UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



NÁVRH ALGORITMOV NA OBJAVOVANIE PODOZRIVÉHO SPRÁVANIA POSKYTOVATEĽOV ZDRAVOTNEJ STAROSTLIVOSTI

DIPLOMOVÁ PRÁCA

2019 Patricia SVITKOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

NÁVRH ALGORITMOV NA OBJAVOVANIE PODOZRIVÉHO SPRÁVANIA POSKYTOVATEĽOV ZDRAVOTNEJ STAROSTLIVOSTI

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr.Henrieta Tulejová, MSc.

Bratislava 2019 **Patricia SVITKOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Patricia Svitková
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Návrh algoritmov na objavovanie podozrivého správania poskytovateľov zdravotnej starostlivosti.
Design of algorithms for detecting suspicious behaviour of health care providers.

Anotácia: Jednou z úloh zdravotnej poisťovne je kontrola efektívneho vynakladania finančných prostriedkov verejného zdravotného poistenia. Na úplnú kontrolu vykázaných dát však poisťovňa nemá dostatočné kapacity, preto môže skontrolovať iba vybranú skupinu dát. V dnešnej dobe je na Slovensku veľký priestor na zlepšovanie kontrolnej činnosti a odhaľovanie podvodov. Cieľom práce bude navrhnúť, vyskúšať a optimalizovať rôzne algoritmy, ktoré uľahčia revíznym lekárom výber dát alebo poskytovateľov, na ktoré sa treba pri kontrole zamerať.

Vedúci: Mgr. Henrieta Tulejová, MSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 25.01.2018

Dátum schválenia: 25.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcela poďakovať svojej školiteľke, pani Mgr. Henriete Tulejovej, MSc. za ponúknutú príležitosť, odborné rady a odporúčania. Ďalej moja veľká vďaka patrí Mgr. Marekovi Ciesarovi za jeho venovaný čas, pomoc a ochotu kedykoľvek zodpovedať moje otázky. Taktiež ďakujem kolektívu v poisťovni Dôvera za umožnené odborné konzultácie. V neposlednom rade veľké poďakovanie patrí mojej rodine a priateľom za ich trpezlivosť, pochopenie a neobyčajnú podporu.

Abstrakt

SVITKOVÁ, Patricia: Návrh algoritmov na objavovanie podozrivého správnia poskytovateľov zdravotnej starostlivosti [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Henrieta Tulejová, MSc., Bratislava, 2019, 66 s.

Cieľom práce bolo navrhnúť spôsob detekovania podozrivého správania poskytovateľov zdravotnej starostlivosti. Konkrétne sme sa venovali preverovaniu všeobecných lekárov. Modely boli skonštruované na predikciu priemerných mesačných nákladov pacientov na jednotlivé skupiny liekov a zdravotníckych pomôcok, pričom sme porovnali viacero metód, konkrétne lineárnu regresiu, rozhodovací strom CART a náhodný les. Výzvou bolo aj spracovanie veľkého množstva údajov, navrhnutie vhodných vstupných premenných, či vymyslenie objektívneho detekovania podozrivého správania outlierov.

Kľúčové slová: podozrivé správanie, podvod, outlier, predikcia nákladov, zdravotná starotlivosť, všeobecný lekár, lineárna regresia, rozhodovací strom, náhodný les

Abstract

SVITKOVÁ, Patricia: Design of algorithms for detecting of suspicious behavior of healthcare providers [Diploma thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Henrieta Tulejová, MSc., Bratislava, 2019, 66 s.

The aim of thesis was to propose a way of detecting suspecious behavior of healthcare providers. Specifically, we have been dedicated to general practitioners and the models were cunstructed to predict the average monthly cost capacity of patients per group of drugs and medical devices. We used and compared more methods, in the concrete linear regression, CART decision tree and random forest algorithm. The challenge was also to process a large amount of data, design appropriate input variables, or devise objective detection of suspicious outlier behavior in the model.

Keywords: suspisious behavior, fraud, outlier, prediction of costs, health care, general practitioner, linear regression, decision tree, random forest

Obsah

Úvod **8**

1 Priblíženie problematiky **10**

1.1 Pojmy 10

1.2 Zameranie práce 11

1.3 Prehľad známych prístupov a metód zo zahraničia 12

1.3.1 Prístup - "podozrivý" alebo "v poriadku" 12

1.3.2 Predikcia nákladov 13

2 Matematické nástroje **15**

2.1 Lineárna regresia 15

2.1.1 Úvod 15

2.1.2 Odhad metódou najmenších štvorcov 15

2.1.3 Gauss - Markovove podmienky 16

2.2 Rozhodovacie stromy 17

2.2.1 Úvod 17

2.2.2 Popis algoritmu 17

2.2.3 Pruning 19

2.2.4 Náhodný les 20

2.3 Metriky na odhadnutie presnosti modelov 21

2.4 K-means zhluková analýza 23

3 Zatriedenie liekov a zdravotníckych pomôcok **24**

3.1 Zatriedenie liekov 24

3.2 Zatriedenie zdravotníckych pomôcok 25

3.3 Výsledné skupiny 25

3.4 Skupiny v tejto práci 26

4 Výber dát a regresných premenných **27**

4.1 Dáta 27

4.2 Regresné premenné 27

4.2.1 Demografické premenné 27

4.2.2 Premenné z VZS 31
4.2.3 Vysvetľované premenné 33

5 Diagnostika lineárneho modelu **35**
5.1 Budovanie modelu 35
5.2 Korelácia vysvetľujúcich premenných 36
5.3 Overenie Gauss-Markových podmienok 38
5.3.1 Nulovosť strednej hodnoty rezíduí 38
5.3.2 Homoskedasticita 38
5.3.3 Normalita 38

6 Optimalizácia rozhodovacích stromov **41**
6.1 CART rozhodovací strom 41
6.2 Náhodný les 43

7 Porovnanie modelov **46**

8 Analýza outlierov **49**

9 Modely s predošlými nákladmi **54**

Literatúra **57**

Úvod

S výrazným rozvojom dátovej vedy v posledných desaťročiach možno pozorovať čím ďalej tým väčší dôraz malých aj veľkých firiem na ukladanie firemných dát a ich následnú analýzu. Dáta nám častokrát môžu odhaliť, v akej sfére daného biznisu by sa mala firma viacej rozvíjať, kde je možno zachytený nejaký problém v systéme, na ktorom odstránení by mala zapracovať alebo ako lepšie zacieliť na potenciálneho zákazníka. Cieľom vždy ostáva zo svojich dát získavať potrebné informácie, poučiť sa z daných zistení a následne naložiť s týmito informáciami tak, aby to viedlo k efektívnejšiemu chodu celej firmy.

Výnimkou by nemal byť ani sektor zdravotníctva, v ktorom sa ročne vyprodukuje nespočetné množstvo dát a obrátia miliardy eur. Práve vďaka správnym analýzam týchto dát sa dá neustále pracovať na zefektívnení nakladania s finančnými prostriedkami, pretože ak sa v tomto sektore niekde neefektívne minú, inde budú chýbať. Aj toto je dôvodom, prečo je potrebné napríklad aj preverovanie podozrivého správania lekárov, či lekární. V minulosti sa zdravotným poisťovniam podarilo odhaliť viaceré nelegálne praktiky lekárov. Môžeme spomenúť napríklad nedávne víťazstvo poisťovní, kedy sa im podarilo vyhrať dlhoročný súd so sieťou spolupracujúcich lekárov a lekární v oblasti podvádzania s konkrétnymi zdravotníckymi pomôckami typu A (vlhké krytie). Tento fakt nás môže len utvrdiť v tom, že táto snaha o preverovanie podozrivých javov má svoje opodstatnenie a zmysel.

V prvej kapitole sa zoznamujeme so zaužívanou terminológiou v zdravotných poisťovniach. Priblížime si s akými typmi podovodv zo strany lekárov sa v zdravotníctve stretávame. Taktiež si predstavíme rôzne nazerania na odhaľovanie podozrivého správania v zahraničí a vysvetlíme, ako funguje tzv. PCG model predikcie nákladov poistencov s chronickými chorobami.

V našej práci sa ďalej zameriame na predikciu uhradených nákladov poistenca na určité skupiny produktov. V druhej kapitole vysvetlíme fungovanie matematických nástrojov a metód, ktoré budú následne prakticky aplikované na naše dáta. Pojde o viacero metód na predikciu sledovaných nákladov pacienta. Dopomôže nám k tomu lineárna regresia, rozhodovacie stromy, či algoritmus náhodných lesov.

V tretej kapitole si vytýčime, ktorým skupinám liekov a pomôcok sa budeme pre

účely tejto práce venovať, prečo vidíme prínosné sa venovať práve im a ako sme si tieto skupiny zadefinovali.

V kapitole 4 predstavujeme náš dátový súbor a premenné, ktoré budeme postupne do modelu vkladať, pričom uvedieme nejaké ich základné vlastnosti a štatistiky. Pojde o dva typy premenných, jedny by sme mohli pomenovať ako demografické a druhé sú spojené s informáciami o diagnózach a čerpaných typoch zdraotnej starostlivosti.

V kapitole 5 postupne budujeme model lineárnej regresie, ktorý následne otestujeme na klasické podmienky ako normalita chýb, či homoskedasticita.

V ďalšej kapitole potom podľa vzoru výsledného modelu z lineárnej regresie aplikujeme na dané dáta rozhodovacie stromy CART a náhodne lesy a zoptimalizujeme hyperpametre, ktoré budeme pri daných motódach zadávať. V kapitole 7 tieto metódy nakoniec porovnáme podľa viacerých kritérií.

V ôsmej kapitole prichádza narad samotná analýza outlierov vrámci sledovaných lekárov. Vysvetlíme jeden z možných spôsob určovania outlierov a na čo si pri tom dávať pozor a ponúkneme výsledné detekovanie pre všetky tri metódy predickie nákladov v oboch sledovaných skupinách.

Na záver v poslednej kapitole navrhneme miernu modifikáciu modelu pridaným predošlých nákladov z jedného roku pred sledovaným rokom a v krátkosti porovnáme získaných outlierov z pôvodného modelu a z tohto modifikovaného, čo nám môže priniesť nový typ infomácie o detekovaných podozrivých správaniach poskytovateľov zdravotnej strarostlivosti..

1 Priblíženie problematiky

V tejto kapitole si bližšie predstavíme terminológiu využívanú vo verejnom zdravotnom poistení, ktorá nás bude touto prácou sprevádzať. Uvedieme základné typy podvodov, s ktorými sa poisťovne dostávajú do kontaktu. Predstavíme konkrétne zameranie našej práce a v krátkosti zhrnieme, aké rôzne prístupy na detekovanie podozrivého správania poskytovateľov existujú v zahraničí.

1.1 Pojmy

Pod pojmom *poskytovateľ zdravotnej starostlivosti* , skrátene PZS chápeme fyzickú alebo právnickú osobu, ktorá má povolenie na poskytovanie zdravotnej starostlivosti. PZS môžeme však chápať na troch úrovniach. V širšom slova zmysle ide o spomínanú fyzickú, či právnickú osobu, ktorá sa zvykne označovať aj ako *hlavná zložka*. Ide napríklad o nemocnicu, polikliniku, lekáreň a pod. Okrem hlavnej zložky má PZS (až na zopár výnimiek spravidla jednu) *vyššiu zložku*. Pod každú vyššiu zložku spadá viacero *nižších zložiek*. Nižšie zložky sú už konkrétne ambulancie, či laboratória vrámci nemocnice, či polikliniky. V našej práci budeme pod pojmom PZS mať na mysli práve jednotlivé nižšie zložky.

Pod *kapitáciou* sa rozumie mesačná paušálna platba, ktorú dostáva lekár od poisťovne za každého registrovaného poistenca bez ohľadu na to, či lekára v danom mesiaci navštívil, alebo nie. Takýto spôsov vyplácania sa týka všeobecných lekárov, gynekológov. Zubné pohotovosti dostávajú túto platbu od počtu poistencov v ich spádovej oblasti.

Pod pojmom vykázaná zdravotná starostlivosť (VZS) sa rozumejú všetky údaje, ktoré získava poisťovňa od svojich zazmluvnených poskytovateľov zdravotnej starostlivosti o poskytnutých výkonoch na svojich poistencoch, vrámci čoho sa nachádzajú informácie o jednotlivých výkonoch (či šlo o odber krvi, vyšetrenie rôznymi prístrojmi, či predpis lieku/pomôcky alebo aj napríklad odbornú radu atď.), na akom pacientovi, v aký deň s akou diagnózou boli vykonané a veľa ďalších.

Pre lepšiu ilustráciu a priblíženie si problematiky našej práce si teraz predstavme najčastejšie typy podvodov, ktorých sa niektorí lekári v súčastnosti môžu dopúšťať.

Do podvodnej činnosti lekárov spadá:

- vykázanie výkonu na pacientovi, ktorý reálne u lekára nikdy nebol
- vykázanie vyššieho množstva výkonov, než bolo reálne na pacientovi prevedených
- vykázanie iných výkonov (drahších) s inými diagnózami v nesúlade so skutočnosťou
- zdieľanie pacientov a ich osobných údajov medzi viacerými poskytovateľmi (s tým spojené aj následné fiktívne vykazovanie výkonov na týchto pacientoch)
- signifikantne vyššie reálne vykonávanie drahších výkonov oproti ostatným lekárom v danej odbornosti bez závažného dôvodu

Problematike zdieľania osobných údajov pacientov medzi poskytovateľmi a hľadanie algoritmu na rátanie spoločných pacientov medzi n-ticami poskytovateľov sa už venovala napríklad diplomová práca [19]. Ktorému typu podvodu sa budeme venovať v našej práci objasníme v nasledujúcej podkapitole.

1.2 Zameranie práce

Rozhodli sme sa zamerať svoju pozornosť na konkrétnu skupinu lekárov - na všeobecných lekárov pre dospelých. Našim cieľom je predikovať, aké by mali mať priemerné mesačné náklady na preskripciu liekov a zdravotníckych pomôcok za rok 2017. Pod zdravotníckymi pomôckami si môžeme predstaviť napríklad korzety na chrbát, rôzne typy bandáží, vložky do topánok, či napríklad plienky pre inkontinetných pacientov. Preventívne prehliadky, rôzne odbery a vyšetrenia do týchto nákladov nebudeme zahŕňať a teda nebudú objektom nášho skúmania.

Našim podozrením je, že niektorí PZS môžu fiktívne vykazovať lieky a pomôcky na svojich pacientov, resp. vykazujú ich vo vyššom množstve ako boli reálne predpísané. Teda našim cieľom v danom algoritme je nejakým spôsobom detektovať takéto druhy správania. Teraz stručne predstavíme jednotlivé etapy, ktorými je potrebné prejsť:

1. rozdelenie liekov a zdravotníckych pomôcok do príbuzenských skupín podľa účelu, na ktorí slúžia

2. predikcia priemerných mesačných nákladov jednotlivých pacientov podľa demografických vlastností a údajov z ich vykázanej zdravotnej starostlivosti na jednotlivé skupiny liekov a pomôcok využitím lineárnej regresie a rozhodovacích stromov

3. predikcia priemerných mesačných nákladov už konkrétnych lekárov nasčítaním predikcií ich pacientov

4. porovnanie predikcie nákladov lekárov s realitou a následná analýza outlierov

1.3 Prehľad známych prístupov a metód zo zahraničia

Vďaka rozsiahlému rozvoju dátovej vedy (po anglicky známy pojem *Data Science*) sa tejto oblasti analýzy dát rôznych firiem začalo venovať mnoho odborníkov. Výnimkou neboli ani poisťovne s množtvom údajov o svojich poistencoch, u ktorých potreba analýz zozbieraných údajov a prínos z nich je nepochybne veľkým krokom k zefektívňovaniu nakladania s peňažnými prostriedkami a mala by viesť k vyššej ponúkanej kvalite služieb. Do tejto oblasti nepochybne patrí aj problematika odhaľovania podvodov a zneužívanie systému. Jeden z prehľadov doposiaľ známej literatúry k danej problematike je ponúknutý napríklad v článku [17], z ktorej budeme čerpať.

1.3.1 Prístup - "podozrivý" alebo "v poriadku"

Existuje veľa metód na odhaľovanie podozrivého správania. Jednou z často využívaných možností, ktorou nemožno opomenúť, je zatrieďovanie skúmaných objektov do kategórie "podozrivý" alebo "v poriadku". V týchto prípadoch je treba podotknúť, že algoritmy fungujú na základe učenia sa podľa vstupných údajov o danom objekte (v našom prípade lekár) rozpoznávať už známe formy podozrivého správania, ktoré sme už v minulosti potvrdili ako podozrivé. Pri takomto prístupe sa často používajú algoritmy spadajúce do oblasti strojového učenia ako kategorické rozhodovacie stromy alebo neurónové siete. Vzhľadom k tomu, že takto označenú vstupnú vzorku podvodníckych poskytovateľov nemáme k dispozícií, týmto smerom rozlišovania na dve kategórie (podozrivý alebo v poriadku) sa v našej práci nebudeme zaoberať.

