inštitút sociálnej politiky
www.employment.gov.sk/isp

Ministerstvo práce, sociálnych vecí a rodiny SR

# Čierna práca sa nevypláca

## Cielenie kontrol nelegálneho zamestnávania s využitím administratívnych dát

**Ján Komadel**

**jún 2019**

## Abstrakt

Ekonomické údaje o zamestnávateľoch sa v súčasnosti nevyužívajú pri cielení kontrol nelegálneho zamestnávania. Kým počet vykonávaných kontrol má rastúci trend, odhalených subjektov je každoročne menej. Navrhujeme metódu strojového učenia, ktorá umožňuje odhaliť vyšší počet nelegálne zamestnávajúcich subjektov pri tretinovom počte kontrol. Model dosahuje najvyššiu úspešnosť v stavebníctve a ubytovacích a stravovacích službách, kde sú úspešnejšie aj súčasné kontroly inšpektorátov práce. Vyššie podozrenie z nelegálneho zamestnávania majú zamestnávatelia so sídlom v bratislavskom kraji, dlhmi, vyšším počtom pracovníkov na dohodu a menším počtom zamestnancov. Vyššia účinnosť kontrol podporuje ochranu zamestnancov, šetrí verejné financie a pozitívne prispieva ku kvalite podnikateľského prostredia.

isp

## Obsah



## Poďakovanie

Za konzultácie a cenné rady autor ďakuje Lucii Fašungovej, Marekovi Plavčanovi a Štefanovi Domonkosovi (Inštitút sociálnej politiky), Karolovi Habinovi, Miroslave Mošonovej a Stanislave Balážovej (Národný inšpektorát práce), Romane Hurtukovej (MPSVR SR), Rastislavovi Gábikovi a Michalovi Beličkovi (Finančná správa), Tomášovi Hellebrandtovi (Útvar hodnoty za peniaze), Martinovi Hulényimu (Inštitút pre stratégie a analýzy), Róbertovi Tóthovi (Tangent Works) a ďalším spolupracovníkom. Za poskytnuté dáta autor ďakuje Národnému inšpektorátu práce, Sociálnej poisťovni, Ústrediu práce, sociálnych vecí a rodiny a spoločnosti FinStat, s. r. o.

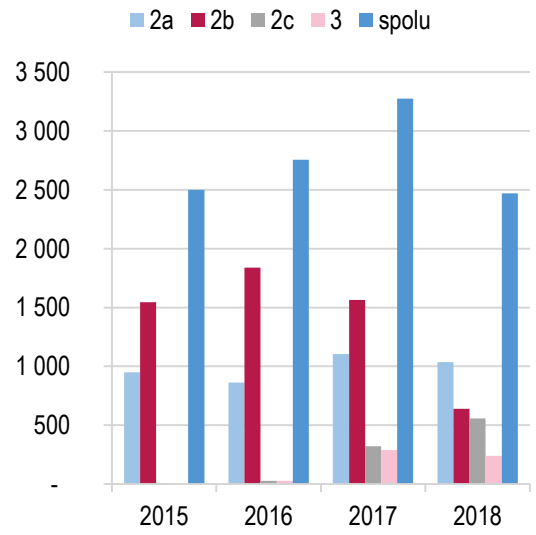
## 1 Čo je nelegálne zamestnávanie?

Slovenská legislatíva[1] pozná štyri druhy nelegálneho zamestnávania. Dva sú spojené so zamestnávaním občanov SR alebo iného členského štátu EÚ[2] a ostatné dva sú spojené so zamestnávaním štátnych príslušníkov tretích krajín:

- 2a zamestnávanie fyzickej osoby bez založenia pracovnoprávneho vzťahu alebo štátnozamestnaneckého pomeru,
- 2b zamestnávanie fyzickej osoby bez prihlásenia do registra poistencov a sporiteľov starobného dôchodkového sporenia[3],
- 2c zamestnávanie štátneho príslušníka tretej krajiny, ktorý má povolenie na iný účel pobytu,
- 3 zamestnávanie štátneho príslušníka tretej krajiny, ktorý nemá povolenie na pobyt.

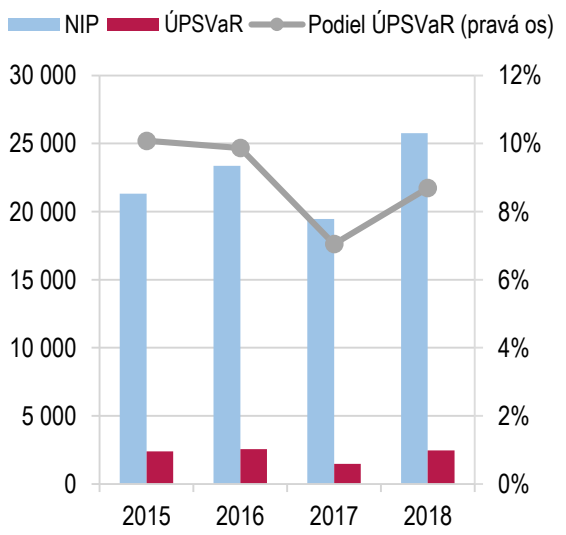
**Za posledné štyri roky klesol podiel nelegálneho zamestnávania typu 2b (neprihlásenie do Sociálnej poisťovne) z takmer dvoch tretín na štvrtinu (Obr. 1).** Jedným z dôvodov poklesu v roku 2018 je, že pred týmto rokom bolo zamestnávanie nelegálne, ak zamestnanec nebol prihlásený do registra Sociálnej poisťovne najneskôr do dňa začatia vykonávania práce. Od roku 2018 je lehota na prihlásenia zamestnanca predĺžená do siedmich dní od začatia vykonávania práce. Druhým dôvodom poklesu podielu nelegálneho zamestnávania typu 2b je výrazný nárast nelegálneho zamestnávania občanov tretích krajín (typy 2c a 3, Obr. 1) od roku 2017.

**Obr. 1 Odhalené prípady nelegálneho zamestnávania podľa druhov**




Zdroj: NIP

**Obr. 2 Vykonané kontroly dodržiavania zákazu nelegálneho zamestnávania**




Zdroj: NIP, ÚPSVaR

**Deväť z desiatich kontrol dodržiavania zákazu nelegálneho zamestnávania vykonajú inšpektoráty práce.** Okrem nich vykonávajú kontroly aj Ústredie práce, sociálnych vecí a rodiny a úrady práce, sociálnych vecí a rodiny, pričom podiel kontrol vykonaných Ústredím a úradmi práce bol v posledných štyroch rokoch od 7 % do 10 % (Obr. 2).[4] V tejto práci sa zameriavame na kontroly inšpektorátov práce, pre ktoré je kontrola dodržiavania zákazu nelegálneho zamestnávania jednou z priorít[5] a ktoré vykonávajú prevažnú väčšinu kontrolnej činnosti. Inšpektoráty práce pri výbere kontrolovaných subjektov vychádzajú z doterajších skúseností alebo z podnetov, avšak nevyužívajú ekonomické údaje o zamestnávateľoch.

[1] Podľa § 2 odsekov 2 a 3 zákona č. 82/2005 Z. z. o nelegálnej práci a nelegálnom zamestnávaní a o zmene a doplnení niektorých zákonov.

[2] Prípadne občanov iných zmluvných štátov Dohody o Európskom hospodárskom priestore alebo Švajčiarska.

[3] Povinnosť prihlásiť zamestnanca do tohto registra Sociálnej poisťovne je do siedmich dní od začatia vykonávania práce. V prípade kontroly nelegálnej práce a nelegálneho zamestnávania, ktorá začala menej ako sedem dní od začatia vykonávania práce, musí byť osoba prihlásená do tohto registra najneskôr do začatia kontroly, inak ide o nelegálne zamestnávanie.

[4] Všetky zobrazené údaje okrem Obr. 2 sa vzťahujú len ku kontrolám inšpektorátov práce.

[5] Svedčí o tom aj zriadenie oddelení kontroly nelegálneho zamestnávania – KOBRA spomínaných v časti 1.1.

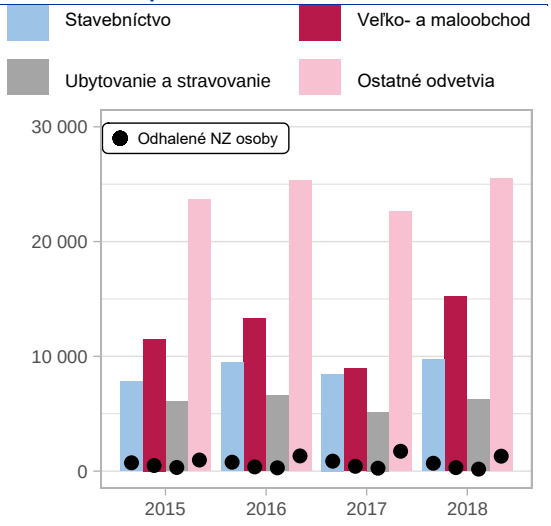
**Box 1 Nelegálna práca a nedeklarovaná práca**

Slovenská legislatíva pracuje s presne definovanými pojmami **nelegálne zamestnávanie** a **nelegálna práca**[6]. Nelegálna práca je závislá práca fyzickej osoby bez založenia pracovnoprávneho vzťahu alebo štátnozamestnaneckého pomeru alebo závislá práca štátneho príslušníka tretej krajiny, ktorý má povolenie na iný účel pobytu. Nelegálna práca je teda spojená s nelegálnym zamestnávaním typu 2a a 2c.

Európska komisia, hlavne prostredníctvom Európskej platformy na riešenie problému nelegálnej práce, používa pojem **nedeklarovaná práca** (*undeclared work*), ktorý definuje ako „každú platenú činnosť, ktorá je zákonná, pokiaľ ide o jej charakter, ale nie je oznámená verejným orgánom, pričom sa berú do úvahy rozdiely v regulačnom systéme členských štátov“.[7] Nedeklarovaná práca zahŕňa nelegálnu prácu a rovnako aj prácu vykonávanú pri ostatných typoch nelegálneho zamestnávania. Navyše zahŕňa aj iné činnosti ako **čiastočne nedeklarovanú prácu** (*under-declared work*), kedy je časť mzdy vyplácaná oficiálne a časť neoficiálne „na ruku“, vykonávanie závislej práce na živnosť alebo poskytovanie tovarov alebo služieb bez dokladu.

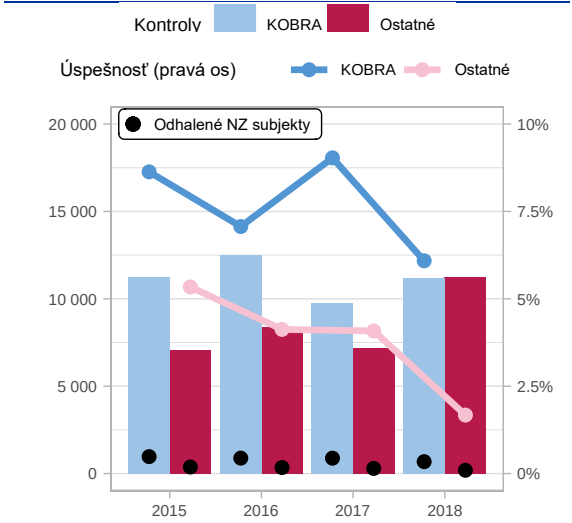
**Viac ako polovica všetkých fyzických osôb kontrolovaných na nelegálne zamestnávanie pracuje v rizikových odvetviach.** Stavebníctvo, veľkoobchod a maloobchod a ubytovacie a stravovacie služby považuje Národný inšpektorát práce za rizikové odvetvia v rámci nelegálneho zamestnávania. V rokoch 2015 - 2018 bola vždy viac ako polovica všetkých kontrolovaných osôb práve z týchto odvetví (Obr. 3). Podiel odhalených nelegálne zamestnávaných osôb z rizikových odvetví na všetkých odhalených nelegálne zamestnávaných osobách v tomto období postupne klesol zo 61 % na 47 %.

