UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Predikce výsledků tenisových utkání pomocí logistické regrese

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek Ph.D.**
Vypracoval(a): **Bc. Pavlína Přibylová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Pavlína Přibylová

Název práce: Predikce výsledků tenisových utkání pomocí logistické regrese

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Cílem této diplomové práce je pomocí logistické regrese vytvořit modely k predikci výsledků tenisových zápasů na vybraných typech tenisových turnajů a posléze ověřit „správnost“ těchto modelů na nových pozorováních. Při vytváření modelů byla použita data ze zápasů mužského tenisu za sezónu 2017 a k následné validaci modelů byla použita nová pozorování ze zápasů za sezónu 2018. Součástí práce je mimo jiné i interpretace parametrů vybraných modelů a popis možných příčin nastalých výsledků.

Klíčová slova: tenis, grandslamové turnaje, tie-break, logistická regrese, klasifikace, ROC křivka

Počet stran: 77

Počet příloh: 1 CD

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Pavlína Přibylová

Title: The logistic regression based predictions of tennis match results

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek Ph.D.

The year of presentation: 2019

Abstract: The aim of this thesis is to create models for predicting tennis match results at certain tennis tournaments and subsequently ensure the models are working "correctly" based on testing new observations. For the model design, male tennis match results of 2017 season were used and for subsequent model validation, male tennis match results of 2018 season were used. This thesis also includes interpretation of chosen model parameters and description of possible causes for results observed.

Key words: tennis, grandslam tournaments, tie-break, logistic regression, classficiation, ROC curve

Number of pages: 77

Number of appendices: 1 CD

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu práce Mgr. Ondřeji Vencálkovi, Ph.D. za věcné připomínky, odborné rady a čas strávený na konzultacích, který mi byl věnován během vypracovávání mé diplomové práce.

Úvod

Co všechno dokáže ovlivnit výsledek tenisového utkání mezi dvěma hráči? Postavení na žebříčku? Úspěšnost ve vzájemných zápasech? Je výška hráče natolik důležitá, že jednoznačně ovlivní výsledek ve prospěch vyššího ze dvou hráčů? Cílem této diplomové práce je predikovat výsledky tenisových utkání pomocí logistické regrese na základě nejvhodnějšího modelu uvažujícího pouze ty důležité faktory. „Správnost“ modelů vytvořených na základě výsledků zápasů za sezónu 2017 si posléze ověříme na datech za sezónu 2018, kdy odhadneme vítěze jednotlivých tenisových zápasů a srovnáme je s nastalými výsledky.

V první kapitole si nejprve shrneme nejdůležitější informace o tenise – co je to žebříček ATP, jak probíhá nasazování hráčů na turnajích a jak vypadá tenisový zápas. Dále bude čtenář seznámen s grandslamovými turnaji a s turnaji Masters, způsoby kvalifikace na tyto turnaje a rozdílnostmi mezi nimi. V neposlední řadě si vytvoříme představu o datových souborech, s kterými budeme v rámci praktické části pracovat. V závěru první kapitoly se zamyslíme nad všemi možnými faktory, které by mohly ovlivnit výsledek zápasu a vysvětlíme, jakým způsobem mohou mít na výsledek zápasu vliv.

V druhé kapitole čtenáře seznámíme s některými rozděleními náhodné veličiny a s regresními modely. Od klasického regresního lineárního modelu přejdeme k zobecněnému lineárnímu modelu, kde se zaměříme na spojovací funkci a její souvislost s logistickou regresí.

Ve třetí kapitole bude čtenář obeznámen s logistickou regresí, způsobem odhadů regresních parametrů a s testováním jejich významnosti. Dále se zaměříme na teoretickou stránku interpretace parametrů a na metody hodnocení kvality výsledných modelů.

Ve čtvrté kapitole se dostáváme k tvorbě modelů pro jednotlivé turnaje, kdy zvolíme různé přístupy tvorby těchto modelů. U vybraných modelů si ukážeme interpretaci vypočtených regresních parametrů.

V poslední kapitole použijeme nová pozorování k validaci modelů, kdy budeme rozhodovat na základě procenta správně zařazených pozorování a na základě velikosti plochy pod ROC křivkou.

Kapitola 1

Game, set, match

V této kapitole si shrneme důležité informace o tenise, grandslamových turnajích a turnajích typu Masters, ukážeme si příklad dat, s kterými budeme dále pracovat v softwaru `R` a uděláme si představu o faktorech, které mohou ovlivnit výsledek zápasu.

1.1. Základní tenisová pravidla

Dlouho před tím, než hráči vůbec vstoupí na tenisový kurt, každý zápas v turnaji je už dopředu pečlivě naplánován. Aby mohli organizátoři vhodně nalosovat dvojice pro zápasy prvního kola, účastníci turnaje musí být „nasazeni“ (anglicky *seeded*). Nejlehčí cestou nasazení hráčů je použít jejich hodnocení v žebříčku ATP – jak je ale tento žebříček určován?

Každý týden mužská asociace profesionálních tenistů (*Association of Tennis Professionals*, ATP) spočítá a aktualizuje hodnocení v žebříčku. Body v žebříčku (*ranking points*) berou v úvahu výsledky na turnajích během posledních 52 týdnů (kalendářní rok), kdy počet bodů je udělován na základě typu turnaje a dosaženého kola, než byl hráč „eliminován“ z turnaje. V následující tabulce jsou vypsány zisky bodů z různých typů turnajů. Následně se hráči seřadí podle dosaženého počtu bodů.

Typ turnaje	Vítěz	Finále	Semifinále	Čtvrtfinále	R16	R32	R64	R128
Grandslam	2000	1200	720	360	180	90	45	10
Masters 1000	1000	600	360	180	90	45	25	10
ATP 500	500	300	180	90	45	20	–	–
ATP 250	250	150	90	45	20	10	–	–

Tabulka 1.1: Přehled bodových zisků na jednotlivých typech turnajů okruhu ATP

Hráči nemají tak velkou volnost v případě výběru turnajů, kterých se chtějí účastnit. Existuje minimální počet soutěžních turnajů, které hráči musí během tenisové sezóny absolvovat – 8 povinných turnajů typu Masters, 4 grandslamové turnaje, ATP Finals[1](pokud se hráč kvalifikoval), 4 turnaje ATP 500, kdy jeden z nich musí být v období po US Open a 2 turnaje ATP 250 nebo nižší úrovně. [1]

Jakmile jsou hráči ohodnoceni dle žebříčku ATP, nasazení hráčů spočívá pouze ve vypsání 32 nejlepších účastníků turnaje – takovýmto způsobem se nasazování hráčů realizuje na Australian Open, French Open a US Open. V případě Wimbledonu se používá speciální rovnice, zohledňující hráčem odehrané turnaje na travnatém povrchu během posledního kalendářního roku, a navíc se bere v úvahu i nejlepší dosažený výsledek (na turnajích travnatého povrchu) během těchto posledních dvanácti měsíců.

Pokud předpokládáme, že hráč musí prohrát alespoň jeden zápas, aby došlo k jeho vyřazení z turnaje, potřebujeme aby se odehrálo alespoň 127 zápasů (pokud začínáme s celkovým počtem 128 hráčů a 127 z nich musí být vyřazeno, abychom mohli určit vítěze). Tento způsob je znám pod názvem vyřazovací systém (anglicky *single-elimination*). Obecně, k určení vítěze je třeba, aby bylo odehráno $\log_2(n)$ kol, kde n značí počet hráčů.

Další věc, kterou musíme uvažovat při „strukturování“ turnaje je, že ideálně by měl turnaj být pro hráče tím obtížnější, čím hlouběji v turnaji postupuje – pravděpodobnost, že hráč narazí na výše postaveného hráče na žebříčku, roste. K tomuto efektu dojde přirozeně, kdy silnější hráči triumfují v počátečních ko-

[1]neboli Turnaj mistrů, vrchol tenisové sezóny odehrávající se v Londýně, kdy vítěz získá 1500 bodů - kvalifikace je podmíněna postavením na žebříčku „ATP Race“, kdy jsou brány v úvahu pouze body získané během stávající sezóny

lech turnaje. Musí se ale dávat pozor na to, aby se někteří z nejvýše nasazených hráčů nepotkali hned z kraje turnaje (a jeden druhého nevyeliminovali). Proto se první a druhý nasazený hráč „nasadí“ na opačné konce eliminačního pavouka, tj. zabezpečí se tak, že pokud první a druhý nasazený vyhrají všechny své zápasy, narazí na sebe až ve finále. Podobná myšlenka je použita i v případě třetího a čtvrtého nasazeného, proto se 4 nejvýše nasazení mohou potkat nejdříve v semifinále (za předpokladu výhry všech svých předchozích zápasů). Stejně tak 8 nejvýše nasazených se nemůže potkat dříve než ve čtvrtfinále, 16 nejvýše nasazených se nemůže potkat dříve než v osmifinále a 32 nejvýše nasazených se nemůže potkat dříve než ve třetím kole (neboli R32).

Nyní se zaměříme na samostatný rozbor zápasu – snad nejvýraznějším rysem tenisu je jeho unikátní bodovací systém. Tenisový zápas je hrán na dva vítězné sety (zápasy na grandslamových turnajích se hrají na tři vítězné sety), první hráč, který vyhraje 2 (případně 3) sety, vyhraje celý zápas. Set je rozdělen na gamy. Konec setu nastane, jestliže hráč vyhrál alespoň 6 gamů, a rozdíl mezi počet jeho získaných gamů a počtem získaných gamů soupeře je nejméně 2. Většinou, pokud obě strany vyhrály 6 gamů v jednom setu, o výsledku setu se rozhodne v tie-breaku (vítěz tie-breaku musí získat alespoň 7 bodů, kdy rozdíl počtu získaných bodů musí být nejméně 2). Tie-breaky nejsou hrány v rozhodujících pátých setech grandslamových zápasů (neplatí pro US Open a nově od roku 2019 i pro Australian Open).

Hra (anglicky *game*) je soubor bodů, kdy jeden hráč podává a druhý „přijímá“. Pokud podávající hráč vyhraje svou hru, říká se, že udržel své podání. Pokud hráč na příjmu vyhraje tuto hru, říká se, že prolomil podání (svého soupeře). K získání setu je nutné alespoň jednou prolomit soupeřovo podání, jinak se pokračuje tie-breakem. Bod (anglicky *point*) je základní hodnotící jednotka v tenise. Bod získá hráč, který trefil úder tak, že ho oponent nedokázal úspěšně vrátit. K zisku gamu je třeba získat alespoň 4 body s rozdílem 2 oproti soupeřovým bodům. Body v tenise jsou následující: 0 (*love*), 15 (*fifteen*), 30 (*thirty*), 40 (*forty*). Jestliže oba dva hráči dosáhnou 40 bodů, nastává shoda (*deuce*). Hráč, který vyhraje následující

bod získává výhodu (*advantage*). Hráč mající výhodu buď získá další bod a vyhraje tím gam nebo bod nezíská a hra se vrátí do shody. Tenisové výsledky se udávají jako počet vyhraných her v jednotlivých setech – například $6-1$ $6-3$, tedy vítěz vyhrál první set šesti hrami ku jedné a druhý set šesti hrami ku třem.

Jedním z hlavních rysů tenisových výsledků je, že na rozdíl od sportů jako je fotbal nebo hokej, ne každý bod ovlivňuje výsledek zápasu. To může vést ke kuriózním situacím, ve kterých hráč může celkově získat více bodů než oponent, nebo dokonce i vyhrát více her než jeho oponent, ale nakonec stejně prohrát zápas. Příkladem může být zápas z Wimbledonu 2010, který je mimo jiné známý rekordní délkou – hrálo se celkem 11 hodin a 5 minut, během 3 dnů. Zápas se odehrál mezi francouzským hráčem Nicolasem Mahutem a americkým hráčem Johnem Isnerem. Výsledné skóre bylo (z pohledu J. Isnera) $6-4$, $3-6$, $6-7$, $7-6$, $70-68$ a to přesto, že Mahut vyhrál celkem 502 bodů, kdežto Isner „jen“ 478. [2]

1.2. Grandslamové turnaje

Grandslamové turnaje jsou nejprestižnější turnaje tenistů za celou sezónu. Jeden grandslam se odehrává v Severní Americe, dva v Evropě a jeden v Austrálii.

Turnaj	Země	Místo konání	Datum konání	Povrch	Počet hráčů
Australian Open	Austrálie	Melbourne	14.–29. ledna	tvrdý	128 (32)
French Open	Francie	Paříž	26. května–9. června	antuka	128 (32)
Wimbledon	Velká Británie	Londýn	1.–14. červenec	tráva	128 (32)
US Open	USA	New York	26. sprna–8. září	tvrdý	128 (32)

Tabulka 1.2: Přehled grandslamových turnajů

Číslo v závorce u počtu hráčů udává počet nasazených hráčů na turnaji. Data konání jsou pouze orientačního charakteru, vždy se jedná o 2 týdny, počínaje nedělí nebo pondělím. Vítěz grandslamového turnaje získá 2 000 bodů do žebříku hodnocení ATP, více viz Základní tenisová pravidla. Zápasy na grandslamových turnajích se hrají na tři vítězné sety, celkový počet setů v zápase tedy může být

tři, čtyři nebo pět.

Účast na těchto turnajích je povinná, prvních 104 tenistů žebříčku ATP automaticky získá účast (pokud je některý tenista zdravotně indisponován, tak jej nahradí tenista v žebříčku o jedno místo níže). Dalších 16 míst je rezervováno pro kvalifikanty, kteří podstoupí menší turnaj – tříkolovou kvalifikaci s vlastním nasazením hráčů. Zbývajících 8 míst je určeno pro hráče, kteří obdrží divokou kartu (anglicky *Wild Card*) – rozhodnutí je na pořadatelích, obvykle vybírají mladé talentované tenisty, kteří ještě neodehráli dostatek zápasů na okruhu ATP a tudíž nemají dostatek bodů na kvalifikaci tradičním způsobem; nebo je divoká karta určena pro hráče, kteří se vrací po delších zdravotních potížích.

Každý grandslamový turnaj je rozdílný oproti ostatním, na začátku sezóny jsou tenisté odpočatí, jejich těla ještě nejsou natolik náchylná ke zraněním jako na konci sezóny. Pro některé tenisty ale může být náročné počasí v Melbourne, v lednu zde vrcholí letní období a teploty povrchu tenisových kurtů mohou dosáhnout až k 70 °C. [3] French Open se hraje na antuce, což je nejpomalejší tenisový povrch, který způsobuje, že se hrají delší výměny a tenisté v pozdějších kolech pociťují větší únavu. Wimbledon patří k nejoblíbenějšímu grandslamu, jak mezi tenisty, tak i mezi fanoušky. Středně rychlý, travnatý povrch dává prostor jak krátkým výměnám, tak i delším zajímavým výměnám, kdy není nouze o kuriózní tenisové prvky. Závěrečný grandslam sezóny, US Open, obvykle přináší nejvíce odstoupení během turnaje (W.O. neboli *Walkover*) nebo zápasů, které skončí předčasně tím, že jeden z hráčů už není schopný pokračovat dále v zápase vlivem zranění nebo nemoci (značí se *Retired*). [4]

1.3. Turnaje typu Masters

Turnaje typu Masters, oficiálně nazývány „ATP Tour Masters 1 000“, představují sérii devíti prestižních tenisových turnajů – čtyři turnaje jsou odehrány v Evropě, čtyři v Severní Americe a jeden v Asii.

Turnaj	Země	Místo konání	Datum konání	Povrch	Počet hráčů
Indian Wells Masters	USA	Indian Wells, CA	5.–18. března	tvrdý	96 (32)
Miami Open	USA	Key Biscayne, FL	19. března–1. dubna	tvrdý	96 (32)
Monte-Carlo Masters	Francie	Roquebrune-Cap-Martin	15.–22. dubna	antuka	56 (16)
Madrid Open	Španělsko	Madrid	7.–14. května	antuka	56 (16)
Italian Open	Itálie	Řím	14.–20. května	antuka	56 (16)
Canadian Open	Kanada	Montreal/Toronto	3.–11. srpna	tvrdý	56 (16)
Cincinnati Masters	USA	Mason, OH	11.–19. srpna	tvrdý	56 (16)
Shanghai Masters	Čína	Šanghaj	7.–14. října	tvrdý	56 (16)
Paris Masters	Francie	Paříž	27. října–4. listopadu	tvrdý	48 (16)

Tabulka 1.3: Přehled Masters turnajů

Číslo v závorce u počtu hráčů udává počet nasazených hráčů na turnaji (tzv. „Seeded“ hráči). Turnaj Canadian Open je specifický v tom, že se místo konání střídá – liché roky se turnaj odehrává v Montrealu, sudé roky v Torontu.

Vítěz turnaje Masters získá 1 000 bodů do žebříčku hodnocení, poražený finalista získá 600 bodů. Zápasy na těchto turnajích se hrají na dva vítězné sety, v minulosti tomu tak ale nebylo – do roku 2007 byly některé finálové zápasy hrány na tři vítězné sety.

