Posudek oponenta diplomové práce

Autor/Autorka	**Bc. Ivana Gabrišková**
Název práce	**Modelování a odhadování výsledků hokejových utkání**
Studijní obor	**Finanční informatika a statistika**
Oponent práce	**RNDr. Blanka Šedivá, Ph.D.**

Splnění cílů práce:

☐ nadstandardně ☐ velmi dobře ☑ splněny ☐ s výhradami ☐ nebyly splněny

Odborný přínos práce:

☐ nové výsledky ☐ netradiční postupy ☑ zpracování výsledků z různých zdrojů ☐ shrnutí výsledků z různých zdrojů ☐ bez přínosu

Matematická (odborná) úroveň:

☐ vynikající ☐ velmi dobrá ☐ průměrná ☑ podprůměrná ☐ nevyhovující

Věcné chyby:

☐ téměř žádné ☐ vzhledem k rozsahu přiměřený počet ☐ méně podstatné, větší množství ☑ podstatnější, větší množství ☐ závažné

Grafická, jazyková a formální úroveň:

☐ vynikající ☐ velmi dobrá ☐ průměrná ☑ podprůměrná ☐ nevyhovující

Slovní hodnocení a dotazy:

Cílem předkládané diplomové práce bylo popsat základní modely používané pro odhadování výsledků sportovních utkání a aplikovat vybrané modely na reálná data výsledků tří hokejových lig. Zadání práce v tomto rozsahu tak bylo splněno.

Vzhledem k charakteru práce musela autorka věnovat zvýšenou pozornost používaným datovým zdrojům. Popisu použitých dat je věnována již druhá kapitola práce a zároveň jsou data obsažena na přiloženém CD. Popis dat však působí místy zmatečně (např. u polské ekstraligy je uvedeno, že počet týmů v jednotlivých sezónách je různý, následně je však napsáno, že se hraje přesně 36 zápasů – strana 3). Také počty „dohledávaných" výsledků uváděné v práci a označené v přiloženém excelovském *data.xlsx* souboru jsou u české extraligy a NHL odlišné.

Odborný přínos práce byl zaměřen především na zpracování několika modelů vhodných pro modelování výsledků sportovních utkání. Autorka ve své práci vychází pouze z velmi omezeného počtu zdrojů literatury, ale přesto by bylo vhodné zařadit část „literature review", která by pomohla vyjasnit důvody, proč byly zvoleny právě vybrané modely. Navržení vlastního modelu je sympatickým pokusem, který se však omezil pouze na použití na českých datech.

Práce obsahuje podstatnější množství faktických a formálních nedostatků, výčet nejzávažnějších pochybení je popsán na závěr tohoto posudku. Struktura práce a navržení jednotlivých kapitol je sice poměrně logické, autorka však v rámci kapitol často strukturu sama narušuje (viz. konkrétní připomínky 3 a 6).

Celá práce působí chaotickým dojmem, v některých částech jsou zbytečně opakovány části textů, např. strana 8 a 12 (zdůvodnění, proč je použito Poissonovo rozdělení), strana 22 (vzorce věrohodnostních funkcí pro různé

modely). Práce též obsahuje řadu technických popisů, ve kterých buňkách programu Excel jsou uloženy jednotlivé hodnoty, a zbytečné printscreeny obrazovky v Excelu.

Vlastní algoritmy řešení a jednotlivé testované hypotézy jsou popsány nepřesně a nejasně (viz. připomínky 2, 4, 5, 13, 16).

Seznam literatury není stylisticky jednotný (chybná citace u zdroje [12]) a obsahuje zdroje, které nejsou citovány ([9],[10]). U bakalářské práce [12] je pouze uvedeno, že „v této práci lze nalézt inspiraci". Naopak v některých částech literatura zjevně chybí, například Bonferroniho korekce (kapitola 3.4).

Práce postrádá významnější vklad vlastní analytické práce, chybí i jen pokus o zhodnocení získaných výsledků, zamyšlení se nad skutečností v čem jsou si vybrané nejvyšší soutěže hokejových lig podobné a v čem jsou naopak rozdílné. (viz. připomínky 2, 7, 9,17, 18).

Uznat práci jako kvalifikační doporučuji pouze v případě, že studentka vyjasní problémy uvedené v příloze. Zejména se jedná o zopakování odhadů parametrů dle algoritmu na straně 15 a dále o vyjasnění připomínek ke kapitole 7 (připomínka 11 – 18).

Předloženou kvalifikační práci hodnotím stupněm NEVYHOVĚLA, v případě, že studentka u obhajoby vyjasní dostatečně problémy uvedených v posudku nevylučuji výsledné hodnocení dobře.

29.8.2016

Datum, jméno a podpis:

Vybrané konkrétní poznámky a připomínky:

1. Statistické pojmy uváděné v kapitole 3 jsou často nepřesně formulovány, vytržené z kontextu a uváděné vzorce u chí-kvadrát testu nezávislosti náhodných veličin (kapitola 3.5) obsahují formální nepřesnosti.

2. Není jasné, co zachycují Tabulky 3-5 na stranách 9-10. Vyjasněte, k jaké přesně hypotéze se vztahují výsledky vyhodnocení testů ve sloupci Test a p-hodnota.

3. V kapitole 5.2.1 je uvedeno, že „ ... Maherův model není příliš vhodný pro hokejová data, což bylo ověřeno na české lize...". Model je však aplikován až v kapitole 5.3 a není zřejmé, na základě čeho bylo rozhodnuto o nevhodnosti tohoto modelu. Pro NHL a polskou extraligu není v této části práce model použit vůbec. Navíc zvolený Model 2 je reparametrizací modelu DP uváděného v kapitole 6.2 a není tedy pravda, že není dále používán.