**Obr. 3 Osoby kontrolované na nelegálne zamestnávanie podľa odvetví**




Zdroj: NIP

**Obr. 4 Úspešnosť kontrol OKNZ – KOBRA je vyššia ako pri ostatných kontrolách**



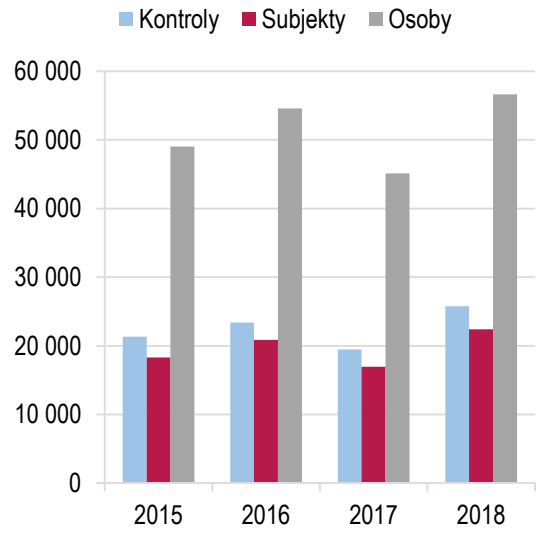

Zdroj: NIP

## 1.1 Úspešnosť odhaľovania nelegálneho zamestnávania

**Inšpektoráty práce zvyšujú počet kontrolovaných subjektov, zatiaľ čo počet odhalených nelegálne zamestnávajúcich subjektov každoročne klesá (Obr. 5 a Obr. 6).** Jedným spôsobom merania úspešnosti kontrol dodržiavania zákazu nelegálneho zamestnávania je porovnanie počtu kontrolovaných subjektov a počtu odhalených nelegálne zamestnávajúcich subjektov. Postupný nárast počtu kontrolovaných subjektov spojený s poklesom odhalených nelegálne zamestnávajúcich subjektov viedol k poklesu úspešnosti odhaľovania zo 7,4 % v roku 2015 na 3,9 % v roku 2018.

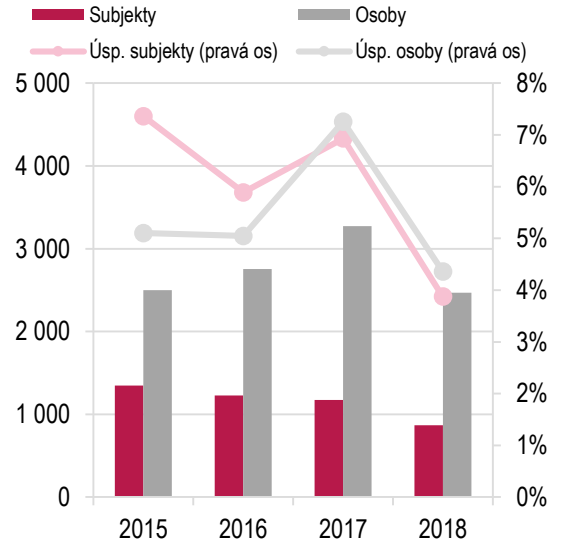
[6] Podľa § 2 zák. č. 82/2005 Z. z. o nelegálnej práci a nelegálnom zamestnávaní a o zmene a doplnení niektorých zákonov.
[7] Európska komisia: https://ec.europa.eu/social/main.jsp?catId=1298&langId=en

**Obr. 5 Vykonané kontroly, kontrolované subjekty a kontrolované osoby**




Zdroj: NIP

**Obr. 6 Nelegálne zamestnávajúce subjekty, nelegálne zamestnávané osoby a úspešnosť odhaľovania**




Zdroj: NIP

**Úspešnosť kontrol OKNZ – KOBRA je vyššia ako úspešnosť ostatných kontrol a rozdiel sa zvyšuje.** V roku 2013 začali na inšpektorátoch práce pôsobiť oddelenia kontroly nelegálneho zamestnávania KOBRA (OKNZ – KOBRA)[8]. Špecifikom OKNZ – KOBRA je výhradné zameranie na nelegálne zamestnávanie a nasadenie aj v neštandardných časoch. V rokoch 2015 – 2018 postupne klesol podiel OKNZ – KOBRA na všetkých kontrolách nelegálneho zamestnávania zo 61 % na 50 %. Kontroly OKNZ – KOBRA sú úspešnejšie v odhaľovaní nelegálne zamestnávajúcich subjektov oproti ostatným kontrolám a v danom období podiel úspešností kontrol OKNZ – KOBRA oproti ostatným kontrolám vzrástol z 1,6-násobku na 3,6-násobok (Obr. 4).

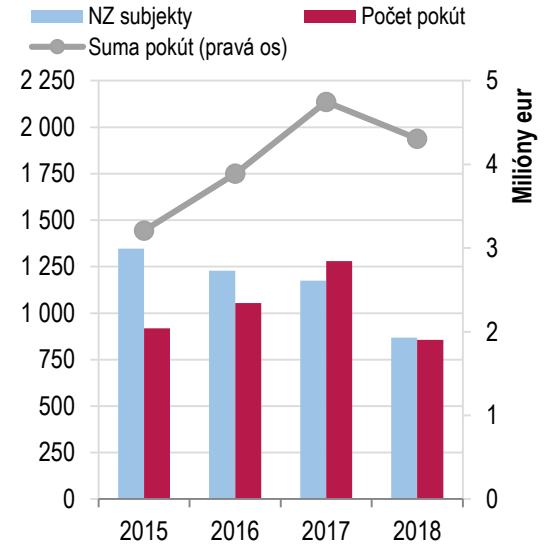
## 1.2 Uložené pokuty

**Celková výška právoplatne uložených pokút nelegálne zamestnávajúcim subjektom rastie napriek tomu, že počet odhalených subjektov klesá.** Za porušenie zákazu nelegálneho zamestnávania ukladá inšpektorát práce odhalenému subjektu pokutu od 2 000 eur do 200 000 eur.[9] Medzi rokmi 2015 až 2017 vzrástol počet uložených pokút o 39 % a celková suma uložených pokút o 48 %, zatiaľ čo počet odhalených nelegálne zamestnávajúcich subjektov klesol o 13 % (Obr. 7). V roku 2018 prišlo k medziročnému poklesu počtu uložených pokút o 33 %, ktorý bol ale spojený s poklesom celkovej sumy uložených pokút len o 9 %. Dôvodom menšej zmeny sumy uložených pokút boli zmeny v štruktúre nelegálneho zamestnávania. Počet odhalení menej závažného nelegálneho zamestnávania typu 2b[10] klesol v dôsledku legislatívnych zmien a naopak, vzrástol počet odhalení nelegálneho zamestnávania štátnych príslušníkov tretích krajín (typy 2c a 3), ktoré sú postihované prísnejšie (Obr. 1).

**Výška uloženej pokuty pre subjekt nie je vždy úmerná počtu odhalených nelegálne zamestnávaných osôb.** Do roku 2016 boli počty odhalených nelegálne zamestnávaných osôb v subjektoch rôznej veľkosti porovnateľné a rovnako aj udeľované pokuty. Od roku 2017 začali inšpektori odhaľovať viac osôb u veľkých zamestnávateľov. V roku 2018 prispôsobili výšku pokuty počtu odhalených osôb, čo zmenšilo rozdiely v ukladaných pokutách prepočítaných na jednu nelegálne zamestnávanú osobu naprieč kategóriami podnikov (Obr. 8). V prepočte na jednu osobu sú pokutami najviac zaťažené subjekty s 10 – 49 zamestnancami.

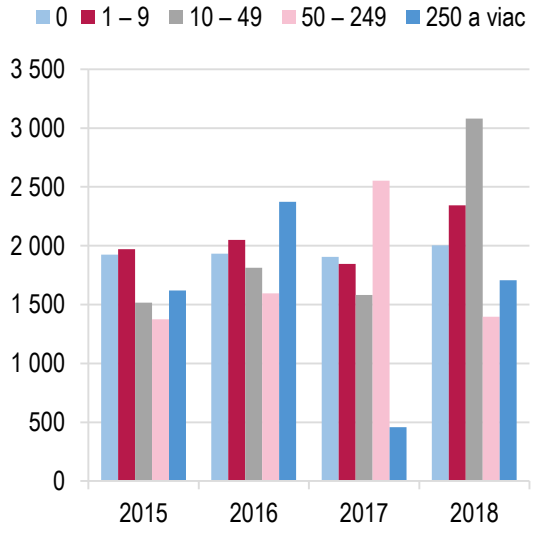
[8] Do 9. 10. 2017 to boli útvary kontroly nelegálneho zamestnávania KOBRA.
[9] Podľa § 19 ods. 2 zákona č. 125/2006 Z. z. o inšpekcii práce a o zmene a doplnení zákona č. 82/2005 Z. z. o nelegálnej práci a nelegálnom zamestnávaní a o zmene a doplnení niektorých zákonov. V prípade nelegálneho zamestnávania viac ako jednej osoby je pokuta najmenej 5 000 eur.
[10] Neprihlásenie zamestnanca do registra Sociálnej poisťovne v stanovenom termíne.