Jak již bylo řečeno v sekci Základní tenisová pravidla, hráči mají povinnost účastnit se 8 Masters turnajů. Do roku 2009 bylo povinných 10 Masters turnajů (kromě zmíněných devíti mezi ně patřil i turnaj v Hamburku) – organizace ATP ale chtěla „odstranit“ turnaje Monte Carlo a Hamburk, což vedlo k protestům jak ze strany hráčů, tak i ze strany organizátorů obou turnajů. Celá situace vyústila podáním žaloby na organizaci ATP. Jako ústupek bylo rozhodnuto, že si turnaj Monte Carlo může ponechat „status“ turnaje ATP 1 000, avšak účast hráčů už nebude povinná, nýbrž dobrovolná. Turnaj v Hamburku stejný osud nepotkal, ztratil status turnaje Masters a klesl na úroveň ATP 500. Z důvodu nepovinné účasti na turnaji Monte Carlo nebude tento podnik ATP předmětem další analýzy. [5]

Součástí těchto turnajů je, stejně jako v případě grandslamových turnajů, kvalifikace na „hlavní“ turnaj, organizována pro hráče, kteří se nedostali do hlavní soutěže pomocí postavení na žebříčku. V případě turnajů o „velikosti“ 96 hráčů

(Indian Wells Masters a Miami Open) se 12 hráčů kvalifikuje prostřednictvím dvoukolové kvalifikace. Pokud se jedná o turnaj o „velikosti“ 48 hráčů (Paris Masters), takto se kvalifikuje jen 6 hráčů. Poslední možností jsou turnaje o „velikosti“ 56 hráčů, kdy z kvalifikace postupuje do hlavní soutěže 7 hráčů. Pokud se stane, že některý z hráčů odstoupí z turnaje ještě před jeho zahájením, tak je nahrazen tzv. *Lucky Loserem (LL)*, kterým bývá hráč, který prohrál v posledním kole kvalifikace.

V případě nasazených hráčů dostane určitý počet nejvýše nasazených tenistů tzv. *Bye* neboli automatický postup do druhého kola bez toho, aby hráč musel absolvovat zápas v prvním kole. Počet hráčů, kteří automaticky postoupí do druhého kola, se liší na základě celkového počtu hráčů v turnaji, jak je ukázáno v následující tabulce. [6]

Počet hráčů na turnaji	**Počet hráčů, kteří obdrží „Bye“**	**Turnaj Masters**
48	16	Paříž
56	8	Monte-Carlo, Madrid, Řím, Toronto/Montreal, Cincinnati, Šanghaj
96	32	Indian Wells, Miami

Tabulka 1.4: Masters turnaje a počet hráčů, kteří obdrží volný vstup do druhého kola

1.4. Představení dat

V této diplomové práci se budeme zabývat analýzou tenisových dat za sezónu 2017. Jednotlivé datové soubory vznikly kombinací informací z webové stránky https://matchstat.com/, která poskytla informace o jednotlivých zápasech, statistikách head-to-head a kurzových sázkách, z webové stránky https://www.atptour.com/, odkud byly čerpány charakteristiky jednotlivých hráčů – věk, postavení na žebříčku a výška hráče a z webové stránky https://www.tenisportal.cz/, kdy byly manuálně vyhledávány informace typu historická úspěšnost na konkrétním turnaji, počet účastí na konkrétním turnaji

a statistika Win/Loss za předchozí sezónu.

Z důvodu případného porušení předpokladů modelu, blíže vysvětleno v sekci Porušení předpokladů, použijeme data z jednotlivých kol zvlášť pro každý turnaj. Naše predikce budeme vytvářet pro 12 různých turnajů, kde každý turnaj bude mít 4 (případně 3) vlastní modely pro jednotlivá kola – R128, R64 a R32 v případě grandslamových turnajů, Indian Wells Masters a Miami Open; R64 a R32 pro zbývající turnaje Masters. Všechny původní datové soubory jsou k dispozici na přiloženém CD, v příslušných složkách `Grandslams` a `Masters`. Pro představu je níže zobrazen základní datový soubor pro první kolo Australian Open. Konečný počet proměnných je vyšší, ale pro základní povědomí o struktuře datových souborů nám postačí následující tabulka.

Name	**KurzV**	**KurzP**	**H2HV**	**H2HP**	**Vysledek**	**Vyska**	**Ranking**
Illya Marchenko	1	29,25	2	0	0	185	93
Yen-Hsun Lu	3,4	1,29	1	1	0	180	84
Quentin Halys	1,33	3,2	1	0	0	191	154
Andy Murray	1	29,25	2	0	1	190	1
Andrey Rublev	3,4	1,29	1	1	1	188	156
Sam Querrey	1,33	3,2	1	0	1	199	31
	Vek	**Seeded**	**Qualifier**	**History**	**Starts**	**MPlayed**	**MWon**
	29	No	No	R64	7	56	0,48
	33	No	No	R32	13	58	0,72
	20	No	No	R64	1	64	0,52
	29	Yes	No	F	11	90	0,89
	19	No	Yes	DNS	0	71	0,54
	29	Yes	No	R32	10	52	0,54

Tabulka 1.5: Ukázka základních dat pro první kolo (R128) Australian Open

Nyní si krátce popíšeme a vysvětlíme, co jednotlivé proměnné v tabulce znamenají. O důvodech zařazení těchto proměnných a bližší analýza jejich možného vlivu na výsledek zápasu bude více řečeno v sekci Faktory, které mohou ovlivnit výsledek tenisového utkání.

Kód v R	Název	Vysvětlení
KurzV	Kurz vítěze	kurz vypsaný sázkovými kancelářemi na výhru tenisty, který nakonec zápas vyhrál
KurzP	Kurz poraženého	kurz vypsaný sázkovými kancelářemi na výhru tenisty, který nakonec zápas prohrál
H2HV	Head-to-head vítěze	počet vyhraných vzájemných zápasů mezi dvěma tenisty, z pohledu vítěze
H2HP	Head-to-head poraženého	počet vyhraných vzájemných zápasů mezi dvěma tenisty, z pohledu poraženého
Vysledek	Výsledek zápasu	1 značí, že tenista zápas vyhrál 0 značí, že tenista zápas prohrál
Vyska	Výška hráče	výška hráče v centimetrech
Ranking	Hodnocení na žebříčku ATP	pořadí tenisty na žebříčku ATP před zahájením konkrétního turnaje
Vek	Věk	věk hráče před zahájením konkrétního turnaje
Seeded	Nasazení	1 značí, že tenista byl nasazen, 0 značí, že tenista nebyl nasazen
Qualifier	Kvalifikant	1 značí, že tenista prošel kvalifikací, 0 značí, že se tenista nemusel kvalifikovat
History	Historická úspěšnost na turnaji	nejlepší výsledek dosažený na konkrétním turnaji
Starts	Počet účastí na turnaji	počet účastí před rokem 2017 na konkrétním turnaji
MPlayed	Počet odehraných zápasů	počet odehraných zápasů na okruhu ATP v předchozí sezóně (rok 2016)
MWon	Počet vyhraných zápasů	počet vyhraných zápasů na okruhu ATP v předchozí sezóně (rok 2016)

Tabulka 1.6: Popis jednotlivých proměnných v datovém souboru

1.5. Faktory, které mohou ovlivnit výsledek tenisového utkání

V této části budeme uvažovat, které vlastnosti tenistů mohou ovlivnit výsledek zápasu. Jelikož neexistuje nějaký univerzální, volně dostupný model, který by odhalil, které faktory zohledňují například sázkové kanceláře, tak jsme využili volně dostupné statistiky hráčů a přemítali, co všechno by mohlo mít vliv na výsledek zápasu na konkrétním turnaji, v konkrétním kole a proti konkrétnímu hráči.
V praktické části můžeme zjistit, že naše modely nejsou vůbec tak složité a výsledek zápasu ovlivňuje jen několik málo faktorů.

1.5.1. Výška tenisty

O tom, zda výška tenisty ovlivňuje jeho výkon (pokud ano, tak do jaké míry), se vedou dlouhé diskuze. Například dle [7], žádný hráč vyšší než 6'4"(193 cm) nedosáhl prvního místa v žebříčku ATP. Pozorujeme zde dva protichůdné efekty – vyšší hráči mohou podávat rychleji než nižší hráči, což značně pomáhá při podání, kdy mají šanci dát více es a tím si zajistit pohodlnější průběh her na podání. Proti tomu, nižší hráči, kteří jsou na okruhu ATP úspěšní, excelují rychlým pohybem a skvělým returnem. Proto je lepší předpokládat, že výška je pro tenistu výhodou, ale jen do určitého bodu.

1.5.2. Postavení na žebříčku ATP – Ranking

Hodnocení na žebříčku ATP (způsob zisku bodů je blíže vysvětlen v sekci Základní tenisová pravidla) objektivně hodnotí výkon tenisty během posledního kalendářního roku. Tuto lehce dostupnou charakteristiku každého hráče jsme bez váhání zařadili mezi faktory, které mají vliv na výsledek zápasu. Při posuzování žebříčku ATP jsme narazili na problém, kdy několik málo hráčů bylo postaveno na žebříčku v třetí stovce a dále. Proto jsme se rozhodli omezit hodnocení na žebříčku dvoustým místem. Tato situace se týká pouze několika tenistů – kvalifikantů a

hráčů, kteří měli delší dobu pauzu ze zdravotních důvodů a zažádali si o takzvané „Entry Protection“[2].

1.5.3. Historická úspěšnost na konkrétním turnaji

Důležitým faktorem, který nesmíme opomenout, je historická úspěšnost hráče na konkrétním turnaji. Jelikož se turnaje hrají na různých površích, tak je přirozené, že někteří hráči budou na konkrétním povrchu vynikat. Například španělský hráč Rafael Nadal patří mezi nejúspěšnější antukové hráče v historii tenisu. V roce 2017 zaznamenal sérii 21 výher na antuce, počínaje French Open 2017 až po Madrid Open 2018. Z důvodu dominance některých hráčů na konkrétních turnajích jsme chtěli zavést proměnnou `TWon`, která by značila počet vyhraných titulů na daném turnaji. Avšak z důvodů popsaných například v [8] jsme od užití této proměnné upustili, jelikož nenulové hodnoty by se objevily pouze u několika hráčů (v případě grandslamů u čtyř hráčů). V úvahu jsme ale vzali alespoň počet startů na konkrétním turnaji, proměnná `Starts`, společně s nejlepším dosaženým výsledkem na turnaji, pomocí proměnné `History`, ve formě R128/R64/R32-/R16/QF/SF/F – neboli kolo, kam nejdále tenista došel.

1.5.4. Nasazení, kvalifikanti a výhoda domácího prostředí

Jak již bylo řečeno v sekci Základní tenisová pravidla, na každém turnaji (ať už se jedná o grandslamový turnaj nebo Masters turnaj) je skupina nejlepších hráčů dle žebříčku ATP nasazena do turnaje. Někteří z nasazených hráčů (ta výše nasazená skupina hráčů) mají v případě Masters turnajů výhodu vynechání prvního kola (více viz sekce Turnaje typu Masters a pojem *Bye*), kdy automaticky postupují do druhého kola. Tím pádem hrají o zápas méně a nemusí být v pozdějších kolech tak unavení. Tento jev budeme sledovat pomocí proměnné `Bye`, kdy 1 bude značit, že hráč neabsolvoval zápas prvního kola turnaje a 0, že hráč absolvoval zápas prvního kola.

[2] více např. na https://metro.co.uk/2018/12/05/protected-ranking-will-andy-murray-seeded-australian-open-8210959/

Kromě výhody volného prvního kola (na grandslamových turnajích k tomuto privilegiu nedochází) mají všichni nasazení hráči výhodu lepšího postavení v rámci celého eliminačního pavouka. Způsob tvorby eliminačního pavouka byl vysvětlen v úvodní sekci této kapitoly. To, jestli je hráč nasazen nebo ne, je zachyceno pomocí proměnné `Seeded`.

Dalším aspektem každého turnaje jsou kvalifikanti. I když se většinou jedná o tenisty, kteří jsou v žebříčku ATP na místech hodně vzdálených od tenisové špičky, bylo by pošetilé je považovat za typově stejné hráče, jako zbytek nenasazených hráčů. U kvalifikantů může nastat situace, kdy tím, že v kvalifikačních zápasech už odehráli dva nebo tři zápasy, tak jsou více rozehraní než tenisté, kteří mohli mít před turnajem i několika týdenní pauzu, a jsou jejich rozehraností zaskočeni. Stejně tak v případě opravdu mladých kvalifikantů je možné pozorovat určitou dravost a postoj, kdy nemají co ztratit a dávají do zápasů opravdu vše. Proto budeme hráče, kteří se kvalifikovali na turnaj podstoupením kvalifikačního „mini turnaje“ sledovat proměnnou `Qualifier`, kdy 1 značí, že se jedná o hráče, který prošel kvalifikačními boji. Příkladem rozehraného kvalifikanta může být za rok 2017 srbský hráč Filip Krajinovic, který absolvoval kvalifikaci na závěrečný Masters turnaj v Paříži, postupně porazil „papírově“ daleko lepší hráče a svého přemožitele potkal až ve finále[3].

Poslední možností, která úzce souvisí s konkrétním turnajem, je zohlednění účasti na turnaji, který se koná na domácí půdě. Tenista je hnán touhou předvést se před domácím publikem a chutí dojít co nejdále. Přece jen vyhrát turnaj, kde si většina fanoušků přeje vaše vítězství, patří mezi důležité milníky v tenisové kariéře. To, zda hráč hraje na domácí půdě, je zachyceno proměnnou `Home`, kdy 1 bude značit, že hráč pochází ze stejné země, kde se daný turnaj odehrává.

[3] více informací na https://www.tenisportal.cz/zpravy/krajinovic-a-sock-jsou-poprve-ve-finale-masters-american-zautoci-na-turnaj-mistru-22037/

1.5.5. Historická úspěšnost vzájemných zápasů

Statistika head-to-head zachycuje výhry a prohry dvou hráčů ve vzájemných zápasech. Původně byla tato statistika v základním tvaru, kdy `H2HV` značilo počet vyhraných zápasů vítězného tenisty a `H2HP` značilo počet vyhraných zápasů poraženého tenisty. Použít statistiku v tomto tvaru se nejevilo jako ideální způsob zachycení všech aspektů odehraných vzájemných zápasů. Proto jsme zvolili variantu uvažování celkového počtu vzájemně odehraných zápasů, `H2HPlayed`, a k tomu počet výher v těchto zápasech pro každého hráče zvlášť, `H2HWon`. Příkladem může být Head-to-head mezi Rafaelem Nadalem a Marcem Baghaditisem. Před druhým kolem (R64) na Australian Open 2017 vyhrál R. Nadal 9 vzájemných zápasů, zatímco M. Baghaditis vyhrál pouze 1 vzájemný zápas. Celkem tedy proti sobě hráli desetkrát, kdy R. Nadal vyhrál 9/10 zápasů a M. Baghaditis pouze 1/10 zápasů. V takovémto případě drtivé dominance ve vzájemných zápasech málokdy očekáváme, že by mohlo dojít k výhře „slabšího“ hráče.

1.5.6. Zkušenost hráče a jeho forma

Bohužel v případě zkušenosti hráče a jeho současné formy (myšleno v době konání jednotlivých turnajů) již nemáme k dispozici tolik dat. Rozhodli jsme se zařadit do možných faktorů Věk (`Vek`), který alespoň trochu zohledňuje zkušenost hráče. V případě současné herní formy hráče nejsme schopni dohledat například vyhrané zápasy odehrávající se před konkrétním turnajem – souhrnné statistiky výher/proher (anglicky *Win/Loss*) se udávají ve tvaru za celou sezónu. Abychom alespoň trochu uvažovali herní rozpoložení tenisty v sezóně 2017, použili jsme statistiku výher/proher za sezónu 2016. Budeme uvažovat pouze zápasy odehrané na okruhu ATP a to proměnnou `MPlayed`, neboli počet odehraných zápasů na okruhu ATP za sezónu 2016. Turnaje nejnižší kategorie jsou pořádány ITF.[4] U hráčů, kteří jsou v žebříčku postaveni až v první či druhé stovce, bychom měli zkreslená čísla, právě z turnajů nejnižší kategorie. Kromě počtu odehraných

[4]mezinárodní tenisová organizace, řídící dění ve světovém tenise, organizuje Davis Cup, Fed Cup, Hopman Cup, všechny 4 grandslamy a turnaje mužského a ženského okruhu ITF – turnaje nejnižší kategorie

zápasů za sezónu nás především zajímá počet vyhraných zápasů za sezónu `MWon` – čím vyšší číslo, tím více zápasů hráč odehrál, dostával se v turnajích daleko, případně se účastnil velkého počtu turnajů a nebyl zdravotně indisponován.

1.5.7. Průběh zápasů hraných na turnaji

Dostane-li se tenista do druhého kola (a dále), tak na jeho výkon má vliv čas strávený na kurtu v předchozích zápasech. Pokud tenista hladce vyhraje ve dvou setech (ve třech v případě grandslamu), bez tie-breaku, tak bude více odpočatý a naběhá menší vzdálenost, než kdyby musel bojovat o postup i ve třetím (v případě grandslamu čtvrtém a pátém) setu. Naneštěstí délky zápasů (čistý časový údaj o délce trvání zápasu) nejsou běžně dostupné. Proto jako indikátor „únavy“ hráče v průběhu turnaje použijeme počet odehraných gamů na turnaji, `GPlayed`.

