Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Studijní program: Kvantitativní metody v ekonomice

Studijní obor: Statistika

ANALYTICKÝ POHLED NA ZMĚNU HERNÍHO SYSTÉMU 1. FOTBALOVÉ LIGY

Diplomová práce

Autor diplomové práce: Bc. Jakub Černý

Vedoucí diplomové práce: doc. RNDr. Ivana Malá, CSc.

Akademický rok 2017/2018

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

V Praze dne 14. 5. 2018

…….
podpis

Poděkování

Na tomto místě bych rád poděkoval paní RNDr. Ivaně Malé CSc. Za odborné vedení diplomové práce a za vstřícnost, kterou mi při psaní práce prokázala.

Jen stěží bych mohl psát diplomovou práci, kdyby mě během celého studia nepodporovali moji rodiče. Zejména mamka tomuto cíli zřejmě obětovala víc než já.

Děkuju. Moc.

Je čím dál složitější se sejít a bude hůř. Davide, Patriku, Zdendo, díky i Vám. A věřím, že posledně zmíněný mi těch 12 milionů promine…

Abstrakt

Práce je zaměřená na tvorbu predikčních modelů, jejich evaluaci a využití pro simulaci nového herního formátu české první fotbalové ligy. Cílem práce je podrobit testu kontroverzní body nového herního systému první ligy. Nejdříve jsou představeny faktory ovlivňující výsledek fotbalového zápasu a jejich kvantifikace za použití Elo ratingu. Zkonstruováno je několik predikčních modelů, z nichž je za pomoci RPS, střední čtvercové chyby a střední absolutní odchylky jako nejlepší vybrán model logistické regrese. Pomocí něj jsou simulovány minulé ročníky první ligy jako by se odehrály v novém herním formátu, který sebou přináší několik kontroverzí. V práci je spočteno, že dojde ke snížení počtu týmů postupujících ze 2. do 1. ligy zhruba o jeden tým za 10 let. Za klíčové lze považovat zjištění, že v novém herním formátu bude docházet k zápasům, ve kterých bude pro tým výhodnější neusilovat o vítězství, ale o prohru, což může poškodit vnímání soutěže v očích veřejnosti.

Klíčová slova: logistická regrese, Elo rating, simulace

Abstract

This diploma thesis is based on predictive modelling, evaluation and the use of models for simulation of new format of Czech first football league. The goal of this paper is to test those parts of new format that face some criticism from fans and experts. It all starts with a description of factors influencing the outcome of the game and their quantification with Elo rating system. Several predictive models are constructed and evaluated using RPS, mean square error and mean absolute error. Logistic regression is chosen as the best model and therefore it is used for simulation of the first football league as if it was played by a new format. Simulation proved that number of teams promoted from second to first tier will on average decrease by 1 team in 10 years. Most importantly, it was proven that under the new format one can expect matches in which one team would more profit from a loss then a win. This can crucially affect public opinion about the competition.

Keywords: logistic regression, Elo rating, simulation

Obsah

Úvod

Český fotbal prochází v posledních letech viditelnou změnou. Investoři se nebojí do svých týmů alokovat větší finanční prostředky, a tak do týmů přicházejí i známí zahraniční hráči. Snahu o zlepšení českého fotbalu – nejen – jako marketingového produktu má podpořit i zavedení videorozhodčího, díky kterému by měl klesnout počet fatálních nepřesných rozhodcovských výroků. Zřejmě největší změnou nadcházející sezóny 2018/19 však bude změna herního systému nejvyšší fotbalové soutěže. Systém se změní poprvé po pětadvaceti letech fungování samostatné české kopané. Od tohoto kroku si sdružení profesionálních fotbalových klubů LFA slibuje zatraktivnění soutěže a s tím i více diváků a peněz.

Nové a inovativní herní formáty se začínají objevovat napříč Evropou. Soutěže, které nemohou konkurovat těm nejprestižnějším a nejsledovanějším ligám, ve kterých hrají týmy s podporou po celém světě, hledají způsob, jak zaujmout fanoušky alespoň ve své zemi. Ale nejen národní soutěže, herní systém změní od roku 2026 i akce ve fotbalovém světě nejsledovanější, FIFA Mistrovství světa.

Když se funkcionáři FIFA rozhodovali o změně herního systému, přihlíželi k analýze porovnávající několik zvažovaných systémů. Přestože se v českém fotbale o změně vedly místy i debaty bouřlivějšího charakteru, o jiném, než ryze pocitovém výběru český fanoušek informován nebyl. Cílem této diplomové práce je nabídnout analytický pohled na změnu hracího systému české nejvyšší fotbalové soutěže. Simulovat minulé ročníky, jako by se odehrály v novém formátu a podrobit testu kritizované body.

Jedná se o práci empirickou, zaměřenou na predikci výsledků fotbalových zápasů. Je zde představeno několik běžně používaných predikčních modelů z oblasti předpovídání sportovních výsledků. Nejlepší model je poté použit pro simulaci soutěže podle nového herního formátu. Veškeré výpočty a vizualizace byly získány pomocí softwaru R a MS Excel.

Úvodní kapitoly popisují starý a nový herní systém a data použitá k analýze. Ve 4. kapitole jsou představeny faktory ovlivňující výsledek zápasu a způsob jejich měření. Na jejich základě následuje navržení prediktorů pro predikční modely popsané v kapitole 6. Mezi nimi je logistická regrese, Poissonova regrese a vícenásobná lineární regrese v kombinaci se standardními statistickými rozděleními jako diskrétní Poissonovo a spojité normální. Z těchto modelů je vybrán ten, který dokáže nejlépe předpovídat výsledky v historii odehraných utkání.

S jeho pomocí je v 7. kapitole simulována sezóna 2016/17, jako by se hrála v novém herním systému. Do pětiletého kontextu jsou tyto výsledky zařazeny v následující kapitole, která shrnuje výsledky nejdůležitějších a nejvíce diskutovaných aspektů nového herní formátu. Poslední výpočetní 9. kapitola je zaměřena na identifikaci zápasů, ve kterých by se týmům v novém herním formátu více vyplatilo prohrát než vyhrát. Ukazuje se, že zápasů, ve kterých týmu záměrná prohra zvedne šanci na účast v evropských pohárech, které jsou spojené s prestiží a finančními odměnami, není vůbec málo.

1 Starý herní systém

Čtvrtstoletí se česká nejvyšší fotbalová soutěž hrála systémem každý s každým dvakrát, jednou na domácím stadionu, jednou na stadionu soupeře. Při účasti 16 týmů každý odehrál 30 utkání, celá sezóna tak nabídla celkem 240 zápasů.

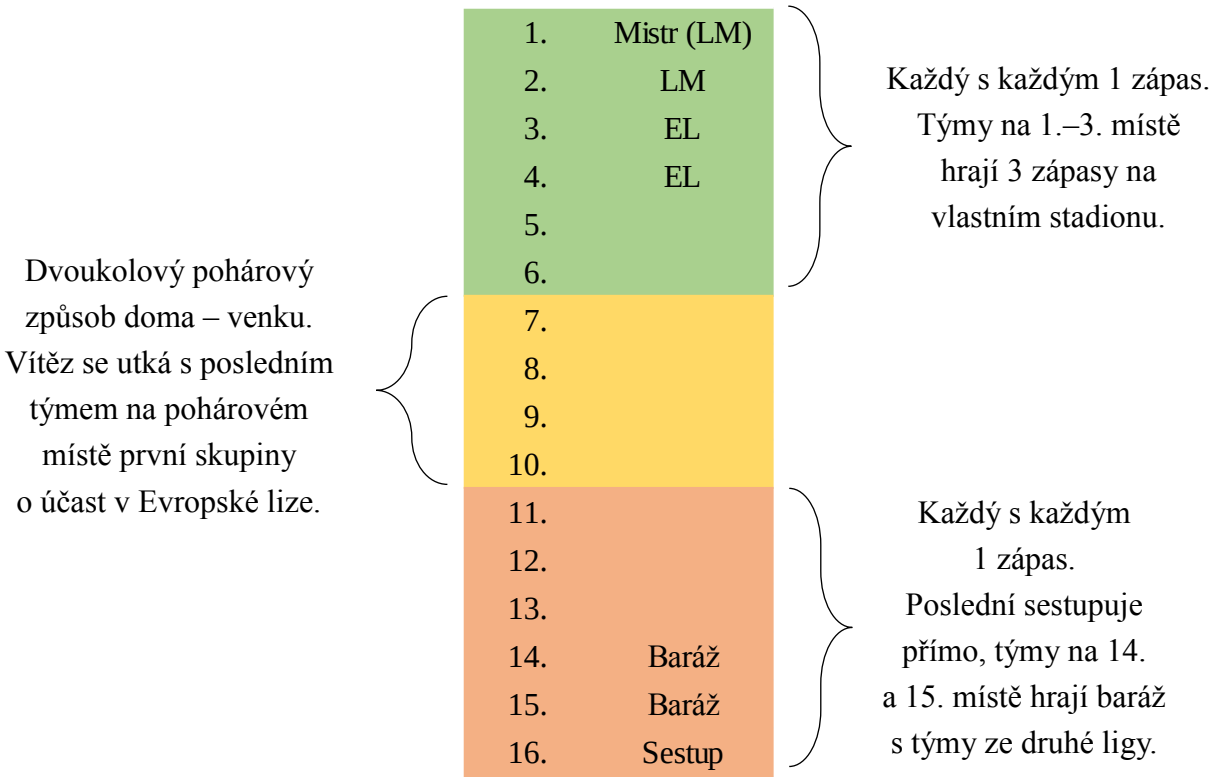
Za výhru si tým připsal 3 body (vyjma první samostatné sezóny 1993/94, kdy to byly body dva), za remízu 1 bod. Tým s nejvíce body získal ligový titul, dva týmy s nejméně body sestupovaly do nižší soutěže. Odtud stejným systémem dva týmy s nejvyšším bodovým ziskem postupovaly do první ligy, aby počet týmů zůstal zachován.

Jedná se o jednoduchý a funkční model, který využívají ty nejlepší soutěže v čele s anglickou Premier League. K té celý svět vzhlíží, minimálně pokud jde o příjmy z televizních práv. Aktuální kontrakt zaručuje roční příjem 1,7 miliardy liber, což je o 71 % více než v letech 2013–2016 (BBC News, 2015). Být majitelem klubu z Premier League může být finančně zajímavé, 17 z 20 klubů dosahovalo již před navýšením příjmu z televizních práv provozní zisk. Dosáhnout ho v českých podmínkách je obtížnější. V sezóně 2015/16 měly české týmy příjmy na úrovni 1,6 % příjmu týmů z nejvyšší anglické ligy (ČTK České noviny, 2017).

Je ovšem nutné dodat, že snažit se v tomto směru srovnávat českou a anglickou ligu, je jako by si Česká republika vzala za vzor HDP Spojených států a snažila se s ním měřit a dostihnout ho. Přesto argument, že ověřený herní model používají ty nejprestižnější ligy, zaznával a podporoval jeho zachování.

2 Nový herní systém

Formát, na který se od příští sezóny přechází, je vlastně rozšířením toho právě dosluhujícího. Po třiceti kolech, kdy týmy budou hrát každý s každým, doma – venku, přibyde takzvaná nadstavbová část (LFA, 2017). Tabulka se rozdělí na tři části.



Obr. 2.1 Rozdělení tabulky na 3 nadstavbové skupiny pro sezónu 2018/19

Týmům v první a třetí skupině na obr. 2.1, tedy ve skupinách o titul a o udržení, pro nadstavbovou část zůstanou všechny body získané v základní části. Mistr a vicemistr budou usilovat o účast v Lize mistrů (LM), třetí a čtvrtý tým čekají kvalifikace Evropské ligy (EL). Na vítěze ligového poháru rovněž čeká Evropská liga. Pokud se vítěz poháru kvalifikuje do evropských pohárů i z ligy, potom toto místo připadne týmu na pátém místě tabulky. Tým na posledním místě postupujícím do Evropy (4. nebo 5. místo, podle vítěze poháru) však o svou účast bude moct přijít na úkor vítěze druhé skupiny, který ho o účast může připravit v pohárovém dvojutkání.

Rozlosování nadstavbové části je fixní a dostupné v (LFA, 2017). Ve zkratce se vychází z toho, aby týmy s lepším umístěním po základní části měly v rámci nově nastaveného formátu

co největší možnou výhodu. Proto první tým po základní části hraje v nadstavbě tři domácí zápasy proti týmům na 2., 3. a 4. místě. O výhodě domácího prostředí bude v této práci pojednáno podrobněji. Obdobně je rozlosována i skupina o udržení. Ve druhé skupině hrané pohárovým způsobem se výhoda lepšího postavení po základní části promítá rovněž v nasazení domácích a hostujících celků. Lépe postavené týmy začnou dvojzápas vždy na hřišti hůře postaveného týmu. Vychází se z toho, že hrát druhé rozhodující utkání před vlastními fanoušky, je výhodné. Důkazem může být například statistika Ligy mistrů UEFA, ve které v historii z osmifinále postoupilo 72 % týmů, které hrály první utkání na hřišti soupeře (Esteva, 2015). I v LM platí pravidlo, že lépe postavený tým po skupinové fázi začíná venku, proto je číslo postupujících týmů hrajících první utkání na hřišti soupeře vysoké, jedná se zpravidla o kvalitnější týmy. Při pohledu na semifinálové zápasy, které se losují náhodně, již výhoda domácího hřiště ve druhém zápase mizí. Možný důvod nabízí (Chris, 2010). Argumentuje tím, že v odvetných zápasech padá více branek. Branka na hřišti soupeře má v pohárovém dvojutkání speciální význam, proto je podle něj výhodné hrát první zápas doma a druhý, ve kterém padá v průměru více branek, na hřišti soupeře.

Zkoumání, jak velkou výhodou je hrát první zápas venku, je v českých podmínkách v podstatě neuskutečnitelné. Dvoukolově se u nás hrály jen některé fáze národního poháru a použití jejich výsledků pro odvození této výhody je složité zejména z toho důvodu, že pro část týmů se jedná o podřadnou soutěž, ve které šetří své opory pro ligové zápasy. Třeba fotbalová Sparta v předminulém ročníku do odvety semifinále nasadila svoje dorostence a juniory (Šedivý, 2016).

2.1 Důvody pro změnu herního systému

Předseda LFA, která profesionální soutěže řídí, Dušan Svoboda v rozhovoru pro iSport řekl: „*Shodli jsme se už dřív, že profesionální soutěže potřebují nový impuls, rozšířit počet zápasů, přidat atraktivní utkání a odbourat ty, kde oběma klubům o nic nejde. Vždycky na každém návrhu může někdo spekulovat. Jsem přesvědčený, že jsme našli nejlepší možný systém z hlediska tradice, ekonomiky klubů i klimatických podmínek.“* (iSport, 2017)

2.2 Dopady

Díky nadstavbové části bude v sezóně odehráno o 42 zápasů více, což představuje nárůst o 17,5 %. Jen dva zápasy navíc přibydou týmům ze druhé skupiny, které hned první dvojzápas prohrají. Naopak tým z první skupiny, který bude na posledním pohárovém místě, sehraje 7 zápasů navíc, tedy o téměř čtvrtinu zápasů na sezónu více než doposud.

Část zápasů, které přibydou, bude navíc velmi zajímavých. Pozornost budou přitahovat nejen zápasy rivalů z čela soutěže, ale atraktivní bude i vyřazovací systém druhé skupiny. Přínos skupiny o udržení bude zejména v barážových zápasech, které mají rovněž předpoklad být pozorně sledovány.

Předseda LFA Dušan Svoboda garantoval týmům i vyšší příjmy. Prvoligové týmy dostanou o 2 miliony více, druhá liga si jako celek polepší o 10 milionů (ČTK České noviny, 2017). Profitovat budou i fanoušci, kteří se dočkají více zápasů v létě, jelikož se zkrátí letní přestávka. Jít na fotbal v tričku a kraťasech je nepochybně příjemnější, než sedět na stadionu v zimní bundě.

2.3 Kritika

Změnu musely odsouhlasit sami kluby. Pro návrh zvedli ruku zástupci 30 z nich, Opava a Znojmo byly proti (iDNES.cz a ČTK, 2017). Koncept jako takový oběma těmto týmům sympatický byl, ale jako zástupci druholigových týmů měly obavy o prostupnosti mezi 1. a 2. ligou. Doteď měly druholigové týmy garantována dvě postupová místa do první ligy. S novým formátem už jen jedno s možností dvou dalších, pokud si je týmy vybojují v baráži.

Z prvoligových týmů měla svoje připomínky Slavia, která ke hlasování vydala tiskové prohlášení. „*Už od samého počátku diskuzí o změnách herního systému první a druhé ligy vyjadřovala SK Slavia Praha své vážné pochybnosti o nutnosti, správnosti a smysluplnosti tohoto kroku.“* (SK Slavia Praha, 2017) Jaké body se Slavii zdají v novém formátu sporné, ve zprávě neuvádí. Konstatuje však, že vzhledem ke shodě mezi ostatními kluby první ligy se k nim i přes své pochybnosti při hlasování přidala a návrh podpořila.

Mezi fanoušky vzbuzuje největší emoce zřejmě druhá nadstavbová skupina, která dává i desátému týmu šanci na účast v Evropské lize na úkor týmu na čtvrté pozici. Spekuluje se, že

před koncem základní části budou některé kluby mít zájem raději se umístit na sedmé pozici a usilovat o účast v EL skrze druhou skupinu, než zůstat na šestém místě. Tým na šestém místě může mít při vyšší bodové ztrátě na pohárovou příčku, jen malou šanci se na ni dostat a potom ji musí ještě obhájit právě proti týmu, který vyhraje druhou nadstavbovou skupinu.

Prostupnost mezi 1. a 2. ligou a záměrné vypouštění zápasů pro zvýšení šancí na získání místa v EL budou mít v následujícím textu práce největší prostor. Jedná se totiž o prvky nového formátu, které mohou mít přímý sportovní, ale i finanční dopad nejen na samotné kluby, ale i na ligu jako celek. V zájmu českého fotbalu by mělo být, aby vysílal do evropských pohárů ty nejsilnější kluby, které pak budou mít reálnou šanci se v nich prosadit. Jsou to často právě peníze z evropských soutěží, ke kterým se majitelé klubů upínají, aby mohli zhodnotit svoji investici do fotbalového klubu.

3 Data a postup modelování

V posledních letech zažívá sportovní statistika obrovský boom. Množství dat sbíraných ze zápasů po celém světě umožnilo konstrukci modelů, o kterých bylo dříve obtížné jen uvažovat. Aktuálním trendem je očekávaný počet branek v zápase (*xG statistics*). Ve zkratce jde o výstup regresního modelu, který odhaduje procentuální šanci na vstřelení gólu z každé vytvořené brankové příležitosti. Osobou, která má velký podíl na popularitě tohoto ukazatele, je holandský bloger 11tegen11, který se na Twitteru blíží k 60 000 uživatelů, kteří jeho výpočty sledují (11tegen11, 2017).

Data potřebná pro výpočet *xG* a jiných pokročilých odhadů, nejen že nejsou veřejně dostupná, ale jejich pořízení je i poměrně nákladné. Pro zpracování této práce připadají v úvahu veřejně přístupná data. Česká republika navíc nikdy nebyla v oblasti sportovních statistik průkopníkem, je to spíše oblast objevená v posledních pěti letech. V současné době už lze na oficiálních stránkách nejvyšší soutěže nalézt zajímavé zápasové i hráčské statistiky, ze kterých se s trochou úsilí dají vytvořit úspěšné predikční modely. Druhá liga je na tom podstatně hůře, přičemž i data z této soutěže jsou pro ověření nového modelu nejvyšší soutěže vzhledem k barážovým zápasům zapotřebí.

3.1 Zdroje

Výpočty v této práci se budou opírat o to nejdůležitější, co každý fotbalový zápas přináší – o jeho výsledek. Výhodou tohoto skromného přístupu je možnost pracovat se stejně rozsáhlými a kvalitními daty pro první i druhou ligu v celém období od vzniku samostatné České republiky.

Prvoligová data pocházejí z oficiálních stránek nejvyšší soutěže (HET Liga, 2017) a zahrnují datum a výsledek zápasu. Pro aktuálnější sezóny i čas zahájení utkání, který může být pro některé výpočty rovněž důležitý. Oficiální stránky druhé ligy bohužel sahají jen několik málo sezón nazpět, takže data pro druhou nejvyšší soutěž pocházejí z projektu CS Fotbal (CS Fotbal, 2017).

Projekt CS Fotbal využívá jako jeden ze zdrojů dat oficiální stránky nejvyšší fotbalové soutěže. Data z těchto dvou zdrojů jsou proto kompatibilní a snadno propojitelná. Jediným důvodem, proč nebyla použita i pro nejvyšší soutěž data z CS Fotbal je ten, že statistiky z oficiálních stránek sbírám již delší čas a zápasy mám ve své soukromé databázi. Pro účely této práce tak stačilo získat navíc jen výsledky zápasů druhé ligy.

3.2 Rozdělení na trénovací a testovací data

Pro první ligu jsou k dispozici výsledky 5 760 zápasů (24 sezón po 240 zápasech). Pro druhou ligu 5 736 zápasů. V sezóně 1994/95 se soutěže účastnilo 18 týmů, tím byl počet zápasů 306. V sezónách 1997/98 a 2015/16 bylo účastníků 15, v sezóně 2004/05 byl jeden klub (FC Bohemians Praha) vyloučen a jeho výsledky anulovány. Počet zápasů v těchto sezónách byl 210.

Trénovací data						
1993/94	1994/95	1995/96	1996/97	1997/98	1998/99	1999/00
2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07

Validační data				
2007/08	2008/09	2009/10	2010/11	2011/12

Testovací data				
2012/13	2013/14	2014/15	2015/16	2016/17

Obr. 3.1 Schéma rozdělení dat podle sezón

Data byla rozdělena přibližně v poměru 60:20:20. Poslední testovací část bude v průběhu konstrukce modelů stát stranou a podle přesnosti predikce na těchto datech bude vybrán konečný model, který poslouží k simulaci nového formátu. Jediným kritériem pro rozdělení dat byla časová posloupnost. Takové rozdělení zaručí, že bude vybrán predikční model, který nejlépe předpoví nejaktuálnější zápasy, což je u sportovních predikcí vždy žádoucí.

3.3 Tvorba predikčních modelů

K naplnění cílů práce je nutné simulovat v historii neodehrané zápasy nadstavbové části, ale i určit pravděpodobnosti ostatních výsledků odehraných zápasů za účelem zodpovězení otázek „Co by se stalo, kdyby…“, např. tým záměrně prohrál. Druhá část úlohy nevyžaduje pouhou předpověď nejpravděpodobnějšího výsledku (klasifikační model), ale odhadnutí pravděpodobnostního rozdělení všech možných výsledků. Rozsáhlý přehled využitelných modelů nabízí např. (Berry a Linoff, 2004).

Bylo by možné představit jeden vybraný model a na jeho základě spočítat požadované výsledky. Takový přístup by nicméně mohl nahrávat úvahám o jeho účelovém vybrání, a tím i ovlivnění výsledků. Namísto jednoho modelu tato kapitola pokryje několik jednoduchých (vzhledem k absenci pokročilých dat to jinak ani nejde) verzí základních modelů, které se pro předpovídání sportovních výsledků používají. Na základě schopnosti výsledky správně predikovat, bude nejlepší model použit pro simulaci nového herního formátu českých profesionálních soutěží.

3.4 Podmínky vyhodnocení modelů

Pokud jde o vyhodnocení kvality předpovědí, pak je často používanou mírou střední čtvercová chyba (*MSE – Mean Squared Error*). (Constantinou a Fenton, 2012) vyzdvihují při hodnocení předpovědí sportovních utkání *RPS – Rank Probability Score*. Tony Corke, bloger provozující populární web MatterOfStats zaměřený na australskou fotbalovou ligu, preferuje střední absolutní odchylku (*MAE – Mean Absolute Error*) a své přesvědčení opírá o jeden ze svých článků (Corke, 2017).

Zajímavé bude spočítat všechna kritéria, a pokud bude jeden model označen jako nejlepší větším počtem těchto vyhodnocovacích kritérií, tím lépe. Pokud ke shodě nedojde, bude vybrán model, který dosáhne nejlepšího průměrného pořadí napříč hodnotícími kritérii.

Každý zápas (i) může skončit třemi různými výsledky – výhrou domácího týmu, remízou nebo výhrou hostujícího týmu – označenými jako $k \in \{1\,(D), 2\,(R), 3\,(H)\}$. Přestože vítěz zápasu je nominální proměnná, v praxi se s ním často zachází jako s proměnnou ordinální, kdy lze výsledek uspořádat. To je důvod, proč jednotlivé výsledky mají přiřazené jak pořadové číslo, tak písmennou zkratku, aby byla zachována čitelnost a názornost výsledků. Ordinální uspořádání bude brzy využito při výpočtu RPS. Hodnota i plní funkci jednoznačného identifikátoru zápasu. Při znalosti dosavadního herního systému (viz kap. 1), lze konkrétní zápas určit participujícími týmy v pořadí domácí – hosté a sezónou, ve které se odehrál. Sezónu lze případně nahradit přesným datem zápasu. Například i = Hradec – Liberec, 2016/17 nebo i = Hradec – Liberec, 22. 04. 2017. Tento zápas 25. kola první ligy skončil vítězstvím hostujícího Liberce 1:0. V novém herním modelu, ve kterém přibydou nadstavbové zápasy, bude muset být využíváno značení s přesným datem, neboť se v jedné sezóně některé týmy utkají vícekrát na jednom stadionu.

