Čierna práca sa nevypláca - komentár

Poučenie z prvého ročníka modelovania výskytu nelegálneho zamestnávania pomocou strojového učenia

Ján Komadel

február 2020

Zhrnutie

Kontroly inšpektorátov práce vykonané u subjektov, ktoré modely strojového učenia ISP označili za podozrivé, potvrdzujú potenciál tohto prístupu pri zefektívnení odhaľovania nelegálneho zamestnávania. Logistická regresia prezentovaná v analýze *Čierna práca sa nevypláca* z júna 2019 dosiahla dvojnásobnú úspešnosť v porovnaní s kontrolami vykonanými v roku 2019. Použitím modernejšej metódy strojového učenia XGBoost sa úspešnosť ešte zvýšila. Samotný model XGBoost prekonal úspešnosť vykonaných kontrol v každom odvetví ekonomiky a celkovo dosiahol 2,5-násobné zlepšenie úspešnosti pri odhaľovaní nelegálnych zamestnávateľov. Kombinácia logistickej regresie a modelu XGBoost viedla k výraznému zúženiu okruhu podozrivých subjektov aj k ďalšiemu zvýšeniu celkovej úspešnosti v odhaľovaní nelegálnych zamestnávateľov.

Poďakovanie

Za konzultácie a cenné rady autor ďakuje Lucii Fašungovej a Štefanovi Domonkosovi (Inštitút sociálnej politiky), Romane Hurtukovej (MPSVR SR) a kolegom z Národného inšpektorátu práce.

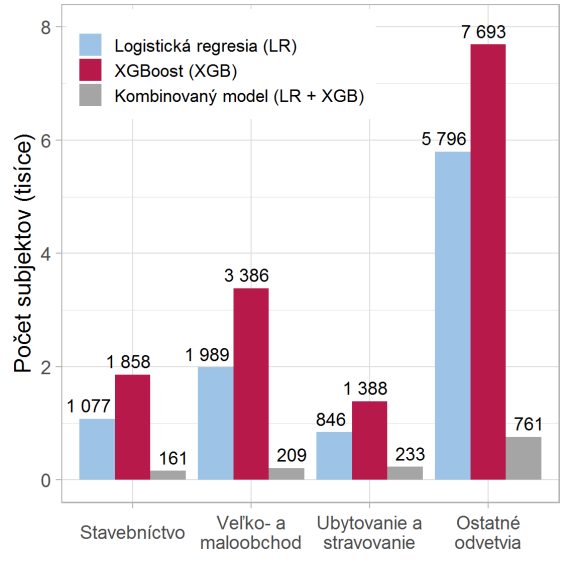
Podozrenia z nelegálneho zamestnávania

V roku 2019 sa ISP v spolupráci s Národným inšpektorátom práce (NIP) začalo venovať problematike cielenia kontrol nelegálneho zamestnávania pomocou administratívnych dát. Aplikovali sme moderné metódy strojového učenia, ktoré na základe dostupných údajov o minulých kontrolách nelegálneho zamestnávania, o počte a štruktúre pracovníkov, o ekonomickej a finančnej situácii subjektov vyhodnocujú, ktoré subjekty sú rizikové z hľadiska pravdepodobnosti, že nelegálne zamestnávajú.

Obr. 1 Doterajšie výstupy ISP k nelegálnemu zamestnávaniu

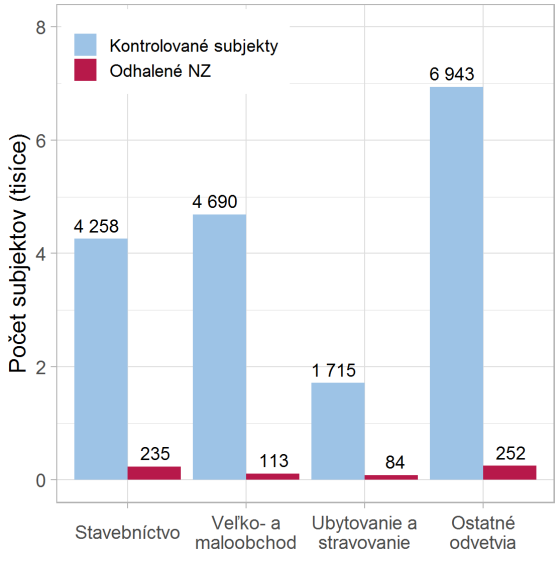
Prvým výsledkom bol zoznam subjektov podozrivých z nelegálneho zamestnávania zhotovený na základe logistickej regresie, ktorý bol poskytnutý NIP v apríli 2019. Použitý klasifikačný model, vstupné údaje a metodika boli v júni 2019 opísané v analýze *Čierna práca sa nevypláca*. V septembri 2019 bol s použitím rovnakých vstupných dát vytvorený nový klasifikačný model XGBoost. Tento model bol využitý na zredukovanie pôvodného zoznamu podozrivých subjektov na tie subjekty, ktoré boli označené za podozrivé obidvomi modelmi a zoznam bol opäť poskytnutý NIP.[1]