1.3.2 Predikcia nákladov

Ďaľšou ponúkanou možnosťou nazerania na danú problematiku je cez optiku predikcie nákladov. V tomto prípade nie je nutnosťou mať vyššie spomínanú vstupnú vzorku s potvrdenými podozrivými subjektami a je to vhodný prístup na spravenie si preverenia tak povediac od základov. Tu však tiež možno dané prístupy rozdeliť na viaceré. Rozdiel je v inom výbere vstupných informácií. V jednom prípade ide o medicínske faktory, teda napríklad aké má chronické choroby atď. Takémuto prístupu sa venoval napríklad Lee a spol. v článku [12]. Ako prediktory zobral demografické informácie, vykázané diagnózy, počet laboratórnych testov a počet podstúpených operácií. Porovnával presnosť neurónových sietí a regresných rozhodovacích stromov, pričom presnejším sa ukázala neurónová sieť. Ďalej môžeme spomenúť Powers a spol. s článkom [18], kde bolo porovnaných viacero regresných modelov - klasická lineárna regresia, regresia s log-transformáciou, 3 dvojkrokové regresie a GLM (generalizovaný lineárny model). Z výsledkov vyšla najpriaznivejšie klasická lineárna regresia.

V druhom prípade sa používajú faktory spojené s nákladovosťou pacienta z minulých období, či nákladovosť v nejakej konkrétnej oblasti. Tento prístup má menšiu nevýhodu v prípade, ak napríklad lekár podvádza na konkrétnom pacientovi systematicky už dlhší čas. Tento typ podvádzania by model potom zachytával výrazne ťažšie. Ideálnou cestou by mohlo mať jeden model aj s predošlými nakladmi a druhý model bez nich a môcť porovnať získané poznatky.

Pri modeloch predikujúcich nákladovosť pacienta nemožno opomenúť medzinárne využívaný model pomenovaný ako *The Pharmacy-based Cost Group Model* (ďalej uz len PCG model), ktorý je na Slovensku implementovaný 6-ty rok. Bol vyvinutý na základe potreby spravodlivejšieho prerozdeľovania peňažných prostriedkov pre jednotlivé poisťovne na základe rizikovosti kmeňa pacientov danej poisťovne. Napríklad poisťovňa, ktorá slúži najmä dôchodcom má prirodzene vyššie priemerné náklady na pacienta ako poisťovna so zdravším kmeňom poistencov. Model sa snaží predikovať nákladovosť poistencov s vybranými chronickými chorobami liečenými určitou farmakoterapiou na základe vybraných a uhradených liekov. Každý pacient sa podľa istých kritérií (napríklad podľa množstva dávok uhradených typov liekov za nejaké stanovené obdobie atď.) priradí maximálne do jednej PCG skupiny odpovedajúcej jednej

danej chronickej chorobe, ktorou pravdepodobne trpí. Je prirodzené, že PCG model sa zaoberá skupinami najmä chronických chorôb, pretože chronicky chorí poistenec má všeobecne výrazne vyššie priemerné náklady ako bežný človek. Viac sa o ňom môžete dozvedieť napríklad v [7] alebo [1]. Celková nákladovosť pacienta sa následne ešte odhaduje aj na základe sociodemografických informácií, ako je pohlavie, vek, či ekonomická aktívnosť.

2 Matematické nástroje

V nasledujúcej kapitole si v úvodných častiach stručne objasníme motiváciu za výberom daného konceptu na predikciu nákladov a následne vysvetlíme matematickú teóriu a základné princípy, na ktorom pracuje. V poslednej časti uvádzame, ktorými metrikami na odhadnutie presnosti jednotlivých modelov budeme používať.

2.1 Lineárna regresia

2.1.1 Úvod

Lineárna regresia je dobre známy základný matematický koncept hľadania lineárnych vzťahov medzi *nezávislými* (alebo *vysvetľujúcimi*) premennými a *závislou* (alebo *vysvetľovanou*) premennou, ktorú máme za cieľ predikovať. Ako sme spomenuli v predošlej kapitole, v niektorých prípadoch môže lineárna regresia výjsť dokonca ako najvhodnejšia. Nám buďe slúžiť ako taký prvý nástrel, základný model a uvidíme, či ju ďalšie modely budú vedieť v presnosti poraziť. Teraz si tento koncept stručne priblížime čerpajúc z [22].

V matematickom zápise možno lineárny model definovať nasledovnou rovnicou:

$$Y = X\beta + \epsilon, \tag{1}$$

kde $X = [\vec{1}, X_1, X_2, ..., X_{k-1}]$ je matica vysvetľujúcich premenných, vektor $\beta = [\beta_0, \beta_1, ..., \beta_{k-1}]^T$ je neznámy vektor parametrov, ktoré popisujú spomínané lineárne závislosti medzi vysvetľujúcimi prememnnými a vysvetľovanou premennou Y a vektor ϵ predstavuje vektor *náhodných odchýlok* pozorovaní od strednej očakávanej hodnoty Y, inak nazývaný aj *chyby* alebo *rezídua* modelu.

2.1.2 Odhad metódou najmenších štvorcov

Pri odhade neznámeho vektora β sa využíva metóda najmenších štvorcov. Jej základným cieľom je minimalizovať súčet druhých mocnín rezíduí:

$$\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta} \epsilon^T \epsilon \quad (2)$$

$$= \min_{\beta} (Y - X\beta)^T (Y - X\beta) \quad (3)$$

$$= \min_{\beta} Y^T Y - 2Y^T X\beta + (X\beta)^T (X\beta) \quad (4)$$

Vďaka konvexnosti účelovej funkcie nám stačí na nájdenie riešenia tejto úlohy položiť deriváciu funkcie podľa β rovnú nule:

$$0 = -2Y^T X + 2X^T X\beta, \quad (5)$$

z čoho vyplýva vzorec pre explicitné riešenie na nájdenie odhadu optimálnych parametrov

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \quad (6)$$

2.1.3 Gauss - Markovove podmienky

Veta 2.1. *Za predpokladov splnenia podmienok*

- $E(\epsilon_i) = 0$ *pre* $\forall i = 1, ..., n$
- $var(\epsilon_i) = \sigma^2$ *pre* $\forall i = 1, ..., n$ *a jednotlivé chyby sú od seba nezávislé, teda* $cov(\epsilon_i, \epsilon_j) = 0$ *pre* $i \neq j$,

možno považovať odhad parametra β získaný metódou najmenších štvorcov za najlepší [1] lineárny nevychýlený [2] odhad, tzv. BLUE (Best Unbiased Linear Estimator) .

Veta 2.2. *Za predpokladov splnenia podmienok z vety 2.1 a pridaním podmienky na normalitu chýb, teda $\epsilon_i \sim N(0, \sigma^2)$ možno považovať odhad parametra β získaný metódou najmenších štvorcov za najlepší nevychýlený odhad, tzv. BUE (Best Unbiased Estimator)*

Dôkazy viet 2.1 a 2.2 možno nájst v [22]

[1] Pod najlepším sa myslí odhad parametra s najmenšou varianciou.

[2] Pod nevychýleným odhadom $\hat{\beta}$ sa rozumie platnosť podmienky $E(\hat{\beta}) = \beta$.

2.2 Rozhodovacie stromy

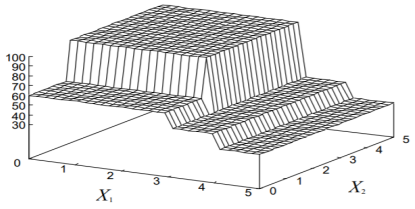
2.2.1 Úvod

Nepochybnou výhodou lineárnej regresie je jednoduchá interpretácia odhadnutých parametrov. Nie vždy je však ideálne aplikovať na dáta danú metódu. Problém môže nastať napríklad v prípade, že závislosti medzi vysvetľujúcimi a vysvetľovanou premennou nie sú lineárneho charakteru alebo ak je nejaká pravdepodobnosť, že v našich dátach budú signifikantné interakcie medzi jednotlivými premennými. Taktiež by v ideálnom prípade mali byť splnené podmienky pre chyby modelu.

Kedže sa v súčastnosti výrazným spôsobom rozmáha dátová veda a strojové učenie, ktorého súčasťou sú aj rozhodovacie stromy, radi by sme otestovali aj túto metódu predikcie a porovnali tento prístupu s klasickou lineárnou regresiou. Ich výhodou je napríklad čiastočné riešenie spomínaného problému interakcií premenných, vďaka spôsobu, akým sa rozhodovacie stromy konštruujú. Čo do ľahkej interpretovateľnosti, stavba modelu je jednoduchá na pochopenie. Problém však môže nastať pri komplexnom strome s veľkou hĺbkou. Taktiež treba optimalizovať jeho hyperparametre, aby nedochádzalo k tzv. pretrénovaniu (po angl. známe pod pojmom *overfitting*). O tom viac v ďalších častiach tejto podkapitoly.

2.2.2 Popis algoritmu

Pri tomto popise algoritmu sa inšpirujeme najmä [25]. Rozhodovacie stromy fungujú na princípe rozdelenia predikčného priestora $\mathbb{R}^k$, kde k je počet vysvetľujúcich premenných na disjunktné *regióny* $R_1, ..., R_J$, viď ilustračný obrázok 1, prevzatý z [25].



Obr. 1: Rozdelenia predikčného priestoru na regióny pre 2 vysvetľujúce premenné.

V prípade regresného problému sa následne všetkým pozorovaniam, ktoré padnú do

regiónu R_j, pridelí hodnota predikcie

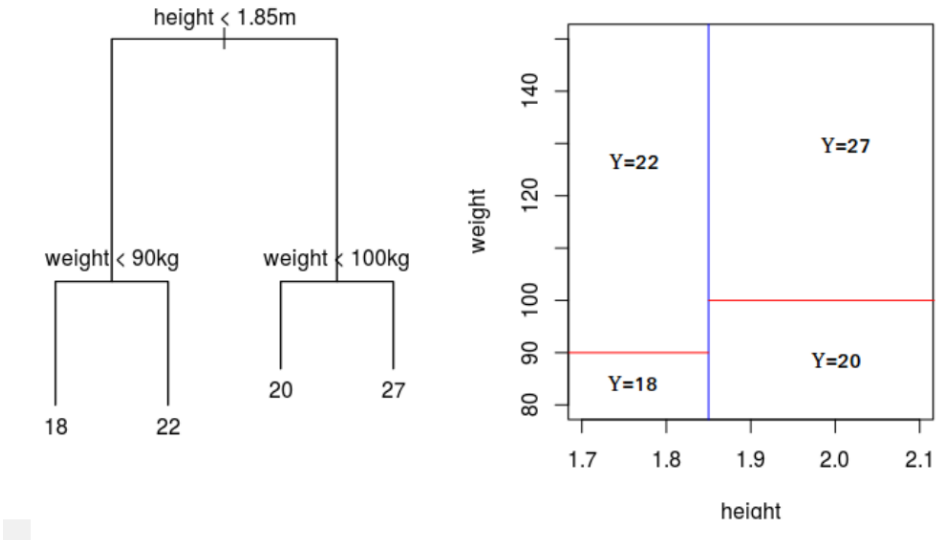
$$y_{Rj} = \frac{1}{n_j} \sum_{R_j} y_i, \tag{7}$$

kde n_j predstavuje počet pozorovaní v danom regióne R_j. Predikovaná hodnota bude teda priemerom predikčných hodnôt pozorovaní, ktoré patria regiónu R_j.

Snahou je nájsť také rozdelenie predikčného, ktoré bude (podobne ako v lineárnej regresií) minimalizovať priemer druhých mocnín chýb:

$$min \quad \frac{1}{n} \sum_{j=1}^{J} \sum_{R_j} (y_i - \hat{y}_{R_j})^2 \tag{8}$$

Prirodzene je nemožné prehľadať všetky možné spôsoby rozdelenia tohto predikčného priestoru. Myšlienka, ako sa dopátrať aspoň k približnému riešeniu, spočíva v *rekurzívnom binárnom delení* (angl. *recursive binary splitting*). Tu sa dostávame k reprezentácií algortimu pomocou rozhodovacieho stromu. Začíname vo vrchole stromu a postupne v každom uzle na základe testu a odpovede "áno" alebo "nie" posúvame dané pozorovonanie do ďalšieho rozhodovacieho (alebo koncového) uzla. Pre lepšiu predstavu uvádzame jednoduchý príklad rozhodovacieho stromu ilustrovaný na obrázku 2 vľavo a rozdelenie na regióny v pravo, prebratý z [10].



Obr. 2: Príklad rozhodovacieho stromu s dvoma vysvetľujúcimi premennými výška a váha

Pre ďalšie účely treba zadefinovať *chybovú funkciu* pre daný uzol j stromu (resp. regiónu R_j) a pre pozorovania (označme $(x, y)_i$, kde $i = 1, ..., n$) do neho patriace ako

$$Err(j) = \frac{1}{n_j} \sum_{R_j} (y_i - \hat{y}_{R_j})^2. \tag{9}$$

V každom uzle (ak sa nejedná o koncový) sa aplikuje test a následné rozdelenie (angl. *split*) v tvare:

$$\text{ak} \quad x_l < s: \quad (x, y)_i \in R_j^{left}(s, l) \qquad \text{inak} \quad (x, y)_i \in R_j^{right}(s, l) \tag{10}$$

teda hľadá sa vhodná voľba premennej X_l, ktorá sa bude porovnávať s *thresholdom* s. Definujme chybovú funkciu pre takéto rozdelenie určené hodnotami s, l v uzle j ako

$$Err(j, s, l) = \frac{n_j^{left}}{n_j} Err(j^{left}) + \frac{n_j^{right}}{n_j} Err(j^{right}), \tag{11}$$

kde hodnota n_j^{left} znamená počet pozorovaní, ktoré sa dostanú do ľavého uzla j^{left}, hodnota n_j^{right} obdobne. Jedná sa teda o vážený priemer priemerných kvadratických chýb v novovzniknutých uzloch. Deliacim kritériom pre vhodné zvolenie hodnôt s a l je prirodzene dosiahnúť čo najnižšiu hodnotu $Err(j, s, l)$, resp.

$$max \quad \triangle Err(j, s, l) := Err(j) - Err(j, s, l), \tag{12}$$

čiže maximalizovať pokles priemeru kvadratických chýb po rozdelení daného uhla. Dá sa to definovať ako *nárast infomatívnosti* (alebo angl. *infomation gain*). Takto definované deliace kritérium a spôsob stavania rozhodovacieho stromu popísaný v tejto časti je zavedený konkrétne v tzv. CART algoritmoch (Classification And Regression Trees) od L. Breiman et al. [2]. Existujú aj iné typy algoritmov ako napríklad ID3 alebo C4.5, ktoré však riešia len klasifikačné úlohy a preto nie sú predmetom záujmu v našej práci.

2.2.3 Pruning

Pre CART algoritmy je zvykom nechať skonštruovať strom do maximálnej veľkosti, až kým sa vrámci všetkých koncových uzloch nachádzajú pozorovania iba s rovnakou

hodnotou predikčnej premennej, alebo s rovnakými vysvetľujúcimi premennými, ktoré už nemožno ďalej deliť. Problémom však je, že pri rozhodovacích stromoch dochádza k *overfittingu*, čo znamená, že strom sa na *trénovacích dátach* (teda dátach, na ktorých sa konštruoval) javí ako veľmi presný. Keď však prídu nové pozorovania, ktoré ešte "nevidel", dosahuje výrazne horšie výsledky. Je to spôsobené tým, že rozhodovací strom sa až príliš podrobne snaží zachytiť všetky informácie z trénovacích dát, v ktorých sa však nachádza aj šum, resp. outliery.

Aby sa tomuto javu zabránilo, často sa používa metóda *pruning* spojená s tzv. *krížovou validáciou*. Náhodne sa rozdelí dátová sada na k rovnako veľkých podmnožín. Spravidla ich to býva do 10. Následne sa určí na akých leveloch hĺbok stromu sa bude presnosť testovať. Na $k-1$ podmnožinách sady sa strom natrénuje a na zostávajúcej sa vypočíta celková chybovosť stromu (podľa predpisu (8)) pre každý level hĺbky. Toto postupne zopakujeme pre všetky podmnožiny sád a získané chybovosti stromov pre jednotlivé levely spriemerujeme. Vyberáme strom s takou hĺbkou, pre ktorú bol tento ukazovateľ najmenší.

