TERMÍN: 08.07.2020

xx27107xx
Recenzia B
Ján Komadel
jan.komadel@employment.gov.sk

Prosím nezasahujte do tejto tabuľky

RECENZENT/KA (meno a priezvisko, pozícia, inštitúcia)**: Ján Komadel**
NÁZOV MATERIÁLU: **Searching for gaps: Bottom-up approach for Slovakia**
TYP VÝSTUPU*[1]: **Analýza**
(pri spoločných výstupoch uviesť aj typy individuálnych vkladov):
ANALYTICKÝ ÚTVAR, REZORT: **Ministerstvo financií SR - Inštitút finančnej politiky**
AUTORI/KY: **Marek Chudý, Jaroslav Bukovina, Lucia Šrámková;**
SPOLUAUTORI/KY: - - ; - - ; - - ; - -
RECENZNÝ FORMÁT*[2]: **2**

PRIPOMIENKY:

P.č.	Pripomienka sa vzťahuje k (strana, odsek):	Text pripomienky*[3]	Odôvodnenie pripomienky	Vysporiadanie sa s pripomienkou*[4]
1	Str. 11, ods. 4 (resp. Tab. B na str. 41)	Odvetvie finančných činností je málo zastúpené v dátach.	Podľa tabuľky B je medzi auditovanými spoločnosťami za všetky tri použité roky dokopy len jedna firma z tohto odvetvia. To sotva ponúka dostatok informácií pre zovšeobecniteľné výsledky	**pripomienka nebola akceptovaná** Ide o kategoriálnu premennú, ktorá vstupuje do binary selection modelu (BSM), aby ovplyvnila prípadný výber firmy na audit. Pre tento typ premenných je aj jedno pozorovanie

			pre toto odvetvie a pri modelovaní by snáď bolo vhodnejšie toto odvetvie zlúčiť s nejakým iným. Finančné činnosti tiež ako jediné z používaných odvetví nie sú uvedené v tabuľke 2.	s hodnotou závysle premennej = 1, cenné. Napr. je bežné, že sa do regresných modelov vkladajú dummy premenné na modelovanie špecifických (napr. extrémnych) pozorovaní. Aj v takom prípade má model k dispozícii len 1 pozorovanie, ktoré ale výrazne ovplyvňuje fit. Pre uvedené odvetvie máme teda síce len 1 firmu vybranú na audit, ale zároveň niekoľko firiem z toho istého odvetvia, ktoré neboli auditované. Tie tiež vstupujú do odhadu BSM. Akurát, pre ne má závisle premenná hodnotu 0. Zlúčenie s iným sektorom je vždy subjektívny krok nie je nevyhnutný. Tabuľka 2 (v aktuálnej verzii je to Tabuľka 3) je súčasťou kapitoly 4.3. kde uvádzame len vybrané charakteristiky. V tabuľke nie je tento sektor uvedený, preto, že tento sektor nevstupuje ako premenná do modelu, v ktorom predikujeme výšky nálezov (outcome equation). V tabuľkách v apendixe tento sektor už nechýba.

2	Str. 13, ods. 1-2	Údaje v texte nekorešpondujú s tabuľkou 2.	Na konci prvého odseku sa tvrdí, že veľké firmy majú najvyššiu *noncompliance* *per firm*. Podľa tabuľky 2 je priemer výrazne vyšší pre stredné podniky (a medián ešte výraznejšie). Na konci druhého odseku sa tvrdí, že 33 % únikov je v obchode, pričom podľa tabuľky 2 je to 47 %. V rovnakej vete sa píše, že *noncompliance per firm* je najvyššia v poľnohospodárstve. Ak je tým myslený medián, tak to sedí s tabuľkou, ale nie je to päťkrát vyššie (ako v odvetví obchodu?). Podobne na vrchu strany 29 v bode (iv) sa tvrdí, že zahraničné firmy sa v priemere vyhýbajú dani o viac ako 50 % viac ako domáce firmy. Podľa tabuľky 7 je to o viac ako 82 %.	**pripomienka bola akceptovaná** V Tabuľke 2 (v aktuálnej verzii Tabuľke 3) je údaj správny. V texte bola chyba. Nastala v dôsledku úpravy kritéria pre zaradenie firmy do jednotlivých kategórií podľa veľkosti. Po zohľadnení tržieb (revenues) sa niektoré veľké firmy presunuli z Large do kategórie Medium. Toto sme zabudli zohľadniť v texte. Súhlasíme s pripomienkou, správne má byť 47%. Áno myslíme medián (v súlade s 15. poznámkou pod čiarou aktuálne na str. 15). 5x vyššia bola hodnota mediánu celej populácie (660Eur) voči hodnote v uvedenom sektore. Formuláciu sme upravili tak, aby neboli pochybnosti. Tvrdili sme, že pri zahraničných firmách je výška nedoplatkov o 50% väčšia ako pri domácich firmách. U zahraničných je táto hodnota 211 Eur, pri domácich je to 117 Eur, čo je cca 50%. Formuláciu sme upravili tak, aby nevznikli pochybnosti.