Obr. 2 Podozrenia podľa jednotlivých modelov



Zdroj: vlastné spracovanie

Obr. 3 Kontroly inšpektorátov práce v roku 2019



Zdroj: NIP

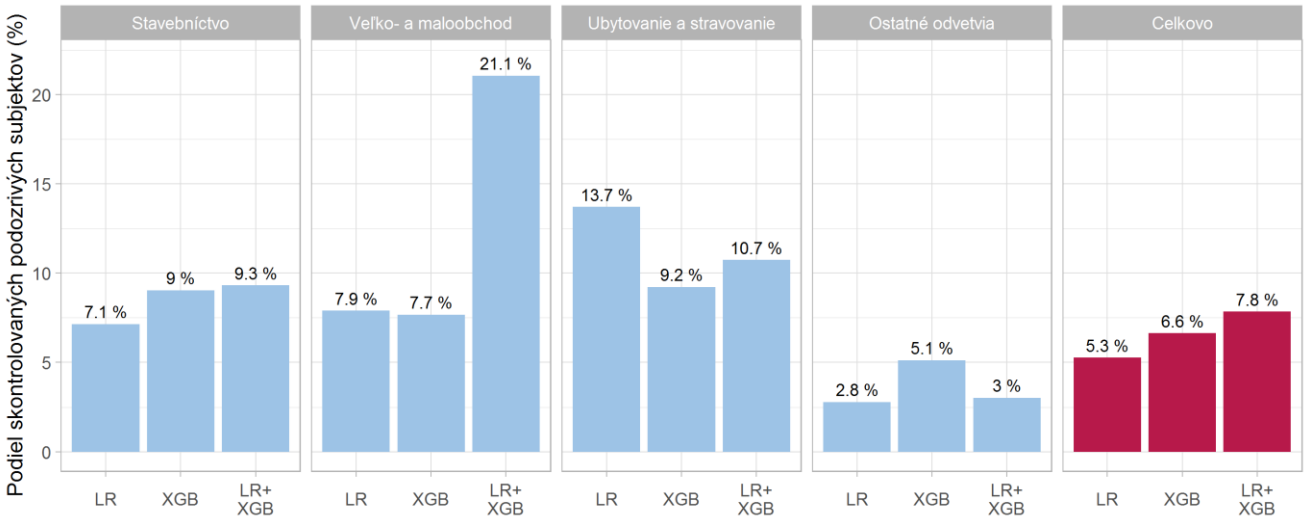
Kým pôvodný zoznam obsahoval 9 708 podozrivých subjektov a XGBoost označil až 14 325 subjektov, zredukovaný zoznam už obsahoval len 1 364 subjektov (Obr. 2). Pre porovnanie, inšpektoráty práce vykonali v roku 2019 kontroly u 17 606 rôznych subjektov[2] a odhalili pritom 684 nelegálne zamestnávajúcich subjektov

[1] Spomedzi podozrivých subjektov boli vylúčené aj tie, ktoré už pravdepodobne neboli aktívne podľa informácií ako vymazanie z Obchodného registra, vypísanie konkurzu alebo reštrukturalizácie, prítomnosť vysokých dlhov (voči Finančnej správe, Sociálnej poisťovni alebo zdravotným poisťovniam) alebo nepodávanie daňových priznaní.

[2] Rôzne subjekty sú chápané ako subjekty s jedinečným platným prideleným IČO v SR.

