UNIVERZITA KOMENSKÉHO V BRATISLAVE

Fakulta matematiky, fyziky a informatiky

Ako dlho trvá nezamestnanosť?

Diplomová práca

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Školiteľ:	Mgr. Ján Komadel, PhD.

Bratislava, 2021 **Bc. Alžbeta Dudová**



Univerzita Komenského v Bratislave

Fakulta matematiky, fyziky a informatiky



ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Alžbeta Dudová

Študijný program: ekonomicko-finančná matematika a modelovanie
(Jednoodborové štúdium, magisterský II. st., denná forma)

Študijný odbor: matematika

Typ záverečnej práce: diplomová

Jazyk záverečnej práce: slovenský

Sekundárny jazyk: anglický

Názov: Ako dlho trvá nezamestnanosť?
How long does unemployment last?

Anotácia: Mesačne je do evidencie uchádzačov o zamestnanie (UoZ) zaradených 18 tisíc nových UoZ. O každom UoZ sú v čase zaradenia do evidencie evidované údaje, ktoré majú potenciál na modelovanie trvania nezamestnanosti daného UoZ. Úspešný model prispeje k predchádzaniu dlhodobej nezamestnanosti tým, že pomôže identifikovať, ktorí UoZ sú rizikoví z hľadiska upadnutia do dlhodobej nezamestnanosti. Tiež môže pomôcť odhaliť, ktoré charakteristiky UoZ najviac vplývajú na dĺžku nezamestnanosti, čo je užitočná informácia aj pri návrhu konkrétnych aktívnych opatrení trhu práce, ktoré by efektívne skracovali trvanie nezamestnanosti.

Vedúci: Mgr. Ján Komadel, PhD.

Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky

Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 17.12.2019

Dátum schválenia: 21.12.2019 prof. RNDr. Daniel Ševčovič, DrSc.

garant študijného programu

.....

študent

.....

vedúci práce

ABSTRAKT

DUDOVÁ, Alžbeta: Ako dlho trvá nezamestnanosť? [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Komadel, PhD.; Bratislava, 2021, 68 s.

Nezamestnanosť je negatívny sociálno-ekonomický jav spojený s trhom práce. Mesačne bolo do evidencie UoZ zaradených priemerne 19 000 uchádzačov (od januára 2015 do mája 2021). Otázka, ako dlho budú uchádzači na úrade práce evidovaní, sa teda týka nezanedbateľného počtu obyvateľov SR. Keď chceme predpovedať dĺžku trvania evidencie konkrétneho uchádzača o zamestnanie, v prvom rade potrebujeme poznať o ňom dostatočné množstvo kľúčových informácií, ktoré na túto dĺžku trvania vplývajú. Takýmito sú napr. dosiahnuté vzdelanie alebo jazykové znalosti. Avšak rovnako dôležité je nájsť model, ktorý na základe množstva dát o minulých uchádzačoch bude schopný čo najpresnejšie určiť predpokladanú dĺžku trvania evidencie pre súčasných uchádzačov o zamestnanie vzhľadom na ich parametre. V našej práci sme sa rozhodli použiť modely strojového učenia (angl. *machine learning*). Konkrétne lineárnu regresiu, náhodný les a XGBoost, ktoré sme implementovali v prostredí R. Ako model s najmenšou chybou predikcie sa ukazuje XGBoost. Zistili sme, že pri odhadovaní dĺžky trvania evidencie jednotlivých uchádzačov sú najdôležitejšie faktory dĺžka poslednej evidencie (pokiaľ UoZ už bol na úrade evidovaný v minulosti), vek, počet znevýhodnení, ktoré uchádzač má, kraj, vzdelanie. Exploratívnou analýzou sme zistili, že uchádzači so žiadnym alebo základným vzdelaním sa ukazujú ako rizikovejší z hľadiska dĺžky trvania evidencie, podobne uchádzači, ktorí neovládajú prácu s počítačom ani na elementárnej úrovni.

Kľúčové slová: Nezamestnanosť, lineárna regresia, náhodný les, XGBoost, trvanie evidencie

ABSTRACT

DUDOVÁ, Alžbeta: How long does unemployment last? [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and statistics; Supervisor: Mgr. Ján Komadel, PhD., Bratislava, 2021, 68 s.

Unemployment is a negative socio-economic phenomenon associated with the labor market. An average of 19,000 job seekers were included in the UoZ register per month (from January 2015 to May 2021). The question of how long job seekers will be registered at the labour office therefore concerns a significant number of inhabitants of the Slovak Republic. When we want to predict the duration of registration of a particular job seeker, we first need to know a sufficient amount of key information about him that affects that duration. Such are e.g. educational attainment or language skills. However, it is equally important to find a model that, based on the amount of data on past job seekers, will be able to determine as accurately as possible the expected duration of registration for current jobseekers with regard to their parameters. In our thesis, we decided to use machine learning models. Specifically, linear regression, random forest, and XGBoost, which we implemented in the R programming environment. XGBoost appears to be the model with the smallest prediction error. We found that the most important factors in estimating the duration of registration of individual job seekers are the length of the last registration (if the UoZ has already been registered at the labour office in the past), age, number of disadvantages that the job seeker has, region, education. By an exploratory analysis we found that job seekers with no or basic education are more risky in terms of the length of registration, similarly to job seekers who do not know how to work with a computer even at the elementary level.

Key words: unemloyment, linear regression, random forest, XGBoost, duration of registration at labour office

Poďakovanie

Na tomto mieste by som sa chcela veľmi poďakovať každému, kto mi pri písaní práce akokoľvek pomohol. Hlavne môjmu školiteľovi Mgr. Ján Komadel, PhD. za pomoc, cenné rady a usmernenie pri písaní tejto práce. Avšak najväčšia vďaka patrí Pánu Bohu.

Obsah

Zoznam použitých skratiek 7

Úvod 8

1. Vymedzenie základných pojmov týkajúcich sa trhu práce 10

1.1. Princípy profilácie uchádzačov o zamestnanie (UoZ) 12

1.1.1. Individuálny prístup 13

1.1.2. Cielené aktívne opatrenia na trhu práce 14

1.1.3. Výhody a nevýhody vybraných prístupov 14

2. Popis závislej premennej a nezávislých premenných použitých pri tvorbe štatistických modelov 16

3. Exploratívna analýza vybraných premenných 35

4. Modely 41

4.1. Lineárna regresia 41

4.2. Ridge a Lasso regresia 42

4.3. Regresné stromy, náhodný les a XGBoost 43

4.3.1. Regresné stromy 43

4.3.2. Náhodný les 44

4.3.3. XGBoost 45

5. Implementácia a výsledky modelov 48

5.1. Lineárna, Ridge a Lasso regresia 48

5.2. Náhodný les 54

5.3. XGBoost 58

6. Diskusia 61

Záver 64

Použitá literatúra 66

Zoznam použitých skratiek

AOTP – aktívne opatrenia trhu práce

EAO – ekonomicky aktívne obyvateľstvo

ISCO-08 – Medzinárodná klasifikácia zamestnaní (angl. International Standard Classification of Occupations)

NACE – štatistická klasifikácia ekonomických činností v Európskej spoločnosti (franc. Nomenclature Statistique des activités économiques dans la Communauté européenne, angl. Statistical Classification of Economic Activities in the European Community)

MS Office – Microsoft Office

MSE tr – stredná kvadratická odchýlka (angl. mean squared error) z tréningovej vzorky dát

OČR – ošetrovanie člena rodiny

OECD – Organizácia pre hospodársku spoluprácu a rozvoj (angl. Organisation for Economic Co-operation and Development)

PC – osobný počítač

PN – práceneschopnosť

SZČO – samostatne zárobkovo činná osoba

ŠÚ SR – Štatistický úrad Slovenskej republiky

UoZ – uchádzač o zamestnanie

ÚPSVaR – ústredie práce, sociálnych vecí a rodiny

VZPS – výberové zisťovanie pracovných síl

Úvod

Definícia nezamestnanosti je založená na tom, že osoba schopná práce je z možnosti pracovať v platenom zamestnaní vyradená. Je to vážny ekonomický a zároveň i sociálny problém. V tejto práci pod pojmom nezamestnaný budeme rozumieť osobu, ktorá je vedená v evidencii uchádzačov o zamestnanie (UoZ). Mesačne do evidencie pribudlo od januára 2015 do mája 2021 priemerne 19 000 UoZ, pričom počas pandémie Covid-19 tieto čísla vzrástli o viac ako 24% (medzi februárom a septembrom 2020). Tento nárast môže mať súvis s reštriktívnymi opatreniami vlády (prísne obmedzenie prevádzok, pohybu ľudí a pod.). Nezamestnanosť sa priamo dotýka veľkej časti obyvateľstva. Krátke obdobie nezamestnanosti je pomerne bežná vec a nemá väčšie následky, avšak dlhšie obdobie nezamestnanosti je nepriaznivé pre jednotlivca aj pre spoločnosť.

Posúdenie rizikovosti UoZ z hľadiska dĺžky nezamestnanosti, patrí vo viacerých krajinách k cenným nástrojom pri nastavení politiky trhu práce. Neuplatnenie sa na trhu práce po viac ako 12 mesiacoch (dlhodobá nezamestnanosť) môže spôsobiť problémy s opätovným sa vrátením do pracovného prostredia (napr. strata pracovných návykov). Úspešné posúdenie rizikových uchádzačov priamo pri evidencii UoZ môže pomôcť lepšie cieliť aktívne opatrenia na trhu práce na tých, ktorí ich najviac potrebujú.

Najlepšia prax pri riešení dlhodobej nezamestnanosti je jej predchádzanie v podobe aktívnych opatrení trhu práce. Dĺžku trvania evidencie UoZ, s využitím údajov o uchádzačoch o zamestnanie dostupných už v čase ich zaradenia do evidencie, umožňuje pomocou matematických a štatistických metód strojového učenia odhadnúť rizikovosť osoby z hľadiska upadnutia do dlhodobej nezamestnanosti pred tým, ako tento problém nastane. Štatistické hodnotenie rizikovosti UoZ na základe dát je moderným prístupom, ktorý tradične prebieha buď na základe subjektívneho posúdenia pracovníka na úrade práce, alebo na základe rozhodovacích pravidiel napríklad podľa veku, pohlavia alebo vzdelania uchádzača.

Ciele práce sú: vyselektovať vhodné údaje na modelovanie závislej premennej dĺžky trvania evidencie UoZ, zvoliť vhodné modely, následne ich aplikovať na dáta a interpretovať výstupné hodnoty. V SR nebol v tejto oblasti publikovaný žiadny výskum.

V tejto práci využívame dáta z evidencie UoZ, ktorú vedie Ústredie práce, sociálnych vecí a rodiny. V prvej kapitole objasňujeme legislatívu Slovenskej republiky, základné pojmy a princípy týkajúce sa profilácie UoZ. V tejto kapitole vychádzame prevažne z nasledujúcich zdrojov [1, 12, 18, 23, 26, 31]. V druhej kapitole popisujeme spracované dáta, premenné, percentuálne zastúpenia v jednotlivých kategóriách premenných a postup spracovania

a odvodenia ďalších premenných, ktoré vo finálnej forme dátového súboru použitom pri modelovaní neboli obsiahnuté. V tretej kapitole uvádzame výsledky exploratívnej analýzy premenných, ktoré sa ukazujú ako významné prediktory na určenie celkovej dĺžky trvania evidencie. V štvrtej kapitole stručne popisujeme matematické pozadie použitých modelov. V tejto kapitole budeme vychádzame prevažne z nasledovných zdrojov [9, 15, 16, 17, 22, 25]. V piatej kapitole uvádzame výsledky predikcií, presnosť modelov, porovnanie modelov a významnosť premenných použitých pri modelovaní v prostredí R. Poslednú kapitolu tvorí diskusia.

1. Vymedzenie základných pojmov týkajúcich sa trhu práce

V tejto kapitole objasňujeme základné pojmy týkajúce sa zamestnanosti a nezamestnanosti definované v zákone o službách zamestnanosti [1] a princípy profilácie UoZ [18, 23, 26].

Uchádzač o zamestnanie (UoZ) je občan, ktorý môže pracovať, chce pracovať, hľadá si zamestnanie a je vedený v evidencii uchádzačov o zamestnanie, a zároveň nie je zamestnaný ani vedený ako samostatne zárobkovo činná osoba (SZČO). Takto definovaný občan sa podľa metodiky Ústredia práce, sociálnych vecí a rodiny (ÚPSVaR) definuje ako *nezamestnaný*. *Nezamestnanosťou* sa rozumie stav, kedy si osoba, ktorá chce a môže pracovať, hľadá zamestnanie sama alebo prostredníctvom príslušného orgánu - úradu práce, sociálnych vecí a rodiny. *Dlhodobo nezamestnaným* je občan vedený v evidencii uchádzačov o zamestnanie najmenej 12 po sebe nasledujúcich mesiacov. Takýto občan je zároveň podľa zákona o službách zamestnanosti zaradený k znevýhodneným uchádzačom o zamestnanie (znevýhodnenie c). Ďalšie znevýhodnené skupiny uchádzačov o zamestnanie sú:

a) občan mladší ako 26 rokov, ktorý ukončil príslušným stupňom vzdelania sústavnú prípravu na povolanie v dennej forme štúdia pred menej ako dvomi rokmi a od jej ukončenia nemal pravidelne platené zamestnanie (absolvent školy),
b) občan starší ako 50 rokov,
d) občan, ktorý dosiahol vzdelanie nižšie ako stredné odborné vzdelanie,
e) občan, ktorý najmenej 12 po sebe nasledujúcich kalendárnych mesiacov pred zaradením do evidencie uchádzačov o zamestnanie nemal pravidelne platené zamestnanie,
f) štátny príslušník tretej krajiny, ktorému bol udelený azyl,
g) občan, ktorý žije ako osamelá dospelá osoba s jednou alebo viacerými osobami odkázanými na jeho starostlivosť alebo starajúca sa aspoň o jedno dieťa pred skončením povinnej školskej dochádzky,
h) občan so zdravotným postihnutím.

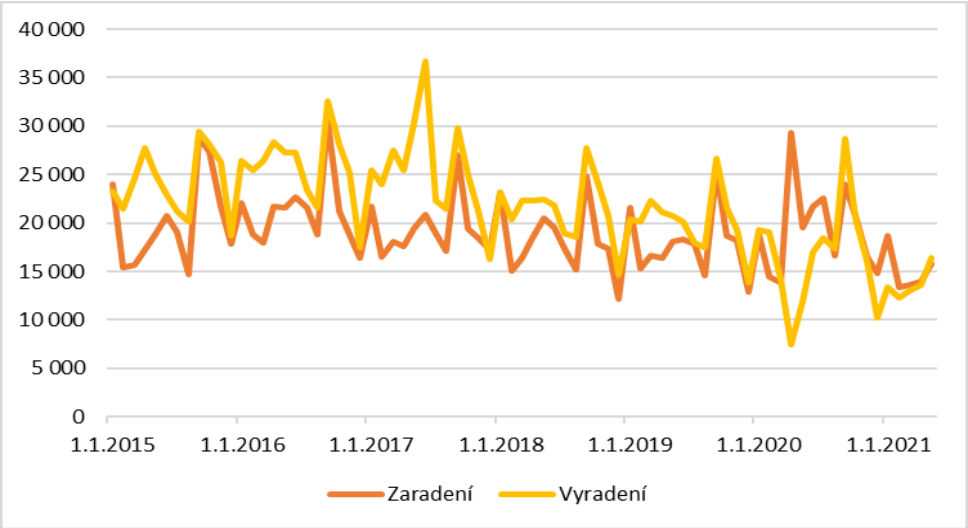
Definujme *mieru nezamestnanosti* ako podiel nezamestnaných k ekonomicky aktívnym osobám (teda voči všetkým pracujúcim aj nezamestnaným).

ÚPSVaR vo svojich mesačných štatistikách vykazuje dve miery nezamestnanosti [31]. Jedna je miera nezamestnanosti vypočítaná z celkového počtu UoZ, kde sa dáva do pomeru

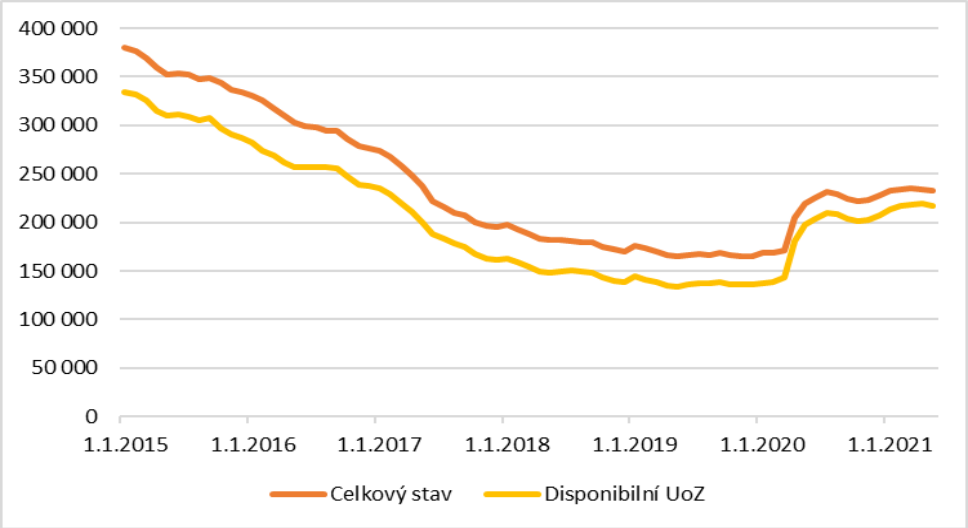
celkový počet UoZ k ekonomicky aktívnemu obyvateľstvu. Ekonomicky aktívne obyvateľstvo (EAO) sa skladá z nezamestnaných a pracujúcich. Údajmi o počte UoZ disponuje priamo ÚPSVaR. Dáta o pracujúcich sú čerpané zo Štatistického úradu Slovenskej republiky (ŠÚ SR). ŠÚ SR na určenie počtu pracujúcich používa metodiku výberového zisťovania pracovných síl (VZPS), ktorá definuje pracujúcich ako všetky osoby vo veku od 15 do 89 rokov, ktoré počas referenčného týždňa:

- vykonávajú aspoň jednu hodinu akúkoľvek prácu za mzdu, plat, odmenu alebo pracovnú činnosť za účelom dosiahnutia zisku (prácu na plný alebo kratší pracovný čas, stálu, dočasnú, príležitostnú alebo sezónnu prácu) vrátane osôb pracujúcich v zahraničí
- do 1 roka, osôb dochádzajúcich za prácou do zahraničia a osôb vykonávajúcich prácu na základe dohôd,
- majú prácu, ale nepracovali z dôvodu choroby, dovolenky, náhradného voľna alebo variabilného pracovného času, materskej alebo rodičovskej dovolenky, školenia súvisiaceho so zamestnaním, mimo sezóny, nedostatku práce u zamestnávateľa, štrajku a výluky, s výnimkou osôb na dlhodobom pracovnom voľne,
- patria sem aj vypomáhajúci (neplatení) členovia rodiny v rodinnom podniku alebo na rodinnej farme, ktorí za svoju činnosť nepoberajú žiadnu mzdu ani odmenu.

Druhou mierou nezamestnanosti, ktorú ÚPSVaR vykazuje je miera evidovanej nezamestnanosti. Táto miera sa používa v rôznych zákonoch SR [2, 4, 5] a i. Pri miere evidovanej nezamestnanosti sa do pomeru k EAO uvažujú len disponibilní UoZ. Disponibilní UoZ sú tí, ktorí nie si nedisponibilní. Nedisponibilní UoZ sú tí, ktorí poberajú PN, OČR, vykonávajú dobrovoľnícku službu, sú na absolventskej praxi a pod. Mesačne do evidencie pribudlo od januára 2015 do mája 2021 priemerne ~19 000 UoZ a priemerne odbudlo ~21 700. Vypočítané na základe dát uvedených na obr. 1.



Obr. 1: Vývoj zaradení a vyradení UoZ v evidencii ÚPSVaR. Počet zaradení a vyradení UoZ je za mesiac.



Obr. 2: Vývoj celkového stavu UoZ v evidencii ÚPSVaR. Počet UoZ je ku koncu mesiaca.

Z obr. 2 je možné vidieť, že počet UoZ od januára 2015 do marca 2020 klesol z 379 757 na 170 867. Po nástupe pandémie COVID-19 na Slovensku, bol zaznamenaný nárast o 34 000 UoZ (v mesiaci marec 2020 bolo evidovaných 170 867 a o mesiac neskôr už 204 480). Priemerne tvorili disponibilní UoZ 85,5 % z celkového počtu UoZ.

1.1. Princípy profilácie uchádzačov o zamestnanie (UoZ)

Odborné moderné prístupy sa už niekoľko rokov používajú ako alternatíva alebo ako komplementárne nástroje k tradičnej profilácii UoZ vo viacerých štátoch OECD. V zásade sa jedná o triedenie UoZ na rizikových, menej rizikových a bezrizikových. Najčastejšie sa jedná o riziko, že sa uchádzač nezamestná v stanovenom období. Môžeme sa na tento problém pozrieť

aj z pohľadu, ako dlho bude trvať dĺžka evidencie na úrade práce, teda ako dlho bude uchádzač nezamestnaný. V súčasnej dobe triedenie UoZ aplikujú krajiny ako napríklad Austrália, USA, Dánsko, Nemecko, Rakúsko a ďalšie [23].

Účel triedenia UoZ by sa dal rozdeliť do dvoch skupín. Prvý sa vzťahuje na individuálnu prácu s klientom a druhý k prideľovaniu programu aktívnych opatrení trhu práce (AOTP). AOTP označujú súbor opatrení určených na pomoc nezamestnaným skôr sa uplatniť na trhu práce. Medzi AOTP patria sprostredkovanie zamestnania, evidenčná činnosť, informačné a poradenské služby, sústavná príprava na povolanie, absolventská prax a iné [1, 28].

1.1.1. Individuálny prístup

Medzi nezamestnanými existuje značná heterogenita. Každý uchádzač má odlišné potreby a očakávania od Úradu práce. Európska stratégia zamestnanosti preto zdôrazňuje potrebu individuálneho prístupu pracovníkov ku klientom.

Vo viacerých krajinách sa ukazuje problém s nedostatočnou obsadenosťou ľudských zdrojov práve na pozíciách poradcov. S veľkou záťažou sa vyrovnávajú nasledujúcimi spôsobmi:

- skrátením dĺžky prípravy,
- skrátením dĺžky stretnutia,
- skupinovým stretnutiami (viac uchádzačov naraz),
- znížením frekvencie návštev na raz mesačne alebo menej často,
- rozdelením uchádzačov na skupiny s ktorými pracujú intenzívnejšie.

Jedným z riešení je tzv. zónové poradenstvo. V prvej zóne, ktorá je určená pre krátkodobých, nerizikových uchádzačov, si ľudia môžu hľadať prácu sami (napríklad za pomoci internetu). Stredne rizikový uchádzači, či uchádzači, ktorí sa do určitej doby nezamestnali spadajú do druhej zóny. V nej prebiehajú konzultácie s poradcom a je poskytnutá asistencia pri hľadaní nového zamestnania. V tretej zóne sú najrizikovejší uchádzači. V tejto zóne sa môžu aplikovať aj nákladnejšie nástroje aktívnej politiky trhu práce. Jedným z dôležitých predpokladov je spoľahlivý mechanizmus, ktorý dokáže uchádzačov rozdeliť do zón. [11, 28]

1.1.2. Cielené aktívne opatrenia na trhu práce

Účinky opatrení nie sú ovplyvnené iba kvalitou programu, dĺžkou, či financovaním, ale aj ďalšími faktormi, ktoré sa týkajú osobných charakteristík a situácie konkrétneho účastníka.

Správny výber uchádzačov, na ktorých sa budú vzťahovať aktívne opatrenia trhu práce, je jedným z kľúčových faktorov ovplyvňujúcim efekty opatrení. Pri tomto výbere sa vyskytujú dve základné chyby výberu. Prvý problém je, že do programu bývajú zaradení uchádzači, ktorí účasť nepotrebujú, pretože si dokážu nájsť prácu ľahko sami. Druhým problémom je, že do programu sú zaradení uchádzači, pre ktorých nie je program adekvátny. Profilácia a kategorizovanie uchádzačov môže slúžiť ako nástroj na spoľahlivejšie cielenie aktívnych opatrení na trhu práce. [11, 23, 28]

Pozíciu uchádzača na trhu práce môžeme odhadnúť viacerými spôsobmi, či ich vzájomnou kombináciou. Jeden spôsob predstavujú zákonom definované rizikové skupiny, či využitie dostupných dát o uchádzačovi a na základe nich odhadnúť dĺžku evidencie uchádzača.

