RECENZNÉ KONANIE **PRIPOMIENKOVACÍ HÁROK**

TERMÍN: 05.11.2020

xx38670xx
Recenzia A
Lukáš Lafférs
lukas.laffers@gmail.com

Prosím nezasahujte do tejto tabuľky

RECENZENT/KA (meno a priezvisko, pozícia, inštitúcia)**: Lukáš Lafférs**
NÁZOV MATERIÁLU: **Profilácia UoZ pomocou strojového učenia**
TYP VÝSTUPU*[1]: **Analýza**
(pri spoločných výstupoch uviesť aj typy individuálnych vkladov):
ANALYTICKÝ ÚTVAR, REZORT: **Ministerstvo práce, sociálnych vecí a rodiny SR - Inštitút sociálnej politiky**
AUTORI/KY: **Ján Komadel, Tadeáš Chujac, Martin Demeter;**
SPOLUAUTORI/KY: - - ; - - ; - - ; - -
RECENZNÝ FORMÁT*[2]: **2**

Ďakujem za možnosť pripomienkovať tento výstup.

Autori prezentujú profiláciu uchádzačov o zamestnanie (UoZ) založenú na historických dátach použitím moderných metód strojového učenia. Úroveň spracovania je podľa môjho názoru vysoká, či už po obsahovej alebo prezentačnej stránke.

Dovoľte mi zhrnúť niektoré pripomienky, ktoré mám k tomuto textu.

-možno najdôležitejšou pripomienkou je, že pri predikcii sa nevyužíva informácia o dĺžke trvania evidencie. Pre UoZ, ktorému alebo ktorej evidencia trvala 13mesiacov alebo 48mesiacov je priradený rovnaký výstup y=1. Takto sa nejaká časť variácia evidencie nevyužíva. Pri veľkom datasete, aký bol použitý, by možno stálo za snahu urobiť nasledovné:
(1) v prvom kroku predikovať na základe prediktorov dĺžku evidencie (rôznymi prístupmi),
(2) v druhom kroku využiť túto informáciu na kvantifikovanie toho, že táto predikovaná hodnota (spolu s odhadom štatistickej neistoty) prekročí hodnotu 12 mesiacov. Takýto prístup nesie so sebou aj výhody aj nevýhody.
Výhodou je, že využíva viacej informácie. Pomôžem si analógiou s futbalom: štatistický model, ktorý sa pozerá na gólový diferenciál môže predikovať lepšie ako binárny klasifikátor, ktorý nerozlišuje medzi zápasmi 6:0 a 2:1.

Okrem iného by sa týmto by sa otváril priestor pre relatívne ľahko interpretovateľné tradičné regresné metódy. Okrem pravdepodobnosti prekročenia hranice 12mesiacov by mali pracovníci na úrade práce okrem tejto pravdedpodobnosti aj očakávanú dĺžku zamestnanosti a neistotu spojenú s týmto odhadom.
Vieme si predstaviť, že strata variácie v dátach binarizáciou môže byť dôležitejšia ako nárast predikčnej sily získaný použitím pokročilých moderných metód strojového učenia.

- Pripomienka nebola akceptovaná.
- Takýto prístup môže byť skutočne prínosný pri profilácii zameranej čisto na rizikovosť uchádzačov z hľadiska upadnutia do dlhodobej nezamestnanosti. Na rozdiel od použitého prístupu ale neumožňuje kvantifikáciu rizikovosti nenájdenia si uplatnenia do jedného roka v prípadoch, kedy je UoZ síce vyradený z evidencie skôr ako po 12 mesiacoch, ale bez pracovného uplatnenia. Takýchto UoZ tiež považujeme za problémových pre trh práce, a preto uprednostňujeme prístup, ktorý ich umožní identifikovať. V spolupráci s jednou študentkou FMFI UK sme sa v jej diplomovej práci začali venovať aj predikciám dĺžky evidencie a radi pri tom vyskúšame aj navrhovaný prístup.