1.5.8. Kurz sázkových kanceláří

Kurz sázkových kanceláří můžeme považovat za kvantifikovaný odhad šance na výhru, otázkou je, zda tento faktor vůbec zahrnout do naší analýzy. Sázkové kanceláře používají podobné, leč složitější úvahy, než jakými se budeme zabývat my. Nevíme, co všechno uvažují sázkové kanceláře, ale pravděpodobně uvažují i zdravotní stav tenisty, zda v minulých zápasech nevyužil MTO – Medical Time Out. Předpokládáme, že zahrnují i dosavadní průběh zápasů v sezóně, kladou důraz na oblíbenost povrchu hráče, detailněji analyzují předchozí účasti na konkrétním turnaji, což vzhledem k omezeným informacím, které jsou volně dostupné na internetu, je v našem případě nemožné, nebo minimálně značně časově náročné a nepraktické. Kromě toho mohou brát v potaz i podrobnosti z hráčova soukromého života – úmrtí v rodině, narození dítěte, sňatek, rozchod, to vše dokáže ovlivnit psychiku hráče a mít spíše (negativní) dopad na jeho výkon. Proměnná `Kurz` je zvláštním typem kvantitativní proměnné, která teoreticky nabývá hodnot od 1 po ∞. Pokud by kurz na výhru hráče 1 a hráče 2 byl 2.0, znamená to, že sázkové kanceláře nedokázali určit favorita a šanci na výhru obou hráčů hodnotí totožně.

Kapitola 2

Regresní modely

K porozumění modelu logistické regrese je vhodné znát zobecněné lineární modely a některá rozdělení náhodné veličiny, s kterými se dále pracuje. Proto se s těmito pojmy nejdříve seznámíme a až poté přistoupíme k samotné logistické regresi.

2.1. Rozdělení náhodné veličiny

2.1.1. Binomické rozdělení

Binomické rozdělení je jedním ze základních typů diskrétních rozdělení. Náhodná veličina X, řídící se binomickým rozdělením $Bi(n, \pi)$, kde $n \in N$, $\pi \in (0, 1)$, může nabývat hodnot $k = 0, 1, \ldots, n$ s pravděpodobnostmi

$$\pi_k = P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Střední hodnotu si odvodíme jako

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} \pi^k (1 - \pi)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} \pi^k (1 - \pi)^{n-k} =$$

$$= n\pi \sum_{k=1}^{n} \binom{n-1}{k-1} \pi^{k-1} (1 - \pi)^{n-k} =$$

$$= n\pi \sum_{j=0}^{n-1} \binom{n-1}{j} \pi^j (1 - \pi)^{n-1-j} = n\pi[\pi + (1 - \pi)]^{n-1} = n\pi$$

Podobným způsobem odvodíme i rovnost $E[X(X-1)] = n(n-1)\pi^2$, kterou využijeme pro odvození rozptylu. Pro rozptyl binomického rozdělení platí

$$\text{var}(X) = E(X^2) - [E(X)]^2 = [n(n-1)\pi^2 + n\pi] - (n\pi)^2 = n\pi(1-\pi).[9]$$

2.1.2. Alternativní rozdělení

Alternativní, také nula-jedničkové rozdělení, je speciálním případem binomického rozdělení, kdy $n = 1$. Náhodná veličina X, která se řídí alternativním rozdělením, nabývá pouze dvou hodnot, $x_1 = 0$, $x_2 = 1$ s pravděpodobnostmi

$$p_1 = P(X = 0) = 1 - \pi, \qquad p_2 = P(X = 1) = \pi,$$

kde parametr $\pi \in (0, 1)$ určuje pravděpodobnost „úspěchu“. Pokud náhodná veličina nabude hodnoty 1, značí to úspěch, tedy v našem případě výhru tenisty. Hodnota 0 značí neúspěch neboli prohru tenisty.

Dále můžeme vypočítat střední hodnotu $E(X)$ následujícím způsobem,

$$E(X) = 0 \cdot (1-\pi) + 1 \cdot \pi = \pi,$$

a rozptyl $\text{var}(X)$ takto,

$$\text{var}(X) = E(X^2) - [E(X)]^2 = [0^2(1-\pi) + 1^2\pi] - \pi^2 = \pi(1-\pi). [9]$$

2.2. Klasický lineární regresní model

Představme si situaci, kdy proměnná Y je lineárně závislá na každé vysvětlující proměnné $x_1, \ldots, x_p$, které jsou vzájemně lineárně nezávislé.

Pak můžeme použít regresní funkci, která je lineární funkcí (vysvětlujících) proměnných $x_1, x_2, \ldots, x_p$,

$$E(Y|(x_1, x_2, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Máme-li k dispozici n nezávislých pozorování, potom pro i-té pozorování platí

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ pro } i = 1, \ldots, n,$$

kde $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ jsou neznámé regresní parametry, $x_1, x_2, \ldots, x_p$ jsou vysvětlující proměnné, $x_{i1}, x_{i2}, \ldots, x_{ip}$, jsou i-té hodnoty těchto p vysvětlujících proměnných a ε_i představuje náhodnou chybu i-tého pozorování. Všech n rovnic pozorování můžeme uspořádat do soustavy rovnic, maticově

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

kde $\mathbf{Y}$ je vektor pozorování, $\mathbf{x}$ je matice plánu (*design matrix*), $\boldsymbol{\beta}$ je vektor regresních koeficientů a $\boldsymbol{\varepsilon}$ je vektor náhodných chyb.
Předpoklady klasického lineárního regresního modelu jsou

- $E(\varepsilon_i) = 0$,
- $\text{var}(\varepsilon_i) = \sigma^2$,
- $cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j, \quad i, j = 1, \ldots, n$,
- $h(\mathbf{X}) = p + 1$, a navíc $p < n$.

V praxi se často potýkáme s problémem splnění těchto předpokladů, navíc někdy přidáváme i předpoklad normality náhodných chyb. Odhady regresních koeficientů se určují pomocí metody nejmenších čtverců (*least squares*). [9]

2.3. Zobecněný lineární model

Pokud jsou některé z předpokladů modelu porušeny, můžeme použít zobecněný lineární model (anglicky *generalized linear model*). Stejně tak přistoupíme k užití zobecněného lineárního modelu, pokud vysvětlovaná proměnná nabývá pouze dvou hodnot 1 a 0 – tenista vyhrál nebo prohrál zápas.

Vrátíme-li se k modelu klasické lineární regrese, můžeme říct, že hodnoty vysvětlované (závislé proměnné) Y jsou tvořeny systematickou a náhodnou složkou, kdy náhodná složka má normální rozdělení,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ pro } i = 1, \ldots, n, \; \varepsilon_i \sim N(0, \sigma^2).$$

Systematickou část vztahu nazveme lineární prediktor (*linear predictor*) a značíme ji η,

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}.$$

Předpokládali jsme, že mezi lineárním prediktorem a modelovanou střední hodnotou je rovnost, tj. platí $E(Y_i) = \mu_i$ a $\mu_i = \eta_i$. V případě zobecněných lineárních modelů již rovnost nepředpokládáme, místo toho zavádíme spojovací funkci (*link function*) g, specifikující vztah mezi střední hodnotou vysvětlované proměnné Y (náhodná složka) a lineárním prediktorem,

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$
$$g^{-1}(g(\boldsymbol{\mu})) = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{X}\beta) = \boldsymbol{\mu} = E(\mathbf{Y}).$$

V případě logistické regrese uvažujeme vysvětlovanou proměnnou Y jako náhodnou veličinu s alternativním rozdělením, $Y \sim Alt(\pi)$. Náhodná složka zobecněných lineárních modelů se skládá z vysvětlované proměnné Y charakterizované nezávislými pozorováními $Y_1, \ldots, Y_n$ s distribucí z třídy rozdělení exponenciálního typu. Hustota třídy těchto rozdělení je charakterizována funkcí

$$f\left(y_i; \theta_i\right) = a\left(\theta_i\right) b\left(y_i\right) e^{y_i Q(\theta_i)},$$

kde $a(\theta_i)$ je libovolná funkce parametru θ_i a $b(y_i)$ je libovolná funkce y_i. Hodnota parametru θ_i se může pro jednotlivé hodnoty vysvětlujících proměnných lišit. Výraz $Q\left(\theta\right)$ se nazývá přirozený parametr. [10]

Kapitola 3

Logistická regrese

V této kapitole si představíme logistickou regresi, stěžejní matematický aparát užitý pro predikci výsledků tenisových utkání. Od svého původního použití v epidemiologickém výzkumu nalézá tato metoda využití ve všech možných oborech – od finančního sektoru, přes zdravotnictví až po lingvistiku.

Obecně tedy máme alternativně rozdělenou proměnnou (subjekt splatí/nesplatí půjčku, tenista vyhraje/nevyhraje zápas), která je vysvětlována p veličinami, jak spojitými (výška tenisty, postavení hráče na žebříčku), tak kategoriálními (zda je hráč nasazen nebo ne).

3.1. Model logistické regrese

Představme si situaci, kdy máme n nezávislých náhodných veličin $Y_1, Y_2, \ldots, Y_n$, s alternativním rozdělením, které má parametry π_i, $i = 1, \ldots, n$ – vysvětlovaná proměnná má pro každou náhodnou veličinu obecně různé parametry (různé hodnoty vysvětlujících proměnných většinou vedou k různým hodnotám parametru π). Ke každé náhodné veličině $Y_1, Y_2, \ldots, Y_n$ náleží $(p+1)$-rozměrný vektor vysvětlujících proměnných $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$.

Platí

$$P(Y_i = 0) = 1 - \pi_i, \qquad P(Y_i = 1) = \pi_i,$$
$$E(Y_i) = \pi_i, \qquad \mathrm{var}(Y_i) = \pi_i(1 - \pi_i).$$

Pokud bychom při hledání vztahu mezi Y_i a X_i použili klasický lineární regresní model

$$E(Y_i) = \pi_i = x_i'\boldsymbol{\beta},$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ je $(p+1)$-rozměrný vektor neznámých regresních parametrů, které chceme odhadnout, výsledkem by byly odhady parametrů, u kterých nemáme zaručeno, že pro libovolné hodnoty vysvětlujících proměnných $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ bude hodnota parametru π_i ležet v požadovaném intervalu $(0, 1)$. Našim úkolem je najít spojovací funkci $g(\mu_i)$ tak, aby zobrazovala interval (0,1) na množinu $\mathbb{R}$. Pro spojovací funkci $g(\mu_i) = \mu_i$, platí $\eta_i = \mu_i$. Tato identická funkce se používá v případě klasického lineárního regresního modelu.

Spojovací funkci, která transformuje průměr na přirozený parametr, nazýváme kanonický link (anglicky *canonical link*). Pro tuto funkci platí

$$g(\mu_i) = Q(\theta_i),$$

$$Q(\theta_i) = \sum_j \beta_j x_{ij}.$$

Nyní musíme převést hustotu alternativního rozdělení do tvaru pro hustotu tříd funkcí exponenciálního typu – přirozený parametr $Q(\theta)$ se nachází v exponentu funkce. Odvození je následující

$$f(y, \pi) = \pi^y (1-\pi)^{1-y} = (1-\pi)\left(\frac{\pi}{1-\pi}\right)^y = (1-\pi)\, e^{\left(y \ln \frac{\pi}{1-\pi}\right)}$$

Tedy $\theta = \pi$, $a(\theta) = (1-\pi)$, $b(y) = 1$ a přirozený parametr $Q(\theta) = \ln(\frac{\pi}{1-\pi})$.
V případě, že máme model logistické regrese s jednou vysvětlující proměnnou, transformace $\pi(x)$, která se nazývá logitová transformace, je definována následovně

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x.$$

Konkrétní podoba modelu logistické regrese, který budeme používat je

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Důležitost logitové transformace spočívá ve vlastnostech funkce $g(x)$. Logit má několik žádoucích vlastností lineárního regresního modelu – $g(x)$ je lineární ve svých parametrech, spojitý a může nabývat hodnot od $-\infty$ po ∞, v závislosti na rozsahu x. Zajímavým rozdílem mezi (klasickým) lineárním regresním modelem a modelem logistické regrese je rozdíl v podmíněném rozdělení vysvětlované proměnné. V případě lineární regrese předpokládáme, že pozorování vysvětlované proměnné je vyjádřeno jako

$$y = E(y|x) + \varepsilon,$$

kde ε představuje chybu, která se řídí normálním rozdělením s nulovou střední hodnotou a konstantním rozptylem. V situaci, kdy vysvětlovaná proměnná nabývá pouze dvou hodnot (1 nebo 0), hodnota vysvětlované proměnné (za daného x) je ve tvaru

$$y = \pi(x) + \varepsilon.$$

Chyba ε může (při pevně dané hodnotě x) nabývat dvou různých hodnot – pokud $y = 1$ (tenista vyhrál zápas), potom je chyba $\varepsilon = 1 - \pi(x)$ s pravděpodobností $\pi(x)$ a pokud $y = 0$ (tenista prohrál zápas), hodnota chyby ε bude rovna $-\pi(x)$ s pravděpodobností $1 - \pi(x)$. Chyba ε se řídí rozdělením s nulovou střední hodnotou a rozptylem $\pi(x)(1 - \pi(x))$. [11]

3.2. Odhad parametrů

Než se pustíme do interpretace parametrů logistické regrese, což je jedna z nejzajímavějších částí celé logistické regrese, seznámíme se s tím, jak se tyto parametry odhadují. Naneštěstí situace již nebude tak jednoduchá, jako v případě lineárního regresního modelu, kde nám stačila metoda nejmenších čtverců – závislost střední hodnoty na regresních parametrech již není lineární. Metoda odhadu parametrů modelu logistické regrese

$$\pi(x_i) = \frac{e^{\sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}}}$$

se nazývá metoda maximální věrohodnosti (anglicky *maximum likelihood estimation*, MLE). Zjednodušeně, tato metoda vrací hodnoty neznámých parametrů, které maximalizují pravděpodobnost získání sledovaného souboru dat. Nejdříve musíme vytvořit věrohodnostní funkci (anglicky *likelihood function*), která uvažuje pravděpodobnost napozorovaných dat jako funkci neznámých parametrů.

Mějme n_i nezávislých pozorování ($n_1 + \cdots + n_N = n$) s fixními hodnotami p vysvětlujících proměnných $x_i = (1, x_{i1}, \ldots, x_{ip})$ pro $i = 1, \ldots, N$, kde y_i značí počet úspěchů (počet přítomností sledovaného znaku). Potom $Y_1, \ldots, Y_N$ jsou nezávislé náhodné veličiny řídící se binomickým rozdělením se střední hodnotou $E(Y_i) = n_i\pi(x_i)$. Sdružená pravděpodobnostní funkce těchto náhodných veličin je přímo úměrná součinu

$$\prod_{i=1}^{N} \pi(x_i)^{y_i}[1-\pi(x_i)]^{n_i-y_i} = \prod_{i=1}^{N} e^{\ln(\frac{\pi(x_i)}{1-\pi(x_i)})^{y_i}} \prod_{i=1}^{N}[1-\pi(x_i)]^{n_i}$$

$$= e^{\sum_{i=1}^{N} y_i \ln \frac{\pi(x_i)}{1-\pi(x_i)}} \prod_{i=1}^{N}[1-\pi(x_i)]^{n_i}.$$

Připomeňme, že i-tý logit můžeme vyjádři jako $\sum_j \beta_j x_{ij}$. Proto je výraz

$$e^{\sum_{i=1}^{N} y_i \ln \frac{\pi(x_i)}{1-\pi(x_i)}} \text{ roven } e^{\sum_j (\sum_i y_i x_{ij})\beta_j}.$$

Dále využijeme vlastnost

$$1 - \pi(x_i) = \tfrac{1}{1+e^{\sum_j \beta_j x_{ij}}}.$$

Po úpravách a zjednodušení pomocí zmiňovaných vlastností dostaneme log věrohodnostní funkci

$$L(\boldsymbol{\beta}) = \sum_j \left(\sum_i y_i x_{ij} \right) \beta_j - \sum_i n_i \ln\left(1 + e^{\sum_j \beta_j x_{ij}}\right).$$

Pro nalezení maxima této věrohodností funkce je třeba provést parciální derivace podle jednotlivých složek β a tyto parciální derivace položit rovno 0,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i x_{ij} \frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{p} \beta_k x_{ik}}}, \quad j = 0, \ldots, p.$$

Regresní parametry $\boldsymbol{\beta}$ odhadujeme z $p+1$ normálních rovnic,

$$\sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i \hat{\pi} x_{ij} = 0 \quad j = 0, 1, 2, \ldots, p,$$

kde jsme využili rovnosti

$$\hat{\pi}_i = \frac{e^{\sum_{k=0}^{p} \hat{\beta}_k x_{ik}}}{1 + e^{\sum_{k=0}^{p} \hat{\beta}_k x_{ik}}}.$$

Označíme-li si matici hodnot x_{ij} jako $\mathbf{X}$, normální rovnice lze zapsat v maticovém tvaru jako

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}},$$

kde $\hat{\boldsymbol{\mu}} = n_i \hat{\pi}_i$, a normální rovnice mají tvar

$$\sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \hat{\mu} x_{ij} = 0 \quad j = 0, 1, 2, \ldots, p.[10]$$

Jelikož jsou normální rovnice nelineární v parametrech, jejich řešení se hledá iteračně – např. pomocí Newton-Raphsonovy metody, která je blíže popsaná v [10] na stranách 143-145.