(Obr. 3). Z kontrolovaných subjektov bolo 61 % z troch rizikových odvetví – stavebníctva, veľkoobchodu a maloobchodu a ubytovacích a stravovacích služieb, na ktoré sa zvlášť zameriavali aj použité modely. Spomedzi odhalených nelegálne zamestnávajúcich subjektov bolo z týchto odvetví 63 %.

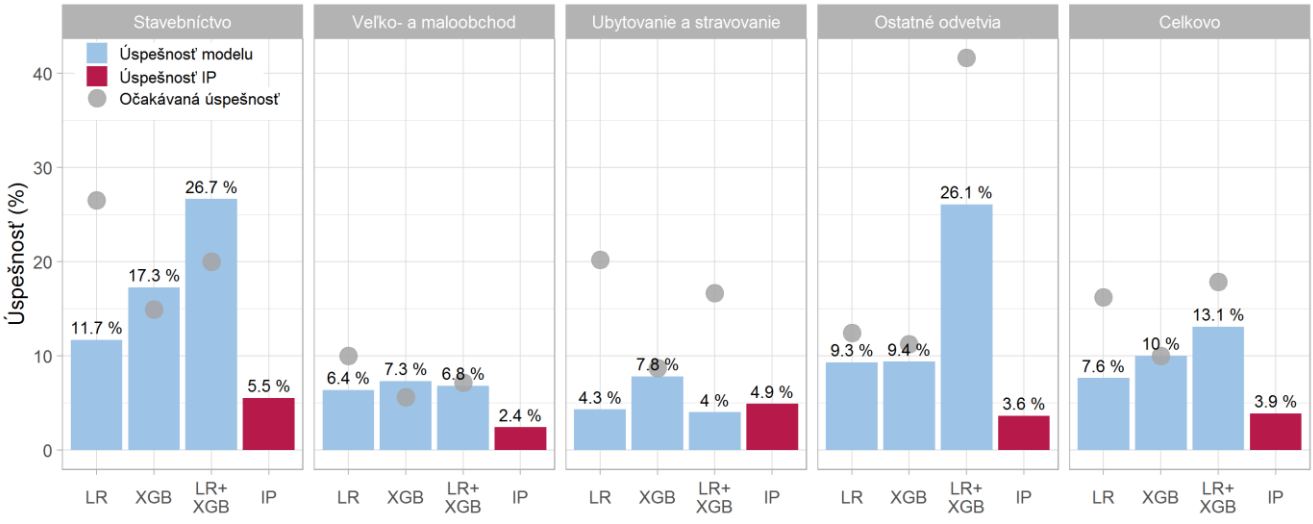
Obr. 4 Z podozrivých subjektov bolo pre jednotlivé modely celkovo skontrolovaných 5 až 8 %



Zdroj: NIP, vlastné spracovanie

Z pôvodného zoznamu takmer 10 tisíc podozrivých subjektov zostaveného podľa logistickej regresie bolo v roku 2019 skontrolovaných 511 subjektov, čo predstavuje 5,3 % (Obr. 4). Zo subjektov podozrivých podľa modelu XGBoost bolo skontrolovaných 950 (6,6 %) a z kombinovaného zoznamu to bolo 107 subjektov (7,8 %). Pre všetky modely bola najmenšia časť podozrivých subjektov skontrolovaná v nerizikových odvetviach, aj keď absolútne počty skontrolovaných podozrivých subjektov tam najnižšie neboli, ale bolo tu najviac podozrení (Obr. 2).