1.1.3. Výhody a nevýhody vybraných prístupov

Preventívne triedenie uchádzačov podľa individuálnej miery zamestnateľnosti sa nazýva včasná identifikácia (angl. *early assessment*). Ide o súbor nástrojov, ktorého základom je posúdenie rizikovosti uchádzača po registrácii na úrade práce. Predikcia a triedenie uchádzačov sa prevádza uvedenými nástrojmi, prípadne ich kombináciou [28]:

- osobným hodnotením poradcu,
- skupinovým screeningom (definovaním rizikových skupín),
- profiláciou (predikcia na základe štatistického modelu),
- stanovením časových hraníc.

S problémom spoľahlivosti odhadu sa môžeme stretnúť pri koncepte včasnej identifikácie rizikových uchádzačov. Môžu nastať dva základné typy chýb, *false positives* a *false negatives* [28]:

- *false positives* – uchádzač je zaradený ako rizikový, ale v skutočnosti nie je. Tomuto uchádzačovi je navrhnutý program, ktorý nepotrebuje, čím sa zvyšujú náklady na včasnú identifikáciu a znižujú jej celkovú efektívnosť,
- *false negatives* – uchádzač je zaradený ako nerizikový, ale prácu si nenájde ani po stanovenej dobe. Tento problém sa dá riešiť zónovým poradenstvom.

Osobné hodnotenie uchádzačov vytvárajú pracovníci na úrade práce či externé osoby. Najčastejšie sa jedná o riadený rozhovor, cieľom ktorého je ohodnotiť uchádzačovu schopnosť

zamestnať sa. Najväčšia výhoda tohto prístupu je možnosť zistiť subjektívne dáta, ktoré sú objektívne veľmi ťažko merateľné napr. schopnosť plynule komunikovať, celkový dojem z uchádzača, vystupovanie a pod. Medzi nevýhody patria vysoké náklady na personál a osobné hodnotenie je závislé na schopnostiach pracovníkov úradu práce [11, 28].

Pri screeningu sú na základe skúseností z praxe vytipované skupiny osôb, ktoré sú na trhu práce rôznym spôsobom znevýhodnené. Tieto znevýhodnenia sme definovali na začiatku kapitoly 1 (str. 10).

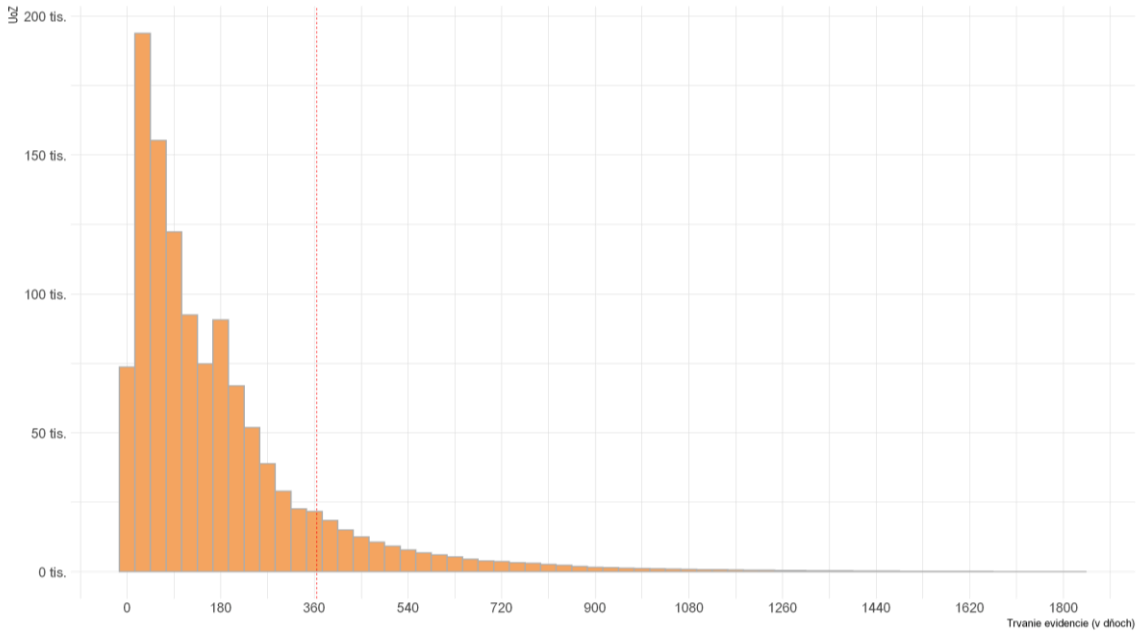
Profilácia kombinuje teóriu a empirické dáta o faktoroch trvania nezamestnanosti. Tieto bývajú hodnotené pomocou rôznych štatistických modelov, ktoré umožňujú hodnotiť mieru individuálneho rizika dlhodobej nezamestnanosti pre jednotlivých UoZ. Medzi ich výhody patrí vysoká spoľahlivosť, nízke náklady, prepájajú teóriu s empirickými dátami, umožňujú zohľadniť informácie o lokálnom trhu práce, a zároveň poskytujú podklad pre objektívne rozhodnutia, ktoré nie sú závislé na subjektívnom posúdení poradcu. Hlavné nevýhody modelov sú, že vytvorenie modelu je komplexná úloha, ktorá si vyžaduje prístup k relevantným štatistickým údajom a pre niektorých nezamestnaných môže byť model stigmatizujúci [28]. Ďalšou nevýhodou je skutočnosť, že niektoré typy modelov môžu obsahovať skryté chyby alebo nedostatky. Ako príklad uvádzame chybné vyhodnocovanie absolventov škôl, u ktorých sa vo väčšine prípadov jedná o sezónnu nezamestnanosť, vyhodnotiť zle, v prípade ak by nastal na trhu práce nesúlad medzi odbormi produkovanými zo škôl a odbormi dopytovanými zamestnávateľmi. Konkrétne ak absolventi školy použití pri budovaní modelu majú priemerné trvanie evidencie 130 dní, zjednodušene môžeme predpokladať, že pre absolventa školy, ktorý sa aktuálne zaregistruje bude mať trvanie evidencie približne 130 dní. Ak však pre tohto absolventa nie voľné pracovné miesto v cieľovej oblasti, kde má záujem pracovať, môže to byť aj dlhšie. Situácie, ktoré sú ojedinelé a výnimočné je vo všeobecnosti ťažké odhadnúť.

Popri ohodnotení poradcom či štatistickým modelom sa v zahraničí (napr. Spojenom kráľovstve) využívajú tzv. časové hranice [23]. Pokiaľ si uchádzač nenájde prácu do dopredu známeho stanoveného termínu (napr. 6 mesiacov), dostáva sa automaticky do zóny s intenzívnejším poradenstvom. Časové hranice sa v praxi samostatne nepoužívajú takmer nikde. Použitie časových hraníc samostatne nie je najvhodnejšie najmä pri uchádzačoch, ktorí sú evidentne veľmi rizikoví a je zbytočné čakať stanovené mesiace na nájdenie práce. Väčšinou časové hranice slúžia ako *záchranná sieť* pre prípad, kedy poradca alebo štatistický model odhadne uchádzača ako nerizikového, ale napriek tomu by upadol do dlhodobej nezamestnanosti. [11, 23, 28]

2. Popis závislej premennej a nezávislých premenných použitých pri tvorbe štatistických modelov

V tejto kapitole uvedieme popis (nezávislých) kategorických a numerických premenných použitých v nasledujúcej časti práce. Finálne sme pracovali so 45 kategorickými premennými (pohlavie, rodinný stav, národnosť, štátna príslušnosť, zdravotné postihnutie, školský odbor, počet detí do 3 rokov, od 4 do 6 rokov, od 7 do 10 rokov a od 11 do 16 rokov, záujem o úväzok, záujem dochádzať do zamestnania, vzdelanie, kraj, oblasť posledného zamestnania, pozícia posledného zamestnania, mesiac zaradenia, jazyk SK, CZ, EN, DE, RU, PL, HU, iný svetový a iný, MS Office a PC iné, titul, SZČO, dobrovoľne nezamestnaný, prísľub zamestnania, úmysel ísť do zahraničia, záujem zamestnania v zahraničí, záujem SZČO, záujem vzdelávať sa, záujem o absolventskú prax, záujem o sústavnú prípravu na povolanie, skupina základný vodičský preukaz, skupina profesionálny vodičský preukaz, znevýhodnenia a), b), d), e) a kategória znevýhodnení f), g) a iné) a so 4 numerickými premennými (počet registrácií za posledné dva roky, posledné trvanie, vek, počet znevýhodnení).

Zo zozbieraných dát, ktoré nám poskytol ÚPSVaR, sme zostrojili (závislú) premennú *y*, ktorá poskytuje informáciu o dĺžke trvania evidencie uchádzačov v dňoch. Tá je vyjadrená ako počet dní, ktorý uplynul od dátumu ich zaradenia do evidencie až po dátum vyradenia z evidencie. Znázornená je na obr. 3, kde sú na osi y nanesené počty UoZ v tisícoch a na osi x je trvanie evidencie v dňoch:

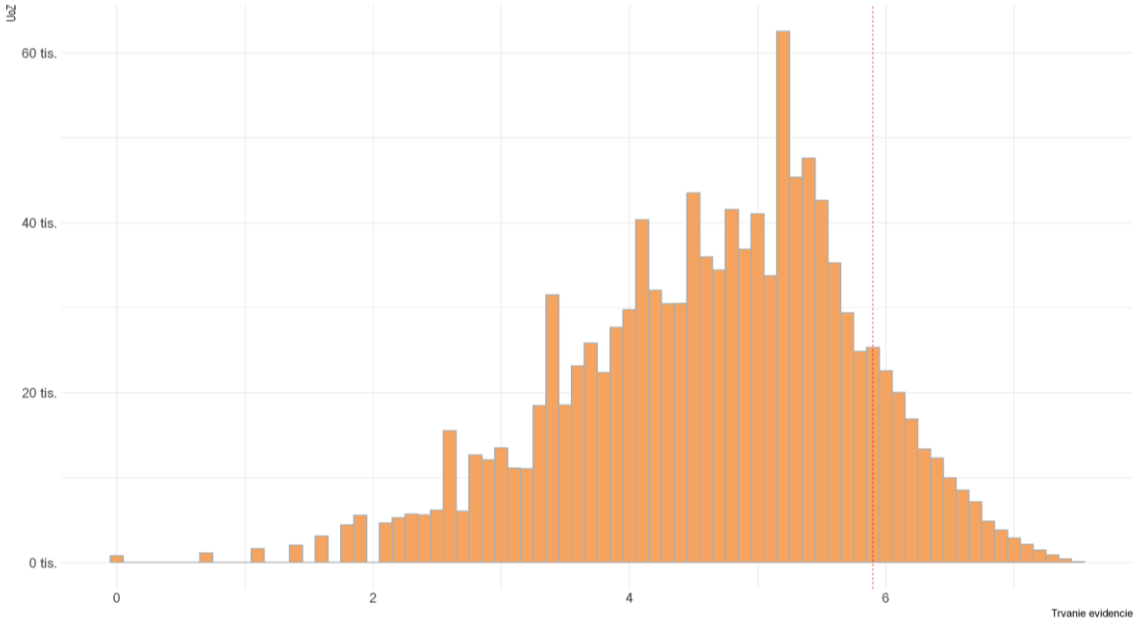


Obr. 3: Frekvenčná analýza dĺžky evidencie UoZ, ktorý boli zaradení do evidencie od 01.01.2015, a zároveň vyradení do 01.03.2018. Priemerná dĺžka evidencie UoZ činila 176,5 dňa. Červená čiara označuje hranicu jedného roku od začiatku evidencie.

Na úpravu šikmých dát sme použili logaritmickú transformáciu, podľa [14, 21]:

$$\tilde{y} = \log(1 + y). \quad (1)$$

Túto transformáciu sme využili v kapitole 5 (str. 48). Výsledná premenná y po transformácii ($\tilde{y}$) je zobrazená na obr. 4, kde na osi y sú počty UoZ v tisícoch a na osi x je trvanie evidencie.

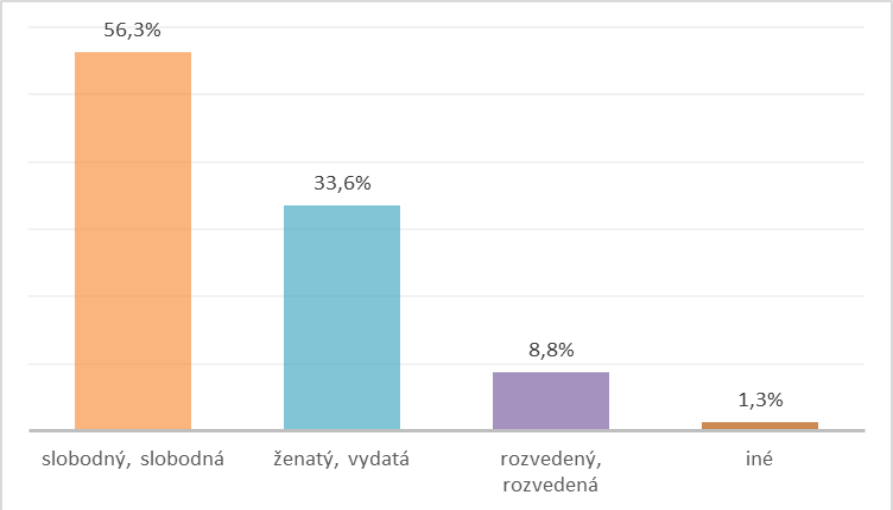


Obr. 4: Frekvenčná analýza trvania evidencie – logaritmická transformácia premennej y. Červená čiara označuje hranicu zlogaritmovaného jedného roku od začiatku evidencie.

Dátový súbor, s ktorým sme pracovali obsahoval celkovo 1,9 milióna zaradení UoZ do evidencie v rokoch 2013 až 2018, ku ktorým prislúchalo 39 unikátnych informácií (nezávislých premenných). V rámci našej práce sme však využívali iba dáta z rokov 2015 až 2018 (1,2 milióna zaradení). Pôvodné dáta sme v prvom kroku zredukovali na údaje od 01.01.2015, aby sme pri každej použitej registrácii UoZ disponovali vždy aj informáciu o dvojročnom období, ktoré jej predchádzalo. To nám umožnilo odvodiť ďalšiu premennú - počet registrácií za posledné dva roky. V druhom kroku sme z dátového súboru vylúčili takých UoZ, ktorí nemali ukončenú evidenciu k 01.03.2018, a teda pre nich nebolo možné určiť dĺžku trvania nezamestnanosti. Z niektorých nezávislých premenných sme potom vyvodili aj niektoré ďalšie, ktoré podrobne popisujeme nižšie. Niektoré iné sme naopak z rôznych dôvodov nepoužili vôbec, napr. premennú dôvod vyradenia uchádzača z evidencie, pretože tento dôvod nie je známy v čase zaradenia do evidencie, kedy chceme modelovať dĺžku nezamestnanosti. Vyradenou premennou je rovnako aj okres prechodného pobytu, nakoľko sa táto informácia vyskytovala v celkovom zastúpení dát zriedkavo a jednotlivé kategórie nemali dostatočné zastúpenie a ďalšie. Premenné, ktoré obsahovali priveľa kategorických zaradení sme

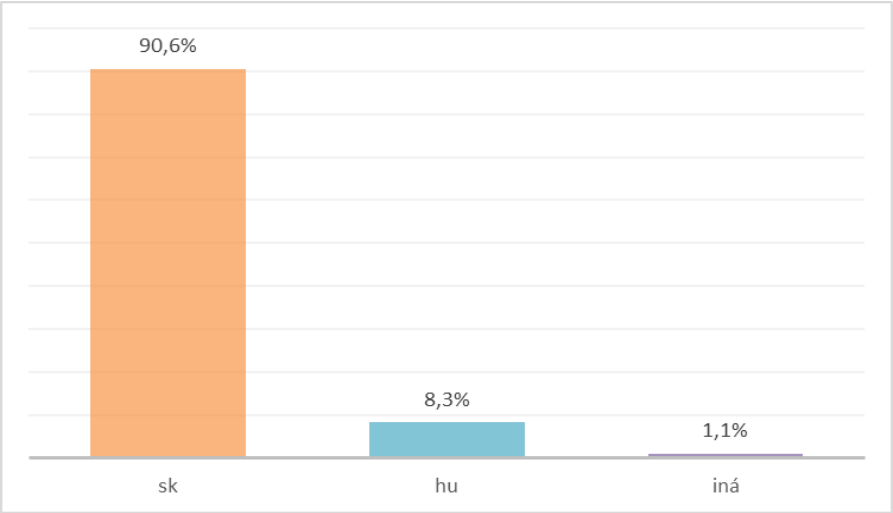
zredukovali na reprezentatívne kategórie (napr. premenná pozícia posledného zamestnania pôvodne obsahovala 10 000 veľmi špecifických kategórií, podľa klasifikácie zamestnaní SK ISCO-08 [6], z ktorých niektoré neboli zastúpené vôbec. Preto sme pre túto premennú použili najvšeobecnejšie kategórie, ktoré zastupujú hlavné triedy profesií.). Podobne sme nepoužili pôvodnú premennú okresu trvalého pobytu uchádzača o zamestnanie, ale kraj, nakoľko v jednotlivých okresoch nebolo dostatočné zastúpenie. K pôvodnému dátovému súboru z údajov UoZ, ktorú vedie ÚPSVaR, sme priradili dodatočné informácie, rovnako z údajov UoZ o tom, aké znevýhodnenia boli aktívne v čase zaradenia, aké konkrétne jazyky a počítačové zručnosti uchádzač ovláda, a aký má typ vodičského preukazu.

V ďalšej časti postupne charakterizujeme nezávislé premenné a percentuálne zastúpenie uchádzačov v jednotlivých kategóriách. Z hľadiska (a) pohlavia medzi UoZ prevažujú muži 52,1 %, ženy tvoria 47,9 %. Ďalšou premennou je (b) rodinný stav UoZ s kategóriami slobodný, rozvedený, ženatý a iné. Do kategórie iné patria vdovec, nezistené a registrované partnerstvo. Najvyššie zastúpenie medzi UoZ majú občania, ktorí sú slobodný - 56,3%. Jednotlivé percentuálne zastúpenia zachytáva obr. 5.



Obr. 5: Percentuálne zastúpenie UoZ na základe ich rodinného stavu.

UoZ sme z hľadiska (c) národnosti roztriedili do 3 kategórií - slovenská, maďarská a iná. Do kategórie iné boli zaradení UoZ minoritnejších národností napr. rómskej, ukrajinskej, českej, rumunskej, poľskej, atď. Podľa očakávania najpočetnejšie zastúpenie - 90,6% tvorili občania slovenskej národnosti (obr. 6).



Obr. 6: Percentuálne zastúpenie UoZ na základe ich národnosti

Podobne pri premennej (d) štátna príslušnosť podľa očakávania dominovali slovenskí občania (99,4%), cudzinci medzi UoZ tvorili iba 0,6 %.

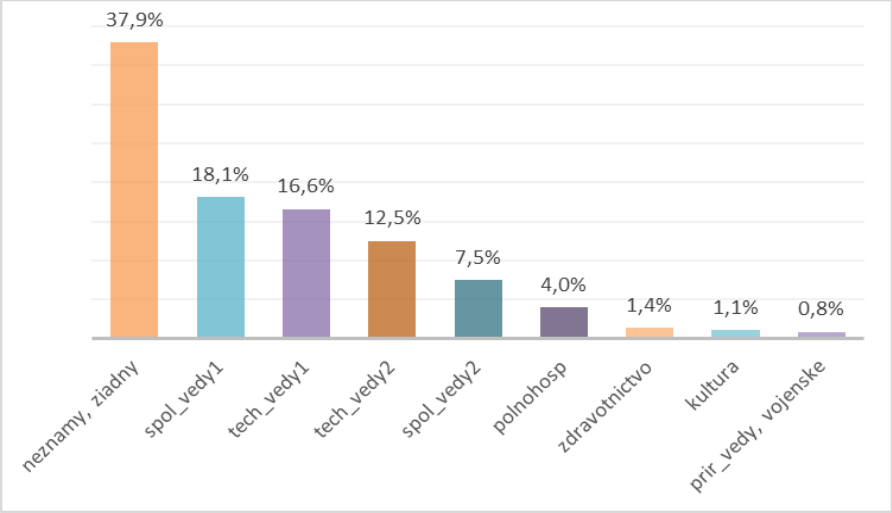
Premennú (e) zdravotné postihnutie sme zadefinovali ako binárnu premennú áno/nie podľa toho, či UoZ mal niektoré zo zdravotných znevýhodnení: telesné, duševné, zrakové, sluchové, viacnásobné alebo iné v kategóriách do 20 %, od 20 % do 40 %, od 41 % do 70 %, nad 70 %. Tzn. ak UoZ uviedol akékoľvek zdravotné postihnutie aj minimálne do 20% bol zaradený do kategórie áno rovnako ako UoZ so znevýhodnením > 70%. Zdraví UoZ/UoZ neuvádzajúci zdravotné znevýhodnenie tvorili 97,8%, Znevýhodnení UoZ 2,2%.

Ďalšia premenná (f) školský odbor poskytla informáciu o odbore, v ktorom UoZ dosiahol najvyššie vzdelanie. Rôzne odbory sme zosumarizovali podľa štatistickej klasifikácie odborov vzdelania [7] do 9 kategórií:

- Kultúra – odbory dejiny a teória umenia, hudby, architektúry, hra na hudobné nástroje, skladba a dirigovanie, výtvarná tvorba, reštaurovanie, herectvo, remeselná výroba a pod.;
- Poľnohospodárstvo – agropodnikanie, záhradníctvo, lesníctvo, chovateľstvo, živočíšna výroba, veterinárne zdravotníctvo, pestovateľ a pod.;
- Zdravotníctvo – lekárske vedy, farmaceutické vedy;
- Technické vedy 1 – baníctvo a banícka geológia, hutníctvo, strojárstvo, informatika a výpočtová technika, elektrotechnika, technická chémia, potravinárstvo a pod.;

- Technické vedy 2 - textil a odevníctvo, spracúvanie kože, plastov, papiera, dreva, gumy, výroba obuvi, hudobných nástrojov, polygrafia, architektúra, stavebníctvo, doprava, automatizácia a riadenie, špeciálne technické odbory a pod.;
- Spoločenské vedy 1 - filozofické, ekonomické, ekonomika a organizácia, obchod a služby, politické a právne vedy a pod.;
- Spoločenské vedy 2 – historické, pedagogické, psychologické vedy, publicistika, knihovníctvo, vedy o telesnej kultúre, učiteľstvo, úplné stredné vzdelanie na gymnáziu a pod.;
- Prírodné a vojenské vedy – fyzikálno-matematické, geologické, geografické, chemické, biologické vedy, ochrana životného prostredia a vojenské odbory, bezpečnostné služby a pod.;
- Neznámy, žiadny – do tejto kategórie patria uchádzači s neuvedeným školským odborom a tí, ktorí nemajú ukončené vzdelanie.

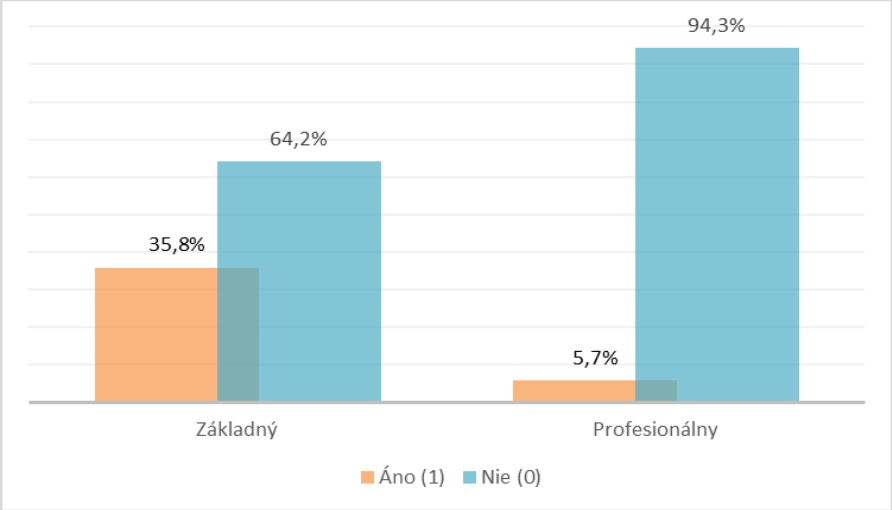
Percentuálne zastúpenie uchádzačov v jednotlivých kategóriách zachytáva obr. 7. Zo zozbieraných údajov vyplýva, že najpočetnejšie zastúpenie má kategória - neznámy, žiadny, ktorá obsahuje až 37,9% všetkých uchádzačov o zamestnanie.



Obr. 7: Percentuálne zastúpenie UoZ na základe školského odboru, v ktorom dosiahli najvyššie vzdelanie

Pôvodne sme hodnotili aj premennú posledný školský odbor, ktorá poskytovala informácie o školskom odbore posledného dosiahnutého vzdelania. Napokon sme ju ale pri modelovaní nepoužili, kvôli veľkej podobnosti s premennou školský odbor. Posledné dosiahnuté vzdelanie je totiž vo väčšine prípadov aj najvyšším dosiahnutým vzdelaním UoZ.

