Posudek oponenta diplomové práce

Autor/Autorka	**Ivana Gabrišková**
Název práce	**Modelování a odhadování výsledků hokejových utkání**
Studijní obor	**Finanční informatika a statistika**
Oponent práce	**Petr Stehlík**

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:

Ivana Gabrišková ve své práci odhaduje výsledky hokejových utkání 3 vybraných lig. K dispozici má datový soubor zhruba s 10.000 zápasy a kurzy sázkových kanceláří na tyto zápasy. Pomocí těchto zápasů následně odhaduje parametry rozšířeného Dixon-Colesova modelu (Dixon, Coles (1997)) o tzv. diagonální rozšíření (Marek, Ťoupal, Šedivá (2014)) a svého mírně rozšiřujícího vlastního modelu.

Práce je přepracovanou verzí diplomové práce, která byla neúspěšně obhajována v loňském roce. Letmým srovnáním je zřejmé, že hlavním rozšíření je rozpracování vlastního modelu v Kapitole 7.

Kromě zpracování velkého množství dat (je nutno ocenit nejen jejich stahování, ale i následné pracné ruční úpravy chybějících dat), autorka studuje statistické modely, jejichž parametry následně odhaduje metodami numerické optimalizace. Vlastní model považuji za rozumné, dobře motivované a zajímavé rozšíření. Stejně tak je srozumitelné použití modelů na různé sázecí strategie.

Zpracování je bohužel velmi slabé.

Text je neopodstaněně rozplizlý na 80 stran a bohužel většinou velmi těžko srozumitelný. Autorka na jedné straně popisuje naprosté zbytečnosti (zadávání hodnot do Excelu, těžko pochopitelné a nešikovné slovní legendy ke grafům...), na druhé straně neuvádí zásadní informace (vazby v modelech, volbu metod, volba lig...). Mnohé odstavce vyvolávají otázky o tom, zda autorka rozumí principům, na kterých je práce postavena (optimalizace, diskontování starších výsledků, závěrečná kapitola o sázení...) a v práci je bohužel mnoho vágních a nekonkrétních formulací odkazů (nespecifické odkazy, některé další modely vyžadují [podmínka], nahrazujeme problematické řešení z [3]...)

Zásadním nedostatkem je pak zcela jistě absence matematických zápisu tvrzení, které je nahrazeno jejich rozmělňování do nesrozumitelných odstavců (navíc, většina matematických vzorců je identicky převzána z Marek, Ťoupal, Šedivá (2014)). Stejně negativně vnímám absenci diskuse do šířky (např. u kritického diskontního parametru ξ, ale i volby alternativních sázecích strategií, či chybějícího přehledu literatury...)

Jazykově a graficky nepovažuji práci také za zdařilou. Kromě pochopitelných překlepů obsahuje i opakující se nedostatky (Poissonovo model, NHL liga...). Grafické ilustrace jsou nekonzistentní, některé tabulky vložené jako obrázky, či vyvolávají neúplný pocit. Mnohé vizualizace jsou zcela nevhodně zvoleny (Obr. 5,11,18), zkreslují (např. Obr. 7), či používají nekonzistentní osy (čas – kolo x dny). Kvalitě textu jistě nepomohou i velmi obecné odkazy typu „dle podmínky z kapitoly 5.1.3".

Za kritický bod Dixon-Colesova přístupu považuji odhad diskontního parametru ξ. Autorka správně naznačuje, že první rozpor spočívá ve faktu, že zatímco modely odhadují skóre zápasů, optimalizace diskontního parametru je prováděna na základě prostých výsledků (domácí, remíza, výhra). Nicméně, srovnání rozšiřujícího (Marek, Ťoupal, Šedivá (2014)) s vlastním modelem autorky ukazuje, že odhad parametru ξ může při volbě různých modelů vycházet velmi odlišně, navíc parametr je volen z (vždy jiných) malých konečných množin. Vyvstává velké množství otázek, které jsou s tímto spojeny. Pro rozumné vlastnosti modelu by stálo za to srovnat výpočet optimální hodnoty ξ nejen v jednom časovém okamžiku pro daný model a daná data. Z teoretického pohledu jde o fakt, zda-li funkce S(ξ) má jediné globální maximum. Pokud ano, pak neprůhledný způsob výběru testovaných hodnot ξ by zcela jistě šel nahradit např. metodou bisekce.

Další drobné připomínky:

A. Autorka v letošní verzi neaktualizovala data o sezónu 2016-17.
B. Není zřejmé, proč byly vybrány právě tyto 3 ligy, zejména volba polská Ekstraligy působí bez komentáře zvláštně.
C. Ve stejném duchu není jasné, proč nejsou diskutovány další modely odhadování výsledků sportovních výsledků. Kromě článku Marek, Ťoupal, Šedivá (2014), končí snaha o přehled literatury v roce 1997.
D. První věta kapitoly 2 je příliš silná. Data dle mého názoru nejsou ani „správná" (to o kurzech, s povahou panelových dat, nelze říci nikdy), ani „úplná" (chybí zápas NHL ze stávkové sezóny 2012-13)
E. Autorka vypouští zápasy play-off, aniž by komentovala důvody.
F. Autorka velmi nešťastně používá v celé práci kolo místo „hrací den". Kolo má v evropských ligách jiný význam, než který používá autorka, a v NHL se vůbec nepoužívá.
G. Definice p-hodnoty na str.5 není správně.
H. Test nezávislosti v kapitole 4.2 je dle mého názoru problematický. Autorka proti postupu v minulém odstavci rozděluje zápasy na sezóny, čímž nezamítá alespoň v některých případech hypotézu o nezávislosti o počtu vstřelených gólů mezi domácím a hostujícím týmem.
I. Místo popisu algoritmu numerického hledání optimálních hodnot parametrů na straně 16, autorka popisuje způsob zadání do konkrétního SW, její překlad GRG algoritmu jako Gradientní metody je nepřesný.
J. Z textu na str.15 bych očekával, že i v algoritmu na str. 16 budou použity odhady (5.10) a (5.11)
K. Váhová funkce (6.11) je diskontní funkce pro log. věrohodnostní funkci, výklad na straně 25 tento fakt spíš zamlžuje, než vysvětluje.
L. Vlastní model (7.1)-(7.2) nedává smysl, pokud není přidána nějaká vazba na nové parametry γ_i
M. Tabulka 25 nic nepotvrzuje, jen naznačuje a jistě by stálo uvést poměry pro všechny sezóny z dané ligy,
N. Kapitola 8 je obecně velmi poškozena znovu vágními, roztáhlými a nepřesnými formulacemi (např. u popisu strategií), chybějícími popisy (volby hodnot L, spojité intervaly) a nešťastnými popisky histogramů.

O. Literatura – nekonzistentní [17], chybějící autor v [8], nešťastné odkazy do wikipedie místo na stránky lig, či na manuály místo na popisy numerických metod.

Otázky k ústní zkoušce

1. Jak dopadnou testy nezávislosti z kapitoly pro celé datové soubory jednotlivých lig?
2. Popište GRG metodu, kterou používáte pro nalezení optimálních hodnot.
3. Pokuste si připravit výpočet optimální hodnoty parametru ξ pro pevně daný model a zvolenou ligu v různých časových okamžicích

Práci doporučuji – ~~nedoporučuji~~ uznat jako kvalifikační (*nehodící se škrtněte*) v případě zodpovězení výše uvedených otázek a *navrhuji hodnocení známkou:*

dobře

Datum, jméno a podpis:

8.6.2018
Petr Stehlík