**Obr. 7 Odhalené nelegálne zamestnávajúce subjekty a právoplatne uložené pokuty**




Zdroj: NIP

**Obr. 8 Priemerná výška uloženej pokuty na jednu odhalenú NZ osobu podľa počtu zamestnancov**

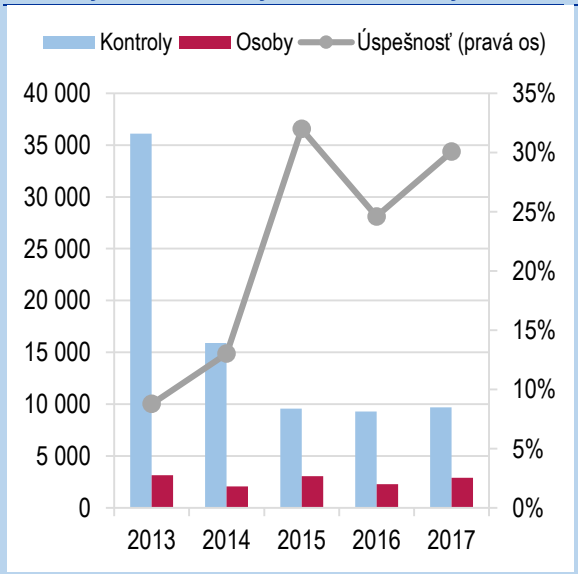



Zdroj: NIP

## Box 2 Inšpirácia z Českej republiky

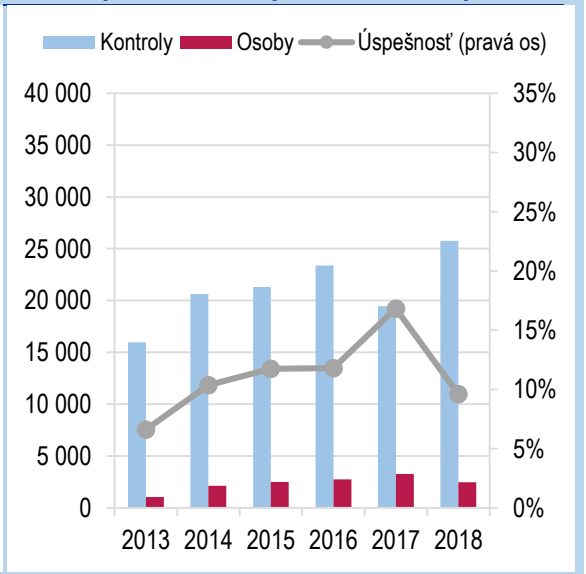
Kým slovenské inšpektoráty zvyšujú počty kontrol nelegálneho zamestnávania (Obr. 10), v Českej republike pristúpili k opačnému kroku. Medzi rokmi 2013 a 2015 klesol počet kontrol v ČR o tri štvrtiny (Obr. 9). Tento pokles bol spojený s nárastom úspešnosti kontrol z hodnôt okolo 10 % na trojnásobok[11]. V rokoch 2015 – 2017 bol v ČR v porovnaní so SR vykonaný menej ako polovičný počet kontrol, ale počty odhalených nelegálne zamestnávaných osôb sa výrazne nelíšia[12].

**Obr. 9 Vykonané kontroly a odhalené osoby v ČR**




Zdroj: SÚIP

**Obr. 10 Vykonané kontroly a odhalené osoby v SR**




Zdroj: NIP

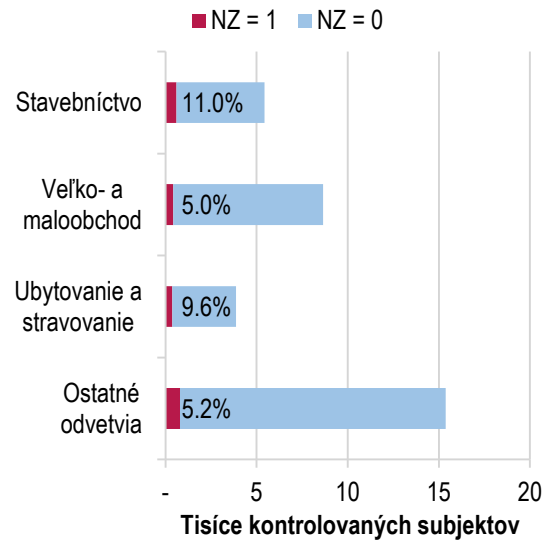
[11] V tejto časti používame inú mieru úspešnosti kontrol, a to podiel odhalených nelegálne zamestnávaných osôb vzhľadom na počet vykonaných kontrol, lebo tieto údaje zverejňuje český Štátny úrad inšpekcie práce v ročných súhrnných správach.
[12] Česká legislatíva sa pri nelegálnej práci odlišuje od slovenskej legislatívy. Preto priame porovnávanie počtu odhalených osôb nie je korektné, ale myšlienka zníženia počtu kontrol spojeného s ich lepším cielením môže byť inšpiráciou aj pre slovenské kontrolné orgány.

## 2 Odhaľovanie nelegálneho zamestnávania s využitím dát

Využitie údajov o zamestnávateľoch by mohlo zlepšiť cielenie kontrol nelegálneho zamestnávania a zvýšiť ich úspešnosť.[13] Databáza, ktorú používame na identifikáciu podozrivých subjektov, pozostáva z údajov o minulých kontrolách (NIP), o počte a štruktúre pracovníkov (Sociálna poisťovňa, ÚPSVaR), o hospodárskych výsledkoch subjektov z finančných výkazov a o finančných ukazovateľoch (FinStat). Bližší popis použitých dát je v prílohe A.

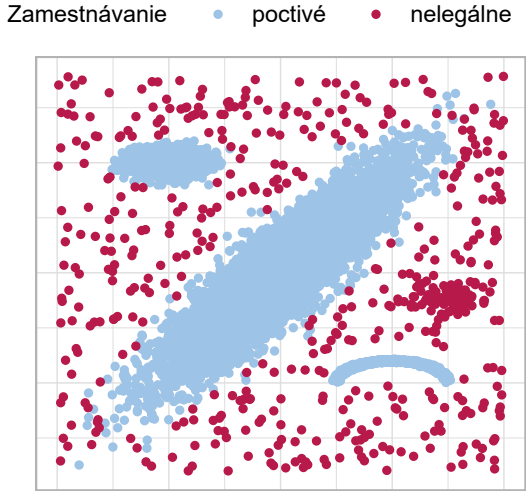
**Identifikácia subjektov podozrivých z nelegálneho zamestnávania nie je typickým príkladom vyhľadávania anomálií ani dobre definovaným klasifikačným problémom.** Metódy na **vyhľadávanie anomálií** sú určené na prípad, že nelegálne zamestnávajúce subjekty sú zriedkavé[14] medzi všetkými kontrolovanými subjektami. Naopak, **klasifikačné metódy** sú určené na prípad, že zastúpenie nelegálne zamestnávajúcich subjektov a poctivo zamestnávajúcich subjektov je vyrovnané. V našej databáze je celkové zastúpenie nelegálne zamestnávajúcich subjektov 6,6 %, pričom najvyšší podiel je v stavebníctve (11 %) a najnižší vo veľkoobchode a maloobchode (5 %, Obr. 11). Preto sme na databáze stavebníckych firiem testovali oba prístupy.

**Obr. 11 Podiel odhalených nelegálne zamestnávajúcich subjektov na kontrolovaných subjektoch**




Zdroj: NIP

**Obr. 12 Ilustrácia anomálií v simulovaných dvojrozmerných dátach**




Zdroj: vlastné spracovanie

**Výsledky testovania siedmich metód[15] vyhľadávania anomálií naznačujú, že väčšina nelegálne zamestnávajúcich subjektov nie je anomáliami vzhľadom na pozorované charakteristiky.** Použitie metód určených na vyhľadávanie anomálií predpokladá, že poctivo zamestnávajúce subjekty sa v mnohorozmernom priestore charakteristík dostupných v našej databáze nachádzajú v zhlukoch, teda že sa svojimi charakteristikami na seba podobajú. Na druhej strane, nelegálne zamestnávajúce subjekty sa svojimi charakteristikami od nich odlišujú a v spomínanom mnohorozmernom priestore sa preto nachádzajú vzdialené od poctivých subjektov (Obr. 12). Vyhľadávanie anomálií ako možný prístup k identifikácii podozrivých subjektov spomína aj Európska platforma na riešenie problému nelegálnej práce (De Wispelaere, a iní, 2017), ale v našich testoch sa vyhľadané anomálie výrazne nezhodujú s nelegálne zamestnávajúcimi subjektami (Obr. 13).

[13] Európska platforma na riešenie problému nelegálnej práce tiež nabáda štáty EÚ k využívaniu dolovania dát na identifikáciu rizikových subjektov – viď napr. (Karaboev, Mineva, & Stefanov, 2019) alebo (De Wispelaere, a iní, 2017).
[14] Často je podiel anomálií menej ako 1 %.
[15] Metódy vyhľadávania anomálií sú stručne opísané v prílohe B.

**Spomedzi testovaných metód[16] pri identifikácii subjektov podozrivých z nelegálneho zamestnávania najlepšie obstála logistická regresia.** Celkovo boli klasifikačné metódy úspešnejšie ako vyhľadávanie anomálií a najvyššiu hodnotu $F_1$ štatistiky dosiahla logistická regresia a jej modifikácia Lasso (Obr. 13).

**Po dodatočnom výbere premenných ostane v modeli 18 – 29 premenných,** v závislosti od odvetvia.[17] Použitú databázu, pôvodne obsahujúcu viac ako 180 vysvetľujúcich premenných, opisuje Box 4 v prílohe A. Vyššia úspešnosť metódy Lasso indikuje, že niektoré z vysvetľujúcich premenných sú v modeli nepotrebné. Pred ďalším použitím logistickej regresie na jednotlivé odvetvia preto aplikujeme dodatočný výber premenných založený na opakovanom náhodnom vynechaní 5 % subjektov z tréningovej vzorky, odhadnutí modelu a zaznamenaní, ktoré z premenných sú signifikantné na hladine významnosti 20 %.[18]

**Box 3 Miery úspešnosti klasifikácie**

Prirodzenou mierou úspešnosti metódy je **správnosť klasifikácie** (*accuracy*), teda aká časť subjektov je klasifikovaná správne – nelegálne zamestnávajúce ako podozrivé a poctivo zamestnávajúce ako nepodozrivé. V značení matice zámen (*confusion matrix*, Tab. 1) je správnosť klasifikácie vyjadrená pomerom správnych pozitívov a správnych negatívov k celkovému počtu pozorovaní.

V databáze kontrolovaných subjektov je však nepomer v zastúpení nelegálne a poctivo zamestnávajúcich subjektov, čo má za dôsledok, že metóda, ktorá označí všetky subjekty ako nepodozrivé, dosiahne správnosť klasifikácie vyššiu ako 93 %, ale bude úplne zbytočná. V takýchto prípadoch je preto vhodné použiť iné miery úspešnosti.

**Pozitívna prediktívna hodnota** (PPH) hovorí, aká časť subjektov klasifikovaných ako podozrivé je skutočne nelegálne zamestnávajúcich. **Senzitivita** vyjadruje, aká časť skutočne nelegálne zamestnávajúcich subjektov je označená ako podozrivé. **$F_1$ štatistika** sa definuje ako harmonický priemer pozitívnej prediktívnej hodnoty a senzitivity

$$F_1 = \left( \frac{PPH^{-1} + senzitivita^{-1}}{2} \right)^{-1} = \frac{2 \times SP}{2 \times SP + FP + FN}.$$

Metóda, ktorá dosahuje vysokú hodnotu $F_1$ štatistiky teda musí mať vysokú PPH (veľká časť subjektov označených ako podozrivé je skutočne nelegálne zamestnávajúcich) aj vysokú senzitivitu (veľká časť skutočne nelegálne zamestnávajúcich subjektov je označená ako podozrivé). Dodatočné miery úspešnosti – kumulatívne zisky a lift – uvádzame v prílohe B.