-z Obrázka 12 sa dá vyčítať, že výkonnosť rôznych modelov je na nerozoznanie. Benefit použitia LightGBM oproti logistickej regresii, čo sa predikcie týka, sa javí ako minimálny (AUC 0.757 vs 0.750). Ak je výpočtová náročnosť dôvodom preferencie LightGBM, potom by možno čitateľ uvítal aspoň jej hrubú kvantifikáciu. Logistická regresia prináša niektoré benefity ako napríklad interpretovateľnosť parametrov. Nie som preto úplne presvedčený o výhodnosti použitia sofistikovanejších metód v tomto prípade.

- Pripomienka bola akceptovaná.
- Odsek 20 o výbere modelu rozšírený o zdôvodnenie voľby LightGBM

Menej dôležité poznámky:
-dodatočným ladením sa zvýšila predikčná sila modelu (Obrázok 13 a 14) - tu mi chýba nejaká (aspoň stručná) diskusia alebo zdôvodnenie, prečo niektoré prediktory boli vynechané z modelu a iné nie.

- Pripomienka bola akceptovaná.
- Doplnená poznámka pod čiarou 16.

-z poznámky pod čiarou 1 vyplýva, že medzi registráciami je viacero UoZ, ktorý sa registrovali viacej krát (viď poznámka 7). Je otázne či je vhodné uvažovať jedného alebo jednu UoZ v modeli viacej ráz.

- Pripomienka bola čiastočne akceptovaná.

- Rozhodli sme sa v modeli nechať viacero zaradení rovnakej osoby, nakoľko nejde o duplicitné pozorovania, keďže mnohé z najdôležitejších premenných sa pri novej registrácii menia (dôvod zaradenia, premenné o pracovnej aktivite za posledné dva roky, vek, miera nezamestnanosti v okrese, údaje o predchádzajúcom zamestnaní a predchádzajúcich evidenciách...). Testovali sme aj modely odhadnuté na dátach bez opakovaných zaradení rovnakých osôb a, s výnimkou poklesu významnosti premenných hovoriacich o predchádzajúcich evidenciách, neprišlo k zásadným zmenám medzi najdôležitejšími premennými.

-v Obrázku 1 možno pozorovať mierny skok medzi 6 a 7 mesiacom. Hypotetizujem, či toto nemôže súvisieť s dĺžkou vyplácania dávok v nezamestnanosti. Je možné, že určitá proporcia UoZ si prácu hľadá len formálne. Ak je ambíciou modelu modelovať schopnosť UoZ nájsť si prácu (a nielen predikovať dĺžku registrácie), takáto podvzorka môže vychyľovať predikcie.

- Pripomienka bola čiastočne akceptovaná.
- Na základe dostupných údajov, žiaľ, nevieme identifikovať, ktorí UoZ si hľadajú prácu naozaj a ktorí len formálne. Dávku v nezamestnanosti (DvN) poberalo len asi 30 % všetkých UoZ v dátovom súbore a z nich znova asi 30 % dávku poberalo 6 mesiacov (9 % všetkých UoZ). Viac ako polovica poberateľov DvN dávku poberala menej ako 6 mesiacov a asi šestina viac ako 6 mesiacov, takže samotná hranica 6 mesiacov netvorí výrazný zlom. Uvažovali sme o skrátení obdobia nezamestnanosti o obdobie poberania DvN, ale to by implicitne predpokladalo, že nikto z poberateľov si počas poberania skutočne nehľadá prácu, čo tiež nepovažujeme za realistický predpoklad. Zo skupiny UoZ, ktorí poberali DvN, sa navyše až 80 % uplatnilo na trhu práce do 12 mesiacov od zaradenia do evidencie. Z uchádzačov, ktorí DvN nepoberali, to bolo len 65 %. To naznačuje, že prípadné obdobie len formálneho hľadania si zamestnania počas poberania DvN by nemal byť významný faktor prispievajúci k rizikovosti dlhodobého pracovného neuplatnenia sa.

-nie je mi jasné, aké boli v dátovom súbore proporcie chýbajúcich pozorovaní a ako sa s nimi autori vysporiadali.