Ako vidíme, ide tu o istý typ optimalizácie *hyperparametra*. Hyperparameter je parameter, ktorý zadávame do modelu ako vstup. Najprv však nepoznáme jeho optimálnu hodnotu a pre každú dátovú sadu môže byť rôzny. Okrem hĺbky stromu môže ísť napríklad o:

- limit na minimálny počet pozorovaní v uzle, aby mohol byť pokus o jeho ďalšie delenie
- limit na minimálny počet pozororvaní v koncovom uzle
- minimálny požadovaný nárast informatívnosti po rozdelení uzla, po anglicky nazývaný aj ako *complexity paramater*

2.2.4 Náhodný les

Ako bolo spomenuté v predošlej časti 2.2.3, jav overfittingu býva veľkou výzvou pri používaní rozhodovacích stromov. Ďalšou možnosťou, ako sa pokúsiť danému javu vyhnúť, je vybudovať väčšie množstvo stromov a za výslednú predikciu považovať priemer z týchto predikcií.

Vyvstáva otázka, ako tieto stromy konštruovať, lebo z rovnakých dát za normálnych okolností vznikne vždy rovnaký rozhodovací strom. Ideou je najprv z dátovej sady vyrobiť pre nový strom novú dátovú sadu pomocou bootstrapovej metódy s opakovaním, teda náhodným ťahaním pozorovaní z existujúcej vzorky, pričom jednotlivé pozorovania sa môžu opakovať. Každý strom je tak natrénovaný na inej, trochu obmenenej vzorke.

Je tu aj ďalšia modifikácia pri výstavbe stromov. V každom rozhodovacom uzle pri hľadaní optimálnej premennej X_l, podľa ktorej sa ďalej testuje a rozdeľuje v danom uzle, sa berie do úvahy len podmnožina zo všetkých vysvetľujúcich premenných. Poväčšinou sa berie druhá odmocnicna z celkového počtu týchto premenných.

Náhodný les sa ukazuje vo veľa prípadoch ako úspešný bojovník za zlepšenie generalizácie modelu a je medzi dátovými analytikmi rozšireným a obľúbeným nástrojom, preto sme sa ho rozhodli vyskúšať a otestovať aj my.

2.3 Metriky na odhadnutie presnosti modelov

Existuje viacero metrík, ktorých cieľom je nejakým spôsobom určiť kvalitu modelu, teda presnosť s akou model správne predpovedá predikovanú prememnú. My sme sa za týmto účelom a pre porovnanie viacerých modelov navzájom medzi sebou rozhodli v práci využívať najmä ***koeficient determinácie*** $\mathbf{R}^2$, ktorý sa jeden z najznámejších a najčastejšie využívaných pri regresných problémoch. Počíta sa vzorcom (13), ktorý dáva do pomeru dve veličiny:

$$R^2 = \frac{ESS}{TSS}. \tag{13}$$

Prvou z nich je tzv. *vysvetlený súčet štvorcov* (po angl. *explanation sum of squares*), čo možno chápať ako kolísavosť predikovaných hodnôt od ich priemeru, teda kolísavosť $\hat{Y}$, ktorá je zachytiteľná modelom a počíta sa ako

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}})^2. \tag{14}$$

Druhou veličinou je tzv. *totálny súčet štvorcov* (po angl. *total sum of squares*), ktorý predstavuje kolísavosť vysvetľovanej premennej od jeho priemeru a počíta sa ako

$$TSS = \sum_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}})^2. \tag{15}$$

Často sa koeficient determinácie počíta aj ako

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{16}$$

kde RSS je tzv. *nevysvetlený súčet štvorcov* (po angl. *rezidual sum of squares*), ktorý reflektuje istú mieru nekvality modelu a definujeme ho ako

$$RSS = \sum_{i=1}^{n} \hat{\epsilon_i}^2. \tag{17}$$

My budeme v práci využívať túto modifikáciu (16). Treba dodať, že čím je koeficient determinácie vyšší, tým by sme mali mať kvalitnejší model.

Ďalej sme si vybrali dve doplňujúce metriky, obe sú ukazovatele miery nekvality modelu, preto sa ich snažíme mať čo najmnešie (blízke nule). Ide o ***priemernú absolútnu odchýlku***, označovanú ako **MAE** (po angl. *mean absolute error*), definovanú ako

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{\epsilon_i}|, \tag{18}$$

ktorá má intuitívnu interpretáciu a síce o koľko sa v priemere model v predikcií "mýli" oproti skutočnej hodnote. Druhou veľmi podobnou metrikou je ***odmocnená priemerná kvadratická odchýlka***, označovaná ako **RMSE** (po angl. *root mean squared error*) počítanú ako

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon_i}^2} \tag{19}$$

ktorej výhodou oproti MAE je vyššia citlivosť na výrazné rezíduá v modeli, teda RSME s novým outlierom porastie výraznejšie ako MSE.

2.4 K-means zhluková analýza

Pri tejto metóde zhlukovej analýzy je cieľom zatriediť n pozorovaní do k klastrov podľa mieri ich vzájomnej podobnosti, resp. vzdialenosti. V tejto metóde sa definuje vydialenosť pozorovaní x a y ako:

$$\rho(x, y) = \sqrt{\sum_{t=1}^{p}(x_t - y_t)^2}, \tag{20}$$

kde p znači počet čŕt, podľa ktorých chceme pozorovania klastrovať. Cieľom je minimalizovať funkciu

$$\sum_{i=1}^{k}\sum_{r\in C_i}\rho^2(x_r, c_i), \tag{21}$$

,

kde c_i nazývaný aj centroid predstavuje stred medzi jednotlivými pozorovaniami priradenými klastru i, teda je určený vzťahom

$$c_i = \frac{1}{|C_i|}\sum_{r\in C_i} x_r. \tag{22}$$

,

V praxi Lloydov algoritmus funguje nasledovne:

- náhodne vytvor vstupné klastre $C_1, C_2, ..., C_k$
- pokiaľ sa neprekročí maximálny počet iterácií alebo nedojde k tomu, že sa už viac klastre nemenia
  - vypočitaj centroidy $c_1, c_2, ..., c_k$
  - pre každé $i = 1, ..., k$ priraď klastru C_i pozorovania, ktoré su príslušnému centroidu c_i najbližšie

Pri výbere počtu klastrov možno využiť "lakťové" pravidlo a pozerať, pri ktorej hodnote k sa hodnota funkcie (21) oproti hodnote tejto funkcie v $k-1$ ešte výrazne znížila. Tento popis K-means metódy sme prebrali z [8].

3 Zatriedenie liekov a zdravotníckych pomôcok

Vďaka rýchlo rozmáhajúcim sa farmaceutickým spoločnostiam existuje nespočetné množstvo liekov. Aj v televízií môžeme každú chvíľu sledovať reklamy propagujúce najrozličnejšie farmaceutické produkty. Lieky so určené na rôzne typy problémov a za každým užitím lieku stojí vlastný príbeh daného pacienta. V tejto časti práce si stručne vysvetlíme etapu zadeľovania liekov a zdravotníckych pomôcok, ktorých nákladovosť sa snažíme predikovať, do príbuzenských skupín.

3.1 Zatriedenie liekov

Našou prirodzenou snahou je zoskupovať lieky určené na podobný problém. Ako sme už spomínali, za každým užitým liekom stojí konkrétny príbeh daného človeka. Prirodzene predpokladáme, že jednotlivé dôvody predpísania a užitia liekov na podobné problémy sú medzi pacientami podobné a teda existujú pre pridružené choroby spoločné rizikové faktory. Našim cieľom je predikcia priemerných mesačných nákladov a teda pre rôzne skupiny chorôb (a teda následnú preskripciu príslušných liekov) budú iné faktory dôležité. Napríklad prirodzene očakávame, že pri preskripcií B-pomôcok (plienky) bude vysoký vek pacienta odohrávať dôležitú rolu, čo však pri takej astme nemusí úplne platiť.

Pri triedení liekov hrajú dôležité úlohy tzv. *ATC skupiny*. Každý liek je podľa svojej hlavnej účinnej látky priradený práve do jednej ATC skupiny. Každá ATC skupina má svoj kód. Kód začína písmenom spravidla určujúcim konkrétnu orgánovu sústavu ľudského tela, ktorej sa účinné látky v danej skupine týkajú. Každý ďalší znak danú skupinu spresňuje. Najdlhší a teda najpresnejší kód ATC skupiny je spravidla 7-miestny.

Na konzultáciach s medicínsky vzdeľanými odborníkmi nám bolo odporučené sa inšpirovať rozdelením liekov, ktoré sa využíva pri spomínanom PCG modeli. My sme sa týmto zatriedením liekov podľa ATC skupín inšpirovali, no zároveň ho modifikovali podľa vlastných potrieb. Využívali sme len 3-mieste ATC kódy, aby sme sa vyhli prílišnej detailizácií a zoskupili väčšie množstvo typov liekov spolu.

3.2 Zatriedenie zdravotníckych pomôcok

Zdravotnícke pomôcky budú tvoriť samostatné skupiny. Každá pomôcka je zaradená do jednej z 10 hlavných kategórií. Určuje ho prvé písmeno kódu danej pomôcky. Túto kategorizáciu možno nájsť v [16].

3.3 Výsledné skupiny

Ideálnymi skupinami zatriedených liekov a pomôcok a následné skúmanie ich nákladovosti sú tie skupiny, z ktorých je vykázaných signifikantne veľký počet produktov. Teda vyberieme si tie skupiny, ktoré majú za rok 2017 veľké množstvo výkazov na pomerne veľký počet pacientov a ideálne pokrývajú čo najväčší objem peňažných prostriedkov, ktoré boli za ne lekárom vyplatené. Vidíme v tom nasledovné výhody:

- Pri vyššom počte výkazov, čiže vyššom množstve údajov bude pravdepodobne možný presnejší odhad - je tu väčšia pravdepodobnosť nájdenia nejakých spoločných ukazovateľov, na základe ktorých možno odhadnúť, či by mal mať daný pacient predpísaný daný liek/pomôcku

- V týchto skupinách produktov s vysokou úrovňou preskripcie taktiež predpokladáme, že môže dochádzať k častým podvodom. Lekár si môže povedať, že pri týchto množstvách si zopár fitkívne vykázaných liekov alebo pomôcok nik nevšimne. Z tohto dôvodu sa môžu javiť ako dobrý kandidáti na hlbšie preverenie.

- V neposlednom rade je nepochybnou výhodou najmä z pohľadu poisťovne, ak je na danú skupinu naviazaný veľký objem peňazí, pretože sa tak prekontroluje správnosť a korektnosť využitia veľkého objemu peňažných prostriedkov, ktoré poisťovňa vynaložila na preplatenie daných liekov a pomôcok.

Náš výber najpočetnejších skupín je zobrazený v tabuľke 1. Druhý a štvrtý stĺpec sú percentuálne podiely zo všetkých nákladov preplatených, resp. produktov predpísaných všeobecnými lekármi za rok 2017. Ako môžeme vidieť, až približne 67%, čo sú dve tretiny všetkých nákladov na preskripciu liekov a pomôcok spadá do týchto siedmych skupín.

Názov skupiny	**Krytie nákladov (%)**	**Počet pacientov**	**Pomer produktov (%)**
Hypertenzia	25,91%	286702	35,00%
Antitrombotiká	16,46%	122817	6,82%
B-pomôcky	10,09%	21236	2,64%
Hypercholesterolémia	4,54%	99926	4,72%
Srdcové choroby	3,55%	39478	3,29%
Diabetes	3,26%	25346	1,38%
Astma	3,12%	34452	1,74%
Spolu	66,93%	940436	55,59%

Tabuľka 1: Zatriedenie liekov a zdravotníckych pomôcok do sledovaných skupín

3.4 Skupiny v tejto práci

Pre účely tejto práce sme sa rozhodli zamerať našu pozornosť najmä na tieto dve:

- **Hypertenzia** (skrátene ďalej **HYP**), reps. lieky proti nej, pretože je to jednoznačne najbežnejšia skupina liekov, ktorá tvorí dokonca o čosi viac ako jednu tretinu zo všetkých predpisaných liekov a pomôcok za daný rok. Po konzultáciach s lekárom do tejto kategórie berieme všetky lieky s ATC skupinami: C02, C03, C07, C08 a C09.

- **B-pomôcky** (skrátene ďalej **B**), čo zahŕňa vložky, plienky, urinárne vrecká a pod. V minulosti sa potvrdilo niekoľko podvodov vrámci tejto oblasti a dostávajú sa do poisťovne informácie z viacerých strán, že sa predpisujú pacientom, ktorí ich nepotrebujú a dostanú ich alebo ani nedostanú. Kuriozitou bol prípad, kedy lekár predpisoval pacientom drahé plienky, lekáreň im následne za poukazy vydala lacnejšiu formu, ktorú potom pacienti šli predávať pred Tesco.

4 Výber dát a regresných premenných

V tejto kapitole si priblížime výber dátovej sady, ktorú sme sa rozhodli v práci analyzovať. Následne predstavíme regresné premenné, ktoré budú vstupovať do modelov a uvedieme ich základné štatistiky a grafy pre lepšiu predstavu.

4.1 Dáta

Predmetom nášeho záujmu budú dospelí poistenci, konkrétne tí, ktorí k dátumu 31.12.2017 dovŕšili minimálne 18 rokov. Zároveň sa do dátovej sady dostali len tí, o ktorých sme mali k dispozícií informáciu, u ktorého všeobecného lekára sú kapitovaní. Opať sme zobrali informáciu aktuálnu ku poslednému dňu daného roka. Z dát sme ďalej vylúčili poistencov s bydliskom v zahraničí. Takýmto spôsobom sa do našich dát dostalo niečo menej ako milión poistencov, konkrétne 940436.

4.2 Regresné premenné

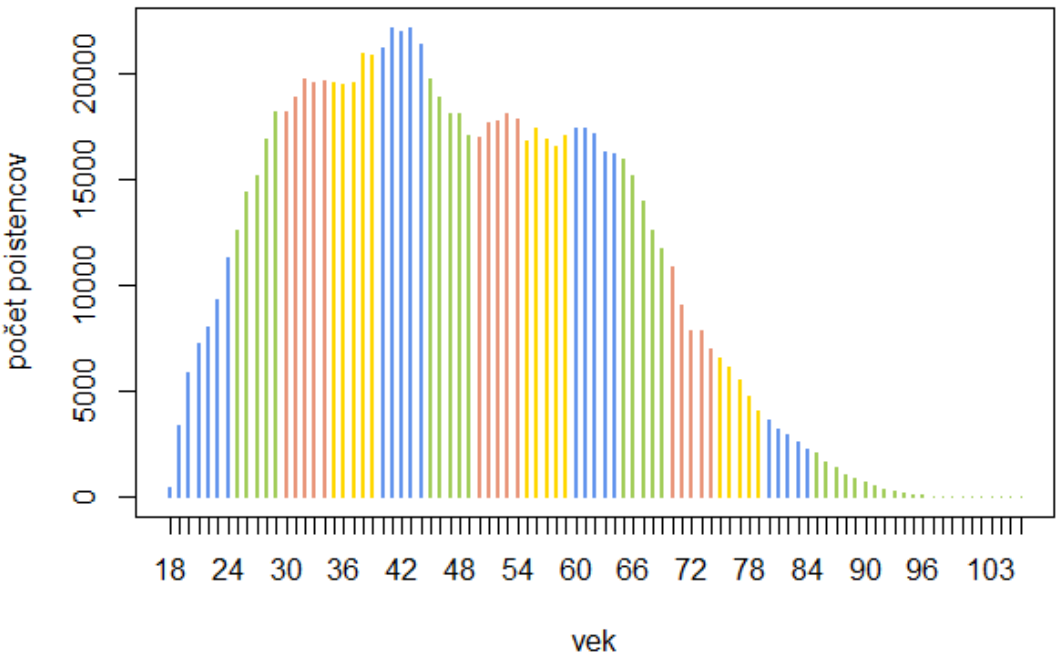
Vysvetľujúce premenné by sme mohli rozdeliť na dve časti. Prvá časť bude pre obe sledované skupiny liekov resp. pomôcok rovnaká. Ide o demografické informácie pacientov. V druhej časti sa pridajú premenné vystupujúce z údajov z ich VZS vykázanej mimo všeobecných lekárov. Pre jednotlivé skupiny liekov a pomôcok sa budú prirodzene trochu líšiť, pretože v každej skupine budú pravdepodobne dôležité informácie z iných odborností. V tretej časti si bližsie predstavíme premenné, ktoré chceme predikovať.