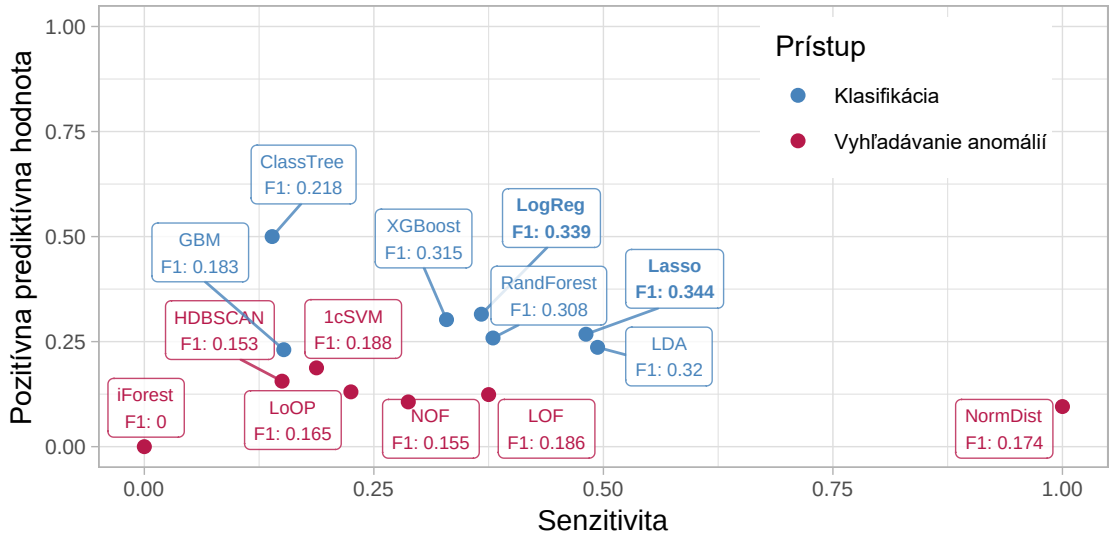
**Tab. 1 Schéma matice zámen**

| | | Odhalené NZ | | |
|---|---|---|---|---|
| | | áno | nie | **Pozitívna prediktívna hodnota (PPH)** |
| **Klasifikácia** | podozrivý | správny pozitív ($SP$) | falošný pozitív ($FP$) | $\boldsymbol{SP/(SP+FP)}$ |
| | nepodozrivý | falošný negatív ($FN$) | správny negatív ($SN$) | $SN/(SN+FN)$ |
| | | $\boldsymbol{SP/(SP+FN)}$ **senzitivita** | $SN/(SN+FP)$ špecificita | Negatívna prediktívna Hodnota (NPH) |

Pozn.: Pozitívna prediktívna hodnota (*positive predictive value*) sa v angličtine nazýva aj *precision*.Senzitivita (*sensitivity*) sa nazýva aj *recall* alebo *true positive rate*. Špecificita (*specificity*) sa nazýva aj *selectivity* alebo *true negative rate*.

[16] Sedem testovaných klasifikačných metód je stručne opísaných v prílohe D.
[17] Zoznam použitých vysvetľujúcich premenných obsahuje Tab. 4 v prílohe A.
[18] V modeli ponecháme premenné, ktoré sú signifikantné vo viac ako polovici zo 400 opakovaní.

**Obr. 13 Klasifikačné metódy sú pri identifikácii stavebníckych subjektov podozrivých z nelegálneho zamestnávania úspešnejšie ako metódy určené na vyhľadávanie anomálií**




Zdroj: vlastné spracovanie. Pozn.: použité metódy sú stručne opísané v prílohách B a D.

## 3 Výsledky

Na modelovanie miery podozrenia $i$-teho subjektu z nelegálneho zamestnávania $p_i$ v jednotlivých odvetviach[19] používame logistickú regresiu

$$ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i} + rok_{2017}\left(\sum_{j\in J}\beta_{n+j}x_{j,i}\right),$$

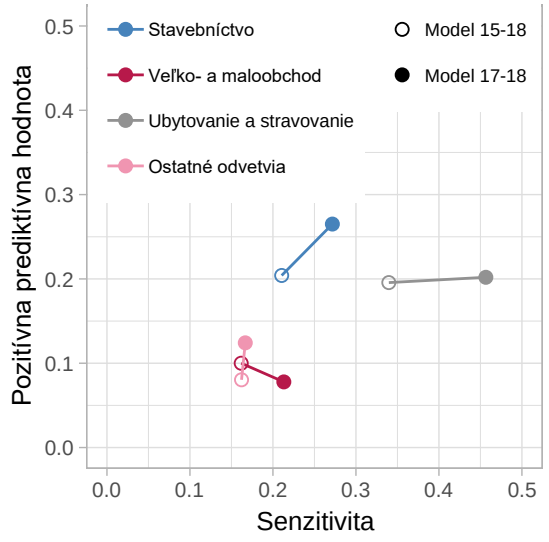
kde $\beta$ sú koeficienty a $x_1, \ldots, x_n$ sú vysvetľujúce premenné. Model obsahuje aj interakcie premenných s indikátorom roku 2017, ktoré zohľadňujú možné zmeny vplyvu vysvetľujúcich premenných na mieru podozrenia v čase.[20] Množina indexov $J \subset \{1, \ldots, n\}$ zodpovedá premenným interagujúcim s rokom v Tab. 4 v prílohe A.

**Úspešnosť modelu hodnotíme porovnaním s výsledkami kontrol v roku 2018.** Natrénovaný model priradí subjektu v danom roku mieru podozrenia z nelegálneho zamestnávania a aj určí hraničnú mieru podozrenia, od ktorej je subjekt klasifikovaný ako podozrivý.[21] Kvôli praktickému využitiu výsledkov pri cielení kontrol nelegálneho zamestnávania vypočítame **celkovú mieru podozrenia subjektu** ako priemer z mier podozrenia v jednotlivých rokoch.[22] Subjekty, ktorých celková miera podozrenia je nad hraničnou mierou, sú klasifikované ako podozrivé.

**Len vo veľkoobchode a maloobchode je výhodné využívať aj staršie údaje.** Z dôvodu zmien v štruktúre nelegálneho zamestnávania v posledných rokoch (Obr. 1) sme okrem modelov využívajúcich všetky údaje z rokov 2015 – 2018 testovali aj modely pracujúce len s údajmi z rokov 2017 a 2018. V modeloch pre stavebníctvo, ubytovacie a stravovacie služby a nerizikové odvetvia vynechanie starších údajov zvýšilo úspešnosť modelu (Obr. 14).

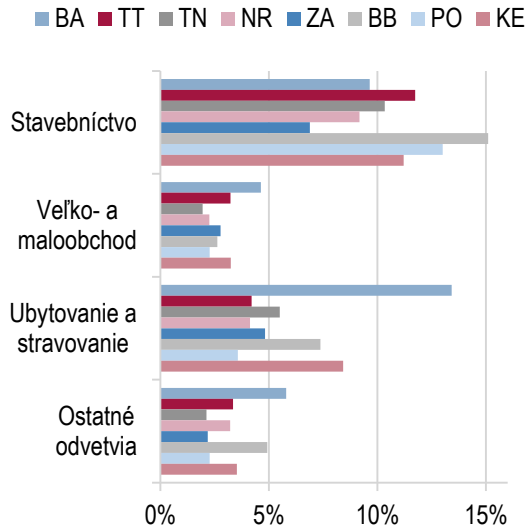
[19] Model aplikujeme zvlášť na každé z troch rizikových odvetví a na ostatné (nerizikové) odvetvia.
[20] Tento model bol použitý pre stavebníctvo, ubytovanie a stravovanie a nerizikové odvetvia, kde sme pracovali s databázou údajov z rokov 2017 – 2018. Pre veľkoobchod a maloobchod bol použitý model pracujúci s údajmi z rokov 2015 – 2018, v ktorom boli navyše zhrnuté aj interakcie s rokmi 2015 a 2016.
[21] Hraničná miera podozrenia je vybraná tak, aby maximalizovala $F_1$ štatistiku na validačnej vzorke (Príloha A).
[22] Kvôli zohľadneniu aktuálnej situácie dostane miera podozrenia z roku 2018 v priemere dvojnásobnú váhu.

**Obr. 14 Pre všetky odvetvia okrem veľkoobchodu a maloobchodu je úspešnejší model, ktorý pracuje len s rokmi 2017 a 2018**




Zdroj: vlastné spracovanie

**Obr. 15 Podiely nelegálne zamestnávajúcich subjektov na všetkých subjektoch kontrolovaných v roku 2018 podľa krajov**




Zdroj: NIP

## 3.1 Prediktory nelegálneho zamestnávania

**Subjekty sídliace v bratislavskom kraji sú najpodozrivejšie.** S výnimkou stavebníckych subjektov z banskobystrického kraja model priraďuje najvyššiu mieru podozrenia z nelegálneho zamestnávania subjektom z bratislavského kraja (Obr. 16a). V odvetví ubytovacích a stravovacích služieb sú nitriansky a prešovský kraj výrazne menej rizikové ako ostatné kraje. Obe tieto pozorovania sú v súlade s výsledkami kontrol v roku 2018 (Obr. 15).

**Subjekty s dlhmi sú podozrivejšie.** Prítomnosť dlhov voči štátu vstupuje do modelov pre všetky rizikové odvetvia a zvyšuje mieru podozrenia (Obr. 16b). Pre nerizikové odvetvia má kladný vplyv na mieru podozrenia prítomnosť dlhov voči Sociálnej poisťovni. Zaujímavosťou je záporný vplyv tohto ukazovateľa na mieru podozrenia v stavebníctve. V tomto odvetví je tiež prítomný kladný vplyv dlhov voči Všeobecnej zdravotnej poisťovni.