3.3. Varianční matice odhadu regresních parametrů

Pro další účely textu je vhodná znalost varianční matice odhadu parametrů. Varianční matici odhadu získáme jako inverzi informační matice $\mathbf{I}$. Prvky informační matice $\mathbf{I}$ jsou dány vztahem $-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}\right)$ pro $r, s = 0, 1, \ldots, p$.

První parciální derivaci $\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_r}$ jsme si již odvodili výše. Přistoupíme k druhé parciální derivaci, tentokrát derivujeme podle proměnné β_s

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{N} x_{ir} x_{is} n_i \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}} \left(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}\right) - e^{\sum_{k=1}^{p} \beta_k x_{ik}} e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{\left(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}\right)^2} =$$

$$= -\sum_{i=1}^{N} n_i x_{ir} x_{is} \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}} \frac{1}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}}.$$

Stejně jako u výpočtu první parciální derivace využijeme rovnosti

$$\pi_i = \frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1+e^{\sum_{k=0}^{p} \beta_k x_{ik}}} \text{ a } 1 - \pi\left(\mathbf{x}_i\right) = \frac{1}{1+e^{\sum_j \beta_j x_{ij}}},$$

a dostaneme výsledný tvar prvků informační matice $\mathbf{I}$

$$\frac{\partial^2 L\left(\boldsymbol{\beta}\right)}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{N} x_{ir} x_{is} n_i \pi_i \left(1 - \pi_i\right).$$

Informační matici tedy zapisujeme ve tvaru

$$\mathbf{I} = \mathbf{X}'\mathbf{W}\mathbf{X}, \qquad \text{kde } \mathbf{W}=\text{diag}\left[n_i \pi_i \left(1 - \pi_i\right)\right].$$

Odhad varianční matice pak získáme jako inverzi informační matice, tedy

$$\widehat{\mathbf{var}}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}. \text{ [11]}$$

3.4. Intervaly spolehlivosti pro parametry β_j a testování významnosti parametrů

Odhad varianční matice odhadu regresních parametrů využijeme k výpočtu intervalů spolehlivosti pro jednotlivé odhady regresních parametrů β_j. Interval spolehlivosti pro regresní parametr β_j využívá asymptotické normality maximálně věrohodného odhadu a má tvar

$$\left\langle \hat{\beta}_j - u_{(1-\frac{\alpha}{2})} \sqrt{\widehat{\text{var}}(\hat{\boldsymbol{\beta}})_{jj}}; \hat{\beta}_j + u_{(1-\frac{\alpha}{2})} \sqrt{\widehat{\text{var}}(\hat{\boldsymbol{\beta}})_{jj}} \right\rangle,$$

kde $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})_{jj}$ značí j-tý diagonální prvek odhadu varianční matice a $u_{(1-\frac{\alpha}{2})}$ je $(1 - \frac{\alpha}{2})$ kvantil normovaného normálního rozdělení $N(0, 1)$.

Poté co odhadneme hodnoty regresních parametrů, dalším krokem je položení si otázky, zda je konkrétní vysvětlující proměnná významná (jestli má její přítomnost v modelu vliv na hodnoty vysvětlované proměnné). Testování významnosti proměnných obvykle zahrnuje formulaci a testování nějaké statistické hypotézy. Pomocí odhadu varianční matice můžeme testovat i významnost parametru β_j,

kdy nulovou hypotézu testujeme ve tvaru $H_0 : \beta_j = 0$. Nulová hypotéza tedy říká, že j-tá vysvětlující proměnná není významná – její hodnoty nemají vliv na hodnoty vysvětlované proměnné. Signifikantnost vysvětlující proměnné testujeme pomocí Waldovy testové statistiky z,

$$z = \frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}})_{jj}}}$$

Za platnosti nulové hypotézy má druhá mocnina z, z^2, asymptoticky χ^2 rozdělení o jednom stupni volnosti. Kritický obor statistiky z^2 je tvaru W=$\langle \chi^2_{1-\alpha}; \infty)$. [10]

V případě, že chceme testovat nulovost všech parametrů v modelu najednou, použijeme analogickou podobu Waldova testu. Nulová hypotéza má tvar $H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$ a testová statistika

$$Z = \hat{\boldsymbol{\beta}}' \left[\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}})\right]^{-1} \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}' \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right) \hat{\boldsymbol{\beta}},$$

má asymptoticky χ^2 rozdělení s $p+1$ stupni volnosti. [11]

Další možností testování významnosti vysvětlující proměnné je použití testu poměru věrohodností (anglicky *Likelihood Ratio Test*). Tento test využívá, jak je zřejmé z jeho názvu, věrohodnostní funkce, popsané v sekci Odhad parametrů. Principem tohoto testu je porovnání dvou modelů – úplného (anglicky *full*) model, který zahrnuje všechny parametry a redukovaného (anglicky *reduced*) model. Redukovaný model získáme tak, že položíme námi „vybrané“ parametry (jejich počet označíme jako r) rovno nule. Testová statistika D (z anglického *deviance*) je tvaru

$$D = -2 \ln L_{reduced} - (-2 \ln L_{full}) = -2 \ln \frac{L_{reduced}}{L_{full}},$$

kde $L_{reduced}$ představuje maximum věrohodnostní funkce redukovaného modelu a L_{full} maximum věrohodnostní funkce úplného modelu. Tato statistika se za platnosti redukovaného modelu řídí χ^2 rozdělením s r stupni volnosti. [12]

3.5. Interpretace parametrů

V případě lineární regrese je interpretace regresních parametrů jednoduchá. Uvažujme situaci, kdy máme pouze jednu vysvětlující proměnnou, tedy model $E(y(x)) = \beta_0 + \beta_1 x$. Koeficient β_1 je roven rozdílu hodnoty vysvětlované proměnné pro $x + 1$ a hodnoty vysvětlované proměnné pro x, $\beta_1 = y\,(x + 1) - y\,(x)$. Interpretace takového koeficientu je přímočará – hodnota β_1 představuje průměrnou změnu hodnoty vysvětlované proměnné při změně vysvětlující proměnné o (jednu) jednotku.

V modelu logistické regrese představuje β_1 změnu logitu při současné změně vysvětlující proměnné o (jednu) jednotku, $\beta_1 = g(x + 1) - g(x)$, kde $g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x$.

Nejprve si představíme interpretaci pro model logistické regrese, kde vysvětlující proměnná x je dichotomická neboli nabývá pouze dvou hodnot – 1 a 0. V rámci praktického použití v této diplomové práci si jako příklad můžeme uvést proměnnou `Seeded` – zda byl hráč nasazen, nebo proměnnou `Qualifier` – zda hráč prošel kvalifikačním turnajem. Rozdíl logitů pro $x = 1$ a $x = 0$ je

$$g\,(1) - g\,(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1.$$

Abychom mohli co nejlépe interpretovat vypočítaný rozdíl logitů, musíme se seznámit s pojmem poměr šancí (anglicky *odds ratio*). Možné hodnoty modelu logistické regrese pro dichotomickou vysvětlující proměnnou jsou zobrazeny v tabulce 3.1.

Šance na „přítomnost“ zkoumaného jevu (tenista vyhrál zápas) mezi objekty, u kterých je hodnota vysvětlující proměnné rovna 1 (tenisté byli nasazeni), je definovaná jako $\frac{\pi(1)}{1-\pi(1)}$. Obdobně, šance na přítomnost zkoumaného jevu mezi objekty, u kterých je hodnota vysvětlující proměnné rovna 0 (tenisté nebyli nasazeni), je definována jako $\frac{\pi(0)}{1-\pi(0)}$.

Vysvětlovaná proměnná (Y)	**Vysvětlující proměnná (X)**	
	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$

Tabulka 3.1: Model logistické regrese pro dichotomickou vysvětlující proměnnou

Potom poměr šancí (OR) můžeme definovat následujícím způsobem

$$\mathrm{OR} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(1+e^{\beta_0+\beta_1}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(1+e^{\beta_0}\right)} =$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

Proto pro model logistické regrese s dichotomickou vysvětlující proměnnou je vztah mezi poměrem šancí OR a regresním koeficientem β_1, $\mathrm{OR} = e^{\beta_1}$. Poměr šancí je hojně využíván například v epidemiologii, uvedeme si jednoduchý příklad. Jestliže y značí přítomnost nebo absenci rakoviny plic a x značí, zda je daný člověk kuřák, potom poměr šancí $\widehat{\mathrm{OR}} = 2$ udává, že rakovina plic nastane u kuřáka s dvakrát větší šancí než u nekuřáka. Při užití poměru šancí narážíme na problém, že odhad $\widehat{\mathrm{OR}}$ má často zešikmenou distribuci. Šikmost je způsobena tím, že škála možných hodnot je od 0 po ∞, ale hodnota $\widehat{\mathrm{OR}} = 1$, která značí neexistenci rozdílu mezi skupinami, představuje takzvanou nulovou hodnotu této distribuce. Teoreticky, pro dostatečně velké soubory, distribuční funkce $\widehat{\mathrm{OR}}$ je normální. Naneštěstí nároky dostatečného počtu pozorování jsou v praxi často nedosažitelné. Proto se často používá spíše $\ln\left(\widehat{\mathrm{OR}}\right) = \hat{\beta}_1$, kdy se distribuční funkce log poměru šancí přibližně řídí normálním rozdělením i pro relativně malé soubory.

Konfidenční interval pro poměr šancí obdržíme tak, že nejdříve vypočítáme koncové body intervalu pro koeficient β_1 a poté tyto hodnoty odlogaritmujeme.

Výsledný $100 \times (1-\alpha)\%$ konfidenční interval má podobu

$$e^{\hat{\beta}_1 \pm u_{\left(1-\frac{\alpha}{2}\right)} \widehat{SE}\left(\hat{\beta}_1\right)}$$

kde $\widehat{SE}\left(\hat{\beta}_1\right)$ značí odhadnutou směrodatnou odchylku odhadu koeficientu β_1 a $u_{\left(1-\frac{\alpha}{2}\right)}$ značí $\left(1-\frac{\alpha}{2}\right)$ kvantil normálního normovaného rozdělení.

Nyní se podíváme na situaci, kdy vysvětlující proměnná může nabývat více než dvou různých hodnot, které nemůžeme vzájemně nijak porovnat. V našem sledovaném datovém souboru takovou proměnnou nemáme, ale mohli bychom mít příklad, pokud bychom brali v úvahu typ povrchu – tvrdý povrch, antuka, tráva. Zvolíme si jednu ze tří hodnot vysvětlující proměnné jako referenční skupinu (např. tvrdý povrch), referenční skupinu snadno identifikujeme – její poměr šancí je roven jedné. Dále musíme vytvořit takzvané pomocné proměnné (anglicky *dummy variables*). Pomocné proměnné by pro náš příklad měly následující tvar

	Pomocné proměnné	
Surface	**Surface_1**	**Surface_2**
Hard (1)	0	0
Clay (2)	1	0
Grass(3)	0	1

Model pro tento teoretický příklad je tvaru $g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, kde x_1 a x_2 představují pomocné proměnné.

Poměr šancí pro antuku „vůči“ referenční skupině tvrdý povrch bychom vypočítali následujícím způsobem

$$\ln\left[\hat{\mathrm{OR}}\left(Clay, Hard\right)\right] = \hat{g}\left(Clay\right) - \hat{g}\left(Hard\right) =$$

$$= \left[\hat{\beta}_0 + \hat{\beta}_1 \times (\text{Surface\_1} = 1) + \hat{\beta}_2 \times (\text{Surface\_2} = 0)\right] -$$

$$- \left[\hat{\beta}_0 + \hat{\beta}_1 \times (\text{Surface\_1} = 0) + \hat{\beta}_2 \times (\text{Surface\_2} = 0)\right] =$$

$$= \hat{\beta}_1$$

Obdobně bychom vyjádřili i poměr šancí pro antuku a trávu.
Posledním případem, kterým se budeme v rámci interpretace parametrů zabývat, je spojitá vysvětlující proměnná. V naší analýze si pod spojitou vysvětlující proměnnou můžeme představit výšku tenisty – proměnnou `Height`. Uvažujme, že výška je jediná vysvětlující proměnná, máme tedy model $g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x$. Bude-li výška tenisty nulová, z modelu nám zůstane pouze absolutní člen (anglicky *intercept*) β_0. Tento parametr je roven logitu šance na výhru oproti prohře, za podmínky, že tenista měří 0 centimetrů,

$$\ln \frac{\pi\left(x_1 = 0\right)}{1 - \pi\left(x_1 = 0\right)} = \beta_0.$$

V této situaci postrádá smysl pracovat s celou škálou výšky tenisty. Jako výchozí „nulovou“ hodnotu můžeme použít například průměr nebo medián výšky. Pro každé pozorování (pro každého tenistu) si vytvoříme novou proměnnou $\widetilde{x}_i = x_i - \bar{x}_i$ pro $i = 1, \ldots, n$. Interpretace parametru β_0 při použití centrované vysvětlující proměnné by byla následující: logit šance na výhru tenisty oproti prohře, za podmínky že tenista měří $\bar{x}$ (průměrná výška). [11]
Interpretaci β_1 si ukážeme pomocí šance. Při zvolené hodnotě x je šance rovna

$$\omega\left(x\right) = \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{\beta_0+\beta_1 x}}} = e^{\beta_0+\beta_1 x}.$$

Porovnáme-li šance výsledku sledovaného jevu (tenista vyhrál/nevyhrál zápas) pro dvě hodnoty vysvětlující proměnné x (výšky tenisty), které se liší o jednotku, dostaneme

$$\frac{\omega\left(x+1\right)}{\omega\left(x\right)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}.$$

Získaný výsledek tedy můžeme interpretovat jako poměr šancí, kdy se hodnota proměnné x změní o jednotku (tj. tenista je vyšší o jednotku (centimetr), případně výška tenisty je o jednu jednotku větší než průměrná výška, případně medián výšky).

Máme-li složitější model s více vysvětlujícími (spojitými) proměnnými, tak

zafixujeme hodnoty všech vysvětlujících proměnných kromě té, jejíž parametr chceme interpretovat (tuto proměnnou si označíme x_i). Vyjádříme si šanci pro zafixované hodnoty p vysvětlujících proměnných při hodnotě x i-té vysvětlující proměnné a šanci pro zafixované hodnoty p vysvětlujících proměnných při hodnotě $x+1$ i-té vysvětlující proměnné. [13]

3.6. Hodnocení kvality modelu

3.6.1. Pearsonova χ^2 statistika a deviance

V lineární regresi měříme vhodnost modelu pomocí rozdílu mezi pozorovanými hodnotami a odhadnutými hodnotami, $y - \hat{y}$. V logistické regresi existuje několik možných způsobů měření rozdílů mezi odhadnutými a napozorovanými hodnotami. Předpokládejme, že náš vytvořený model obsahuje p nezávislých proměnných, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ a J značí počet různých napozorovaných hodnot $\mathbf{x}$. Pokud mají některá pozorování stejné hodnoty $\mathbf{x}$, tak $J < n$. Počet pozorování s hodnotami $\mathbf{x} = \mathbf{x}_j$ označíme m_j pro $j = 1, 2, 3, \ldots, J$, počet všech pozorování potom je $\sum m_j = n$. Počet výskytů pozorovaného znaku, $y = 1$, mezi m_j pozorováními s hodnotami $\mathbf{x} = \mathbf{x}_j$ si označíme jako y_j. Celkový počet pozorování s výskytem pozorovaného znaku označíme n_1, kde $\sum y_j = n_1$.

Abychom zdůraznili, že v logistické regresi jsou odhadnuté hodnoty počítány pro každou kombinaci regresorů v datovém souboru, odhadneme proto pravděpodobnost těchto kombinací pomocí odhadnutých regresních parametrů a odhadnutou hodnotu pro j-tou kombinaci regresorů označíme jako

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1 + e^{\hat{g}(\mathbf{x}_j)}},$$

kde $\hat{g}(\mathbf{x}_j)$ je odhad logitu. Pearsonova rezidua konkrétního souboru hodnot regresorů jsou definována následovně

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \quad pro = 1, \ldots, J.$$

Souhrnná statistika založená na těchto reziduích je Pearsonova χ^2 statistika

$$X^2 = \sum_{j=1}^{J} r\left(y_j, \hat{\pi}_j\right)^2 .$$

Odchylka reziduí je definována jako

$$d\left(y_j, \hat{\pi}_j\right) = \pm\sqrt{2\left(y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + \left(m_j - y_j\right)\ln\left(\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right)\right)},$$

kde znaménko + nebo − je stejné jako znaménko $(y_j - m_j\hat{\pi}_j)$. Laicky řečeno, odchylku reziduí modelujeme pomocí očekávaného počtu jedniček a nul dle modelu a pomocí skutečného počtu jedniček a nul.