Pro potřeby snazších výpočtů byl každému zápasu přiřazen i jednoznačný číselný identifikátor, který odpovídá pořadí, v jakém se zápasy odehrály. Pro ukázkový zápas lze napsat $i = 11\,397$.

Skutečně pozorovaný výsledek zápasu bude označen ${}_i s_k \in \{0,1\}$ a bude nabývat hodnoty jedna, pokud *i*-tý zápas skončil *k*-tým možným výsledkem. Odhad pravděpodobnosti k-tého možného výsledku pro i-tý zápas bude ${}_i p_k \in [0,1]$.

Pro zmíněný zápas Hradce s Libercem je tím pádem skutečný výsledek

$${}_{11397}s_D = 0, \qquad {}_{11397}s_R = 0, \qquad {}_{11397}s_H = 1$$

a odhad pravděpodobnosti výsledku podle benchmarkového modelu, který je popsán v kapitole 4.2.7

$${}_{11397}p_D = 0{,}298, \qquad {}_{11397}p_R = 0{,}329, \qquad {}_{11397}p_H = 0{,}373.$$

Se zavedeným značením lze psát

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{k \in \{D,R,H\}} \left({}_i s_k - {}_i p_k \right)^2 , \tag{3.1}$$

nebo podle alternativního označení výsledku zápasu

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{k=1}^{3} \left({}_i s_k - {}_i p_k \right)^2 , \tag{3.2}$$

kde *n* je počet zápasů.

Výpočet průměrného *RPS* je pro fotbalové zápasy dán vzorcem

$$RPS = \frac{1}{n} \sum_{i}^{n} \frac{1}{2} \sum_{k=1}^{3} \left(\sum_{j=1}^{k} {}_i s_{k,j} - \sum_{j=1}^{k} {}_i p_{k,j} \right)^2 , \tag{3.3}$$

kde *j* určuje v případě *s* pozici pro výpočet kumulativního výsledku v ordinálně seřazených výsledcích a v případě *p* určuje odhad kumulativní pravděpodobnosti na stejné pozici. Je-li $j = 2$, potom *s* určuje, zda domácí tým vyhrál nebo alespoň remizoval a *p* přiřazuje tomuto jevu odhad pravděpodobnosti. Střední absolutní odchylka se podle zde používaného zápisu spočte jako

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{k \in \{D,R,H\}} \left| {}_i s_k - {}_i p_k \right|, \tag{3.4}$$

nebo podle alternativního označení výsledku zápasu,

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{k=1}^{3} \left| {}_i s_k - {}_i p_k \right|. \tag{3.5}$$

4 Faktory ovlivňující výsledek zápasu

Problematikou předpovídání výsledků sportovních událostí se zabývají například (Stekler a kol., 2009). Ve své práci rozebírají různé druhy predikcí podle toho, kdo predikuje, co predikuje atd. Pro účel této práce je zapotřebí předpovědět kompletní pravděpodobností rozdělení výsledku. Prvním krokem bude identifikace vlivů, které výsledek zápasu determinují.

Při odhlédnutí od různých korupčních skandálů, kterými si prošly snad všechny světové soutěže, zde budou uvažovány tři základní skupiny faktorů ovlivňující výsledek fotbalového zápasu:

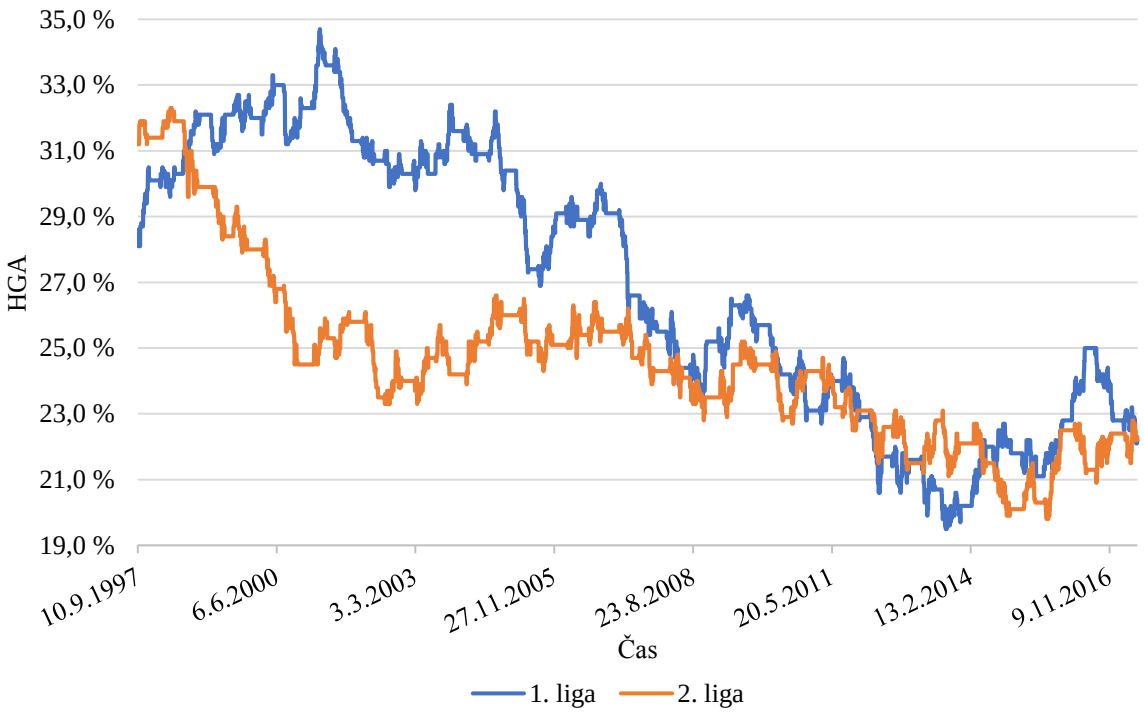
1) kvalita týmů,
2) výhoda domácího prostředí,
3) ostatní.

Obdobně definují faktory např. i (Dixon a Coles, 1997). Je nesporné, že čím kvalitnější a sehranější mužstvo je, tím jeho šance na úspěch rostou. Do pojmu „*kvalita týmu*“ zde bude skryto vše, co přímo souvisí s hráči. Vzhledem k tomu, že data nenabízejí možnost pracovat s hráči jako jednotlivci, není možné usuzovat o šancích týmu v zápase například na základě absence důležitých hráčů. Kvalitnější tým má i kvalitnější náhradníky, proto faktor kvality týmu implicitně zahrnuje i schopnost absentující hráče nahradit.

V knize *Statistical Thinking in Sports* je výhodě domácího prostředí (HGA – Home Ground Advantage) věnována celá kapitola (Albert a Koning, 2007). Autoři identifikují tři hlavní pilíře, na kterých výhoda domácího prostředí stojí:

1) únava a psychologický efekt cestování hostujícího týmu,
2) psychologická výhoda domácího týmu plynoucí z podpory diváků,
3) taktická výhoda známého prostředí.

Jednoduchým porovnáním počtu vítězství domácích a hostujících týmů vyměřili výhodu domácího prostředí ve fotbale na 21,7 procentního bodu (p. b.). Větší výhodu mají jen hráči ragby. Nejen, že se HGA liší sport od sportu, ale i v rámci jednotlivých zemí. Rozdíly mohou být i napříč různými soutěžemi v jedné zemi.



Obr. 4.1 Výhoda domácího prostředí v českých soutěžích

Na obr. 4.1 je vyčíslená výhoda domácího prostředí tak, jak je popsána ve zmíněné knize pomocí jednoduchého klouzavého průměru (*SMA – Simple Moving Average*) o délce 1 000 zápasů uspořádaných podle data. *SMA* se v kontextu této práce dá pro *i*-tý zápas $(i \geq 1\,000)$ zapsat jako

$$ {}_{i}SMA = \frac{\sum_{i-999}^{i} \left({}_{i}s_{D} - {}_{i}s_{H} \right)}{1\,000} \cdot 100\% . \tag{4.1}$$

Největší rozdíl mezi první a druhou ligou byl v dubnu 2001, kdy v první lize vyhrály domácí týmy v průměru za předchozích 1 000 zápasů 53,3 % zápasů a hosté pouze 18,6 %. Ve druhé lize to bylo 49,3 % vítězných zápasů pro domácí a 24,3 % pro hosty. Rozdíl v prví lize tak byl 34,7 p. b. Pro druhou ligu byl tento rozdíl 25 p. b. Od sezóny 2007/08 je vidět, že HGA má pro první i druhou ligu velmi podobný trend a v posledních zápasech, které jsou v datasetu k dispozici, je *SMA* totožný pro první i druhou ligu 22,2 %. To je pouze o půl procentního bodu více, než ve své práci zjistili Albert a Koning.

Je jasné, že minimálně 2 ze 3 výše definovaných pilířů výhody domácího prostředí nejsou totožné pro všechny stadiony. Únava hostujícího týmu z cestování by měla růst s tím, jak

vzdálený je stadion, na který cestují. Výhoda podpory diváků bude zase tím větší, čím hojněji jsou zápasy domácími diváky navštěvovány. I to jsou faktory, které mohou hrát roli v rozdílu mezi soutěžemi, který je vidět na posledním obrázku. Hlubší pohled nabízí např. (Pohlkamp, 2014), který zkoumá třeba i to, nakolik je ovlivněno rozhodování rozhodčího na stadionu s a bez atletické dráhy kolem hřiště. Ukazuje se, že atletická dráha zřejmě tlumí tlak, který na rozhodčího vytvářejí plné ochozy. Je proto výhodnější hrát domácí zápasy na stadionu bez atletické dráhy.

U faktorů, které ovlivňují výsledek utkání a byly označeny jako *ostatní*, bude nadále uvažována nulová střední hodnota a tyto faktory nebudou modelovány. Zavádí se tím předpoklad, že neovlivňují šanci týmu na vítězství. Pro představu lze uvést například vliv počasí, u kterého lze vyslovit domněnku, že ovlivní například celkový počet branek v zápase, ale nikoli šance jednoho z týmů v zápase zvítězit. Pokud by to tak bylo, potom by jeden tým zvládal například deštivé počasí lépe než druhý, to už pokrývá faktor kvality týmu.

4.1 Možnosti hodnocení kvality týmu

S výše nastavenými předpoklady vyvstává otázka, jak kvalitu jednotlivých týmů kvantifikovat, když data, která jsou pro zpracování této práce k dispozici, nabízí pouze výsledky zápasů.

Inspiraci je možné najít ve světě šachu. Žebříček nejlepších šachistů světa se sestavuje na základě kvality hráčů měřené v tzv. Elo bodech (dále bude na tuto hodnotu odkazováno jako na „Elo“). Tento koncept vymyslel maďarský fyzik Arpad Elo a dnes tento *rating system*, jak se metodám na hodnocení kvalitativní úrovně účastníků hry říká, v nějaké podobě využívají například i populární online hry. Více o autorovy i Elo hodnocení nabízí článek na webu ChessBase od Daniela Rosse z univerzity v Indianě (Ross, 2007).

S alternativou k Elo hodnocení přišel Mark Glickman, člen Americké Statistické Společnosti přednášející statistiku na Harvardu. Základním rozdílem oproti Elo je zavedení ukazatele, který měří spolehlivost ratingu (kvality) hráče (Glickman, 2017).

Porovnání obou systémů přinesl člen Světové šachové federace Michalis Kaloumenos (Kaloumenos, 2012). Pro hodnocení kvality fotbalových týmů v českých soutěžích bude použit Elo rating system. Bylo dokázáno, že samotné Elo lze velmi dobře použít pro predikci výsledku

bez aplikace dalších pokročilých metod. V porovnání s ostatními možnostmi hodnocení kvality fotbalových týmů vychází rovněž nejlépe (Lasek, 2012).

4.2 Konstrukce Elo

Hodnocení kvality týmu měřené Elo ratingem je číselný ukazatel, který je tím vyšší, čím lepších výsledků tým dosahuje. Výsledky týmu jsou v tomto kontextu považovány za indikátor kvality, neboť lepší tým dosahuje zpravidla lepších výsledků.

Elo týmu se aktualizuje po každém zápase, který tým odehraje. Roste, když tým uhraje lepší než očekávaný výsledek, v opačném případě klesá. Elo vítězného týmu vzroste o stejný počet bodů, o který klesne Elo týmu, který prohraje. Průměr Elo napříč týmy tak zůstává konstantní. Při použití značení (Kirill, 2017), vítěze srovnání kvality předpovědí podle (Lasek, 2012), se Elo rating spočte jako

$$R_n = R_o + K \cdot GD \cdot (W - W_e), \tag{4.2}$$

kde R_n je nový pozápasový Elo rating a R_o ten předzápasový. K je konstanta určující, jak rychle Elo rating reaguje na nejaktuálnější výsledky. S menší hodnotou K bude Elo rating ukazovat dlouhodobou kvalitu týmu, zatímco s vysokou hodnotou spíše aktuální formu. GD je koeficient zohledňující rozdíl ve výsledku. Je tím větší, čím větším počtem branek tým zvítězí. W je výsledek zápasu a nabývá hodnoty 1, když tým v zápase zvítězí, 0,5 v případě remízy a 0, pokud tým zápas prohraje. W_e je v konceptu Elo ratingu označení pro očekávaný výsledek vypočtený vzorcem

$$W_e = \frac{1}{10^{-(dr+HGA)/400} + 1}, \tag{4.3}$$

kde dr je rozdíl v Elo ratingu týmů a HGA výhoda domácího prostředí. Hodnota HGA je kladná pro domácí tým (zvyšuje šanci na výhru) a záporná pro hostující celek (snižuje šanci na výhru).

Dále se v textu rozlišuje HGA jako zkratka pro výhodu domácího prostředí a HGA jako proměnná tuto výhodu kvantifikující.

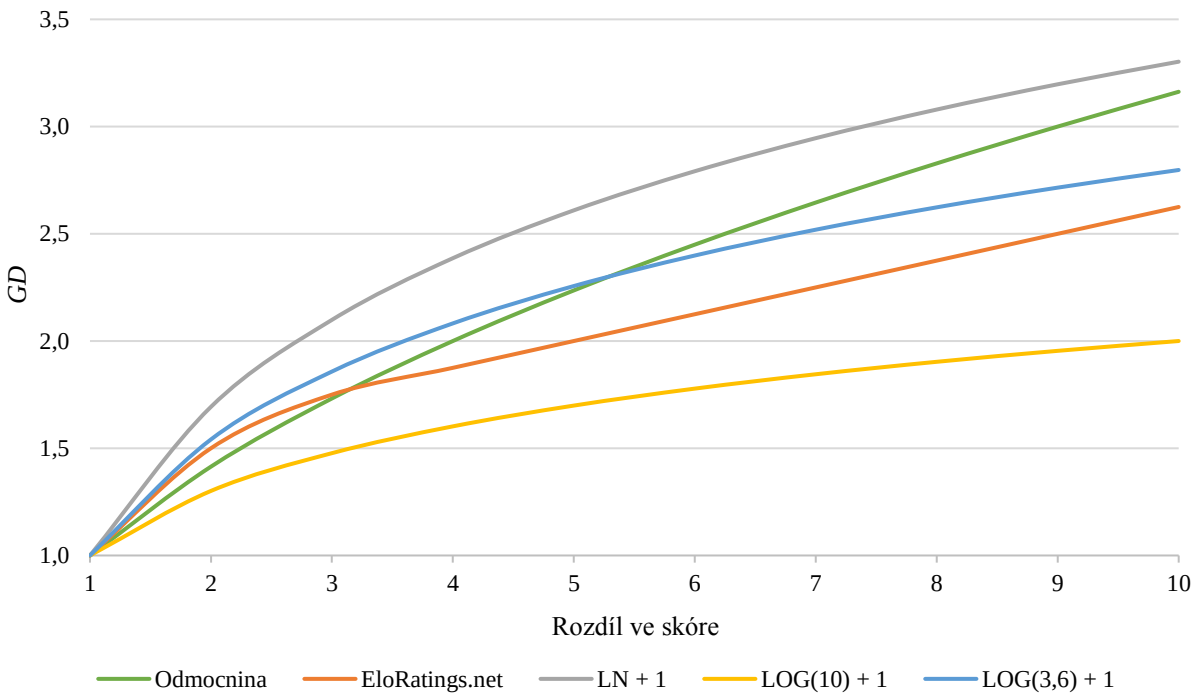
S takto zavedeným výpočtem vyvstává několik otázek, které je zapotřebí vyřešit a není na ně univerzální odpověď. Je potřeba zvolit hodnoty K, GD, HGA a tzv. *starting values*, neboli

počáteční hodnoty Elo, se kterými týmy vstoupí do prvního zápasu. Dalším potenciálním problémem je fakt, že 1. a 2. liga nejsou uzavřené. Každou sezónu do druhé ligy postupují nejúspěšnější týmy z nižší soutěže a opačným směrem jdou nejhorší týmy druhé ligy. Je nutné najít způsob, který přiřadí novým týmům ve druhé lize takové Elo hodnocení, které nenaruší celý systém podceněním nebo přeceněním jejich kvality.

4.2.1 Volba *GD*

Důvod, proč v modelu zohledňovat rozdíl ve skóre, je zřejmý. Tým, který porazí svého soupeře 4:0, pravděpodobně podal v zápase lepší výkon než tým, který stejného soupeře porazí pouze 1:0. Nemusí tomu tak být samozřejmě pokaždé, ale v dlouhodobém horizontu je to oprávněný předpoklad. V praxi uplatňovaný koncept má i druhou rovinu, a sice že každý další gól rozdílu skóre je pro určení kvality týmu od určité hranice stále méně důležitý. Rozlišit výhru 2:1 a 3:1 je v tomto kontextu důležitější než rozlišit vítězství 7:1 a 8:1.

Některé možnosti pro výpočet *GD* z rozdílu ve skóre jsou na obr. 4.2. Je zde třeba vidět, že funkce používaná (Kirill, 2017) nesplňuje požadavek na stále klesající význam dalšího gólu, jelikož *GD* roste od rozdílu tří branek lineárně. Odmocnina, která je v praxi rovněž hojně používána, má první diferenci sice klesající, ale pokles je velmi pozvolný. Použití přirozeného a dekadického logaritmu je také problematické, první zasahuje do výpočtu Elo velmi silně, druhý mnohem méně.



Obr. 4.2 Funkce použitelné pro výpočet *GD*

Při výpočtu *GD* bude použit logaritmus o základu 3,6. Je vidět, že se jedná o jakýsi kompromis mezi nabídnutými možnostmi. Volba je také podložena mojí dobrou osobní zkušeností s touto funkcí. Výpočet se dá s využitím logiky zavedeného značení zapsat jako

$$GD = \log_{3,6}\left|goly_D - goly_H\right| + 1. \tag{4.4}$$

4.2.2 Volba *HGA*

Očekávaná změna Elo ratingu je pro každý tým v každém zápase rovna nule. Pokud rating ukazuje skutečnou kvalitu týmu, potom se v dlouhodobém horizontu očekávané výsledky W_e rovnají skutečným výsledkům W a v rovnici (4.2) zbyde pouze $R_n = R_o$. Z toho plyne, že se rating týmu nezmění. Důležité je zdůraznit, že W_e je forma vyjádření očekávaného výsledku pouze v rámci výpočtu Elo hodnocení a do modelů v 6. kapitole může vstupovat jako prediktor (viz tab. 5.1).

Kdyby nebyla v rovnici (4.3) zohledněna výhoda domácího prostředí, potom by W_e pro domácí týmy bylo podhodnocené. Z obr. 4.1 je patrné, že domácí týmy vyhrávají více zápasů než hosté a v zápase dvou vyrovnaných týmů tak má větší šanci na vítězství ten domácí. Při výpočtu Elo

ratingu by to znamenalo vysoké přesuny bodů na stranu domácích týmů, které by v dlouhodobém horizontu nebyly nulové, ale kladné a vyčíslovaly by výhodu domácího prostředí. Hodnota *HGA* z rovnice (4.3) by tak měla na kumulaci bodů domácími týmy reagovat a zajišťovat dlouhodobou nulovou střední hodnotu.

Z popsaného problému je zřejmé, že volba (Kirill, 2017) není zrovna šťastná, jelikož používá konstantní *HGA* = 100 při průměrném Elo 1 500. Na obr. 4.1 je vidět, že výhoda domácího prostředí se v čase mění a může být různá pro 1. i 2. ligu, proto bude aktualizována po každém zápase a pro každou soutěž zvlášť. Pro *i*-tý zápas v *L*-té (1. nebo 2.) lize lze *HGA* zapsat jako

$$ {}^{L}_{i}HGA = {}^{L}_{i-2}HGA + \Delta Elo_{i-1} \cdot c \,, \tag{4.5} $$

kde ΔElo je druhý sčítanec vzorce (4.2) z pohledu domácího týmu a c je konstanta, která určuje, jak rychle se bude *HGA* přizpůsobovat aktuálním výsledkům.

4.2.3 Konstanta *K*

Na stránkách, ze kterých je přebrána velká část značení výpočtu Elo používají několik hodnot *K* podle důležitosti zápasu. Vzhledem k tomu, že se jedná o mezinárodní fotbal, je to volba vcelku logická. Mistrovství světa je pro týmy důležitější než přátelské utkání, mělo by proto mít v hodnocení větší váhu. V ligovém fotbale se každý zápas hraje o 3 body a dělat mezi nimi rozdíly sice možné je, ale výpočet Elo by se zkomplikoval a výrazné zlepšení modelu se od tohoto kroku očekávat nedá. Po zavedení nového formátu budou nadstavbové zápasy pro úspěch týmu určitě důležitější než ty v základní fázi a ve výši *K* by se to mohlo zohlednit. Pro zjednodušení modelu bude uvažována hodnota *K* jako konstantní celou sezónu pro všechny zápasy. Oblíbená hodnota *K* je u blogerů, zabývajících se fotbalovým Elo hodnocením, kolem 20, pokud volí pro počáteční hodnoty průměrné Elo 1 500.

4.2.4 Počáteční hodnoty

Několikrát již byla v textu zmíněna průměrná hodnota Elo 1 500. Znamená to, že průměrně kvalitní tým má Elo rating 1 500 bodů. Z této nejčastěji používané hodnoty budou vycházet i výpočty v této práci. Je zapotřebí přiřadit počáteční hodnoty Elo ratingu pro 32 týmů v sezóně 1993/94, která je v datasetu první. Týmy první ligy by měly mít vyšší Elo než týmy druholigové, jelikož první ligu hrají lepší týmy. I mezi týmy uvnitř soutěží je kvalitativní rozdíl

a bylo by výhodné se hned na úvod trefit co nejpřesněji do jejich ohodnocení Elo body. Nastavit je nutné i hodnotu *HGA* pro první zápas v obou soutěžích.

Jelikož se ve volném čase předpovídání sportovních výsledků intenzivně věnuji a Elo rating je jedním ze základních stavebních kamenů mé práce, přišel jsem s postupem, který z dat odhadne přibližné počáteční hodnoty. Mít standardizovaný postup je vhodné, jelikož počáteční hodnoty se mění pokaždé, když se podaří získat nová data ze staršího období. Postup je popsán v podkapitole 4.2.6.

4.2.5 Elo postupujících týmů

Přesuny mezi první a druhou ligou jsou bezproblémové, týmy si svoje hodnocení přenášejí do nové soutěže. Problém působí týmy, které postoupí do 2. ligy z nižší soutěže. Těm bude vždy přiřazeno průměrné Elo týmů, které se ve skončené sezóně umístily na posledních 3 místech tabulky. Lze namítnout, že v historii zpravidla sestupovaly 2 týmy a logická volba by bylo průměrné Elo dvou posledních týmů. Ne nutně ale sestupují týmy s nejnižším Elo ratingem. Hodnota přiřazená druholigovým nováčkům vypovídá o kvalitě týmů, které se pohybují u dna druholigové tabulky a v dlouhodobém horizontu je totožná s Elo ratingem sestupujících týmů.

4.2.6 Algoritmus pro výpočet počátečních hodnot

Při výpočtu počátečních hodnot příliš nezáleží – alespoň pokud nejsou zvoleny naprosto nevhodně – na konstantách *K* a *c*. Výpočet se provádí ve dvou krocích

1) zjištění průměrného Elo v soutěži,

2) určení Elo jednotlivých týmů.

Pro bod 1) je důležitých úvodních 6 sezón, které jsou k dispozici, tedy sezóny 1993/94 až 1998/99. Další sezóny se k výpočtu nevyužívají zejména proto, že kvalita soutěže (měřená průměrným Elo ratingem týmů v nich působících) se může v čase měnit. Prioritou je být co nejblíže kvalitě soutěže v úvodní sezóně, pro kterou se počáteční hodnoty určují.

Když je odhadnuto průměrné Elo soutěže, pro odhadnutí kvality týmů se použije 55 % zápasů první sezóny. To jsou přibližně zápasy před zimní přestávkou a přestupním obdobím, které dokáže skrze nákupy a prodeje hráčů kvalitu týmů výrazně ovlivnit.