4. Z algoritmu postupu odhadu parametrů modelu (strana 15) není jasné, proč se postup hledání numerického řešení rovnic (5.5) a (5.6) opakuje vícekrát. Můj pokus o zopakování nalezení parametrů, dle algoritmu na straně 15 v sw Microsoft Office Professional Plus 2016 skončil nezdarem a nalezené hodnoty byly odlišné od hodnot uváděných v práci. Docházelo při hledání řešení k nějakým numerickým problémům?

5. V kapitole 5.3.3 není zřejmé, zda testujeme shodu očekávaných počtů gólů pro každý tým nebo souhrnně pro celou ligu. Také není zřejmé, zda v tomto případě byla použita Bonferroniho korekce hladiny významnosti.

6. V kapitole 6.5.1 (strana 21) je zmiňováno, že váhová funkce byla zavedena v předešlých kapitolách. V práci je však váhová funkce zavedena až v následujícím textu.

7. V kapitolách 6.6.1, 6.7.1, 6.8.1 a 6.9.1 je při hledání „optimální" hodnoty parametru ξ je pokaždé použita jiná množina, ze které jsou hodnoty vybírány (konkrétní rozdíly jsou zřetelné v Tabulce 10 – strana 26, Tabulce 15 – strana 31, Tabulkách 16-18 – strana 35 a Tabulce 20 – strana 40). V textu nejsou uvedeny důvody pro takto rozdílný přístup.

8. Vzhledem k tomu, že modely v kapitole 6 využívají „faktor zapomínání" závislý na čase, měly by být u grafů na Obrázku 5 a dále vždy na x-ové ose datumy, kdy se uvedená kola konala, a grafy by měly mít charakter po částech konstantních grafů.

9. Volby vhodných modelů v kapitole 6 a získané výsledky nejsou presentovány v jednotné formě a v některých okamžicích si i protiřečí. Například u české extraligy se píše „odhad parametru výhody domácího prostředí se téměř nemění, protože popisují celou ligu" (strana 28) a naopak na str.33 je pro polskou ligu vývoj tohoto parametru vykreslen a komentován. Existuje tedy rozdíl ve vnímání domácího prostředí v české a v polské extralize?

10. V kapitole 6 jsou často zbytečně uváděny technické popisy postupu v programu Excel (jakou buňku nastavit, obrázky nastavení řešitele, apod.).

11. V kapitole 6.9 je popsán vlastní model, kdy je pro každý tým uvažován rozdílný parametr výhody domácího prostředí, tím se však autorka „pouze" vrací k modelům popsaným v kapitole 5. Zároveň by bylo vhodné více diskutovat porovnání modelů s ohledem na různý počet odhadovaných parametrů tak, jak je to rozpracováno i v citované Maherově práci (například statisticky porovnat logaritmicko věrohodnostní funkce modelů apod.).

12. V úvodu kapitoly 7 je chybně analyzovaná a vysvětlena hodnota parametru L. Interpretaci L jako minimální hranice pro očekávanou hodnotu zisku považuji za chybné. Následně je zavedena podmínka L>1, ale v další části textu se pracuje s hodnotami L=1,0;1,1 ... . Dále není zřejmé o jakou teorii se opírá tvrzení „teoreticky by bylo nejlepší volit co největší L".

13. V kapitole 7.1 Sázení pro Extraligu je nejprve uvažováno, že je třeba 2 940 Kč, což odpovídá sázení na 91% zápasů. Následně je v kapitole 7.1.1 provedena simulace při sázení na 100%, resp. 60%, resp. 80% zápasů a uvažovaná částka je 1000 Kč. Není tedy jasné, zda jsou porovnávány porovnatelné modely. Například pro modely pro data Extraliga CZE na Obrázku 25 na straně 45 porovnáváme hodnoty pro L=1,07, což podle Tabulky 21 odpovídá počtu zápasů 146 (38%) s náhodným sázením na 80% zápasů.

14. Dále je zde uvedeno, že při výši 1000 Kč vkladu, nenastala situace, kdy by sázející zbankrotoval. Jak si však vysvětlit hodnoty histogramu zisk/ztráta menší než 1000 ?

15. Na straně 43 je uvedeno, že „model je ziskový již pro hodnoty parametru L 1,07 a vyšší". V Tabulce 21 je však pro L=1,21 uveden čistý zisk -8,00.

16. Algoritmus náhodného sázení na straně 44 je nejasný. Funkce randbetween v programu Excel vrací celá čísla (Microsoft Office Professional Plus 2016). Jak je pak myšlena následující podmínka „v případě vygenerování čísla menšího než 0,2"? Pokud je algoritmus náhodného sázení myšlen tak, že na všechny možné výsledky zápasu sázíme se stejnou pravděpodobností, pak tento koncept odpovídá nereálnému předpokladu, že všechny výsledky zápasů (výhra domácích, remíza, výhra hostů) jsou stejně pravděpodobné.

17. V práci není proveden ani pokus o vysvětlení, dle mého názoru zajímavého, chování závislosti ziskovosti sázení na volbě parametru L. Například Obrázek 26 na straně 46 naznačuje, že počet sázek s rostoucí hodnotou L monotónně klesá, ale průběh zisku má výrazně komplikovanější tvar. Můžete vysvětlit, čím je situace způsobena.

18. Stejně práce neobsahuje zamyšlení nad tím, proč například strategie sázení na maximální kurzy je nadprůměrně úspěšná v české extralize a naopak propadající u dalších dvou soutěží.