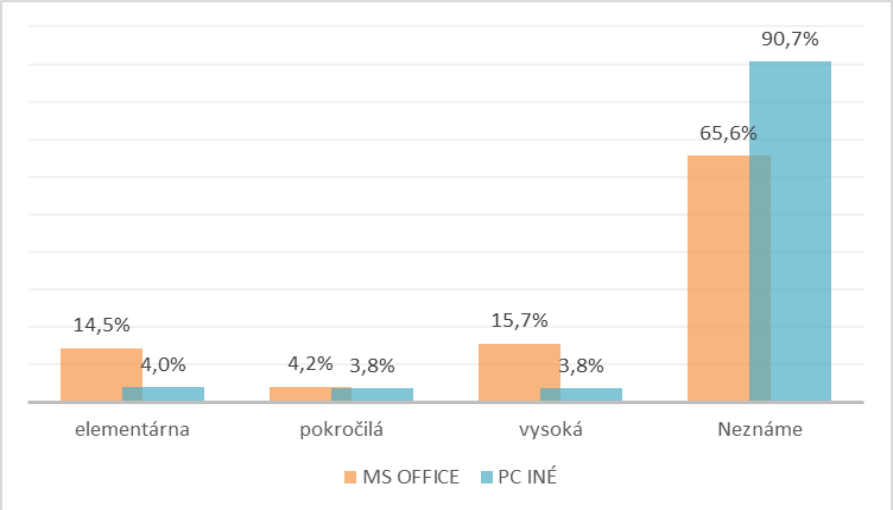
Pôvodne binárnou premennou bol aj (g) vodičský preukaz s hodnotami 0/1 (nie/áno). Túto premennú sme nahradili konkrétnymi skupinami základného a profesionálneho vodičského preukazu. Výsledne sme tak pracovali s 2 premennými, ktoré nám poskytli detailnejšiu informáciu o vodičských schopnostiach UoZ. Do (g1) základnej skupiny vodičského oprávnenia patrili všetci uchádzači, ktorý mali aspoň jedno z oprávnení A, A1, A2, AM, B, B1 a do skupiny (g2) profesionálneho vodičského oprávnenia tí, ktorý má aspoň jedno z oprávnení BE, C, C1, C1E, D, DE, D1, D1E, E, T. Percentuálne zastúpenie v jednotlivých skupinách zachytáva obr. 8.



Obr. 8: Percentuálne zastúpenie UoZ podľa schopnosti viesť motorové vozidlo a type vodičského oprávnenia

Podobný postup ako pre premennej (g) sme použili aj pre binárne premenné (h) práca s počítačom, s kategóriami 0/1 (neovláda prácu s počítačom/ovláda prácu s počítačom), a (i) cudzí jazyk, tiež s kategóriami 0/1 (neovláda cudzí jazyk/ovláda cudzí jazyk). Ak UoZ ovládal prácu s počítačom, tak sme mu priradili informáciu o tom, čo konkrétne z tejto oblasti ovláda, a ďalej sme pracovali už iba s touto rozšírenou informáciou. Keďže pôvodných kategórií bolo veľmi veľa (boli veľmi vyšpecifikované, a zároveň viaceré z nich neboli UoZ zastúpené vôbec, alebo len veľmi málo), rozhodli sme sa vytvoriť z údajov iba dve premenné (h1) MS Office a (h2) PC iné. V rámci premennej Microsoft Office boli zaradení UoZ, ktorí uviedli schopnosť používať aspoň jeden program z balíku Microsoft Office (Outlook, PowerPoint, Excel, Word, Lotus notes). Do premennej PC iné patrili UoZ, ktorý ovládajú aspoň jednu z uvedených pokročilejších znalostí: grafické nástroje (Adobe Photoshop, Illustrator, Corel Draw, Gimp, DTP a pod.), správa systémov (Microsoft Windows, MSSQL, Active Directory, ORACLE SQL, Linux a pod.), tvorba webu (HTML, XHTML, AJAX, Javascript a pod.), programovanie

(C, C++, Java, Delphi, SQL, Visual Basic a pod.), databázy, CAD, a ekonomické programy. Úroveň znalostí okrem toho rozlišujeme na elementárnu, pokročilú, vysokú a neznámu. Do kategórie neznáme patrili uchádzači, ktorí nemajú znalosti v danej oblasti a do kategórie elementárna sme zaradili aj tých uchádzačov, ktorí uviedli, že znalosť ovládajú, ale už neuviedli stupeň znalosti. Zo získaných údajov vyplýva, že iba 34,6 % UoZ ovláda MS Office minimálne na elementárnej úrovni, pričom iba 9,3% UoZ je schopná používať aj nejaký pokročilejší softvér (obr.9).



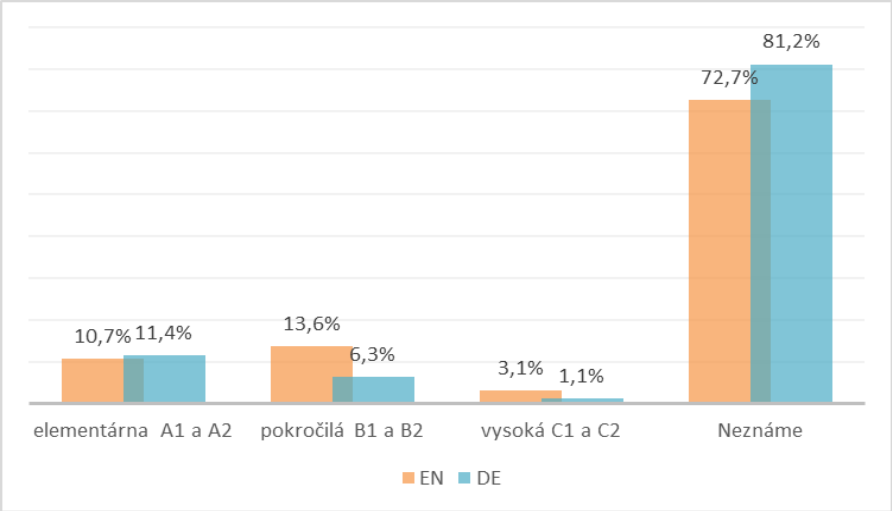
Obr. 9: Percentuálne zastúpenie UoZ na základe schopností práce s PC

Pri premennej cudzí jazyk sme vytvorili kategórie slovenský, český, anglický, nemecký, ruský, poľský, maďarský, iný svetový jazyk a iný jazyk. Do kategórie iné svetové jazyky patria francúzsky, taliansky, španielsky a pod. Do kategórie iné jazyky patria rumunský, maltský, slovinský, chorvátsky a pod. Tieto premenné pôvodne obsahovali úrovne jazyka, ktorí UoZ ovládajú. Konkrétne elementárna A1 a A2, pokročilá B1 a B2, vysoká C1 a C2. Avšak z dôvodu nedostatočného zastúpenia v spomínaných úrovniach jazyka sme ponechali iba informáciu o tom, či UoZ ovláda daný jazyk (1) alebo nie (0). Jednotlivé percentuálne zastúpenia zachytáva tab. 1. Uviedli sme v nej aj slovenský jazyk, nakoľko sme predpokladali, že pre občanov iných národností by mohla byť jeho neznalosť prediktorom pre horšie uplatnenie na trhu práce. Uvedená kategória však poskytla problematické údaje, väčšina UoZ ju zjavne vníma automaticky, pričom iba 1,8% uchádzačov ju uviedlo ako znalosť. Jazyky anglický a nemecký uvádzame osobitne (obr. 10) nakoľko sme pri nich sledovali aj úrovne znalostí UoZ – elementárnu A1 a A2, pokročilú B1 a B2, vysokú C1 a C2 a neznámu. Na obr. 11 môžeme vidieť, že UoZ, ktorí ovládajú anglický alebo nemecký jazyk aspoň na elementárnej

úrovni A1 a A2, tvoria 27,3 % (EN) a 18,8 % (DE). Do elementárnej úrovne A1 a A2 sme zaradili aj UoZ, ktorí uviedli, že daný jazyk ovládajú, ale neuviedli na akej úrovni.

Tab. 1: Percentuálne zastúpenie UoZ v premenných týkajúcich sa jazykových zručností

Jazyk	Nie (0)	Áno (1)
SK	98,2 %	1,8 %
CZ	99,1 %	0,9 %
RU	94,6 %	5,4 %
PL	99,2 %	0,8 %
HU	94,6 %	5,4 %
Iný svetový jazyk	97,0 %	3,0 %
Iný jazyk	94,9 %	6,1 %



Obr. 10: Percentuálne zastúpenie UoZ na základe ovládanie anglického alebo nemeckého jazyka

Pre premenné počet detí do 3 rokov, od 4 do 6 rokov, od 7 do 10 rokov a od 11 do 16 rokov k dátumu zaradenia do evidencie sme vytvorili kategórie 0 (žiadne dieťa) a 1 a viac. Percentuálne zastúpenia sú v uvedené tab. 2.

Tab. 2: Percentuálne zastúpenie UoZ v premenných týkajúcich sa počtu detí v čase zaradenia do evidencie

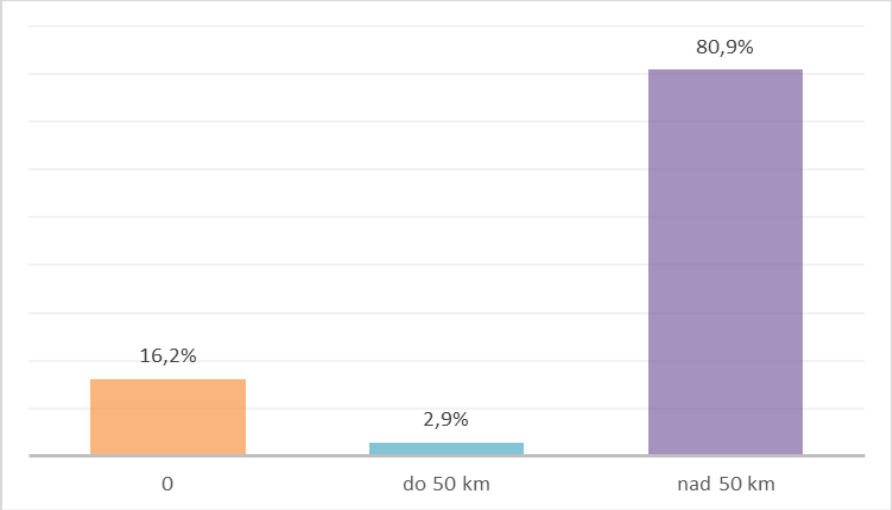
Počet detí	0	1 a viac
do 3 rokov	96,6 %	3,4 %
od 4 do 6 rokov	96,7 %	3,3 %
od 7 do 10 rokov	96,5 %	3,5 %
od 11 do 16 rokov	96,7 %	3,3 %

Ďalšie premenné sú binárne s hodnotami 0/1 (nie/áno), ktoré nesú informáciu o záujmoch , úmysloch a minulosti uchádzača. Napríklad premenná (k) SZČO a (l) dobrovoľne nezamestnaný hovoria o tom, či daný uchádzač niekedy v minulosti mal živnosť alebo bol dobrovoľne nezamestnaný. K záujmovým premenným patria (m) záujem o vykonávanie samostatnej zárobkovej činnosti, (n) záujem pracovať aj mimo územia SR, (o) záujem vzdelávať sa s cieľom zvýšiť uplatnenie na trhu práce, (p) záujem o absolventskú prax a (q) záujem o sústavnú prípravu na povolanie. K ďalším binárnym premenným patrí informácia o (r) úmysle odísť do zahraničia a (s) prísľub zamestnania. (t) Premennú titul sme vytvorili na základe akademického titulu pred menom. Percentuálne zastúpenie v jednotlivých kategóriách spomínaných premenných je v tab. 3.

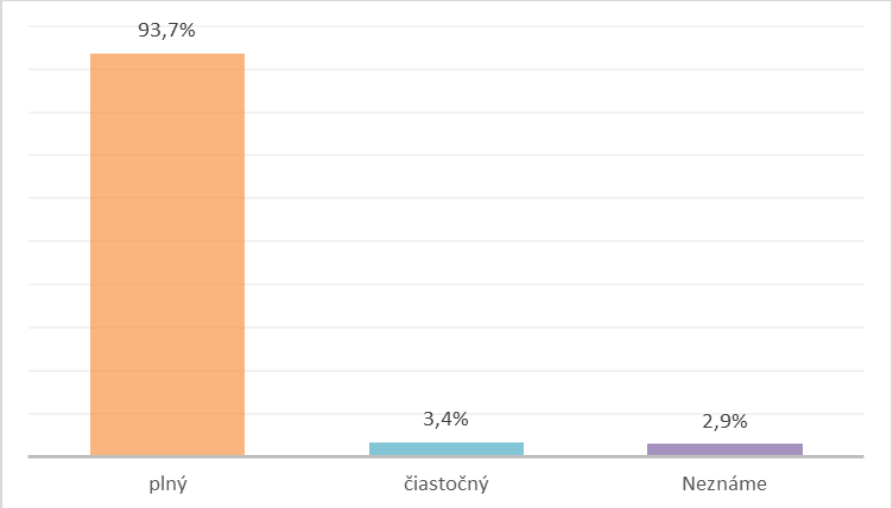
Tab. 3: Percentuálne zastúpenie UoZ podľa ich záujmov, úmyslov a ich minulosti

Premenná	0 (Nie)	1 (Áno)
(k) SZČO	99,8 %	0,2 %
(l) Dobrovoľne nezamestnaný	39,9 %	60,1 %
(m) Záujem SZČO	94,5 %	5,5 %
(n) Záujem zamestnania v zahraničí	79,3 %	20,7 %
(o) Záujem vzdelávať sa	74,1 %	25,9 %
(p) Záujem o absolventskú prax	95,9 %	4,1 %
(q) Záujem o sústavnú prípravu	97,1 %	2,9 %
(r) Úmysel ísť do zahraničia	97,2 %	2,8 %
(s) Prísľub zamestnania	99,8 %	0,2 %
(t) Titul	81,8 %	18,8 %

Ďalšími záujmovými premennými sú (u) ochota/záujem dochádzať do zamestnania, ktorá má 3 kategórie 0, do 50 km a nad 50 km a (v) typ úväzku o aký má UoZ záujem (plný, čiastočný alebo neznámy), čo je znázornené na obr. 11 – 12.

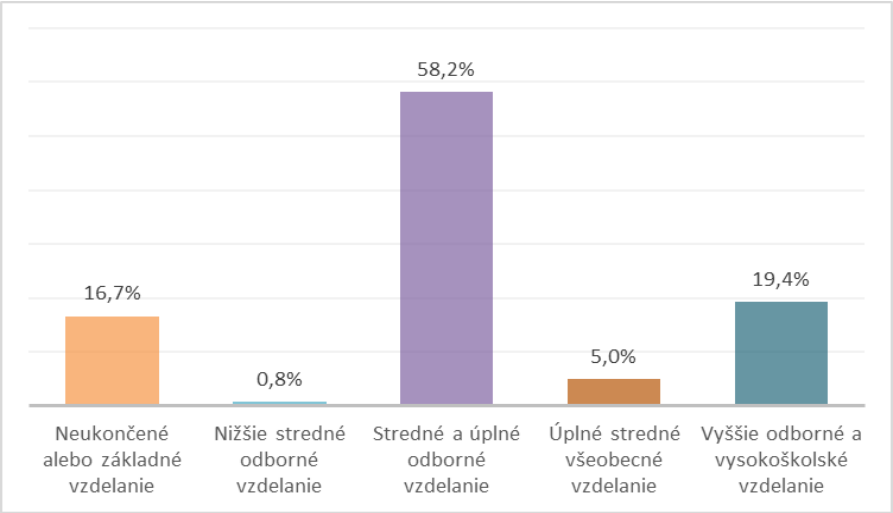


Obr. 11: Percentuálne zastúpenie UoZ rozdelené podľa ochoty akceptovať dochádzanie do zamestnania.



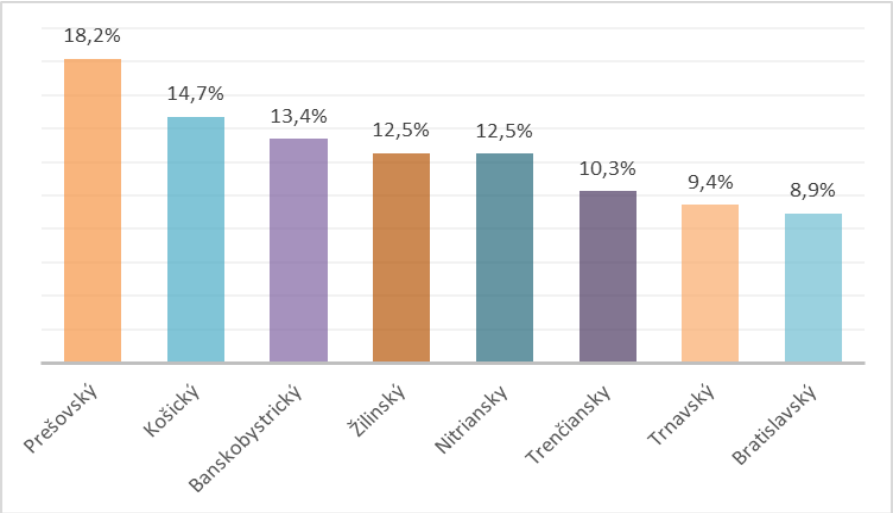
Obr. 12: Percentuálne zastúpenie UoZ podľa záujmu o konkrétny typ úväzku

Premenná (x) vzdelanie poskytuje informáciu o stupni dosiahnutého vzdelania. Obsahuje 5 kategórií neukončené alebo základné vzdelanie, nižšie stredné odborné vzdelanie, stredné alebo úplné stredné odborné vzdelanie, úplné stredné všeobecné vzdelanie a vyššie odborné vzdelanie alebo vysokoškolské vzdelanie prvého, druhého alebo tretieho stupňa. Porovnanie pre jednotlivé kategórie v premennej vzdelanie môžeme vidieť na obr. 13.



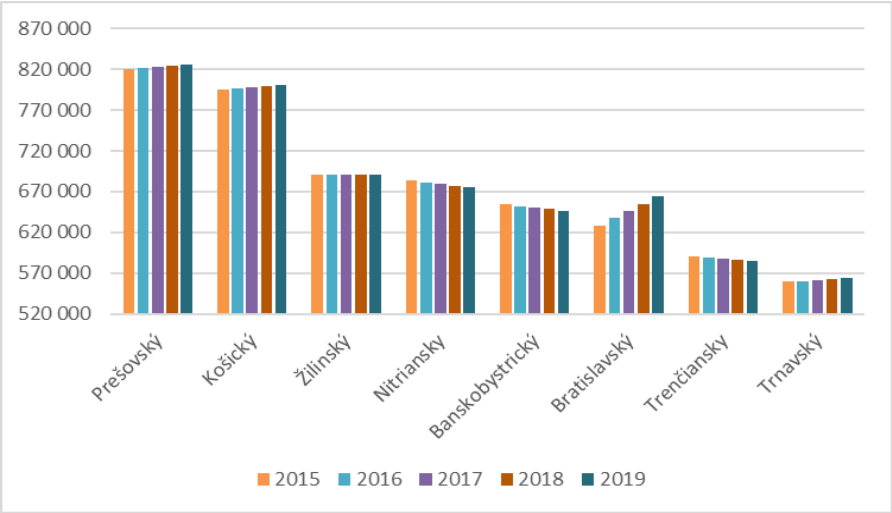
Obr. 13: Percentuálne zastúpenie UoZ na základe ich najvyššieho dosiahnutého vzdelania

Ďalšou informáciou o uchádzačovi bol (y) kraj trvalého pobytu. Najväčšie zastúpenie uchádzačov v dátach je v Prešovskom kraji 18,2 %. Najmenšie zastúpenie je v Bratislavskom kraji 8,9 %. Percentuálne zastúpenie UoZ podľa krajov zachytáva obr. 14.



Obr. 14: Percentuálne zastúpenie UoZ podľa ich kraja (podľa trvalého pobytu)

Podľa údajov zo štatistického portálu SR o priemernom počte obyvateľov spracovaný bilančnou metódou k 1.júlu daného roka [29], je najväčší kraj Prešovský, čo by mohlo byť dôvodom najvyššieho percentuálneho zastúpenia medzi uchádzačmi. Na obr. 15 vidíme, že najväčší percentuálny nárast priemerného počtu obyvateľstva je v Bratislavskom kraji 5,4 % (od začiatku roka 2015 do konca roka 2019).



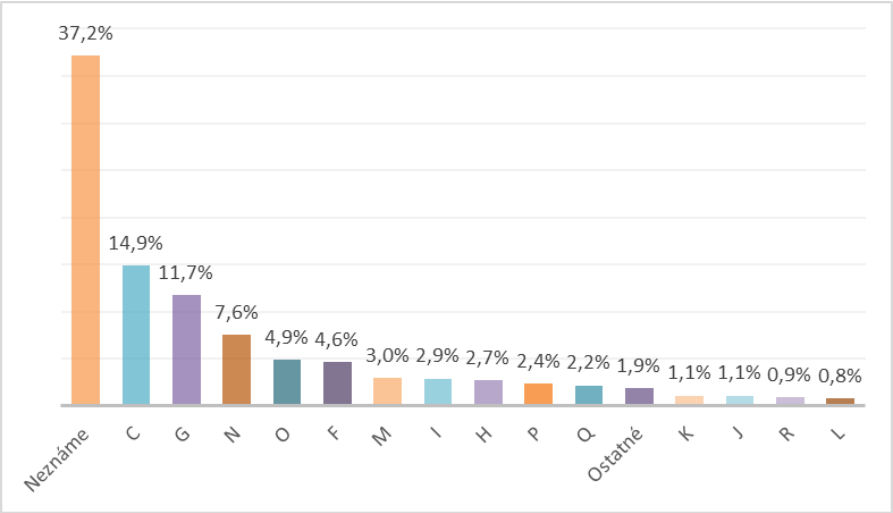
Obr. 15: Priemerný počet obyvateľov SR podľa krajov v rokoch 2015 - 2019, zoradené od najvyššieho počtu po najmenší podľa údajov z roku 2015

Premenná (z) oblasť posledného zamestnania obsahuje celkovo 16 kategórií, ktoré sú rozdelené podľa SK NACE Rev.2 klasifikácie odvetví [3] nasledovne:

- C - Priemyselná výroba: výroba potravín, nápojov, tabakových výrobkov, textilu, odevov, kože a kožených výrobkov, koksu a rafinovaných ropných produktov, chemikálií, elektrických zariadení, strojov, nábytku, papiera a pod.,
- F - Stavebníctvo: výstavba budov, inžinierske stavby, špecializované stavebné práce a pod.,
- G - Veľkoobchod a maloobchod: veľkoobchod a maloobchod a oprava motorových vozidiel a motocyklov,
- H - Doprava a skladovanie: pozemná, vodná, letecká doprava a doprava potrubím, skladové a pomocné činnosti v doprave, poštové služby a služby kuriérov a pod.,
- I - Ubytovacie a stravovacie služby: ubytovanie, činnosti reštaurácií a pohostinstiev,
- J - Informácie a komunikácia: nakladateľské činnosti, výroba filmov, videozáznamov a televíznych programov, príprava a zverejňovanie zvukových nahrávok, činnosti pre rozhlasové a televízne vysielanie, informačné služby, počítačové programovanie, poradenstvo a súvisiace služby a pod.,
- K - Finančné a poisťovacie činnosti: finančné služby, poistenia, zaistenia a dôchodkového zabezpečenia okrem povinného sociálneho poistenia, pomocné činnosti finančných služieb a poistenia a pod.,
- L - Činnosti v oblasti nehnuteľností,

- M - Odborné, vedecké a technické činnosti: právne a účtovnícke činnosti, vedenie firiem, architektonické a inžinierske činnosti, vedecký výskum a vývoj, reklama a prieskum trhu, veterinárne činnosti,
- N - Administratívne a podporné služby: prenájom a lízing, sprostredkovanie práce, činnosti cestovných agentúr, rezervačné služby cestovných kancelárií a súvisiace činnosti, bezpečnostné a pátracie služby, činnosti súvisiace s údržbou zariadení a krajinnou úpravou, administratívne, pomocné kancelárske a iné obchodné pomocné činnosti a pod.,
- O - Verejná správa a obrana: verejná správa a obrana, povinné sociálne zabezpečenie a pod.,
- P - Vzdelanie,
- Q - Zdravotníctvo a sociálna pomoc,
- R - Umenie, zábava a rekreácia: tvorivé, umelecké a zábavné činnosti, činnosti knižníc, archívov, múzeí a ostatných kultúrnych zariadení, činnosti herní a stávkových kancelárií, športové, zábavné a rekreačné činnosti,
- Ostatné (A, B, D, E, T, U, S): dodávka elektriny, plynu, pary, studeného vzduchu a vody, činnosti domácností ako zamestnávateľov, čistenie a odvod odpadových vôd, zber, spracúvanie a likvidácia odpadov, ozdravovacie činnosti a ostatné činnosti nakladania s odpadom, činnosti extrateritoriálnych organizácií a združení, oprava počítačov, osobných potrieb a potrieb pre domácnosti a pod.,
- Neznáme - uchádzači, ktorí neuviedli oblasť posledného zamestnania alebo predtým nikde nepracovali

Na obr. 16 pozorujeme, že 62,8 % UoZ uviedlo informáciu o odvetví posledného zamestnania. Naopak o 37,2 % UoZ je oblasť posledného zamestnania neznáma. Druhá najviac zastúpená kategória je odvetvie priemyselnej výroby, s celkovým zastúpením 14,9 %. Odvetvia, v ktorých je percentuálne zastúpenie UoZ menej ako 1 %, sú činnosti v oblasti nehnuteľností a umenie, zábava a rekreácia.