Pro soubory hodnot regresorů pozorování, u kterých se nevyskytl pozorovaný znak, $y_j = 0$, je odchylka reziduí následující

$$d\left(y_j, \hat{\pi}_j\right) = -\sqrt{2m_j|\ln\left(1-\hat{\pi}_j\right)|}$$

a odchylka reziduí pro $y_j = m_j$ je

$$d\left(y_j, \hat{\pi}_j\right) = \sqrt{2m_j|\ln\left(\hat{\pi}_j\right)|}.$$

Souhrnná statistika založená na rozptylu reziduí je deviance

$$D = \sum_{j=1}^{J} d\left(y_j, \hat{\pi}_j\right)^2 .$$

Statistika X^2 a D se za platnosti predikovaného modelu řídí χ^2 rozdělením s $J - (p+1)$ stupni volnosti. Statistika D je vlastně test poměru věrohodností saturovaného modelu s J parametry vůči modelu s p+1 parametry. Problém nastane když $J \approx n$, kdy počet parametrů roste stejnou měrou jako velikost sledovaného vzorku – p-hodnoty vypočítané pro obě statistiky jsou chybné. Tomuto problému se můžeme vyhnout, když data seskupíme do m menších skupin (většinou $m = 10$) – příkladem mohou být Hosmer-Lemeshowy testy, více si můžete přečíst například v [11] na stranách 147 až 156.

3.6.2. Akaikeho informační kritérium

Obdobou deviance definované v sekci Intervaly spolehlivosti pro parametry β_j a testování významnosti parametrů je Akaikeho informační kritérium AIC. Kritérium počítá s věrohodnostní funkcí modelu a oproti devianci zohledňuje i počet parametrů p tohoto modelu. AIC je dáno vztahem

$$AIC = -2(\ln L - p),$$

kde L značí maximální hodnotu věrohodnostní funkce příslušeného modelu. V případě, kdy máme možnost si vybrat z více modelů, vybíráme model s co nejnižší hodnotou tohoto kritéria, tj. s co největší věrohodností a co nejmenším počtem parametrů. [10]

3.6.3. Klasifikační tabulka

Dá se říci, že v podstatě řešíme klasifikační úlohu, kdy chceme jednotlivá pozorování zařadit do jedné ze dvou kategorií – hráč zvítězí nebo hráč prohraje. Pokud pro nás nejsou moc důležité hodnoty regresních parametrů, ale zajímá nás především počet správně a špatně zařazených pozorování, můžeme přistoupit k použití klasifikační tabulky. Při zařazování pozorování do dvou kategorií (standardní situace u logistické regrese) musíme určit takzvaný „cut point“ c. Pokud odhadnutá pravděpodobnost přesahuje hodnotu c, potom odhadneme hodnotu vysvětlované proměnné na 1, v opačném případě 0. Nejčastěji se používá hodnota 0.5, ale ve specifických situacích je možné použití i jiné hodnoty. Klasifikační tabulka má následující podobu

		Predikovaný výsledek		
		Výhra (1)	Prohra (0)	
Skutečný výsledek	Výhra (1)	30 (TP)	5 (FN)	35
	Prohra (0)	10 (FP)	24 (TN)	34
		40	29	

Tabulka 3.2: Příklad klasifikační tabulky

V následující tabulce si vysvětlíme, co znamenají jednotlivé zkratky užité v příkladu klasifikační tabulky. S využitím těchto zkratek zavedeme následující pojmy.

Zkratka	**Název**	**Vysvětlení**
TP	True Positive	správně zařazená pozorování, u kterých nastal výskyt sledovaného znaku
TN	True Negative	správně zařazená pozorování, u kterých nenastal výskyt sledovaného znaku
FN	False Negative	špatně zařazená pozorování, u kterých nastal výskyt sledovaného znaku
FP	False Positive	špatně zařazená pozorování, u kterých nenastal výskyt sledovaného znaku

Tabulka 3.3: Vysvětlení pojmů souvisejících s klasifikační tabulkou

Odhad senzitivity se vypočítá jako $\mathrm{TP}/\left(\mathrm{TP}+\mathrm{FN}\right)$, senzitivita udává schopnost správně určit pozorování, u kterých nastal sledovaný znak.
Odhad specificity vypočítáme jako $\mathrm{TN}/\left(\mathrm{FP}+\mathrm{TN}\right)$, senzitivita udává schopnost správně určit pozorování, u kterých sledovaný znak nenastal. Dalším vhodným ukazatelem, který se u klasifikačních tabulek používá, je *Positive predictive value* (PPV) a *Negative predictive value* (NPV). Tyto hodnoty vypočítáme jako

$$\mathrm{PPV}=\mathrm{TP}/\left(\mathrm{TP}+\mathrm{FP}\right),\qquad \mathrm{NPV}=\mathrm{TN}/\left(\mathrm{FN}+\mathrm{TN}\right).$$

PPV odhaduje pravděpodobnost, že sledovaný jev opravdu nastane, pokud byl predikován a NPV odhaduje pravděpodobnost, že sledovaný jev nenastane, pokud nebyl predikován. Všechny tyto pojmy jsou hojně využívány v zdravotnické statistice.

V příkladu z tabulky 3.2 bylo správně zařazeno 78 % pozorování (součet prvků na diagonále, $30+24$ vydělený počtem všech pozorování, 69). Hodnoty dalších ukazatelů vyšly následovně.

Ukazatel	Výpočet	Hodnota
Senzitivita	$30/(30+5)$	0.86
Specificita	$24/(24+10)$	0.71
PPV	$30/(30+10)$	0.75
NPV	$24/(24+5)$	0.83

Problémem u většiny metod klasifikace je, že pokud bychom použili stejná pozorování jako při tvorbě modelu, tak stejná data použijeme dvakrát (soubor pozorování si můžeme rozdělit na trénovací a testovací část, přijdeme tím ale o část informace). V praxi to vede ke sklonu k „lepším“ odhadům, než kdybychom použili naprosto nová pozorování. Nová pozorování tedy použijeme k validaci našeho modelu. Součástí této diplomové práce je validace (a hledání) nejlepšího modelu pomocí nových pozorování – zápasů za tenisovou sezónu 2018. [14]

3.6.4. ROC křivka

Senzitivita a specificita závisí na jediném „cut pointu“, který rozhoduje o klasifikaci pozorování. Další popis přesnosti klasifikace přináší plocha pod ROC (Receiver Operating Characteristic) křivkou. Tato křivka, původem z teorie detekce signálu, ukazuje, jak přijímač detekuje existenci signálu za přítomnosti šumu. Křivka vykresluje pravděpodobnost detekce „pravého“ signálu (*senzitivita*) a falešného signálu (*1-specificita*) pro škálu všech možných cut pointů.

Plocha pod touto křivkou (nabývá hodnot od 0 po 1) popisuje míru schopnosti modelu rozlišovat mezi přítomností sledovaného znaku a jeho nepřítomností – budeme používat pojem diskriminace, který je pro schopnost „rozlišování“ modelu používán. Chceme-li najít optimální cut point pro účely klasifikace, měli bychom zvolit cut point, který maximalizuje součet hodnot senzitivity a specificity. Tento bod se nachází v místě, kde jsou si senzitivita a specificita rovny. V následující tabulce vidíme, jak dobré diskriminace jsme dosáhli, na základě velikosti plochy pod ROC křivkou.

Hodnota ROC	**Výsledek**
$= 0.5$	značí nepřítomnost diskriminace (stejného výsledku dosáhneme, pokud budeme házet mincí)
$0.7 \leq \text{ROC} < 0.8$	tento výsledek představuje přijatelnou diskriminaci
$0.8 \leq \text{ROC} < 0.9$	tento výsledek představuje výbornou diskriminaci
≥ 0.9	tento výsledek představuje výjimečnou diskriminaci

V praxi je téměř nemožné narazit na model s výjimečnou diskriminací. Další možností pochopení plochy pod ROC křivkou je následující; označme si jako n_1 počet přítomností sledovaného znaku, n_0 počet nepřítomností sledovaného znaku, vytvoříme $n_1 \times n_0$ dvojic – každé pozorování, kde se vyskytl sledovaný znak spárujeme s každým pozorováním, kde se sledovaný znak nevyskytl. Z těchto $n_1 \times n_0$ dvojic určíme míru, kdy pozorování, u kterého nastal pozorovaný znak, mělo vyšší pravděpodobnost. Tato míra odpovídá ploše pod ROC křivkou. [11]

Kapitola 4

Hledání modelu

Nyní se dostáváme k hledání jednotlivých modelů logistické regrese, které na základě vybraných statistik dokáží co nejlépe předpovědět, který ze dvou tenistů vyhraje tenisový zápas. Základní informace týkající se jednotlivých datových souborů jsme si představili v úvodní kapitole této práce. Celkem tedy máme k dispozici dvanáct datových souborů (pro 12 turnajů), kdy jednotlivé datové soubory obsahují dvě nebo tři menší části – datové soubory pro jednotlivá kola. Grandslamových turnajů se v roce 2017 zúčastnilo 193 různých hráčů a turnajů Masters se v roce 2017 zúčastnilo 145 různých hráčů. K hledání modelů byl použit software `R`, v textu naleznete ukázky použitého kódu. Skripty obsahující hledání modelů naleznete na přiloženém CD v příslušných složkách `NahodnyVyber` a `Dohromady`.

Většina datových souborů obsahovala všechny hodnoty, pouze v několika případech chyběl údaj o výšce tenisty, což ale nepředstavovalo závažný problém. Naopak problematická pozorování, která byla z každého datového souboru odstraněna, nastala kvůli jedné ze tří možných příčin: nebyl znám kurz sázkových kanceláří, výsledek zápasu byl rozhodnut odstoupením jednoho z hráčů během utkání nebo jeden z hráčů odstoupil z turnaje před začátkem zápasu. Celkové počty pozorování (jak z pohledu vítěze zápasu, tak z pohledu poraženého) jsou zaznamenány v následující tabulce.

Turnaj/Kolo	R128	R64	R32
Australian Open	112	60	30
French Open	122	60	22
Wimbledon	114	62	28
US Open	120	62	28
Indian Wells Masters	62	62	30
Miami Open	62	60	28
Madrid Open	–	46	32
Italian Open	–	46	32
Canadian Open	–	48	32
Cincinnati Masters	–	48	32
Shanghai Masters	–	48	32
Paris Masters	–	32	32

Tabulka 4.1: Počty pozorování pro jednotlivé datové soubory sezóny 2017

Různé počty zápasů (a kol) jsou způsobeny odlišnostmi jednotlivých turnajů a počtu nasazených hráčů, kteří obdrželi přímý postup do druhého kola (viz Game, set, match).