Obr. 5 Úspešnosť modelov a kontrol inšpektorátov práce v roku 2019



Zdroj: NIP, vlastné spracovanie

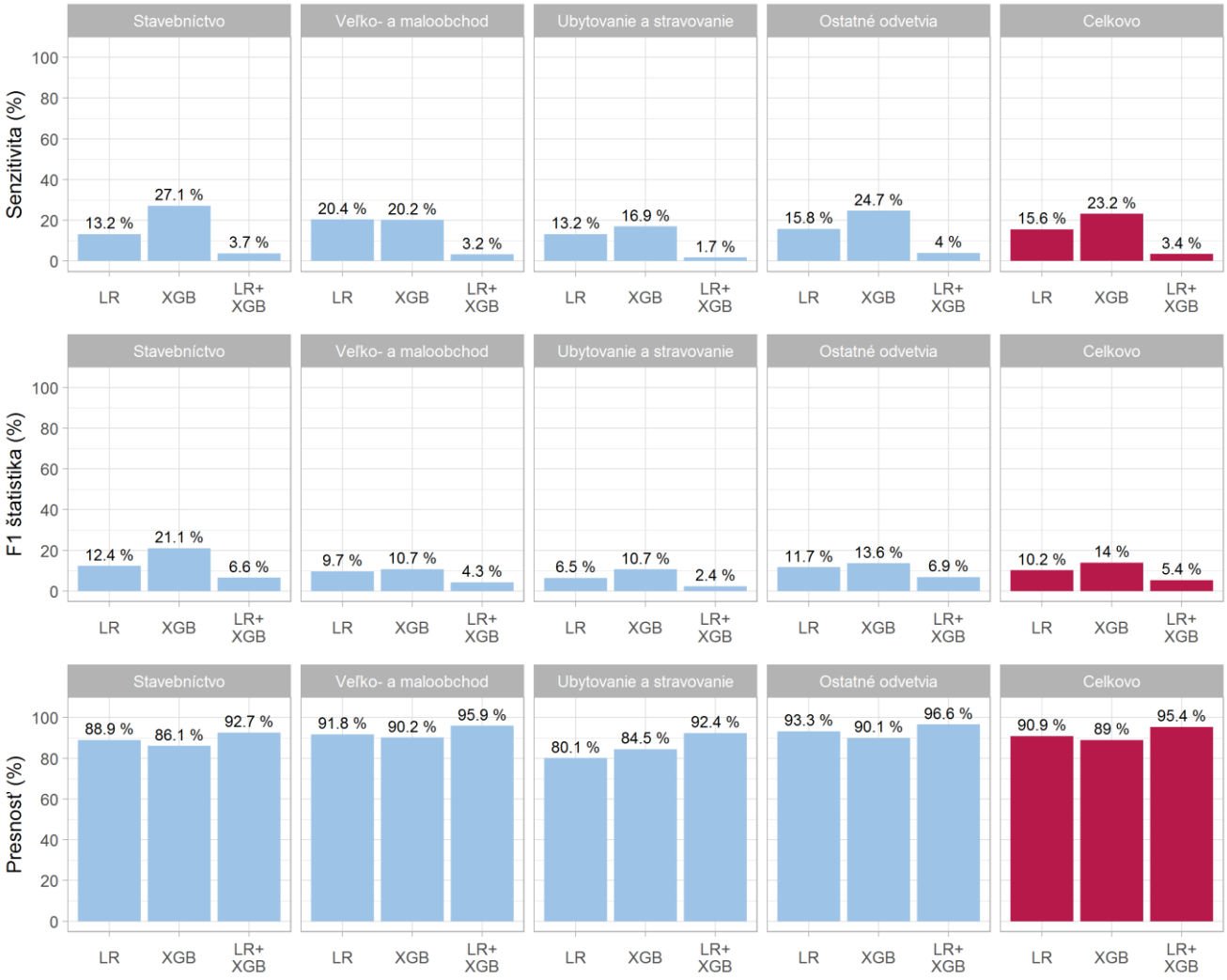
Hodnotenie úspešnosti modelov

Úspešnosť modelov pri identifikácii nelegálne zamestnávajúcich subjektov bola dvoj- až trojnásobná oproti úspešnosti kontrol inšpektorátov práce, ktoré celkovo zistili nelegálne zamestnávanie u 3,9 % kontrolovaných subjektov (Obr. 5). Logistická regresia dosiahla v porovnaní s kontrolami inšpektorátov práce dvojnásobnú

úspešnosť,[3] aj keď dosiahnutá miera úspešnosti bola nižšia, ako sa odhadovalo v analýze *Čierna práca sa nevypláca*. Najnižšiu úspešnosť tento model dosiahol v odvetví ubytovacích a stravovacích služieb, kde bol horší ako kontroly inšpektorátov práce. Nenaplnenie očakávanej úspešnosti môže súvisieť aj s legislatívnymi zmenami platnými od roku 2019, ktoré zjednodušujú legálne zamestnávanie občanov tretích krajín.[4] Tieto zmeny sa prejavili medziročným poklesom počtu odhalených nelegálne zamestnávaných občanov tretích krajín o 57 %.