Zjištění průměrného Elo v soutěži

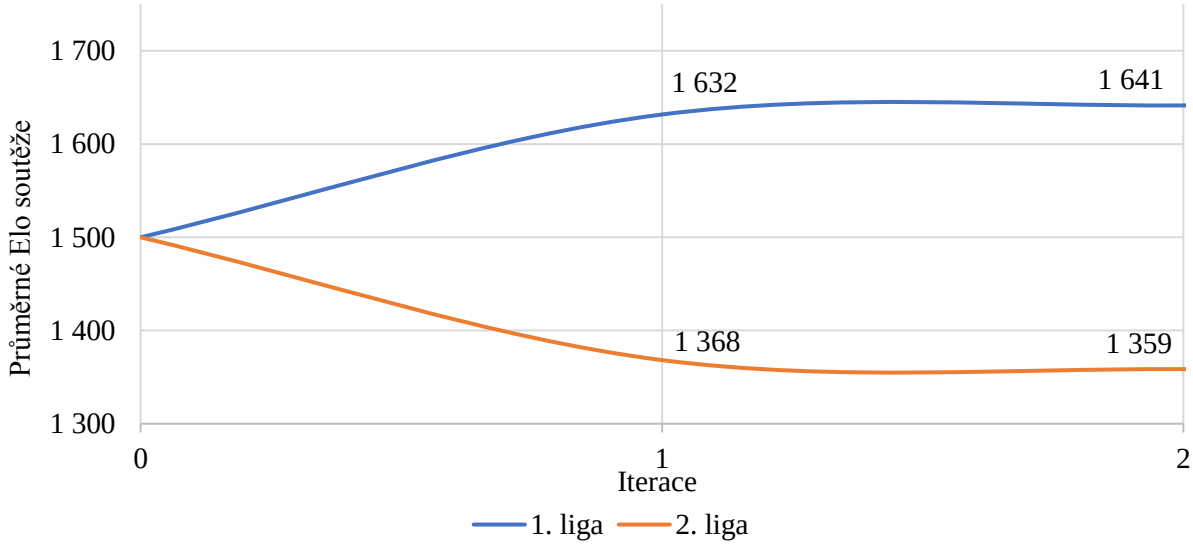
A. Určení výpočetních konstant, úvodní hodnoty *HGA* a průměrného Elo napříč soutěžemi.

   $K = 20$; $c = 0{,}075$; $HGA = 75$; $Elo = 1\,500$.

B. Počítej Elo pomocí vzorců (4.2) až (4.5) pro prvních 6 sezón.

C. Zjisti *HGA* a průměrné Elo uvnitř soutěží v poslední spočtené sezóně (1998/99).

D. Použij zjištěné Elo a *HGA* jako počáteční hodnoty pro týmy uvnitř této soutěže v první sezóně (1993/94) a opakuj body B. až D., dokud je Manhattanská vzdálenost (Pecáková, 2011) počátečních hodnot týmů dvou po sobě jdoucích iterací větší než 20.



Obr. 4.3 Hledání kvality průměrného týmu v soutěži

Jak je vidět na obr. 4.3, rozdíl v kvalitě první a druhé ligy se dal v sezóně 1993/94 odhadovat přibližně na 282 Elo bodů. Výhoda domácího prostředí *HGA* popsaným výpočtem vychází na 157 bodů pro 1. ligu a 118 bodů pro 2. ligu. Získané výsledky se ani při použití jiné míry vzdálenosti, než je Manhattanská, příliš neliší. Hodnota 20 vychází z mé dobré zkušenosti. Většinou zaručuje rychlý výpočet a uspokojivé výsledky.

Určení Elo jednotlivých týmů

A. Použij výsledky z bodu 1 a spočítej Elo pomocí vzorců (4.2) až (4.5) pro prvních 55 % zápasů první sezóny.

B. Získané Elo hodnoty jednotlivých týmů použij jako jejich počáteční hodnoty.

Konkrétní inicializační hodnoty Elo týmů v první sezóně jsou v příloze 1. Za nejkvalitnější týmy byly označeny Sparta a Slavia, zatímco na opačném konci jsou Kladno a Znojmo.

4.2.7 Optimalizace výpočetních konstant

Pro výpočet Elo ratingu, který bude použit jako odhad kvality týmu v každém konkrétním zápase, který odehraje, už jsou k dispozici počáteční hodnoty týmů i *HGA*. Pořád ale zbývá určit klíčovou konstantu *K* a konstantu úpravy výhody domácího prostředí *c*, které jsou potřebné pro výpočet rovnic (4.2) a (4.5).

Protože do predikčních modelů v kapitole 6 jako prediktory vstupují proměnné odvozené z Elo hodnot, budou konstanty optimalizovány tak, aby samotný Elo rating byl schopný věrohodně výsledky odhadnout. Níže popsaný model bude použit i jako benchmark přesnosti předpovědí v podkapitole 6.1.

$$ {}_i p_D = \frac{\sum_{i=1}^{n} {}_i s_D}{n} \Big| W_e = {}_i W_e \ , \tag{4.6} $$

$$ {}_i p_R = \frac{\sum_{i=1}^{n} {}_i s_R}{n} \Big| W_e = {}_i W_e \ , \tag{4.7} $$

$$ {}_i p_H = \frac{\sum_{i=1}^{n} {}_i s_H}{n} \Big| W_e = {}_i W_e \ , \tag{4.8} $$

kde *n* je počet zápasů, které byly ohodnoceny po zaokrouhlení na dvě desetinná místa stejným W_e jako předpovídaný zápas. Pravděpodobnost výhry domácích, remízy a výhry hostujícího týmu je relativní četností těchto jevů mezi zápasy, ve kterých se střetly – po započtení výhody domácího prostředí – stejně kvalitní týmy.

Zjevnou nevýhodou tohoto přístupu je fakt, že pro zápasy, ve kterých se potkává jasný favorit s outsiderem, nemusí být v historii takový počet zápasů, aby se předpověď dala považovat za dostatečně přesnou. Je-li *n* malé, mohou rozhodovat pouze jednotky pozorování. Proto budou konstruovány i další predikční modely, které v takové situaci mají lepší vlastnosti. Přestože se jedná o na první pohled vážný problém, pro předpovědi nadstavbové části tomu tak nebude. Na obr. 2.1 je znázorněno, že nadstavbovou část budou mezi sebou hrát týmy z určité části tabulky, kde by se měly pohybovat týmy podobné kvality a k hraničním zápasům jasných favoritů s outsiderem v nich docházet nebude.

V této části výpočtů nebudou využita data označená na obr. 3.1 jako *testovací*, přesnost předpovědí bude spočítána ze zápasů spadajících do kategorie *validační*.

Elo se aktualizuje po každém zápase, vyjádřit jednoduše vliv růstu nebo poklesu K a c proto nelze. Co nejlepší kombinaci parametrů je nutné určit empiricky. Byly spočítány předpovědi všech zápasů pro všechny kombinace parametrů $K \in \{10, 11, 12, ..., 29, 30\}$ a $c \in \{0{,}050; 0{,}055; 0{,}060; ..., 0{,}095; 0{,}100\}$. Následně byly odhadnuty metodou nejmenších čtverců modely vícenásobné regrese (James a kol., 2013), do kterých jako vysvětlovaná proměnná vstupovala vyhodnocovací kritéria z podkapitoly 3.4, a jako vysvětlující oba parametry jako polynom 3. stupně

$$RPS = \beta_0 + \beta_1 K + \beta_2 K^2 + \beta_3 K^3 + \beta_4 c + \beta_5 c^2 + \beta_6 c^3,$$

$$MSE = \beta_0 + \beta_1 K + \beta_2 K^2 + \beta_3 K^3 + \beta_4 c + \beta_5 c^2 + \beta_6 c^3,$$

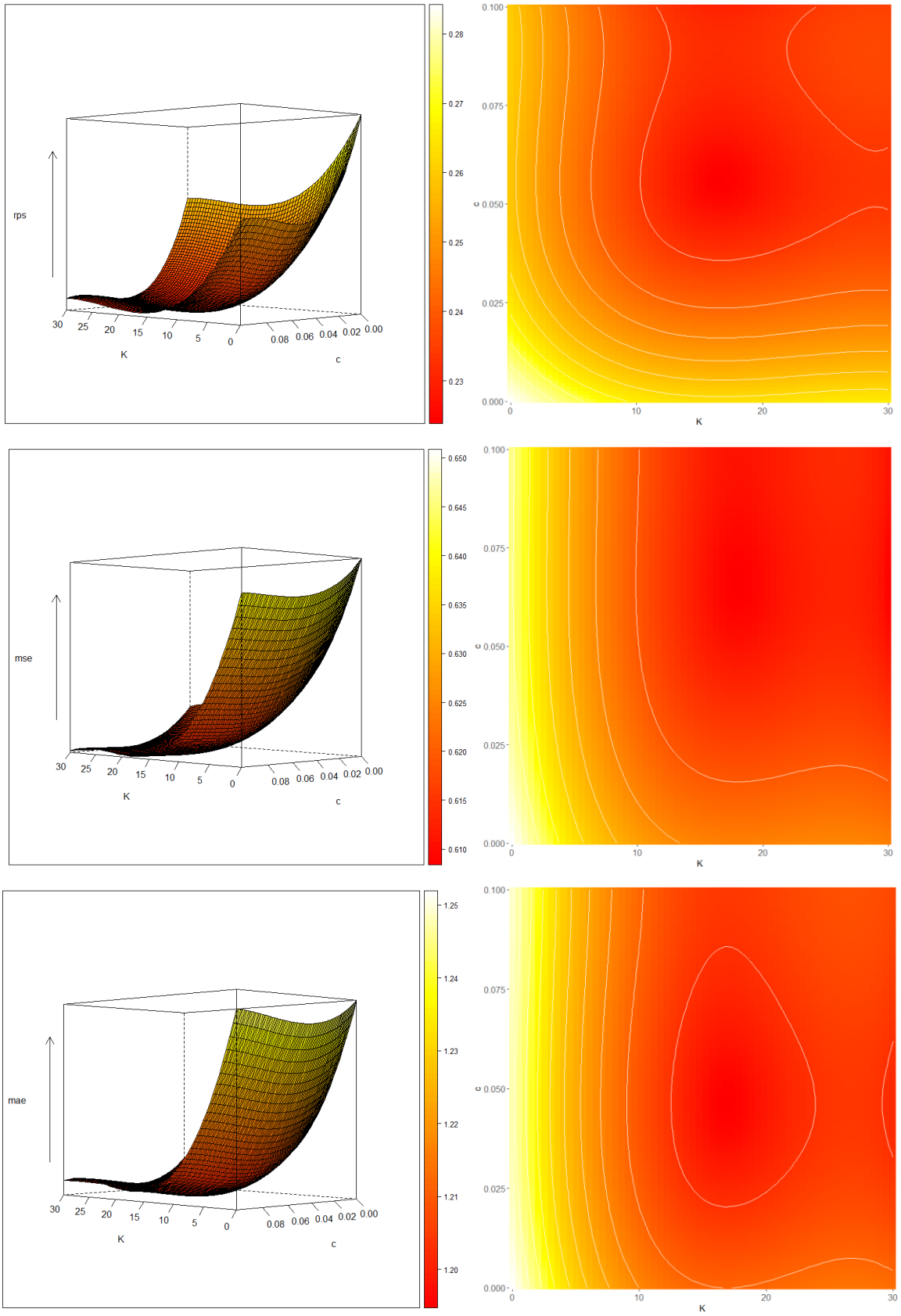
$$MAE = \beta_0 + \beta_1 K + \beta_2 K^2 + \beta_3 K^3 + \beta_4 c + \beta_5 c^2 + \beta_6 c^3.$$

Na obr. 4.4 si lze všimnout, že odhadnuté funkce pro jednotlivá kritéria se zásadně neliší, přestože rozdíly se pochopitelně identifikovat dají. Zobrazen je průběh funkcí pouze pro takové hodnoty K a c, které mohou být pro výpočet Elo vhodné. Důležitým poznatkem je fakt, že ani jedna z funkcí nemá v takto nastavených logických mezích globální minimum. Určit ideální kombinaci obou parametrů na základě většího počtu lokálních extrémů všech funkcí uspokojivě nelze. Jako rozumný nápad se proto jeví zafixovat konstantu K nebo c na zvolené hodnotě a tu zbývající optimalizovat.

Tab. 4.1 Nejlepších 5 kombinací parametrů *K* a *c*

K	***c***	**RPS**		**MSE**		**MAE**		**Průměr pořadí**
		hodnota	pořadí	hodnota	pořadí	hodnota	pořadí	
15	0,0650	0,2078	8	0,6087	2	1,1957	2	4,0000
22	0,0550	0,2076	3	0,6097	15	1,1953	1	6,3333
29	0,0550	0,2072	1	0,6088	3	1,1982	20	8,0000
29	0,0600	0,2074	2	0,6092	5	1,1982	21	9,3333
21	0,0550	0,2078	6	0,6101	20	1,1969	5	10,3333

Hodnoty *K* a *c* se řádově výrazně liší, proto je vhodné pro rozhodnutí o tom, jaký parametr zafixovat, použít variační koeficient (Řezanková a Löster, 2009). Ten vyjadřuje variabilitu proměnné relativně k jejímu průměru. Z 231 spočtených kombinací parametrů je variační koeficient pěti nejlepších kombinací, prezentovaných v tab. 4.1, pro *K* roven 0,229 a pro *c* 0,069. Koeficient úpravy domácího prostředí *c* je stabilnější a bude proto zafixován na hodnotě 0,058 (průměr z 5 nejlepších kombinací).



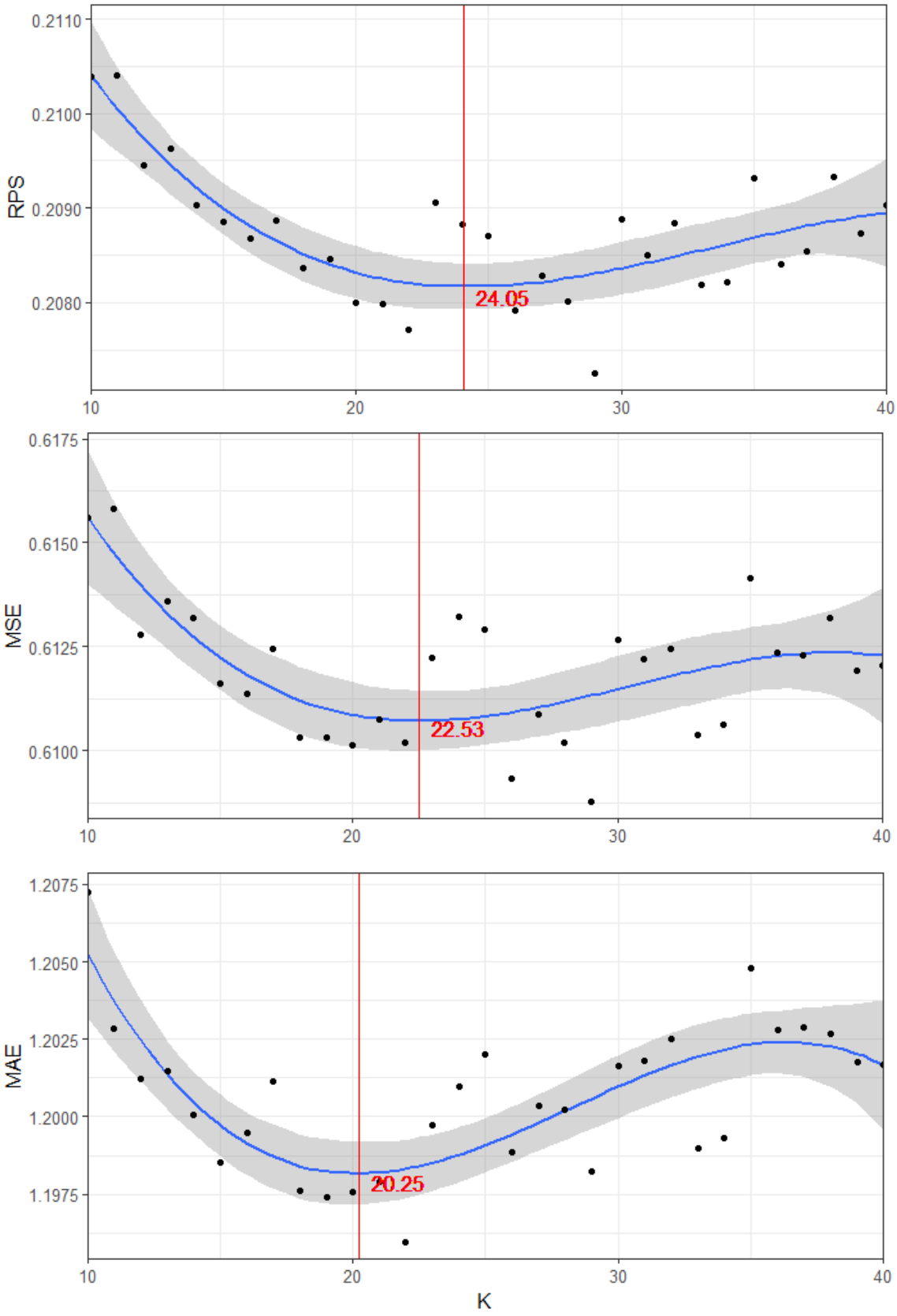
Obr. 4.4 Závislost přesnosti předpovědí na *K* a *c* (nahoře RPS, uprostřed MSE, dole MAE)

Po nastavení hodnoty *c* na 0,058 byl znovu odhadnut metodou nejmenších čtverců lineární model, kde vysvětlované proměnné byly vyhodnocovací kritéria přesnosti predikce a vysvětlující proměnnou byla hodnota *K* jako polynom 3. stupně.

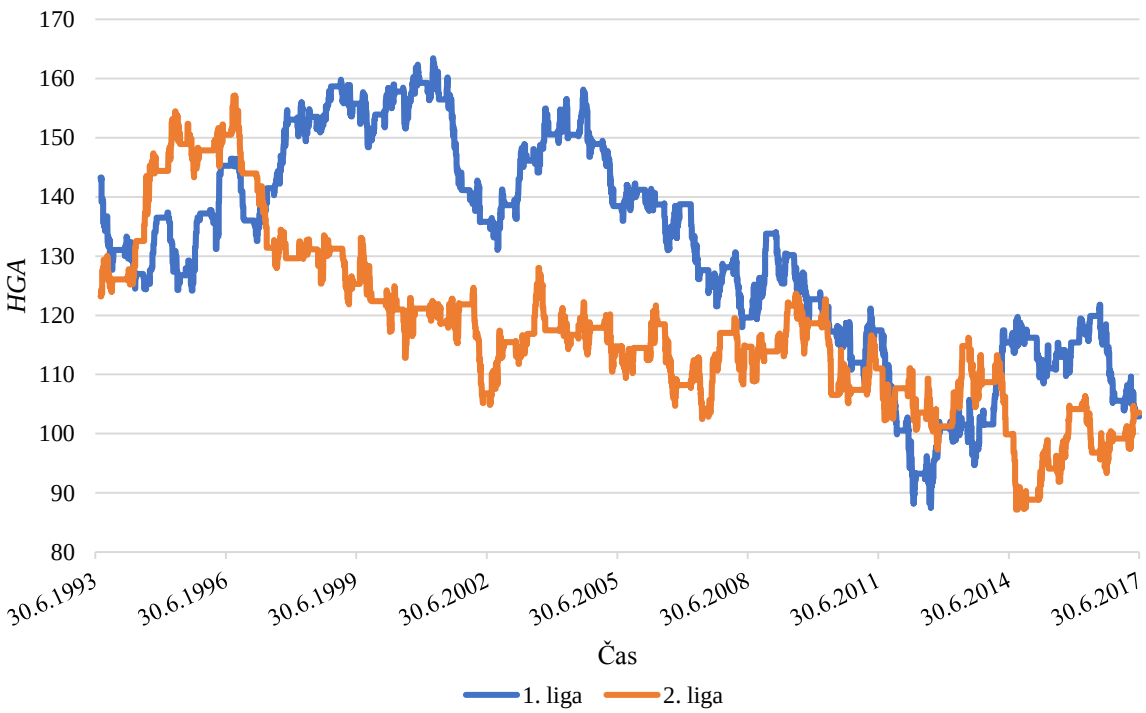
Výsledky jsou prezentovány na obr. 4.5. Zobrazeny jsou i intervaly spolehlivosti, přestože nebyla zkontrolována korektnost lineárních modelů. V kontextu prováděné analýzy to není nutné, neboť se zde nepracuje s rozptylem odhadů parametrů, ale využíván je pouze odhad metodou nejmenších čtverců, na který žádné nároky kladeny nejsou. Zobrazení intervalů spolehlivosti tak může přísný čtenář brát jen jako grafický prvek.

Výsledky vysvětlují, proč je u blogerů využívajících Elo rating oblíbená hodnota *K* kolem 20. Při zprůměrování minimálních hodnot, které jsou na obr. 4.5 zapsány červenou barvou, je výsledek 22,27. Nemá velký význam pracovat s desetinným číslem a hodnota *K* bude proto nastavena na 22.

Získáním této nejdůležitější konstanty je zakončeno hledání výpočtu co nejlepšího ukazatele kvality týmu. Elo rating v sobě implicitně zahrnuje i výhodu domácího prostředí, která je také vyjádřena v Elo bodech. Pro zajímavost a pro porovnání s obr. 4.1 je takto získaná výhoda zobrazena na obr. 4.6.



Obr. 4.5 Přesnost předpovědí v závislosti na *K*, *c* = 0,058
(nahoře RPS, uprostřed MSE, dole MAE)



Obr. 4.6 Výhoda domácího prostředí měřená v Elo bodech

5 Konstrukce prediktorů

Při vytváření prediktivních modelů by sice stačilo použít pouze Elo hodnoty kvality týmu a výhody domácího prostředí, ale budou-li tyto hodnoty použity pro výpočet dalších odvozených prediktorů, mohou být z nich zkonstruované modely přesnější.

Abych toto tvrzení podložil, je vhodné uvést jako příklad Poissonovu regresi. S její pomocí je výsledek zápasu předpovídán za využití odhadu průměrného počtu vstřelených branek oběma týmy. Jen při použití základních Elo hodnot by model nebyl schopen rozlišit gólovou potenci týmu a odhadoval by stejný průměrný počet branek všem týmům se stejným Elo hodnocením. Při zkonstruování prediktorů, které vyjadřují gólovou potenci týmů, bude model schopen mezi týmy najít jemnější rozdíly a – ideálně – nabídnout i lepší predikce.

Smysluplných prediktorů lze odvodit stovky, tisíce a možná i více. V rámci této práce není vysoký počet nutný. Bude vyzkoušeno více modelů, a ten, který nabídne nejpřesnější odhady výsledků, bude použit k simulaci ligy podle nového formátu. Velké množství prediktorů se zkoumá zejména při využívání předpovědí za účelem kurzového sázení. V něm může každá setina přesnosti výrazně ovlivnit ziskovost sázkaře. Ti se tak snaží nalézt „zázračný“ prediktor, který jim v tom pomůže.

V tab. 5.1 je přehled prediktorů, které budou k dispozici pro konstrukci modelů. První čtyři lze označit za základní proměnné, které vycházejí pouze z Elo hodnocení popsaném v podkapitole 4.2. Ostatní jsou buď průměry nebo mediány hodnot za poslední (až) 4 zápasy nebo (až) 8 zápasů. Pokud například tým postoupí do druhé ligy a odehraje v ní 2 zápasy, potom pro předpověď 3. zápasu mají ukazatele posledních 4 zápasů (značené předponou *L4*) a 8 zápasů (*L8*) stejnou hodnotu. Důvod je co nejmenší ztráta záznamů, které bude možné použít v regresních modelech, které nepovolují chybějící hodnoty. V případě, že tým v těchto zápasech vstřelil celkem 4 branky, potom atributy *L4goly* a *L8goly* mají hodnotu 2. Dva horizonty jsou uvažovány k zachycení krátkodobé (*L4*) a středně až dlouhodobé výkonnosti (*L8*).

V tab. 5.1 jsou proměnné zobrazeny z pozice jednoho týmu, například domácího. Proměnné, které nesou hodnoty těchto ukazatelů pro soupeřův tým, mají příponu *sp*. *L4soupMediansp* je tím pádem ukazatel toho, na jak silné protivníky soupeř týmu narazil v posledních 4 zápasech.

Může se zdát, že každý model může být postaven až na 32 prediktorech. Ve skutečnosti tomu tak nebude, jelikož mezi některými je funkční závislost. Například *pelo*, *elodif* a *HGA* jsou proměnné ze vzorce (4.3). U každého modelu bude uveden způsob, jakým byly prediktory pro ten konkrétní model vybrány.