4.1. Porušení předpokladů

V datových souborech máme k dispozici výsledky zápasů (a vlastnosti tenistů) jak z pohledu vítěze, tak z pohledu poraženého. Pokud bychom obě pozorování použili pro hledání modelu, porušili bychom základní předpoklad logistické regrese – nezávislost jednotlivých pozorování. Například pro první kolo Australian Open bychom při použití všech pozorování (56 zápasů z pohledu vítěze i poraženého, tj. 112 záznamů) dostali model s jinými signifikantními parametry než při použití pouze 56 záznamů, jak je ukázáno na následujícím obrázku. Horní model byl vytvořen ze 112 pozorování a oba parametry popisující závislost pravděpodobnosti vítězství jsou významné (p-hodnota Waldovy testové statistiky je menší než hodnota 0,05). Dolní model vznikl použitím pouze 56 pozorování, kdy jsme vybrali jedno pozorování z každého zápasu, a významný parametr je jen jeden – proměnná `H2HWon`.

```
Call:
glm(formula = Vysledek ~ H2HWon + Kurz, family = binomial(),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.26742  -0.08162   0.02097   0.08005   2.03532

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1216     1.6844  -2.447 0.014408 *
H2HWon       12.0026     3.2753   3.665 0.000248 ***
Kurz         -0.8456     0.3985  -2.122 0.033860 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
glm(formula = Vysledek ~ H2HWon + Kurz, family = binomial(),
    data = data2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28337  -0.07766  -0.00063   0.07866   1.50496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2523     2.3468  -1.812  0.06999 .
H2HWon       11.9545     4.5696   2.616  0.00889 **
Kurz         -0.7260     0.5268  -1.378  0.16814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obrázek 4.1: Porovnání dvou modelů – první model využívá všechna pozorování, druhý model pouze jedno pozorování z každého zápasu

Jako řešení tohoto problému, kdy nemůžeme použít všechna dostupná pozorování, jsme uvažovali dva přístupy. První přístup spočívá v použití pouze jednoho záznamu pro každý zápas, kdy náhodně vybereme popis výsledku zápasu buď z pohledu vítěze nebo z pohledu poraženého. Tím ale přicházíme o část informace, obsažené ve statistikách protihráče. Druhý přístup využívá všechny dostupné informace, náhodně vybereme „domácího“ hráče (hráč číslo 1) a výsledek zápasu budeme posuzovat z jeho pohledu, ale zároveň budou v pozorování obsaženy i statistiky týkající se „hostujícího“ hráče (hráč číslo 2). Druhý přístup ale obstál jen v teoretické rovině, při hledání jednotlivých významných parametrů byla p-hodnota Waldovy testové statistiky téměř vždy výrazně vyšší než hodnota 0,05 jak pro charakteristiky „domácího“ hráče, tak pro charakteristiky „hostujícího“ hráče.

Další komplikací, kterou není vhodné opomenout, je skutečnost, že kdybychom vzali zápasy všech kol na jednom turnaji dohromady, tak v jednotlivých pozorováních budou vystupovat stejní hráči. Museli bychom uvažovat náhodný efekt jednotlivých hráčů, čímž by výrazně vzrostla složitost modelů. Proto jsme se

rozhodli primárně vytvářet modely pro jednotlivá kola na turnajích a sledovat, jaké faktory se stávají významnými, čím „hlouběji“ hráči postupují turnajem. Dále se pokusíme vytvořit modely pro jednotlivá kola jednotlivých typů turnajů (grandslamové turnaje a turnaje typu Masters s 56 hráči) – zde dojde k porušení nezávislosti, kdy se jednotlivých turnajů mohou účastnit stejní hráči. Porušení nezávislosti je ale méně závažnější než při uvažování všech kol na jednom turnaji dohromady. Výsledky těchto modelů bychom proto měli brát s rezervou, avšak větší rozsahy datových souborů nám mohou přinést lepší modely, což si snadno ověříme v kapitole Validace modelu.

4.2. Postup hledání modelu

Jelikož faktory, které mohou ovlivnit výsledek zápasu, jsme si stanovovali (a dohledali) sami, nemáme zaručeno, že opravdu mají vliv na výsledek zápasu. Proto jsme nejdříve hledali jednotlivé významné proměnné a poté jsme zkoušeli jejich vzájemné kombinace. V softwaru `R` hledáme model logistické regrese pomocí příkazu `fit=glm(`$y \sim x_1 + x_2 + \ldots$`,data=data,family=binomial)`, kdy jako výstup dostaneme odhady koeficientů regresních parametrů, p-hodnotu test významnosti parametrů, Akaikeho informační kritérium (AIC) a další statistiky spojené s modelem. Při posuzování významnosti parametrů se orientujeme pomocí p-hodnoty náležící Waldově testové statistice z. Za významné parametry považujeme parametry, pro něž je p-hodnota testové statistiky z menší než hodnota 0,05.

I když se některé proměnné nabízely k hledání interakcí (např. proměnné `History+Starts`), tak ve většině případů byl model s interakcemi horší než model bez interakcí, který uvažoval pouze tu „významnější“ proměnnou z dané dvojice. Při hledání nejlepšího modelu jsme se orientovali převážně podle Akaikeho informačního kritéria a dále i podle plochy pod ROC křivkou. Hodnoty plochy pod ROC křivkou jsme ale uvažovali pouze orientačně, jelikož „validaci“ provádíme na stejných datech, na kterých jsme model vytvořili, tudíž dochází ke zkreslení

odhadu kvality klasifikace.

4.3. Náhodný výběr jednoho pozorování z každého zápasu

Nejdříve budeme vytvářet modely tak, že náhodně vybereme jedno z dvojice pozorování pro každý zápas. Tuto náhodnost jsme vytvořili vygenerováním n hodnot z alternativního rozdělení, kde n značí počet zápasů v jednotlivých kolech (poté, co jsme odstranili neúplná pozorování – zápasy, které nebyly dohrány a zápasy, které se vůbec neodehrály).

4.3.1. Grandslamové turnaje

Jako první se zaměříme na grandslamové turnaje, kde hledáme modely pro první tři kola turnaje – R128, R64 a R32. Významné parametry pro modely bivariantní analýzy prvních tří kol všech grandslamových turnajů jsou zaznamenány v následující tabulce. V závorce jsou udány p-hodnoty Waldovy testové statistiky z, hodnoty v tabulce jsou seřazeny podle p-hodnoty.

Australian Open		
R128	R64	R32
H2HWon (0,00693)	Seeded (0,00154)	H2HWon (0,0031)
Kurz (0,0089)	H2HWon (0,00196)	Kurz (0,036)
Seeded (0,00733)	Ranking (0,01128)	
MWon (0,0317)	MPlayed (0,0132)	
MPlayed (0,0454)	Vek (0,0294)	
	MWon (0,0349)	
	Starts (0,037)	
	Kurz (0,0545)	
French Open		
H2HWon (0,00154)	MPlayed (0,00246)	GPlayed (0,122)
Ranking (0,00509)	Ranking (0,00696)	
Seeded (0,00567)	Seeded (0,00797)	
Kurz (0,0242)	H2HWon (0,0519)	
MPlayed (0,0577)	GPlayed (0,0708)	
MWon (0,0936)		
Qualifier (0,0974)		
Wimbledon		
H2HWon (0,000334)	H2HWon (0,00426)	H2HWon (0,068)
Kurz (0,001233)	Seeded (0,0107)	MPlayed (0,0701)
Seeded (0,0032)	Kurz (0,0152)	H2HPlayed (0,0925)
MWon (0,00748)	Ranking (0,01816)	
MPlayed (0,00963)	MPlayed (0,0186)	
Ranking (0,0359)	MWon (0,0413)	
US Open		
Kurz (0,00802)	H2HWon (0,0101)	Ranking (0,0635)
MPlayed (0,0368)	MWon (0,0242)	
Ranking (0,0445)	MPlayed (0,0463)	
Seeded (0,0721)	Ranking (0,0505)	
MWon (0,0899)		

Tabulka 4.2: Významné parametry bivariantní analýzy pro jednotlivá kola grandslamových turnajů

Z tabulky lehce vyčteme, že proměnná `Kurz` patří mezi významné parametry u všech prvních kol grandslamových turnajů. Avšak v druhém kole už se objevila jen u dvou ze čtyř turnajů a ve třetím kole jen u jednoho grandslamového turnaje. Ve významnosti jednoznačně vede proměnná `H2HWon`, která vznikla jako podíl `H2HV` (případně `H2HP` u tenisty, co prohrál zápas) a `H2HPlayed`, což je celkový počet vzájemně odehraných zápasů mezi dvěma tenisty.

U třetího kola můžeme pozorovat nevýznamnost většiny proměnných, nastalá situace může být způsobena menším počtem odehraných zápasů a tudíž zápasy, které jsme použili k tvorbě modelu, mohly být více ovlivněny náhodou. „Nevýznamnost“ parametrů v třetím kole může být způsobena i tím, že do třetího kola běžně postupují vesměs dobří hráči, kdy rozdíly mezi protihráči jsou (až na výjimky) minimální. I přes to, že se při tvorbě modelů pro třetí kola nemáme čeho chytit, zkusíme modely vytvořit.

Přidaná proměnná v dalších kolech turnajů, `GPlayed`, se projevila pouze u druhého kola a potažmo třetího kola French Open. Jako zajímavé se rozhodně jeví druhé kolo Australian Open, kdy se mezi významné parametry zařadily proměnné `Vek` a `Starts`. V prvním kole French Open můžeme pozorovat i slabý vliv toho, zda hráč prošel kvalifikací – má ale tato skutečnost pozitivní nebo negativní dopad na šanci na výhru tenisty? Výsledný model (kdy bychom rozhodovali jen na základě toho, zda hráč prošel kvalifikací nebo ne) by byl následujícího tvaru

$$\pi\left(\mathbf{x}\right) = \frac{e^{0,1542-1,4069\cdot x}}{1+e^{0,1542-1,4069\cdot x}},$$

kde x je proměnná, která nabývá hodnot 1 nebo 0 (hráč absolvoval kvalifikační boje nebo hráč neabsolvoval kvalifikační boje). Poměr šancí pro hráče, který prošel a neprošel kvalifikací, je roven $\widehat{\mathrm{OR}} = e^{-1,4069}$ – tedy šance hráče, který prošel kvalifikací, na výhru zápasu je 5krát menší (převrácená hodnota k $e^{-1,4069} = 0,245$) než šance na výhru zápasu u hráče, který neprošel kvalifikací. Výsledný negativní efekt je překvapivý, předpokládali jsme, že hráči, kteří prošli kvalifikací mají výhodu rozehranosti a dokážou soupeře překvapit. V části Interpretace parametrů si ukážeme další interpretace vybraných (a složitějších) modelů.

Nyní si představíme čtveřici nejlepších modelů pro tři analyzovaná kola jednotlivých grandslamových turnajů, které použijeme v následující kapitole k validaci. Modely jsme vybírali na základě Akaikeho informačního kritéria (AIC), kdy čím nižší hodnota Akaikeho informačního kritéria, tím lepší model. Hodnoty AIC jsou uvedeny v závorce. Všechny modely byly vytvářeny na stejných náhodných výběrech, kdy jsme využili funkce `set.seed`.

Australian Open		
R128	R64	R32
`H2HWon+Kurz` (26,17)	`H2HWon` (21,406)	`H2HWon` (14,566)
`H2HWon` (26,21)	`H2HWon+Ranking` (22,594)	`H2HWon+Vyska` (16,551)
`Ranking+H2HWon` (28,13)	`H2HWon+Seeded` (23,089)	`H2HWon+MWon` (16,566)
`Seeded+H2HWon` (28,21)	`H2HWon+Vek` (23,175)	`Vyska+Ranking` (17,067)
French Open		
`H2HWon` (26,272)	`H2HWon+Seeded` (13,887)	`Kurz` (13,609)
`H2HWon+Qual.` (27,549)	`H2HWon` (13,91)	`GPlayed+MPlayed` (14,03)
`MPlayed+H2HWon` (27,87)	`H2HWon+GPlayed` (14,073)	`Kurz+MPlayed` (14,522)
`MWon+H2HWon` (28,101)	`H2HWon+Starts` (15,764)	`GPlayed+Kurz` (14,727)
Wimbledon		
`H2HWon+Seeded` (26,171)	`H2HWon+Ranking` (25,356)	`MPlayed+Vyska` (11,968)
`H2HWon` (26,991)	`H2HWon+Seeded` (25,621)	`H2HWon+H2HPlayed` (12,993)
`H2HWon+Seeded+` `MWon` (27,728)	`H2HWon+Kurz` (26,401)	`H2HPlayed+Kurz` (14,604)
`H2HWon+Seeded+` `MPlayed` (27,909)	`H2HWon` (26,445)	`H2HWon+Vyska+` `H2HPlayed` (14,87)
US Open		
`H2HWon+Kurz` (39,746)	`H2HWon` (20,917)	`Ranking` (16,206)
`H2HWon+Ranking` (40,614)	`H2HWon+Kurz` (21,607)	`Ranking+Kurz` (17,027)
`H2HWon` (40,838)	`H2HWon+MPlayed` (22,569)	`Ranking+Kurz+` `MWon` (17,542)
`H2HWon+Kurz+` `Ranking` (41,158)	`H2HWon+Ranking` (22,81)	`Ranking+Seeded` (18,078)

Tabulka 4.3: Nejlepší modely pro jednotlivá kola grandslamových turnajů

4.3.2. Turnaje Masters

Nyní se zaměříme na hledání modelů pro turnaje typu Masters. Situace se zde komplikuje, jelikož existují různé typy turnajů Masters. Pro turnaje Indian Wells Masters a Miami Open hledáme modely pro první tři kola – R128, R64 a R32. Pro zbývající turnaje hledáme modely pouze pro první dvě kola, tj. R64 a R32. Jako první si ukážeme významné parametry bivariantních modelů pro turnaje Indian Wells Masters a Miami Open, v závorce jsou uvedeny p-hodnoty Waldovy testové statistiky z.

Indian Wells Masters		
R128	**R64**	**R32**
`H2HWon` (0,0301)	`H2HWon` (0,00178)	`H2HWon` (0,0811)
`H2HPlayed` (0,0318)	`GPlayed` (0,00657)	`Vyska` (0,108)
	`Bye` (0,00702)	
	`Kurz` (0,00716)	
	`Ranking` (0,01662)	
	`MPlayed` (0,0384)	
	`MWon` (0,0865)	
	`H2HPlayed` (0,087)	
Miami Open		
`H2HWon` (0,0071)	`H2HWon` (0,00496)	`H2HWon` (0,058)
`MPlayed` (0,0138)	`Ranking` (0,0451)	`Vyska` (0,0721)
`Kurz` (0,0282)	`Kurz` (0,0507)	`MWon` (0,0957)
`Starts` (0,0989)	`MWon` (0,0578)	

Tabulka 4.4: Významné parametry bivariantní analýzy pro turnaje Indian Wells Masters a Miami Open

Jak vidíme, tak významné parametry těchto turnajů se mezi sebou značně liší. Zajímavá je absence proměnných `Bye` (určitá část nasazených hráčů obdrží volný vstup do druhého kola, tudíž nehraje zápas v prvním kole) a `GPlayed` (počet odehraných her (gamů) v předchozích kolech turnaje) ve druhém kole Miami Open, což může značit nepřítomnost vlivu odehraných (`GPlayed`) a neodehraných (`Bye`) zápasů. Stejně jako u grandslamových turnajů, vliv vzájemných zápasů mezi tenisty jednoznačně dominuje významným parametrům.

Podíváme se, jaká je situace u šesti zbývajících turnajů typu Masters.

Madrid Open	
R64	**R32**
H2HWon (0,0592) MWon (0,105)	MWon (0,0499) Vek (0,0572) Kurz (0,0585) H2HWon (0,063) MPlayed (0,0852) GPlayed (0,0975)
Italian Open	
H2HWon (0,0154) Ranking (0,0535) MPlayed (0,0769)	Ranking (0,044) MWon (0,0539) MPlayed (0,0601) Vek (0,086) H2HWon (0,0919)
Canadian Open	
Kurz (0,0594)	H2HWon (0,126)
Cincinnati Masters	
H2HWon (0,014)	Kurz (0,117)
Shanghai Masters	
Kurz (0,0985)	MPlayed (0,0301) MWon (0,055) H2HWon (0,0569) Kurz (0,0733) H2HPlayed (0,074) Vek (0,0789) Ranking (0,080)
Paris Masters	
MWon (0,0519) H2HWon (0,0662) MPlayed (0,0805)	H2HWon (0,0627) MPlayed (0,0949)

Tabulka 4.5: Významné parametry bivariantní analýzy pro zbývající turnaje Masters

Je patrné, že významné parametry se mezi jednotlivými turnaji liší. Překvapením je přítomnost proměnné `Vek` u třech turnajů. Zajímavé je také to, že počet odehraných her, `GPlayed`, se objevil jen u jednoho turnaje a volný postup do druhého kola, `Bye`, se mezi významné parametry vůbec nedostal.
Nyní si představíme trojici nejlepších modelů pro jednotlivá kola Indian Wells Masters a Miami Open. Číslo v závorce udává hodnotu Akaikeho informačního kritéria.

Indian Wells Masters		
R128	**R64**	**R32**
`H2HWon+H2HPlayed` (15,244)	`H2HWon` (27,901)	`H2HPlayed` (17,253)
`H2HWon` (15,259)	`H2HWon+Kurz` (25,593)	`H2HWon+H2HPlayed` (17,67)
`H2HWon*H2HPlayed` (17,18)	`H2HWon+MPlayed` (29,691)	`H2HWon` (18,784)
Miami Open		
`H2HWon` (17,575)	`H2HWon` (23,632)	`Kurz` (8,817)
`H2HWon+Kurz` (18,303)	`H2HWon+Kurz` (25,13)	`H2HWon` (10,633)
`H2HWon+Starts` (19,262)	`H2HWon+Ranking` (25,265)	`Vyska+MWon` (13,328)

Tabulka 4.6: Nejlepší modely pro turnaje Indian Wells Masters a Miami Open

Rozhodli jsme se volit pouze tři modely, protože u některých turnajů byl problém najít vůbec nějaké významné parametry, tudíž čtvrtý model by často dosahoval velice nízké úspěšnosti klasifikace. V poslední části kapitoly Náhodný výběr jednoho pozorování z každého zápasu si představíme trojici nejlepších modelů pro zbylé turnaje Masters.

Madrid Open	
R64	**R32**
H2HWon (13,407)	Vek+Kurz (12,945)
H2HWon+MWon (14,753)	H2HWon+GPlayed (12,953)
H2HWon+Ranking (15,092)	Kurz+GPlayed (13,223)
Italian Open	
H2HWon+Ranking+MPlayed (18,122)	H2HWon (10,277)
H2HWon+Ranking (20,062)	Ranking+H2HWon (12,233)
H2HWon+Kurz (20,348)	MPlayed+H2HWon (12,258)
Canadian Open	
H2HWon (11,348)	H2HWon+Vyska (21,661)
H2HWon+Vyska (11,352)	Vyska (21,923)
Kurz+H2HWon (13,073)	Starts (22,17)
Cincinnati Masters	
H2HWon (21,877)	H2HWon (9,967)
H2HWon+Starts (22,248)	Kurz+H2HWon (10,446)
H2HWon+MWon (23,826)	H2HWon+Ranking (11,685)
Shanghai Masters	
Kurz (32,74)	H2HWon (10,218)
Starts+Kurz (33,035)	MPlayed+H2HWon (11,892)
Starts+Home (33,047)	H2HWon+Kurz (11,984)
Paris Masters	
MWon+H2HWon (12,256)	H2HWon (16,987)
H2HWon+MPlayed (13,4)	H2HWon+Kurz (18,078)
MWon+H2HWon+MPlayed (14,196)	H2HWon+MPlayed (18,206)

Tabulka 4.7: Nejlepší modely pro zbývající turnaje Masters

4.4. Souhrnné modely pro jednotlivá kola vybraných turnajů

V této části kapitoly Hledání modelu se budeme zabývat hledáním modelů pro jednotlivá kola vybraných typů turnajů. Budeme uvažovat dva typy turnajů, grandslamové turnaje a turnaje typu Masters s 56 účastníky, kterých je celkem pět[5], viz tabulka Masters turnaje a počet hráčů, kteří obdrží volný vstup do druhého kola. Použijeme opět náhodný výběr jednoho pozorování pro každý zápas (buď z pohledu vítěze nebo z pohledu poraženého), ale tato pozorování z každého turnaje sloučíme dohromady, tudíž dojde k porušení jednoho z předpokladů modelu, nezávislosti pozorování. Hlavním cílem těchto souhrnných modelů je zjistit, jak moc se výsledné modely liší oproti modelům pro jednotlivé turnaje a zda-li je porušení předpokladů vůbec závažné. Stejně jako v předchozí části kapitoly si vygenerujeme n hodnot z alternativního rozdělení (n značí počet zápasů konkrétního kola na turnaji). Po vygenerování jednotlivých souborů pozorování (kdy v případě hodnoty 1 zvolíme zápas z pohledu vítěze, v případě hodnoty 0 zápas z pohledu poraženého) sloučíme soubory do jednoho velkého souboru pro konkrétní kolo.

4.4.1. Grandslamové turnaje

Jako první se zaměříme na grandslamové turnaje, kdy budeme hledat tři modely pro tři kola – R128, R64, R32. V tabulce níže najdeme nejvýznamnější parametry bivariantní analýzy těchto tří kol, v závorce jsou uvedeny p-hodnoty Waldovy testové statistiky.

[5]technicky vzato je těchto turnajů šest, turnaj Monte Carlo ale není součástí analýzy

R128	R64	R32
H2HWon $(6,91 \cdot 10^{-13})$	H2HWon $(1,85 \cdot 10^{-8})$	H2HWon $(3,9 \cdot 10^{-4})$
Seeded $(3,09 \cdot 10^{-7})$	Seeded $(4,12 \cdot 10^{-7})$	Kurz $(2,77 \cdot 10^{-3})$
Kurz $(3,58 \cdot 10^{-7})$	MPlayed $(6,71 \cdot 10^{-7})$	Ranking $(7.34 \cdot 10^{-3})$
Ranking $(1,21 \cdot 10^{-5})$	Ranking $(8,17 \cdot 10^{-7})$	MWon $(7.78 \cdot 10^{-3})$
MPlayed $(1,9 \cdot 10^{-5})$	Kurz $(1,15 \cdot 10^{-5})$	MPlayed $(3.39 \cdot 10^{-2})$
MWon $(4.37 \cdot 10^{-5})$	MWon $(9,06 \cdot 10^{-4})$	
	Qualifier $(7,63 \cdot 10^{-3})$	
	Starts $(3,16 \cdot 10^{-2})$	

Tabulka 4.8: Významné parametry bivariantní analýzy pro souhrnné modely grandslamových turnajů

Mezi významnými parametry najdeme téměř stejné proměnné, jako u modelů pro jednotlivé grandslamové turnaje. Zajímavá je přítomnost proměnné `Qualifier` a `Starts` – jejich interpretaci si ukážeme v následující sekci. V následující tabulce najdeme přehled čtyř nejlepších modelů pro jednotlivá kola všech zápasů odehrávajících se na grandslamových turnajích. V závorce najdeme hodnoty Akaikeho informačního kritéria, na jehož základě jsme modely vybírali.

R128	R64	R32
H2HWon*H2HPlayed+Kurz+ Seeded (103,8)	H2HWon+Ranking (67,258)	H2HWon+Kurz (37,756)
H2HWon+Kurz (107,94)	H2HWon+MPlayed+ Ranking (68,886)	H2HWon (37,797)
H2HWon*H2HPlayed+Kurz+ Seeded+MPlayed*MWon (108,38)	H2HWon+Kurz+ Ranking (69,245)	H2HWon+Kurz+ MWon (39,259)
H2HWon+Kurz+ Seeded (109,6)	H2HWon+MPlayed (70,547)	H2HPlayed+Ranking (39,452)

Tabulka 4.9: Nejlepší modely pro jednotlivá kola všech grandslamových zápasů

4.4.2. Turnaje typu Masters

Předmětem další analýzy budou zápasy na turnajích typu Masters s 56 účastníky, zaměříme se pouze na dvě kola – R64 a R32. V následující tabulce jsou společně s nejdůležitějšími parametry vyobrazeny p-hodnoty Waldova testu.

R64	R32
H2HWon $(7,26 \cdot 10^{-7})$	H2HWon $(4,74 \cdot 10^{-6})$
Seeded $(1,93 \cdot 10^{-3})$	Kurz $(4,23 \cdot 10^{-4})$
Kurz $(8,93 \cdot 10^{-3})$	MWon $(5,85 \cdot 10^{-4})$
Ranking $(1,53 \cdot 10^{-2})$	MPlayed $(1,3 \cdot 10^{-3})$
MPlayed $(9,75 \cdot 10^{-2})$	Ranking $(2,07 \cdot 10^{-3})$
MWon $(1,09 \cdot 10^{-1})$	Seeded $(5,1 \cdot 10^{-3})$
	Qualifier $(8,79 \cdot 10^{-3})$
	Bye $(2,11 \cdot 10^{-2})$

Tabulka 4.10: Významné parametry bivariantní analýzy pro souhrnné modely Masters turnajů s 56 účastníky

V prvním kole se mezi významné parametry zařadily stejné proměnné jako v případě grandslamových turnajů. V druhém kole vybraných turnajů Masters se mezi významné proměnné přidaly proměnné Qualifier a Bye, všechny proměnné z prvního kola zůstaly mezi významnými proměnnými. Přehled čtyř nejlepších modelů pro dvě první kola turnajů typu Masters nalezneme v následující tabulce, kdy číslo v závorce udává hodnotu Akaikeho informačního kritéria.

R64	R32
H2HWon+Seeded+Ranking (66,024)	H2HWon+Kurz+Seeded (56,386)
H2HWon+Seeded+Ranking+Kurz (66,951)	H2HWon (57,334)
H2HWon+Kurz+Seeded (67,078)	H2HWon+Seeded (57,636)
H2HWon (67,576)	H2HWon+Kurz+Seeded+MWon (58,257)

Tabulka 4.11: Nejlepší modely pro jednotlivá kola vybraných Masters turnajů

4.5. Interpretace parametrů

V poslední části této kapitoly si ukážeme interpretaci vybraných zajímavých modelů představených na předchozích stránkách. Začneme interpretací jednoduchých modelů s jednou vysvětlující proměnnou, kde nás zajímá, zda má proměnná pozitivní nebo negativní vliv na výsledek zápasu, a jak velký tento efekt je.

První model (z části modely pro všechny grandslamové turnaje dohromady), kdy by jedinou vysvětlující proměnnou byla proměnná `Qualifier`, by byl následujícího tvaru:

$$\pi\left(\mathrm{x}\right) = \frac{e^{0,7621-1,8608\cdot x}}{1+e^{0,7621-1,8608\cdot x}},$$

kde x značí, zda hráč prošel kvalifikací (1) nebo ne (0). Poměr šancí pro hráče, který prošel a neprošel kvalifikací, je roven $\widehat{\mathrm{OR}} = 0,155$ – šance hráče, který prošel kvalifikací, na výhru zápasu je více než 6krát menší (převrácená hodnota $0,155$), než šance na výhru zápasu u hráče, který kvalifikaci neabsolvoval.

Porovnáme-li poměr šancí u druhého modelu (rovněž z části pro všechny zápasy z grandslamových turnajů dohromady) s jedinou vysvětlující proměnnou `Starts`, dostaneme následující rovnost

$$\frac{\omega\left(x+1\right)}{\omega\left(x\right)} = \frac{e^{-0,14687+0,09973\cdot(x+1)}}{e^{-0,14687+0,09973\cdot x}} = e^{0,09973} = 1,105.$$

Šance na výhru u tenisty, který má počet minulých účastí na turnaji větší (o jednu účast) než druhý tenista je přibližně o 10 procent vyšší – patrně rozdíl jedné účasti na turnaji má sice kladný vliv na výhru tenisty, ale velikost tohoto vlivu není příliš velká.

V modelu pro pět vybraných turnajů typu Masters se v druhém kole mezi důležité proměnné zařadila proměnná `Bye`, která je pro turnaje typu Masters typická. Model s touto vysvětlující proměnnou by byl následujícího tvaru

$$\pi(\mathbf{x}) = \frac{e^{0,1699+1,5647\cdot x}}{1+e^{0,1699+1,5647\cdot x}},$$

kde x značí, zda hráč obdržel volný postup do druhého kola (1) nebo musel absolvovat zápas prvního kola (0). Poměr šancí $\widehat{\mathrm{OR}} = e^{1,565} = 4,78$ – šance na výhru hráče, který obdržel volný postup do druhého kola, je téměř 5krát větší než šance na výhru hráče, který volný vstup do druhého kola neobdržel. Tento kladný efekt volného postupu do druhého kola jsme předpokládali, jelikož hráči jsou více odpočatí, a především se tato skutečnost týká hráčů, kteří jsou nasazení, tedy špičky světového tenisu.

Pro model `H2HWon+Seeded+Ranking`, který jsme stanovili jako jeden z možných modelů pro první kolo vybraných Masters turnajů, kde $\mathbf{x} = (x_1, x_2, x_3)$ představuje vektor regresorů, dostaneme následující rovnost

$$\ln \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = -4,72 + 10,06 \cdot x_1 + 2,31 \cdot x_2 - 0,02 \cdot x_3,$$

kde x_1 náleží proměnné `H2HWon` a nabývá hodnot z intervalu $\langle 0; 1\rangle$, x_2 náleží proměnné `Seeded` a nabývá hodnoty 1 pro nasazené hráče a hodnoty 0 pro nenasazené hráče a x_3 náleží proměnné `Ranking`, která nabývá hodnot od 1 až po 200.

Nejdříve se podíváme na hodnotu absolutního členu β_0, jeho odhad $\hat{\beta}_0 = -4,72$ je roven logaritmu šance na výhru tenisty oproti prohře za situace, kdy jsou všechny vysvětlující proměnné nulové. Šance při nulovosti regresorů by byla rovna 0,0089, což znamená, že prohra by byla přibližně 112krát pravděpodobnější než výhra.

Hodnota parametru β_1, s odhadem $\hat{\beta}_1 = 10,06$ představuje logaritmus poměru šancí mezi nasazenými hráči s totožnými hodnotami proměnné `Ranking`, kdy se poměr vyhraných vzájemných zápasů vůči všem odehraným vzájemným zápasům liší o 1 (tedy hráč vyhrál všechny vzájemné zápasy). Poměr šancí na výhru při rozdílu 1 u proměnné `H2HWon` je rovna $e^{10,06}$, výhra hráče, který vyhrál všechny

vzájemné zápasy by byla více než 20 000krát pravděpodobnější než výhra hráče, který všechny vzájemné zápasy prohrál.

Pro hodnotu parametru β_2, jejíž odhad $\hat{\beta}_2 = 2,31$ je roven logaritmu poměru šancí mezi hráči se stejnou hodnotou proměnných `H2HWon` a `Ranking`, kdy rozdíl mezi hráči je způsoben tím, zda je hráč nasazen nebo ne[6]. Šance na výhru při rozdílu nasazenosti hráčů je rovna $e^{2,31}$, tedy výhra nasazeného hráče je přibližně 10krát pravděpodobnější než výhra nenasazeného hráče.

Poslední parametr β_3, s odhadem $\hat{\beta}_3 = -0,02$ je roven logaritmu poměru šancí mezi hráči se stejnou hodnotou proměnných `H2HWon` a `Seeded`, s rozdílem postavení na žebříčku o jednu pozici. Šance na výhru při rozdílu postavení na žebříčku o jednu pozici je rovna $e^{-0,02}$, tedy výhra hráče, který je postaven níže na žebříčku než jeho protihráč, se s každým dalším rozdílem v postavení sníží o 2 %.

[6]V praxi taková situace logicky nastat nemůže, neexistují dva hráči se stejným postavením na žebříčku a navíc „nasazenost“ přímo souvisí s pořadím na žebříčku

Kapitola 5

Validace modelu

V této kapitole se dostáváme k validaci jednotlivých modelů, tudíž k zjištění, jak „dobře“ naše modely dokázaly předpovědět vítěze tenisových zápasů. V předcházející kapitole jsme vytvořili tři nebo čtyři modely pro jednotlivé turnaje, případně pro turnaje stejného typu. Tyto modely (odhady regresních parametrů těchto modelů) použijeme na nová pozorování, tenisové zápasy za sezónu 2018. Opět máme k dispozici 12 datových souborů, kdy jednotlivé datové soubory obsahují dvě nebo tři menší části pro jednotlivá kola. Z datových souborů byly odstraněny neúplné záznamy, které nastaly kvůli stejným příčinám, jak bylo vysvětleno v předchozí kapitole. Celkově počty pozorování pro jednotlivé datové soubory týkající se sezóny 2018 jsou zaznamenány v následující tabulce.

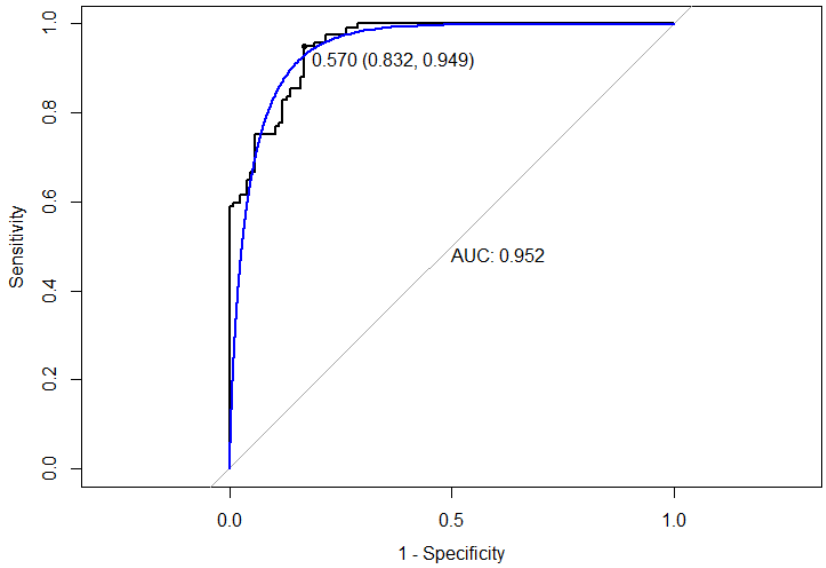
Turnaj/Kolo	**R128**	**R64**	**R32**
Australian Open	126	62	32
French Open	126	64	32
Wimbledon	124	64	32
US Open	110	58	32
Indian Wells Masters	62	64	28
Miami Open	64	64	32
Madrid Open	–	48	32
Italian Open	–	46	26
Canadian Open	–	46	32
Cincinnati Masters	–	48	32
Shanghai Masters	–	48	32
Paris Masters	–	30	26

Tabulka 5.1: Počty pozorování pro jednotlivé datové soubory sezóny 2018

Při předpovědi vítěze zápasu jsme zvolili přístup, který jsme využili už při tvorbě modelů – náhodně jsme vybrali jednoho ze dvou hráčů tenisového zápasu, a u vybraného hráče jsme na základě napozorovaných hodnot odhadli pravděpodobnost jeho vítězství. V případě našich modelů by nebyl součet pravděpodobnosti výhry u vítěze a u poraženého roven jedné, proto jsme zvolili přístup náhodného výběru. Odhad pravděpodobnosti výhry v zápase jsme získali pomocí příkazu