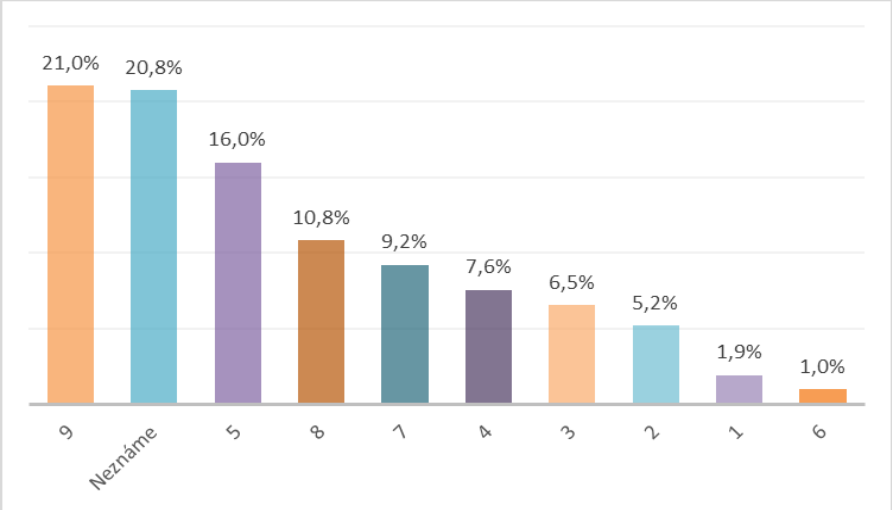
Obr. 16: Percentuálne zastúpenie UoZ na základe oblasti posledného zamestnania

Premenná (ab) pozícia posledného zamestnania obsahuje 10 kategórií (označenia kategórií sú hlavné triedy SK ISCO-08 [6]):

- 1 - Zákonodarcovia, riadiaci pracovníci: členovia vlády, prezident, riadiaci pracovníci (manažéri) administratívnych, podporných a obchodných činností, rektor, dekan a pod.,
- 2 - Špecialisti: vedci a odborníci vo fyzikálnych, chemických, matematických, štatistických a v príbuzných odboroch, chemici, geológovia a geofyzici, biológovia, botanici, zoológovia, projektanti stavieb a areálov, stavební inžinieri, lekári, farmaceuti, učitelia, umelecký zamestnanci, novinári a redaktori, sudcovia a pod.,
- 3 - Technickí, Zdravotnícki, Pedagogickí Zamestnanci a Zamestnanci v príbuzných odboroch,
- 4 - Administratívni pracovníci: všeobecní administratívni pracovníci a zapisovatelia, zamestnanci na prípravu dát výpočtovej techniky, sekretárky, úradníci v knižniciach, na poštách a v príbuzných odboroch, pokladníci (v bankách, poisťovniach, na pošte a pod.), krupiér recepční a pod.,
- 5 - Pracovníci v službách a obchode: obsluhujúci pracovníci v leteckej a lodnej doprave, sprievodca, vlakvedúci, revízor, horský vodca, sprievodca v cestovnom ruchu a pri turistických zájazdoch, kuchár, čašník, barman, opatrovateľka, ošetrovateľ, sanitár, masér, kaderník, tetovač, hasič, policajt, plavčík, predavač a pod.,
- 6 - Kvalifikovaní pracovníci v poľnohospodárstve, lesníctve a rybárstve,

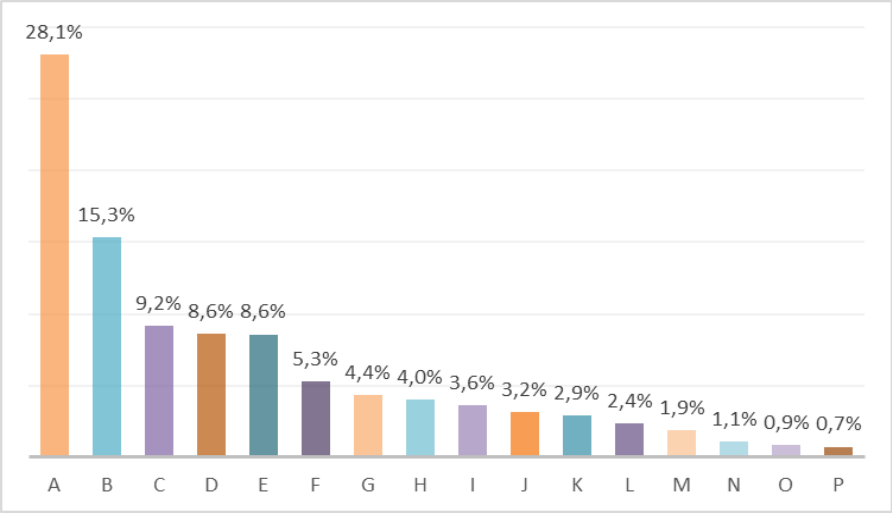
- 7 - Kvalifikovaní pracovníci a remeselníci: murár, tesár, inštalatér, maliar, kominár, zvárač, elektrikár, automechanik, opravári bicyklov a podobných vozidiel, zlatník, krajčír, hrnčiar, konzervovači ovocia, zeleniny a príbuzných produktov, spracovatelia tabaku a výrobcovia tabakových výrobkov, cukrár, pekár a pod.,
- 8 - Operátori a montéri strojov a zariadení: baník, operátori zariadení na spracovanie nerastov, studniari, vŕtači a podobní pracovníci, oceliar, obsluha pecí na výrobu skla a keramiky, na výrobu výrobkov z gumy, papiera a plastov, stolár, chemik, žeriavnik a pod.,
- 9 - Pomocní a nekvalifikovaní pracovníci: upratovači a pomocníci, pouliční predavači, chyžná, umývači motorových vozidiel, dezinfektor (likvidátor chorôb), vrátnik, chatár, školník, kostolník, kuriér, kŕmič v živočíšnej výrobe, manipulačný robotník na stavbách, dokladač tovaru, smetiar a pod.,
- Neznáme - uchádzači, ktorí neuviedli oblasť posledného zamestnania alebo predtým nikde nepracovali

Ako môžeme vidieť na obr. 17 má najväčšie zastúpenie kategória Pomocní a nekvalifikovaní pracovníci 21,0 %, naopak najmenšie zastúpenie má kategória Kvalifikovaní pracovníci v poľnohospodárstve, lesníctve a rybárstve 1,0 %. Do prvých troch kategórií s najväčším percentuálnym zastúpením (Pomocní a nekvalifikovaní pracovníci: upratovači a pomocníci, Neznáme a Pracovníci v službách a obchode) patrí viac ako 50 % UoZ. Kategórie Administratívni pracovníci a Technickí, Zdravotnícki, Pedagogickí Zamestnanci a Zamestnanci v príbuzných odboroch majú medzi sebou približne 1 % rozdiel.



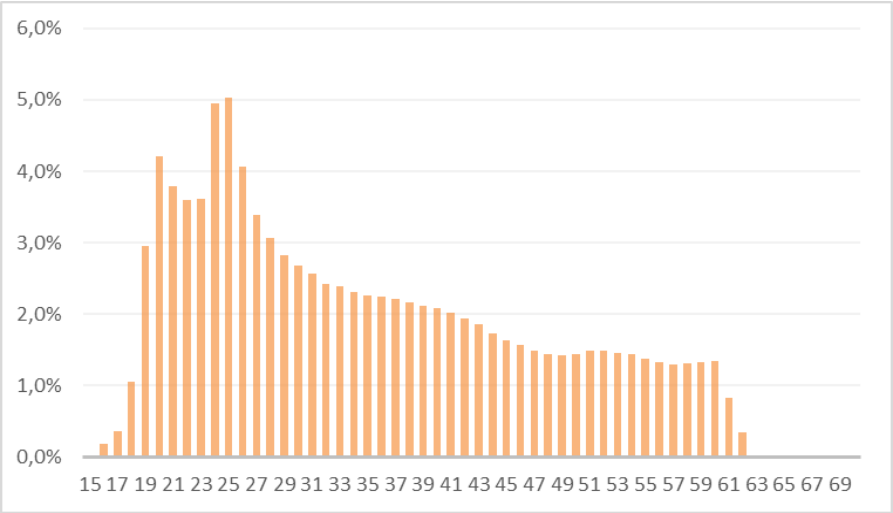
Obr. 17: Percentuálne zastúpenie UoZ na základe pozície posledného zamestnania

Ďalšou premennou (ac) je dôvod zaradenia do evidencie, ktorá obsahuje 16 kategórií nepracoval/dobrovoľne nezamestnaný (A), bez udania dôvodu (B), ostatné dôvody (C), uplynutím doby určitej (D), v skúšobnej dobe (E), ukončenie vysokoškolského štúdia (F), skončenie dočasnej PN (G), skončenie zárobkovej činnosti v zahraničí (H), pozastavenie, zánik alebo zrušenie živnosti (I), ukončenie základného vzdelávania (J), výpoveď zo strany zamestnanca (K), (ne)poberanie rodičovského príspevku do 3/6 rokov (L), nadbytočnosť zamestnanca (M), disciplinárne problémy (N), zdravotné dôvody (O), skončenie poberania príspevku (P). Do kategórie ostatné dôvody patrí osoba, ktorá dovŕšila dôchodkový vek a nevznikol jej nárok na dôchodok, nemôže vykonávať prácu bez vážneho ohrozenia zdravia, skončila dočasná PN, invalidita, osobná asistencia, výkon väzby, zrušenie miesta a pod. Do kategórie disciplinárne problémy patrí neuspokojivé plnenie úloh, odsúdenie za úmyselný trestný čin, podmienečné zastavenie trestného stíhania skončenie výkonu väzby, závažné porušenie disciplíny. Na obr. 18 môžeme vidieť, že najpočetnejšia kategória v tejto premennej je nepracoval/dobrovoľne nezamestnaný 28,1 %. Kategórie s menej ako 2 % zastúpením sú: nadbytočnosť zamestnanca, disciplinárne problémy, zdravotné dôvody a skončenie poberania príspevku. Kategórie, ktoré majú približne rovnaké percentuálne zastúpenie 9 % sú: uplynutím ostatné dôvody, doby určitej a v skúšobnej dobe.



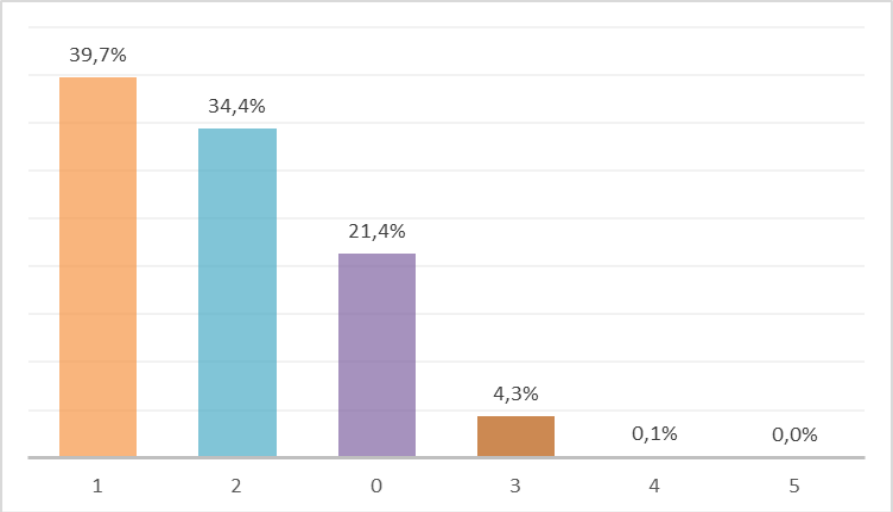
Obr. 18: Percentuálne zastúpenie UoZ na základe dôvodu zaradenia do evidencie

Premenná (ad) vek hovorí o veku uchádzača k dátumu zaradenia do evidencie. Viac ako polovica uchádzačov je vo vekovej kategórii 19 - 33 ročných. Najpočetnejšia kategória je 25 ročných, konkrétne 5 %. S pribúdajúcim vekom je uchádzačov menej. Percentuálne zastúpenie UoZ na základe veku uchádzačov je znázornené na obr. 19.



Obr. 19: Percentuálne zastúpenie UoZ na základe veku uchádzačov

Premenná (ae) počet znevýhodnení obsahuje informáciu o počte zákonom definovaných znevýhodnení uvedených v kapitole 1. Na obr. 20 vidíme, že najpočetnejšia kategória je tých, ktorí majú jedno znevýhodnenie 39,7 %. Kategória UoZ, ktorí nemajú žiadne znevýhodnenie je až na treťom mieste z hľadiska početnosti s 21,4 percentným zastúpením. Prípady, kedy má uchádzač 4 a viac znevýhodnení sú ojedinelé.



Obr. 20: Percentuálne zastúpenie UoZ na základe počtu aktívnych znevýhodnení v čase zaradenia do evidencie

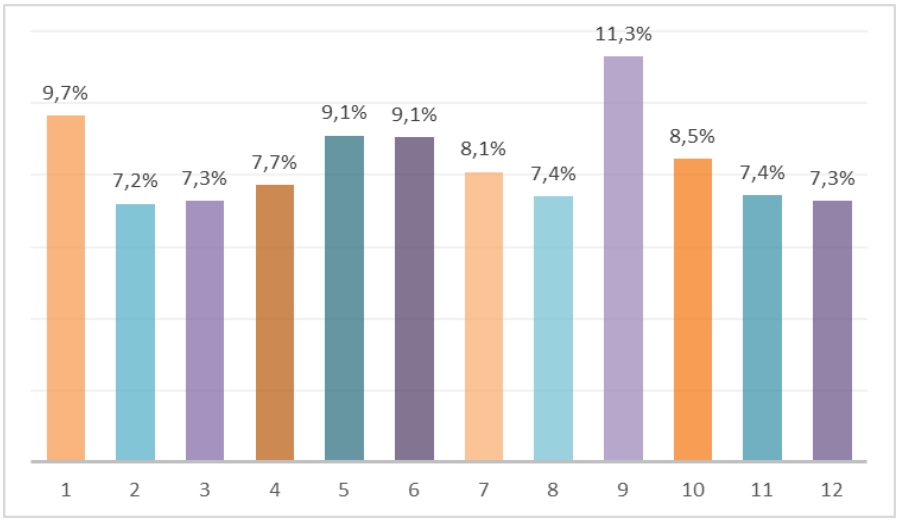
Ďalšie premenné sú binárne 0/1 (nie/áno), ktoré obsahujú (af) informáciu o tom, či uchádzač má konkrétne znevýhodnenie 1 (áno) alebo nemá 0 (nie). Znevýhodnenie c) dlhodobo nezamestnaný občan sme vyradili nakoľko pri zaradení do evidencie nevieme, či uchádzač upadne do dlhodobej nezamestnanosti alebo nie. V tab. 4 môžeme vidieť, že najväčšie zastúpenie je tých UoZ (58,6 %), ktorí za posledných 12 mesiacov nemali pravidelne platené

zamestnanie. Druhou skupinou z hľadiska početnosti sú UoZ starší ako 50 rokov (15,7 %). Vzdelanie nižšie ako stredné odborné má 15,0 % uchádzačov, čo vyplýva z percentuálneho zastúpenia v premennej Znevýhodnenie d).

Tab. 4: Percentuálne zastúpenie UoZ v premenných týkajúcich sa znevýhodnení

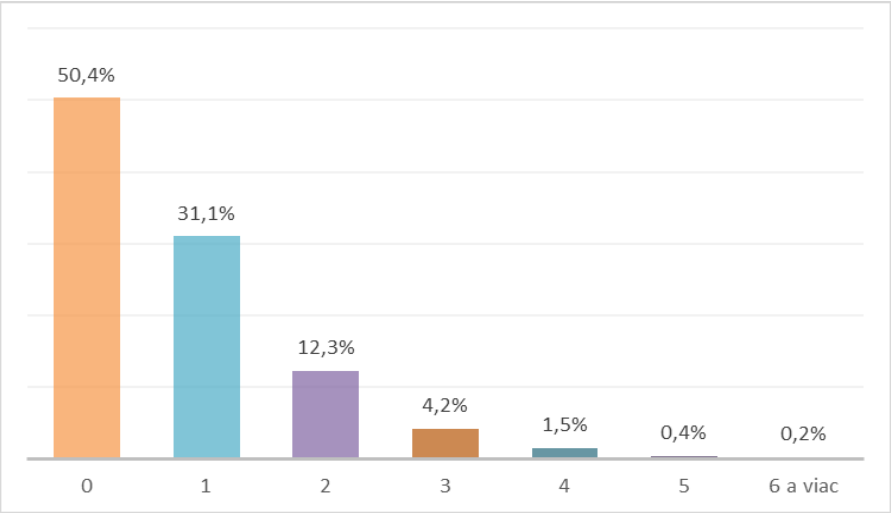
Premenná	Nie (0)	Áno (1)
Znevýhodnenie a) absolvent školy	87,3 %	12,7 %
Znevýhodnenie b) občan starší ako 50 rokov	84,3 %	15,7 %
Znevýhodnenie d) občan - vzdelanie nižšie ako stredné odb.	85,0 %	15,0 %
Znevýhodnenie e) občan - 12 kal. mes. nemal pravidelne plat. zam.	41,4 %	58,6 %
Znevýhodnenie f), g) a iné	99,8 %	0,2 %

Na základe dátumu zaradenia sme odvodili premennú (ag) mesiac zaradenia. Na obr. 22 môžeme pozorovať vysoké zastúpenie v mesiaci september, čo môže byť spôsobené absolventami stredných škôl a v mesiaci máj a jún absolventami vysokých škôl. Mesiace február, marec, apríl, august, november a december je zastúpenie UoZ pomerne rovnomerné, v rozmedzí od 7,2 % do 7,7 %.



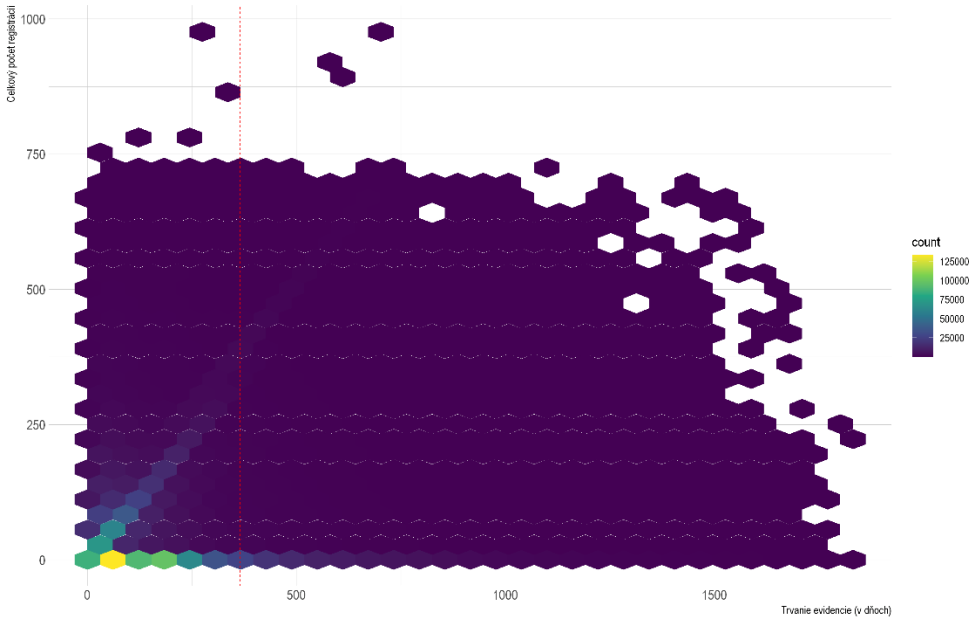
Obr. 21: Percentuálne zastúpenie UoZ na základe mesiaca zaradenia

Podobne sme odvodili premennú (ah) počet registrácií za posledné dva roky, čím sa zredukovali dáta a zaradenia v rokoch 2013 - 2014 sme používali iba ako náhľad do minulosti. Kategórie sú numerické s rozsahom 0 - 13 počtu registrácií za posledné dva roky. Na obr. 22 môžeme vidieť, že viac ako polovica UoZ je v evidencii zaregistrovaná prvý krát za posledné dva roky. Približne 2 % je takých, ktorí boli v evidencii 4 a viac krát.



Obr. 22: Percentuálne zastúpenie UoZ na základe počtu registrácií za posledné dva roky.

K tejto premennej sme pridali ďalšiu, ktorá nesie (ai) informáciu o poslednom trvaní evidencie v dňoch za posledné dva roky od dátumu zaradenia do evidencie. Pokiaľ týchto evidencií bolo za posledné dva roky viac, do tejto premennej sa priradí priemerné trvanie týchto evidencií. V prípade ak je UoZ zaradený prvýkrát je hodnota premennej posledné trvanie nulová. Na obr. 23 môžeme pozorovať istú lineárnu závislosť medzi trvaním evidencie v aktuálnom zaradení a trvaním evidencie, ktoré UoZ mal za posledné dva roky pred zaradením do evidencie.

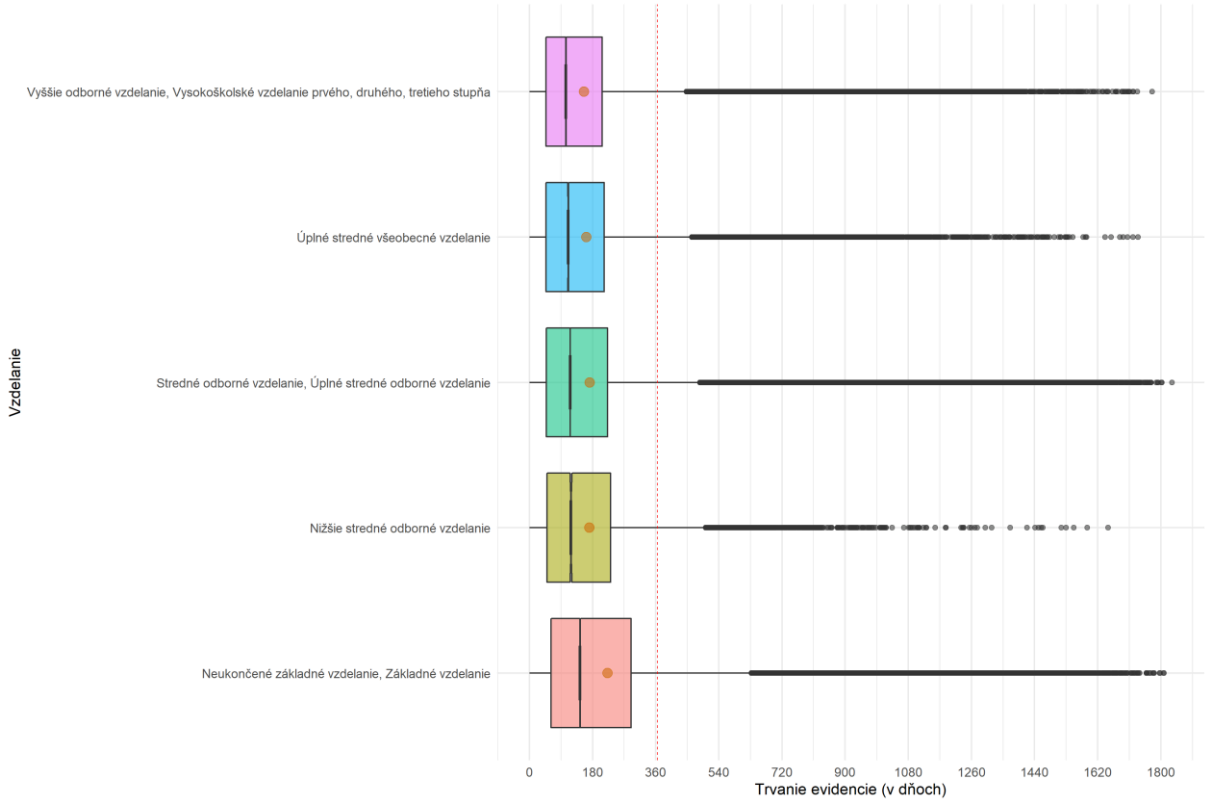


Obr. 23: Rozloženie UoZ podľa dĺžky posledného trvania a trvania evidencie aktuálne skúmaného zaradenia. Na osi x je aktuálne trvanie evidencie. Na osi y je trvanie evidencie pred posledným zaradením.