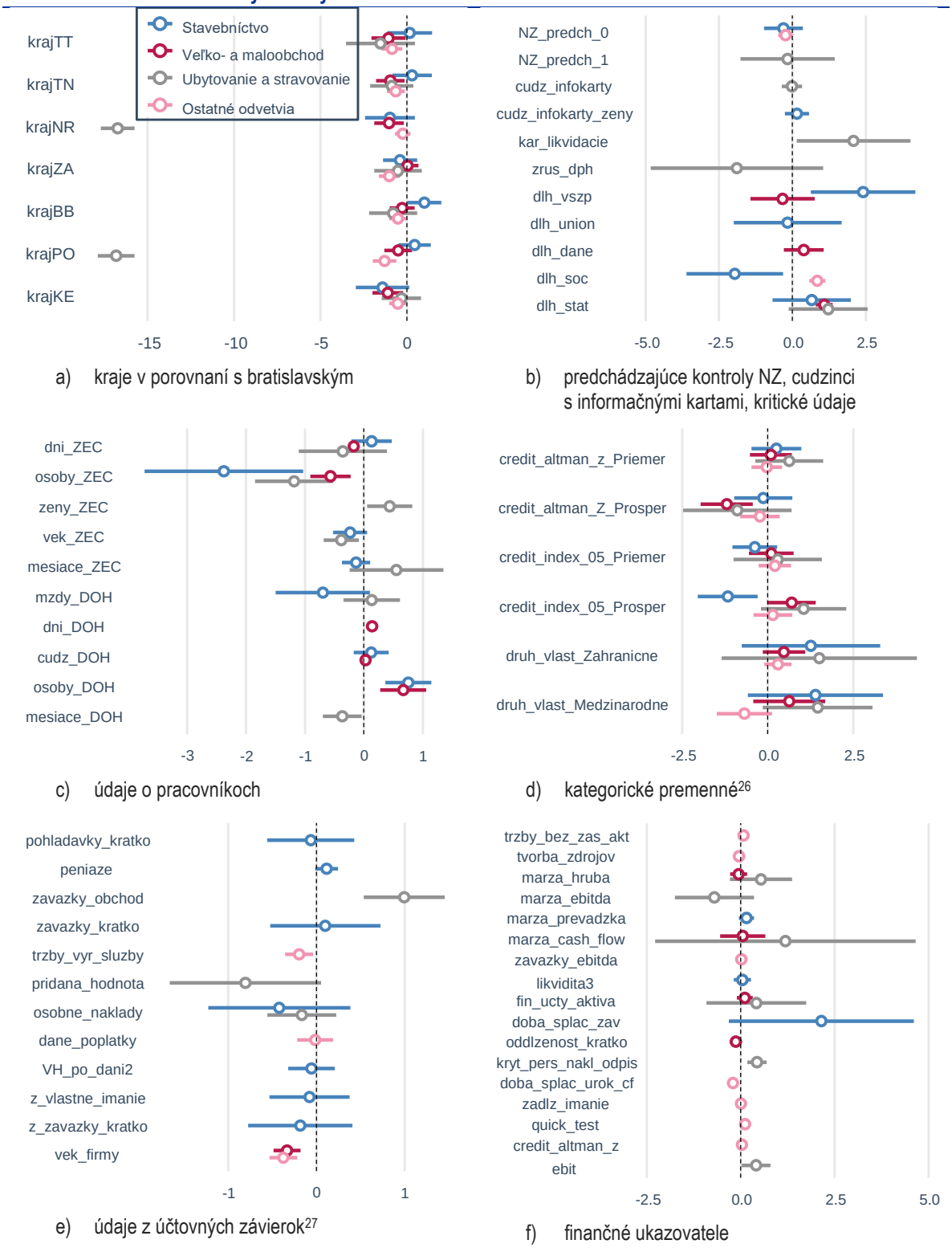
**Podozrivejšie sú subjekty s nižším počtom zamestnancov a s vyšším počtom pracovníkov na dohodu.** V ubytovacích a stravovacích službách sú navyše podozrivejšie subjekty, ktoré zamestnávajú vyšší podiel žien, zamestnávajú mladších zamestnancov a zamestnávajú pracovníkov na dohodu iba časť roku (Obr. 16c). Vo veľkoobchode a maloobchode sú podozrivejšie subjekty, ktoré majú zamestnancov na menej dní v mesiaci a pracovníkov na dohodu na viac dní.

**Domáce subjekty sú menej podozrivé ako zahraničné a medzinárodné** (Obr. 16d). Vo veľkoobchode a maloobchode aj v nerizikových odvetviach modely navyše priraďujú vyššiu mieru podozrenia mladším subjektom (Obr. 16e).

**Viaceré premenné prekvapivo neovplyvňujú výslednú mieru podozrenia z nelegálneho zamestnávania.** Patrí sem indikátor odhaleného nelegálneho zamestnávania v predchádzajúcom roku[23], premenné hovoriace o cudzincoch medzi pracovníkmi[24] alebo informácia o mzdách, ktoré subjekt vypláca.[25]

[23] Vystupuje v modeli pre ubytovanie a stravovanie, ale nemá signifikantný vplyv.
[24] Premenné s údajmi z informačných kariet, resp. premenná hovoriaca o pomere cudzincov medzi pracovníkmi na dohodu boli vybrané do modelov pre rizikové odvetvia, ale ich vplyv nie je významný.
[25] V modeli pre stavebníctvo je prítomný vplyv nízkych miezd pre pracovníkov na dohodu.

**Obr. 16 Odhadnuté koeficienty a ich významnosť**




a) kraje v porovnaní s bratislavským

b) predchádzajúce kontroly NZ, cudzinci s informačnými kartami, kritické údaje

c) údaje o pracovníkoch

d) kategorické premenné[26]

e) údaje z účtovných závierok[27]

f) finančné ukazovatele

Zdroj: vlastné spracovanie. Pozn.: Koeficienty sú pre lepšiu porovnateľnosť škálované. Zobrazená štatistická signifikancia je na hladine významnosti 20 % pre robustné smerodajné odchýlky.

[26] Pre premenné *credit_altman_z* a *credit_index_05* je referenčná kategória „Neprosperujúca“. Pre druh vlastníctva je referenčná kategória „Domáce".

[27] Z obrázku je kvôli prehľadnosti vynechaný koeficient zmeny výrobnej spotreby, ktorý vystupuje len v modeli pre ubytovacie a stravovacie služby a je výrazne záporný a signifikantný.

**Tab. 2 Úspešnosť modelov vzhľadom na výsledky kontrol nelegálneho zamestnávania v roku 2018**

**Stavebníctvo**

| | | Odhalené NZ áno | Odhalené NZ nie | spolu | PPH |
|---|---|---|---|---|---|
| Klasifikácia | podozrivý | 22 | 61 | 83 | **26.5 %** |
| | nepodozrivý | 59 | 1 181 | 1 240 | 95.2 % |
| | spolu | 81 | 1 242 | | NPH |
| | | **27.2 % senzitivita** | 95.1 % špecificita | | |

| | | | |
|---|---|---|---|
| Celkový počet subjektov: | 1 323 | Správnosť klasifikácie: | 90.9 % |
| $F_1$ štatistika: | 26.8 % | McFadden pseudo $R^2$: | 0.2345 |

**Ubytovanie a stravovanie**

| | | Odhalené NZ áno | Odhalené NZ nie | spolu | PPH |
|---|---|---|---|---|---|
| Klasifikácia | podozrivý | 21 | 83 | 104 | **20.2 %** |
| | nepodozrivý | 25 | 695 | 720 | 96.5 % |
| | spolu | 46 | 778 | | NPH |
| | | **45.7 % senzitivita** | 89.3 % špecificita | | |

| | | | |
|---|---|---|---|
| Celkový počet subjektov: | 824 | Správnosť klasifikácie: | 86.9 % |
| $F_1$ štatistika: | 28.0 % | McFadden pseudo $R^2$: | 0.3149 |

**Veľkoobchod a maloobchod**

| | | Odhalené NZ áno | Odhalené NZ nie | spolu | PPH |
|---|---|---|---|---|---|
| Klasifikácia | podozrivý | 11 | 99 | 110 | **10.0 %** |
| | nepodozrivý | 57 | 2 383 | 2 440 | 97.7 % |
| | spolu | 68 | 2 482 | | NPH |
| | | **16.2 % senzitivita** | 96.0 % špecificita | | |

| | | | |
|---|---|---|---|
| Celkový počet subjektov: | 2 550 | Správnosť klasifikácie: | 93.9 % |
| $F_1$ štatistika: | 12.4 % | McFadden pseudo $R^2$: | 0.1353 |

**Ostatné odvetvia**

| | | Odhalené NZ áno | Odhalené NZ nie | spolu | PPH |
|---|---|---|---|---|---|
| Klasifikácia | podozrivý | 19 | 134 | 153 | **12.4 %** |
| | nepodozrivý | 95 | 3 697 | 3 792 | 97.5 % |
| | spolu | 114 | 3 831 | | NPH |
| | | **16.7 % senzitivita** | 96.5 % špecificita | | |

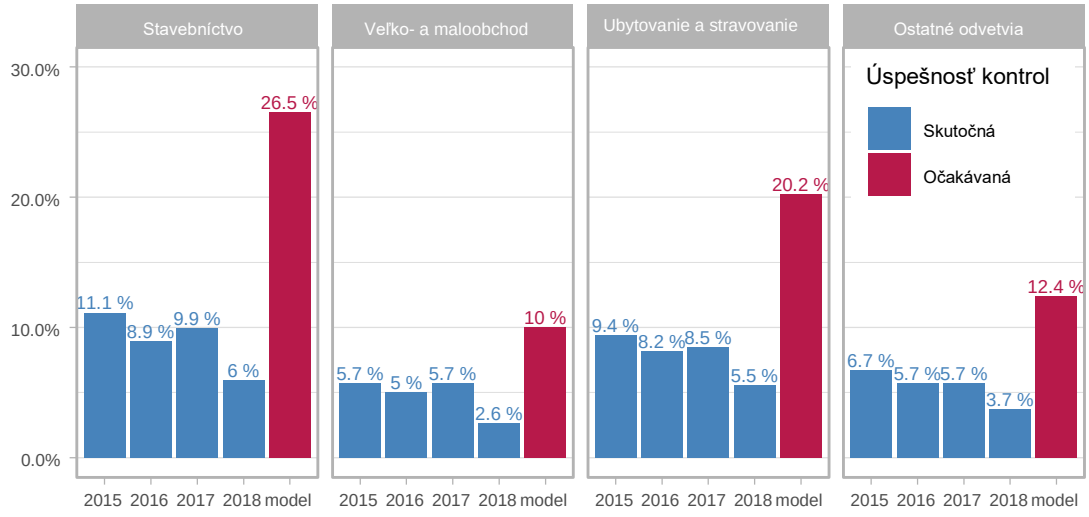
| | | | |
|---|---|---|---|
| Celkový počet subjektov: | 3 945 | Správnosť klasifikácie: | 94.2 % |
| $F_1$ štatistika: | 14.2 % | McFadden pseudo $R^2$: | 0.1147 |

## 3.2 Úspešnosť modelov

**Použitý model logistickej regresie lepšie sedí na odvetvia stavebníctva a ubytovacích a stravovacích služieb.** Na základe McFaddenovej pseudo-$R^2$ štatistiky sú modely pre tieto odvetvia kvalitnejšie ako tie pre veľkoobchod a maloobchod a pre nerizikové odvetvia (Tab. 2). Jedným z dôvodov môže byť vyššia úspešnosť predchádzajúcich kontrol nelegálneho zamestnávania, čo má za následok menej nevyrovnané zastúpenie poctivo a nelegálne zamestnávajúcich subjektov v tréningovej vzorke (Obr. 11). Lepšia zhoda modelu pre stavebníctvo a ubytovanie a stravovanie vedie aj k vyššej úspešnosti identifikácie podozrivých subjektov ($F_1$ štatistika, pozitívna prediktívna hodnota, senzitivita v Tab. 2 a Obr. 14).

**Očakávaná úspešnosť modelov je troj- až štvornásobkom úspešnosti kontrol v roku 2018 (Obr. 17).** Pre stavebníctvo je to takmer 4,5-násobok. Na odhalenie podobného počtu nelegálne zamestnávajúcich subjektov by preto mohlo stačiť vykonať menej ako tretinu kontrol.

**Obr. 17 Očakávaná úspešnosť modelov prevyšuje súčasnú úspešnosť kontrol pre všetky odvetvia**




Zdroj: NIP, vlastné spracovanie

# 4 Záver

**Cielením kontrol na základe administratívnych dát možno stroj- až zoštvornásobiť úspešnosť odhaľovania nelegálneho zamestnávania.** V posledných rokoch slovenské kontrolné orgány zvyšujú počty kontrol dodržiavania zákazu nelegálneho zamestnávania, zatiaľ čo počty odhalených nelegálne zamestnávajúcich subjektov klesajú. S využitím údajov o predchádzajúcich kontrolách, o počte a štruktúre pracovníkov a o hospodárskych výsledkoch subjektov navrhujeme spôsob, ktorý umožní odhaliť vyšší počet nelegálne zamestnávajúcich subjektov pri tretinovom počte kontrol.