4.2.1 Demografické premenné

V podobných modeloch najmä v oblasti zdravotníctva je bežnou praxou, až nevyhnutnosťou brať do úvahy vek a pohlavie pacienta. Ďalej sme sa inšpirovali PCG modelom a podobne ako v ňom budeme prihliadať aj na informáciu o ekonomickej aktívnosti pacienta. Je na zváženie, či by bolo vhodné pri tvorbe modelu vziať do úvahy aj okres bydliska poistenca (o tom viac neskôr). Ako poslendú sme pridali dummy premennú, či daný pacient v nasledujúcom roku 2018 zomrel. Teraz si tieto prediktory priblížime detailnejšie:

- **Vek**

Vzhľadom k tomu, že vek s nákladmi na zdravotnú starostlivosť častokrát nevykazuje lineárnu závislosť, ale skôr polynomiálny až exponenciálny rast, voľbou dummy premenných namiesto jednej spojitej premennej sa tohto problému jednoducho a elegantne zbavíme, pretože parametrom pri dummy premenných odpovedajúcich vyšším vekom sa môžu odhadnúť výrazne vyššie hodnoty. Ďalej sme sa kvôli veľkému počtu možných hodnôt, ktoré nadobúdajú veky pacientov [3] (od 18 do 106 rokov) rozhodli pre zlúčenie viacerých do jednej kategórie. Spôsob zlúčenia je farebne ilustrovaný na obrázku 3. Pripomíname, že v našich dátach figurujú len dospelí poistenci. Začíname kategóriou 18 až 24, ďalej vždy po piatich rokoch. Posledná kategória je 85+.



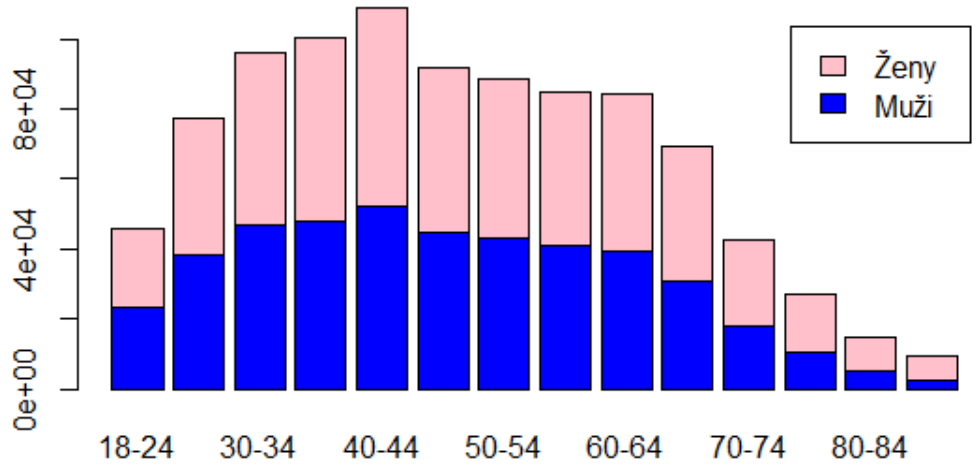
Obr. 3

Obr. 4: Histogram pre vek poistencov

[3]Pre konzistentnosť sme sa rozhodli zobrať vek, ktorí nadobúdali poistneci na konci roku 2017.

- **Pohlavie**

Náš datový súbor je pomerne vyrovnaní. Figuruje v ňom približne $47,12\%$ mužov a $52,88\%$ žien. Na obrázku 5 môžeme vidieť, ako sa s rastúcim vekom zväčšuje pomer žien k mužom, čo zodpovedá známemu faktu, že sa ženy v priemere dožívajú dlhšieho veku ako muži.



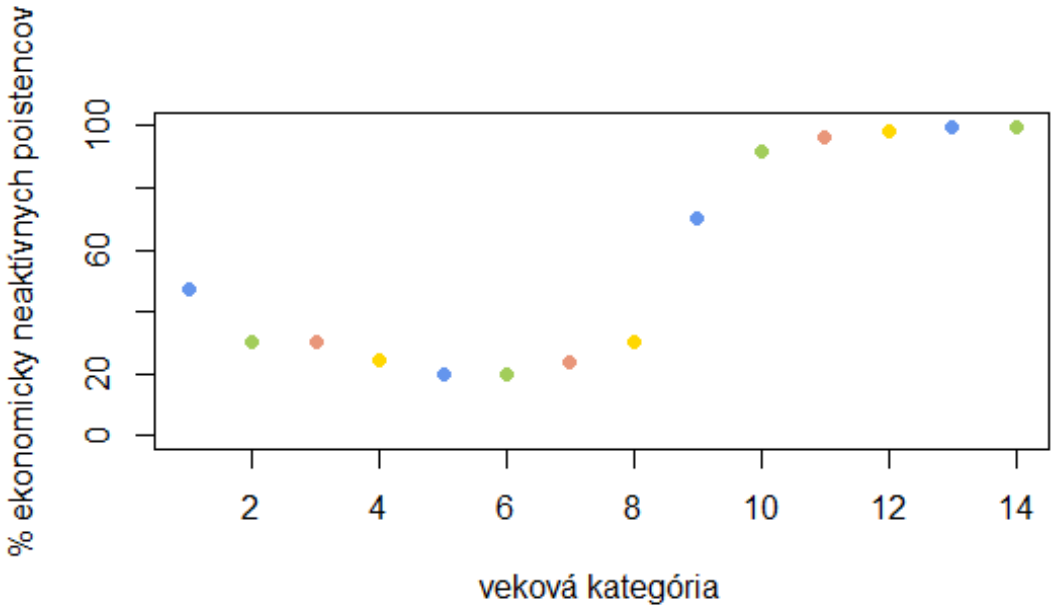
Obr. 5

Obr. 6: Porovnanie početností mužov a žien v jednotlivých vekových kategóriach

- **Ekonomická aktívnosť**

Ide tu o informáciu, či je daný poistenec zárobkovo činný, alebo zaňho platí poistné štát (napr. študenti, dôchodcovia, či ženy na materskej). Táto premenná by mohla byť do istej mieri korelovaná s mozgovou aktivitou, ktorú jednotliví poistenci vykonávajú počas pracovných dní a to sa môže v niektorých prípadoch (najmä väčšej nečinnosti) odzrkadlovať na ich mentálnom zdraví. To môže ovplyvniť výskyt istých chorôb (napríklad parkinsovonova choroba, či alzheimer) a následné prepojenie s nárastom nákladov napríklad na B-pomôcky. Taktiež sa dá z opačného pohľadu povedať, že chorľavý človek sa vzhľadom na svoj zdravotný stav s väčšou pravdepodobnosťou stane

poistencom štátu. Na obrázku 7 ilustrujeme percentuálne rozloženie ekonomicky neaktívnych ľudí v konkrétnych vekových kategóriach. Prirodzene s vyšším vekom tento podiel narastá.



Obr. 7: Podiel ekonomicky neaktívnych poistnecov pre vekové kategórie v %

- **Okres**

Ďalšou vstupnou informáciou, ktorou sa v demografickej časti budeme zaoberať, je zaradenie dummy premennných pre okres, v ktorom poistenec býva [4]. Tu prichádza k istému rozporu. Ako sa hovorí, iný kraj, iný mrav. Na jednej strane môže byť táto informácia dobrá na modelovanie "štruktúry lekárov" v danom okrese - môže nastať situácia, že v okrese chýba veľa lekárov určitej kľúčovej odbornosti pre naše skupiny liekov a pomôcok a preto všeobecný lekár "preberá" úlohu tohto odborníka a na jeho odporúčanie predpisuje pacientom dané produkty sám a teda jeho náklady oproti iným okresom rastú. Na druhej strane sa môže tiež napríklad stať, že v okrese s celkovo malým množstvom všeobecných lekárov, kde sa lekári navzájom poznajú a komunikujú medzi sebou, sú dohodnutí na nejakom type podvodu, ktorý väčšina z daného okresu

[4]Tu máme na mysli jeho korešpondenčnú adresu. Ak nie je uvedená, tak sa myslí prechodné bydlisko. Ak nie je uvedené ani to, berie sa trvalý pobyt poistenca.

praktizuje. V takomto prípade by sme informácie o okrese zahrnúť do modelu nechceli, pretože nechceme tento typ správania v modeli vysvetliť (priradením vyššej hodnoty pre odhad parametra odpovedajúcemu danému okresu). Chceli by sme ho v ideálnom prípade detekovať tak, že sa títo lekári vyskytnú medzi outliermi a teda sa ukáže, že sa ich správanie vymyká normám. Ďalším problémom je veľký počet premenných, konkrétne 71 okresov[5], čo znamená 71 dummy premenných, ktoré by takto do modelu pribudli. K tejto dileme sa ešte vrátime neskôr v v podkapitole 5.1.

- **Úmrtie**

Hovorí sa, že 90% nákladov na zdravotnú starostlivosť človek vynaloží posledný rok pred svojou smrťou. Táto jednoduchá myšlienka bola pre nás zaujímavá a hodná preskúmania, tak sme sa rozhodli ju zahrnúť do nášeho modelu. Pribudla tak nová premenná úmrtie, ktorá hovorí, či daný pacient v nasledujúcom roku 2018 zomrel. V našej dátovej sade sa to konkrétne týkalo 10110 poistencov.

4.2.2 Premenné z VZS

Samotná tabuľka obsahujúca vykázanú zdravotnú starostlivosť len za rok 2017 a len na tomto výbere poistencov má dokopy vyše 60 miliónov riadkov. Vyťahovanie informácií z nej nebolo jednoduché, preto sú aj dané počty jednotlivých premenných mierne obmedzené. rozhodli sme sa pre dva typy informatívnych premenných.

- **Typy VZS**

Prvým typom premenných budú dummy premenné, ktoré hovoria, či daný poistenec čerpal (aspoň raz) v roku 2017 daný typ zdravotnej starostlivosti. Do typov ZS patrí napríklad všeobecná ambulantná starostlivosť, špecializovaná zdravotná starostlivosť, magnetická rezonancia, záchranná zdravotná služba, lekáreňská starostlivosť, očná optika a iné. Keďže však špecializovaná zdravotná starostlivosť (teda špecialisti) sú obšírnou skupinou, tento typ ZS sme rozdelili aj podľa daných odborností, ktorých sa táto starostlivosť týkala. Následne sme dostali 120 našimi poistencami čerpaných

[5]Všetky bratislavské okresy sme zlúčili do jedného. Podobne sme urobili aj s košickými.

typov ZS a pri špecializovanej starostlivosti aj odborností (ďalej už len typov ZS) za daný rok.

Vzhľadom k veľkým rozmerom tabuľky VZS a teda jej náročnému prehľadávaniu sme sa rozhodli urobiť predselekciu, ktoré premenné z takto vzniknutých skupín z typov ZS a odborností by mohli mať výrazný informatívny charakter pre naše predikcie nákladov. Urobili sme menšiu analýzu, ktorá vyzerala nasledovne:

Porovnávali sme vždy dve skupiny, v jednej boli poistenci, ktorí mali kladné náklady na sledovanú skupinu (HYP, resp. B), v druhej boli všetci poistenci. Pre každý typ ZS sme zistili, koľkí ho v danom roku čerpali z prvej skupiny a koľkí ho čerpali celkovo. Ideálnymi prediktormi by prirodzene boli tie, u ktorých pomer týchto dvoch počtov bol čo najväčší. Tu treba mať na mysli, aký je pomer celkovo týchto dvoch skupín, teda pomer počtu poistencov s kladnými nákladmi na sledovanú skupinu ku všetkým poistencom v našom datasete. Tento pomer je napríklad pre skupinu B približne niečo nad 2%. To znamená, že za predpokladu, že daný typ ZS nejakým spôsobom nesúvisí s nákladmi na B-pomôcky, očakávali by sme, že tento pomer pre daný typ ZS ostane približne 2%. Ak sa však tento ukazovateľ zvýši na napríklad 20%, znamenalo by to, že až 10-krát častejšie sa v tejto vzorke ľudí využívajúcich tento typ ZS nachádzajú poistenci s nákladmi na B-pomôcky.

Ďalším faktorom bolo, aby boli vybrané dummy premenné dostatočne informatívne. Z tohto dôvodu sme si určili hranicu, pod akú početnosť poistencov, ktorých sa daný typ ZS týkal, nejdeme. Po následných konzultáciach s odborníkom sme vybrali 7 prediktorov pre skupinu B.

Obdobnú analýzu sme spravili aj pre skupinu HYP. Vzhľadom k tomu, že hypertenzia (alebo vysoký krvný tlak) je často vyskytujúcim sa civilizačným ochorením, v tejto skupine je pomer pacientov s predpísanými liekmi na hypertenziu voči všetkým približne až 30%. Kvôli danému faktu máme na percentuálny ukazovateľ pomeru počtov v danom type ZS výrazne vyššie postavené nároky. Vybrané prediktory v oboch skupinách uvádzame v prílohe A.

- **Diagnózy**

V druhej časti premenných vychádzajúcich z informácie z VZS, budú také, ktoré

opisujú, aké mali naši pacienti za rok 2017 vykázané diagnózy. Ku každému výkonu, ktorý PZS poistovni vykáže sa uvádza kvôli akej diagnóze bol vykonaný. Opäť pojde o dummy premenné - mal alebo nemal. Tu však budeme brať do úvahy diagnózy iba tých výkonov, ktoré neboli vykázané všeobecnými lekármi, ktorých správanie sa snažíme preveriť. Každá diagnóza má pridelený svoj kód. Využívame kategorizáciu chorôb z MKCH-10- SK-2016 [14]. Podobne ako pri ATC kóde, dlhší kód je detailnejší. V tomto prípade štvor až päťznakové kódy diagnóz sú najdlhšie. My sme brali do úvahy trojznakové kvôli väčšej generalizácií. Pri výbere diagnóz sme postupovali podobným spôsobom ako pri typoch ZS a po konzultáciach s odborníkom s lekárskym vzdelaním uvádzame výsledné diagnózy v tabuľkách **??** a **??** v prílohe B.

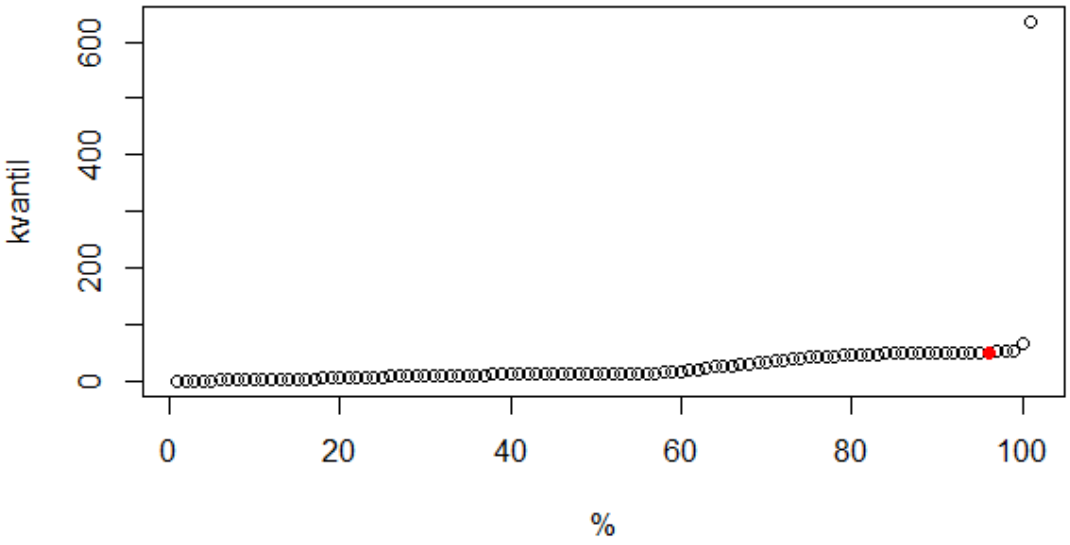
4.2.3 Vysvetľované premenné

Našim cieľom je odhadovať priemerné mesačné náklady na B-pomôcky (resp. na lieky proti hypertenzií) na daného pacienta. Hodnoty sú uvádzané v eurách.

Na B-pomôcky prispela poisťovna v sledovanom roku približne sumou 6 miliónov eur. V priemere je to mesačne okolo 500 000 eur. Približne 20 000 poistencov malo tieto pomôcky predpísané, teda ak mal poistenec predpísané B-pomôcky, mal v priemere priemerné mesačné náklady okolo 25 eur. Maximálne priemerné mesačné náklady istého poistenca boli však v hodnote 635,74 eur. Ako aj z obrázku 8 možno spozorovať, počet poistencov s výrazne vyššími nákladmi je zanedbateľný a ide skôr o akési anomálie.