3. Exploratívna analýza vybraných premenných

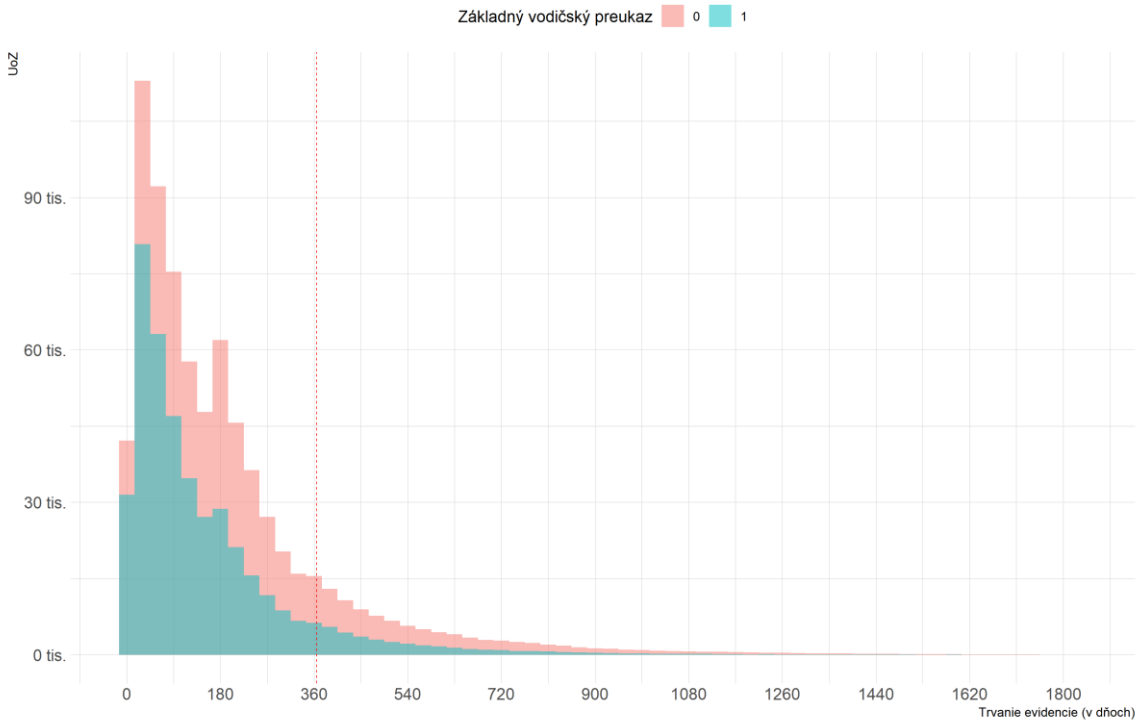
V rôznych skupinách populácie pozorujeme rozdielne zastúpenie uchádzačov, ktoré sme podrobne popísali v kapitole 2.

Uchádzači so žiadnym alebo základným vzdelaním sa ukazujú ako rizikovejší z hľadiska dĺžky trvania evidencie. Priemerné trvanie evidencie týchto uchádzačov je o približne 52 dní viac ako tých, ktorí majú nižšie stredné odborné vzdelanie a o 68 dní viac ako tých, ktorý majú vyššie odborné alebo vysokoškolské vzdelanie. S vyšším dosiahnutým vzdelaním počet dní evidencie klesá. Porovnanie môžeme vidieť na obr. 24.



Obr. 24: Boxplotová analýza premennej vzdelania v jednotlivých kategóriách. Oranžovou bodkou je znázornené priemerné trvanie v kategórii. Červená čiara označuje hranicu jedného roku od začiatku evidencie.

Podobne pre uchádzačov, ktorí nedisponujú ani základným vodičským preukazom je čas strávený na úrade priemerne 191 dní, čo je o 42 dní viac ako tí, ktorí majú aspoň základný vodičský preukaz. Pri premennej profesionálny vodičský preukaz nie je rozdiel medzi kategóriami taký výrazný (pre kategóriu 0 (nie) je priemer 177 dní a pre 1 (áno) 162). Rozloženie trvania evidencie v čase na základe kategórie je znázornené na obr. 25.

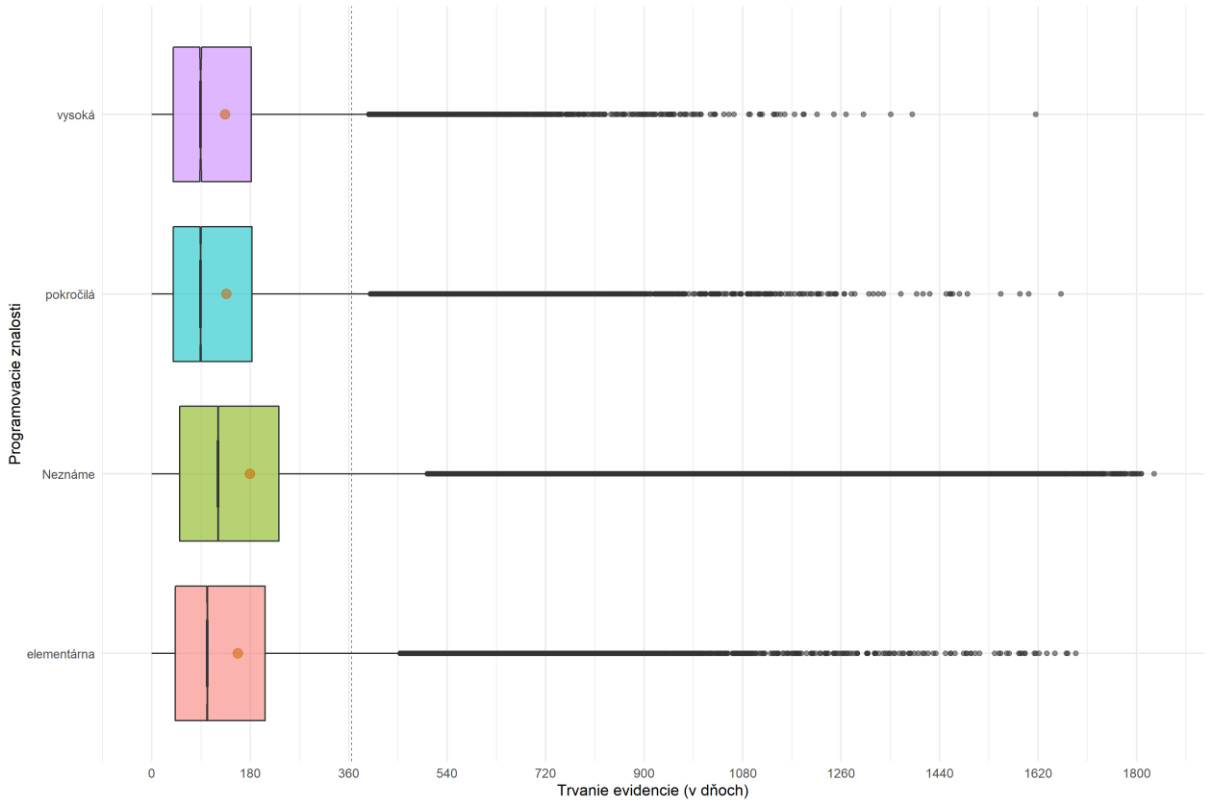


Obr. 25: Frekvenčná analýza dĺžky evidencie UoZ podľa kategórií základného vodičského preukazu. Červená čiara označuje hranicu jedného roku od začiatku evidencie.

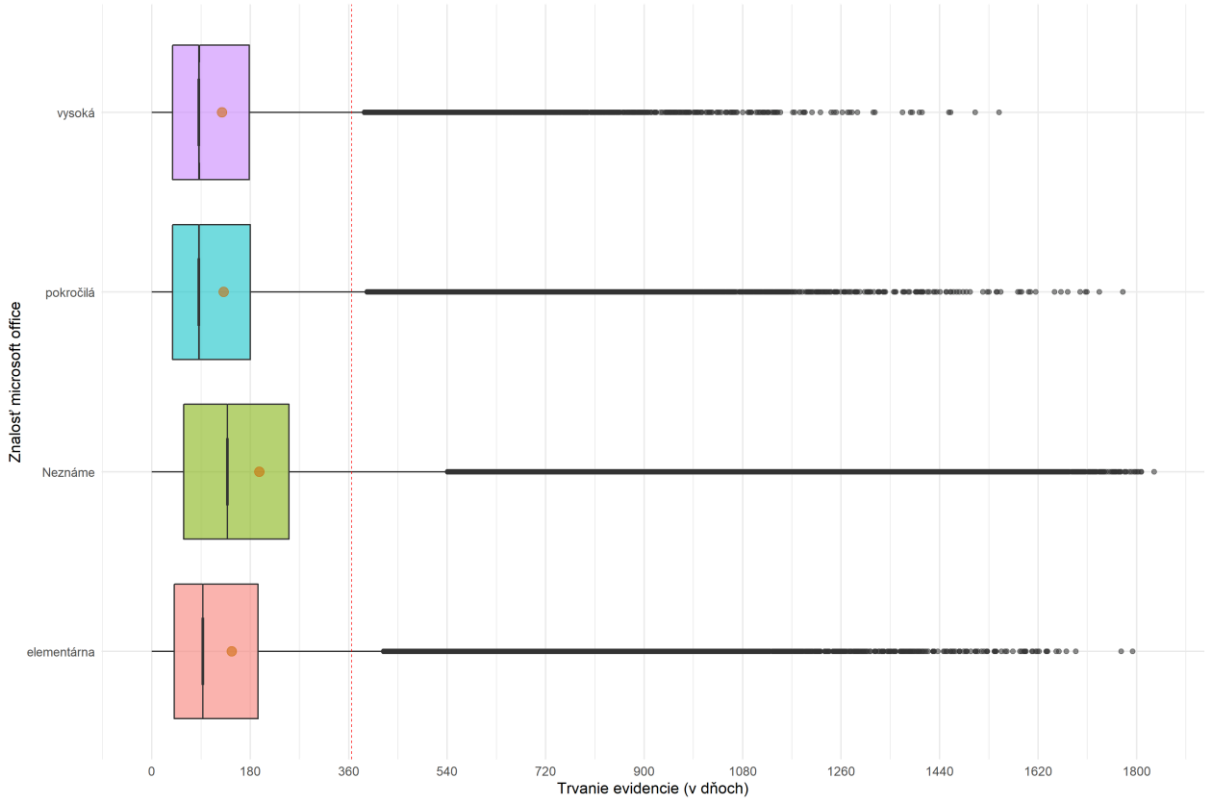
UoZ, ktorí ovládajú prácu s počítačom aspoň na elementárnej úrovni majú výrazne kratšie trvanie evidencie. Pre premennú práca s MS Office je v kategórii neznáme priemer 197 dní a pre premennú PC iné v rovnakej kategórii je to 180 dní. Avšak už pri elementárnom stupni znalosti MS Office sa priemer výrazne znížil na 146 dní a pri PC iné na 157 dní, pričom pri vyššej kategórii sa trvanie znižuje. Priemerná trvania pre ostatné kategórie sú uvedené v tab. 5.

Tab. 5: Priemerné trvanie evidencie pre premenné MS Office a PC iné v jednotlivých kategóriách.

Kategória/Premenná	MS Office	PC iné
Neznáme	197	180
elementárna	146	157
pokročilá	128	136
vysoká	132	134



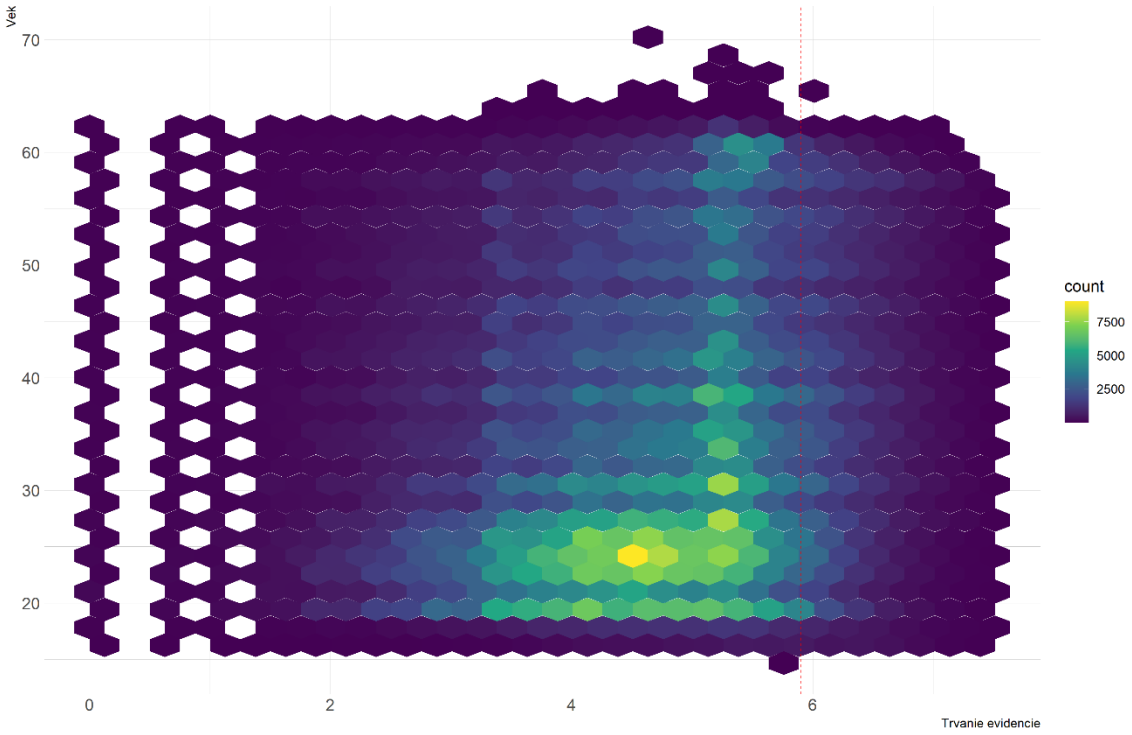
(a) Boxplotová analýza premennej práca s pc iné v jednotlivých kategóriách



(b) Boxplotová analýza premennej ovládanie MS Office v jednotlivých kategóriách

Obr. 26: Oranžovou bodkou je znázornené priemerné trvanie v kategórii. Červená čiara označuje hranicu jedného roku od začiatku evidencie.

Smerodajnou charakteristikou UoZ pri určovaní trvanie evidencie jednotlivých UoZ sa ukazuje vek. Najnižšiu priemernú hodnotu trvania evidencie má skupina UoZ vo veku od 20 do 25 rokov približne 133 dní, pričom s rastúcim vekom sa priemerné trvanie zvyšuje. Tento trend si môžeme všimnúť aj na obr. 27. Najvyšší priemer trvania evidencie je v skupine UoZ od 55 do 60 rokov približne 236 dní. Najstarší UoZ v evidencii mal 70 rokov a trvanie evidencie 105 dní, naopak najmladší mal 15 rokov a trvanie evidencie 282 dní.



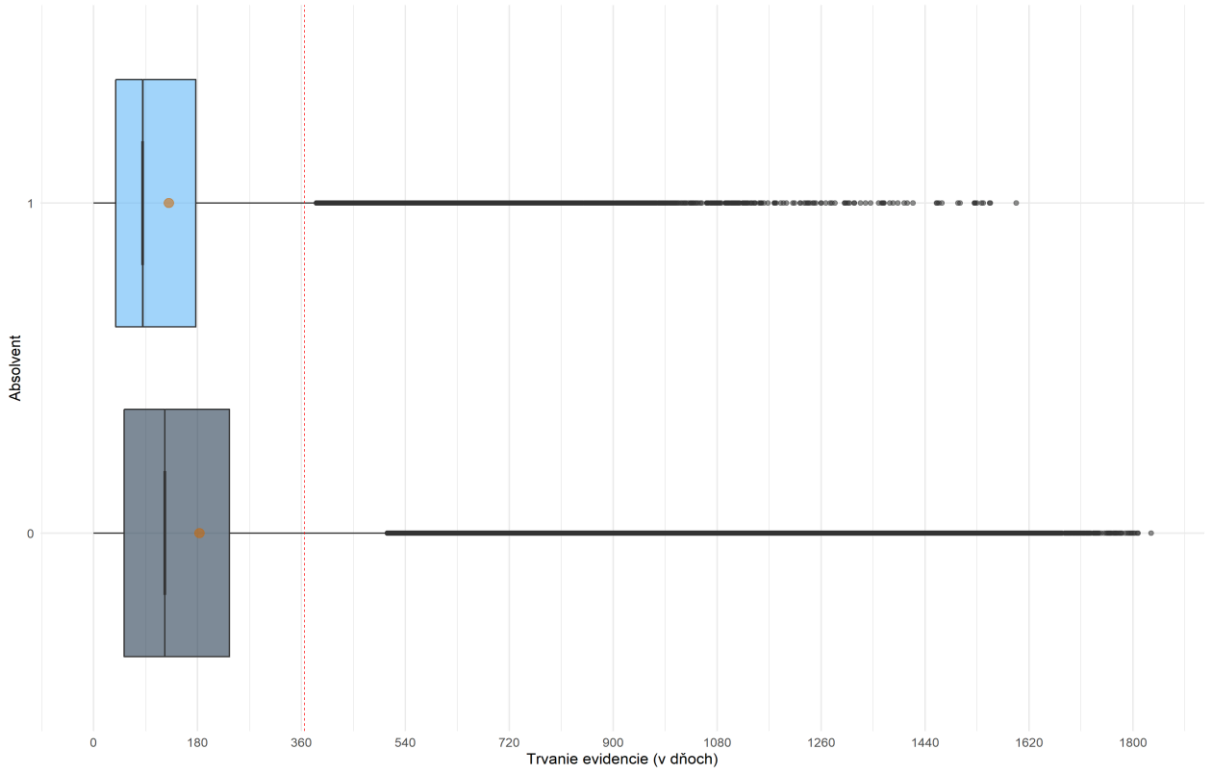
Obr. 27: Rozloženie UoZ podľa veku a trvania evidencie. Červená čiara označuje hranicu jedného roku od začiatku evidencie.

Z regionálneho hľadiska najvyšší priemerný počet dní strávený na úrade práce je v Košickom kraji 209 dní, potom nasleduje Prešovský kraj 197 dní a tretí v poradí je Banskobystrický kraj 195 dní. Najnižší priemerný počet dní v evidencii je v Trnavskom kraji 145 dní. Priemerné trvanie všetkých krajov je v tab. 6.

Tab. 6: Priemerné trvanie evidencie v premennej kraj podľa jednotlivých kategórií.

Kraj	Priemerné trvanie evidencie (v dňoch)
Banskobystrický	195
Bratislavský	156
Košický	209
Nitriansky	163
Prešovský	197
Trenčiansky	151
Trnavský	145
Žilinský	161

Medzi znevýhodneniami sa ukázali aj také, ktoré nie sú priťažujúce z hľadiska priemerného času strávenom na úrade práce. Konkrétne sa jedná o absolventov škôl. Priemerné trvanie evidencie so znevýhodnením absolvent školy je 130 dní a pre UoZ, ktorí toto znevýhodnenie nemajú je to 183 dní. Premenná znevýhodnenie e) nemá v jednotlivých kategóriách (0/1) takmer žiadny rozdiel v priemernom trvaní evidencie. Priemerné trvania všetkých znevýhodnení sú v tab. 7.



Obr. 28: Boxplotová analýza znevýhodnenia absolvent školy v jednotlivých kategóriách. Červená čiara označuje hranicu jedného roku od začiatku evidencie. Oranžovou bodkou je znázornené priemerné trvanie.

Tab. 7: Priemerné trvanie evidencie v premenných týkajúcich sa znevýhodnení UoZ (skrátené názvy premenných).

Premenná	Priemerné trvanie evidencie (v dňoch) 0/1
Znevýhodnenie a)	183/130
Znevýhodnenie b)	168/223
Znevýhodnenie d)	166/235
Znevýhodnenie e)	174/178
Znevýhodnenie f), g) a iné	176/285

Binárne premenné, ktoré sa týkajú záujmov a minulosti UoZ a ich priemerné trvania, sú uvedené v tab. 8. Môžeme si všimnúť, že UoZ s prísľubom zamestnania priemerne strávia na úrade priemerne 135 dní.

Tab. 8: Priemerné trvanie evidencie v premenných, ktoré sa týkajú záujmov a minulosti UoZ.

Premenná	Priemerné trvanie evidencie (v dňoch) 0/1
Prísľub zamestnania	176/135
Úmysel odísť do zahraničia	179/99
Záujem zamestnať sa v zahraničí	186/140
Záujem SZČO	177/166
Záujem vzdelávať sa	182/159
Záujem o absolventskú prax	178/137
Záujem o sústavnú prípravu na povolanie	177/149
Dobrovoľne nezamestnaný	169/181
SZČO	176/204
Zdravotné postihnutie	175/230

4. Modely

V tejto kapitole popíšeme použité modely a výsledky. Budeme prevažne vychádzať zo zdrojov [9, 15, 16, 17, 22, 25]. V štatistickej analýze sa bežne používajú regresné modely pri skúmaní závislostí dvoch a viacerých premenných. Cieľom analýzy je objasniť vzťah medzi meranou, vysvetľovanou závislou premennou y (v našom prípade sa jedná o premennú trvanie) a vysvetľujúcimi premennými veličinami x (v našom prípade sa jedná o maticu, ktorá obsahuje informácie o uchádzačoch). Jedná sa vlastne o použitie týchto modelov na predikciu pravdepodobného výsledku, na základe dát dostupných z minulosti.

Najskôr sa pozrieme na jednoduchú lineárnu regresiu. Tá však ako štandardná regresná metóda na množinu premenných, určených na predikciu y, môže mať tendenciu viesť k preučeniu (angl. overfitting), pokiaľ ide o počet premenných, ktoré sú v konečnom dôsledku zahrnuté do modelu. Toto spôsobuje, že model síce veľmi dobre sadne na trénovaciu sadu dát, ale pre testovacie dáta vznikne veľká odchýlka, prípadne je model veľmi zložitý, kvôli veľkému množstvu parametrov. Na riešenie týchto problémov možno použiť rôzne penalizačné alebo regularizačné regresné techniky. Preto sme ďalej porovnávali Lasso (angl. Least Absolute Shrinkage and Selection Operator) regresiu a ridge regresiu (angl. ORR, Ordinary Ridge Regression). Nakoniec vyskúšame modely so súborových metód strojového učenia náhodný les a XGBoost.

4.1. Lineárna regresia

Pri lineárnej regresii máme náhodné veličiny $y_1, \dots, y_n$ a maticu $X = (x_{ij})$ typu $n \times s$, kde $s < n$. Vysvetľovaná premenná je y, ktorej hodnoty modelujeme pomocou vysvetľujúcich premenných X, s je počet vysvetľujúcich premenných a n je počet pozorovaní. Predpokladáme, že pre náhodný vektor $y = (y_1, \dots, y_n)^T$ platí:

$$y = X\beta + \varepsilon, \tag{2}$$

kde $\beta = (\beta_1, \dots, \beta_s)^T$ je vektor neznámych parametrov a $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ je náhodný vektor, ktorý spĺňa podmienky $E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2 I_n$. Pritom σ^2 je taktiež neznámy parameter. Tento model budeme nazývať regresným. Pretože y závisí na β lineárne, hovoríme o lineárnom regresnom modeli alebo jednoducho o lineárnej regresii. Pre regresné modely je typické, že sa nezavádzajú nadbytočné vysvetľujúce premenné x_{ij}. Požaduje sa, aby mala matica X lineárne nezávislé stĺpce. Parametre $\beta_1, \dots, \beta_s$ sa odhadujú metódou najmenších štvorcov (3). [9]

4.2. Ridge a Lasso regresia

Cieľom Ridge a Lasso regresie je identifikovať premenné a zodpovedajúce regresné koeficienty, ktoré vedú k modelu, ktorý minimalizuje chybu predikcie. To sa dosiahne zavedením obmedzenia parametrov modelu, ktoré „zmenšujú“ regresné koeficienty, v prípadoch, kedy je to potrebné až k nule.

