RECENZNÉ KONANIE **PRIPOMIENKOVACÍ HÁROK**

TERMÍN: 6. júna 2019

NÁZOV MATERIÁLU: Čierna práca sa nevypláca: Cielenie kontrol nelegálneho zamestnávania s využitím administratívnych dát

TYP VÝSTUPU*[1]: analýza

AUTOR(I): Ján Komadel

ANALYTICKÝ ÚTVAR, REZORT: ISP, MPSVR SR

RECENZNÝ FORMÁT*[2]: 2

RECENZENT: Róbert Tóth

PRIPOMIENKY:

Pripomienka sa vzťahuje k (strana, odsek):	Text pripomienky*[3]	Odôvodnenie pripomienky	Vysporiadanie sa s pripomienkou*[4]
7, 2	V štúdii chýba využitie algoritmov ako https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf, kde síce ide o klasifikáciu, ale do úvahy je braná aj disproporčnosť dát. Delenie na klasifikáciu a detekovanie anomálií dáva zmysel, ale ani pri balanse pozorovaní menšom ako 1 percento, netreba automaticky siahať po detekčných algoritmoch. Naopak, správne použitie klasifikačných algoritmov pre nevybalansované vzorky dát môže priniesť oveľa lepšiu presnosť.		Pripomienka neakceptovaná. Využitie klasifikačných metód určených na nevyvážené dáta je určite dobrý nápad. Z časových dôvodov túto analýzu nebudeme rozširovať, ale v budúcnosti pri riešení podobných úloh otestujeme aj tento typ metód. Výhodou logistickej regresie oproti náhodnému lesu je možnosť priamo určiť vplyv prediktorov na odhadovanú mieru podozrenia.
14, 4	Prečo nie je lepšie vytvoriť jeden veľký model a odvetvie použiť ako premennú? Nestráca sa takto spoločná sila prediktorov naprieč odvetviami? Možno ide o nerovnakú dostupnosť prediktorov pre dané odvetvia alebo o nemožnosť v praxi všetky kontroly posielať iba do jedného odvetvia ak by model takto vybral?		Pripomienka neakceptovaná. Testovali sme aj model pre všetky odvetvia spolu, ale v porovnaní dopadol horšie ako samostatné modely. Veľký spoločný model si navyše vyžaduje aj veľa interakčných členov, keďže významné prediktory pre jednotlivé odvetvia sa líšia (hovoria

[1] Podľa parametrov analytických výstupov opísaných v materiáli Recenzný postup.
[2] Podľa materiálu Recenzný postup.
[3] Do tabuľky značiť pripomienky zásadného metodologického a obsahového charakteru (nie štylistické či gramatické opravy).
[4] Pripomienka bola akceptovaná / pripomienka nebola akceptovaná a zdôvodnenie/ pripomienka bola čiastočne akceptovaná a zdôvodnenie.

			o tom skúsenosti NIPu aj naše výsledky). Použitie samostatných modelov umožnilo identifikáciu sektorovo špecifických prediktorov. Napriek tomu sme pozorovali aj spoločné prediktory pre viaceré odvetvia.
9, 2	Bolo by vhodné presnejšie popísať výber najlepšieho modelu. Bola „hraničná“ pravdepodobnosť vybraná pre každý algoritmus rovnako – t.j. cross-validovaním na roku 2018? Ak áno, napriek tomu, že sa to môže zdať „fér“, nie je dobrá praktika do výberu najlepšieho modelu miešať aj dáta odložené bokom. Pri tomto spôsobe sa môže stať, že víťazný algoritmus nebude ten, ktorý by v praxi fungoval najlepšie, ale ten, ktorý dokázal najlepšie „podvádzať“ pri výbere kritickej pravdepodobnostnej hranice vidiac odloženú vzorku dát.		Pripomienka akceptovaná. Pridaný popis výberu hraničnej miery podozrenia. Hraničná miera je vyberaná na validačnej vzorke a úspešnosť je porovnávaná na testovacej vzorke, takže „podvádzanie“ na testovacích dátach nie je. Preto aj neakceptujeme poslednú pripomienku nižšie.
8, Box 3	Bolo by možno vhodné použiť aj metriku https://en.wikipedia.org/wiki/Lift_(data_mining), ktorá viac reflektuje mieru pravdepodobnosti priradenej jednotlivým subjektom. Pre štatistiku F1 je v zásade jedno, či je pravdepodobnosť NZ 90 percent alebo 70, ak sú obe nad kritickou hranicou. Pritom v praxi by sa na kontrolu zrejme upredňostňovali subjekty s vyššou pravdepodobnosťou. Ak modely z nejakého dôvodu lepšie triafajú subjekty, ktorým prisúdili menšiu pravdepodobnosť (kvôli overfittingu), je dobré o tom vedieť. Rovnako by táto miera poskytla lepší odhad pomeru počtu kontrol a odhalených NZ.		Pripomienka akceptovaná. Pridaná príloha s kumulatívnymi ziskami a liftom.
12, 2	Tento odhad môže byť nepresný, viď poznámku vyššie.		Pripomienka neakceptovaná.

RECENZNÉ KONANIE **PRIPOMIENKOVACÍ HÁROK**

CELKOVÉ HODNOTENIE: V oblasti machine learningu vždy bude existovať veľa alternatív pre algoritmy a miery ich vyhodnotenia a aj keď táto štúdia neprechádza cez celý ich rozsah, je veľmi kvalitným vypracovaním danej problematiky a dáva jasný dôvod si myslieť, že zavedenie algoritmu do praxe môže výrazne znížiť náklady štátu.

SCHVÁLIŤ*[5]: X odporúčam

neodporúčam

---

podpis recenzenta

Súhlasím* s uvedením svojho mena ako mena recenzenta v recenzovanej publikácii:

X ÁNO

NIE

Súhlasím* so zverejnením tohto pripomienkovacieho hárka:

X ÁNO

NIE

5 Hodiace sa označte krížikom.