Tab. 5.1 Přehled prediktorů

Označení	Popis	Motivace
elo	Elo hodnocení kvality týmu.	Ukazatel kvality týmu.
elodif	Rozdíl v Elo hodnocení týmů.	Jaký je mezi týmy kvalitativní rozdíl.
HGA	Vyčíslení HGA v Elo bodech.	Výhoda, jakou tým získává na domácím stadionu.
pelo	W_e z rovnice (4.3)	Modely budou moct vycházet z predikcí samotného Elo.
ziskElo	Průměrná změna Elo týmu.	Kvantifikace aktuální formy.
soupMedian	Medián Elo hodnocení posledních soupeřů týmu.	Vyjádření náročnosti posledních zápasů.
goly	Průměr vstřelených branek.	Gólová potence týmu.
golys	Průměr inkasovaných branek.	Náchylnost k inkasování gólu.
golyNAelo	Průměr z počtu branek po vynásobení Elo soupeře.	Gólová potence se zohledněním síly soupeře.
golysNAelo	Průměr z počtu inkasovaných branek po vynásobení převrácenou hodnotou Elo soupeře.	Náchylnost k inkasování gólu se zohledněním síly soupeře.

6 Predikční modely

Potřebné hodnoty prediktorů byly spočítány, zavedeny byly i podmínky vyhodnocení modelů, nezbývá než vystavět několik predikčních modelů a vybrat mezi nimi ten, který bude nejlépe předpovídat výsledky nejvyšších dvou českých fotbalových soutěží. Pro vyhodnocení modelů v této kapitole budou použita data označená na obr. 3.1 jako *testovací*. Nebude-li konstruovaný model vyžadovat optimalizaci parametrů, bude odhadnut z dat trénovacích i validačních. V opačném případě bude odhadnut na trénovacích datech a optimální parametry budou hledány za použití validačních dat. To bude vlastně jen případ vícenásobné lineární regrese v podkapitole 6.4.

6.1 Benchmark model

Tento model vychází pouze z rozdílu v kvalitě týmů vyjádřené Elo body, byl již představen v podkapitole 4.2.7. Odhadnut byl na trénovacích a validačních datech.

Tab. 6.1 Přesnost předpovědí benchmarkového modelu

Model	RPS	MSE	MAE
Benchmark	0,1991	0,5879	1,1752

Při porovnání hodnot z tab. 4.1 a tab. 6.1 je vidět, že se přesnost předpovědí podle všech kritérií zlepšila. Je nepochybně dobrým signálem, že s více pozorováními (pro trénování modelu přibyla validační data) došlo ke zlepšení předpovědí.

6.2 Logistická regrese

Možností, jak zkonstruovat model logistické regrese je několik. Každý zápas může skončit třemi různými výsledky – výhrou domácího týmu, remízou nebo výhrou hostujícího týmu. Budou-li data na vstupu připravena v takové podobě, že každý zápas bude v datové matici dvakrát, zvlášť z pohledu domácího a hostujícího týmu, lze celé pravděpodobnostní rozdělení výsledku zápasu odhadnout jednorovnicovým modelem (6.1). Ten určí odhady pravděpodobnosti vítězství obou týmů a odhad remízy se určí jako doplněk do jedné.

Tento přístup má tu zásadní chybu, že zřejmě porušuje předpoklad vzájemné nezávislosti pozorování. Vzhledem k tomu, že tato práce je empirická a zaměřená na kvalitu předpovědí, bude i přes porušení předpokladů vyzkoušen i tento přístup. Zjednodušení modelu ze dvou – potažmo tří – rovnic na jedinou může najít v praxi své využití, zejména bude-li dávat uspokojivé výsledky.

Pokud bude v datové matici každý zápas právě jednou, bude odstraněno porušení předpokladu nezávislosti pozorování. Model však bude o něco složitější. Bude zapotřebí odhadnout minimálně dva modely logistické regrese pro dva různé výsledky a třetí určit doplňkem do jedné. Například bude odhadnut zvlášť model pro předpověď výhry domácích (6.2) a zvlášť model pro předpověď pravděpodobnosti remízového výsledku (6.3). Doplněk do jedné potom určí odhad pravděpodobnosti výhry hostujícího celku. Tímto způsobem lze dojít ke třem různým modelům podle toho, pro jaké výsledky bude odhadnut regresní model, a jaký výsledek bude dopočítán.

U dvojrovnicových modelů nehrozí, že by součet dvou odhadů pravděpodobností jednotlivých výsledků byl větší než jedna. V modelu pro odhad pravděpodobnosti výhry domácího týmu (6.2) je proměnná vyjadřující rozdíl ve kvalitě týmů *elodif* s kladným koeficientem. V modelu pro hostující tým (6.4) je proměnná *pelo*, která vyjadřuje totéž jinou formou, se záporným koeficientem. Remízový výsledek se velmi špatně predikuje a odhad pravděpodobnosti remízy jen zřídka překročí hodnotu 0,3. Remízový odhad je tím pádem vždycky nízký a provázanost proměnných v ostatních modelech zaručí, že součet dvou vybraných hodnot je menší než jedna.

Poslední možností je třírovnicový přístup, kdy bude odhadnut regresní model pro všechny výsledky. Takto zavedený přístup už však nezaručí součet odhadů pravděpodobností z jednotlivých modelů rovný jedné. K napravení tohoto nedostatku bude využita funkce softmax (Polamuri, 2017). Funkce se používá i ve výstupní vrstvě neuronových sítí a namapuje hodnoty z jednotlivých modelů tak, aby byl jejich součet roven jedné.

Tab. 6.2 Přesnost předpovědí logistické regrese

Model	RPS		MSE		MAE	
	hodnota	rozdíl	hodnota	rozdíl	hodnota	rozdíl
Benchmark	0,1991	–	0,5879	–	1,1752	–
Jednorovnicový	0,1978	– 0,0012	0,5848	– 0,0032	1,1763	0,0011
Dvojrovnicový H	0,1981	– 0,0009	0,5871	– 0,0009	1,1720	– 0,0032
Dvojrovnicový R	0,1974	– 0,0017	0,5834	– 0,0045	1,1635	– 0,0118
Dvojrovnicový D	0,1989	– 0,0002	0,5885	0,0006	1,1756	0,0004
Třírovnicový	0,2135	0,0144	0,6192	0,0312	1,2758	0,1006

Pozn.: Písmeno u dvojrovnicových modelů určuje dopočítávanou pravděpodobnost. (H – hosté, R – remíza, D – domácí)

Z tab. 6.2 je patrné, že největší zlepšení oproti benchmarkovému modelu přinesl dvojrovnicový model s dopočítanou remízou. Tento výsledek se dá považovat za víceméně očekávaný, neboť správně předpovědět remízu je velmi obtížné. Model tak profituje ze snazší predikce vítězství domácích a hostujících týmů.

Jako nejhorší se ukázal třírovnicový model, který nedokázal vylepšit benchmarkový model ani v jednom kritériu.

6.2.1 Jednorovnicový model

Všechny regresní modely v této práci byly odhadnuty v programu R (R Core Team, 2018) za použití balíčku My.stepwise (International-Harvard Statistical Consulting Company, 2017), který vybírá proměnné modelu metodou stepwise založené na testu věrohodnostním poměrem. Výstupem je rovněž faktor zvětšení rozptylu (VIF – Variance Inflation Factor), vše popsané v (Hebák, 2015). O výběru nejlepší podmnožiny prediktorů je rozhodováno na pětiprocentní hladině významnosti.

Jak již bylo zmíněno, tento model odhaduje pravděpodobnost výhry domácího a hostujícího týmu jedinou rovnicí

$$\log\frac{p_D}{1-p_D} = \log\frac{p_H}{1-p_H} = -0{,}902 + 0{,}005\cdot HGA + 0{,}003\cdot elodif + \\ + 0{,}115\cdot L8golyssp + 0{,}230\cdot L8goly - 0{,}128\cdot L4golys - \\ - 0{,}161\cdot L8golysp + 0{,}146\cdot L4golyssp, \tag{6.1}$$

kde je pro lepší čitelnost potlačen index *i* určující konkrétní zápas. Při srovnání výstupu s přehledem prediktorů v tab. 5.1 je vidět, že tým má větší šanci na výhru, pokud:

- hraje na domácím stadionu (*HGA*),
- je kvalitnější než jeho soupeř (*elodif*),
- vstřelil v posledních 8 zápasech hodně branek (*L8goly*),
- soupeř hodně branek v posledních 4 a 8 zápasech inkasoval (*L4golyssp* a *L8golyssp*).

Šance naopak klesá, pokud tým v posledních 4 zápasech často inkasoval (*L4golys*) a jeho soupeř v posledních 8 zápasech často skóroval (*L8golysp*). Všechny tyto výsledky dávají logický význam a ve zkratce říkají, že lepší tým, který střílí branky, a hraje proti často inkasujícímu týmu, má větší šanci zvítězit.

6.2.2 Vícerovnicové modely

Všechny vícerovnicové modely, jak byly popsány výše, vycházejí z regresních modelů pro jednotlivé výsledky fotbalového zápasu. Ty jsou všechny zapsány z důvodu snazší manipulace s daty z pohledu domácího týmu. Pro výhru domácího týmu

$$\log\frac{p_D}{1-p_D} = -0{,}102 + 0{,}003\cdot elodif + 0{,}182\cdot L4golyssp + 0{,}00013\cdot L8golyNAelo - \\ - 0{,}154\cdot L8golysp - 1{,}138\cdot L4golys, \tag{6.2}$$

remízu

$$\log \frac{p_R}{1-p_R} = -0{,}712 + 0{,}366 \cdot pelosp - 0{,}219 \cdot L8golyssp - 0{,}00009 \cdot L8golyNAelo \quad (6.3)$$

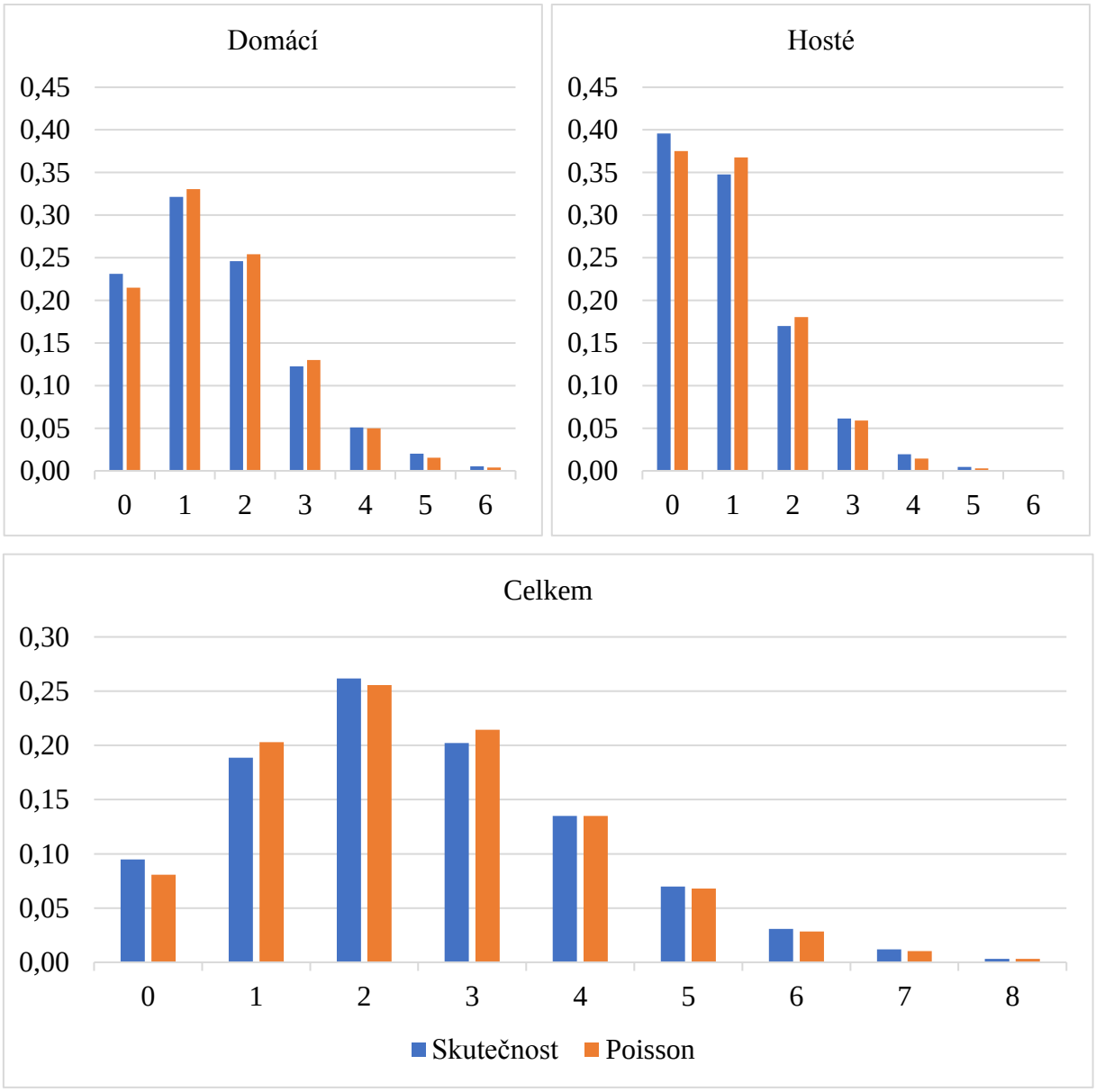
a prohru domácího – tj. výhru hostujícího – týmu

$$\begin{aligned} \log \frac{p_H}{1-p_H} = {} & 0{,}432 - 2{,}929 \cdot pelo - 0{,}010 \cdot L4ziskElo + 0{,}268 \cdot L8golysp + \\ & + 0{,}201 \cdot L8golys - 0{,}143 \cdot L4golyssp - 0{,}0001 \cdot L8golyNAelo - \\ & - 0{,}019 \cdot L8ziskElosp . \end{aligned} \quad (6.4)$$

Model remízy z rovnice (6.3) byl oproti výstupu My.stepwise upraven. Původně do něj vstupovaly i proměnné *elodif* a *L4goly*. Důsledkem přítomnosti *elodif* spolu s *pelosp* byla vysoká multikolinearita, která znemožňovala správnou interpretaci výsledků. VIF pro obě proměnné se blížil 30, o něco větší byl pro proměnnou *elodif*, ta byla proto odebrána. Tento zásah do modelu způsobil na stanovené 5 % hladině významnosti statistickou nevýznamnost *L4goly*, proto byla odebrána i tato proměnná.

6.3 Poissonův model

Myšlenka předpovídání fotbalových výsledků za použití Poissonovy regrese založené na Poissonovu rozdělení vychází z toho, že počet branek v zápase má přibližně Poissonovo rozdělení. Smysluplnost použití tohoto rozdělení byla nejdříve ověřena pomocí popisné statistiky na celém datovém souboru.



Obr. 6.1 Podíly zápasů podle počtu vstřelených branek
(odhad parametru Poissonova rozdělení = průměr vstřelených branek)

Nejdříve bude pomocí Poissonovy regrese (Daróczi, 2015) předpovězen střední počet vstřelených branek obou týmů v zápase. Pomocí Poissonova rozdělení (Marek, 2012) bude potom určen odhad pravděpodobnosti přesného počtu branek, které týmy vstřelí. Odhad pravděpodobnosti konečného skóre se určí jako součin těchto odhadů. Součtem odhadů pravděpodobností výsledků, které znamenají výhru domácích, hostů a remízy, se získají konečné předpovědi. Podle příkladu v tab. 6.3 má Hradec 34,51 % šanci na výhru, Liberec 36,56 % a předpověď remízy má 28,93 %. Zápas skončil 0:1, což byl 2. nejpravděpodobnější výsledek (znázorněno červeně).

Tab. 6.3 Odhad pravděpodobnostního rozdělení výsledku zápasu

Hradec – Liberec, 22. 4. 2017

Pravděpodobnost			Góly hosté						
			0	1	2	3	4	5	…
			0,322	0,365	0,207	0,078	0,022	0,005	0,001
Góly domácí	0	0,336	0,108	0,123	0,069	0,026	0,007	0,002	0,000
	1	0,366	0,118	0,134	0,076	0,029	0,008	0,002	0,000
	2	0,200	0,064	0,073	0,041	0,016	0,004	0,001	0,000
	3	0,073	0,023	0,027	0,015	0,006	0,002	0,000	0,000
	4	0,020	0,006	0,007	0,004	0,002	0,000	0,000	0,000
	5	0,004	0,001	0,002	0,001	0,000	0,000	0,000	0,000
	…	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Pozn.: Výstup jednorovnicového modelu (popsán níže), kde střední hodnota vstřelených branek domácím týmem je 1,0916 a hostujícím 1,1325. Zeleně výhra domácích, oranžově výhra hostů.

Takto popsaný model zavádí předpoklad vzájemné nezávislosti branek vstřelených domácím a hostujícím týmem. To znamená, že ať například domácí tým vstřelí v zápase jednu nebo osm branek, na počet gólů, které inkasuje od hostujícího celku, to nebude mít vliv. Jedná se o běžnou praxi popsanou např. v (Cronin, 2017). Korelace mezi počtem vstřelených branek domácími a hostujícími týmy v celých datech je přibližně – 0,05. Čistě frekventistickým pohledem se na vzorku přesahujícím 11 000 zápasů hovoří o statistické významnosti lineárního vztahu na libovolně malé hladině významnosti. Takový rychlý závěr by zřejmě nebyl úplně správný.

Při pohledu na počet vstřelených branek jako na proměnnou ordinální lze sestavit kontingenční tabulku a použít chí-kvadrát test nezávislosti (Hebák, 2015). Bohužel ani ten nelze vyhodnotit zcela objektivně, jelikož pro splnění předpokladu tohoto testu (očekávané četnosti alespoň 5 ve více než 80 % polích) je nutné sáhnout ke spojení některých kategorií. To je dáno malým zastoupením zápasů s vysokým skóre. Počet zápasů podle skóre nabízí příloha 2.

Před spojením kategorií – které vede ke splnění předpokladů – byla p-hodnota pro zápasy první ligy 0,80, pro druhou ligu 0,16. Společně pro zápasy obou soutěží klesla na 0,04. Na tom lze i přes porušení předpokladů opět doložit fakt velkého vzorku ovlivňující p-hodnotu. Při spojení málo zastoupených kategorií se p-hodnoty testu dostanou prakticky na nulu, při jejich odseknutí se pohybují okolo jedné tisíciny. Společným jmenovatelem, který podporuje zamítnutí

hypotézy nezávislosti, jsou standardizovaná rezidua výsledků 1:0 a 0:1, které nastávají méně často než za platnosti hypotézy (standardizovaná rezidua menší než – 3). Častěji naopak končí zápasy remízou 1:1. Lze to vysvětlit zvýšenou snahou týmu, který prohrává pouze o gól o vyrovnání skóre. Velmi často se mu to povede, nebo při riskantním stylu hry naopak inkasuje a prohraje o dvě branky (standardizovaná rezidua větší než 2).

Bez nesouladu ve zmíněných případech by hypotéza nezávislosti zamítnuta nebyla. Možností, jak tento nedostatek překonat, může být například použití dvojrozměrného Poissonova rozdělení (Karlis a Ntzoufras, 2005) nebo upravený výpočet remízových výsledků (Cortis, 2018). V této práci bude předpoklad nezávislosti respektován. Stále se jedná o běžnou praxi a cílem této práce není konstrukce komplexních modelů, které nabízejí jen nepatrné vylepšení. Namísto toho tato práce zkoumá více jednodušších – a pro účely simulace dostačujících – modelů.

Tab. 6.4 Přesnost předpovědí Poissonova modelu

Model	RPS		MSE		MAE	
	hodnota	rozdíl	hodnota	rozdíl	hodnota	rozdíl
Benchmark	0,1991	–	0,5879	–	1,1752	–
Jednorovnicový	0,1979	– 0,0011	0,5847	– 0,0033	1,1793	0,0041
Dvojrovnicový	0,1976	– 0,0015	0,5843	– 0,0037	1,1681	– 0,0072

Stejně jako v případě logistické regrese lze zkonstruovat jednorovnicový model, který předpoví zároveň střední počet branek domácího i hostujícího týmu. I tentokrát si tento model vede oproti benchmarkovému modelu ve 2 ze 3 vyhodnocovacích kritérií lépe a pozadu zůstává pouze podle střední absolutní chyby. Dvojrovnicový model se nicméně i zde ukazuje jako lepší.

6.3.1 Jednorovnicový model

Při konstrukci Poissonovy regrese stojí za zmínku ještě předpoklad shody průměru s rozptylem v Poissonově rozdělení. Jde-li o branky domácích týmů, průměr je 1,537 a rozptyl 1,703, pro hosty jsou tato čísla 0,981 a 1,098. Pro jednorovnicový model jsou nicméně relevantní čísla bez rozlišení domácího a hostujícího týmu, průměrný počet branek je 1,259 a rozptyl 1,478.

Rozptyl je větší než průměr z důvodu několika málo zápasů, ve kterých padne hodně branek. Rozdíl to ale není nikterak velký, odhad disperzního parametru pro model (6.5) je 1,021, čímž je změna chybových odhadů a p-hodnot minimální. Použití Poissonovy regrese v základní podobě je proto v pořádku

$$\begin{aligned} \ln goly_D = \ln goly_H = &-0{,}133 + 75{,}240 \cdot L4golysNAelosp + \\ &+ 0{,}124 \cdot L8goly + 0{,}002 \cdot HGA + 0{,}081 \cdot L8golyssp + \\ &+ 0{,}001 \cdot elodif - 0{,}023 \cdot L4golys\,. \end{aligned} \tag{6.5}$$

Oproti výstupu My.stepwise byl tento model z důvodu multikolinearity upraven. Stejně jako do rovnice (6.3) i sem vstoupily proměnné *pelo* a *elodif*. Větší VIF (46,9) měla proměnná *pelo*, po jejím odstranění nebyly další zásahy do modelu nutné.

Z modelu vyplývá, že tým dá více gólů (a tím má větší šanci zápas vyhrát), pokud

- je kvalitnější (*elodif*),
- hraje na domácím stadionu (*HGA*),
- vstřelil v posledních 8 zápasech hodně branek (*L8goly*),
- soupeř v posledních 8 zápasech často inkasoval (*L8golyssp*),
- soupeř v posledních 4 zápasech inkasoval i od slabších týmů (*L4golysNAelosp*).

Šanci na skórování naopak snižuje vysoká gólová potence soupeře v posledních 4 zápasech (*L4golys*). Proti těmto závěrům nejde z logiky věci nic namítnout a podle výsledků v tab. 6.4 model zřejmě funguje poměrně spolehlivě.

6.3.2 Dvojrovnicový model

Dvojrovnicový model je založen na Poissonově regresním modelu zvlášť pro počet branek domácích a hostujících týmů. Pro domácí tým

$$\ln goly_D = 0{,}138 + 0{,}001 \cdot elodif + 0{,}119 \cdot L8goly + 0{,}128 \cdot L8golyssp - \\ - 0{,}038 \cdot L8golys \tag{6.6}$$

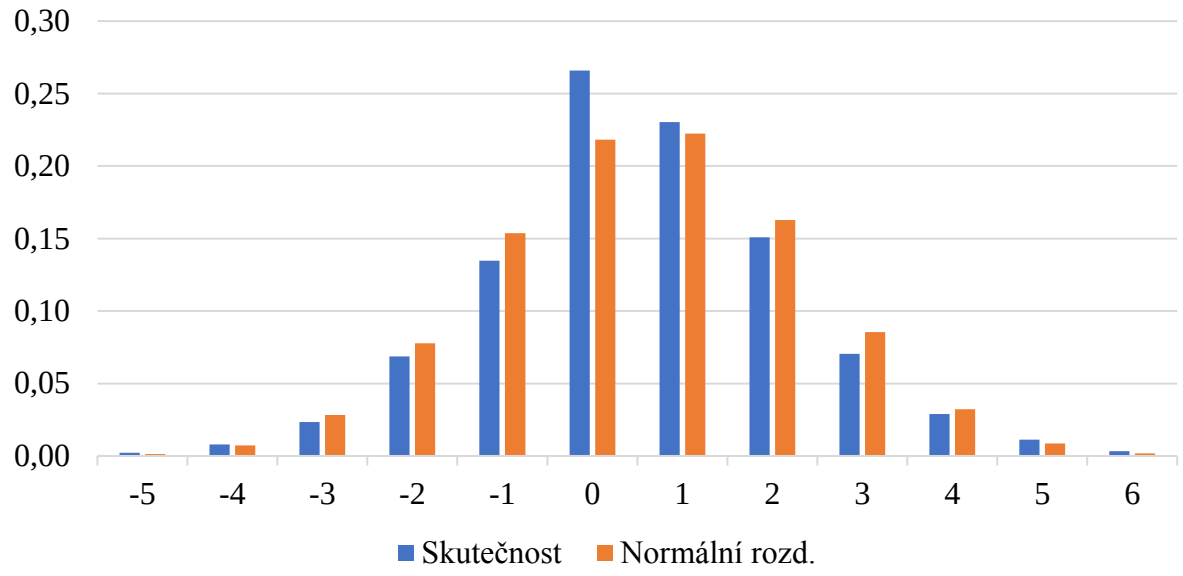
a hostující tým

$$\ln goly_H = -0{,}637 + 0{,}923 \cdot pelo + 184{,}340 \cdot L4golysNAelosp + 0{,}190 \cdot L8goly - \\ - 0{,}014 \cdot L8ziskElo - 0{,}076 \cdot L4golys - 0{,}055 \cdot L8golysp\,. \tag{6.7}$$

V rovnici (6.7) se poprvé objevuje proměnná s jiným než očekávaným znaménkem. *L8ziskElo* udává formu týmu v posledních 8 zápasech. Dalo se předpokládat, že tým, který v minulých zápasech podával lepší než očekávané výsledky, bude ve zlepšených výkonech pokračovat. Namísto toho se proměnná objevuje s minusem, což se dá interpretovat tak, že čím více hrál tým v minulých zápasech nad své možnosti, tím spíše se v zápase následujícím vrátí ke svým standardním (horším) výkonům.