Výrazné zníženie počtu kontrol a ich zefektívnenie prinesie **úsporu verejných financií**, resp. vytvorí priestor pre lepšiu alokáciu interných kapacít a zdrojov**.** Zároveň pri vyššom počte odhalených nelegálne zamestnávajúcich subjektov možno očakávať vyššiu celkovú sumu uložených pokút, čím sa kladný vplyv na verejné financie ešte umocní. Efektívnejšie kontroly prinesú aj **pozitívne sociálne vplyvy** v podobe riadnej evidencie zamestnancov a plnenia pracovnoprávnych a odvodových povinností, čím pozitívne prispejú aj ku **kvalite podnikateľského prostredia**.

**V testoch 14 metód strojového učenia je najúspešnejšia logistická regresia**. Túto metódu aplikujeme na každé z troch rizikových odvetví – stavebníctvo, veľkoobchod a maloobchod a ubytovacie a stravovacie služby – a na ostatné odvetvia. Modely dosahujú najvyššiu úspešnosť pre stavebníctvo a ubytovacie a stravovacie služby, kde sú úspešnejšie aj súčasné kontroly inšpektorátov práce.

Modely pre jednotlivé odvetvia vykazujú niekoľko podobných znakov. **Miera podozrenia z nelegálneho zamestnávania je vyššia** pre subjekty, ktoré majú

- sídlo v bratislavskom kraji,
- dlhy voči štátu,
- nižší počet zamestnancov,
- vyšší počet pracovníkov na dohodu,
- zahraničných alebo medzinárodných vlastníkov,
- sú mladšie.

Naopak, na mieru podozrenia nemajú výrazný vplyv výsledky kontrol nelegálneho zamestnávania z predchádzajúceho roku, informácie o cudzincoch medzi pracovníkmi ani mzdy, ktoré subjekt vypláca.

**Medzi hlavné obmedzenia modelu patrí nedostupnosť dát o pobočkách a prevádzkach zamestnávateľov**.[28] Plánovaná evidencia miesta výkonu práce Sociálnou poisťovňou od roku 2022[29] v budúcnosti umožní identifikáciu podozrení na úrovni prevádzok subjektov,[30] čo tento nedostatok zredukuje. Presnosť modelu taktiež ovplyvňuje zmena definície nelegálneho zamestnávania od januára 2018, avšak pri opakovanom použití modelu v budúcnosti bude efekt tejto legislatívnej zmeny zanedbateľný.

Za účelom zlepšenia odhaľovania nelegálneho zamestnávania boli výstupy z analytického modelu poskytnuté Národnému inšpektorátu práce. Popri využití dostupných údajov pri cielení kontrol odporúčame systematicky zbierať štruktúrované údaje o vykonávaných kontrolách, vrátane údajov o prevádzkach zamestnávateľov. Prínosom môže byť integrácia informačných systémov inšpekcie práce na informačný systém Sociálnej poisťovne a Finančnej správy, a tiež rozvoj interných analytických kapacít v rámci Národného inšpektorátu práce.

[28] Tento problém je najvýraznejší v stavebníctve, kde miesto výkonu práce z princípu nie je v sídle zamestnávateľa.
[29] Doplnenie § 232a do zákona č. 461/2003 Z. z. o sociálnom poistení pomocou Čl. I bodu 101. zákona č. 317/2018 Z. z.
[30] Presnejšie subjektov v obci. V prípade viacerých prevádzok v rovnakej obci stále nebude možné tieto prevádzky rozlíšiť.

## Literatúra

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

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Breunig, M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD Record, 29*(2), 93-104.

Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization,. *ACM Transactions on Knowledge Discovery from Data, 10*(1), 1-51.

De Wispelaere, F., Pacolet, J., Rotaru, V., Naylor, S., Gillis, D., & Alogogianni, E. (2017). Data mining for more efficient enforcement: A practitioner toolkit from the thematic workshop of the European Platform Undeclared Work. *Brussels: European Commission*.

Dobson, A. J., & Barnett, A. (2008). *An Introduction to Generalized Linear Models.* Boca Raton, FL, United States: Taylor & Francis Inc.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189-1232.

Huang, J., Zhu, Q., Yang, L., & Feng, J. (2015). A non-parameter outlier detection algorithm based on Natural Neighbor. *Knowledge-Based Systems, 92*, 71-77.

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.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer-Verlag New York.

Karaboev, S., Mineva, D., & Stefanov, R. (2019). Toolkit on risk assessments for more efficient inspections as a means to tackle undeclared work. *Brussels: European Commission*.

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). LoOP: Local Outlier Probabilities. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data, 6*(1), 1-39.

Národný inšpektorát práce. (2019). *Informatívna správa o vyhľadávaní a potieraní nelegálnej práce a nelegálneho zamestnávania za rok 2018.*

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation, 13*(7), 1443-1471.

Státní úřad inspekce práce. (2018). *Roční souhrnná zpráva o výsledcích kontrolních akcí provedených inspekcí za rok 2017.*

# Prílohy

## A. Použité dáta

Základom databázy, ktorú používame na identifikáciu subjektov podozrivých z nelegálneho zamestnávania, sú údaje o **minulých kontrolách dodržiavania zákazu nelegálneho zamestnávania**, ktoré v rokoch 2014 - 2018 vykonali inšpektoráty práce.[31] Konkrétne využívame informáciu o tom, ktorý subjekt bol v danom roku kontrolovaný a či bolo pri kontrole zistené nelegálne zamestnávanie alebo nie.

Z databázy individuálnych poistných vzťahov Sociálnej poisťovne sme zistili údaje o **počte a štruktúre pracovníkov** v jednotlivých subjektoch. Pre každý subjekt v každom roku sme spočítali, zvlášť pre zamestnancov a pracovníkov na dohodu,[32] priemernú mesačnú hrubú mzdu[33], priemerný počet odpracovaných dní v mesiaci, priemerný mesačný podiel žien a cudzincov, priemerný počet pracovníkov v mesiaci, priemerný vek pracovníka a počet mesiacov, kedy mal subjekt nejakých pracovníkov.

**Cudzinci** pracujúci na Slovensku sa dajú rozdeliť do troch skupín: občania EÚ/EHP, občania tretích krajín s informačnou kartou bez povolenia na zamestnanie[34] a občania tretích krajín s povolením na zamestnanie. Údaje o prvých dvoch skupinách[35] sme zahrnuli do databázy vo forme informácie o priemernom mesačnom podiele týchto cudzincov medzi všetkými pracovníkmi subjektu a podiele žien medzi týmito cudzincami.

Hospodárske výsledky subjektov sme zohľadnili pomocou údajov z **účtovných závierok**.[36] Údaje zo súvahy a z výkazu ziskov a strát boli normované na jedného pracovníka a údaje z výkazu ziskov a strát navyše aj na 365 dní.[37] Okrem samotných údajov z účtovnej závierky v danom roku sme použili aj percentuálnu zmenu oproti predchádzajúcemu roku.

Posledným zdrojom údajov je databáza **finančných ukazovateľov** doplnených o **kritické údaje**, ako sú napríklad dlhy, konkurzy alebo platobné rozkazy.[38]

### Box 4 Spojená databáza

Prepojením spomínaných databáz cez IČO subjektu a rok sme získali databázu so štruktúrou, ktorú ilustruje Tab. 3. Odvetvie je buď jedno z troch rizikových odvetví (stavebníctvo, veľkoobchod a maloobchod, ubytovanie a stravovanie), alebo ostatné odvetvia. Premenná $NZ$ má hodnotu 1, ak bolo v danom subjekte v danom roku odhalené nelegálne zamestnávanie. V prípade, že subjekt bol kontrolovaný, ale nebolo zistené nelegálne zamestnávanie, je $NZ = 0$.[39] Ak subjekt v danom roku nebol kontrolovaný, tak hodnota nie je vyplnená.

Stĺpce označené NIP označujú ďalšie dve binárne premenné získané z databázy minulých kontrol nelegálneho zamestnávania. Sú to indikátory označujúce subjekty, u ktorých bolo v predchádzajúcom roku

[31] Tieto údaje ISP poskytol Národný inšpektorát práce.
[32] Medzi pracovníkmi na dohodu sú pracovníci na dohodu o vykonaní práce, dohodu o pracovnej činnosti a dohodu o brigádnickej práci študenta.
[33] Odhadnutá z vymeriavacieho základu na úrazové poistenie. Individuálne hodnoty boli pred počítaním priemerov normované na celý mesiac (30,5 dní poistenia) a transformované na odchýlky od priemeru pre danú vekovú kategóriu a pohlavie pracovníka, okres sídla zamestnávateľa a divíziu SK-NACE.
[34] Táto skupina zahŕňa aj občanov tretích krajín pracujúcich prostredníctvom agentúr dočasného zamestnávania a zahraničných Slovákov.
[35] Tieto údaje ISP poskytlo Ústredie práce, sociálnych vecí a rodiny.
[36] Tieto údaje, dostupné aj v Registri účtovných závierok, ISP poskytla firma FinStat, s. r. o.
[37] Pre numerické premenné boli namiesto samotných hodnôt použité odchýlky od priemeru pre danú kategóriu veľkosti podniku, okres a divíziu SK-NACE.
[38] Tieto údaje zhromažďuje a počíta z dostupných údajov firma FinStat, s. r. o., ktorá ich poskytla ISP.
[39] Pre väčšinu subjektov teda na základe jednej kontroly konkrétnej prevádzky určíme, či bol subjekt v daný rok poctivý alebo nelegálne zamestnával.

odhalené nelegálne zamestnávanie, a subjekty, ktoré boli v predchádzajúcom roku kontrolované, ale nelegálne zamestnávanie nebolo zistené.[40]

Stĺpce SP a Infokarty označujú vyššie spomínaných 14 premenných o zamestnancoch a pracovníkoch na dohodu získaných z poistných vzťahov v Sociálnej poisťovni a dve premenné o cudzincoch s informačnou kartou. Údaje z účtovných závierok obsahujú 90 premenných, kde okrem samotných údajov zo súvahy a výkazu ziskov a strát sú obsiahnuté aj informácie o adrese sídla subjektu, právnej forme, druhu vlastníctva, SK NACE kóde ekonomickej činnosti a o veku subjektu.

Finančné ukazovatele obsahujú premenné ako rôzne marže, ukazovatele zadlženosti alebo kreditné a bonitné modely. Okrem nich sú sem zaradené aj kritické údaje ako dlhy voči zdravotným poisťovniam a voči sociálnej poisťovni, pohľadávky štátu, daňové nedoplatky alebo konkurzy a reštrukturalizácie. Dokopy je v tejto časti 74 premenných.

Spojená databáza obsahuje pozorovania subjektov v rokoch 2015 – 2018, pričom každý z viac ako 430 000 záznamov patrí jednému subjektu v konkrétnom roku.[41] Celkovo je v databáze zastúpených takmer 142 000 rôznych subjektov a pozorovaných premenných je 182. Po vylúčení riedko vyplnených premenných[42] a tých, ktoré nie sú vhodné pre použitie v modeloch[43], ich ostane viac ako 130.[44] Databáza len za roky 2017 a 2018, ktorá je vhodnejšia pre väčšinu odvetví, obsahuje takmer 229 000 pozorovaní viac ako 126 000 rôznych subjektov.

