Posudek oponentky na diplomovou práci

Autor: Bc. Ondřej Lepša

Název: Využití bootstrapu a křížové validace v odhadu predikční chyby regresních modelů

Předložená práce se zabývá tématem posuzování predikční kvality regresního modelu. Téma považuji za užitečné, zajímavé a velmi vhodné, vyžadující od autora široké spektrum dovedností. Často se setkáváme i u uživatelů ve statistice vzdělaných s tím, že je zaměňována kvalita modelu (popis dat) a jeho predikční schopnost. Celkově téma tak, jak ho uchopil autor, považuji za obtížné až příliš obtížné.

Autor odvedl při přípravě textu velké množství práce a je jasně patrné jeho zaujetí řešeným problémem. Ukazuje dobrou znalost problematiky lineárních regresních modelů a zaujetí pro výpočetní statistiku a programování.

Práce obsahuje velmi kvalitní práci s literaturou, je vidět, že autor opravdu četl a nad obsahem článků se zamýšlel. V textu vhodně cituje prameny, ilustruje jimi svoje úvahy a používá je také k diskusi. V tomto smyslu považuji také seznam použité literatury za velmi vhodný, jen mi není jasné, k čemu je publikace první (sbírka příkladů). Při četbě jsem narazila jen na jeden odkaz na tuto knihu, který by se dal určitě nalézt i v jiných použitých materiálech. Také publikace Hebák (2005) tak, jak je uvedená v seznamu, neexistuje. Je třeba také citovat program R a použité balíčky, které často vycházejí z článků.

Dále je třeba vysoce ocenit výpočetní schopnosti a uživatelskou znalost programu R na velmi dobré úrovni. Autor práce s přehledem naprogramoval (napsal skript) potřebné postupy a dal přednost tomu, napsat svůj postup, před již hotovými balíčky. Popis skriptu v textu je jasný a přehledný. Je mu věnována, podle mého názoru, příliš velká pozornost na úkor závěrů a prezentace výsledků. Je ale jasné, že autor předkládá svou práci.

Autor se dále nevyvaroval chyb. Z formálních bych chtěla zdůraznit, že vzorec je součástí textu a nemůže být tečka za větou před ním. Dále je třeba používat stejné písmo ve vzorcích i textu. V českém textu také nezůstává na konci řádku jednopísmenná předložka. Převážně anglická literatura s sebou přináší problémy s českou terminologií, z tohoto důvodu by stálo za úvahu podobné práce psát anglicky. Dále při psaní se občas projevují (patrně) automatické opravy používající na základě angličtiny velká písmena tam, kde v češtině nejsou. Obdobné problémy má autor v případě skloňování nepřeložených názvů a podle mého názoru by autor našel lepší překlad než „selekce modelu“.

Z hlediska čtenáře je nepříjemné to, že autor důsledně nerozlišuje vektory a náhodné veličiny, tato tendence v celé práci vrcholí v příloze tvrzením, že rozptyl … je vlastně kovarianční matice. Nemohu souhlasit s textem omezujícím použití kritérií AIC a BIC, obě kritéria jsou široce používána pro srovnání modelů využívajících různých rozdělení a nejenom pro regresní problémy. Domnívám se, že autor měl na mysli pouze své definice uvedené v práci, které normalitu vyžadují, nejsou ale obecnými používanými pro maximálně věrohodný odhad. Z textu dále není jasné, jak zkoumaný problém ovlivňuje rozptyl residuí (autor volí jednotkový), lze u souboru s feliciemi takovou věc předpokládat? Bohužel se mi zdá, že v místech, kde jsou autorem prezentovány výsledky, je práce trochu zmatená a čtenář se velmi těžko orientuje.

Otázky:

1. V textu se věnujete také volbě počtu výběrů v bootstrapu. V některé literatuře se uvádí řádově větší počty výběrů, než uvádíte. Zkoušel jste více možností volby B a vliv této volby na výsledky?

2. V popisu procedur máte vždy znova napsané generování modelů. Dělal jste to ve skutečnosti tak, že jste generoval model a použil ho pro všechny úvahy (případně jste to nedělal najednou, ale generované hodnoty řídil volbou parametrů generátoru)?

Předloženou práci považuji za vysoce nadstandardní a zdařilou, problémy a nedokonalosti, z nichž některé jsem zmínila v posudku, vyplývají z povahy textu a lze je autorovi prominout. Proto práci doporučuji k obhajobě a v případě úspěšného průběhu obhajoby doporučuji hodnotit **výborně.**

V Praze Ivana Malá