Model XGBoost aj kombinácia logistickej regresie a XGBoost boli úspešnejšie ako pôvodná logistická regresia. XGBoost dosiahol celkovú úspešnosť 10 %, ktorá sa zhoduje s ex-ante odhadom úspešnosti a viac ako 2,5-násobne prevyšuje úspešnosť kontrol inšpektorátov práce. Ako jediný z modelov XGBoost prekonal kontroly inšpektorátov práce aj v ubytovacích a stravovacích službách, ale tiež bol v tomto odvetví najmenej úspešný. Celkovo najúspešnejší bol kombinovaný model s celkovou úspešnosťou 13,1 %. Tento model dosiahol v stavebníctve a nerizikových odvetviach výrazne vyššiu úspešnosť ako ostatné modely alebo kontroly inšpektorátov práce, ale v ubytovaní a stravovaní bola jeho úspešnosť naopak najnižšia.

Obr. 6 Iné miery úspešnosti klasifikácie



Zdroj: vlastné spracovanie

[3] Úspešnosť modelu vyjadrujeme ako pozitívnu prediktívnu hodnotu (PPH), t. j. podiel tých podozrivých subjektov, kde bolo zistené nelegálne zamestnávanie, na všetkých subjektoch, ktoré model označil za podozrivé a boli kontrolované.

[4] 1. januára 2019 nadobudla účinnosť novela zákona č. 5/2004 Z. z. o službách zamestnanosti aj novela zákona č. 404/2011 Z. z. o pobyte cudzincov.

Pozitívna prediktívna hodnota (PPH) sleduje, aká časť podozrivých subjektov sa ukáže ako skutočne nelegálne zamestnávajúce. Inou mierou úspešnosti klasifikácie je napríklad senzitivita, ktorá meria, aká časť skutočne nelegálne zamestnávajúcich subjektov bola modelom označená ako podozrivé. V tejto miere nad ostatnými modelmi dominuje XGBoost, ktorý označil za podozrivé takmer štvrtinu nelegálne zamestnávajúcich subjektov (Obr. 6). Naopak, kombinovaný model, ktorý mal najvyššiu PPH, je ďaleko najhorší so senzitivitou na úrovni troch percent. Tieto výsledky súvisia s tým, že XGBoost označil za podozrivé viac ako 10-násobne viac subjektov ako kombinovaný model (Obr. 2).

Pri sledovaní PPH a senzitivity súčasne pomocou F_1 štatistiky[5] je tiež najúspešnejší XGBoost a kombinovaný model je najmenej úspešný. Naopak, pri presnosti klasifikácie je výsledok najlepší pre kombinovaný model, ktorý správne klasifikuje[6] viac ako 95 % subjektov, kým XGBoost má presnosť len na úrovni 89 %. Výhodou úzkeho zoznamu podozrivých subjektov z kombinovaného modelu je menšia záťaž na kontrolné kapacity. Naopak, pomerne široký zoznam z modelu XGBoost zachytí medzi podozreniami väčšiu časť nelegálne zamestnávajúcich subjektov.

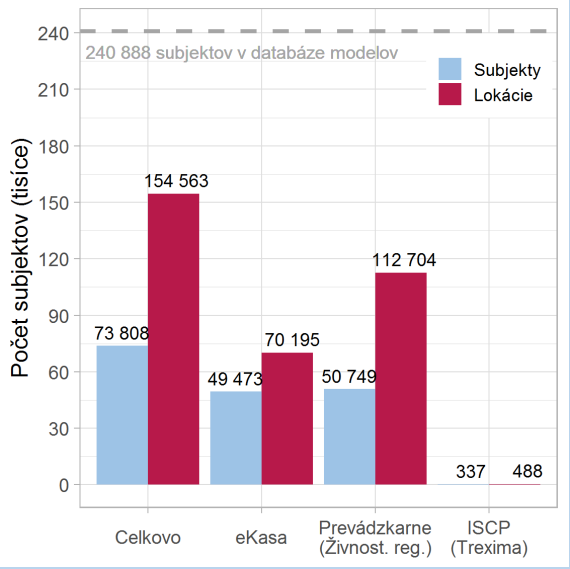
Box 1 Kde sa firmy naozaj nachádzajú?