Definícia 4.1. [16, 25]

Definícia minimalizačných úloh pre odhad parametrov β lineárnej, ridge a lasso regresie:

Metóda najmenších štvorcov:
$$\min_{\beta_0,\beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{s} \beta_j x_{ij} \right)^2, \tag{3}$$

Ridge regresia:
$$\min_{\beta_0,\beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{s} \beta_j x_{ij} \right)^2 + \lambda \|\beta\|_2, \tag{4}$$

Lasso regresia:
$$\min_{\beta_0,\beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{s} \beta_j x_{ij} \right)^2 + \lambda \|\beta\|_1, \tag{5}$$

kde $y_1, \dots, y_n$ sú náhodné veličiny a $X = (x_{ij})$ typu $n \times s$ je matica vysvetľujúcich premenných, $\beta = (\beta_1, \dots, \beta_s)^T$ je vektor neznámych parametrov a $\| \|_1$ je L_1 norma (p-norma pre $p = 1$) a $\| \|_2$ je L_2 norma (p-norma pre $p = 2$). P-norma je definovaná rovnicou (6):

L_p:
$$\|\beta\|_p = \left(\sum_{i=1}^{s} |\beta_i|^p \right)^{1/p}. \tag{6}$$

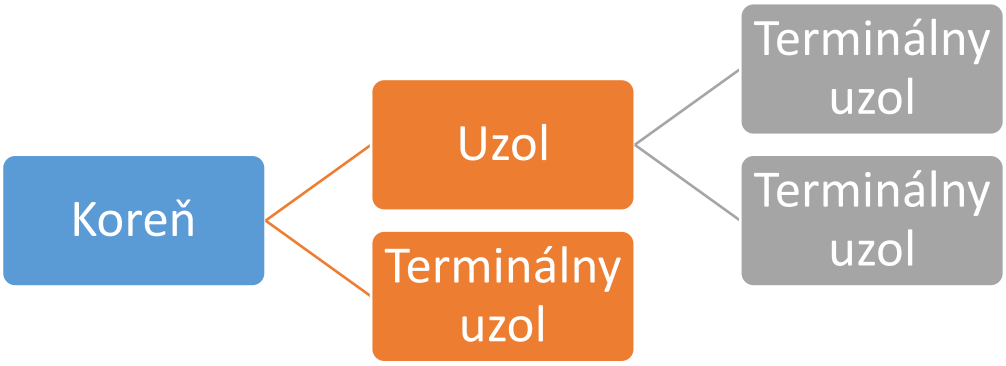
Prakticky dochádza k obmedzeniu zložitosti modelu. Voľba λ sa často robí prístupom automatizovanej k-násobnej krížovej validácie. Pre tento prístup je množina údajov náhodne rozdelená na k vzoriek. Kým k-1 čiastkových vzoriek sa používa na odhad koeficientov predikčného modelu, zostávajúca čiastková vzorka sa používa na overenie platnosti tohto modelu. Tento postup sa vykonáva k-krát, pričom každá z k čiastkových vzoriek sa použije na validáciu a ostatné na vývoj modelu. Celkový výsledok sa získa kombináciou k samostatných výsledkov validácie pre rozsah λ hodnôt a výberom preferovanej hodnoty λ, ktorá sa potom použije na určenie finálneho modelu.

4.3. Regresné stromy, náhodný les a XGBoost

Rozhodovacie stromy tvorí sada hierarchicky usporiadaných rozhodovacích pravidiel. Terminológia bola prevzatá z prírody na základe podobnosti s reálnymi stromami. Rozhodovací strom sa skladá z koreňa, ktorý predstavuje celý súbor a postupne sa vetví do ďalších uzlov. Uzol, ktorý sa ďalej nevetví sa nazýva terminálny. [22] V tejto podkapitole budeme prevažne vychádzať z [15, 17, 22, 28, 34].

4.3.1. Regresné stromy

Existujú mnohé algoritmy na vytváranie stromov. Medzi najznámejšie patrí CART (*angl. Classification and regression trees*). Na začiatku tvorby stromu patria všetky pozorovania súboru do jedného uzla (koreňa). Následne sú tieto pozorovania rozdelené do ďalších dcérskych uzlov na základe rozdeľovacieho kritéria (*angl. spliting criterion*). Tieto uzly sa ďalej delia až kým neprídu do terminálneho uzlu. Na obr. 31 je znázornená schéma jednoduchého rozhodovacieho stromu. Stromy, ktoré sa z jedného uzla rozdeľujú práve do dvoch ďalších vetiev sa nazývajú binárne stromy.



Obr. 29: Schéma jednoduchého rozhodovacieho stromu.

Existuje niekoľko rozdeľovacích kritérií, ktoré sa líšia podľa toho, či sa jedná o klasifikačný alebo regresný strom. Rozdeľovacie kritérium meria homogenitu v uzloch (*angl. node impurity*). Predpokladajme, že máme regresný strom rozložený do určitého počtu terminálnych uzlov a predikovanú hodnotu závislej spojitej premennej y chceme vyjadriť ako konštantu pre každý z nich. Ak použijeme kritérium, ktoré minimalizuje strednú kvadratickú odchýlku, najlepším odhadom tejto konštanty bude priemer. Snažíme sa nájsť rozdelenie závislej premennej y, ktoré bude minimalizovať rovnicu (8).

Definícia 4.2. [21, 32]

Kritérium minima kvadratickej odchýlky (*angl. least square deviation*) $Q(T)$:

$$\bar{y}_t = \frac{1}{N_t}\sum y_{i(t)}\,, \tag{7}$$

$$Q_t(T) = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \bar{y}_t)^2\,, \tag{8}$$

kde N_t je počet pozorovaní v uzle *t*, $y_{i(t)}$ sú hodnoty závislej premennej v uzle *t* a *T* označuje náhodný strom.

Pri regresných stromoch je finálnemu uzlu priradený priemer hodnôt závislej premennej v danom uzle. Takto sa prichádza o pôvodný rozsah hodnôt závislej premennej. Počet výsledných predikovaných hodnôt bude rovný počtu terminálnych uzlov. Výsledná nespojitá plocha je veľkou nevýhodou regresných stromov. Túto nespojitosť možno odstrániť napríklad preložením dát lineárnym regresným modelom v každom terminálnom uzle alebo použitím kombinácie náhodných stromov (náhodný les) a pod. Regresné stromy nemajú požiadavky na rozdelenie dát, na rozdiel od modelov z častí 4.1 a 4.2.

4.3.2. Náhodný les

Náhodný les sa skladá zo súboru stromov $T_1, \ldots, T_N$, ktorých regresnú funkciu môžeme označiť ako $h(X, \Theta_1), \ldots, h(X, \Theta_N)$, kde h je funkcia, X je prediktor a $\Theta_1, \ldots, \Theta_N$ sú nezávislé rovnako rozdelené náhodné vektory. Pre náhodný les sa používajú binárne stromy typu CART. Celková vzorka sa rozdelí na testovací a testovací súbor, pričom trénovacie súbory pre jednotlivé stromy T_i sú tzv. bootstrapové výbery z dátového súboru *L*. Bootstrapové výbery sú náhodné výbery s opakovaním veľkosti *p*. Tieto výbery nie sú nezávislé, nakoľko niektoré pozorovania sú zahrnuté do testovacieho súboru viackrát a niektoré naopak vôbec. Pozorovania, ktoré sú v *i*-tom bootstrapovom výbere L_i sa používajú pri tvorbe stromu T_i a tie, ktoré sa tam nedostali sa používajú na odhad chyby nazývanej *obb* (angl. *out-of-bag*). Pre regresiu je predikcia každého pozorovania priemerom zo všetkých stromov:

$$\hat{f}_{rf}(x) = \frac{1}{N}\left(\sum_{i=1}^{N} T_i(x)\right), \tag{9}$$

kde $T_i(x)$ je výsledná predikcia pre vložené premenné *x* *i*-teho stromu. Stromy, ktoré sa tvoria pri výbere s opakovaním sú závislé a teda výsledné stromy budú korelované. Zníženie tejto

korelácie sa dosiahne náhodným výberom iba z určitého počtu prediktorov. Pre každý strom sa tak najlepšie vetvenie pre daný uzol hľadá iba z *m* prediktorov $X_1, \ldots, X_m$.

4.3.3. XGBoost

Pri matematických modeloch je dôležité nájsť také, ktoré si dokážu efektívne poradiť s komplexnými údajmi a veľkým objemom týchto dát. Z tohto dôvodu sú súborové metódy (angl. *ensemble learning method*) veľmi efektívnym nástrojom na zlepšenie výkonu viacerých existujúcich modelov. Tieto metódy sa spoliehajú hlavne na randomizačné techniky, ktoré spočívajú v generovaní mnohých rozmanitých riešení daného problému alebo na postupoch adaptívneho dôrazu (napr. zosilnenie, angl. *boosting*). Jednou z nich je aj metóda s názvom eXtreme Gradient Boosting alebo XGBoost, ktorá je založená na zvýšení gradientu (angl. *gradient boosting*). Je veľmi obľúbenou medzi najlepšími uchádzačmi v súťažiach Kaggle [13].

Algoritmus XGBoost, ktorý navrhli Chen a Guestrin v roku 2016 [17], patrí medzi veľmi rýchle a dobre integrovateľné algoritmy rozhodovacieho stromu. Algoritmus využíva CART ako základný klasifikátor a rozhoduje spoločne s viacerými s rozhodovacími stromami, pričom vstupná vzorka nasledujúceho rozhodovacieho stromu súvisí s výsledkami a predikciou predchádzajúceho rozhodovacieho stromu. XGBoost je vysoko flexibilný a všestranný nástroj, ktorý dokáže vyriešiť väčšinu regresií a klasifikácií, ako aj účelové funkcie vytvorené používateľom. Za predpokladu, že samotný model XGBoost je zložený z N stromov, je možné ho vyjadriť nasledovne [17]:

$$\Phi(x_i) = \widehat{y_i} = \sum_{k=1}^{N} f_k(x_i), \tag{10}$$

kde $\widehat{y_i}$ predstavuje predikovanú hodnotu, f_k predstavuje k-tý strom, $f_k(x_i)$ je umiestnenie i-tej vzorky v k-tom strome, x_i je i-tý vstupný údaj. XGBoost je podobný väčšine modelov strojového učenia a jeho optimalizačnou funkciou je súčet stratovej funkcie a regulačnej funkcie, ktorá riadi presnosť a zložitosť modelu. Konkrétny predpis účelovej funkcie je nasledovný:

$$L(\Phi) = l(\Phi) + \Omega(\Phi) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{11}$$

kde

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 , \tag{12}$$

pričom y_i je pozorovaná hodnota, l je stratová funkcia, ktorá sa zvyčajne používa na meranie rozdielu medzi y_i a $\hat{y}_i$, druhý člen Ω je regularizačný člen, γ predstavuje zložitosť každého terminálneho uzlu, T predstavuje celkový počet terminálnych uzlov v rozhodovacom strome, λ je parameter uplatňujúci sa na škálovanie penalizácie terminálneho uzla a ω_j predstavuje umiestnenie j-tého terminálneho uzla.

Výpočet rovnice (11) je iteračný a to tak, že $\hat{y}_i^{(t)}$ je predikcia pre *i*-te pozorovanie na *t*-tej iterácii. Cieľ je minimalizovať nasledujúci výraz:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) . \tag{13}$$

Optimalizačný proces rovnice (13) je aproximovaný Taylorovým rozvojom (14). Príklad optimalizácie v kroku t je v rovnici (15).

$$f(x + \Delta x) = f(x) + f(x)'\Delta x + f(x)''\Delta x, \tag{14}$$

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right)\right] + \Omega(f_t) , \tag{15}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) ,$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) ,$$

kde g_i a h_i sú príslušné derivácie stratovej funkcie. Po úpravách sa dosiahne nasledovné zjednodušenie:

$$L^{(t)} = \sum_{i=1}^{n} \left[g_i, f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) . \tag{16}$$

Parametre rozhodovacieho stromu sú substituované do účelovej funkcie a množina vzorky definujúca j-tý terminálny uzol je nasledovná: $I_j = \{i \mid q(x_i) = j\}$, kde q označuje štruktúru každého stromu, ktorý sleduje index terminálneho uzlu pre prislúchajúcu vstupnú hodnotu. Prepis rovnice (16) po doplnení funkcie Ω vyzerá nasledovne:

$$L^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T\,. \tag{17}$$

Pre pevnú štruktúru $q\,(x)$ je možné optimálnu váhu ω^*_j terminálneho uzla j vypočítať podľa nasledujúceho vzorca:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}.$$

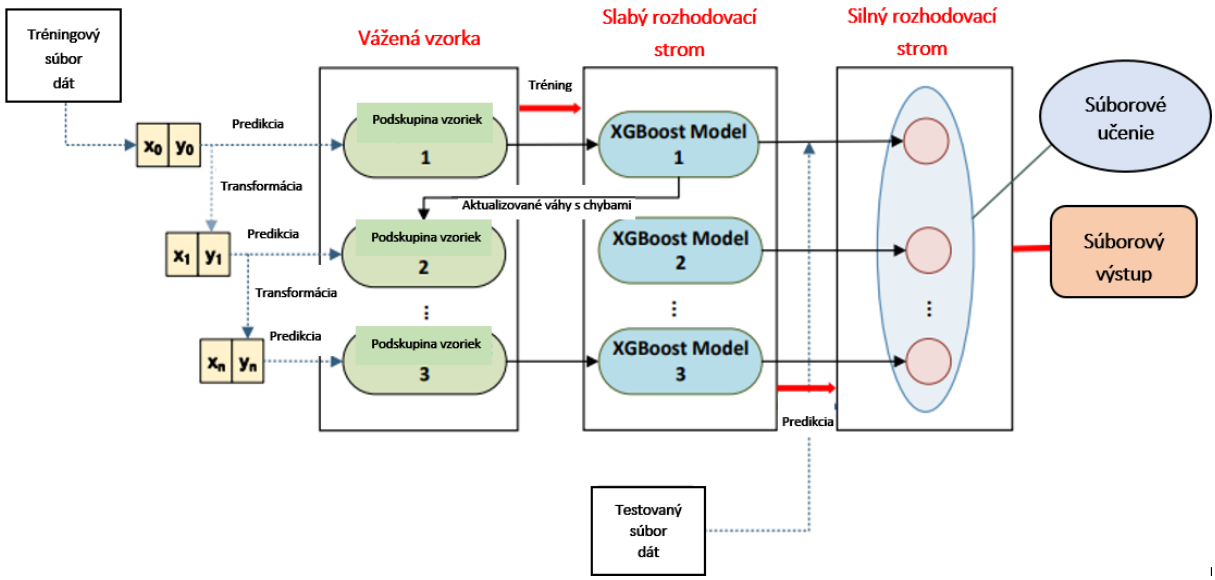
A potom vypočítať korešpondujúcu optimálnu hodnotu podľa:

$$L^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T.$$

Vo všeobecnosti nie je možné vypočítať všetky možné stromové štruktúry q. Miesto toho algoritmus začína jedným uzlom a potom sa k stromu pridávajú vetvy. Ak I_L a I_R sú množiny ľavého a pravého uzla po rozdelení a množina $I = I_L \cup I_R$, potom je rovnica zníženia straty po rozdelení nasledovná:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma.$$

Na nasledujúcom obrázku je znázornená štruktúra algoritmu XGBoost.



Obr. 30: Štruktúra algoritmu XGBoost (prevzaté a upravené) [24].

5. Implementácia a výsledky modelov

V tejto kapitole popíšeme praktickú implementáciu modelov v softvéri R. Vo všetkých modeloch sme si dáta rozdelili na trénovaciu a testovaciu sadu dát. Trénovacia sada obsahovala 80 % dát a testovacia zvyšných 20 %. Na porovnávanie modelov sme používali metriky MAE (angl. Mean Absolute Error), MSE (angl. Mean Squared Error), RMSE (angl. Root Mean Squared Error), definované nasledujúcimi rovnicami:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|\,, \tag{18}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2\,, \tag{19}$$

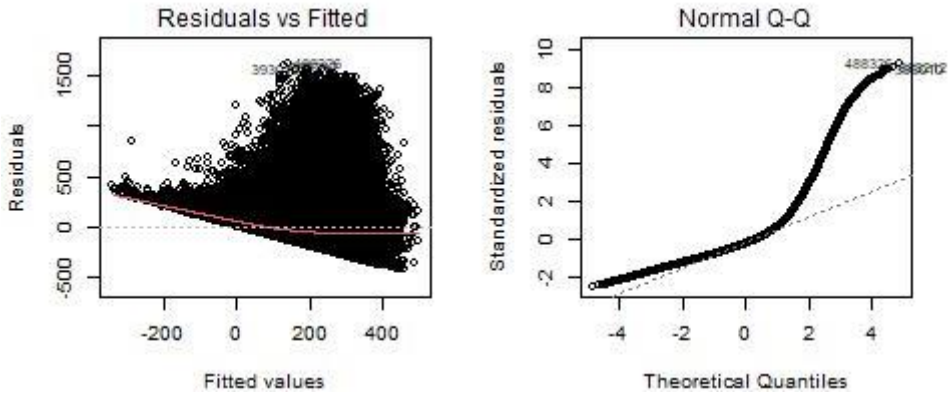
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}\,, \tag{20}$$

kde y sú konkrétne pozorovania (v našom prípade dĺžka evidencie, teda premenná trvanie) a $\tilde{y}$je modelom predikovaná hodnota skutočného y.

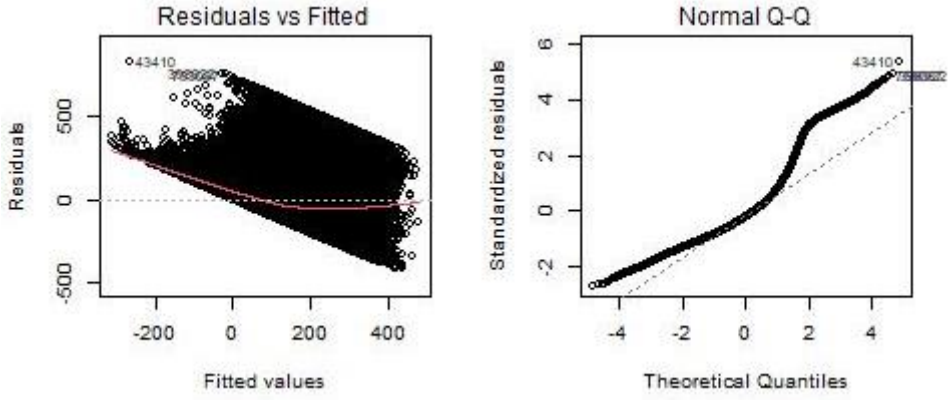
5.1. Lineárna, Ridge a Lasso regresia

Tieto typy modelov sme aplikovali na troch rôznych podobách skúmanej premennej trvanie. Pôvodné trvanie, trvanie s logaritmickou transformáciou a trvanie s maximálnou dĺžkou evidencie dva roky (teda každej evidencii, ktorá trvala viac ako dva roky sa priradilo trvanie presne dva roky). Rôzne formy premennej trvania sme skúšali pretože pôvodné dáta nadobúdajú priveľa extrémnych hodnôt, čo pre použitie lineárnej regresie nie je vhodné (musí byť zachovaná normalita rezíduí).

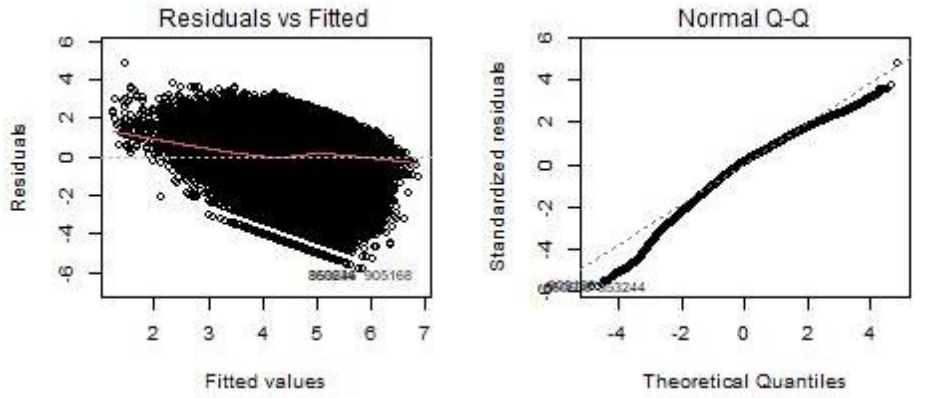
Výsledky vizuálneho overenia homoskedasticity a normality rezíduí sme porovnali pre všetky typy premennej trvania, čo môžeme vidieť na obr. 31. Na obrázku vľavo je červenou čiarou naznačený lineárny vzťah v regresii. Odchýlky od horizontálnej prerušovanej čiernej čiary naznačujú nelinearitu. Logaritmická transformácia premennej trvanie pomohla aspoň pri vizuálnych testoch dáta viac priblížiť k normálnemu rozdeleniu. Na exaktné overenie sme použili Breusch-Pagan test (funkciu *bptest()* z knižnice *lmtest*). Na overenie normality rezíduí sme porovnávali viac testov pomocou funkcie *ols_test_normality()*, z knižnice *olsrr*. Pre žiadnu z uvedených troch podôb premennej trvanie nám testy nepotvrdili normalitu ani homoskedasticitu.



(a) Pôvodná premenná trvanie

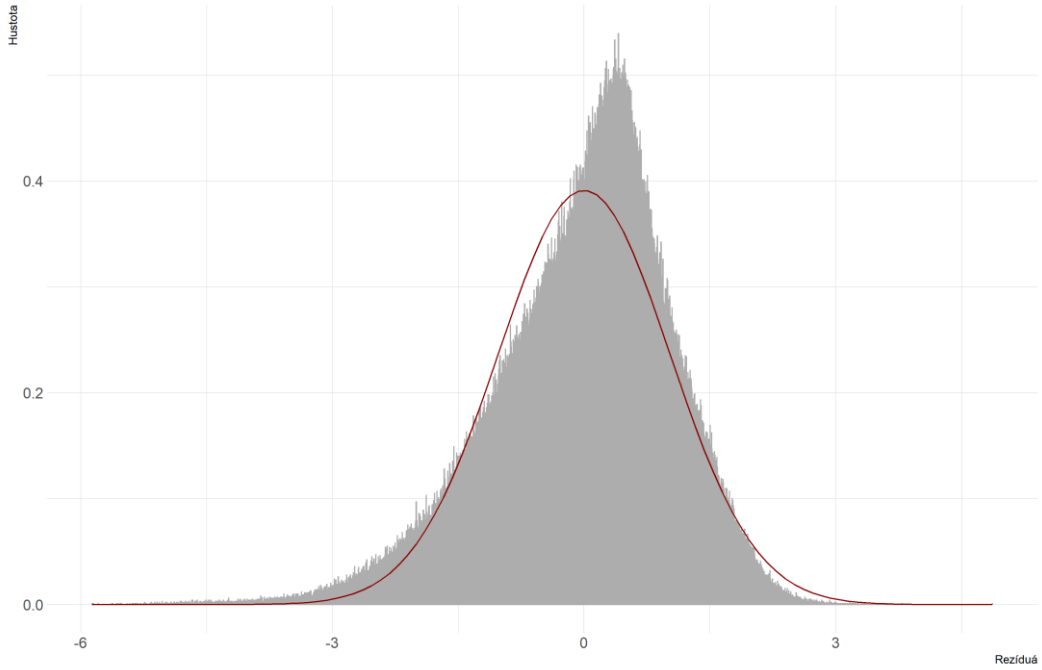


(b) Premenná trvanie pre maximálnu dĺžku evidencie dva roky



(c) Logaritmická transformácia premennej trvanie

Obr. 31: Výsledky pre overenie lineárneho vzťahu v lineárnej regresii, Q-Q.



Obr. 32: Frekvenčná analýza rezíduí pre model, kde bola použitá logaritmická transformácia premennej trvanie. Červená čiara označuje distribúciu normálneho rozdelenia.

Na odhad parametrov lineárnej regresie sme použili funkciu $lm()$ z knižnice *stats*. Výsledky pre tento typ regresie a rôzne typy trvaní sú uvedené v tab. 9.