```
glm.probs=predict(fit1,newdata=data, type="response"),
```

kde `fit1` představuje vytvořený model na základě dat za sezónu 2017. Poté bylo potřeba určit takzvaný „cut-point“, c: pokud vypočtená pravděpodobnost výhry v zápase přesáhne hodnotu c, tak daného tenistu označíme jako vítěze zápasu. Pokud je vypočtená pravděpodobnost výhry v zápase menší než c, tak tenistu označíme jako poraženého. K určení této prahové hodnoty jsme užili ROC křivky, kdy jsme využili balíčku `pROC`[15]. Ukázku ROC křivky, s prahovou hodnotou c 0,57 a s plochou pod ROC křivkou (AUC) 0,952 naleznete níže. Vyhlazená ROC křivka vznikla užitím binormálního vyhlazování.



Obrázek 5.1: Ukázka ROC křivky pro souhrnný model prvního kola grandslamových turnajů

Poté, co určíme nejvhodnější cut point, jednoduše rozhodneme, které tenisty jsme určili jako vítěze, a které jako poražené, užitím například následujícího příkazu

```
glm.pred=ifelse(glm.probs >= 0.57, "Win", "Loss").
```

Abychom zjistili, jak dobré klasifikace jsme dosáhli, vytvoříme si klasifikační tabulku příkazem

```
tabulka=table(glm.pred, data$Vysledek),
```

kdy výsledná tabulka má následující podobu:

