Oponentský posudek diplomové práce

Analýza cen pojištění pomocí strojového učení

Autorka práce: Bc. Barbora Pánková [ČVUT v Praze, FJFI]

Předložená diplomová práce se zabývá analýzou cen pojištění odpovědnosti z provozu vozidla prováděnou prostřednictvím zobecněných lineárních modelů a neuronových sítí. Obsahově je vlastní text rozčleněn do čtyř kapitol. První kapitola stručně připomíná základní principy pojišťovnictví a v obecné rovině představuje studovaný problém odhadování (konkurenčních) cen pojištění. Druhá kapitola uvádí obecný teoretický rámec zobecněných lineárních modelů a souvisejících statistických konceptů. Třetí kapitola zevrubně popisuje konstrukci vlastních zobecněných lineárních modelů cen pojištění odpovědnosti z provozu vozidla dvou konkurenčních pojišťoven (s využitím log-link gama regrese), jejich validaci a vzájemné srovnání. Čtvrtá kapitola se zaměřuje na aplikaci neuronových sítí v kontextu předmětné úlohy (v obdobném rozsahu jako ve třetí kapitole) a na komparaci jednotlivých prezentovaných modelových přístupů.

První dvě kapitoly diplomové práce mají vesměs kompilační charakter. Ve třetí a čtvrté kapitole je patrný vlastní příspěvek autorky v podobě podrobné explorační analýzy vstupních dat, identifikace a konstrukce různých modelů cen pojištění, jejich celkového zhodnocení a porovnání jejich predikčních schopností. Práce má tedy zřetelný praktický přesah.

Samotné vymezení i zpracování tématu diplomové práce je vyhovující. Autorka zadané téma kvalifikovaně uchopila, dokázala se podle všeho orientovat v širším spektru odborné literatury a syntetizovat dosažené poznání. Po formální stránce je práce bez problémů. Podmínky kladené na rozsah splňuje. Obsahuje přiměřené množství matematických či metodologických nepřesností, resp. stylistických či pravopisných pochybení. Úprava práce je adekvátní.

V textu se ovšem také objevily některé nedostatky, resp. nejasnosti. Uveďme vybrané:

Obecné

- Je vždy třeba korektně a komplexně zavést, resp. vysvětlit používané odborné termíny a značení (kupříkladu $\widehat{m}$ v definici 2.2.1, funkce $f(\cdot)$ v rovnicích (2.3) a (2.5), obory hodnot parametrů α a β v rovnici (2.5), pojem *kanonické spojovací funkce* v sekci 2.3, *Akaikeho informační kritérium* v sekci 3.4.4, $F^{-1}(\alpha)$ v definici 4.2.2 apod.).
- Při zápisu vektorů je vhodné dodržovat jednotný přístup k transponování (viz strana 12, odstavec 4; strana 14, poslední odstavec aj.).
- Některá (vesměs drobná) typografická / pravopisná pochybení: (*i*) ponechávání vybraných předložek samostatně na koncích řádků (cf. https://prirucka.ujc.cas.cz/?id=880), (*ii*) čísla uváděná v tabulkách se obvykle zarovnávají k desetinnému oddělovači, (*iii*) je třeba vhodně rozlišovat mezi „5%“ a „5 %“, „90 kW“ a „90kW“ apod.

Kapitola 2

- Tato kapitola je pojata relativně stručně, některé pasáže mohly být rozšířeny, resp. zpracovány detailněji (například (*i*) zcela absentuje zmínka o metodice odhadu neznámých parametrů zobecněných lineárních modelů a (*ii*) sekce 2.5 a 2.6 jsou velmi obecné).
- Definice 2.4.1: Není uvedena podmínka existence konečných druhých momentů náhodných veličin X, Y.

Kapitola 3

- Sekce 3.1.2 a 3.4.2: Byly v rámci explorace dat nějak ošetřeny chybějící záznamy, neobvyklá či chybná pozorování? Cf. například rozsah první kategorie na obrázcích 3.7, 3.8, 3.9, 3.10, 3.28 a 3.29.

- Sekce 3.1.2 a 3.4.2, odstavec *Okres bydliště*: Bylo by vhodnější uvést taxativní výčet okresů v jednotlivých kategoriích, eventuálně zvolenou kategorizaci zaznamenat přímo do mapy ČR.
- Sekce 3.1.3, tvrzení *„Je třeba ověřit, zda jsou obě proměnné v celkovém modelu významné. Pokud by nebyly, jednu proměnnou bychom do modelu nezahrnuli.“* (podobně pak sekce i v 3.4.3): Pokud by (obě) silně korelované spojité proměnné byly významné a současně by byly ponechány v modelu, mohl by být potenciálně identifikován problém s multikolinearitou.
- Sekce 3.1.4 a 3.4.4: Není zřejmé, jaký typ residuí (standardizovaná, devianční, parciální, …) byl v rámci prováděné diagnostiky uvažován. Samotná residuální diagnostika je vesměs základní.
- Sekce 3.6: Jakým způsobem byla párována data obou pojišťoven (tj., jak proběhlo přiřazení pojistného bíle pojišťovny k datové sadě černé pojišťovny)? Není diskutováno.
- Byly při tvorbě modelů uváženy možné interakce mezi vysvětlujícími proměnnými, popř. užití funkčních transformací vysvětlujících proměnných? Není diskutováno.
- Výpisy z výpočetního software (speciálně ty, které zachycují odhady parametrů log-link gama regresních modelů a související modelové charakteristiky) nebyly řádně vysvětleny (interpretovány, formalizovány). Autorka implicitně spoléhá na erudici čtenáře.
- Ve výpisech zachycujících odhady parametrů log-link gama regresních modelů by podle všeho měly být zobrazovány *z*-poměry, nikoliv *t*-poměry (a k nim příslušné *p*-hodnoty).
- Jaké softwarové procedury byly aplikovány v rámci konstrukce modelů? Software **`R`**, funkce **`glm`**?

Kapitola 4

- Sekce 4.3 a 4.4: Jak byly stanovovány hodnoty hyperparametrů analyzovaných neuronových sítí (expertní výběr vs. *grid search*, resp. kombinace obojího)? Byla testována neuronová síť s více než jednou skrytou vrstvou? Z uvedeného totiž není zcela zřejmé, zda byly tyto klíčové aspekty při konstrukci modelů neuronových sítí korektně uváženy.

Literatura

- Některé deklarované reference použité při přípravě předložené kvalifikační práce jsou s ohledem na jejich povahu přinejmenším diskutabilní, viz například položky [3], [11], [13], [15] a [16], strany 67 a 68. Nadto, bývá zvykem seznam referencí řadit podle abecedy a uvádět datum návštěvy odkazované webové stránky.
- Je-li citována knižní publikace, je vždy lépe specifikovat kapitolu či alternativně rozsah stran, na něž se odkazuje.

Závěr: Předloženou diplomovou práci **doporučuji** přijmout k obhajobě a navrhuji ji klasifikovat stupněm **C**.

V Pardubicích dne 17. července 2020 Radek Hendrych

Kontakt:

RNDr. Mgr. Radek Hendrych, Ph.D.

Univerzita Karlova, Matematicko-fyzikální fakulta, Katedra pravděpodobnosti a matematické statistiky

Sokolovská 83, 186 75 Praha 8, Česká republika, *e-mail:* hendrych@karlin.mff.cuni.cz