Tab. 9: Porovnanie MAE, MSE, RMSE pre rôzne trvania evidencie lineárnej regresie (MAE, MSE a RMSE je vypočítané na testovacej vzorke dát)

	MAE	MSE	RMSE
Pôvodné trvanie	119,72	31 410,54	177,23
Log trvanie	0,80181	1,03904	1,01933
Trvanie max 2 roky	112,29	23 639,09	153,75

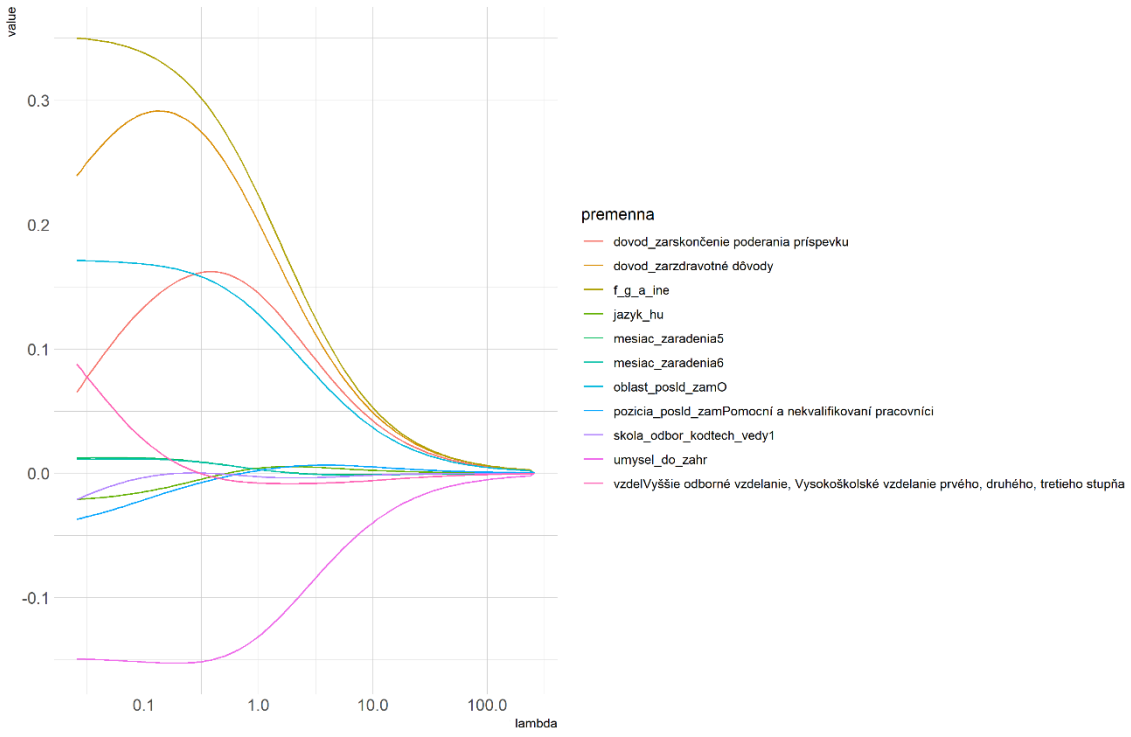
Pre určenie optimálnej hodnoty parametra λ pre regularizačné ridge a lasso regresie sme použili predvolenú 10-súborovú krížovú validáciu, pomocou funkcie $cv.glmnet()$ z knižnice $glmnet$. Na porovnávanie sa používala MSE tr (teda MSE, ktorá vznikla z rozdelenia trénovacej sady dát pri krížovej validácii). Optimálne parametre λ, sme zvolili ako výstupy spomínanej funkcie $cv.glmnet()$, konkrétne $lambda.1se$. Hodnota $lambda.min$, minimalizuje MSE tr a $lambda.1se > lambda.min$, ktorá vedie k silnejšej regularizácii a zároveň nezvýši MSE tr o viac ako jednu smerodajnú odchýlku oproti minimu. Parameter $lambda.min$ sme nepoužili kvôli hodnotám blízkym 0 (teda by sa jednalo o model blízky lineárnej regresii bez penalizácie premenných, naviac by sme dostali rovnako zložitý model,

ktorý by sa mohol príliš naviazať na tréningovú sadu dát a spôsobiť veľkú odchýlku pri testovaní na nových dátach). Hodnoty pre nás zvoleného optimálneho parametra λ sú uvedené v tab. 10. Tento parameter sme následne použili pri definovaní modelu a predikciách pri funkcii $predict()$, z knižnice *car*.

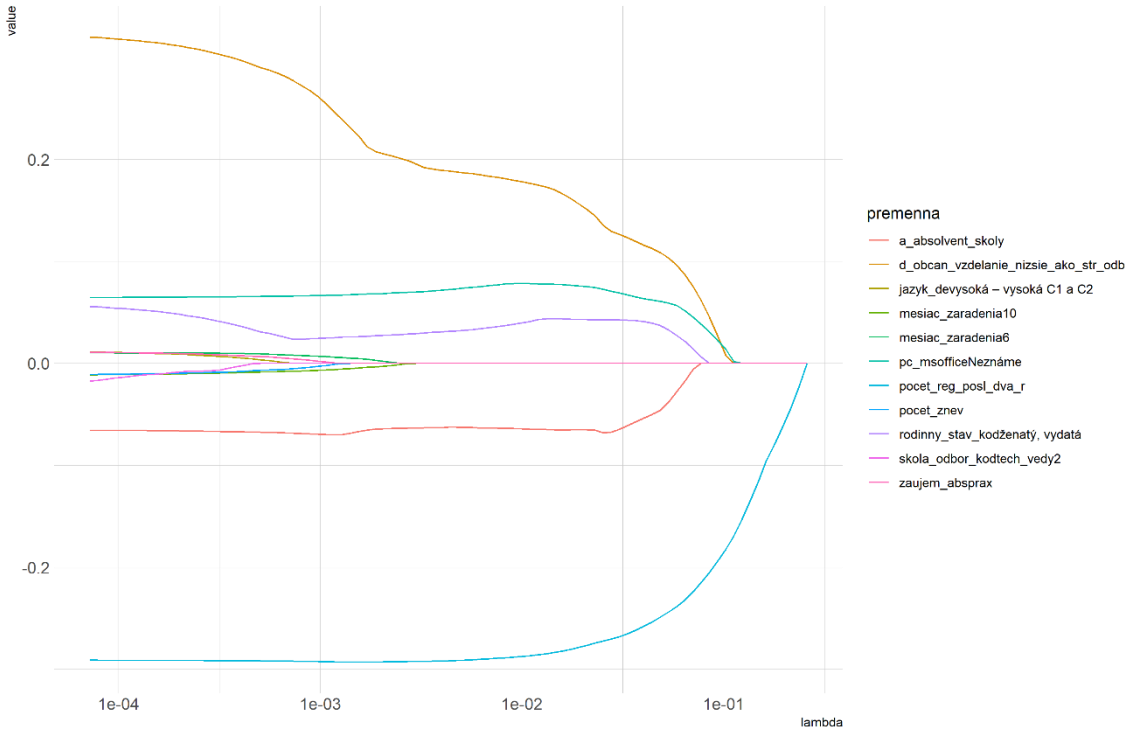
Tab. 10: Optimálne hodnoty parametra λ, pre ridge a lasso regresiu, pre rôzne hodnoty premennej trvania evidencie

	$\lambda - ridge$	$\lambda - lasso$
Pôvodné trvanie	21,58	0,29
Log trvanie	0,07	0,00
Trvanie max 2 roky	12,54	0,20

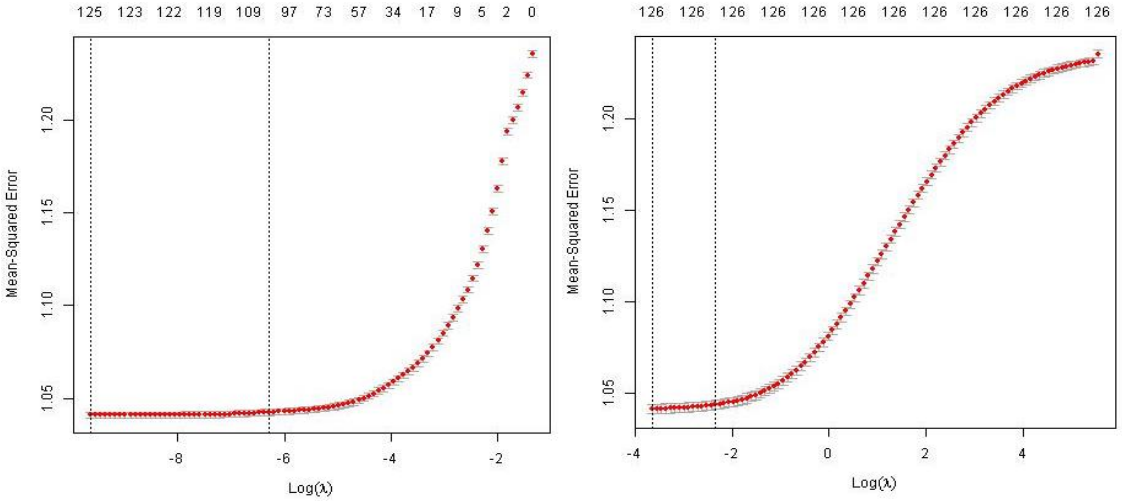
Môžeme si všimnúť, že pri lasso regresii je hodnota regularizačného parametra výrazne menšia oproti ridge regresii. Na obr. 33 - 34 je znázornených 10 premenných, ktoré ani pri vysokých hodnotách parametra λ neostali nulové a v prípade ridge regresie blízke nule.



Obr. 33: Ridge regresia: Premenné modelu pre zlogaritmované trvanie a postupný vývoj hodnoty β pri zvyšujúcej sa penalizácii λ pre vybrané parametre.



Obr. 34: Lasso regresia: Premenné modelu pre zlogaritmované trvanie a postupný vývoj hodnoty β pri zvyšujúcej sa penalizácii λ pre vybrané parametre.



(a) Ridge regresia (b) Lasso regresia

Obr. 35: Vývoj MSE tr pre zväčšujúcu hodnotu parametra $log(\lambda)$. Vo vrchnej časti grafu je počet parametrov modelu. Použité bolo zlogaritmovaná premenná trvania. Prvou prerušovanou čiarou je vyznačená $\lambda.min$ a druhou prerušovanou čiarou je vyznačená $\lambda.1se$.

Výsledné hodnoty pre zvolené metriky pre modely ridge a lasso regresie sú uvedené v tab. 11. Podľa tab. 9 a tab. 11 vidíme, že najnižšia chyba bola pre takmer všetky metriky a všetky podoby premennej trvanie pre lineárny model.

Tab. 11: Porovnanie MAE, MSE, RMSE pre rôzne trvania evidencie ridge a lasso regresie

	Pôvodné trvanie	Log trvanie	Trvanie max 2 roky
MAE (ridge)	119,65	0,80365	112,35
MSE (ridge)	31 542,64	1,04109	23 707,65
RMSE (ridge)	177,60	1,02034	153,97
MAE (lasso)	119,72	0,80257	112,35
MSE (lasso)	31 467,53	1,04061	23 669,68
RMSE (lasso)	177,39	1,02010	153,84

Pre porovnanie rôznych podôb premennej trvanie sme sa rozhodli použiť R^2, definovanú nasledovne:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{21}$$

kde y sú konkrétne pozorovania (v našom prípade dĺžka evidencie, teda premenná trvanie) a $\tilde{y}$je modelom predikovaná hodnota skutočného y a $\bar{y}$ je priemer z y. Najvyššia hodnota vyšla pre logaritmickú transformáciu premennej trvanie zaokrúhlene 0,16 pre každý model, pre pôvodné trvanie 0,14 a pre trvanie ohraničené na maximálne dva roky 0,15. Toto porovnanie spolu s výhodami dát blízkych normálnemu rozdeleniu nás viedlo k rozhodnutiu v ďalších častiach práce používať logaritmickú transformáciu premennej trvanie.

5.2. Náhodný les

Pre model náhodného lesa sme použili logaritmickú transformáciu premennej trvanie a všetky premenné podrobne opísané v kapitole 2.

Pre správne fungovanie algoritmu a výslednej predikcie s minimálnou chybou je potrebné zvoliť vhodné parametre modelu. Medzi tie patrí určenie počtu stromov náhodného lesa (*ntrees*), počet prediktorov (*mtries*), maximálna hĺbka stromu tzn. najdlhšia cesta medzi koreňovým a terminálnym lisovým uzlom (*max_depth*), minimálny počet pozorovaní v uzle, aby sa mohol rozdeliť (*min_rows*), parameter na rozdelenie počiatočného datasetu (*nbins*) a posledný parameter, ktorý sme zvolili na zlepšenie výslednej predikcie, je počet pozorovaní v trénovacom dátovom súbore (*sample_rate*).

Algoritmus tvorby lesa môžeme popísať nasledovne [22]:

1. Vytvorí sa bootstrapový podsúbor L_i o veľkosti N – trénovacích súborov.
2. Náhodne sa vyberie m prediktorov.
3. Vytvorí sa strom T_i na bootstrapovom súbore L_i s použitím náhodne vybraných m prediktorov. Rast stromu sa zastaví, keď dosiahne maximálnu hĺbku stromu (prípadne ak je v uzle zadaný minimálny počet pozorovaní, aby sa uzol mohol rozdeliť).
4. Pomocou vytvoreného stromu sa vypočítajú predikcie pre pozorovania z testovacieho. Krok 1 – 4 sa opakuje do konečného počtu stromov v náhodnom lese.
5. Spočíta sa celkový výsledok predikcie celého lesa priemerovaním.

Na implementáciu v softvéri R sme sa rozhodli využiť knižnicu *h2o*. Základné funkcie z knižnice *randomForest* a *ranger* (čo je rýchlejšia implementácia náhodného lesa) sme nevyužili pre neefektivitu výpočtov.

Najlepšie modely sme hľadali spomedzi viacerých s rôznymi parametrami. Najskôr sme vytvorili mriežku, ktorá sa použila ako hyper parameter vo funkcii *h2o.grid* z knižnice *h2o*. Rozsah parametrov mriežky je v tab. 12. Celkovo máme 2025 kombinácií a teda skupinu 2025 potenciálne dobrých modelov. Neprehľadávali sme všetky modely, ale použili sme na hľadanie *Random Discrete*, ktoré je popísané v [32, 33]. Dôležitý parameter je obmedzenie celkového času výpočtu, ktorý sme zvolili podľa potreby (maximálne však na 10 hodín). Parametre sme pre účely hľadania po mriežke zvolili menšie, nakoľko pre viac stromov a väčšiu hĺbku boli výpočty technicky náročné ak sme chceli súčasne porovnávať viac modelov takýmto spôsobom. Najviac technicky náročné výpočty boli na pamäť počítača. Práve nedostatočná pamäť spôsobila, že maximálny počet modelov na jedno spustenie bol 12.

Tab. 12: Rozsah parametrov použitých pri hľadaní najlepšieho modelu náhodného lesa.

parameter	minimálna hodnota	maximálna hodnota	krok/všetky hodnoty
ntrees	100	200	50
mtries	15	25	5
max_depth	5	25	5
min_rows	1	5	2
nbins	10	30	5
sample_rate	0,632	0,8	[0,632; 0,7; 0,8]

V tab. 13 môžeme vidieť poradie prvých troch najlepších modelov pre prvú skupinu 12 modelov. Pri zoradení modelov, podľa MSE tr, od najmenšej po najväčšiu sme mohli pozorovať aké hodnoty parametrov sú pri vývoji modelu dôležité. Pre parameter hĺbka stromu (*max_depth*) sú v najlepších modeloch nižšie hodnoty (10 a 15). Pre parameter počet stromov (*ntrees*) sa ukazuje, že čím viac ich je tým je výsledný model presnejší.

Tab. 13: Prvé tri najlepšie modely zoradené podľa MSE tréningovej vzorky dát (MSE tr).

max depth	min rows	mtries	nbins	ntrees	sample rate	MSE tr
10	5	15	25	200	0,700	0,989
10	3	15	20	150	0,632	0,998
10	5	20	30	100	0,800	1,036

Na základe toho sme vybrali parametre pre ďalšie modely takým spôsobom, že sme postupne nastavovali vybrané parametre v mriežke vždy po dvojiciach. Najskôr sme hodnoty ntrees nastavili na 100, 150 a 200 a mtries na hodnoty 15, 20 a 25, čo výsledne dáva deväť modelov s rôznymi parametrami. Následne sme použili parametre ntrees a mtries z najlepšieho modelu a v ďalšom kroku sme nastavovali parametre max_depth v rozsahu 10, 15 a 20 a min_rows v rozsahu 1, 3 a 5. Nakoniec sme dostali optimálne parametre ntrees, mtries, max_depth, min_rows, ktoré sme použili v poslednom kroku pri nastavovaní zvyšných dvoch parametrov nbins v rozsahu 15, 25 a 30 a sample_rate 0,632, 0,7 a 0,8. Následne sme dodatočne otestovali ďalšie tri modely, ktorým sme nastavili väčšie rozsahy parametrov. Výsledky a parametre sú uvedené v tab. 14. Týmto parametrom sme následne vybrali väčší rozsah zvolených hodnôt.

Tab. 14: Umiestnenie troch vybraných modelov zoradené podľa MSE tréningovej vzorky dát (MSE tr)

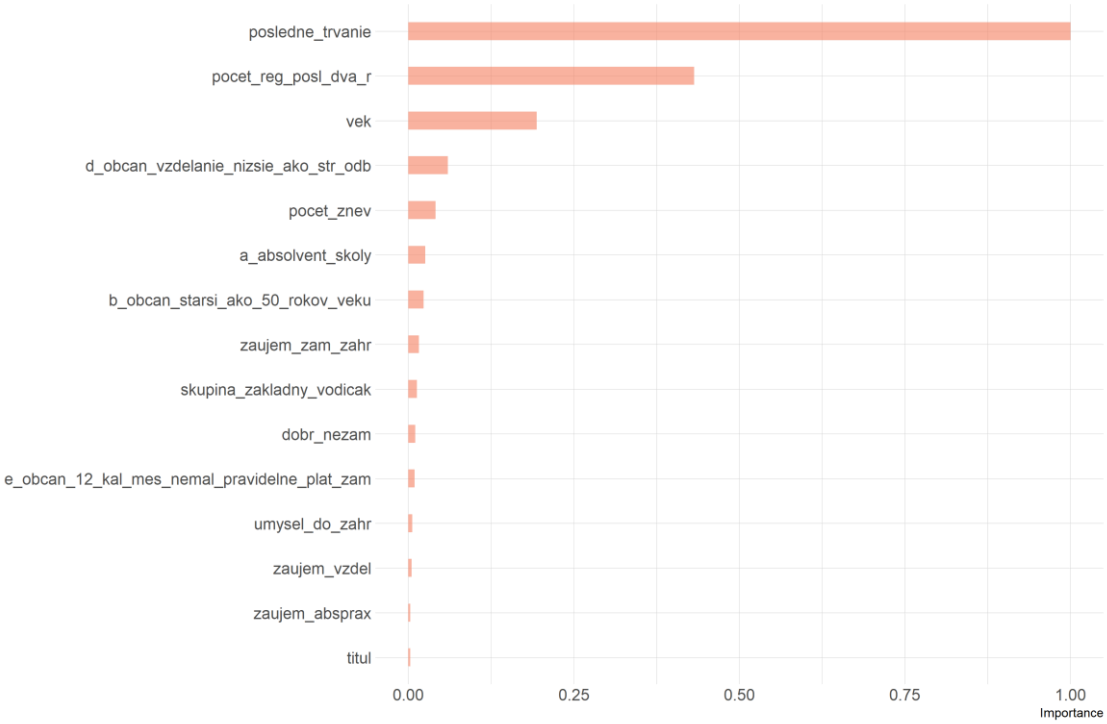
max depth	min rows	mtries	nbins	ntrees	sample rate	MSE tr
10	5	15	25	500	0,700	0,98924
10	5	15	25	300	0,700	0,98927
10	1	15	25	150	0,800	0,98943

Z tab. 15 je vidieť, že pri miernych úpravách parametrov dokážeme MSE tr ovplyvniť až na štvrtom desatinnom mieste. Nakoniec sme pre tieto modely (prvé tri najlepšie z vyhľadávania po mriežke v tab. 13 a tri modely s nami zadanými parametrami v tab. 14) vypočítali predikcie na testovacej sade dát. Výsledky MSE, MAE a RMSE sú v tab. 14 uvedené pre najlepší model. Niektoré modely sa líšili aj pri týchto uvedených metrikách až na štvrtom desatinnom mieste.

Tab. 15: MSE, MAE, RMSE porovnanie najlepšieho modelu pre logaritmickú transformáciu premennej trvanie.

max depth	min rows	mtries	nbins	ntrees	sample rate	MAE	MSE	RMSE
10	5	15	25	500	0,700	0,763	0,985	0,992

Náhodné lesy sú vhodné pre úlohy, v ktorých je veľa prediktorov. Nie všetky nesú významnú informáciu. Najčastejšie sa využíva významnosť. Postup merania významnosti premenných (angl. *importance*) sa dá opísať nasledovne: pre každý strom sa zaznamená chyba predikcie (MSE ob) z *obb* dát, ktoré neboli použité pri trénovaní modelu. Potom sa to isté zopakuje pre permutáciu každého prediktora. Týmto spôsobom sa zistia hodnoty MSE ob pre každý prediktor, ktorý určí jeho významnosť. Toto meranie je často vyjadrené ako percento a je štandardizované na maximálnu hodnotu najvýznamnejšieho prediktoru. Popísaný proces sa opakuje pre všetky prediktory. Ak zámena hodnôt premennej nemá žiadny vplyv na výsledok, tak táto premenná nemá význam. Čím väčší je rozdiel medzi náhodou a skutočnosťou, tým väčší je aj vplyv premennej. [22] Pre model, ktorého parametre sú uvedené v tabuľke 15, je významnosť parametrov zobrazená na obr. 36. Pre ostatné modely bola významnosť parametrov veľmi podobná. Ako významné prediktory sa ukazujú premenné dĺžka posledného trvania, počet registrácií za posledné dva roky, vek, nižšie vzdelanie ako stredné odborné vzdelanie a počet znevýhodnení.



Obr. 36: Významnosť premenných pre logaritmickú transformáciu premennej trvanie. Parametre modelu: max_depth = 10, min_rows = 5, mtries = 15, nbins = 25, ntrees = 500, sample_rate = 0,700.

5.3. XGBoost

Pre tento model existuje viacero parametrov, ktorých hodnoty sa dajú optimalizovať, tak aby výsledná predikcia bola s čo najmenšou chybou. Príkladom takých parametrov sú všeobecné parametre, ktoré sa týkajú zosilovacej techniky (*booster*), parametre vybraného zosilovača, parametre úlohy učenia a iné. Skúmali sme viacero parametrov súvisiacich s nastavenou metódou *gbtree* (*booster* parameter, ktorý je založený na náhodných stromoch).

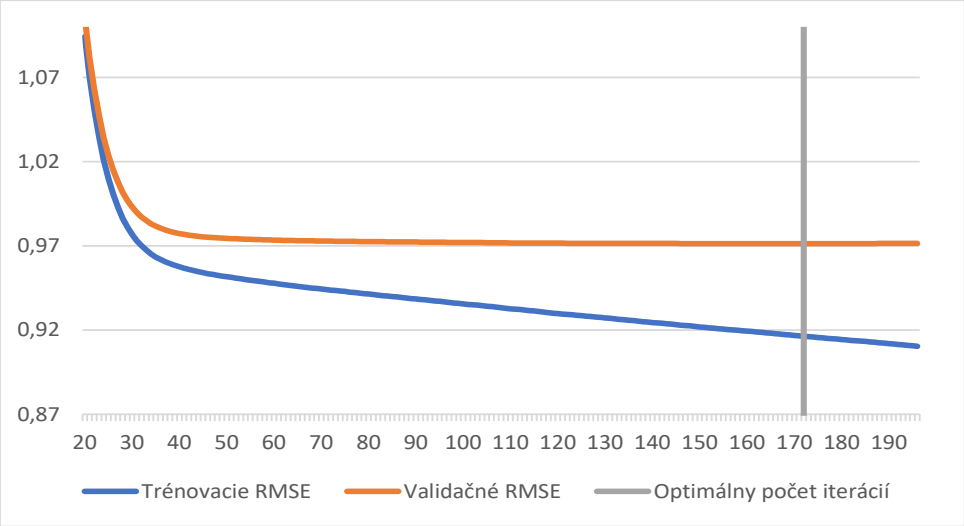
Parametre XGBoost sú maximálna hĺbka stromu (*max_depth*), rýchlosť učenia (*eta*), počet pozorovaní dodaných v strome (*subsample*), počet premenných použitých v strome vyjadrené v percentách pre každú iteráciu (*colsample_bytree*), počet stromov (*nrounds*) a iné. Rozsahy prípustných hodnôt pre vybrané parametre sú uvedené v tab. 16.