**Tab. 3 Štruktúra spojenej databázy**

| Rok | IČO | Odvetvie | NZ | NIP | SP | Infokarty | Účtovné závierky | Finančné ukazovatele |
|---|---|---|---|---|---|---|---|---|
| 2018 | 12345678 | stavebníctvo | 1 | ... | ... | ... | ... | ... |
| 2015 | 87654321 | ostatné | 0 | ... | ... | ... | ... | ... |
| 2017 | 12344321 | obchod | ? | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Zo všetkých viac ako 130 premenných sme ešte odstránili viac ako polovicu na základe slabej korelácie s indikátorom nelegálneho zamestnávania $NZ$ alebo silnej korelácie s lepším prediktorom . Po tomto výbere databáza obsahuje okolo 50 vysvetľujúcich premenných.[44]

Posledná fáza výberu premenných, po ktorej v modeli ostane 18 – 29 premenných (Tab. 4), je špecifická pre logistickú regresiu. Kritériom výberu je významnosť premennej pri opakovanom odhadnutí modelu na náhodne zvolených 95 % tréningovej vzorky. Ponechané sú premenné, ktoré sú signifikantné na hladine významnosti 20 % vo viac ako polovici zo 400 opakovaní.[45]

[40] Takto konštruovaným premenným vieme priradiť hodnotu pre všetky subjekty, preto sú vhodnejšie ako použitie premennej NZ posunutej o rok.
[41] Údaje z účtovných závierok a finančných ukazovateľov sú použité s ročným oneskorením kvôli ich neskoršej dostupnosti oproti ostatným zdrojom. Dáta o kontrolách nelegálneho zamestnávania a o pracovníkoch v danom roku sú teda spojené s finančnými údajmi z predchádzajúceho roku.
[42] Údaje z účtovných výkazov sú prevedené na jednotnú štruktúru, ale rôzne subjekty podávajú rôzne typy finančných výkazov, čím vznikajú chýbajúce hodnoty. Vylúčené boli premenné, kde chýba viac ako 10 % pozorovaní.
[43] Napríklad dátumové premenné, kód obce alebo duplicitné kódy ekonomickej činnosti z rôznych zdrojov.
[44] Presný počet premenných závisí od konkrétneho odvetvia.
[45] Vyššia hladina významnosti bola zvolená, aby sme predišli eliminácii premenných, ktoré samostatne nie sú až tak významné, ale potenciálne môžu kvalitu modelu zvýšiť v súčinnosti s inými premennými.

**Tab. 4 Vysvetľujúce premenné použité v modeloch pre jednotlivé odvetvia**

| Premenná | Popis premennej | Stavebníctvo | Veľko- a maloobchod | Ubytovanie a stravovanie | Ostatné odvetvia |
|---|---|---|---|---|---|
| NZ_predch_0 * | Neodhalené NZ v predchádzajúcom roku | ✓✓ | | | ✓ |
| NZ_predch_1 * | Odhalené NZ v predchádzajúcom roku | | | ✓ | |
| dni_ZEC | Priemerný počet dní zamestnaneckého pomeru v mesiaci | ✓✓ | ✓ | ✓✓ | |
| osoby_ZEC | Priemerný počet zamestnancov v mesiaci | ✓ | ✓ | ✓ | |
| zeny_ZEC | Priemerný podiel žien medzi zamestnancami | | | ✓✓ | |
| vek_ZEC | Priemerný vek zamestnanca | ✓✓ | | ✓ | |
| mesiace_ZEC | Počet mesiacov v roku, kedy mal subjekt zamestnancov | ✓✓ | | ✓✓ | |
| mzdy_DOH | Priemerná hrubá mesačná mzda dohodára | ✓✓ | | ✓✓ | |
| dni_DOH | Priemerný počet dní dohodárskeho pomeru v mesiaci | | ✓ | | |
| cudz_DOH | Priemerný podiel cudzincov medzi dohodármi | ✓✓ | ✓ | | |
| osoby_DOH | Priemerný počet dohodárov v mesiaci | ✓✓ | ✓ | | |
| mesiace_DOH | Počet mesiacov v roku, kedy mal subjekt dohodárov | | | ✓✓ | |
| cudz_infokarty | Podiel cudzincov s infokartami medzi pracovníkmi | | | ✓✓ | |
| cudz_infokarty_zeny | Priemerný podiel žien medzi cudzincami s infokartami | ✓✓ | | | |
| pohladavky_kratko | Krátkodobé pohľadávky súčet | ✓✓ | | | |
| peniaze | Peniaze | ✓ | | | |
| zavazky_obchod | Záväzky z obchodného styku | | | ✓ | |
| zavazky_kratko | Krátkodobé záväzky súčet | ✓✓ | | | |
| trzby_vyr_sluzby | Tržby z predaja vlastných výrobkov a služieb | | | | ✓ |
| pridana_hodnota | Pridaná hodnota | | | ✓✓ | |
| osobne_naklady | Osobné náklady súčet | ✓✓ | | ✓✓ | |
| dane_poplatky | Dane a poplatky | | | | ✓ |
| VH_po_dani | Výsledok hospodárenia za účtovné obdobie po zdanení | ✓✓ | | | |
| z_vlastne_imanie | Zmena - Vlastné imanie | ✓ | | | |
| z_zavazky_kratko | Zmena - Krátkodobé záväzky súčet | ✓✓ | | | |
| z_spotreba_vyroba | Zmena - Výrobná spotreba | | | ✓✓ | |
| vek_firmy | Vek firmy v rokoch | | ✓ | | ✓ |
| trzby_bez_zas_akt | Tržby očistené o Zásoby a Aktiváciu | | | | ✓ |
| tvorba_zdrojov | Hrubá tvorba zdrojov z prevádzkovej činnosti | | | | ✓ |
| marza_hruba | Hrubá marža | | ✓ | ✓✓ | |
| marza_ebitda | EBITDA marža | | | ✓ | |
| marza_prevadzka | Prevádzková marža | ✓✓ | | | |
| marza_cash_flow | Marža účtovného peňažného toku | | ✓✓ | ✓ | |
| zavazky_ebitda | Záväzky/EBITDA | | | | ✓ |
| likvidita3 | Likvidita 3. stupňa | ✓ | | | |
| fin_ucty_aktiva | Finančné účty/Aktíva | | ✓✓ | ✓✓ | |
| doba_splac_zav | Doba splácania záväzkov | ✓✓ | | | |
| oddlzenost_kratko | Krátkodobá toková oddlženosť | | ✓✓ | | |
| kryt_pers_nakl_odpis | Krytie personálnych nákladov a odpisov | | | ✓✓ | |
| doba_splac_urok_cf | Doba splácania dlhov z čistého peňažného toku | | | | ✓ |
| zadlz_imanie | Miera zadlženosti vlastného imania | | | | ✓ |
| quick_test | Quick test | | | | ✓ |
| credit_altman_z | Credit scoring - Altmanovo Z skóre | | | | ✓ |
| ebit | Zisk pred zdanením a úrokmi (EBIT) | | | ✓ | |
| kar_likvidacie * | Konkurzy a reštrukturalizácie (KaR), Likvidácie | | | ✓ | |
| zrus_dph * | Zrušenie registrácie platiteľa DPH | | | ✓ | |
| dlh_vszp * | Dlh - VšZP | ✓✓ | ✓✓ | | |
| dlh_union * | Dlh - ZP Union | ✓✓ | | | |
| dlh_dane * | Dlh - Daňový nedoplatok | | ✓✓ | | |
| dlh_soc * | Dlh - Soc. poisť. | ✓✓ | | | ✓ |

| Premenná | Popis premennej | Stavebníctvo | Veľko- a maloobchod | Ubytovanie a stravovanie | Ostatné odvetvia |
|---|---|---|---|---|---|
| dlh_stat * | Dlh - Pohľadávky štátu | ✓✓ | ✓ | ✓✓ | |
| divizia ** | Divízia SK-NACE | ✓✓ | ✓✓ | ✓✓ | |
| trieda ** | Trieda SK-NACE | | | | ✓✓ |
| kraj ** | Kraj z adresy sídla subjektu | ✓✓ | ✓✓ | ✓✓ | ✓✓ |
| credit_altman_z_ind ** | Credit scoring - Altmanovo Z skóre indikácia | ✓✓ | ✓✓ | ✓✓ | ✓✓ |
| credit_index_05_ind ** | Credit scoring – INDEX 05 indikácia | ✓✓ | ✓✓ | ✓✓ | ✓✓ |
| druh_vlast ** | Druh vlastníctva | ✓✓ | ✓✓ | ✓✓ | ✓✓ |
| rok ** | Rok | ✓ | ✓ | ✓ | ✓ |
| Počet premenných | | 29 | 19 | 28 | 18 |

Pozn.: symbol ✓ znamená, že daná premenná bola v modeli pre dané odvetvie zahrnutá len samostatne; symbol ✓✓ znamená, že bola zahrnutá aj v interakcii s rokom.
* indikátorová premenná (0/1), ** Kategorická premenná

Pri odhade modelov pre jednotlivé odvetvia boli subjekty v danom roku kontrolované na nelegálne zamestnávanie rozdelené na tréningovú (60 % údajov, Tab. 5), validačnú (20 %) a testovaciu vzorku (20 %). Na tréningovej vzorke boli odhadnuté koeficienty. Odhadnutý model bol použitý na výpočet miery podozrenia v danom roku pre subjekty z validačnej vzorky a bola zvolená hraničná miera podozrenia tak, aby maximalizovala $F_1$ štatistiku. Úspešnosť modelu s koeficientami odhadnutými na tréningovej vzorke subjektov a hraničnou mierou podozrenia vybranej na validačnej vzorke bola následne testovaná na testovacej vzorke. Nakoniec sme pre vybrané modely testovali úspešnosť klasifikácie vzhľadom na výsledky kontrol v roku 2018.