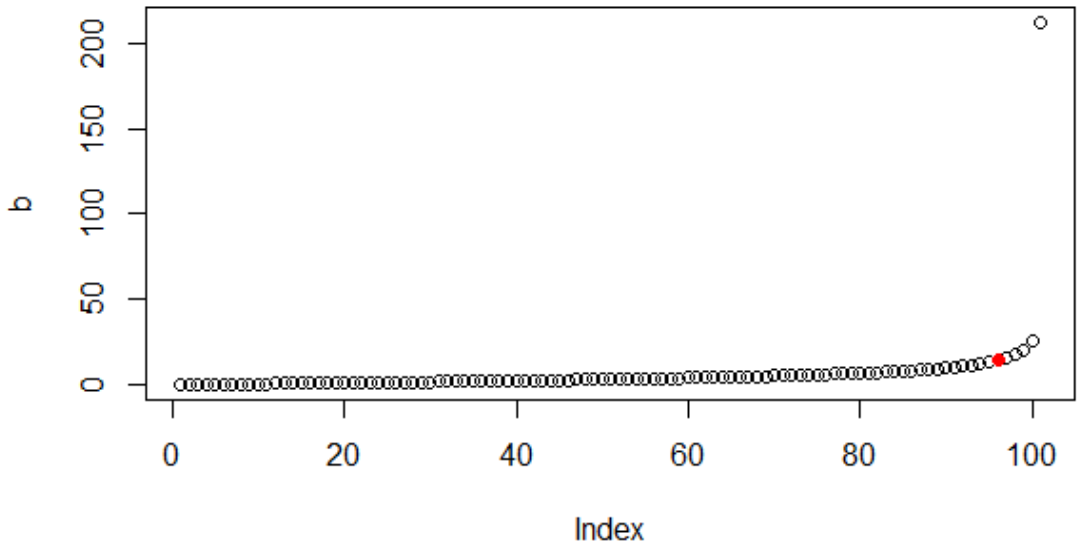
S veľkou pravdepodobnosťou sa však náklady týchto zopár jedincov v modeli neodhadnú správne a príslušným lekárom by to v následnej analýze výrazne pohoršilo. V nákladoch na zdravotnú starostlivosť sa však v ojedinelých prípadoch stáva, že má pacient v danej oblasti výrazne vyššie náklady a neradi by sme príslušných lekárov za jedného takéhoto pacienta tak výrazne penalizovali. Z tohto dôvodu sme sa rozhodli všetkým hodnotám nad 95%-ný kvantil, priradiť 95%-ný kvantilovú hodnotu, čo je 51,74 eur. Na obrázku 8 je táto hodnota zvýraznená červenou farbou.

Na lieky proti hypertenzií prispela poisťovna v sledovanom roku približne 16 miliónami eur, čo je mesačne okolo 1,3 miliónu. Je známe, že toto ochorenie postihuje veľkú časť populácie, čomu zodpovedá aj fakt, že tieto lieky malo aspoň raz predpísaných takmer



Obr. 8: Kvantilová funkcia pre náklady na B-pomôcky

300 000, teda skoro jedna tretina sledovaných poistencov. Takýto poistenec mal v priemere priemerné mesačné náklady okolo 4 eur. Opäť sa však stretávame s výraznými outliermi, viď obrázok 9. Z rovnakého dôvodu sme obdobne upravili aj tieto náklady. Tentokrát hodnota 95%-ného kvantilu dosahovala úroveň 14,30 eur.



Obr. 9: Kvantilová funkcia pre náklady na lieky proti hypertenzií

5 Diagnostika lineárneho modelu

5.1 Budovanie modelu

V tejto časti si pre motiváciu a lepší náhľad, ako vybranné premenné pomohli zlepšovať presnosť modelu, uvedieme tabuľku s koeficientami determinácie R^2 z viacerých modelov. Pre tento účel zatiaľ využívame metódu najmenších štvorcov, teda lineárnu regresiu.

Aby sme vedeli, čo možno od tohto ukazovateľa očakávať, je potrebné si uvedomiť kontext problému. Pri predikcií nákladov na zdravotnú starostlivosť nie je prekvapením, že modely nedosahujú vysoké R^2. O tomto fenoméne pojednáva napríklad [4], kde autor udáva, že R^2 je v takýchto modeloch poväčšinou na úrovni približne 20%. Pre porovnanie je napríklad v [26] možno dočítať sa, že PCG model v Holandsku s premennými pohlavie a vek dosahoval koeficient determinácie okolo 6%.

Na overenie a demonštráciu dostatočnej všeobecnosti modelov sme našu dátovú sadu rozdelili náhodným výberom na trénovaciu (75%) a testovaciu sadu (25%). V tabuľke 2 uvádzame výsledné hodnoty R^2 pre nasledovné modely:

	HYP			**B**		
Premenné	**Počet premenných**	R^2 **train**	R^2 **test**	**Počet premenných**	R^2 **train**	R^2 **test**
demo	16	24,61%	24,45%	16	9,66%	9,46%
demo + typZS	25	33,65%	33,56%	22	16,37%	16,28%
demo + diag	37	36,81%	36,76%	37	23,87%	23,36%
demo + typZS + diag	46	37,35%	37,30%	43	25,36%	25,04%
demo + typZS + diag + okres	117	37,80%	37,73%	114	25,44%	25,09%

Tabuľka 2: Výsledné hodnoty R^2 pre jednotlivé modely

V tabuľke 2 sa pod demo majú na mysli demograficke premenné ako pohlavie, veková kategória, ekonomická aktívnosť a úmrtie.

Vďaka vysokému počtu objektov v našej dátovej sade (takmer milión) sa modelom darí predikovať s podobnou odhadnutou presnosťou ako pre trénovacie, tak aj pre testovacie dáta. Možno teda konštatovať, že lineárne modely z hľadiska všeobecnosti nestrácajú na podobnej presnosti ani pri nových dátach.

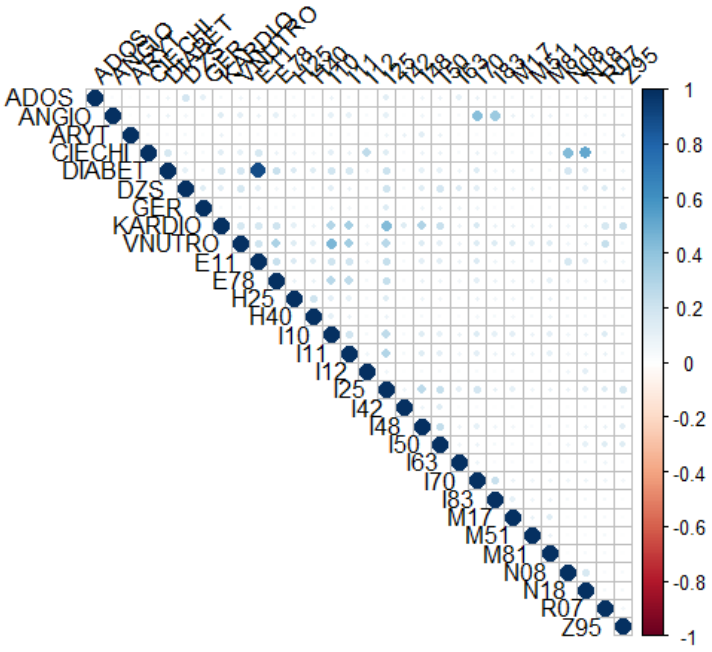
Ďalšou zaujímavosťou je, že po pridaní informácie o okrese, v ktorom daný poistenec býva (celkovo tak pribudlo 71 dummy premenných) sa R^2 zvýšilo len o necelé pol percento u HYP a necelých $0,10\%$ v triede s B-pomôckami. Rozhodli sme sa z hladiska *trade off*-u medzi počtom prememnných (čiže komplexnosťou modela) a presnosti tieto prememnné vynechať.

Zistili sme, že v skupine HYP vyšla jedna diagnóza H25 (Starecký zákal šošovky) nesignifikantná. Podobne aj v B skupine vyšla jedna diagnóza - I67 (Iné cievne choroby mozgu) nesignifikantná. Tieto diagnózy sme teda z modelu vylúčili. Mohlo ísť aj napriek našej snahe o nízku početnosť v pozorovaniach, či poprípade sa tieto premenné dali vysvetliť inými.

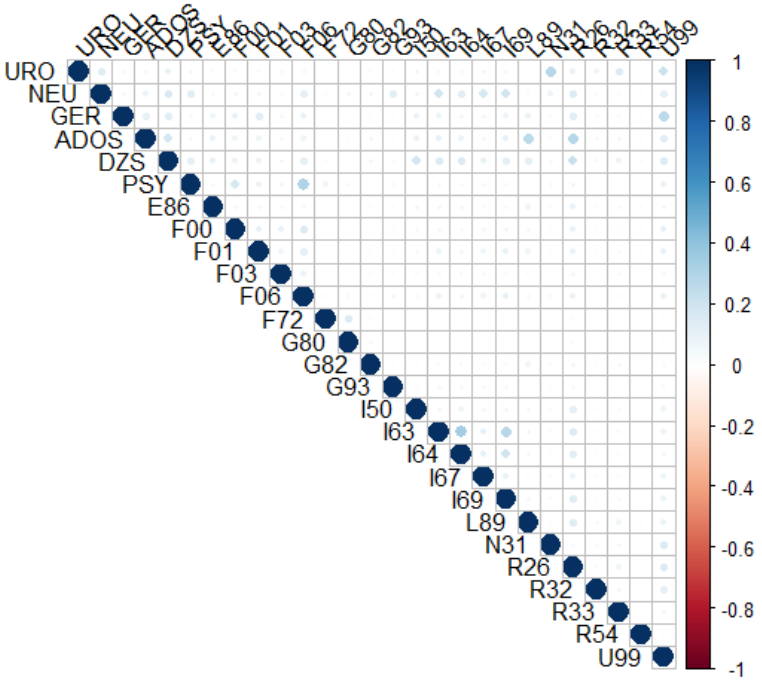
Pozornému oku čitateľa taktiež z tabuľky 2 neujde, že so zmenou z modelu s demografickými premennými a diagnózami na model s pridaním prememnných o typoch ZS v prípade hypertenzie narástlo R^2 taktiež len o približne pol percenta. Môže to byť spôsobené napríklad výraznou koreláciou medzi typmi ZS a diagnózami, čo overíme v nasledovnej časti.

5.2 Korelácia vysvetľujúcich premenných

Je prirodzené, že najmä v medicínsky dátach môže dochádzať k zvýšenej korelácií medzi jednotlivými premennými. Taktiež sa v našom prípade môže veľmi jednoducho stať, že určitá diagnóza je vykazovaná vo veľkej miere na jednom danom type ZS, resp. odbornosti. Pre lepší náhľad ponúkame na obrázkoch 10 a 11 vizualizácie výberových korelačných matíc pre obe sledované skupiny HYP a B.



Obr. 10: Vizualizácia výberovej korelačnej matice premenných pre HYP



Obr. 11: Vizualizácia výberovej korelačnej matice premenných pre B

Z obrázka 10 možno spozorovať vysokú koreláciu medzi diabetológiou a diagnózou E11, čo je *diabetes mellitus nezávislý od inzulínu*. Táto korelácia nie je koniec koncov prekvapujúca. Dosahuje približne hodnotu $89,6\%$. Všetky ostatné korelácie sú pod úroveň 45%. Pre preverenie ďalšej môžnej signifikantnej korelácií sme aplikovali analýzu VIF (*Variance Inflation Factor*). Podrobné vysvetlenie metódy je nad rámec našej práce, viac o sa o nej dá dočítať napríklad v [11]. Ide o metódu odhalenia kolinearity medzi vysvetľujúcimi premennými. Ako aj v článku pojdenávajú, treba si zvoliť treshold nad akú hranicu je výsledný ukazovateľ danej premennej neprípustný. Zväčša ide o hodnotu 10, 5 či 3. V našom prípade vyšli všetky ukazovatele premenných pod 2, len pre spomínane E11 a DIABET vyšlo číslo nad 5. Rozhodli sme sa dané premenné jednoducho zlúčiť, teda pacient mal priradenú hodnotu jedna, ak mal vykázaný nejaký výkon z diabetológie alebo na diagnózu E11. Ostatné premenné nechávame v pôvodnom stave, pretože nám prinášajú dostatočnú doplňujúcu informáciu oproti ostatným.

5.3 Overenie Gauss-Markových podmienok

5.3.1 Nulovosť strednej hodnoty rezíduí

Nulovosť strednej hodnoty chýb sa dá jedine predpokladať, pretože pri modeloch s interceptom je priemer odhadnutých chýb rovný nule vďaka samotnému spôsobu konštrukcie modelov.

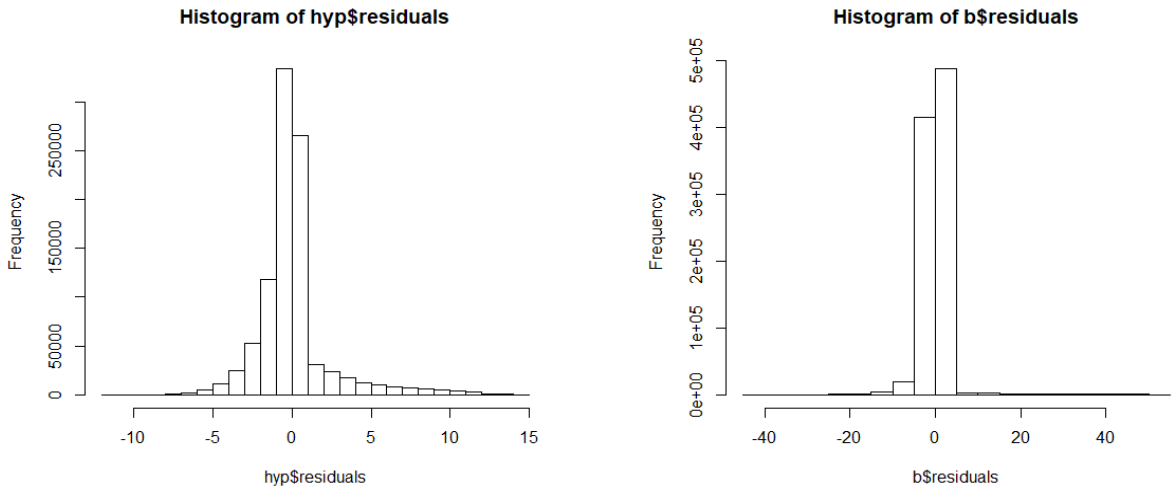
5.3.2 Homoskedasticita

Na otestovanie homoskedasticity použijeme Breusch-Pagan test [22]. Hodnota p-value nám vyšla veľmi blízka nule - hypotézu o homoskedasticite zamietame v oboch sledovaných modeloch. Poďla teórie však parametrické testy fungujú exaktne v prípade, že máme overenú normalitu dát, resp. nezamietnutú.

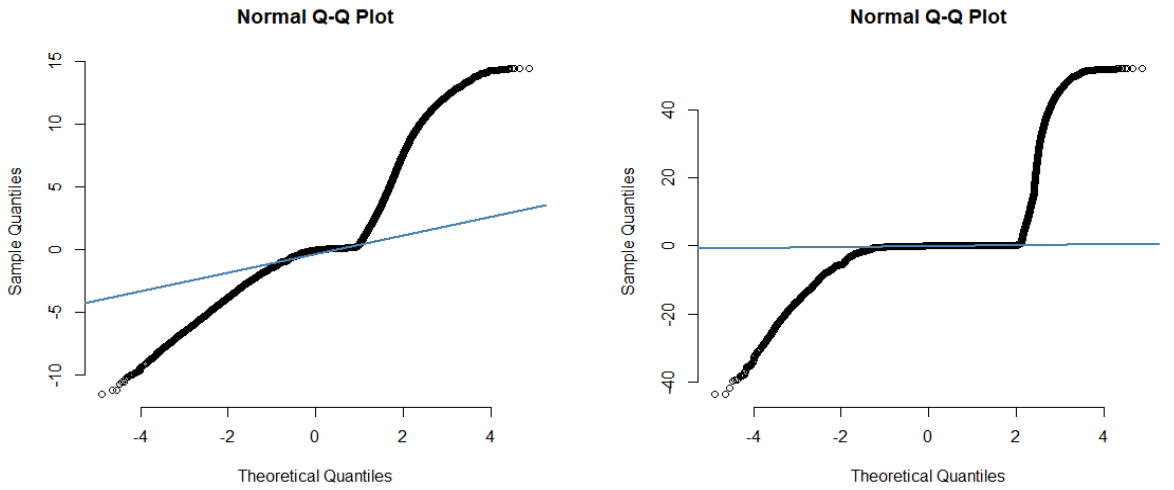
5.3.3 Normalita

Na overenie normality použijeme primárne grafické znázornenie, využijúc histogramy na obrázku 12 a Q-Q ploty na obrázku 13. Na histogramoch sa dá spozorovať istú nesymetrickosť. Tento jav tiež výrazne dokazujú aj Q-Q ploty. V prípade normality

sa očakáva graf v tvare rovnej čiary. To však nie je náš prípad. Zdá sa, že rezíduá modelov nemožno považovať za normálne. Treba však pripomenúť, že normalita rezíduí je najmä v prípade vysoko početných dátových sadách veľmi obtiažné dosiahnúť a ide skôr o akýsi ťažko dosiahnuteľný ideál. V praxi sa však často stáva, že táto podmienka nie je splnená. Samotnú metódu najmenších štvorcov však možno aplikovať aj bez predpokladu na normalitu. Pre zaujímavosť sa tomuto problému venuje napríklad článok [13].