Tab. 16: Rozsahy vybraných parametrov pre model XGBoost.

max depth	eta	subsample	min child weight	subsample bytree
(0, ∞)	(0,1)	(0,1)	(0, ∞)	(0,1)

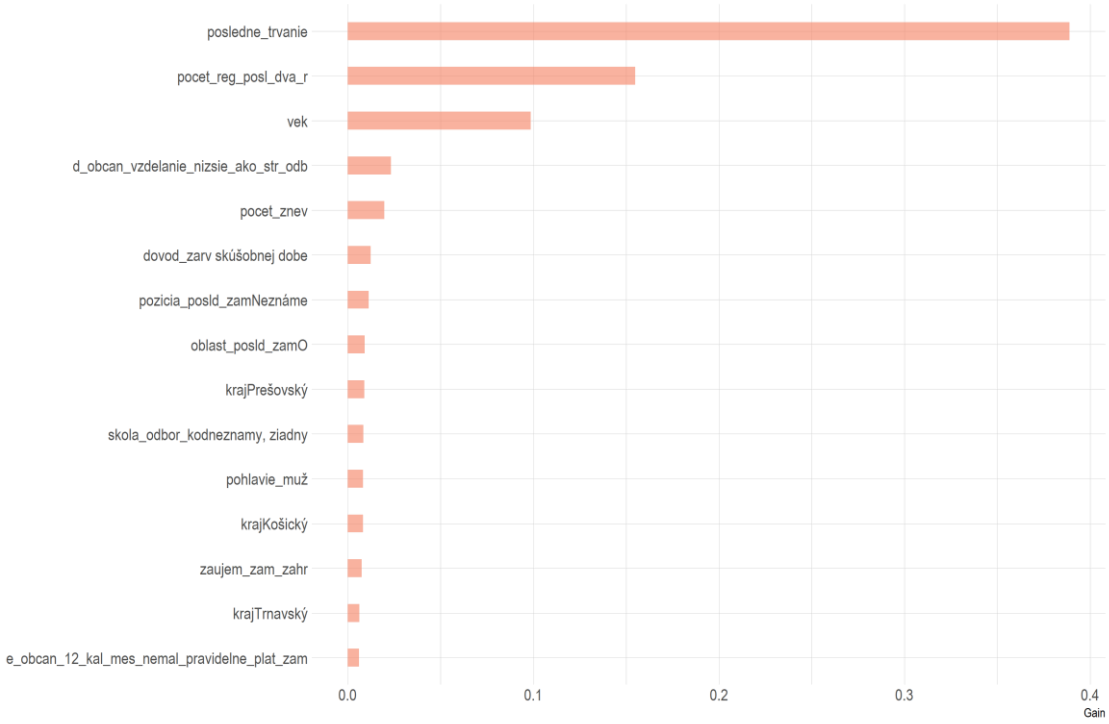
Pre implementáciu v softvéri R sme použili funkciu *xgb.train*(), z knižnice *xgboost*. Najskôr sme pretransformovali dátový súbor z triedy *data.table* na triedu *xgb.DMatrix*. Následne sme zvolili parametre pre 5-súborovú krížovú validáciu (*5-fold cross-validation*) pre nájdenie optimálneho parametra *nrounds*, pomocou funkcie *xgb.cv*(), z knižnice *xgboost*. 10-súborovú krížovú validáciu (*10-fold cross-validation*) sme v tomto prípade nepoužili kvôli technickej a časovej náročnosti výpočtov. Parametre modelu sme zvolili *eta* = 0,1, *gamma* = 0, *max_depth* = 10, *min_child_weight* = 1, *subsample* = 1, *colsample_bytree* = 1. Čím je parameter *eta* bližšie k nule tým sú výpočty časovo náročnejšie. Experimentovali sme aj so zväčšením hĺbky stromu, ale neprinieslo to zlepšienie.

Pre rastúci počet stromov trénovacie RMSE klesá. Toto RMSE by stále klesalo až do konečného počtu stromov. Avšak v istom momente sa validačné RMSE ustáli na nejakej úrovni a už neklesá ale naopak začína narastať. To je spôsobené pretrénovaním modelu, teda model je príliš naviazaný na vstupné dáta a keď mu poskytneme nejaké iné dáta chyba predikcie je vyššia. Optimálny počet iterácií je práve tam, kde je validačné RMSE najmenšie, čo je v našom prípade pri 173 iteráciách. Celkový vývoj trénovacieho a validačného RMSE znázornené na obr. 37.



Obr. 37: Vývoj trénovacieho a validačného RMSE a optimálny počet iterácií (173) pre model XGBoost s parametrami eta = 0,1, gamma = 0, max_depth = 10, min_child_weight = 1, subsample = 1, colsample_bytree = 1, nrounds = 250. Na osi x je uvedený počet iterácií

Na obr. 38 môžeme vidieť výstup z funkcie *xgb.importance*(), z knižnice *xgboost.* Premenné sú zoradené podľa *Gain* (čo je nastaviteľný parameter funkcie *xgb.importance*). Predstavuje prínos každej premennej do modelu. Pre stromové modely sa berie do úvahy prínos každej premennej každého stromu a následne sa z jednotlivých "ziskov" spravý priemer, aby sme získali informáciu pre celý model. Najvýznamnejšími prediktormi sa ukázali premenné posledné trvanie (teda informácia o počte dní posledného trvania evidencie v prípade, ak už bol uchádzač evidovaný na úrade práce za posledné dva roky pred dátumom zaradenia do evidencie), počet registrácií za posledné dva roky, vek, znevýhodnenie d) občan, ktorý dosiahol vzdelanie nižšie ako stredné odborné vzdelanie a počet znevýhodnení. Podľa modelu tieto informácie najviac ovplyvňujú to, ako dlho bude daná osoba nezamestnaná.



Obr. 38: Významnosť premenných pre logaritmickú transformáciu premennej trvanie. Parametre modelu: eta = 0,1, gamma = 0, max_depth = 10, min_child_weight = 1, subsample = 1, colsample_bytree = 1, nrounds = 250.

Nakoniec sme pomocou nášho modelu a funkcie *predict*(), z knižnice *stats* vypočítali predikcie na testovacej sade dát. Porovnanie všetkých modelov v tab. 17 ukazuje, že model XGBoost vykazuje vo všetkých metrikách najmenšiu chybu.

Tab. 17: Sumarizačná tabuľka výsledkov MAE, MSE, RMSE pre najlepšie modely. Parametre modelu náhodného lesa: max_depth = 10, min_rows = 5, mtries = 15, nbins = 25, ntrees = 500. Parametre modelu XGBoost: eta = 0,1, gamma = 0, max_depth = 10, min_child_weight = 1, subsample = 1, colsample_bytree = 1, nrounds = 250.

Model	MAE	MSE	RMSE
Lineárna regresia	0,80181	1,03904	1,01933
Ridge regresia	0,80365	1,04109	1,02034
Lasso regresia	0,80257	1,04061	1,02010
Náhodný les	0,76393	0,98503	0,99248
XGBoost	0,74006	0,93507	0,96699

6. Diskusia

Prvým dôležitým čiastkovým výsledkom, ktorý sme dosiahli bolo spracovanie a úprava vstupných dát (nezávislých premenných) tak, aby boli vhodné ako vstupy do použitých modelov. Pôvodné dáta v niektorých prípadoch poskytovali príliš špecifické informácie o uchádzačoch (napr. klasifikácie zamestnania SK ISCO-08, kde sa nachádza až 10 000 kategórií), inde boli prítomné takmer totožné informácie (napr. školský odbor posledného dosiahnutého vzdelania a najvyššieho dosiahnutého vzdelania). V prípade takmer totožných informácií o UoZ sme vybrali jednu premennú a druhú sme nepoužili. V prípade špecifických premenných sme využívali všeobecnú kategóriu ostatné alebo iné, prípadne sme zlúčili kategórie, ktoré nám poskytli obsahovo podobnú informáciu. Napr. do kategórie svetových jazykov sme zlúčili francúzsky, taliansky, španielsky a pod. Ďalším príkladom je typ úväzku, o ktorý má UoZ záujem, kde sme v kategórii čiastočný zlúčili rôzne typy čiastočných úväzkov (polovičný, tretinový, atď).

Ako druhý čiastkový výsledok uvádzame niektoré premenné, ktoré primárne slúžia pre ďalšiu analýzu v štatistických modeloch, bližšie popísaných v kapitole 4, avšak majú samé o sebe zaujímavú výsledkovú hodnotu. Ich význam umocňuje to, že pochádzajú z veľmi robustného dátového súboru, ktorý obsahoval údaje približne o 1,2 milióna zaradeniach UoZ v rokoch 2015 až 2018. Priemerná dĺžka UoZ činila 176,5 dňa. Väčšie percentuálne zastúpenie mali muži 52,1 %, ženy boli zastúpené o 4,2 percentuálneho bodu menej. Uvedené rozdiely medzi pohlaviami, teda neboli výrazné. Medzi UoZ bola väčšina z rodinným stavom slobodný (56,3 %), pričom ženatých/vydatých bolo iba 33,6 %. Jedným z možných vysvetlení je, že celkový počet sobášov v SR a počet detí narodených v manželstve z roka na rok klesá [29], a teda aj celkový počet manželstiev je nižší oproti slobodným občanom, čo môže čiastočne odzrkadľovať aj situáciu v evidencii UoZ. Ďalšie možné vysvetlenie je, že osoby žijúce v manželstve majú záväzky voči iným (manželke/manželovi, deťom). Nemať žiadny zdroj príjmu je pre nich oveľa závažnejšia skutočnosť ako pre tých, ktorí tieto záväzky nemajú a sami sú nútení aktívnejšie a zodpovednejšie pristupovať k hľadaniu zamestnania. Avšak z hľadiska priemernej dĺžky trvania evidencie UoZ slobodný uchádzač strávi na úrade 155 dní a ženatý/vydatá 208 dní. Táto skutočnosť môže súvisieť s tým, že slobodní UoZ sú variabilnejší v hľadaní práce ako ženatý/vydatá. Väčšina UoZ (90,6 %) má slovenskú národnosť, nasleduje maďarská národnosť (8,3 %) a iná (1,1 %). Nízke percentuálne zastúpenie v kategórii iná môže byť spôsobené tým, že občania iných národností z pohraničných oblastí (česko, poľsko, ukrajina, rakúsko) pracujú v susedných krajinách. Taktiež niektorí, z týchto občanov, môžu

považovať uvádzanie slovenskej národnosti za benefit na trhu práce. Pri hodnotení vzdelania UoZ sme zistili, že 16,7 % nemá ukončené vzdelanie alebo má ukončené iba základné vzdelanie. Tento jav môže mať negatívne dôsledky v tom, že nebude schopný porozumieť pracovným postupom a pod. Štandardne sa na trhu práce môžu vyžadovať určité znalosti UoZ. Konkrétne sa jedná o ovládanie angličtiny ako cudzieho jazyka alebo používanie balíka MS Office. Z dát vyplýva, že 65,6 % uchádzačov neovláda MS Office ani na elementárnej úrovni a 72,7 % neovláda anglický jazyk. Preto ostáva otázkou či UoZ skutočne disponujú takou nízkou počítačovou gramotnosťou alebo sa jedná len o malý dôraz pri vypĺňaní dotazníka na úrade práce. Z hľadiska počtu detí iba 13,5 % UoZ má deti do 16 rokov. Veríme, že za týmto je skutočnosť, že rodičia celkovo viac predchádzajú situácii aby sa na úrad práce vôbec dostali. Iným zaujímavým zistením je, že najviac UoZ je z Prešovského a Košického kraja a najmenej z Trnavského a Bratislavského, čo odráža centralizáciu pracovných príležitostí viac na západ. Čo sa týka dôvodov zaradenia do evidencie, je najčastejším dôvodom dobrovoľná nezamestnanosť (28 % UoZ). Aj iné premenné poskytujú zaujímavé informácie, ich bližšia diskusia presahuje rozsah tohto školského diela. Hlavným účelom spomínaných a aj iným premenných je ich použitie pri vytváraní štatistických modelov, ktoré slúžia na predikciu dĺžky trvania nezamestnanosti.

Tretí výsledok predstavuje exploratívna analýza, kde sme skúmali priemerné dĺžky trvanie jednotlivých UoZ, sme zistili, že UoZ, ktorí nedisponujú ani základným vodičským preukazom strávia na úrade priemerne 191 dní, čo je o 42 dní viac ako tí, ktorí majú aspoň základný vodičský preukaz. To môže byť spôsobené tým, že osobám s vodičským oprávnením sa otvára väčšia škála pracovných príležitostí napr. kuriér, osobný vodič a pod. Ďalej sme zistili, že niektoré typy znevýhodnení nie sú priťažujúcou okolnosťou z hľadiska priemernej dĺžky trvania evidencie. Priemerné trvanie evidencie so znevýhodnením absolvent školy je 130 dní a pre UoZ, ktorí toto znevýhodnenie nemajú je to 183 dní. Čo ukazuje na to, že absolventi škôl sú schopní pomerne rýchlo a efektívne reagovať na pracovné príležitosti.

Dôležitou skutočnosťou je, že pri konštruovaní väčšiny modelov sme pracovali s logaritmickou transformáciou premennej trvanie. Pri lineárnych regresných modeloch sme narazili, kvôli silným predpokladom normalite rezíduí a heteroskedasticity. Pôvodná premenná trvanie pochádzala zo šikmého rozdelenia, čo viedlo k tomu, že rezíduá nepochádzali z normálneho rozdelenia. Tento problém sme sa pokúsili vyriešiť tak, že sme vytvorili tri podoby premennej dĺžky trvania evidencie: ponechali sme pôvodnú premennú, ďalej sme sa pozreli na logaritmickú transformáciu premennej trvanie a trvanie ohraničené na maximálne dva roky evidencie (ak UoZ bol na úrade viac ako dva roky, priradili sme mu trvanie práve dva roky).

Pre všetky tri podoby premennej trvanie sme zostrojili model klasickej lineárnej regresie, ridge regresie a lasso regresie a okrem metrík MAE, MSE a RMSE sme porovnali aj R^2. Táto hodnota nám vyšla najlepšie pre logaritmickú transformáciu premennej trvanie. Čo znamená, že spomedzi zvolených podôb premennej trvanie najviac chyby vysvetlenej modelom poskytujú dáta po logaritmickej transformácii.

Najdôležitejším zistením je, že zo všetkých skúmaných modelov má najnižšiu chybu predikcie vo všetkých porovnávaných metrikách XGBoost (MAE: 0,74006, MSE: 0,93507, RMSE: 0,96699). Stromové metódy dosiahli lepšie výsledky ako rôzne druhy lineárnej regresie. To môže mať súvis s tým, že nemajú silné predpoklady ako napr. spomínaná lineárna regresia (normalita rezíduí) a že sú schopné zachytiť aj nelineárne vzťahy medzi premennými a interakcie, ktoré by sme pri regresii museli zadať manuálne.

Pri stromových modeloch (náhodný les a XGBoost) sa ako najvýznamnejšie prediktory ukázali posledné trvanie evidencie, počet registrácií za posledné dva roky, vek, či UoZ má nižšie ako stredné odborné vzdelanie a počet znevýhodnení. To nám hovorí o tom, že tieto premenné sú dôležité na určenie výslednej dĺžky trvania evidencie. Pri lineárnych modeloch boli medzi dôležitými prediktormi znevýhodnenia absolvent školy, mesiac zaradenia, MS Office a tiež aj počet znevýhodnení a počet registrácií za posledné dva roky. Vidíme, že aj toto poukazuje na to, že predpovedať ako dlho bude niekto nezamestnaný, je zložitý problém, na ktorý vplývajú rôzne faktory. K týmto faktorom patrí aj výber vhodného modelu.

Záver

V predkladanej práci sme sa zaoberali modelovaním dĺžky trvania evidencie uchádzačov o zamestnanie. Potrebné dáta nám poskytlo Ústredie práce, sociálnych vecí a rodiny. Použité premenné sme podrobne popísali v druhej kapitole našej práce.

Metódy strojového učenia sa používajú bežne v rôznych oblastiach od predikcie vývoja ceny Bitcoinu až po predpovedanie počasia. Na výber sme mali mnoho existujúcich modelov, avšak nedá sa jednoznačne povedať aké modely zvoliť a otestovať. Vzhľadom na povahu našich dát a charakter závislej premennej sme sa rozhodli porovnávať dve triedy modelov lineárnu regresiu a stromové metódy. Pri lineárnej regresii sme porovnávali tri modely: klasickú lineárnu regresiu, ridge regressiu a lasso regresiu. Pri stromových metódach sme použili modely náhodného lesa a XGBoost.

Lineárna regresia má dôležitý predpoklad normality rezíduí, čo je väčšinou pri dátach z reálneho sveta ťažko dosiahnuteľné. Toto nám potvrdili aj vizuálne a exaktné testy na overenie normality spomínaných rezíduí. Preto sme sa pomocou logaritmickej transformácie premennej trvanie pokúsili priblížiť normálnemu rozdeleniu podobne ako v [14, 21]. Pre ridge regresiu sme zistili, že nezávislé premenné, ktoré ani pri vysokej penalizácii λ neostali blízke nule (teda sú pri určovaní predikcie dôležité), sú ukončené vysokoškolské vzdelanie, dôvod zaradenia do evidencie (zdravotné dôvody a skončenie poberania príspevku), mesiac zaradenia máj a jún a ďalšie. Pre lasso regresiu sme zistili, že parametre, ktoré neostali nulové sú počet registrácií za posledné dva roky, nemecký jazyk na vysokej úrovni, počet znevýhodnení a iné. Najlepší model spomedzi lineárnych modelov pre všetky zvolené metriky MAE, MSE, RMSE je klasický lineárny model.

Použitie náhodného lesa nekladie podmienky na rozdelenie chýb predikcie ani na vstupné dáta. Po odskúšaní rôznych parametrov zvolenými postupmi sme našli model s najmenšou chybou na testovacej vzorke dát. Parametre tohto modelu sú uvedené v tab. 14. Ako významné prediktory sa pre tento model ukazujú posledné trvanie, počet registrácií za posledné dva roky, vek, znevýhodnenie d) občan, ktorý dosiahol nižšie ako stredné odborné vzdelanie, počet znevýhodnení a iné.

Pre model XGBoost sme na začiatok použili vybrané parametre z najlepšieho modelu náhodného lesa. Napríklad parameter maximálna hĺbka stromu nastavená na hodnotu 10. Keď sme skúšali nastaviť vyššiu hodnotu, konkrétne 15, neviedlo to k zlepšeniu. Pre určenie parametra optimálneho počtu stromov sme použili 5-súborovú krížovú validáciu. Optimálny počet iterácií pre najmenšie validačné RMSE je pri 173 iteráciách. Najvýznamnejšími

prediktormi sa ukázali premenné posledné trvanie (teda informácia o počte dní posledného trvania evidencie v prípade, ak už bol uchádzač evidovaný na úrade práce za posledné dva roky pred dátumom zaradenia do evidencie), počet registrácií za posledné dva roky, vek, znevýhodnenie d) občan, ktorý dosiahol vzdelanie nižšie ako stredné odborné vzdelanie a počet znevýhodnení. Podľa modelu tieto informácie najviac ovplyvňujú to, ako dlho bude daná osoba nezamestnaná.

Môžeme si všimnúť, že pre lineárne regresné modely sa ukázali ako dôležité premenné mierne odlišné informácie o UoZ ako pri stromových modeloch. Táto skutočnosť poukazuje na to, že predpovedať ako dlho bude niekto nezamestnaný, je zložitý problém, na ktorý vplývajú rôzne faktory. K týmto faktorom patrí aj výber vhodného modelu.

Na základe výsledkov tejto práce je model s najmenšou chybou podľa metrík MAE, MSE a RMSE XGBoost. Tento model sa môže aplikovať pri registrácii UoZ na úrade práce a môže pomôcť pri identifikovaní potenciálnych dlhodobých nezamestnaných.

Použitá literatúra

[1] 5/2004 Z. z., ZÁKON, zo 4. decembra 2003 o službách zamestnanosti a o zmene a doplnení niektorých zákonov, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2004/5/20060101.html

[2] 150/2013 Z. z., ZÁKON, z 15. mája 2013 o Štátnom fonde rozvoja bývania, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2013/150/20160101.html

[3] 306/2007 Z. z., VYHLÁŠKA Štatistického úradu Slovenskej republiky z 18. júna 2007, ktorou sa vydáva Štatistická klasifikácia ekonomických činností, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2013/150/20160101.html

[4] 336/2015 Z. z., ZÁKON, z 11. novembra 2015 o podpore najmenej rozvinutých okresov a o zmene a doplnení niektorých zákonov, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2015/336/

[5] 396/2012 Z. z., ZÁKON, z 5. decembra 2012 o Fonde na podporu vzdelávania, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2012/396/20190901

[6] 449/2020 Z. z., VYHLÁŠKA Štatistického úradu Slovenskej republiky z 18. decembra 2020, ktorou sa vydáva štatistická klasifikácia zamestnaní, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2020/449/

[7] 571/2006 Z. z., VYHLÁŠKA Štatistického úradu Slovenskej republiky z 2. októbra 2006, ktorou sa vydáva Štatistická klasifikácia odborov vzdelania, dostupné na internete (30.6.2021): https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2006/571/20061101.html

[8] Albada S.J., Robinson P.A., 2007. Transformation of arbitrary distributions to the normal distribution with application to EEG test–retest reliability, Journal of Neuroscience Methods, vol. 161, pp 205-211. ISSN 01650270, dostupné na internete (30.6.2021): arXiv:1801.01748

[9] Anděl J., 2005. Základy matematické statistiky. MATFYZPRESS vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze. ISBN: 8086732401

[10] Bali R., Sarkar D., Lantz B., Lesmeister C., 2016. R: Unleash Machine Learning Techniques. Packt Publishing. ISBN: 9781787127340

[11] Bakošová A., Bulla R., Kozárová E., Bugár Š., 2017, „Nezamestnateľní“nezamestnaní, IA MPSVR SR, ISBN 9788089837304

[12] Bartlett M. S., 1947. The use of transformations. International Biometric Society, vol. 3, pp. 39–52. Doi:10.2307/3001536

[13] Bentéjac C., Csörgő A., Martínez-Muñoz G., 20219. A Comparative Analysis of XGBoost, vol. 54, pp. 1937-1967, Artificial Intelligence Review, Springer Science and Business Media LLC, ISSN=15737462, dostupné na internete (30.6.2021): arXiv:1911.01914

[14] Box G. E. P., Cox D. R., 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), vol. 26, pp. 211 – 243.

[15] Breiman L., 2001. Random Forests. Machine Learning, vol. 45, pp. 5 – 32., dostupné na internete (): Doi: 10.1023/A:1010933404324

[16] Chan-Lau J., 2017, Lasso Regressions and Forecasting Models in Applied Stress Testing, IMF Working Paper

[17] Chen T., Guestrin C., 2016. XGBoost: A Scalable Tree Boosting System., dostupné na internete (): ArXiv: 1603.02754v3, doi: 10.1145/2939672.2939785

[18] Desiere S., Langenbucher K., Struyven L., 2019. Statistical profiling in public employment services: An international comparison. OECD Social, Employment and Migration Working Papers, No. 224, OECD Publishing., dostupné na internete (30.6.2021): https://doi.org/10.1787/b5e5f16e-en.

[19] Friedman J., Hastie T., 1998, Tibshirani R., Regularization Paths for Generalized Linear Models via Coordinate Descent., J Stat Softw., dostupné na adrese (30.6.2021): https://pubmed.ncbi.nlm.nih.gov/20808728/

[20] Hastie T., Tibshirani R., Friedman J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer, New York. ISBN: 9780387848587

[21] Keene O. N., 1995. The log transformation is special. Statistics in Medicine, vol. 14, pp. 811 – 819., dostupné na internete (): doi: 10.1002/sim.4780140810

[22] Komprdová K., 2012. Rozhodovací stromy a lesy. Akademické nakladatelství CERM, s.r.o. Brno. ISBN: 9788072047857

[23] Loxha A., Morgandi M., 2014. Profiling the unemployed: a review of OECD experiences and implications for emerging economics. Social Protection Discussion Papers and Notes 91051, The World Bank., Paris, dostupné na internete (30.6.2021): https://openknowledge.worldbank.org/handle/10986/20382

[24] Ma M., Zhao G., He B., Li Q., Dong H., Wang S., Wang Z., 2021. XGBoost-based method for flash flood risk assessment. Vol. 598, pp. 1 – 12, dostupné na internete (): doi: 10.1016/j.jhydrol.2021.126382

[25] Řehák P., 2020, Poznámky k úvodu do funkcionální analýzy, Brno

[26] Scoppetta A., Buckenleib A., 2018. Tackling Long-Term Unemployment through Risk Profiling and Outreach. ESF Transnational Platform, Brusel., dostupné na internete (30.6.2021): https://www.euro.centre.org/publications/detail/3166

[27] Shi X., Li Q., Qi Y., Huang T., Li J., 2017. An accident prediction approach based on XGBoost, 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 1 – 7., dostupné na internete (): doi: 10.1109/ISKE.2017.8258806

[28] Soukup T., Michalička L., Kotíková J., 2009. Třídění uchazečů na úřadech práce - řešení problematiky cílení APZ a poradenství, Výzkumný ústav práce a sociálních věcí, v.v.i., Praha, ISBN 9788074160271

[29] ŠÚ SR: Prehľad pohybu obyvateľstva - SR, oblasti, kraje, okresy, mesto, vidiek (ročne). Webové sídlo Štatistického úradu Slovenskej republiky, Databázy, STATdat., Demografia a sociálne štatistiky, Obyvateľstvo a migrácia, dostupné na adrese (30.6.2021): http://www.statistics.sk/

[30] Troya Í. M., Chen R., Moraes L. O., Bajaj P., Kupersmith J., Ghani R., Brás N. B., Zejnilovic L., 2018. Predicting, explaining, and understanding risk of long-term unemployment. 32nd Conference on Neural Information Processing Systems, Montreal.

[31] Ústredie práce, sociálnych vecí a rodiny, Nezamestnanosť - mesačné štatistiky, dostupné na internete (30.6.2021): https://www.upsvr.gov.sk/statistiky/nezamestnanost-mesacne-statistiky.html?page_id=1254

[32] Wright M. N., Ziegler A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, vol. 77., dostupné na internete (): ArXiv: 1508.04409v2, doi: 10.18637/jss.v077.i01

[33] Zumel N., Mount J., 2019. Practical Data Science with R (2nd ed.). Manning Publications. ISBN: 9781617295874

[34] Zwillinger D., 2002. CRC Standard Mathematical Tables and Formulae (31st ed.) (Advances in Applied Mathematics). Chapman and Hall/CRC. ISBN: 9781584882916