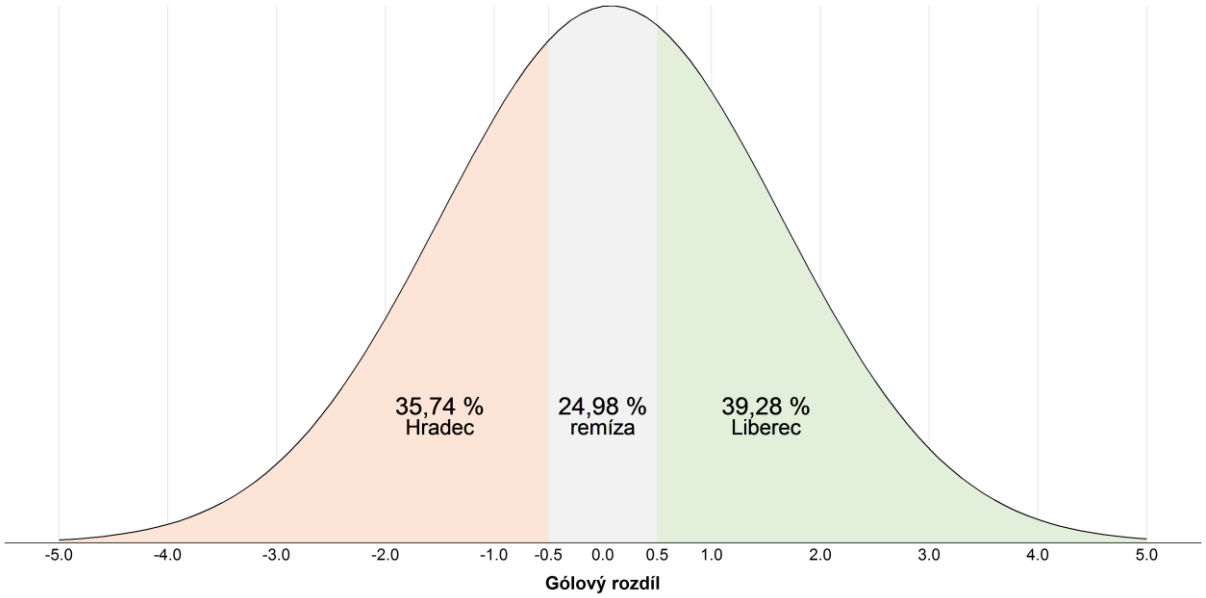
6.4 Vícenásobná lineární regrese

Brankový rozdíl, tedy rozdíl mezi brankami vstřelenými domácím a hostujícím týmem (nebo opačně) se blíží normálnímu rozdělení. S konceptem, jak toho využít přichází například (Cortins, 2015). Na obr. 6.2 jsou pro zachycení normálního rozdělení v diskretizované podobě použity intervaly: gólový rozdíl ± 0,5. Průměrný gólový rozdíl je 0,557 a rozptyl 2,935. Je vidět, že nejvíce je podhodnocen remízový výsledek, tedy nulový gólový rozdíl. Jedná se vlastně o analogii s Poissonovým modelem, který remízové výsledky také lehce podhodnocoval.



Obr. 6.2 Podíly zápasů podle gólového rozdílu z pohledu domácího týmu (odhad parametrů normálního rozdělení = průměr a rozptyl gólového rozdílu)

Výstavba modelu začne vícenásobnou lineární regresí opět za pomoci balíčku My.stepwise. Regresní model bude odhadnut pouze na trénovací podmnožině dat. Výstup modelu v podobě předpovědi gólového rozdílu a směrodatné chyby předpovědi bude použit pro odhad pravděpodobností výsledku zápasu ilustrované na obr. 6.3.



Obr. 6.3 Odhad rozdělení gólového rozdílu Hradec – Liberec, 22. 4. 2017 (průměr 0,073, odchylka 1,569)

Z obr. 6.3 je zjevné, že předpověď výrazně ovlivňuje to, jak jsou nastaveny hranice určující pravděpodobnost remízy. Model se tak dá správným nastavením těchto hranic dále vylepšit. Při použití intervalu od – 0,5 do 0,5 model předpovídá průměrnou šanci výhry domácího týmu 52,15 %, podíl remízových výsledků by měl být pouze 21,91 %. Skutečné hodnoty jsou 48,14 % a 26,08 %, výhra hostů je predikována vcelku přesně.

Nabízí se tedy otázka, podle jakého kritéria optimalizovat nastavení intervalu určujícího remízu. K optimalizaci bude použita část dat označená na obr. 3.1 jako *validační* a Řešitel MS Excel. Interval by se měl pohybovat okolo nuly, proto byly nastaveny omezující podmínky pro dolní a horní mez intervalu. Dolní mez musí být větší nebo rovna – 1,5, horní mez menší nebo rovna 1,5. Aby bylo zabráněno úplnému ignorování remízových výsledků, musí být délka intervalu určující remízu alespoň 0,5.

Optimalizačními kritérii (účelovou funkcí), budou vyhodnocovací kritéria MSE, MAE a RPS. Pouze při konstrukci tohoto modelu je použita optimalizace vůči vyhodnocovacím kritériím, čímž může být oproti ostatním lehce zvýhodněn. Na druhou stranu základ modelu v podobě regrese byl oproti ostatním modelům odhadnut na menším počtu zápasů, čímž byl naopak znevýhodněn.

Tab. 6.5 Optimální meze pro výpočet pravděpodobnosti remízy podle vyhodnocovacích kritérií

Kritérium	Dolní mez	Horní mez
RPS	– 0,5117	0,6724
MSE	– 0,5005	0,6733
MAE	– 0,5000	0,0000

Tab. 6.6 Přesnost předpovědí modelu s vícenásobnou lineární regresí

Model	RPS		MSE		MAE	
	hodnota	rozdíl	hodnota	rozdíl	hodnota	rozdíl
Benchmark	0,1991	–	0,5879	–	1,1752	–
Optimální MSE	0,1977	– 0,0013	0,5840	– 0,0039	1,1744	– 0,0009
Optimální MAE	0,2095	0,0104	0,6318	0,0439	1,1091	– 0,0662
Optimální RPS	0,1977	– 0,0013	0,5840	– 0,0039	1,1744	– 0,0009

Z tab. 6.5 a tab. 6.6 lze mimo jiné vyčíst jistý vztah mezi vyhodnocovacími kritérii. Zatímco MSE a RPS jsou si velmi podobné, střední absolutní chyba se chová odlišně. Nejnižší MAE by bylo dosaženo ignorováním remízového výsledku, tomu nicméně zamezily omezující podmínky.

Již bylo ukázáno, že předpovědět remízové výsledky je velmi složité. Šance remízového výsledku se nejčastěji pohybuje mezi 20 % a 30 %. Absolutní odchylka předpovědi, pokud k remíze skutečně dojde, je tím pádem vždy vysoká. Střední absolutní odchylka predikce remízy se u většiny dříve prezentovaných modelů pohybovala okolo 0,38. Při konstantní předpovědi nulové šance na remízu se tato hodnota rovná podílu remízových výsledků, tedy zhruba 0,26. Nárůst chyb ostatních výsledků potom není natolik velký, aby se při optimalizaci skutečně nevyplatilo toto iracionální rozhodnutí o ignorování remízy učinit. Použití MAE jako vyhodnocovacího kritéria pro fotbalové zápasy se tak nejeví jako příliš správné.

6.4.1 Regresní model

Základem předpovědí je regresní model v podobě

$$GR_D = 0{,}459 + 0{,}003 \cdot elodif + 0{,}183 \cdot L4golyssp + 0{,}0002 \cdot L8golyNAelo - \\ - 238{,}100 \cdot L4golysNAelo - 0{,}0001 \cdot L8golyNAelosp, \tag{6.8}$$

kde GR_D je gólový rozdíl z pohledu domácích, tedy $goly_D - goly_H$. Do tohoto modelu byly zařazeny proměnné zohledňující kvalitu týmu a počet vstřelených branek. Brankový rozdíl – tím i šance na výhru domácího týmu – je tím větší, pokud

- je tým kvalitnější (*elodif*),
- dokázal vstřelit branky i silným soupeřům v posledních 8 zápasech (*L8golyNAelo*),
- soupeř v posledních 4 zápasech často inkasoval (*L4golyssp*).

Opačný efekt má pokud

- tým v posledních 4 zápasech inkasoval i od slabších týmů (*L4golysNAelo*),
- soupeř dokázal vstřelit branky i silným soupeřům (*L8golyNAelosp*).

Jak bylo vidět v tab. 6.6, předpovědi modelu byly velmi slušné a nabídly oproti benchmarkovému modelu jisté zlepšení. Přestože předpovědi jsou hlavním předmětem této části práce, vzhledem k využití nejen předpovědí, ale i směrodatné chyby předpovědí, není od věci podívat se na jednoduchou diagnostiku modelu. Příloha 3 nabízí graf rozdělení reziduí, které velmi pěkně kopíruje normální rozdělení. Také jsou zde rezidua vykreslena proti vyrovnaným hodnotám a proměnným modelu. Splnění předpokladu homoskedasticity se zdá být hraniční. Fakt, že předpovědi modelu si vedou velmi slušně, podporuje rozhodnutí, že model je pro svůj účel vhodný.

6.5 Srovnání modelů

Celkem bylo za pomoci logistické regrese (LR), Poissonovy regrese (PR) a vícenásobné lineární regrese (VlR) navrženo 10 modelů. Byly porovnávány s benchmarkovým modelem obsahujícím pouze Elo hodnocení a 8 z nich benchmarkový model, ve smyslu podmínek stanovených v podkapitole 3.4, překonalo.

Tab. 6.7 Srovnání predikčních modelů

Model	RPS		MSE		MAE		Průměr pořadí
	hodnota	pořadí	hodnota	pořadí	hodnota	pořadí	
LR 2r R	0,1974	1	0,5834	1	1,1635	2	1,33
PR 2r	0,1976	2	0,5843	4	1,1681	3	3,00
VlR MSE	0,1977	4	0,5840	2	1,1744	5	3,67
VlR RPS	0,1977	3	0,5840	3	1,1744	6	4,00
LR 2r H	0,1981	7	0,5871	7	1,1720	4	6,00
LR 1r	0,1978	5	0,5848	6	1,1763	9	6,67
PR 1r	0,1979	6	0,5847	5	1,1793	10	7,00
VlR MAE	0,2095	10	0,6318	11	1,1091	1	7,33
Benchmark	0,1991	9	0,5879	8	1,1752	7	8,00
LR 2r D	0,1989	8	0,5885	9	1,1756	8	8,33
LR 3r	0,2135	11	0,6192	10	1,2758	11	10,67

Pozn.: Název modelu zkracuje dříve zavedené názvy. První část tvoří typ regresního modelu, další části určují počet rovnic modelu a případně bližší specifikaci v podobě zavedené dříve.

Jak je vidět v tab. 6.7, na prvních třech místech jsou zástupci všech typů regresních modelů. Dá se říci, že model postavený na dvojrovnicové – rovnice (6.2) a (6.4) – logistické regresi s určením remízového výsledku doplňkem do jedné je výrazně lepší než ostatní. Tento model nabízí nejnižší RPS, MSE a druhou nejlepší MAE. Druhé místo v tomto kritériu je zapříčiněno modelem vícenásobné lineární regrese s optimalizací právě střední absolutní chyby. Důvod velmi nízké hodnoty při optimalizaci MAE byl zmíněn výše.

Pro simulaci ligy podle nového herního systému bude použit dvojrovnicový model logistické regrese popsaný v podkapitole 6.2. Odhad pravděpodobnosti výhry domácího týmu p_D určuje rovnice (6.2), p_H rovnice (6.4) a $p_R = 1 - p_D - p_H$. Hodnoty prediktorů pro simulované zápasy jsou spočteny po posledním reálně odehraném zápase. První a třetí nadstavbová skupina mohou být simulovány v jediném kroku, zatímco skupina o Evropskou ligu je simulována po jednotlivých dvojutkáních.

7 Nadstavba v sezóně 2016/17

Sezónu 2016/17 ovládla pražská Slavia před druhou Plzní a třetí Spartou. Kvalifikaci Ligy mistrů hrála Slavia s Plzní a do Evropské ligy postoupily týmy Sparty, Mladé Boleslavi a Zlína. Zlín si vybojoval přímé místo v základní skupině Evropské ligy skrze Pohár FAČR. Tato práce ligový pohár do úvahy nebere a soustředí se čistě na simulaci ligových zápasů. Následující řádky tak nebudou brát na účast Zlína v EL zřetel a budou se zabývat pouze šancí vybojovat si účast v evropských pohárech výkony v ligových zápasech. Jak by sezóna dopadla, kdyby se již minulou sezónu hrála nadstavbová část, bylo spočítáno na 100 000 simulacích průběhu ligy.

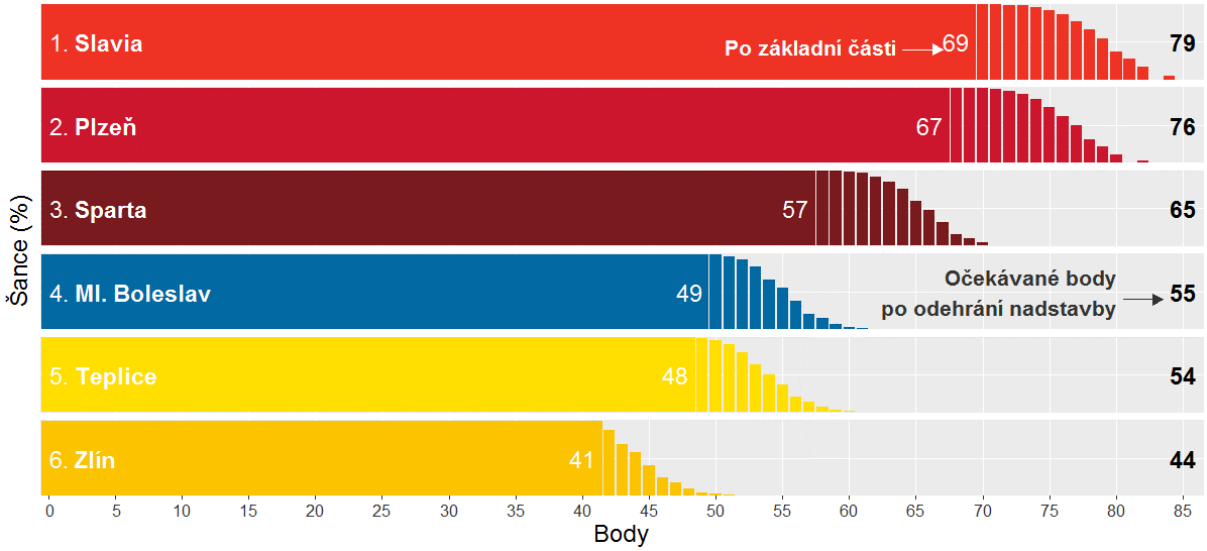
Tab. 7.1 Výsledná tabulka 1. ligy sezóny 2016/17

Pořadí	Klub	Z	V	R	P	G+	G–	GR	B
1.	Slavia	30	20	9	1	65	22	43	69
2.	Plzeň	30	20	7	3	47	21	26	67
3.	Sparta	30	16	9	5	47	26	21	57
4.	Ml. Boleslav	30	13	10	7	47	37	10	49
5.	Teplice	30	13	9	8	38	25	13	48
6.	Zlín	30	11	8	11	34	35	– 1	41
7.	Dukla	30	11	7	12	39	35	4	40
8.	Jablonec	30	9	12	9	43	38	5	39
9.	Liberec	30	10	9	11	31	28	3	39
10.	Karviná	30	9	7	14	39	49	–10	34
11.	Brno	30	6	14	10	32	45	–13	32
12.	Slovácko	30	6	14	10	29	38	– 9	32
13.	Bohemians 1905	30	7	7	16	22	39	–17	28
14.	Jihlava	30	6	9	15	26	47	–21	27
15.	Hradec	30	8	3	19	29	51	–22	27
16.	Příbram	30	6	4	20	29	61	–32	22

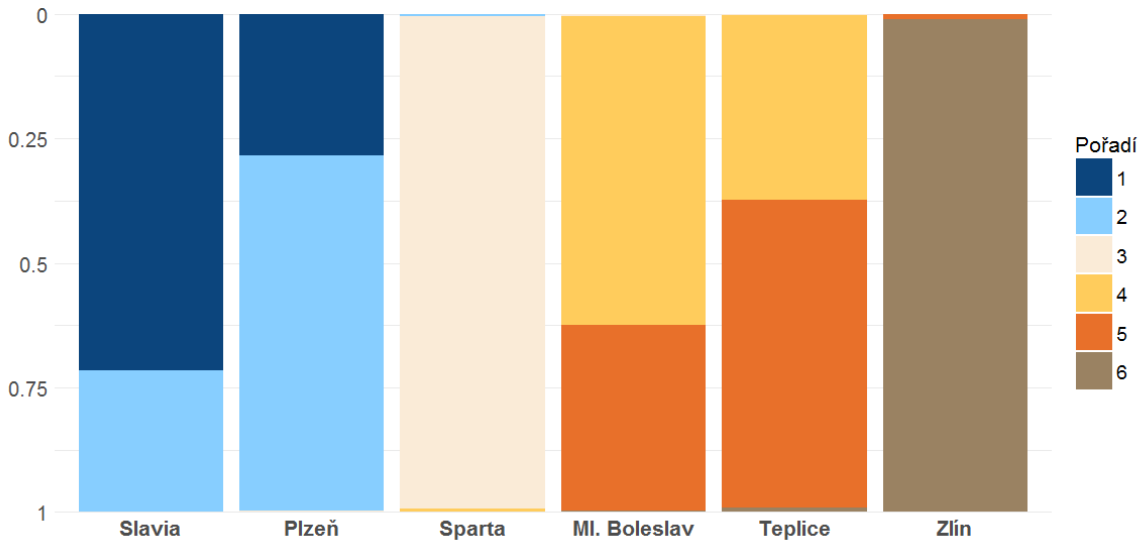
Pozn.: Z – zápasy, V – výhry, R – remízy, P – prohry, G+ – vstřelené branky, G– – inkasované branky, GR – gólový rozdíl, B – počet bodů.

7.1 Skupina o titul

O titul hraje prvních šest týmů po základní části, v tab. 7.1 tomu odpovídá zelená barva. Na obr. 7.1 je vykreslena šance týmu na každý bodový zisk po započítání simulovaných nadstavbových zápasů a odhad očekávaného bodového zisku. Je vidět, že Slavia by v průměru zvýšila bodový náskok nad Plzní o 1 bod. Sezónu zakončila s 69 body a v nadstavbové části by přidala v průměru dalších 10 bodů, Plzeň pouze bodů 9.



Obr. 7.1 Projekce bodového zisku v nadstavbové části skupiny o titul



Obr. 7.2 Odhad pravděpodobnosti umístění v nadstavbové části skupiny o titul

Na obr. 7.2 je vidět, že Slavia by v nadstavbě mohla přijít o titul na úkor Plzně. Žádný jiný tým už by se do boje o titul nezapojil. Plzeň by slavila titul zhruba ve 28,42 % případech, kdy by se za totožné situace nadstavba skutečně hrála. V tomto směru by tedy k zatraktivnění soutěže zřejmě došlo, neboť zápasy těchto dvou týmů by byly nesmírně důležité, obzvlášť by to platilo o vzájemném zápase. Za nelogický krok potom lze zřejmě považovat to, že tento vzájemný zápas je na programu 2. nadstavbového kola. Pokud by Slavia Plzeň porazila a ani v 1. kole by nezaváhala, poslední 3 kola nadstavby by už vzhledem ke slávistickému náskoku tak dramatická – z pohledu boje o titul – nebyla. Za nejatraktivnější by se dal považovat vzájemný zápas prvního s druhým v posledním kole. Je ovšem pravda, že ve 2. kole tento zápas přitáhne pozornost vždy, zatímco při jeho zařazení do posledního kola už může být rozhodnuto a požadovaný efekt by se dostavit nemusel vůbec.

Negativním přínosem nadstavby v minulé sezóně by byly zápasy Sparty a Zlína. Ani jeden tým už vlastně neměl o co hrát. Spartě by 3. místo zůstalo v 98,94 % případů a Zlín by s podobnou šancí 98,89 % zůstal na 6. místě. Nadstavba by se tak omezila na boj Slavie s Plzní o titul a Boleslavi s Teplicemi o 4. místo zaručující kvalifikační dvojutkání o EL. Mladá Boleslav by 4. místo na úkor Teplic ztratila v 37,19 % simulacích.

Zlín je výborným příkladem týmu, který ze 6. místa má takřka nulovou šanci na evropské poháry. Osmibodovou ztrátu na Mladou Boleslav a sedmibodovou na Teplice by se Zlínu podařilo smazat jen se šancí 1,12 ‰. Postup do Evropské ligy by slavil pouze v 0,26 ‰ případů. Je nutné dodat, že takto vzácné výsledky jsou ovlivněny náhodou simulace a nemusí být přesná. Podstatné sdělení, že Zlín na 6. místě je prakticky bez šance zapojit se do boje o evropské poháry, je však bez pochyby platné.

Při pohledu na poslední kolo, ve kterém Zlín porazil venku Příbram, se přímo nabízí otázka, jaké by byly šance Zlína, kdyby v tomto zápase prohrál a bojoval o poháry skrze druhou nadstavbovou skupinu. Tématu záměrného vypouštění zápasů se bude věnovat kapitola 9.

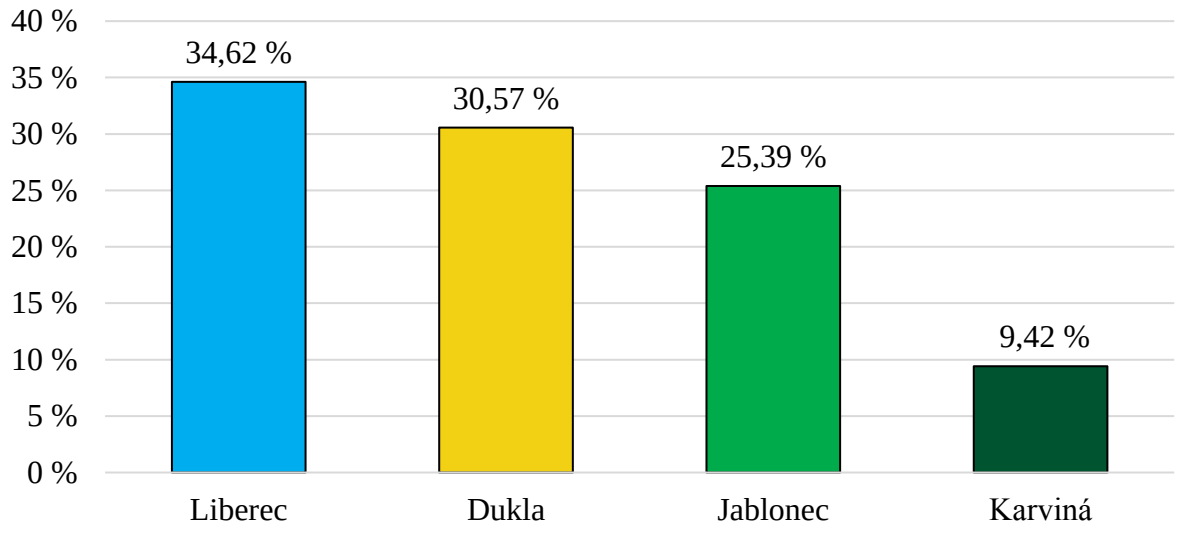
7.2 Skupina o účast v Evropské lize

Na začátek je důležité ujasnit, jak je tato skupina (a později i barážové zápasy) simulována. Jedná se o vyřazovací dvojutkání doma – venku. V nich je důležité celkové skóre a počet branek vstřelený na hřišti soupeře. Predikční model, který byl vybrán, nepředpovídá skóre, ale pouze

vítěze zápasu. O vítězi dvojutkání rozhoduje celkové skóre, kdy každý tým hraje jednou na svém stadionu a jednou na hřišti soupeře. Vzhledem k nemožnosti rozhodnout, zda je výhodnější začínat dvojutkání venku nebo doma (diskutováno v kap. 2), dojde vlastně k vyrušení výhody domácího prostředí.

Vítěze dvojutkání je tím pádem možné určit jako vítěze jediného zápasu hraného bez výhody domácího prostředí, ve kterém je ignorována remíza. Stejný přístup volí např. (Schiefler, 2015). Model je nicméně navržen tak, že předpovídá zvlášť šanci na výhru domácího a hostujícího týmu. Kromě ignorování výhody domácího prostředí, tak bude šance na postup určena jako průměr šancí na postup pro šance spočítané tak, že do rovnic (6.2) a (6.4) vstoupí oba týmy jako domácí i jako hosté.