Obr. 12: Histogram rezíduí pre skupinu HYP vľavo a pre skupinu B vpravo



Obr. 13: Q-Q plot pre skupinu HYP vľavo a pre skupinu B vpravo

Ďalšou možnosťou by bolo použiť logaritmickú transformáciu. V našom prípade by to však nebolo ideálne z dvoch dôvodov. Vo vysvetľovaných premenných sa nachádza veľa núl, čo sa dá ešte obísť pričítaním nejakej zvolenej konštanty k k nákladom každého pacienta. Takéto počínanie sa podľa [15] neodporúča napríklad kvôli nejasnosti zvolenej konštanty. Taktiež sa môžu pozorovania s nulovými nákladmi inak správať oproti kladným a po transformácií by sa tento jav ťažšie zachytával. Druhým dôvodom je nepraktickosť pracovania s logaritmami pre naše účely. Existuje síce spôsob spätnej tranformácie pomocou tzv. *smearing estimator* [5], no nedosahoval požadované výsledky. My sa logaritmickej transformácií ďalej venovať nebudeme.

6 Optimalizácia rozhodovacích stromov

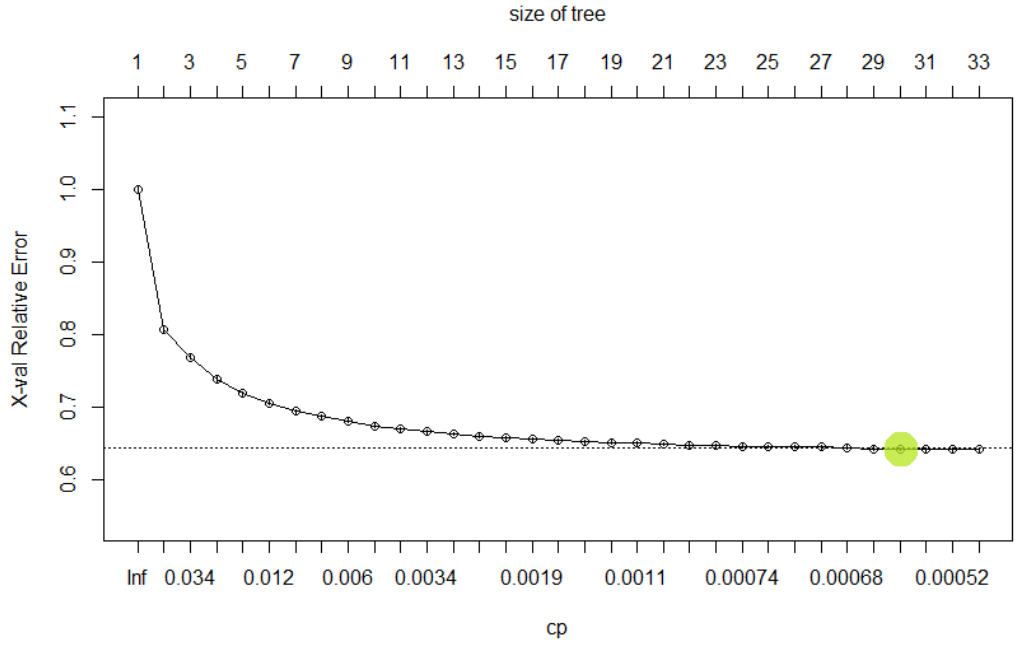
Nasleduje otestovanie, ako sa s danou úlohou popasujú rozhodovacie stromy. Zaujímvaým pozorovaním môže byť taktiež porovnanie modelu s jedným stromom a či a o koľko sa zlepší predikcia pri vyžití náhodného lesa. Pri budovaní stromov budeme využívať tie premenné, ktoré sme nechali na záver v lineárnej regresií. Pri optimalizácií budeme využívať metódu rozdelenia dátového súboru a trénovaciu a testovaciu, podobne ako sme to aplikovali v podkapitole 5.1.

6.1 CART rozhodovací strom

Pri použití algoritmu s jedným rozhodovacím stromom (ďalej už len RS) využívame funkciu `rpart()` v softvéri R [24]. Pôvodnou myšlienkou pri optimalizovaní RS bolo v prvej etape spraviť strom do maximálnej možnej hĺbky, pokým sa mali aké dáta od seba odseparovať. Táto myšlienka sa však využíva skôr pri klasifikačných problémoch, kde sa predikuje kategória, do ktorej pozorovanie padne. V regresných problémoch by to technicky znamenalo vybudovať taký strom, ktorý sa nezastaví pokým sa dáta v koncových uzloch rozlišujú vo vysvetľujúcich premenných a ich skutočná hodnota predikovanej premennej nie je stále totožná. V našom prípade to skúšať nejdeme, pretože sa nám to pri takom vysokom počte pozorovaní nejaví ako zmysluplná cesta.

Algoritmu sa dajú nastaviť rôzne stopovacie parametre, spomínané v časti 2.2.3. Pre veľký rozsah dátovej sady sme stopovacie pravidlá typu minimálny počet pozorovaní v uzle nebrali do úvahy. Najväčšia možná maximálna hĺbka stromu v softvéri R je limitovaná na hodnotu 30. Avšak pre nás bude kľúčovým hyperparametrom minimálny nárast informatívnosti po rozdelení uzla - `cp`, ktorý je štandardne nastavený na úroveň $0,01$. Existuje viacero palcových pravidiel, ako optimálne vybrať hodnotu tohto hyperparametra. Využívajú sa pri tom informácie z krížovej validácie (priemer odhadnutej chyby z validačných sád a štandardná odchýlka pre daný level cp) . Často je zaužívaný prístup spomenutý v časti 2.2.3. Zoberem takú úroveň cp, pri ktorom je priemerná chybovosť z krížovej validácie najmenšia. V tomto prípade sa pod chybovosťou myslí doplnok koeficientu determinácie do 1. V oboch prípadoch však s klesaním cp klesá aj priemerná chybovosť a nedostávame bod zlomu. Rozhodli sme sa teda využiť za-

budovanú funkciu `plotcp()` [23], podľa ktorej zoberem najväčšie cp, pre ktoré jeho priemerná chybovosť z krížovej validácie spadá do celkovo najmenšej nadobudnutej priemernej chybovosti z krížovej validácie plus jej štandardná odchýlka. Dostávame tak obrázky 14 a 15.

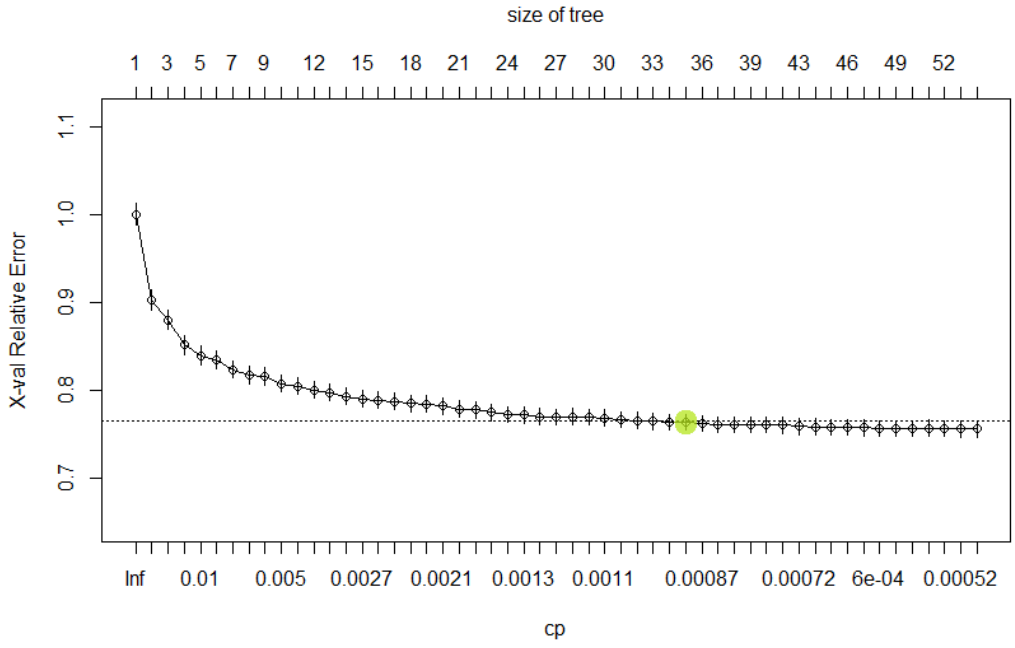


Obr. 14: Optimalizácia hyperparametra cp pre skupinu HYP

Zvýraznili sme približne odhadnuté miesto a rozhodli sme sa nastaviť cp na úroveň 0,00055 pre skupinu HYP a 0,0009 pre skupinu B. Zhrnutie vlastností a dosiahnutých R^2 rozhodovacích stromov so zvolenými parametrami cp ponúkame v tabuľke 2.

	HYP	**B**
R^2 train	35,68%	24,46%
R^2 valid	35,64%	23,44%
počet koncových uzlov	29	36
maximálna hĺbka	7	11
cp	0,00055	0,0009

Tabuľka 3: Vlastnosti a odhadnutá presnosť výsledných stromov pre skupinu HYP a B



Obr. 15: Optimalizácia hyperparametra cp pre skupinu B

6.2 Náhodný les

Pri aplikácií takéhoto komplexného algoritmu budeme používať jazyk Python a implementovanú knižnicu *Scikit-learn*. Opäť nastáva otázka, ako správne nastaviť parametre daného algoritmu. Využívaná funkcia `RandomForestRegressor()` [21] neponúka nastavenie parametra cp. V prípade náhodných lesov našu pozornosť zameriame najmä na optimalizovanie hĺbky stromov, pretože tento hyperparameter má na komplexnosť algoritmu exponenciálny vplyv. Opäť sme využívali metódu krížovej validácie využijúc funkciu `cross_validate()` [20]. Zistenia z krížových validácií ponúkame v tabuľkách 4 a 5.

HYP		
maximálna hĺbka	R^2 train	R^2 valid
7	34,19%	33,81%
10	37,91%	36,22%
12	40,44%	37,01%
15	44,36%	37,07%

Tabuľka 4: Optimalizácia maximálnej hĺbky stromu pre skupinu HYP

Na skupinu HYP sme začali testovaním maximálnej hĺbky na hodnote 7. V tomto prípade sme dosiahli R^2 na úrovni okolo 34%. V porovnaní s výsledkom z validačných dát je možno konštatovať, že ešte nedochádza k pretrénovaniu. Zvýšením maximálnej hĺbky na hodnotu 10 pozorujeme, že došlo k zvýšeniu ukazovateľa na validačných dátach približne o $2,5\%$. Pri maximálnej hĺbke 12 sa však so zvýšením R^2 na validačných dátach o necelé 1% výrazne zvýšilo R^2 z validačných dát. Možno pozorovať istú mieru pretrénovania. Po zvýšení maximálnej hĺbky na úroveň 15 sa nám nepridalo signifikantné zlepšenie na validačných dátach. Z daných informácií sme sa rozhodli pre model s úrovňou hĺbky stromov 10.

B		
maxímálna hĺbka	R^2 test	R^2 valid
5	23,14%	22,18%
7	27,35%	25,65%
9	31,79%	26,43%
11	36,42%	26,90%
13	41,06	26,75%

Tabuľka 5: Optimalizácia maximálnej hĺbky stromu pre skupinu B

Z výsledkov R^2 dosiahnutých náhodnými lesmi predikujúcimi nákladovosť na B-pomôcky si možno všimnúť nasledovné. Od úrovne 9 dochádza k výraznejšiemu trendu pretrénovania sa, pričom R^2 na validačných dátach ostáva približne na rovnakej úrovni. Rozhodli sme za vhodnú hĺbku stromov v náhodnom lese zvoliť hodnotu 7.

Okrem maximálnej hĺbky stromov sme zisťovali, ako dôležitý je faktor počtu stromov v náhodnom lese. Začali sme na úrovni 30-tich stromov. Predošlá optimalizácia bola robená s týmto nastavením. Následne sme po zvolení vhodnej maximálnej hĺbky vyskúšali, ako sa sa zmenia R^2 pre nastavenie 200 stromov v náhodných lesoch. V oboch sledovaných skupinách sme došli k záveru, že tento faktor nepomohol k signifikantnému zlepšeniu (v oboch skupinách došlo k zlepšeniu R^2 na validačných dátach len okolo $0,2\%$) a pridŕžame sa pôvodného nastavania parametra na 30 stromov. Parametre ako minimálny počet pozorovaní v uzle, či minimálny počet pozorovaní, aby

sa daný uzol mohol ďalej deliť sme nechali na hodnote 1, resp. 2. Vzhľadom na vysoký pomer počtu pozorovaní v našej vzorke ku koncovým uzlom pre zvolené hĺbky stromov, tieto limity s veľkou pravdepodobnosťou nehrajú až takú dôležitú rolu.

7 Porovnanie modelov

Po malej modifikácií premennej a optimalizácií rozhodovacích stromov CART a náhodných lesov, ktoré prebiehalo na trénovacej sade, si následne všetky tri metódy otestujeme na testovacej sade (25% odložených dát) tak, ako to bolo v časti 5.1 u lineárnej regresie. Dosiahnuté skóre vo viacerých metrikách uvádzame v tabuľke **??**. Metrika MAE má napríklad dobrú interpretovateľnosť. V našom prípade nám hovorí, o ± koľko eur sa priemerne pri predikcií daných nákladov pacienta pomýlime. Pre úplnosť a na sledovanie možného pretrénovania uvádzame v tabuľke **??** dosiahnuté skóre aj na trénovacích dátach.

	HYP			**B**		
Metóda	**R2**	**MAE**	**RMSE**	**R2**	**MAE**	**RMSE**
LR	37,05%	1,2649	2,2986	25,04%	0,8272	3,7510
CART	35,60%	1,2529	2,3251	24,39%	0,7213	3,7674
NL	36,50%	1,2532	2,3096	25,67%	0,7352	3,8124

Tabuľka 6: Dosiahnuté skóre pre jednotlivé metódy na testovacích dátach

	HYP			**B**		
	R2	**MAE**	**RMSE**	**R2**	**MAE**	**RMSE**
LR	37,13%	1,2631	2,2969	25,36%	0,8231	3,7411
CART	35,68%	1,2506	2,3233	24,46%	0,7184	3,7636
NL	38,11%	1,2436	2,2785	27,24%	0,7054	3,6679

Tabuľka 7: Dosiahnuté skóre pre jednotlivé metódy na trénovacích dátach

V tabuľkách je pod skratkou LR myslená lineárna regresia, pod CART rozhodovací strom typu CART a NL označuje náhodný les. Ako môžeme vidieť z testovania, pre všetky tri metódy dostávame približne podobné R^2. Vzhľadom k náhodnosti výberu testovacej sady nepovažujeme tieto rozdiely za signifikantné. Všetky metódy dosahujú podobné uspokojenie. Čo možno podotknúť je už spomínaný fakt, že pri náhodných lesoch v oboch skupinách dochádza k malému pretrénovaniu. Osobne vidíme ako vhodnú

voľbu metódy lineárnu regresiu, ktorá je svojou komplexnosťou najjednoduchšia.

Porovnali sme dosahované presnosti jednotlivých metód pri predikcií nákladov v daných skupinách na pacienta. Pre nás by však bolo prínosné sa taktiež pozrieť na dosahovanú presnosť pri predikcií nákladov už na jednotlivých PZS. Nasčítali sme predikcie nákladov sledovaných produktov pacientov, ktorí sú kapitovaní u daného PZS a to sa stáva odhadom nákladovosti daného PZS. V tomto kroku si uvedomujeme istý šum dát, pretože sa stávalo, že nejaký pacient mal sledovaný produkt predpísaný aj iným všeobecným lekárom, ako tým, u ktorého bol na konci roku 2017 kapitovaný. Avšak vzhľadom k tomu, že sme chceli čo najvernejšie zachytiť reálnu spotrebu daných produktov daného poistenca, nechceli sme tieto dáta odtrániť. Mohlo dôjsť len k jedinému prípadu, kedy by tento jav nepriaznivo ovplyvnil našu analýzu. Ak sme nespravodlivo priradili nejaké náklady danému PZS, teda on ich reálne nepredpísal, no na základe toho nám vyjde ako outlier. Na tento problém pri preverovaní podozrivého PZS musíme myslieť a overiť to dodatočnou analýzou. Z praxe nám však vychádzali tieto šumy len okolo 6% celkovej nákladovosti daného PZS a zatiaľ sme sa nestretli s takým, kde by to bol ozaj dôvod vyčlenenia z podozrivej vzorky.