Výzvou pri praktickom využití výstupov z modelov pri kontrolách nelegálneho zamestnávania je slabá informácia o tom, kde zamestnanci subjektu reálne vykonávajú svoju činnosť. Miesto výkonu práce by od roku 2023 malo byť pre každého zamestnanca evidované Sociálnou poisťovňou[7], ale v súčasnosti takáto informácia chýba.

V **Obchodnom registri** je evidovaná adresa sídla subjektu, ktorá ale nezriedka slúži len na prijímanie korešpondencie. V **Živnostenskom registri** sa evidujú aj adresy prevádzkarní, ale tie má len časť subjektov (asi pätina subjektov z našej databázy, Obr. 7). Ďalšia informácia o adresách organizačných jednotiek subjektov je v **Informačnom systéme ceny práce**, ktorý spravuje firma Trexima, s. r. o., ale v tejto databáze je zastúpená len určitá vzorka subjektov obsiahnutá vo výberovom zisťovaní.

Iným možným zdrojom informácie o tom, kde subjekty prevádzkujú svoju činnosť je evidencia registračných pokladníc v systéme **eKasa**, na ktorý od 1. júla 2019 musia byť napojení všetci podnikatelia. Systém eKasa síce nie je primárne určený na evidenciu prevádzok subjektov, ale poskytuje určitú informáciu o adresách, kde subjekt vykonáva činnosť. Špecifické je stavebnícke odvetvie, v ktorom činnosť z princípu nie je realizovaná v prevádzkach subjektu, čo cielené kontroly podozrivých subjektov komplikuje.

Obr. 7 Adresa z iného zdroja ako sídlo subjektu v OR SR je známa pre 30 % subjektov z databázy



Zdroj: Finančná správa SR, Živnostenský register, Trexima

[5] Je definovaná ako ich harmonický priemer, formálna definícia napr. v analýze *Čierna práca sa nevypláca*, Box 3.
[6] Pod správnou klasifikáciou rozumieme skutočne nelegálne zamestnávajúce subjekty klasifikované ako podozrivé a subjekty, ktoré boli kontrolované bez zistenia nelegálneho zamestnávania, ako nepodozrivé.
[7] Od 1. januára. 2023 nadobúda účinnosť § 232a zákona č. 461/2003 Z. z. o sociálnom poistení, podľa ktorého budú evidované analytické údaje zamestnancov.

Záver

Použité modely majú potenciál prispieť k zefektívneniu odhaľovania nelegálneho zamestnávania, ako jeden z faktorov pri výbere subjektov, ktoré majú byť kontrolované. Pri subjektoch označenými pôvodnou logistickou regresiou za podozrivé je podiel skutočne nelegálne zamestnávajúcich subjektov dvojnásobný oproti úspešnosti kontrol inšpektorátov práce v roku 2019. Pri modeli XGBoost je to viac ako 2,5-násobok a pri kombinovanom modeli takmer 3,5-násobok.

Má zmysel používať aj modernejšie metódy ako logistická regresia, ktorá bola použitá v analýze *Čierna práca sa nevypláca*. Napriek tomu, že pri žiadnom z modelov nebolo skontrolovaných viac ako 8 % podozrivých subjektov, dá sa hodnotiť ich úspešnosť na tejto vzorke, aj keď skutočnú úspešnosť modelov by bolo možné posúdiť len v prípade skontrolovania celého zoznamu. Model XGBoost, prípadne jeho kombinácia s logistickou regresiou porazili pôvodný model v ukazovateľoch pozitívnej prediktívnej hodnoty, senzitivity aj v presnosti klasifikácie. Tiež sa ukázalo ako dôležité z hľadiska efektívnosti kontrol odstrániť spomedzi podozrivých subjektov tie, ktoré javia známky neaktivity.