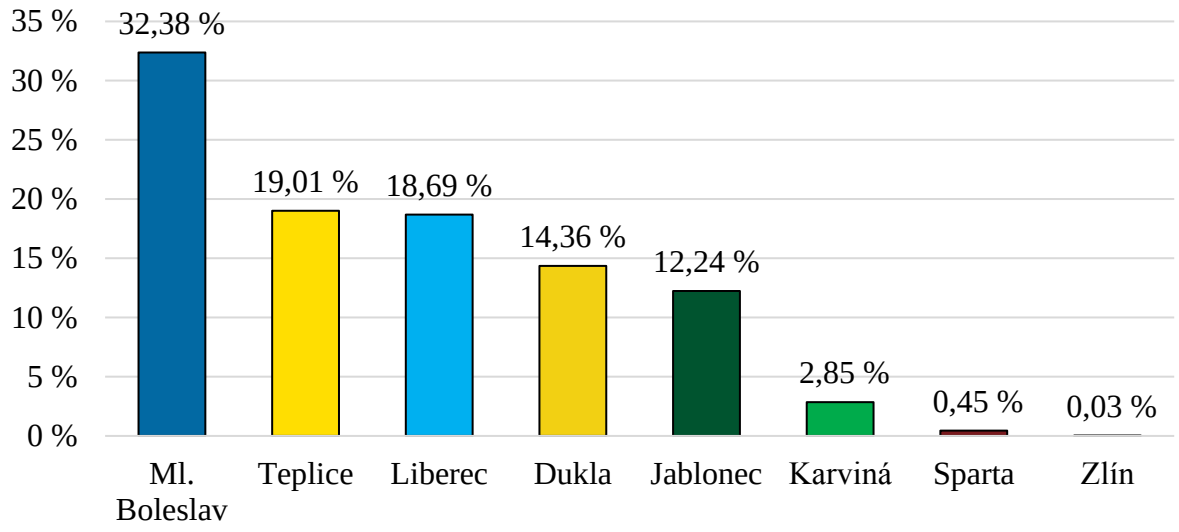
Prvním dvojutkáním by byl zápas Dukly s Karvinou se šancí na postup v poměru 68:32. Druhé dvojutkání by přineslo atraktivní podještědské derby Liberec – Jablonec s šancemi 56:44. Ve finále této skupiny by tak nejčastěji došlo k zápasu Dukla – Liberec. Šance všech týmů na vítězství ve skupině jsou na obr. 7.3.



Obr. 7.3 Šance na vítězství ve skupině o účast v EL

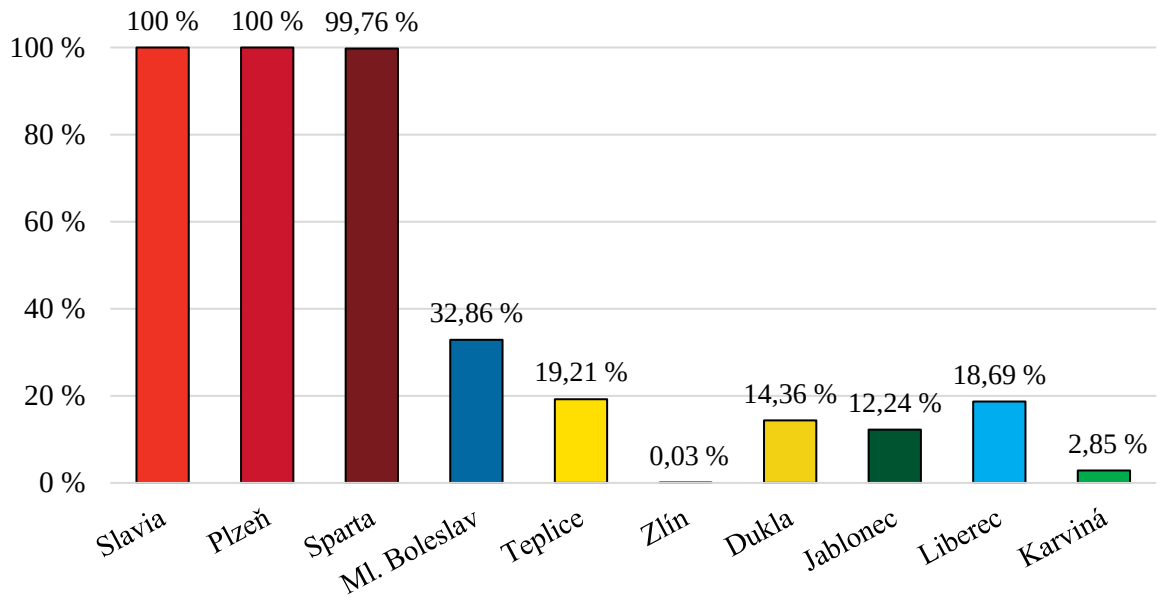
7.3 Účast v evropských pohárech

Ve skupině o titul se mohly na 4. místě umístit 4 týmy (Ml. Boleslav, Teplice, Sparta a Zlín), přičemž šance Sparty a Zlína byly skoro zanedbatelné. Šance týmů ze druhé nadstavbové skupiny jsou na obr. 7.3. Těchto 8 týmů se potenciálně mohlo dostat až do rozhodujícího dvojutkání o účast v Evropské lize. Šance probojovat se do evropského poháru právě skrz toto dvojutkání shrnuje obr. 7.4.



Obr. 7.4 Šance na postup do EL skrz nadstavbovou kvalifikaci

Na obr. 7.4 jsou kromě Sparty pouze ty týmy, pro které byla cesta kvalifikací největší nadějí zahrát si Evropskou ligu. Dále jen Mladá Boleslav a Teplice měly šanci na přímou účast (4,87 ‰ a 2,02 ‰). Sparta měla největší šanci na evropské poháry přímo umístěním na 1.–3. místě (99,31 %). Jinými slovy, Sparta měla šanci 4,49 ‰, že se ze 3. místa propadne na 4. a následně zvítězí ve dvojutkání o EL.



Obr. 7.5 Šance na účast v evropských pohárech (postupují 4 týmy)
(týmy seřazeny podle umístění v základní části)

Na obr. 7.5 jsou týmy záměrně seřazeny podle umístění v základní části, aby bylo vidět přerozdělení šancí na to nejcennější, o co se v české lize hraje – evropské poháry. Liga mistrů by připadla téměř jistě dvojici Slavia, Plzeň. Šanci 3,68 ‰ na druhé místo zaručující kvalifikaci Ligy mistrů měla ještě Sparta, nikdo jiný šanci neměl. Alespoň Evropskou ligu však Slavia s Plzní měla skutečně jistou.

První tři týmy po základní části měly i největší šanci na evropské poháry. Spíše než nastaveným herním systémem, to však bylo v důsledku jejich velkého bodového náskoku na ostatní týmy již po základní části.

Čtvrtou největší šanci na poháry měla Mladá Boleslav, která skončila na 4. místě, přesto se mohla na EL těšit jen v necelé třetině simulací, což oproti původnímu hernímu systému, kdy měla účast jistou, představuje pokles o 67 p. b. Jen pětinovou naději měly páté Teplice, za kterými byl jen s minimálním odstupem až devátý Liberec. Tomu by tak nový herní systém pomohl nejvíce. Zajímavou naději by měla také pražská Dukla a severočeský Jablonec. Karviná byla nejslabším týmem druhé nadstavbové skupiny. Na devátý Liberec po základní části ztrácela 5 bodů a výkonnostně blíže tak měla spíše ke skupině o záchranu, od které ji dělily jen 2 body. Přesto i Karviná by měla 89krát větší šanci na účast v EL než šestý Zlín, jehož bodový handicap byl ve skupině o titul prakticky nesmazatelný.

Bylo jasné, že nový herní systém pomůže týmům na 7.–10. místě, aby měly šanci dostat se do Evropské ligy. V minulé sezóně by to znamenalo, že se šance týmů, které se v dlouhodobé soutěži umístily do šestého místa sníží zhruba o 48 p. b. ve prospěch týmů s horší výkonností. Je tedy možné očekávat, že ve slabé polovině ligových ročníků hraných novým systémem přijdou týmy z horní šestky tabulky o Evropskou ligu na úkor hůře postavených týmů. Ne vždy se však dá očekávat tak velký pokles šancí předních týmů. V minulé sezóně se ve druhé skupině ocitnul velmi kvalitní Liberec, který se s předními týmy může velmi dobře měřit a šance druhé nadstavbové skupiny tak vytáhnul zřejmě o něco výše, než bude typicky běžné.

7.4 Skupina o záchranu

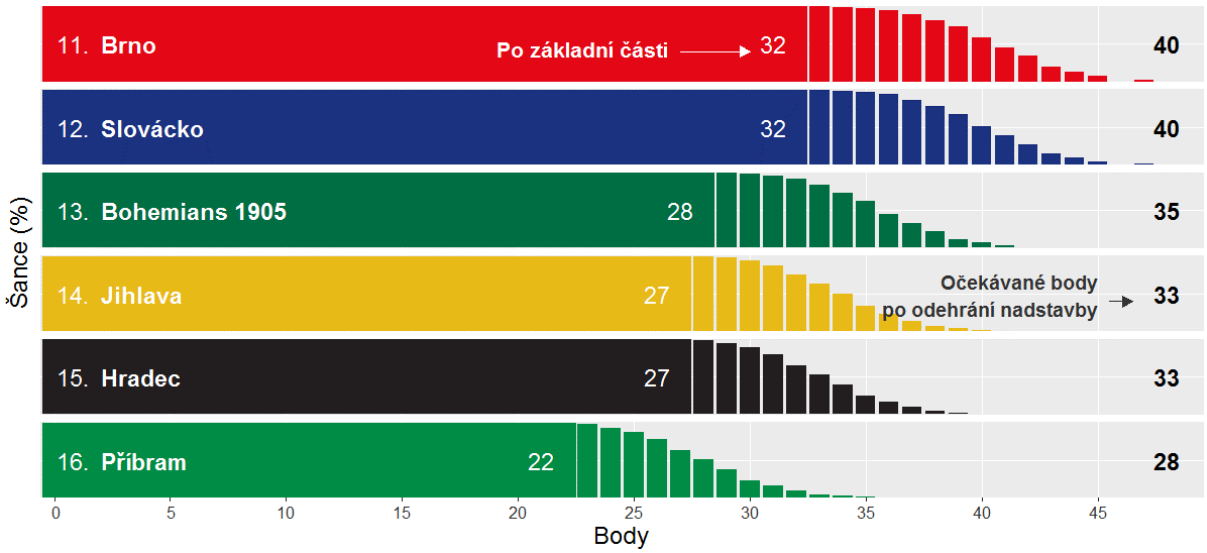
Boj o titul a evropské poháry je bezpochyby to příjemnější, co může týmy v sezóně potkat. Pro týmy druhé ligy bude stejným vrcholem sezóny boj o postup do první ligy v barážových utkáních. Jejich účastníky určí ještě třetí skupina nadstavbové části pro týmy na 11.–16. místě.

Tab. 7.2 Top 5 týmů konečné tabulky 2. ligy sezóny 2016/17

Pořadí	Klub	Z	V	R	P	G+	G–	GR	B
1.	Olomouc	30	21	6	3	59	22	37	69
2.	Ostrava	30	18	10	2	48	20	28	64
3.	Opava	30	19	6	5	61	33	28	63
4.	Vlašim	30	16	6	8	61	34	27	54
5.	Budějovice	30	12	10	8	39	31	8	46

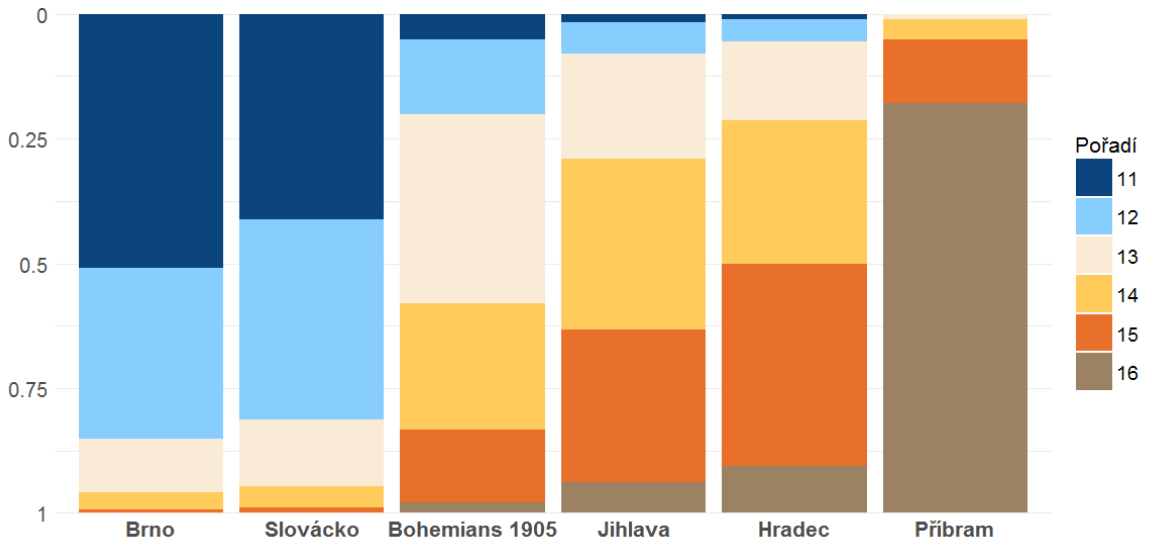
Pozn: Z – zápasy, V – výhry, R – remízy, P – prohry, G+ – vstřelené branky, G– – inkasované branky, GR – gólový rozdíl, B – počet bodů.

Jistotu postupu měla Olomouc, která skončila ve druhé lize na prvním místě. Ostravu s Opavou by čekala barážová utkání s 15., respektive 14. týmem první ligy, tedy (nejspíše) s Hradcem a Jihlavou. Jak je vidět v tab. 7.2, zmíněná trojice druholigových týmů si na čele tabulky vypracovala velký bodový náskok na ostatní týmy. Situace na chvostu prvoligové tabulky byla o něco zamotanější než v její skupině o titul, čímž se účast týmů v baráži odhadovala o něco hůře.



Obr. 7.6 Projekce bodového zisku v nadstavbové části skupiny o záchranu

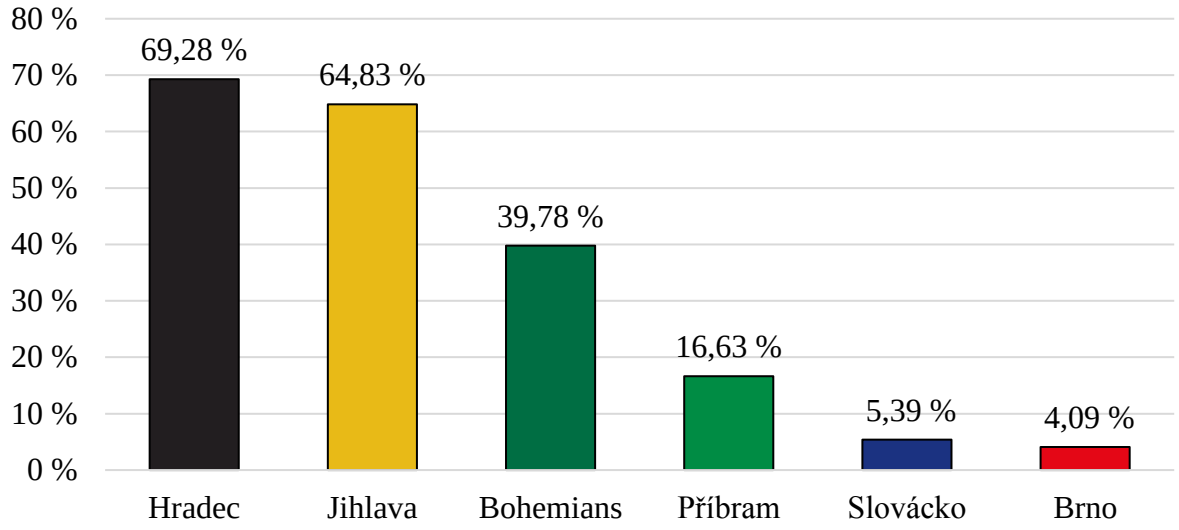
Jak bylo vidět již v tab. 7.1, a na obr. 7.6 je to zopakováno, větší ztrátu na posledním místě měla Příbram a s největší pravděpodobností by jí tak čekal přímý sestup. Rovnou do 2. ligy by šla v 82,22 % případů, Hradec by se na poslední místo propadnul v 9,45 %, Jihlava v 6,13 % a Bohemians v 2,16 %. Slovácko s Brnem přímým sestupem s šancí pod půl promile reálně ohroženy nebyly.



Obr. 7.7 Odhad pravděpodobnosti umístění v nadstavbové části skupiny o záchranu

Nejvýhodnější pozici před přidanými zápasy by měly týmy Brna a Slovácka, které měly pětibodový náskok na barážové příčky. Na obr. 7.7 jsou barážová umístění znázorněna žlutou

a oranžovou barvou. Je vidět, že největší šanci by měl Hradec s Jihlavou a Bohemians. Konkrétní čísla pro všechny týmy jsou na obr. 7.8, kde je vidět, že Slovácko s Brnem by se ani baráže příliš obávat nemusely.



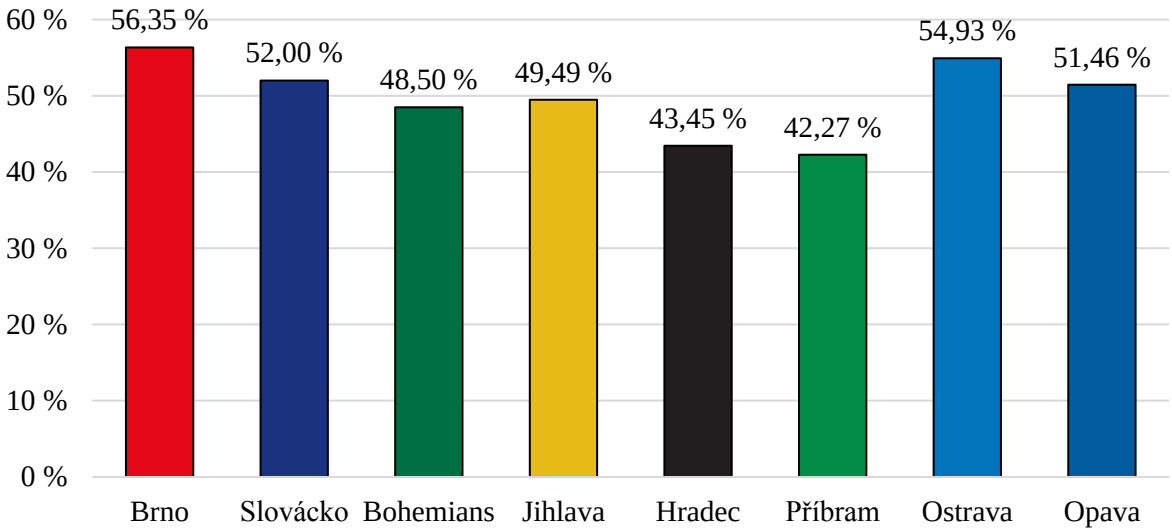
Obr. 7.8 Šance na účast v baráži (účastní se 2 týmy z 1. ligy)

7.5 Baráž

Jak bylo zmíněno v podkapitole 2.3, barážové zápasy byly pro Opavu a Znojmo tak zásadní překážkou, že hlasovaly proti návrhu nového herního systému. Měly obavu, že druholigové týmy budou mít menší šanci v barážových zápasech uspět a v průměru tak bude do první ligy postupovat méně druholigových týmů než dosavadní dva.

Účast Ostravy a Opavy v baráži by v novém systému byla jistá. Jací by byli jejich protivníci, je na obr. 7.8. Pro potvrzení nebo vyvrácení obav druholigových týmů je potřeba vyjádřit pravděpodobnost úspěchu v baráži. Ta je pro všechny zainteresované týmy shrnuta na obr. 7.9. Řazení týmů je podle dosažené pozice po základní části, kde druholigové týmy následují prvoligové.

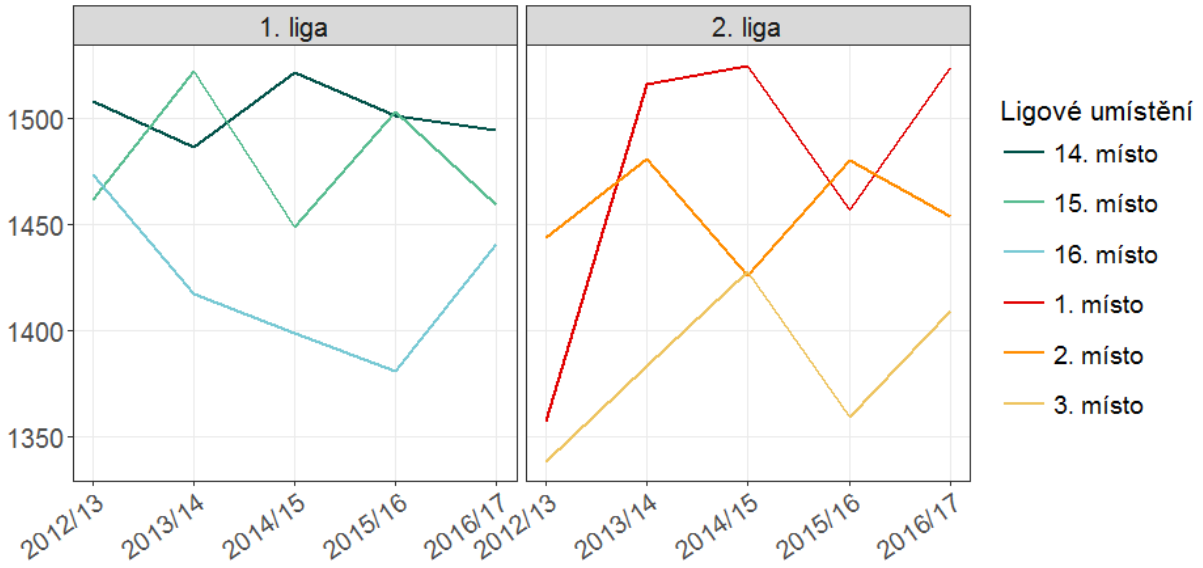
Je možné si všimnout, že Ostrava s Opavou mají více než 50 % šanci si skrze barážové utkání vybojovat účast v první lize. Úspěšnější by v baráži bylo jen Brno se Slovákem. Nejhorší vyhlídky by měla Příbram, která i kdyby se zázračně zvedla z posledního místa, v baráži by měla menší než poloviční šanci první ligu udržet.



Obr. 7.9 Šance na úspěch v baráži – dvojrovnicový model logistické regrese

7.5.1 Benchmark model

Obr. 7.9 naznačil potenciální problém. Je skutečně možné, aby týmy ze druhé ligy měly větší šanci uspět v baráži než většina týmů z první ligy, kterých se baráž může týkat? V kapitole 4 byly určeny 3 faktory ovlivňující výsledek zápasu – kvalita týmu, výhoda domácího prostředí a ostatní. Výhoda domácího prostředí není v barážových zápasech uvažována, hlavním faktorem určující vítěze zápasu by tak měla být kvalita týmu. Ta je v této práci měřena Elo ratingem.



Obr. 7.10 Elo hodnocení kvality týmů na konci sezóny

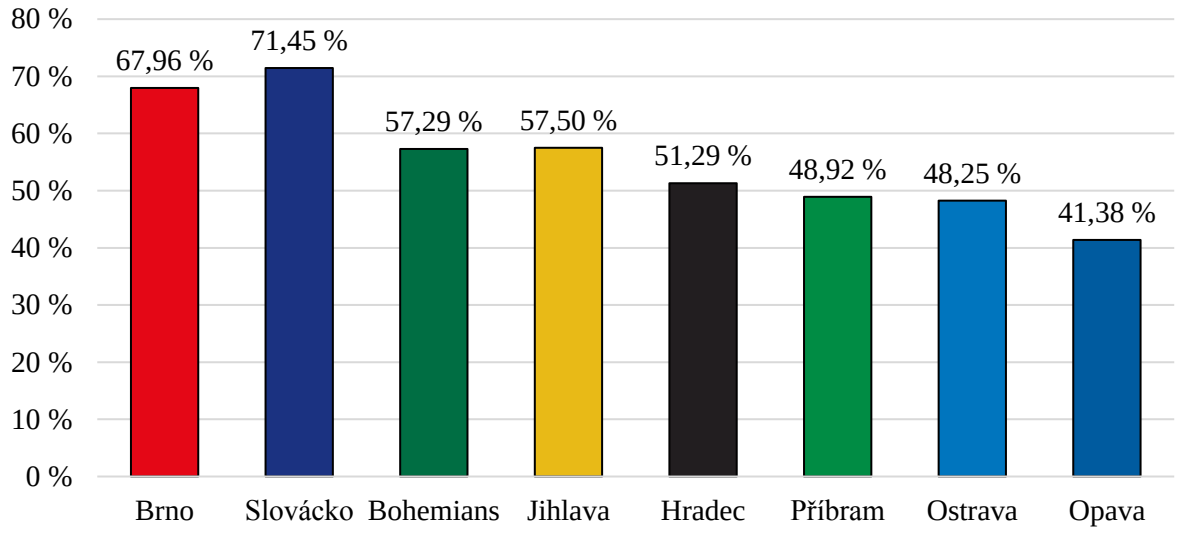
Na obr. 7.10 je vidět, že Olomouc jako vítěz druhé ligy byla výkonnostně lepší než chvost první ligy. Tento jev se v historii nezřídka opakuje. Pro vyhodnocení barážových utkání je nicméně zapotřebí porovnávat zelenou křivku 15. místa první ligy s oranžovou křivkou 2. místa druhé ligy. Ve všech zobrazených sezónách byla kvalita druholigového týmu o něco menší než kvalita potenciálního prvoligového soupeře. V minulé sezóně byl tento rozdíl nejtěsnější, v zápase Ostravy (2. místo druhé ligy v sezóně 2016/17) s Hradcem (15. místo první ligy v sezóně 2016/17) by se zřejmě favorit laickým pohledem hledal složitě.