```
          Prediction
glm.pred   0   1
    Loss 111  21
    Win    6 104
```

Procento správně zařazených pozorování získáme jako součet prvků na diagonále vydělený součtem všech prvků v tabulce. Všechny modely byly validovány na stejných datových souborech, kdy při náhodném výběru pomocí alternativního rozdělení byla použita funkce `set.seed`.

5.1. Modely s náhodným výběrem jednoho pozorování

V této části se podíváme, jak dobré klasifikace jsme dosáhli u modelů, které využívaly náhodný výběr jednoho pozorování a podíváme se, který model (případně modely) dosáhly nejlepších výsledků. Nejdříve se budeme věnovat grandslamovým turnajům, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC. Příslušné kódy pro validaci modelů jsou k dispozici na přiloženém CD ve složce `NovaData`.

Australian Open		
R128	**R64**	**R32**
H2HWon+Kurz 90,47 % (0,963)	H2HWon 100 % (1)	H2HWon 100 % (1)
H2HWon 88,88 % (0,956)	H2HWon+Ranking 100 % (1)	H2HWon+Vyska 100 % (1)
Ranking+H2HWon 88,88 % (0,948)	H2HWon+Seeded 100 % (1)	H2HWon+MWon 100 % (1)
Seeded+H2HWon 88,88 % (0,958)	H2HWon+Vek 100 % (1)	Vyska+Ranking 56,25 % (0,6)
French Open		
H2HWon 80,95 % (0,890)	H2HWon+Seeded 96,77 % (0,991)	Kurz 100 % (1)
H2HWon+Qual. 80,95 % (0,900)	H2HWon 96,77 % (0,995)	GPlayed+MPlayed 87,50 % (0,917)
MPlayed+H2HWon 80,95 % (0,891)	H2HWon+GPlayed 96,77 % (0,986)	Kurz+MPlayed 81,25 % (0,867)
MWon+H2HWon 80,95 % (0,895)	H2HWon+Starts 96,77 % (0,991)	GPlayed+Kurz 93,75 % (0,983)
Wimbledon		
H2HWon+Seeded 84,12 % (0,937)	H2HWon+Ranking 90,32 % (0,977)	MPlayed+Vyska 75 % (0,700)
H2HWon 84,12 % (0,942)	H2HWon+Seeded 90,32 % (0,977)	H2HWon+H2HPlayed 81,25 % (0,808)
H2HWon+Seeded+MWon 87,30 % (0,940)	H2HWon+Kurz 83,87 % (0,927)	H2HPlayed+Kurz 37,50 % (0,500)
H2HWon+Seeded+MPlayed 85,71 % (0,934)	H2HWon 90,32 % (0,986)	H2HWon+Vyska+H2HPlayed 81,25 % (0,867)
US Open		
H2HWon+Kurz 98,18 % (0,983)	H2HWon 96,55 % (0,967)	Ranking 75 % (0,830)
H2HWon+Ranking 96,36 % (0,981)	H2HWon+Kurz 96,55 % (0,967)	Ranking+Kurz 81,25 % (0,850)
H2HWon 94,54 % (0,966)	H2HWon+MPlayed 96,55 % (0,961)	Ranking+Kurz+MWon 56,25 % (0,517)
H2HWon+Kurz+Ranking 96,36 % (0,983)	H2HWon+Ranking 100 % (1)	Ranking+Seeded 87,50 % (0,900)

Tabulka 5.2: Validace modelů pro jednotlivá kola grandslamových turnajů, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC (plochu pod ROC křivkou)

Z tabulky vyčteme, že některými modely jsme dosáhli přesné klasifikace. Jedná se především o turnaj Australian Open, kdy modely v druhém a třetím kole sice obsahovaly více vysvětlujících proměnných, ale vliv proměnné `H2HWon` byl natolik silný, že vypočtené pravděpodobnosti byly ovlivněny zbylými vysvětlujícími proměnnými minimálně.

V případě třetího kola French Open byl použit model s jedinou vysvětlující proměnnou `Kurz` a bylo dosaženo 100% úspěšné klasifikace. Někdy se může stát, že sázkové kanceláře přesně odhadnou vítěze zápasu. V případě 100% klasifikace druhého kola US Open byl použit model s proměnnými `H2HWon+Ranking`. Je patrné, že nejhorší (a především nejrozmanitější) výsledky nastaly ve třetích kolech grandslamových turnajů. Všechny modely prvních a druhých kol grandslamových turnajů v sobě zahrnovaly proměnnou `H2HWon`, a výsledná „diskriminace“ téměř vždy dosahuje hodnot vyšších než 0,90. Jakmile ale k predikci použijeme jiné proměnné, tak se úspěšnost klasifikace zhorší, ale například u druhého modelu třetího kola French Open, kdy jsme použili model s proměnnými `GPlayed+MPlayed`, jsme dosáhli téměř stejně dobrých výsledků jako při užití modelů s proměnnou `H2HWon`.

Indian Wells Masters		
R128	**R64**	**R32**
`H2HWon+H2HPlayed` 93,54 % (0,966)	`H2HWon` 93,75 % (0,989)	`H2HPlayed` 78,57 % (0,578)
`H2HWon` 93,54 % (0,977)	`H2HWon+Kurz` 90,62 % (0,987)	`H2HWon+H2HPlayed` 78,57 % (0,711)
`H2HWon*H2HPlayed` 93,54 % (0,966)	`H2HWon+MPlayed` 96,87 % (0,991)	`H2HWon` 85,71 % (0,822)
Miami Open		
`H2HWon` 90,62 % (0,978)	`H2HWon` 90,62 % (0,961)	`Kurz` 87,50 % (0,783)
`H2HWon+Kurz` 93,75 % (0,983)	`H2HWon+Kurz` 87,50 % (0,965)	`H2HWon` 87,50 % (0,95)
`H2HWon+Starts` 87,50 % (0,957)	`H2HWon+Ranking` 90,62 % (0,970)	`Vyska+MWon` 56,25 % (0,55)

Tabulka 5.3: Validace modelů pro turnaje Indian Wells Masters a Miami Open, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC (plochu pod ROC křivkou)

V tabulce 5.3 jsme se přesunuli k turnajům typu Masters, kde se nejdříve zaměříme na turnaje, kde jsme analyzovali tři kola. Stejně jako u modelů pro grandslamové turnaje i tady byly téměř vždy použity modely, které obsahovaly proměnnou `H2HWon`. Dvě výjimky představují první a třetí model třetího kola Miami Open, kde u prvního modelu byla použita pouze proměnná `Kurz` a dosáhli jsme uspokojivého výsledku, a u třetího modelu byly použité proměnné `Vyska+MWon` a výsledek klasifikace nebyl moc příznivý. Nejlepších výsledků bylo dosaženo kombinací proměnné `H2HWon` a některou z proměnných `Kurz, Ranking, Starts` nebo `MPlayed`.

Situace pro zbývající turnaje Masters, kde jsme analyzovali jen dvě kola, je zaznamenána v tabulce 5.4. Za zajímavý úkaz můžeme považovat první model u druhého kola Madrid Open, který vznikl pomocí proměnných `Vek+Kurz` a dosáhli jsme více než 93% úspěšnosti klasifikace, zatímco druhý model, který uvažoval `H2HWon+GPlayed`, dosáhl úspěšnosti pouze 56 %. Je možné, že v dalších kolech Madrid Open rozhodují více zkušenosti (věk) než jen vzájemné zápasy.

V případě prvních kol Italian Open a Shanghai Masters modely nedosáhly ani na úspěšnost 80 % a to jsme u Italian Open použili `H2HWon`, která u ostatních turnajů vedla k nadprůměrným výsledkům. První model pro Shanghai Masters využil proměnnou `Kurz` a dosáhl úspěšnosti jen 66 %, je tedy možné, že došlo k nestandardnímu průběhu turnaje, kdy ani sázkové kanceláře nedokázaly správně určit vítěze. Velice dobré předpovědi vznikly pro turnaj Cincinnati Masters, avšak druhý model pro druhé kolo tohoto turnaje, který využil proměnné `Kurz+H2HWon`, dosáhl o téměř 20 % menší úspěšnosti klasifikace než zbylé modely, které využily proměnné `H2HWon` a `H2HWon+Ranking`. Tudíž je možné, že sázkové kanceláře odhadly vítěze zápasů hůře než námi uvažovaná (a vytvořená) statistika `H2HWon`.

Madrid Open	
R64	**R32**
H2HWon 87,50 % (0,911)	Vek+Kurz 93,75 % (0,967)
H2HWon+MWon 83,33 % (0,926)	H2HWon+GPlayed 56,25 % (0,383)
H2HWon+Ranking 83,33 % (0,919)	Kurz+GPlayed 31,25 % (0,717)
Italian Open	
H2HWon+Ranking+MPlayed 73,91 % (0,683)	H2HWon 88,90 % (0,889)
H2HWon+Ranking 69,56 % (0,643)	Ranking+H2HWon 80 % (0,907)
H2HWon+Kurz 73,91 % (0,841)	MPlayed+H2HWon 86,67 % (0,889)
Canadian Open	
H2HWon 82,60 % (0,873)	H2HWon+Vyska 87,50 % (0,950)
H2HWon+Vyska 78,26 % (0,865)	Vyska 75 % (0,742)
Kurz+H2HWon 82,60 % (0,889)	Starts 37,50 % (0,617)
Cincinnati Masters	
H2HWon 91,66 % (0,922)	H2HWon 100 % (1)
H2HWon+Starts 91,66 % (0,93)	Kurz+H2HWon 81,25 % (0,75)
H2HWon+MWon 91,66 % (0,926)	H2HWon+Ranking 100 % (1)
Shanghai Masters	
Kurz 66,66 % (0,652)	H2HWon 93,75 % (0,925)
Starts+Kurz 37,50 % (0,570)	MPlayed+H2HWon 93,75 % (0,917)
Starts+Home 37,50 % (0,559)	H2HWon+Kurz 93,75 % (0,917)
Paris Masters	
MWon+H2HWon 86,66 % (0,889)	H2HWon 100 % (1)
H2HWon+MPlayed 93,33 % (0,944)	H2HWon+Kurz 92,30 % (0,925)
MWon+H2HWon+MPlayed 86,66 % (0,870)	H2HWon+MPlayed 92,30 % (0,950)

Tabulka 5.4: Validace modelů pro zbývající turnaje Masters, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC (plochu pod ROC křivkou)

5.2. Souhrnné modely pro jednotlivá kola vybraných turnajů

Nyní se zaměříme na modely vytvořené pro jednotlivá kola grandslamových turnajů a turnajů Masters s 56 účastníky. Tyto modely porušují jeden z předpokladů logistické regrese, více viz Porušení předpokladů. Nejprve se podíváme na validaci modelů pro první tři kola grandslamových turnajů. K dispozici máme mnohem větší soubor pozorování, větší vzorek zápasů, takže bychom mohli dosáhnout lepších výsledků, než když se soustředíme pouze na konkrétní turnaje. Procenta vyjadřují úspěšnost klasifikace a hodnota v závorce udává AUC, neboli plochu pod ROC křivkou. Příslušné kódy pro validaci modelů jsou k dispozici na přiloženém CD ve složce `NovaDataDohromady`.

R128	R64	R32
`H2HWon*H2HPlayed+Kurz+Seeded` 88,84 % (0,952)	`H2HWon+Ranking` 97,58 % (0,996)	`H2HWon+Kurz` 90,74 % (0,940)
`H2HWon+Kurz` 89,25 % (0,954)	`H2HWon+MPlayed+Ranking` 97,58 % (0,997)	`H2HWon` 87,03 % (0,932)
`H2HWon*H2HPl.+Kurz+Seeded+MPl.*MWon` 87,19 % (0,95)	`H2HWon+Kurz+Ranking` 97,58 % (0,996)	`H2HWon+Kurz+MWon` 92,59 % (0,951)
`H2HWon+Kurz+Seeded` 88,01 % (0,951)	`H2HWon+MPlayed` 97,58 % (0,997)	`H2HPlayed+Ranking` 74,07 % (0,775)

Tabulka 5.5: Validace modelů pro jednotlivá kola všech grandslamových zápasů, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC (plochu pod ROC křivkou)

Nejlepším modelem pro první kola grandslamových turnajů je model `H2HWon+Kurz`, který dosáhl téměř 90% úspěšnosti klasifikace. V druhém kole jsme dosáhli téměř 98% úspěšnosti klasifikace, kdy všechny modely uvažovaly proměnnou `H2HWon` a k ní jednu nebo dvě ze skupiny proměnných `Ranking, MPlayed` nebo `Kurz`. Silný vliv proměnné `H2HWon` ale neumožnil ostatním proměnným, aby ovlivnily vypočtenou pravděpodobnost na výhru, a proto byla úspěšnost klasifikace u všech modelů totožná. Pro třetí kola grandslamových turnajů nejlépe obstál model `H2HWon+Kurz+MWon`, kdy proměnnou `MWon` hodnotíme úspěšnost hráče za minulou sezónu.

V poslední části této kapitoly se budeme zabývat validací souhrnných modelů pro první dvě kola Masters turnajů s 56 účastníky. Úspěšnost klasifikace a oblast pod ROC křivkou, AUC, jsou zaznamenány v následující tabulce.

R64	**R32**
`H2HWon+Seeded+Ranking` 82,20 % (0,885)	`H2HWon+Kurz+Seeded` 88,60 % (0,929)
`H2HWon+Seeded+Ranking+Kurz` 82,20 % (0,886)	`H2HWon` 81,01 % (0,920)
`H2HWon+Kurz+Seeded` 83,05 % (0,886)	`H2HWon+Seeded` 79,74 % (0,919)
`H2HWon` 82,20 % (0,891)	`H2HWon+Kurz+Seeded+MWon` 88,60 % (0,929)

Tabulka 5.6: Validace modelů pro jednotlivá kola vybraných Masters turnajů, procenta udávají počet správně zařazených pozorování a číslo v závorce udává AUC (plochu pod ROC křivkou)

Nejlepším modelem pro první kolo Masters turnajů se stal model `H2HWon+Kurz+Seeded`, ale i ostatní modely, které obsahovaly proměnné `Seeded, Ranking` nebo `Kurz` dosáhly téměř stejné úspěšnosti. V druhém kole pro vybrané turnaje typu Masters byla úspěšnost klasifikace o něco málo vyšší. Dva nejlepší modely jsou modely `H2HWon+Kurz+Seeded` a `H2HWon+Kurz+Seeded+MWon`. Patrně, vliv proměnné `MWon` nebyl příliš velký, jelikož neovlivnil výsledky klasifikace.

Můžeme zde pozorovat zajímavý jev – na rozdíl od situace, kdy jsme měli k dispozici pouze zápasy z jednotlivých turnajů a modely obsahující proměnnou `H2HWon` a nějakou další proměnnou dosahovaly totožných výsledků, nyní se do popředí dostávají složitější modely s více proměnnými. V případě zápasů druhého kola je rozdíl mezi nejlepším modelem a modelem obsahujícím pouze proměnnou `H2HWon` více jak 7 procent.

Závěr

Cíle této diplomové práce bylo predikovat výsledky tenisových utkání pomocí logistické regrese na základě nejvhodnějšího modelu, který uvažuje pouze důležité faktory. Modely byly vytvářeny na základě údajů z tenisových utkání za sezónu 2017. Správnost vytvořených modelů jsem posléze ověřila na nových pozorováních, na tenisových utkáních za sezónu 2018.

V první kapitole jsem čtenáře seznámila s tenisem a jeho pravidly, s turnaji, které jsou předmětem analýzy, předvedla jsem ukázku jednoho z datových souborů a vysvětlila, které faktory (proměnné) chci pro tvorbu modelů uvažovat.
Ve druhé a třetí kapitole, které tvoří teoretickou podstatu této práce, jsem čtenářům přiblížila regresní modely, a především logistickou regresi, s důrazem na interpretaci parametrů a na hodnocení kvality modelu z pohledu řešení klasifikačního problému.

Ve čtvrté kapitole jsem se věnovala vytváření jednotlivých modelů, kdy jsem přiblížila problém porušení předpokladů logistické regrese, který měl značný vliv na výslednou podobu modelů. Dále jsem popsala, jakým způsobem jsem modely tvořila. Posléze jsem se věnovala samotné tvorbě modelů, kdy jsem zvolila dva odlišné přístupy. První přístup spočíval v náhodném výběru jednoho z dvojice pozorování po každý zápas jednotlivých kol u jednotlivých turnajů. Druhý přístup spočíval v tvorbě souhrnných modelů pro jednotlivá kola grandslamových turnajů a vybraných druhů Masters turnajů. Pro každý grandslamový turnaj jsem na základě Akaikeho informačního kritéria vybrala čtveřici nejlepších modelů a pro každý turnaj Masters jsem vybrala trojici nejlepších modelů. V poslední části čtvrté kapitoly jsem se věnovala interpretaci parametrů, jak u jednoduchých mo-

delů s jednou vysvětlující proměnnou, tak jednoho složitějšího modelu, s třemi vysvětlujícími proměnnými.

V páté, závěrečné kapitole jsem na základě čtyř (případně tří) vytvořených modelů provedla validaci na nových pozorováních. Výsledky klasifikace hodnotím jako nadprůměrné, zejména modely pro jednotlivá kola grandslamových turnajů vykazují značný potenciál, kdy více pozorování pomohlo ke zpřesnění klasifikace. Jednoznačná dominance v parametrech výsledných modelů náleží proměnné `H2HWon`, kdy se tato proměnná nacházela téměř ve všech vytvořených modelech. Významnost vzájemně odehraných zápasů předčila i významnost kurzu daného sázkovými kancelářemi, což je zajisté překvapivé zjištění. Dále se v modelech často objevovaly proměnné `Ranking, Seeded, MPlayed`, v několika málo případech i proměnné `Vyska, Vek, Starts`. Nejlepších výsledků ale bylo dosaženo pomocí modelů, v kterých se vyskytovala proměnná `H2HWon`.

Pevně věřím, že tato práce, zejména její praktická část, je dostatečně srozumitelná i pro jedince, kteří nemají příliš velké matematické povědomí a že dokáže zaujmout i některé z řad sportovních fanoušků.

Literatura

[1] Reem Abulleil, *Is it finally time to rethink the number of mandatory tournaments in tennis?* [online]. 2018, [cit. 2019-02-23]. dostupné z: https://sport360.com/article/tennis/263784/is-it-finally-time-to-rethink-the-number-of-mandatory-tournaments-in-tennis.

[2] Cheryl Praeger, *Maths point—the mathematics of tennis* [online]. 2018, [cit. 2019-02-23]. dostupné z: https://www.science.org.au/curious/everything-else/tennis-maths.

[3] Scott Spits, *Heated reaction: Australian Open revamps extreme heat policy* [online]. 2018, [cit. 2019-02-23]. dostupné z: https://www.smh.com.au/sport/tennis/heated-reaction-australian-open-revamps-extreme-heat-policy-20181229-p50oqk.html.

[4] ITF Tennis – ABOUT [online]. [cit. 2019-02-19]. dostupné z: http://www.itftennis.com/about/grand-slam%C2%AE/overview.aspx

[5] Tandon Kamakshi, *Posing 10 ATP questions for 2009* [online]. 2008, [cit. 2019-02-26]. dostupné z: http://www.espn.com/sports/tennis/columns/story?columnist=kamakshi_tandon&id=3686016.

[6] *Official Site of Men's Professional Tennis* [cit. 2017-02-19]. dostupné z: https://www.atptour.com/en.

[7] Jeff Sackmann, *How Much Does Height Matter in Men's Tennis?* [online]. 2017, [cit. 2019-02-23]. dostupné z: http://www.tennisabstract.com/blog/2017/09/04/how-much-does-height-matter-in-mens-tennis/

[8] Sreemanta Sinha *Federer, Nadal and Djokovic are ridiculously good, but are the Big 3 better than every other era in tennis history?* [online]. 2018, [cit. 2019-02-23]. dostupné z: https://www.sportskeeda.com/tennis/how-dominant-are-big-3-federer-nadal-djokovic

[9] Hron K., Kunderová P., Základy počtu pravděpodobnosti a matematické statistiky. Univerzita Palackého. Přírodovědecká fakulta, 2013. ISBN: 978-80-244-3396-7

[10] Agresti, Alan. *Categorical Data Analysis*, Wiley, 2002. ISBN: 0-471-36093-7.

[11] Hosmer D. W., Lemeshow S. *Applied Logistic Regression, Second Edition*, Wiley, 2005. ISBN: 0-471-35632-8.

[12] Dokoupil, Petr: *Diplomová práce – Aplikovaná logistická regrese* [online]. 2012, dostupné z: https://theses.cz/id/a5zu7h/Diplomov_prce_-_Dokoupil_-Petr.pdf.

[13] Zvára, Karel. *Regrese, 1.vydání*, MatfyzPress, 2008. ISBN: 978-80-737-8041-8.

[14] Jeffrey S., Simonoff. *Analyzing Categorical Data*, Springer-Verlag New York Inc., 2003. ISBN: 0-387-00749-0.

[15] Robin, X.: *Package 'pROC'*, [online]. 2019, [cit. 2019-03-15]. dostupné z: https://cran.r-project.org/web/packages/pROC/pROC.pdf

Seznam tabulek