3	Str. 18, ods. 1-2	Skutočne je c = 50 najnižšie také, ktoré zaručuje citlivosť aspoň 50 %?	Z textu vyznieva, že táto hodnota bola vybratá ako najnižšie celé číslo, ktoré zaručí *out-of-sample* citlivosť aspoň 50 %. V tabuľke 3 je ale pri c = 50 citlivosť najmenej 63 % pre rok 2015. Naozaj by to pre c = 49 vyšlo už pod 50 %? Taktiež tvrdenie, že z hľadiska AUC nie je takmer žiadny rozdiel medzi c = 1 a c = 50, nie je veľmi šťastne zvolené. Nárast AUC o niekoľko percentuálnych bodov je štandardne považovaný za výrazné zlepšenie.	**pripomienka bola akceptovaná** Hodnota c=50 bola získaná ako maximum z $\{c_{2014}, c_{2015}, c_{2016}\}$ t. j., z čísiel $\{43, 38, 50\}$ Opravili sme aj tvrdenie o zanedbateľnom zlepšení v súlade s výhradou recenzenta.
4	Sekcia 2	Lineárny model môže byť obmedzujúci.	Použité modely predpokladajú lineárne vzťahy medzi premennými, čo, pochopiteľne, nemusí korešpondovať s realitou. Bolo by prínosné vyskúšať aj modely, ktoré nerobia takéto predpoklady. Autori v [1] napríklad pri odhade *VAT tax gap* používajú dvojkrokový *gradient boosting*, ktorému sa darí lepšie ako klasickému Hackmanovmu modelu.	**pripomienka bola čiastočne akceptovaná** Rozumieme obavám, že použitie lineárnych modelov je reštriktívne. Na druhej strane ide o bežnú prax, ktorá má svoje teoretické aj praktické dôvody. Pre naše účely rozlišujeme predpoklad linearity v binary selection modeli (BSM) a v outcome modeli. Použitím semi-parametrického odhadu BSM modelu sme odstránili predpoklad linearity v tejto časti. Porovnaním

			[1] Tagliaferri, Scacciatelli, Di Loro: *VAT tax gap prediction: a 2-steps Gradient Boosting approach*, preprint, arXiv:1912.03781	výsledku s parametrickým Probit modelom sme zistili, že predikčné kvality modelov sú rovnaké. Pre výstupnú regresiu (outcome) model predpokladá lineárny vzťah. Je faktom, že skutočný DGP nepoznáme, ale je veľmi nepravdepodobné, že by bol lineárny v charakteristikách X. Na základe čoho je však možné konštatovať, že postup zvolený v Tagliaferri et al. je lepší? Podľa nás nič nezaručuje, že dostaneme lepšiu aproximáciu DGP, či lepšiu predikciu out-of-sample. Argumentom pre zachovanie súčasného prístupu podľa nás je: 1. jednoduché metódy často predpovedajú mimo vzorky lepšie, než vysoko sofistikované a parametrizované prístupy (vo financiách to platí napr. pre random-walk model). Predpoklad linearity v nami zvolenom prístupe nevnímame ako chybu, len tým obmedzujeme