Porovnání druhé barážové dvojice už tak optimisticky z pohledu druholigových týmů nevypadá. Tmavě modrá křivka 14. místa v první lize je od žlutého 3. místa ve druhé lize už poměrně vzdálená. Pro minulou sezónu graf porovnává Jihlavu s Opavou, rozdíl je necelých 100 Elo bodů.

Vysvětlení vysokých šancí Ostravy a Opavy z obr. 7.9 je nutné hledat přímo v modelu, který je pro předpovídání používán. Do rovnice (6.2), používané pro předpověď pravděpodobnosti výhry domácího týmu, vstupují proměnné *L4golyssp*, *L8golysp*, *L4golys*. Jedná se o počty vstřelených branek, které nejsou váženy kvalitou týmu, kterému byl gól buď vstřelen, nebo od kterého byl inkasován. V rovnici (6.4), používané pro předpověď pravděpodobnosti výhry hostujícího týmu, jsou proměnné *L8golysp*, *L8golys*, *L4golyssp* se stejnou vlastností. Týmy z čela druhé ligy v lize dominují, střílí hodně branek a málo jich inkasují. Prvoligové týmy

na konci tabulky to mají přesně opačně – v lize nejsou úspěšné, často inkasují a skórují spíše sporadicky.

Přestože v rovnicích (6.2) a (6.4) jsou i proměnné zohledňující kvalitu týmů, výše zmíněné proměnné nadhodnocují šanci týmů z druhé ligy. Jedním řešením může být použití benchmarkového modelu. Ten vychází pouze z kvality týmů a jeho predikční schopnosti nejsou špatné. V tab. 6.7 se ve srovnání sice umístil až na 9. místě, absolutní odchylky vyhodnocovacích kritérií od vítězného modelu však nejsou velké.



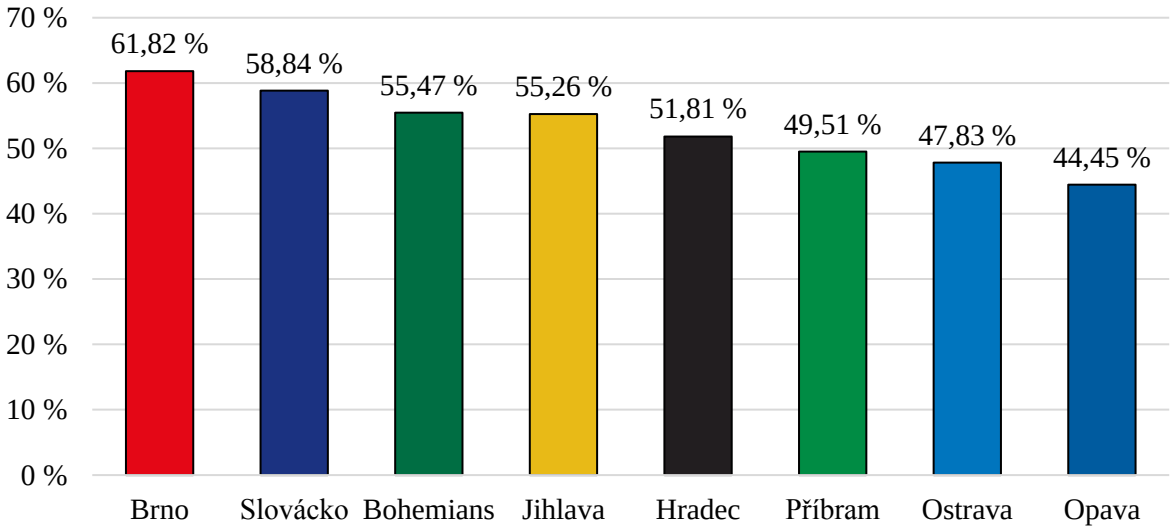
Obr. 7.11 Šance na úspěch v baráži – benchmark model

Obr. 7.11 ukazuje šance na úspěch v baráži podle benchmarkového modelu. Druholigové týmy mají v souladu s obr. 7.10 menší než 50 % šanci na úspěch. To znamená, že si skutečně pohorší. Namísto 2 týmů by za stejných podmínek jako v minulé sezóně postoupilo v průměru 1,90 druholigového týmu (jistě postupující Olomouc, Ostrava s šancí 48,25 % a Opava s 41,38 %, tím $1 + 0{,}4825 + 0{,}4138 = 1{,}8963$). Srozumitelněji lze ekvivalentně tvrdit, že v desetiletém horizontu postoupí do první ligy v průměru o jeden tým méně než doposavad.

Jeden tým za 10 let zní jako rozumná cena za 20 velmi atraktivních dvojzápasů, které tato změna za stejnou dobu přinese. Jak bylo zmíněno ve 2. kapitole, nový herní systém se dotkne druhé ligy i zvýšením financí o 10 milionů Kč na sezónu. I to lze považovat za jistou kompenzaci této sportovní ztráty.

7.5.2 Upravený model logistické regrese

Jinou možností, jak se vypořádat s nedostatkem vybraného modelu dvojrovnicové logistické regrese, je úprava vstupů modelu. Nastavením hodnot proměnných, které nezohledňují kvalitu týmů, na nulu, dojde k vypuštění těchto proměnných z modelu a výsledek jimi nebude zkreslen. Oproti benchmarkovému modelu bude tento model o poznání bohatší, díky rovnici (6.4) a proměnné *L4ziskElo* bude například zohledňovat i aktuální formu týmů.



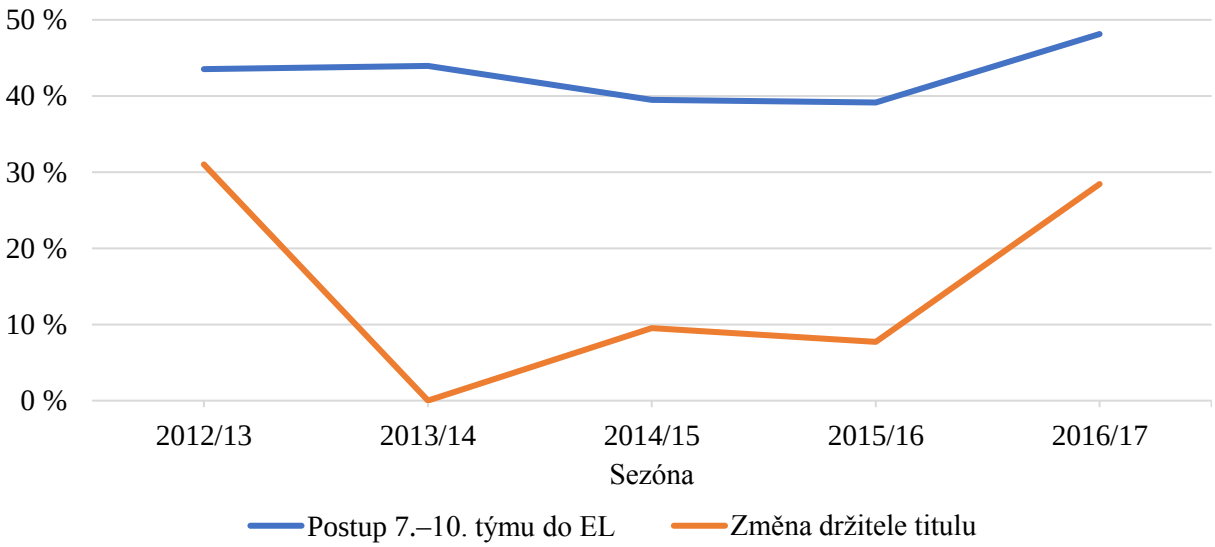
Obr. 7.12 Šance na úspěch v baráži – upravený dvojrovnicový model logistické regrese

Vývoj šancí týmů seřazených podle pořadí v základní části na obr. 7.12 je z prezentovaných variant výpočtu barážových utkání intuitivně nejlogičtější. Opavě dokonce přisuzuje o 3 p. b. větší šanci na postup do 1. ligy než benchmarkový model. Celkem se tak šance druholigových týmů oproti výsledkům prezentovaných na obr. 7.11 ještě o něco zvýší.

Druholigové týmy mají velmi slušnou naději se do první ligy probojovat. Zhoršení jejich pozice, měřené průměrným počtem postupujících týmů, je jen velmi malé. Pokles z 2,00 na 1,92 představuje průměrnou ztrátu 0,08 postupujícího týmu na sezónu. Baráž tak bude pro prvoligová mužstva strašákem. Jak je vidět na obr. 7.12, ani týmy na 11. (Brno) a 12. místě (Slovácko) by v baráži výrazně nedominovaly. Zařazení barážových utkání lze považovat za marketingově podařený tah bez větších negativních sportovních dopadů.

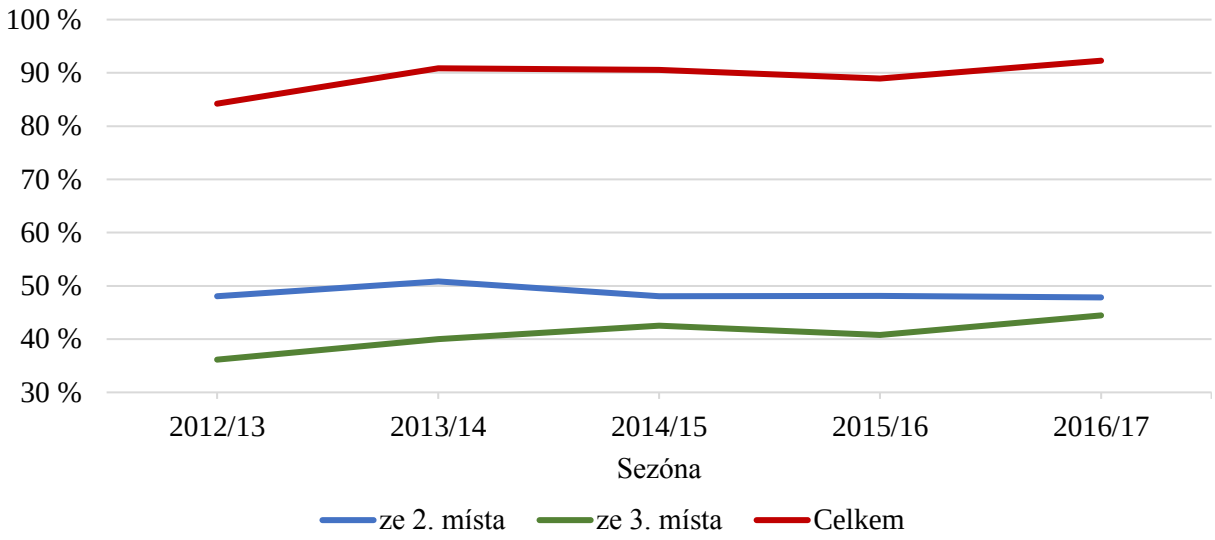
8 Důležité hodnoty z posledních 5 sezón

Všechny důležité ukazatele již byly představeny a okomentovány v předešlé kapitole pro sezónu 2016/17. Nadstavbové zápasy mohou změnit držitele mistrovského titulu, dávají naději týmům na horších pozicích na postup do Evropské ligy a až 3 týmy mohou do 1. ligy postoupit z nižší soutěže. V této kapitole je vždy uvažováno, že se hraje podle schématu na obr. 2.1, včetně rozdělení pozic evropských pohárů. To i přesto, že se počty zástupců kvalifikujících se skrze ligovou soutěž mohly lišit. Důvodem je především srovnatelnost výsledků s aktuální situací.



Obr. 8.1 Šance na postup týmu ze 2. nadstavbové skupiny do EL a na změnu držitele titulu

Na obr. 8.1 je vidět, že šance týmů na 7.–10. místě na vybojování účasti v Evropské lize je celkem stabilně mezi 39 % a 44 %. Minulá sezóna s 48 % byla spíše výjimkou díky silnému Liberci, který skončil na 9. místě. Zatímco o účasti v EL rozhoduje pouze jeden dvojzápas, o titulu rozhodne dalších 5 nadstavbových zápasů. Vybuduje-li si tak tým na 1. místě dostatečný bodový náskok, nemusí mít v baráži příliš velké starosti svoji pozici obhájit. Jasným důkazem je sezóna 2013/14, ve které si Sparta na prvním místě vypracovala náskok 13 bodů před Plzní.



Obr. 8.2 Šance druholigových týmů na úspěch v baráži

O baráži bylo již poměrně dlouze diskutováno, na obr. 8.2 jsou shrnuty šance druholigových týmů na postup do 1. ligy v posledních pěti sezónách. Pouze Hradec v sezóně 2013/14 by měl s 50,84 % nadpoloviční šanci na úspěch. Celkově jsou ale šance v čase poměrně vyrovnané a v průměru postoupí 1,89 týmu za sezónu. Dříve vyslovené závěry o povedeném marketingovém tahu s minimálním sportovním dopadem tak zůstávají v platnosti.

9 Záměrná prohra jako cesta k úspěchu

Jedním z cílů práce bylo zjistit, zda někdy může nastat situace, kdy se týmu více vyplatí zápas prohrát než vyhrát. Vlastně už v 7. kapitole (konkrétně na obr. 7.5) byla na tuto otázku nalezena odpověď. Je zde jasně vidět, že Zlín v minulé sezóně neměl z 6. místa reálnou naději do Evropské ligy postoupit. Průměrný tým druhé nadstavbové skupiny měl šanci 376krát větší. V posledním kole hrál Zlín na půdě poslední Příbrami a zvítězil 2:0. Kdyby prohrál, měl by v konečném účtování 38 bodů a skončil by na 9. místě.

Propad až na 9. místo by byl způsoben vyrovnaností tabulky (viz tab. 7.1). Svůj poslední zápas vyhrály i týmy Dukly, Jablonce a Liberce. Kterémukoli z těchto týmů mohla zůstat nevýhodná pozice v podobě 6. místa, stačilo jen, aby tým zvítězil a ostatní nikoliv. Všechny týmy byly v situaci, kdy jim výhra mohla více ublížit než pomoci. Je tedy možné, že v novém herním formátu by se v posledním kole ve 4 zápasech z 8 snažily týmy prohrát.

Aby bylo možné identifikovat zápasy, ve kterých se jednomu nebo oběma týmů více vyplatí prohrát, musela tomu být upravena i simulace. Pro každý zápas byl simulován vývoj soutěže v případě výhry domácího týmu, remízy, výhry hostujícího týmu a za předpokladu výsledku podle predikčního modelu. Zohledněna byla i časová posloupnost zápasů. Zápasy se stejným časem výkopu byly simulovány najednou, aby nedošlo ke zkreslení výsledků. Takto formulovanou simulaci je velmi obtížné výpočetně optimalizovat, jelikož se pro každý požadovaný zápas skutečně musí udělat projekce budoucího vývoje ligy pro 4 různé výsledky.

Z důvodu rozumné časové náročnosti je simulace počítána až od 16 ligového kola. Spekulovat o záměrných prohrách v první polovině soutěže stejně nemá velký význam, jelikož si týmy teprve svoji pozici budují. Počet simulací pro každý zápas a každý uvažovaný výsledek byl pro zbytek soutěže 2 000. Pro 5 sezón po 120 zápasech a 4 možných výsledcích to dělá 4 800 000 simulovaných průběhů sezóny s časovou náročností simulace blížící se 40 hodinám.

Tab. 9.1 Počet zápasů po sezónách, ve kterých by bylo alespoň pro jeden tým výhodnější prohrát

Sezóna	Celkem		Domácí		Hosté	
	všechny	>= 20	všechny	>= 20	všechny	>= 20
2012/13	10	6	4	2	6	4
2013/14	13	6	5	3	8	3
2014/15	6	2	4	2	2	0
2015/16	24	10	11	5	13	5
2016/17	21	11	9	4	12	7

Pozn.: Jako „>= 20“ jsou označeny zápasy, ve kterých byl rozdíl mezi počtem účastí v EL při prohře a výhře alespoň 20 z 2 000 simulací (1 p. b.)

Z tab. 9.1 je zjevné, že zápasy, ve kterých se vyplatí alespoň jednomu týmu prohrát, nejsou úplnou raritou. Bez omezujících podmínek, které by takový zápas měl splňovat, se v každé z posledních dvou sezón odehrálo přes 20, což jsou vlastně téměř 3 kompletní ligová kola. Existuje ale několik objektivních důvodů, proč zvažovat, zda se opravdu jedná o zápas, který se týmu prokazatelně vyplatí prohrát. Zaprvé je to bezpochyby chyba plynoucí z podstaty simulačního výpočtu. Pokud výsledek zápasu šance týmu ovlivní jen minimálně, může se vlivem náhody stát, že bude označen jako vhodný k záměrné prohře.

Druhým důvodem je vyřazení čistě spekulativních zápasů. Například zápas Karviné s Jabloncem v posledním kole minulé sezóny by se vyplatil prohrát oběma týmům. (Takové zápasy, které by se vyplatily prohrát oběma týmům, byly v simulaci nalezeny 4. S omezující podmínkou žádný.) Při bližším pohledu, proč by tomu tak bylo, si lze povšimnout, že tým, který by zápas prohrál, by se vzhledem k postavení v tabulce vyhnul v 1. kole nadstavbového dvojzápasu ve skupině o účast v Evropské lize silnému Liberci. Nelze popřít, že i takto budou mít možnost týmy přemýšlet a na hřišti se podle toho chovat. Mým cílem je ale zaměřit se především na zápasy, ve kterých je výhoda záměrné prohry nepopiratelná a pramenící zejména z umístění na hraně mezi 1. a 2. nadstavbovou skupinou.

Tab. 9.2 Počet zápasů po kolech, ve kterých by bylo alespoň pro jeden tým výhodnější prohrát

Kolo	Celkem		Domácí		Hosté	
	všechny	>= 20	všechny	>= 20	všechny	>= 20
19–23	19	6	9	3	10	3
24	8	3	2	1	6	2
25	4	1	1	0	3	1
26	7	2	3	1	4	1
27	6	3	2	1	4	2
28	11	6	7	4	4	2
29	8	7	5	4	3	3
30	11	7	4	2	7	5

Pozn.: Jako „>= 20“ jsou označeny zápasy, ve kterých byl rozdíl mezi počtem účastí v EL při prohře a výhře alespoň 20 z 2 000 simulací (1 p. b.)

Dalo se očekávat, že zkoumaných zápasů bude s koncem sezóny přibývat. Prvních 15 ligových kol nebylo vůbec simulováno, což se ukázalo jako správné rozhodnutí. První podezřelý zápas byl vždy zaznamenán nejdříve v 19. kole. Kromě sezóny 2015/16 by se alespoň jeden takový zápas odehrál v posledním 30. kole. Situace minulé sezóny už diskutována byla, prohra by se skutečně vyplatila čtveřici Zlín, Dukla, Jablonec a Liberec.

Tab. 9.3 Šance na účast v evropských pohárech podle výsledku zápasu 30. kola

Tým	Prohra	Výhra	Zvýšení šance		Remíza	Model
			absolutně	relativně		
Zlín	2,60 %	0,05 %	2,55 p. b.	5 100,00 %	1,55 %	1,50 %
Dukla	12,70 %	3,00 %	9,70 p. b.	323,33 %	12,40 %	11,30 %
Jablonec	15,00 %	8,45 %	6,55 p. b.	77,51 %	13,55 %	13,10 %
Liberec	22,25 %	11,40 %	10,85 p. b.	95,18 %	21,20 %	17,00 %

Pozn.: Absolutní zvýšení šance je rozdílem šancí při prohře a výhře. Relativní zvýšení je podíl těchto hodnot vyjádřený v procentech.

K výsledkům v tab. 9.3 je dobré zopakovat, že simulace každého zápasu proběhla nezávisle na výsledcích zápasů ostatních, jelikož výkop všech zápasů posledního kola je ve stejný čas. V simulaci tak o výsledcích všech ostatních zápasů kromě zkoumaného zápasu rozhodovala náhoda založená na popsaném prediktivním modelu. O postavení v tabulce tím pádem vždy spekuluje pouze jeden tým, ostatní hrají naplno podle svých modelovaných schopností.

Kromě zkoumaného rozdílu mezi výhrou a prohrou zápasu je dobré se podívat i na poslední sloupec tabulky, kde je šance týmu za předpokladu, že všechny týmy hrají naplno bez spekulací o výsledku. To je nicméně myšlení, které lidský mozek jen velmi špatně chápe. Hráči před vyběhnutím na hřiště budou srovnávat možnost prohry a výhry a podle toho se budou na hřišti chovat. Všechny týmy svoje zápasy vyhrály, porazily Příbram, Teplice, Karvinou a Mladou Boleslav (popořadě jako v tabulce). Hrálo-li by se podle nového formátu, pak by takovéto výsledky zřejmě nečekal ani největší sportovní optimista.

Příloha 4 nabízí přehled zápasů, ve kterých by se vyplatilo prohrát domácímu týmu a příloha 5 ty zápasy, ve kterých totéž platí pro hosty. Z důvodu zkrácení seznamu je zde aplikována stejná podmínka jako např. v tab. 9.1 a zobrazeny jsou jen ty zápasy, ve kterých tým prohrou zvýší svoji šanci na evropské poháry alespoň o 1 p. b.

10 Závěr

Český fotbal čeká po pětadvaceti letech velká změna herního systému. Diskuze byly dlouhé. Zřejmě i proto, že nebyly podložené průkaznou analýzou podobné takové, jakou nabízí tato diplomová práce.

V jejím úvodu je nový herní systém popsán. V další části byl optimalizován Elo rating system pro výpočet ukazatele kvality týmu, který byl zaveden jako hlavní faktor ovlivňující výsledek fotbalového zápasu. Pomocí prediktorů odvozených od kvality týmů a jejich gólové potence byly zkonstruovány a pečlivě otestovány predikční modely potenciálně použitelné pro simulaci soutěže podle nového formátu. Mezi benchmarkovým modelem stojícím pouze na Elo kvalitě týmů, logistickou regresí, Poissonovou regresí a vícenásobnou lineární regresí vždy v několika variacích se jako nejlepší ukázal model logistické regrese o dvou rovnicích – jedné pro výpočet pravděpodobnosti výhry domácího týmu a druhé pro výpočet pravděpodobnosti výhry hostujícího týmu. Pravděpodobnost remízy se určila jako doplněk do jedné.

Vedlejším produktem této práce je zjištění, že střední absolutní odchylka není vhodná jako kritérium vyhodnocení predikčních schopností modelu předpovídajícího výsledky fotbalových utkání. Při optimalizaci modelu na jejím základě dochází k naprostému ignorování remízových výsledků.

Vybraný regresní model popsaný v podkapitole 6.2 byl použit pro simulaci soutěže podle nového systému. Pro predikci nově zavedených barážových utkání byl z důvodu zahrnutí nevhodných proměnných upraven do podoby okomentované v podkapitole 7.5.2.

Právě barážová utkání mohla projekt nového herního systému úplně zablokovat, neboť druholigové týmy měly obavy o prostupnost mezi 1. a 2. ligou. Obávaly se menší šance se do první ligy probojovat. V čele odpůrců stály zástupci Opavy a Znojma, kteří z tohoto důvodu hlasovali proti zavedení nového herního formátu. Pomocí simulace byla v 7. a 8. kapitole tato jejich obava kvantifikována. Zatímco ve starém systému postupovaly vždy 2 týmy, v tom novém to bude v průměru 1,89 týmu. Ke snížení tedy dojde o zhruba 1 postupující tým v horizontu 10 let.

Opava se Znojmem by preferovaly zachování dvou přímých postupujících s jedním barážovým zápasem mezi 3. týmem 2. ligy a 14. týmem 1. ligy. To by znamenalo v průměru asi 2,4 postupujících týmů za sezónu. Ze dvou zvažovaných alternativ tak menší změnu přináší přijatý

návrh. Snížení nadějí na sportovní úspěch je navíc doprovázeno kompenzací v podobě navýšení příjmů druholigových týmů. Zvolená forma nadstavbových utkání o záchranu a následné barážové zápasy lze v tomto kontextu považovat za velmi povedený tah, který nejspíše přispěje k atraktivitě nejen 1., ale i 2. ligy.