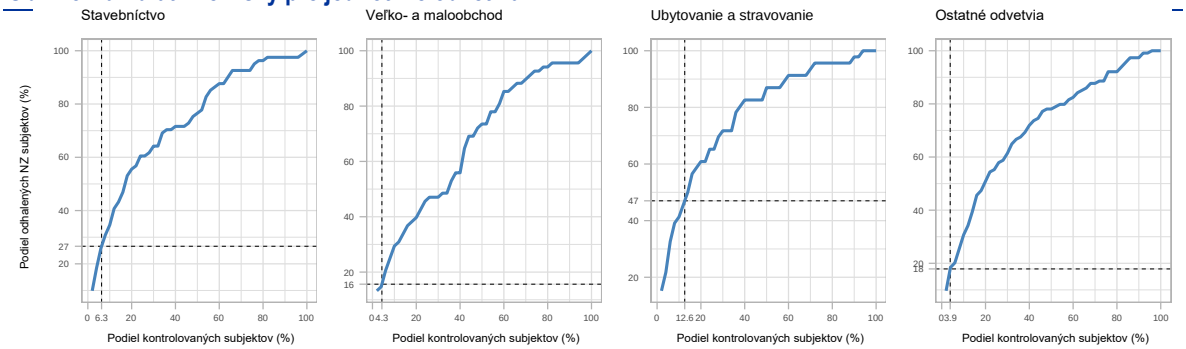
**Tab. 5 Počty subjektov v tréningovej, validačnej a testovacej vzorke pre jednotlivé odvetvia**

| | 2015-2018 | | | | 2017-2018 | | | |
|---|---|---|---|---|---|---|---|---|
| | Tréning (60 %) | Validácia (20 %) | Testovanie (20 %) | Celkom | Tréning (60 %) | Validácia (20 %) | Testovanie (20 %) | Celkom |
| Stavebníctvo | 2 943 | 981 | 981 | 4 905 | 1 413 | 471 | 470 | 2 354 |
| Veľko- a maloobchod | 4 913 | 1 637 | 1 638 | 8 188 | 2 293 | 765 | 766 | 3 824 |
| Ubytovanie a stravovanie | 2 030 | 677 | 677 | 3 384 | 900 | 301 | 301 | 1 502 |
| Ostatné sektory | 8 555 | 2 851 | 2 851 | 14 257 | 4 319 | 1 440 | 1 439 | 7 198 |

## B. Kumulatívne zisky a lift

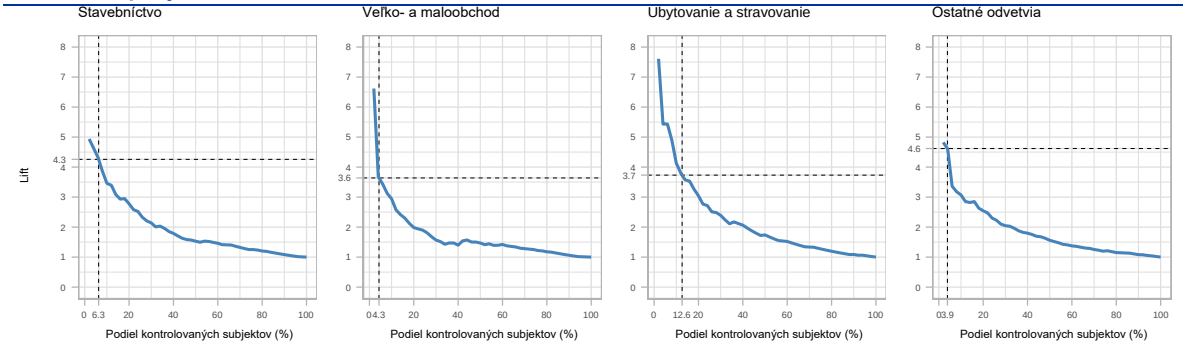
**Kumulatívne zisky** hovoria o tom, aká časť všetkých nelegálne zamestnávajúcich subjektov bude odhalená, ak inšpektoráty skontrolujú zvolený podiel najpodozrivejších subjektov (Obr. 18). Napríklad pre stavebníctvo náš model odporúča skontrolovať 6,3 % všetkých subjektov, pričom bude odhalená viac ako štvrtina všetkých nelegálne zamestnávajúcich subjektov. V prípade skontrolovania pätiny subjektov s najvyššou mierou podozrenia by bola odhalená viac ako polovica nelegálne zamestnávajúcich subjektov.

**Obr. 18 Kumulatívne zisky pre jednotlivé odvetvia**




Zdroj: vlastné spracovanie

**Obr. 19 Lift pre jednotlivé odvetvia**




Zdroj: vlastné spracovanie

**Lift** dáva do pomeru podiel skontrolovaných subjektov a podiel odhalených nelegálne zamestnávaných subjektov (Obr. 19). Pre stavebníctvo je pri skontrolovaní 6,3 % najpodozrivejších subjektov odhalených 27 % nelegálne zamestnávajúcich subjektov, čo predstavuje lift $27/6{,}3 \doteq 4{,}3$.

## C. Metódy vyhľadávania anomálií

**Aproximácia normálnym rozdelením (NormDist).** Dáta sa aproximujú viacrozmerným normálnym rozdelením s vektorom stredných hodnôt daným výberovými priemermi premenných a kovariančnou maticou danou výberovou kovariančnou maticou. Za anomálie sú označené body, pri ktorých je pravdepodobnosť výskytu pod zvolenou hodnotou.

**Local outlier factor (LOF)** je metóda založená na meraní lokálnej odchýlky daného bodu od susedných bodov. Každému dátovému bodu je priradený Local outlier factor (LOF) na základe jeho vzdialenosti od bodov v jeho okolí s ohľadom na hustotu bodov v danej oblasti. Za anomálie sú označené body s LOF nad zvolenou hodnotou (Breunig, Kriegel, Ng, & Sander, 2000).

Metóda **Local outlier probability (LoOP)** je odvodená od metódy LOF. Local outlier probability (LoOP) je vďaka použitiu lokálnych štatistík menej citlivá na voľbu parametra *k* označujúceho počet použitých susedných bodov. Za anomálie sú označené body s LoOP nad zvolenou hodnotou (Kriegel, Kröger, Schubert, & Zimek, 2009).

**Natural outlier factor (NOF)** je neparametrická metóda založená na algoritme prirodzeného okolia (*natural neighborhood*). Porovnáva hustotu dátového bodu a jeho susedov. Za anomálie sú označené body s NOF nad zvolenou hodnotou (Huang, Zhu, Yang, & Feng, 2015).

Metóda **Isolation forest (iForest)** je založená na izolácii bodov pomocou rozhodovacích stromov. Za anomálie sú označené body, ktoré sa dajú izolovať od ostatných po menšom priemernom počte (náhodne zvolených) rozhodnutí (Liu, Ting, & Zhou, 2012).

Pri metóde **One-class support vector machines (1cSVM)** sa dáta transformujú do priestoru s vyššou dimenziou (*kernel trick*), kde sa separujú nadrovinou s maximálnou vzdialenosťou od dát. Vďaka transformácii sa 1cSVM dá použiť aj na nelineárnu separáciu (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001).

**Hierarchical density-based spatial clustering of applications with noise (HDBSCAN)** je rozšírením metódy DBSCAN, ktorá na základe lokálnej hustoty hľadá v dátach zhluky za prítomnosti šumu. Každému bodu je priradená miera anomality a body s touto mierou vyššou ako zvolená hranica sú označené za anomálie (Campello, Moulavi, Zimek, & Sander, 2015).

## D. Klasifikačné metódy

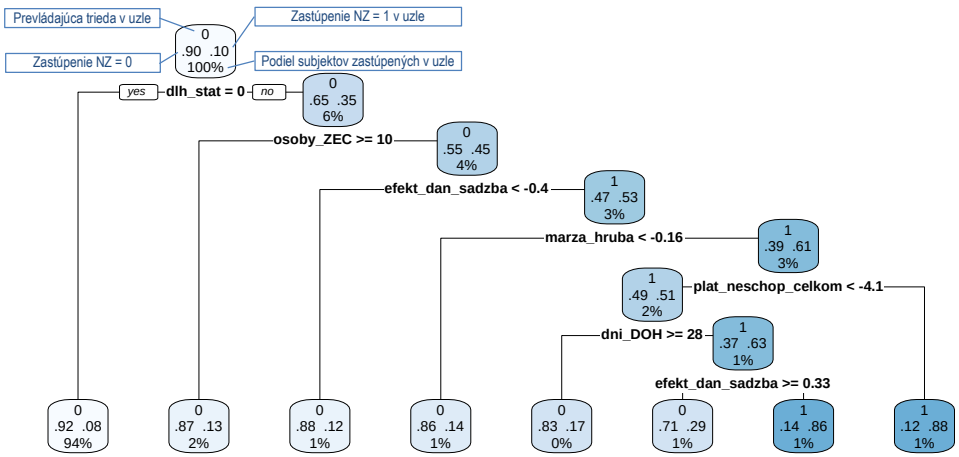
**Logistická regresia (LogReg)** modeluje pravdepodobnosť $p$, že subjekt je nelegálne zamestnávajúci, pomocou logistickej funkcie $p = \frac{1}{1+e^{-x^T\beta}}$, kde $x = (1, x_1, \ldots, x_n)^T$ je vektor vysvetľujúcich premenných,

$\beta = (\beta_0, \beta_1, \dots, \beta_n)^T$ je vektor koeficientov (Dobson & Barnett, 2008). Koeficienty modelu $\beta$ sa metódou iteratívne prevažovaných najmenších štvorcov odhadnú z tréningovej vzorky subjektov, kde je známe, či subjekt nelegálne zamestnával[46]. Daný subjekt je klasifikovaný ako podozrivý, ak odhadnutá pravdepodobnosť nelegálneho zamestnávania je nad zvolenou hranicou (Aggarwal, 2015).

**Lasso logistická regresia (Lasso)** odhaduje parametre modelu logistickej regresie pomocou metódy penalizovanej maximálnej vierohodnosti. Oproti klasickej logistickej regresii vedie lasso k jednoduchšiemu modelu, ktorý vylúči menej potrebné vysvetľujúce premenné[47].

**Klasifikačný strom (ClassTree)** funguje na princípe postupného delenia subjektov na základe hodnoty jednej z vysvetľujúcich premenných na dve skupiny tak, aby boli oddelené nelegálne zamestnávajúce subjekty od poctivých. Klasifikačný strom na Obr. 20 oddelí dve skupiny subjektov, v ktorých je zastúpenie nelegálne zamestnávajúcich subjektov viac ako 85 %, ale v každej z týchto skupín je len stotina všetkých subjektov z tréningovej vzorky. Natrénovaný klasifikačný strom sa dá použiť na priradenie pravdepodobnosti nelegálneho zamestnávania (Izenman, 2008).

**Obr. 20 Klasifikačný strom pre stavebnícke subjekty**




Zdroj: vlastné spracovanie

Metóda **náhodný les (RandForest)** využíva naraz vyšší počet klasifikačných stromov, ktorých výsledky priemeruje. Tento postup pomáha eliminovať častý problém klasifikačných stromov, ktorým je pretrénovanie na tréningovej vzorke (Breiman, 2001).

**Lineárna diskriminačná analýza (LDA)** je založená na hľadaní lineárnej kombinácie vysvetľujúcich premenných, ktorá separuje nelegálne zamestnávajúce subjekty od poctivých. V porovnaní s ostatnými použitými metódami má LDA pomerne siné predpoklady ako normalitu dát a homoskedasticitu (Izenman, 2008)

**Gradient boosting machine (GBM)**, podobne ako náhodný les, využíva viacero jednoduchých klasifikačných stromov. Na rozdiel od náhodného lesa, pri GBM nie sú jednotlivé stromy použité naraz ale postupne tak, aby nasledujúci strom zohľadnil výsledky tých predchádzajúcich (Friedman, 2001).

**Extreme gradient boosting (XGBoost)** je efektívnejšia implementácia princípu gradient boosting zovšeobecnená na použitie pri optimalizácii ľubovoľnej diferencovateľnej funkcie a rozšírená o regularizáciu, ktorá pomáha predísť pretrénovaniu modelu (Chen & Guestrin, 2016).

[46] Za pravdepodobnosť $p$ je dosadená premenná $NZ \in \{0, 1\}$ spomínaná v Box 4.
[47] Názov lasso je skratkou anglického *least absolute shrinkage and selection operator* a táto metóda sa vo všeobecnosti používa aj na výber premenných do modelu (Dobson & Barnett, 2008, s. 114--115).