Po nasčítaní predikcií k priradeným PZS a následným porovnaním s reálnymi nákladmi poistencov daného PZS možno očakávať väčšiu presnosť odhadu. Viac sa tejto myšlienke venujeme v kapitole 8. Pre úplne porovnanie jednotlivých metód uvádzame teda nielen dosahované skóre presnosti počítané na pacienta, ale aj dosiahnuté skóre presnosti na PZS, viď tabuľka 8. To je koniec koncov to, čo nás zaujíma viac.

	HYP			**B**		
Metóda	**R2**	**MAE**	**RMSE**	**R2**	**MAE**	**RMSE**
LR	90,15%	98,68	169,78	68,65%	96,68	178,43
CART	90,10%	100,33	170,13	68,41%	97,05	179,12
NL	89,52%	102,43	175,16	64,70%	102,08	189,34

Tabuľka 8: Dosiahnuté skóre presnosti odhadov nákladovosti PZS

Vidíme, že presnosť sa podľa očakávania výrazne zlepšila a taktiež môžeme pozorovať, že sa lineárnej regresií darilo vo všetkých metrikách najviac. Naopak prekvapivo

algoritmy náhodných lesov vykázali najslabšie dosiahnuté skóre presnosti. Z hodnoty RMSE najmä v skupine B možno šípiť, že pri predikcií náhodným lesom sa vytvorilo viacero alebo výraznejších outlierov v odchýľkach modelu. Čo sa týka hodnoty MAE, tentokrát možno povedať,že predikcie priemerných mesačných nákladov PZS na sledované produkty sa v oboch prípadoch od tých reálnych pohybujú v priemere ± 100 eur.

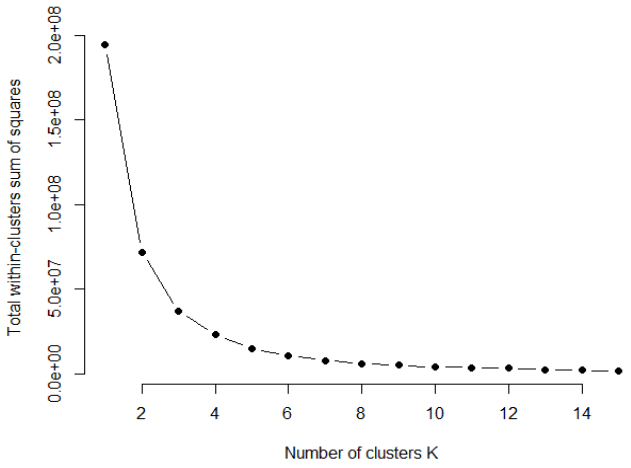
8 Analýza outlierov

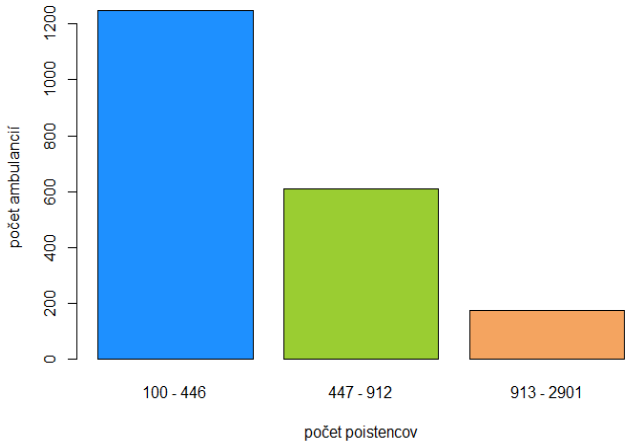
V tejto kapitole sa dostávame postupne k cieľu našej diplomovej práce. Hlavnou myšlienkou vybudovaných modelov bolo zachytiť populačné vlastnosti poistencov a dostať tak celkový obraz, aké náklady by sme priemerne od takého a takého pacienta v danej vekovej kategórií, daného pohlavia a s danými rizikovými faktormi z vykázanej zdravotnej starostlivosti mohli približne očakávať. Je pravdou, že modely vysvetľujú okolo 36%, resp. 26% variability nákladov, teda modely pri predikcií nákladov na samotných pacientov nedosahujú vysokú presnosť. Ako sme sa však v predošlej kapitole mohli presvedčiť, po nasčítaní predikcií na jednotlivých PZS sa tieto rozdiely budú istým spôsobom vykrývať a teda pri samotnej predikcií nákladov na PZS dosahujeme lepšiu relatívnu presnosť. Myšlienku za tým si môžeme predstaviť na zjednodušenom prípade. Predstavme si, že ženám vo vekovej kategórií od 40 do 44 rokov s danými výkazmi z VZS sa priemerne očakávajú priemerné mesačné náklady okolo 2 eur. V realite skôr však ide o to, že napríklad 9 z 10 žien má nulové náklady a zvyšná má 20. Keď však má PZS vo svojom kmeni veľký počet takejto skupiny populácie, v priemere za predpokladu "normálneho" správania lekára a predpokladu istej mieri "podobnosti" jeho kmeňa s celkovou populáciou, na ktorom táto analýza stavia, sa tieto chyby predikcie po nasčítaní do určitej mieri vykryjú.

Tu si však treba uvedomiť dôležitý fakt. Čím viac pacientov má daný lekár, tým prirodzene očakávame, že sa spomínané odchýlky priemerných nákladov na pacienta vykryjú vo vyššej miere. Preto je veľmi dôležité neporovnávať v analýze lekára so 100 pacientami s lekárom, ktorí ich má v našej poisťovni poistených 1000. Pre nás bol inšpiráciou článok [3], kde sa pri podobnom probléme nevyrovnanosti dostupných informácií rozhodli použiť klastrovú analýzu. Následne sa lekári porovnávali iba vrámci klastrov.

V našom prípade sme sa rozhodli porovnávať PZS nie ako konkrétnych lekárov, ale poďla nižších zložiek. Často je totiž zvykom vrámci jednej ambulancie jednoducho zameniť ošetrujúceho lekára, resp. neodhlásiť kolegu zo systému a vykázať výkon pod jeho menom. V našich dátach vystupuje 2276 ambulancií. Rozhodli sme sa však preveriť iba tie, ktoré u nás majú minimálne 100 poistencov. Takto sa do našej ďalšej analýzy dostalo 2230 ambulancií. Využili sme k-means metódu zhlukovej analýzy. Pri danej

metóde sme zhlukovali podľa atribútu počet poistencov danej ambulancie v našom dátovom súbore. Pomocou lakťového pravidla, viď obrázok 16, sme určili optimálny počet klastrov na 3. Dostali sme tak klastre opísané na obrázku 17.





Obr. 16: Výber optimálneho počtu klastrov **Obr. 17:** Počet PZS v jednotlivých klastroch

Ako sme už vyššie načrtli, pri hodnotení PZS sa zameriame na priemernú odchýlku medzi reálnymi a predikovanými nákladmi jeho pacientov, definovanú vzťahom:

$$Err_j = \frac{1}{n_j} \sum_{i \in PZS_j} (y_i - \hat{y}_i), \tag{23}$$

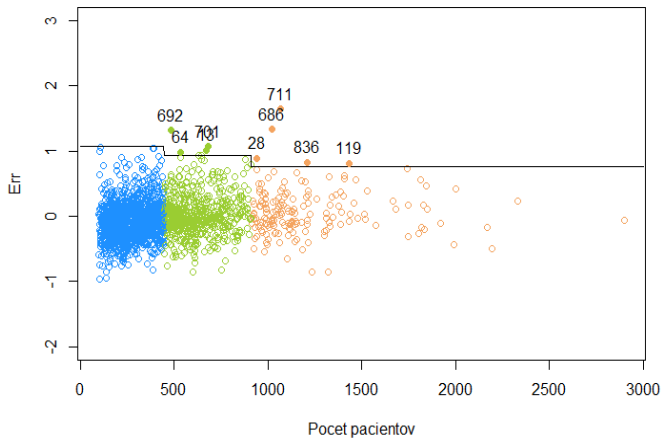
kde n_j predstavuje počet poistencov daného PZS. Existujú rôzne spôsoby ako určiť, či dosiahnutú priemernú odchýlku daného PZS možno považovať za signifikantne vyššiu v porovnaní s ostatnými z klastra. Jednou jednoduchou možnosťou je napríklad považovať za podozrivých tých PZS, ktorých tento ukazovateľ patrí do najvyšších k percent. Tento spôsob je vhodný vtedy, keď dopredu vieme, aké množstvo podozrivých chceme označiť.

My sme vyskúšali známu ideu pomocou **Z-scóre**, využitú napríklad aj v spomínanom článku [3]. Každému PZS sme vyrátali hodnotu

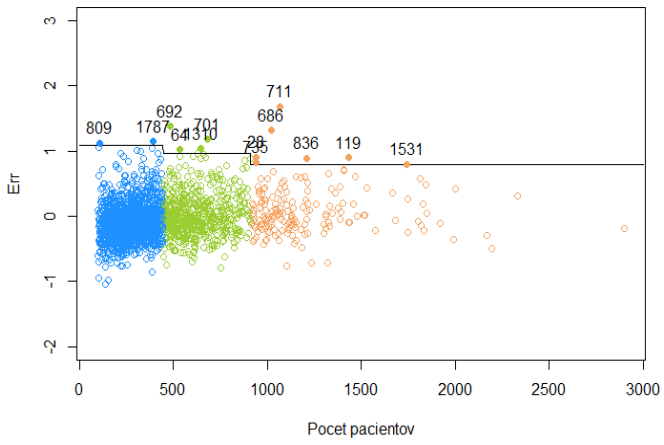
$$z_j = \frac{(x_j - \hat{x})}{S}, \tag{24}$$

kde x_j predstavuje sledovaný ukazovateľ Err_j, $\hat{x}$ priemer týchto odchýlok v danom klastri a S sa rovná jeho štandarnej odchýlke. Následne za outlierov určíme tých, ktorí dosiahli z_j väčšie ako nami stanovená hodnota, označme L. Často sa v literatúre uvádza

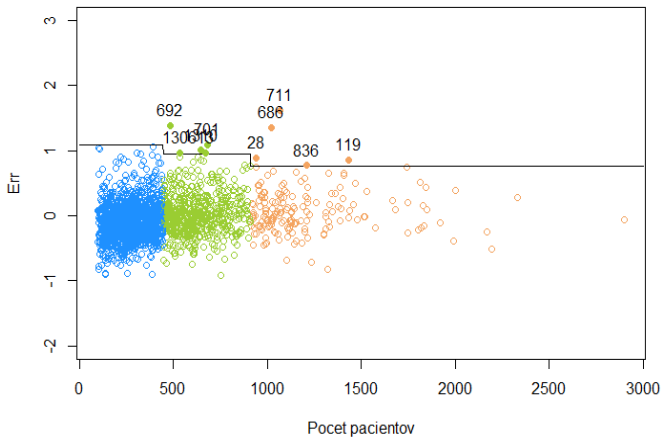
hodnota 3. Pre porovnanie napríklad v článku [3] bola táto hodnota nastavená na 2,33. Taktiež dochádza v datasetoch s menším počtom pozorovaní (u nás počet PZS v klastri) k tomu, že je vrámci takéhoto datasetu ťažké dosiahnúť vysoké Z-scóre [9]. Kvôli danému javu a pretože sme chceli ostať konzistentní tvrdeniu, že pri PZS s menším počtom pacientov by sme chceli docieliť menšiu prísnosť ako u tých s väčším počtom, nastavili sme hranicu L postupne na 4 v prvom klastri, 3 v druhom a 2 v treťom. Výsledky tejto analýzy pre skupinu liekov proti hypertenzií ponúkame na obrázkoch 18, 19 a 20.



Obr. 18: Analýza outlierov HYP - LR



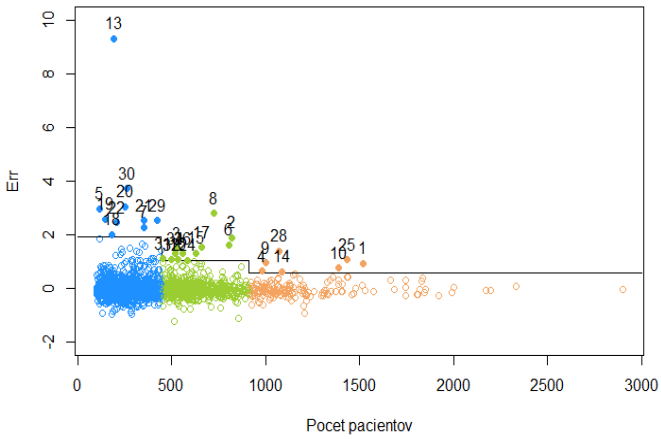
Obr. 19: Analýza outlierov HYP - NL



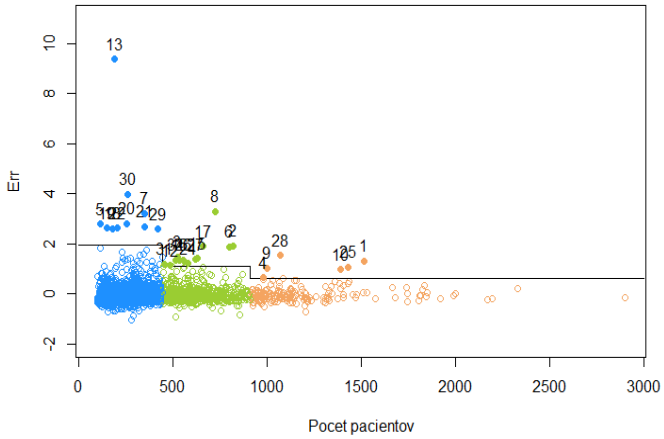
Obr. 20: Analýza outlierov HYP - CART

Ako však môžeme pozorovať, využijúc akúkoľvek metódu na predikciu, výsledky sú s malou variabilitou veľmi podobné. Modely sa viac menej zhodujú na rovnakých PZS. V takejto analýze samozrejme veľa záleží od nastavenia hranice L. Na hľadanie outlierov sa môže použiť napríklad aj DBscan [6]. V praxi sa prirodzene táto analýza častokrát odohráva metódou pozriem a vidím, nakoľko sme si dáta upravili do dvoch dimenzií a sú ľahko vizualizovateľné. Toto je v praxi vždy výhodou, aby človek dostal dobrý náhľad, ako to v klastroch vyzerá a čo sa nám môže javiť podozrivo.

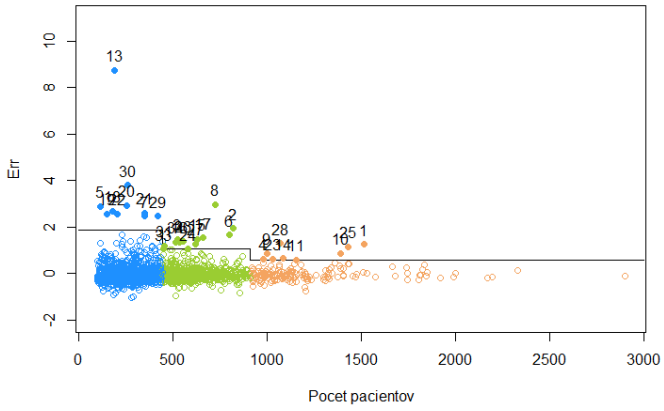
Zaujímavým faktom je, že napríklad s PZS označeným ako 1531 boli vrámci isté preverovania ajv poisťovni Dôvera aj v minulosti.