Problémom pri využití zoznamov subjektov podozrivých z nelegálneho zamestnávania pri kontrolnej činnosti ostáva **slabá informácia o mieste výkonu práce zamestnancov**. Čiastočne (pre asi 30 % subjektov z našej databázy) je možné túto informáciu, ktorá by mala byť od roku 2023 evidovaná Sociálnou poisťovňou pre každého zamestnanca, doplniť pomocou údajov zo Živnostenského registra a z evidencie registračných pokladníc (Box 1). Výzvou zostáva aj nastavenie rovnováhy medzi kapacitami inšpektorátov a rozsahom zoznamu v zmysle pokrytia významnej časti nelegálne zamestnávajúcich subjektov.

Literatúra

Aggarwal, C. C. (2015). *Data Mining: The Textbook.* Cham: Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, (s. 785–794). San Francisco, CA, USA.

Komadel, J. (2019). Čierna práca sa nevypláca: Cielenie kontrol nelegálneho zamestnávania s využitím administratívnych dát. *Inštitút sociálnej politiky*. Dostupné na Internete: https://www.employment.gov.sk/sk/ministerstvo/vyskum-oblasti-prace-socialnych-veci-institut-socialnej-politiky/analyticke-komentare/cierna-praca-sa-nevyplaca.html

Národný inšpektorát práce. (2020). *Informatívna správa o vyhľadávaní a potieraní nelegálnej práce a nelegálneho zamestnávania za rok 2019.* V príprave.

Príloha – úspešnosť modelov

Tab. 1 Úspešnosť - logistická regresia

Stavebníctvo

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	9	68	77	**11.7 %**
nepodozrivý	59	1 013	1 072	94.5 %
spolu	68	1 081	1 149	NPH
	13.2 % senzitivita	93.7 % špecificita		

Ubytovanie a stravovanie

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	5	111	116	**4.3 %**
nepodozrivý	33	575	608	94.6 %
spolu	38	686	724	NPH
	13.2 % senzitivita	83.8 % špecificita		

Veľkoobchod a maloobchod

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	10	147	157	**6.4 %**
nepodozrivý	39	2 077	2 116	98.2 %
spolu	49	2 224	2 273	NPH
	20.4 % senzitivita	93.4 % špecificita		

Ostatné odvetvia

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	15	146	161	**9.3 %**
nepodozrivý	80	3 132	3 212	97.5 %
spolu	95	3 278	3 373	NPH
	15.8 % senzitivita	95.5 % špecificita		

Tab. 2 Úspešnosť - XGBoost

Stavebníctvo

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	29	139	168	**17.3 %**
nepodozrivý	78	1 310	1 388	94.4 %
spolu	107	1 449	1 556	NPH
	27.1 % senzitivita	90.4 % špecificita		

Ubytovanie a stravovanie

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	10	118	128	**7.8 %**
nepodozrivý	49	903	952	94.9 %
spolu	59	1 021	1 080	NPH
	16.9 % senzitivita	88.4 % špecificita		

Veľkoobchod a maloobchod

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	19	241	260	**7.3 %**
nepodozrivý	75	2 896	2 971	97.5 %
spolu	94	3 137	3 231	NPH
	20.2 % senzitivita	92.3 % špecificita		

Ostatné odvetvia

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	37	357	394	**9.4 %**
nepodozrivý	113	4 235	4 348	97.4 %
spolu	150	4 592	4 742	NPH
	24.7 % senzitivita	92.2 % špecificita		

Tab. 3 Úspešnosť - logistická regresia + XGBoost

Stavebníctvo

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	4	11	15	**26.7 %**
nepodozrivý	103	1 438	1 541	93.3 %
spolu	107	1 449	1 556	NPH
	3.7 % senzitivita	99.2 % špecificita		

Ubytovanie a stravovanie

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	1	24	25	**4.0 %**
nepodozrivý	58	997	1 055	94.5 %
spolu	59	1 021	1 080	NPH
	1.7 % senzitivita	97.6 % špecificita		

Veľkoobchod a maloobchod

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	3	41	44	**6.8 %**
nepodozrivý	91	3 115	3 206	97.2 %
spolu	94	3 156	3 250	NPH
	3.2 % senzitivita	98.7 % špecificita		

Ostatné odvetvia

Klasifikácia	Odhalené NZ: áno	Odhalené NZ: nie	spolu	PPH
podozrivý	6	17	23	**26.1 %**
nepodozrivý	144	4 575	4 719	96.9 %
spolu	150	4 592	4 742	NPH
	4.0 % senzitivita	99.6 % špecificita		