Jako pádnější argument proti přijetí návrhu na nový herní systém se ukazují hlasy obyčejných fanoušků, kteří poukazují na jistou nespravedlnost 2. nadstavbové skupiny, která dává i 10. týmu po základní části šanci na účast v Evropské lize. V té se hraje o prestiž, finanční bonusy i národní koeficient, podle kterého jsou jednotlivým národním asociacím přidělována místa v Lize mistrů a Evropské lize.

Kapitola 9, příloha 4 a příloha 5 dokládají, že tato obava je naprosto opodstatněná. V minulé sezóně by se v posledním kole hrály dokonce 4 zápasy, ve kterých by se jednomu týmu více vyplatilo zápas prohrát než vyhrát.

Diskuze o těchto zápasech budou bouřlivé a k zápasům a lize to pozornost bezpochyby přitáhne. Otázkou je, kdo vyrazí na stadion, na kterém bude hrát tým, který bude usilovat o prohru. Může dojít i k situaci, kdy tým záměrně inkasuje v posledních minutách a jeho vlastní fanoušci se budou radovat. O tom, zda se na něco takového bude mít nezaujatý – například televizní – divák chuť dívat, lze s úspěchem pochybovat. Například Zlín by svoji šanci na Evropskou ligu prohrou v posledním kole minulé sezóny zvýšil o 5 100 %, Dukla o více než 300 %, to je velmi slušná motivace, když ne k záměrnému inkasování, tak alespoň k vypuštění důležitých soubojů, které k brance soupeře stejně povedou.

Kdyby se druhá nadstavbová skupina nehrála vyřazovacím systémem, ale stejně jako ostatní dvě skupiny každý s každým, byl by efekt podobný. Zájem na prohře by mělo méně týmů, o to by však byl silnější. Navíc by nedošlo k takovému úbytku nevýznamných zápasů, jelikož týmy ze spodní části této skupiny by nemusely mít velkou šanci se posunout na 7. místo. Byly by vlastně v situaci analogické k situaci Zlína v minulé sezóně, jen o skupinu níže.

Nový systém bezpochyby přinese zajímavé zápasy. Druhá skupina však může představovat časovanou bombu, která může v některých sezónách (jako byla třeba právě ta minulá) přinést na české fotbalové stadiony spíše fotbalovou frašku než kvalitní kopanou. Pro přežití a usazení nového systému bude důležité, aby se to v několika následujících sezónách nestalo. Bude důležité, aby z 2. ligy postoupil více než jeden tým, a aby nenastala situace na pomezí prvních

dvou nadstavbových skupin, která by bývala nastala v minulé sezóně. Stane-li se tak, je možné, že hlasy proti nastolenému systému zesílí natolik, že se bude velmi brzy znovu měnit.

Zejména z důvodu omezeného prostoru se tato práce omezila na diskuzi sporných částí nového herního systému. Jen s minimálním úsilím by se dal formulovaný model a naprogramované simulace upravit pro další komplexní analýzy, které by mohly pomoci, když ne přímo vybrat optimální herní systém, tak alespoň zástupcům jednotlivých týmů při rozhodování, zda nový herní systém přijmout nebo nikoliv. Je s podivem, že žádná podobná analýza zpracována nebyla. Dokud se o důležitých věcech bude v českém fotbale i nadále rozhodovat pocitově bez reálných podkladů, k jeho zkvalitnění ani nový herní systém příliš nepomůže.

Bibliografie

11tegen11. (2017). *About 11tegen11…*. Získáno 22. Srpna 2017, z 11tegen11 - Tactical analysis of Dutch football...: http://11tegen11.net/about/

Albert, J., & Koning, R. H. (2007). *Statistical thinking in sports*. Boca Raton, Florida, Spojené státy americké: Chapman.

BBC News. (2015). *Premier League in record £5.14bn TV rights deal*. Získáno 21. Srpna 2017, z BBC News: http://www.bbc.com/news/business-31379128

Berry, M. J., & Linoff, G. (2004). *Data mining techniques* (2nd ed.. vyd.). Indianapolis, Ind.: Wiley Pub.

Constantinou, A. C., & Fenton, N. E. (2012). Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. *Journal of Quantitative Analysis in Sports*(issue 1). doi:10.1515/1559-0410.1418

Corke, T. (2017). *Tipping Accuracy vs MAE as a Footy Forecaster Metric*. Získáno 23. Srpna 2017, z MatterOfStats: http://www.matterofstats.com/mafl-stats-journal/2017/6/9/tipping-accuracy-vs-mae-as-a-footy-forecaster-metric

Cortis, D. (2015). *How to use standard deviation for betting*. Získáno 19. Března 2018, z Pinnacle: https://www.pinnacle.com/en/betting-articles/Betting-Strategy/how-to-use-standard-deviation-for-betting/P8724GE57FBZWD3F#height

Cortis, D. (2018). *Inflating or deflating the chance of a draw in soccer*. Získáno 13. Března 2018, z Pinnacle: https://www.pinnacle.com/en/betting-articles/Soccer/inflating-or-deflating-the-chance-of-a-draw-in-soccer/CGE2JP2SDKV3A9R5

Cronin, B. (2017). *Poisson Distribution: Predict the score in soccer betting*. Získáno 13. Března 2018, z Pinnacle: https://www.pinnacle.com/en/betting-articles/Soccer/how-to-calculate-poisson-distribution/MD62MLXUMKMXZ6A8

CS Fotbal. (2017). *Přehled sezón*. Získáno 22. Srpna 2017, z CS Fotbal: http://www.csfotbal.cz/druha-liga

ČTK České noviny. (2017). *Fotbalová liga bude mít od sezony 2018/19 nadstavbovou část*. Získáno 22. Srpna 2017, z ČTK České noviny: http://www.ceskenoviny.cz/zpravy/fotbalova-liga-bude-mit-od-sezony-2018-19-nadstavbovou-cast/1471898

ČTK České noviny. (2017). *Příjmy v první fotbalové lize dosáhly 1,6 procenta Premier League*. Získáno 21. Srpna 2017, z ČTK České noviny: http://www.ceskenoviny.cz/zpravy/prijmy-v-prvni-fotbalove-lize-dosahly-1-6-procenta-premier-league/1505800

Daróczi, G. (2015). *Mastering Data Analysis with R*. Birmingham: Packt Publishing.

Dixon, M. J., & Coles, S. G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Applied Statistics*, stránky 265-280. Získáno 24. Října 2017, z http://www.math.ku.dk/~rolf/teaching/thesis/DixonColes.pdf

Esteva, R. (2015). *History favours home sides for second legs*. Získáno 22. Srpna 2017, z UEFA: http://www.uefa.com/uefachampionsleague/news/newsid=2219279.html

Glickman, M. (2017). *About Mark Glickman*. Získáno 21. Září 2017, z Mark Glickman's World: http://www.glicko.net/bio.html

Hebák, P. (2015). *Statistické myšlení a nástroje analýzy dat*. Praha: Informatorium.

HET Liga. (2017). *Historie ligy*. Získáno 22. Srpna 2017, z HET Liga: http://www.hetliga.cz/historie-landing

Chris. (2010). *Champions League ties – home or away first?* Získáno 22. Srpna 2017, z ZonalMarking - Football Tactics: http://www.zonalmarking.net/2010/03/19/champions-league-away-goals-home-advantage/

iDNES.cz a ČTK. (2017). *Opavě a Znojmu se nelíbí prostupnost mezi první a druhou ligou*. Získáno 22. Srpna 2017, z iDNES Sport: http://fotbal.idnes.cz/opava-znojmo-vyhrady-system-ligy-dlg-/fotbal.aspx?c=A170410_170820_fotbal_ten

International-Harvard Statistical Consulting Company. (2017). *Package 'My.stepwise'*. Získáno 13. Března 2018, z r-project: https://CRAN.R-project.org/package=My.stepwise

iSport. (2017). *Šéf ligy o změnách: Sezonu natáhneme do léta, kluby dostanou víc peněz*. Získáno 22. Srpna 2017, z iSport: http://isport.blesk.cz/clanek/fotbal-1-liga-rocnik-2016-17/299556/sef-ligy-o-zmenach-sezonu-natahneme-do-leta-kluby-dostanou-vic-penez.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer. Získáno 18. Září 2017, z http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf

Kaloumenos, M. (2012). *A brief comparison guide between the ELO and the Glicko rating systems*. Získáno 21. Září 2017, z English Chess: http://www.englishchess.org.uk/wp-content/uploads/2012/04/Elo_vs_Glicko.pdf

Karlis, D., & Ntzoufras, I. (2005). Bivariate Poisson and Diagonal Inflated Bivariate. *Journal of Statistical Software*. Získáno 13. Března 2018, z https://tolstoy.newcastle.edu.au/R/e8/help/att-6544/karlisntzuofras05.pdf

Kirill. (2017). *The World Football Elo Rating System*. Získáno 22. Září 2017, z EloRatings: http://eloratings.net/system.html

Lasek, J. (2012). *Football team rankings*. Master's thesis in Mathematics, VU University Amsterdam, Faculty of Sciences, Amsterdam. Získáno 21. Září 2017, z http://lasek.rexamine.com/master.pdf

LFA. (2017). *Herní formát 1. ligy od sezóny 2018/19*. Získáno 22. Srpna 2017, z LFA: http://lfafotbal.cz/upload/file/Hern%C3%AD%20form%C3%A1t%20I_%20ligy%20-final+.pdf

Marek, L. (2012). *Pravděpodobnost*. Praha: Professional Publishing.

Pecáková, I. (2011). *Statistika v terénních průzkumech*. Praha: Professional Publishing.

Pohlkamp, S. (2014). *Are football referees really neutral (or do they have prejudices)?* Hamburg. Získáno 31. Října 2015, z https://www.wiso.uni-hamburg.de/fileadmin/vwl/industrieoekonomik/Team/Steffi/prejudice.pdf

Polamuri, S. (2017). *Difference Between Softmax Function and Sigmoid Function*. Získáno 8. Března 2018, z Dataaspirant: http://dataaspirant.com/2017/03/07/difference-between-softmax-function-and-sigmoid-function/

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ross, D. (2007). *Arpad Elo and the Elo Rating System*. Získáno 21. Září 2017, z ChessBase: http://en.chessbase.com/post/arpad-elo-and-the-elo-rating-system

Řezanková, H., & Löster, T. (2009). *Úvod do statistiky*. Praha: Oeconomica.

Schiefler, L. (2015). *The system or how this works*. Získáno 26. Března 2018, z ClubElo: http://clubelo.com/System

SK Slavia Praha. (2017). *Slavia k hlasování o novém herním systému*. Získáno 22. Srpna 2017, z SK Slavia Praha Fotbal: http://slavia.cz/clanek.asp?id=Slavia-k-hlasovani-o-novem-hernim-systemu-14576

Stekler, H. O., Sendor, D., & Verlander, R. (2009). *Issues in Sports Forecasting*. Získáno 23. Srpna 2017, z Research Program on Forecasting: http://www.gwu.edu/~forcpgm/2009-002.pdf

Šedivý, P. (2016). *Sparta - Jablonec 1:2, domácí mladíci bojovali, ale do finále jde soupeř*. Získáno 22. Srpna 2017, z iDNES Sport: http://fotbal.idnes.cz/sparta-jablonec-odveta-semifinale-domaciho-fotbaloveho-poharu-p7p-/fot_dsouteze.aspx?c=A160504_122233_fot_dsouteze_pes

Seznam tabulek a obrázků

Tabulky

Obrázky

Přílohy

Příloha 1: Počáteční Elo týmů pro sezónu 1993/94

Příloha 2: Počet zápasů podle skóre

Příloha 3: Diagnostika reziduí vícenásobné lineární regrese

Příloha 4: Zápasy, ve kterých domácí tým prohrou zvýší svoji šanci na účast v evropských pohárech alespoň o 1 p. b.

Příloha 5: Zápasy, ve kterých hostující tým prohrou zvýší svoji šanci na účast v evropských pohárech alespoň o 1 p. b.

Příloha 1: Počáteční Elo týmů pro sezónu 1993/94

1. liga

Tým	Elo
Sparta	1 756
Slavia	1 724
Ostrava	1 691
Budějovice	1 691
Liberec	1 682
Olomouc	1 663
Plzeň	1 661
Zlín	1 650
Cheb	1 647
Žižkov	1 634
Brno	1 626
Drnovice	1 611
Bohemians 1905	1 601
Hradec	1 588
FC Karviná	1 549
Dukla	1 489
Průměr	**1 641**

2. liga

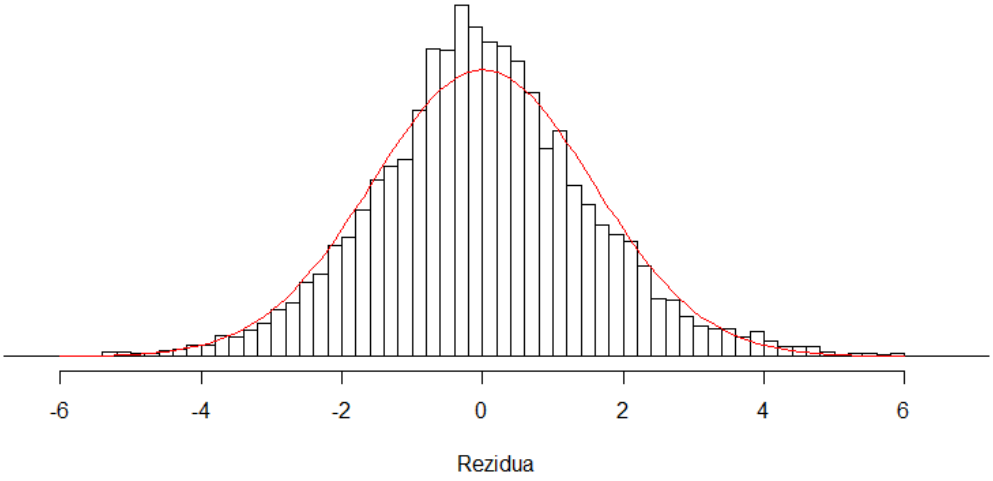
Tým	Elo
Jablonec	1 454
Benešov	1 452
Blšany	1 429
MFK Frýdek-Místek	1 415
FK Fotbal Třinec	1 408
Opava	1 399
Teplice	1 380
SK Pardubice	1 362
FK Baník Havířov	1 340
SK ALFA Brandýs nad Labem	1 324
FK Český ráj Turnov	1 321
SC Xaverov Horní Počernice	1 316
1. SK Prostějov	1 311
FC Coring Bohumín	1 283
Znojmo	1 274
Kladno	1 269
Průměr	**1 359**

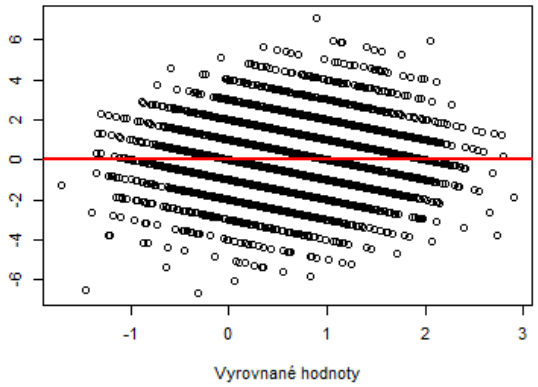
Příloha 2: Počet zápasů podle skóre

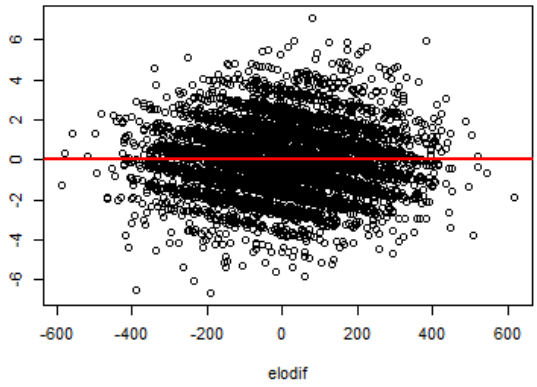
1. liga		Góly hosté								
		0	1	2	3	4	5	6	7	8
Góly domácí	0	546	379	248	83	38	6	0	1	0
	1	710	708	312	127	27	7	1	0	1
	2	565	481	248	80	22	4	1	0	0
	3	292	249	126	37	12	0	0	0	0
	4	124	101	49	19	4	0	0	0	0
	5	48	40	17	2	2	1	0	0	0
	6	14	10	5	0	0	0	0	0	0
	7	6	3	2	0	0	0	0	0	0
	8	1	0	0	0	0	0	0	0	0
	9	0	1	0	0	0	0	0	0	0

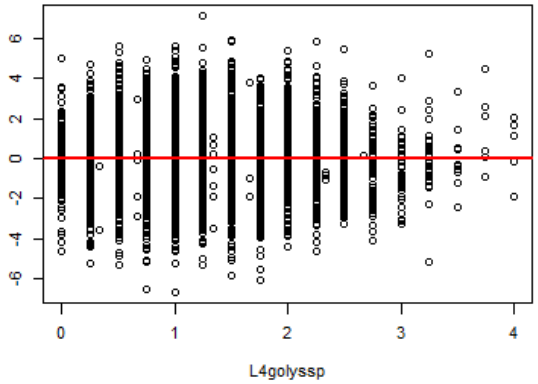
2. liga		Góly hosté								
		0	1	2	3	4	5	6	7	9
Góly domácí	0	542	415	242	101	36	15	4	1	0
	1	664	680	277	121	45	8	3	0	1
	2	567	510	239	73	28	9	1	0	0
	3	270	252	123	42	2	3	0	0	0
	4	118	108	42	14	7	0	0	0	0
	5	55	42	20	6	0	0	0	0	0
	6	18	12	1	0	0	0	0	0	0
	7	5	4	4	1	1	0	0	0	0
	8	0	3	0	0	0	0	0	0	0
	10	1	0	0	0	0	0	0	0	0

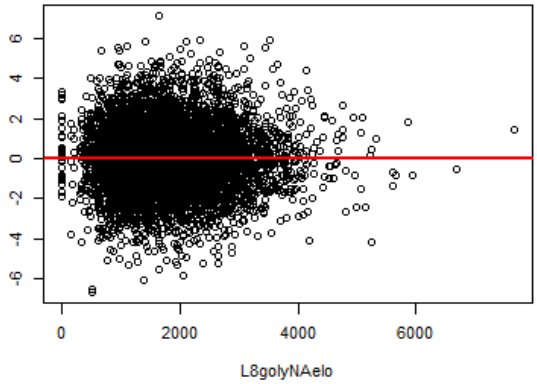
Příloha 3: Diagnostika reziduí vícenásobné lineární regrese

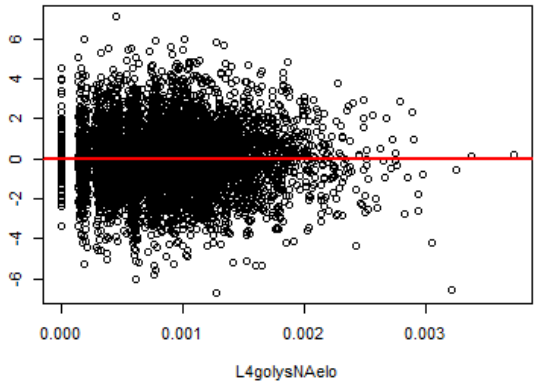


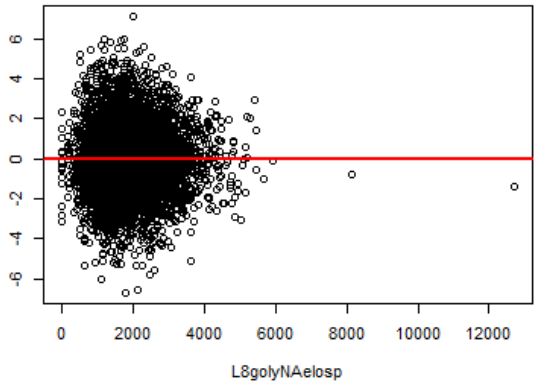












Příloha 4: Zápasy, ve kterých domácí tým prohrou zvýší svoji šanci na účast v evropských pohárech alespoň o 1 p. b.

Sezóna	Kolo	Datum	Domácí	Hosté	Výsledek	Prohra	Výhra	Zvýšení šance	
								absolutně	relativně
2012/13	28	22.5.2013	Slavia	Budějovice	3:0	20,55 %	17,35 %	3,20 p. b.	18,44 %
2012/13	29	26.5.2013	Slavia	Hradec	3:0	21,75 %	17,65 %	4,10 p. b.	23,23 %
2013/14	27	2.5.2014	Dukla	Slovácko	1:1	13,00 %	9,20 %	3,80 p. b.	41,30 %
2013/14	28	10.5.2014	Slovácko	Znojmo	2:0	7,05 %	4,20 %	2,85 p. b.	67,86 %
2013/14	29	25.5.2014	Dukla	Liberec	0:1	15,35 %	12,60 %	2,75 p. b.	21,83 %
2014/15	29	23.5.2015	Dukla	Teplice	5:1	12,90 %	9,50 %	3,40 p. b.	35,79 %
2014/15	30	30.5.2015	Teplice	Hradec	1:3	10,10 %	5,70 %	4,40 p. b.	77,19 %
2015/16	21	11.3.2016	Jablonec	Ostrava	1:1	13,90 %	12,80 %	1,10 p. b.	8,59 %
2015/16	24	8.4.2016	Brno	Bohemians	2:1	6,45 %	4,65 %	1,80 p. b.	38,71 %
2015/16	19	20.4.2016	Dukla	Sparta	1:2	10,60 %	9,20 %	1,40 p. b.	15,22 %
2015/16	26	24.4.2016	Brno	Příbram	2:0	3,15 %	1,35 %	1,80 p. b.	133,33 %
2015/16	28	8.5.2016	Jablonec	Teplice	1:0	16,65 %	15,50 %	1,15 p. b.	7,42 %
2016/17	19	4.3.2017	Jablonec	Bohemians	0:0	14,35 %	13,30 %	1,05 p. b.	7,89 %
2016/17	28	13.5.2017	Zlín	Jihlava	0:3	1,75 %	0,25 %	1,50 p. b.	600,00 %
2016/17	29	20.5.2017	Jablonec	Teplice	0:0	14,10 %	8,55 %	5,55 p. b.	64,91 %
2016/17	30	27.5.2017	Liberec	Ml. Boleslav	4:0	22,25 %	11,40 %	10,85 p. b.	95,18 %

Příloha 5: Zápasy, ve kterých hostující tým prohrou zvýší svoji šanci na účast v evropských pohárech alespoň o 1 p. b.

Sezóna	Kolo	Datum	Domácí	Hosté	Výsledek	Prohra	Výhra	Zvýšení šance	
								absolutně	relativně
2012/13	26	5.5.2013	Brno	Ml. Boleslav	1:1	7,55 %	5,60 %	1,95 p. b.	34,82 %
2012/13	27	11.5.2013	Plzeň	Slavia	0:1	19,45 %	18,35 %	1,10 p. b.	5,99 %
2012/13	29	26.5.2013	Dukla	Ml. Boleslav	2:1	8,45 %	4,20 %	4,25 p. b.	101,19 %
2012/13	30	1.6.2013	Jihlava	Slavia	3:1	21,05 %	17,35 %	3,70 p. b.	21,33 %
2013/14	23	8.4.2014	Slavia	Jablonec	0:0	6,85 %	4,95 %	1,90 p. b.	38,38 %
2013/14	28	9.5.2014	Bohemians	Dukla	3:2	13,10 %	9,80 %	3,30 p. b.	33,67 %
2013/14	30	31.5.2014	Sparta	Jihlava	4:1	9,45 %	7,95 %	1,50 p. b.	18,87 %
2015/16	21	12.3.2016	Slovácko	Brno	2:1	9,00 %	7,00 %	2,00 p. b.	28,57 %
2015/16	22	19.3.2016	Ml. Boleslav	Jablonec	1:0	13,30 %	11,15 %	2,15 p. b.	19,28 %
2015/16	24	9.4.2016	Slavia	Jablonec	0:0	17,50 %	14,45 %	3,05 p. b.	21,11 %
2015/16	25	15.4.2016	Dukla	Jablonec	6:1	17,00 %	13,50 %	3,50 p. b.	25,93 %
2015/16	29	11.5.2016	Plzeň	Jablonec	1:2	16,85 %	12,95 %	3,90 p. b.	30,12 %
2016/17	24	15.4.2017	Brno	Jablonec	2:0	14,00 %	12,05 %	1,95 p. b.	16,18 %
2016/17	27	7.5.2017	Sparta	Liberec	1:0	16,75 %	14,70 %	2,05 p. b.	13,95 %
2016/17	28	13.5.2017	Slavia	Dukla	2:2	12,15 %	8,35 %	3,80 p. b.	45,51 %
2016/17	29	20.5.2017	Dukla	Zlín	1:0	1,70 %	0,20 %	1,50 p. b.	750,00 %
2016/17	30	27.5.2017	Teplice	Dukla	0:1	12,70 %	3,00 %	9,70 p. b.	323,33 %
2016/17	30	27.5.2017	Příbram	Zlín	0:2	2,60 %	0,05 %	2,55 p. b.	5100,00 %
2016/17	30	27.5.2017	Karviná	Jablonec	0:2	15,00 %	8,45 %	6,55 p. b.	77,51 %