				schopnosť modelu fiknúť trénovacie dáta, avšak, na prospech transparentnosti odhadu ako takého. 2. Navrhovaný Gradient boosting má, podobne ako ďalšie moderné nadstavby regresných modelov , metódou, ktorá má veľké praktické uplatnenie v situáciách, keď nie je potrebné vysvetliť, ako sa model k výsledku dopracoval, ale ide len o konečnú presnosť fitu. 3. Problémom týchto prístupov je, že ich teoreticky optimálne vlastnosti sú zaručené technickými a ťažko overiteľnými predpokladmi. V praxi (v R package-och) je potrebné implementovať ich pomocou heuristických algoritmov. Tie závisia na veľkom počte hyperparametrov (napr. počet bootstrap replikácií,

				váhy, ktoré sú prikladané minulým chybám modelu, na báze ktorých sa model učí), ktorých hodnoty treba uhádnuť pomocou trénovacích dát a nie je zaručené, že sú optimálne aj pre out-of-sample. 4. *Apriori* nie je nijak odôvodniteľný výber konkrétnej machin learning metódy. Ak nás obmedzuje konkrétne problém s možnou nelinearitou DGP, existujú desiatky možných tried ensemble metód a ich podtried. Existujú tiež modely neurónových sietí, ktoré sú vhodnejšie na out-of-sample predpovede (Vid. Hastie et al. The Elements of Statistical Learning). Nič však nezaručuje, že ak nejaká konkrétna metóda funguje na konkrétnych dátach lepšie ako iná, bude to platiť aj na naše dáta.ä 5. Kôli závyslosti ich kvalít na konkrétnom type

				predikčného problému a na veľkom množstve hyperparametrov je potrebne metódy navzájom porovnávať a na to je potrebne mat dostatočne veľkú testovaciu vzorku. To však nie je nás prípad. Práca, ktorú recenzent cituje má k dispozícii 18 000 auditov, zatiaľ čo my len cca 1100. 6. Porovnanie záverov Tagliaferri et al, s našou prácou bolo pre nás prínosom, lebo náa motivovalo k zamysleniu sa nad možnými budúcimi prístupmi, ktoré bude možné uplatniť pri aktuálnom odhade medzery na DPH zdola. Avšak, ich práca je nepublikovaná, zrejme neprešla recenzným konaním a má niekoľko nedostatkov. Napr: Niektoré motivačné tvrdenia autorov o nedostatkoch

				Heckmanovho modelu sú veľmi nejasne formulované a v literatúre boli mnohé nedostatky (týkajúce sa napr. striktného predpokladu normality) prekonané. Ďalej, autori sa odkazujú na článok, ktorý aplikoval selection bias correction pre rôzne machinelearning techniky, avšak išlo o klasifikačne algoritmy, ktoré sú sčasti robustné voči bias sample selection a korekcia je potrebná hlavne pri vyhodnocovaní ich presnosti (pomocou MSPE). Nikde však nie je spomenutý gradient boosting a hlavne, nejde o rovnaký setup, lebo ich output je kategoriálna premenná. Stredná hodnota takejto premennej je príslušná pravdepodobnosť, zatiaľ čo u nás (a aj u Tagliaferri et al.) je to reálna veličina, ktorej stredná hodnota je

				kombináciu hodnoty a pravdepodobnosti, t.j., teoretická platnosť korekcie nie je dokázaná pre náš setup. V závere autori porovnávajú predikčné schopnosti modelov pomocou štatistiky R2, čo však nie je bežnou praxou, najmä ak ide o predikciu agregovanej veľkosti noncompliance, nie o individuálne hodnoty. Ďalej autori spomínajú, že v implementácii zvolili váženie jednotlivých weak learners cez bagging, nie boosting. Bagging je naivnejšou verziou ensemble metód, kde model nekoriguje svoje predošlé chyby pomocou váh, preto by sa článok mal volať *VAT tax gap prediction: a 2-steps Bagging approach*

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

Analýzu považujem za kvalitne spracovanú a bez pochýb predstavuje cenný „opačný“ pohľad na daňovú medzeru oproti tradičnému *top-down* prístupu, ktorý okrem iného odhadu celkovej daňovej medzery umožňuje skúmať aj štruktúru daňovej medzery z hľadiska charakteristík individuálnych firiem. Práve toto môže byť mimoriadne nápomocné pri efektívnejšom cielení auditov.

---

[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.