
-0,2652	0,0703	-3,7737	0,0002
0,0195	0,0052	3,7531	0,0002
-0,2254	0,0221	-10,2093	0,0000

Tabulka B.8: Vybrané statistiky z matice záměn pro mapu Dust2

Původní pozitivní	Původní negativní
5809	1670
1367	5573

Tabulka B.9: Statistiky z matice záměn pro mapu Dust2

statistika	hodnota
Přesnost	0,7894
Senzitivita	0,7694
Specificita	0,8095

B.4 Mapa Inferno

Tabulka B.10: Výstup z programu R pro logistický model na mapě Inferno

Estimate	Std. Error	z value	Pr(>|z|)
2,3696	0,0661	35,8444	0,0000
0,1867	0,0024	79,2581	0,0000
0,2845	0,0047	60,6496	0,0000
-0,3793	0,0036	-106,1823	0,0000
-0,1817	0,0703	-2,5861	0,0097
0,0165	0,0052	3,1944	0,0014
-0,2221	0,0221	-10,0599	0,0000

Tabulka B.11: Vybrané statistiky z matice záměn pro mapu Inferno

Původní pozitivní	Původní negativní
5833	1579
1416	5591

Tabulka B.12: Statistiky z matice záměn pro mapu Inferno

statistika	hodnota
Přesnost	0,7923
Senzitivita	0,7798
Specificita	0,8047

B.5 Mapa Nuke

Tabulka B.13: Výstup z programu R pro logistický model na mapě Nuke

Estimate	Std. Error	z value	Pr(>|z|)
2,3640	0,0658	35,9237	0,0000
0,1850	0,0023	78,8206	0,0000
0,2825	0,0047	60,5275	0,0000
-0,3765	0,0036	-106,0337	0,0000
-0,2195	0,0704	-3,1198	0,0018
0,0242	0,0052	4,6867	0,0000
-0,2099	0,0220	-9,5225	0,0000

Tabulka B.14: Vybrané statistiky z matice záměn pro mapu Nuke

Původní pozitivní	Původní negativní
5865	1582
1344	5628

Tabulka B.15: Statistiky z matice záměn pro mapu Nuke

statistika	hodnota
Přesnost	0,7971
Senzitivita	0,7806
Specificita	0,8136

B.6 Mapa Overpass

Tabulka B.16: Výstup z programu R pro logistický model na mapě Overpass

Estimate	Std. Error	z value	Pr(>|z|)
2,3160	0,0659	35,1465	0,0000
0,1858	0,0024	78,9978	0,0000
0,2835	0,0047	60,5120	0,0000
-0,3762	0,0036	-105,9345	0,0000
-0,1678	0,0702	-2,3882	0,0169
0,0210	0,0052	4,0633	0,0000
-0,2048	0,0220	-9,2959	0,0000

Tabulka B.17: Vybrané statistiky z matice záměn pro mapu Overpass

Původní pozitivní	Původní negativní
5839	1561
1376	5643

Tabulka B.18: Statistiky z matice záměn pro mapu Overpass

statistika	hodnota
Přesnost	0,7963
Senzitivita	0,7833
Specificita	0,8093

B.7 Mapa Train

Tabulka B.19: Výstup z programu R pro logistický model na mapě Train

Estimate	Std. Error	z value	Pr(>|z|)
2,3477	0,0659	35,6435	0,0000
0,1854	0,0023	79,1325	0,0000
0,2845	0,0047	60,6243	0,0000
-0,3755	0,0035	-105,9025	0,0000
-0,2395	0,0703	-3,4044	0,0007
0,0217	0,0052	4,2023	0,0000
-0,2064	0,0220	-9,3709	0,0000

Tabulka B.20: Vybrané statistiky z matice záměn pro mapu Train

Původní pozitivní	Původní negativní
5884	1552
1398	5585

Tabulka B.21: Statistiky z matice záměn pro mapu Train

statistika	hodnota
Přesnost	0,7954
Senzitivita	0,7825
Specificita	0,8080