- Pripomienka bola akceptovaná.
- Z dôvodu chýbajúcich údajov o mieste trvalého pobytu alebo o veku uchádzača bolo odstránených asi 100 zaradení do evidencie. Vzhľadom na zanedbateľný počet takýchto pozorovaní sa im v práci nevenujeme. Pri viacerých premenných sme chýbajúce hodnoty zohľadnili ako samostatnú kategóriu (napr. neznámy dôvod zaradenia, neznáme vzdelanie alebo neznáme odvetvie predchádzajúceho zamestnania), nakoľko nevyplnenie údaju môže dávať tiež určitú informáciu o uchádzačovi.

-v Obrázku 14 sa mi v hornej časti krivka presnosti javí ako nediferencovateľná, čo ide proti mojej intuícii

- Pripomienka nebola akceptovaná.
- Pri vysokých hodnotách hraničnej rizikovosti ρ nie je krivka presnosti veľmi pekná a pre $\rho = 1$ dokonca ani nie je presnosť definovaná, lebo žiaden UoZ nie je klasifikovaný ako rizikový. Pri hodnotách blízko jednotky, kedy je označených len zopár UoZ s najvyššou rizikovosťou, môže pri zvýšení hraničnej rizikovosti dochádzať aj k poklesu presnosti. V našom prípade napríklad pri $\rho = 98{,}9$ % máme 7 správne pozitívnych a jedného falošne pozitívneho, čo vedie k presnosti 87,5 %. Pri $\rho = 99{,}0$ % nám počet správne pozitívnych klesne

na 4 a stále máme jedného falošne pozitívneho, čo vedie k presnosti 80,0 %. Pri ρ = 99,1 % už máme len jedného správne pozitívneho a nijakých falošne pozitívnych, takže presnosť je 100 %, a pri vyšších hodnotách hraničnej rizikovosti nie je nikto klasifikovaný ako rizikový. Takéto vysoké hodnoty hraničnej rizikovosti však nepovažujeme za vhodné pri praktickom využití modelu, a preto sa v analýze týmto extrémnym prípadom nevenujeme.

-medzi citovanými zdrojmi som si nevšimol softérovú implementáciu použitých metód strojového učenia alebo SHAP

- Pripomienka bola akceptovaná.
- Pridané odkazy na uverejnené články autorov použitých knižníc, ktoré uvádzajú ako odporúčanú formu citácie.

-oceňujem vyhodnotenie predikčnej schopnosti modelu, kalibráciu parametrov a prezentáciu.
-SHAP je modernou odpoveďou na kvantifikovanie a vizualizovanie vysvetľovacej sily jednotlivých prediktorov pri komplexných modeloch. SHAP vizualizácie podobné Obrázkom 18 a 20 by mohli byť užitočným komplementom pre pracovníkov na Úrade práce. Treba však dať pozor na to, aby neboli interpretované kauzálne - pri vysokej korelácii medzi prediktormi toto môže byť zavádzajúce.

Ďakujem ešte raz za dôveru a možnosť pripomienkovať Váš výstup. V prípade nejasností sa neváhajte na mňa obrátiť.

CELKOVÉ HODNOTENIE (recenzent/ka vyplní túto časť po vysporiadaní sa s pripomienkami analytickou jednotkou):

Ďakujem za reakciu aj upresnenie niektorých nejasností.

Považujem odpoveď za dostatočnú.

[1] Výber medzi: 1. analýza (komplexný analytický materiál s návrhmi konkrétnych systémových opatrení); 2. komentár (rozsahovo menší analytický materiál venujúci sa konkrétnemu čiastkovému problému); 3. manuál (metodické usmernenie vyplývajúce z potreby zjednotenia procesov a postupov v konkrétnej oblasti).

[2] Formát 1 pre komentár/manuál (2 recenzenti bez povinného odborného workshopu); Formát 2 pre analýzu (3 recenzenti a povinný odborný workshop).

[3] Do tabuľky značiť pripomienky zásadného metodologického a obsahového charakteru (nie štylistické či gramatické opravy).

[4] Vyplní analytická jednotka: pripomienka bola akceptovaná / pripomienka nebola akceptovaná a zdôvodnenie / pripomienka bola čiastočne akceptovaná a zdôvodnenie.