Obr. 21: Analýza outlierov B - LR



Obr. 22: Analýza outlierov B - NL



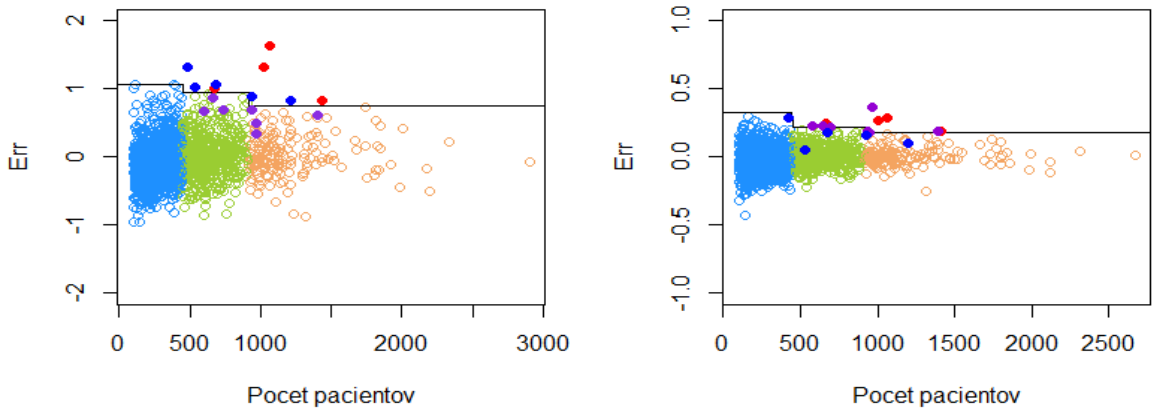
Obr. 23: Analýza outlierov B - CART

Totožnú analýzu sme aplikovali aj pre skupinu B-pomôcok, viď obrázky 21, 22 a 23. Oku isto neunikne výrazný outlier, ktorého priemerná odchýlka na pacienta prevyšuje priemer priemerných odchýlok skoro 10-násobne. Po prvom náhľade usudzujúc z adries jeho pacientov vyzerá, že daný lekár obsluhuje viacero domovov sociálnych služieb. Takýto lekári majú spravidla vyššie náklady na B-pomôcky. Je však na uváženie, či je to aj vždy v celej miere opodstatné. Do určitej mieri je pravdou, že ľudia v týchto zariadeniach sú na tom spravidla zdravotne horšie, ako rovnako starý človek, ktorý je dostatočne samostatný a nie je odkázaný na domov sociálnych služieb. Do budúcna vidíme možnosť rozšírenia modelu o informáciu, či je daný pacient z domovu dôchodcov. Avšak na druhej strane sa môže stať, že to lekári v týchto kruhoch výrazne zneužívajú a predpisujú plienky bez reálnej potreby pacienta. Vidíme teda praktické využitie aj modelu bez tejto premennej.

9 Modely s predošlými nákladmi

Ako zaujímavou modifikáciou našich modelov vidíme pridanie informácie o minulých nákladoch na sledované produkty z roku pred tým, teda pridaním premennej s priemernými mesačnými nákladmi za rok 2016. V tejto analýze sme sa rozhodli využívať lineárnu regresiu, pretože vykazovala porovnateľné výsledky s rozhodovacími stromami, no je výpočtovo najjednoduchšia a netreba hľadať optimálne vstupné hyperparametre. Ako sme mohli v predošlých častiach vidieť, lineárna regresia sa najmä pri takom veľkom rozsahu dátového súboru nemá tendenciu pretrénovať, preto v tejto kapitole budeme trénovať modely na všetkych dátach. V tejto analýze sme však očistili dáta od pacientov, ktorí na konci roku 2016 neboli ešte poistení v našej poisťovni a teda nemali sme k dispozícií ich dáta z toho roku. Taktiež sme odstránili tých poistencov, ktorí boli ku koncu roku 2016 zapísaní u iného všeobecného lekára oproti roku 2017. Stručne si predstavíme dosiahnuté výsledky.

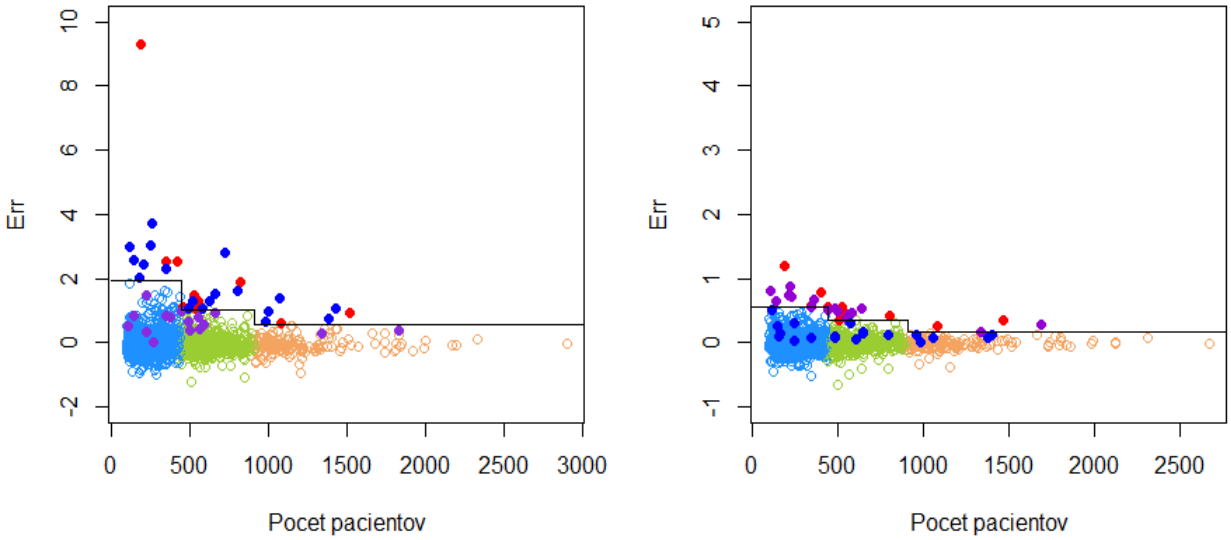
V prvom rade si treba uvedomiť, že pridaním takejto premennej do modelu signifikantne zvýšime presnosť modelov. Tento fakt potvrdzuje aj zvýšený koeficient determinácie z pôvodných cca 37% na **85,28%** v prípade sledovanej skupiny HYP a z pôvodných približne 25% na **83,25%** v prípade B skupiny. Pre porovnanie modelov s a bez predošlých nákladov a náhľad, ako sa detekovaní outlieri zmenili, ponúkame obrázok 24.



Obr. 24: Porovnanie detekcie outlierov v modeli bez predošlých nákladov (vľavo) a s predošlými nákladmi (vpravo) - skupina HYP

Na obrázku 24 sú červenou farbou zvýraznení outlieri, ktorých detekovali oba modely, tmavomodrou farbou sú označení tí, ktorí sa javili podozrivo v modeli bez predošlých nákladov, no v modeli s predošlými nákladmi sa do podozrenia (u niektorých len tesne) nedostali. Fialovou sú naznačení tí, ktorých model zahrňujúci predošlé náklady označil za podozrivých, no v druhom modeli toto označenie nezískali. Môžeme pozorovať, že sa nám takýmto porovnaním detekoval zaujímavý typ outlierov, napríklad pre PZS označených fialovou farbou platí, že vzhľadom na výkazy a demografický stav jeho poistencov nemá celkovo vysoké náklady, no oproti roku predtým mu tieto náklady výraznejšie stúpli.

Podobným spôsobom sme postupovali aj pri skupine B- pomôcok, kde náhľad do detekcie outlierov v modeloch na predikciu tejto skupiny pomôcok ponúkame na obrázku 25



Obr. 25: Porovnanie detekcie outlierov v modeli bez predošlých nákladov (vľavo) a s predošlými nákladmi (vpravo) - skupina B

Pozorujeme, že v skupine B-pomôcok sa oba modely zhodli na 10 PZS, no ďalších 19 (fialových) a 16(modrých) boli detekovaní rôzne. Taktiež vidíme pokles priemernej odchylky na pacienta spomínaného výrazného outliera, teda vo veľkej miere mu nový model s predošlými nákladmi vysvetlil jeho vysokú nákladovosť.

Záver

Cieľom našej práce bolo navrhnúť spôsoby na objavovanie podozrivého správania poskytovateľov zdravotnej starostlivosti. Zvolili sme si zamerať našu pozornosť na predikciu nákladov poistencov na skupiny liekov a zdravotníckych pomôcok predpísaných všeobecnými lekármi. Konkrétne sme sa v práci venovali liekom proti hypertenzií a B-pomôckam.

Navrhli sme typ premenných vstupujúcich do modelu a porovnali viacero druhov metód predikcie očakávaných nákladov poistenca. Porovnali sme metódu najmenších štvorcov, rozhodovací strom CART a následne aj generalizujúcejší algoritmus náhodný les. Po zobratí všetkých plusov a mínusov jednotlivých metód a dosiahnutých výsledkov nám lineárna regresia prišla ako najjednoduchšia a najvhodnejšia voľba. Následne sme po nasčítaní týchto predikcií na daného poskytovateľa zdravotnej starostlivosti určili spôsob detekcie podozrivého správania. Navrhli sme tak metodológiu odhaľovania anomálií v sfére nákladovosti liekov a zdravotníckych pomôcok a prakticky zrealizovali a otestovali na dvoch spomenutých skupinách.

Jedným z prínosov tejto práce vidíme aj napriek v našom prípade nie úplne uspokojivým a nie výrazne lepším výsledkom oproti klasickej lineárnej regresií, teraz v dobe rozmáhajúcej sa dátovej vedy, naštudovanie fungovania rozhodovacích stromov a náhodných lesov, ktoré si často v zdravotníctve nájdu svoje uplatnenie [17]. Vďaka tomu sme mohli pre náš problém otestovať a porovnať hneď niekoľko metód.

Ako hlavný prínos vidíme veľkú využiteľnosť navrhnutého postupovania v praxi. Snahou bolo položiť základ podobným analýzam, ktoré by mohli prispievať k zefektívneniu nakladania finančných prostriedkov v zdravotníctve. V budúcnosti by sa mohla navrhnutá metodológia rozšíriť na viaceré skupiny liekov a pomôcok s pokrytím čo najväčšieho objemu finančných prostriedkov, ktoré zdravotná poisťovna vynakladá na preplatenie daných produktov. Počas postupu práce a pri analyzovaní získaných outlierov sme prišli na možné zlepšenia a rozšírenia modelov. Napríklad v prípade B-pomôcok by možno pomohlo dobudúcna nejaký spôsobom zachytiť, či daný poistenec býva v domove sociálnych služieb. Taktiež je tu možnosť rozšíriť pozorované obdobie z jedného roku na dlhšie obdobie, čo by mohlo priniesť prirodzene vyššiu mieru správneho odhadnutia rizikovosti daného poistenca.

Literatúra

[1] Bodóová, A.: *Aplikácia data miningu v zdravotníctve*, diplomová práca, Univerzita Komenského v Bratislave (2018)

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: *Classification and Regression Trees*, Chapman and Hall/CRC, (1984)

[3] van Capelleveen, G. et al: *Outlier detection in healthcare fraud: A case study in theMedicaid dental domain* International Journal of Accounting Information Systems (2016), Volume 21, 18-31

[4] Diehr, P. et al.: *Methods for analzying health care utilization and costs*, Annual Review of Public Health, Volume 20 (1999), 125-144

[5] Duan, N.: *Smearing Estimate: A Nonparametric Retransformation Method*, Journal of the American Statistical Association (1983), Volume 78 (383), 605-610

[6] Ester, M. et al.: *A DensityBased Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, The Second International Conference on Knowledge Discovery and Data Mining (1996), 226–231.

[7] Hajíčková, T.: *The Pharmacy-based Cost Group Model: Application in the Czech Health Care System*, Master Thesis, Charles University in Prague (2015)

[8] Harman, R.: *Prednáška o analýze zhlukov*, dostupné na internete (10.5.2019): `http://www.iam.fmph.uniba.sk/ospm/Harman/teaching.htm`

[9] Iglewicz, B., Hoaglin D.: *How to Detect and Handle Outliers*, The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.(1993), Volume 16

[10] InsightR: How Random Forests improve simple Regression Trees? dostupné na internete (10.5.2019): `https://insightr.wordpress.com/2017/09/23/how-random-forests-improve-simple-regression-trees/`

[11] Kock, N., Lynn, G.: *Lateral Collinearity and Misleading Results in Variance-Based SEM: An Illustration and Recommendations*, ournal of the Association for Information Systems (2012), Volume 13 (7)

[12] Lee S.M., Kang J.O., Suh Y.M.: *Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models*, J Korean Med Sci. (2004) Volume 19(5), 677-81.

[13] Lumley, T. et al.: *The Importance of the Normality Assumption in Large Public Health Data Sets*, Annual Review of Public Health (2002), Volume 23 , 151-169

[14] Medzinárodná klasifikácia chorôb - MKCH, dostupné na internete (10.5.2019):

`http://www.nczisk.sk/Standardy-v-zdravotnictve/Pages/Medzinarodna-klasifikacia-chorob-MKCH-10.aspx`

[15] Mihaylova, B. et al.: *Review of statistical methods for analysing healthcare resources and costs*, Health Economics (2011), Volume 20, Issue 8 , 897-916

[16] Ministerstvo zdravotníctva: kategorizácia pomôcok, dostupné na internete (10.5.2019):

`http://www.health.gov.sk/Clanok?zkzp-201904`

[17] Morid, M. A. et al.:*Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation*, AMIA Annu Symp Proc. (2017), 1312–1321

[18] Powers C.A. et al: *Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques* Med Care. (2005),Volume 43(11), 1065-72.

[19] Santrová, A,: *Návrh a optimalizácia algoritmov na odhaľovanie fraudov v dátach zdravotnej poisťovne*, diplomová práca, Univerzita Komenského v Bratislave (2017)

[20] Scikit-learn - dokumentácia: Cross Validate, dostupné na internete (10.5.2019):

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html#sklearn.model_selection.cross_validate

[21] Scikit-learn - dokumentácia: Random Forest Regressor, dostupné na internete (10.5.2019):

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[22] Somorčík, J., Ekonometria, prednášky, Fakulta matematiky, fyziky a informatiky UK, 2017

[23] R - dokumentácia: plotcp, dostupné na internete (10.5.2019):

https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/plotcp

[24] R - dokumentácia: rpart, dostupné na internete (10.5.2019):

https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart

[25] Torgo, L.,F.: *Inductive Learning of Tree-based Regression Models*, PhD. Thesis, University of Porto (1999)

[26] van Kleef, et al.: *Risk equalization in The Netherlands: an empirical evaluation* Expert Review of Pharmacoeconomics & Outcomes Research (2013), Volume 13 (6)

Príloha A

Skratka	**Názov**	**Kód typu ZS**	**Kód odbornosti**
ADOS	Agentúra domácej ošetrovateľskej starostlivosti	104	-
ANGIO	Angiológia	200	056
ARYT	Arytmia a koronárna jednotka	200	278
CIECHI	Cievna chirurgia	200	068
DIABET	Diabetológia, poruchy látkovej premeny a výživy	200	050
DZS	Dopravná zdravotná služba	800	-
GER	Geriatria	200	060
KARDIO	Kardiológia	200	049
VNUTRO	Vnútorné lekárstvo	200	001

Tabuľka 9: Vybrané typy ZS (a odbornosti) pre skupinu HYP

Skratka	**Názov**	**Kód typu ZS**	**Kód odbornosti**
ADOS	Agentúra domácej ošetrovateľskej starostlivosti	104	-
DZS	Dopravná zdravotná služba	800	-
GER	Geriatria	200	060
NEU	Neurológia	200	004
PSY	Psychiatria	200	005
URO	Urológia	200	012

Tabuľka 10: Vybrané typy ZS (a odbornosti) pre skupinu HYP

Príloha B

Kód	Názov diagnózy
E11	Diabetes mellitus nezávislý od inzulínu
E78	Poruchy metabolizmu lipoproteínov a iné lipidémie
H25	Starecký zákal šošovky - cataracta senilis
H40	Glaukóm
I10	Esenciálna hypertenzia
I11	Hypertenzná choroba srdca
I12	Hypertenzná choroba obličiek
I25	Chronická ischemická choroba srdca
I42	Kardiomyopatie
I48	Predsieňová fibrilácia a flater
I50	Srdcové zlyhanie
I63	Mozgový infarkt
I70	Ateroskleróza
I83	Kŕčové žily dolných končatín
M17	Artróza kolenného kĺbu
M51	Iné poruchy medzistavcových platničiek
M81	Osteoporóza bez patologickej fraktúry
N08	Glomerulárne choroby pri chorobách zatriedených inde
N18	Chronické zlyhanie obličiek
R07	Bolesť v hrdle a v hrudníku
Z95	Prítomnosť srdcových a cievnych implantátov a štepov

Tabuľka 11: Vybrané diagnózy pre skupinu HYP

Kód	**Názov diagnózy**
E86	Dehydratácia, hypovolémia
F00	Demencia pri alzheimerovej chorobe
F01	Vaskulárna demencia
F03	Demencia, bližšie neurčená
F06	Iné duševné poruchy zapríčinené poškodením a dysfunkciou mozgu a somatickou chorobou
F72	Ťažká duševná zaostalosť (mentálna retardácia)
G80	Detské mozgové ochrnutie
G82	Paraplégia a tetraplégia (kvadruplégia)
G93	Iné poruchy mozgu
I50	Srdcové zlyhanie
I63	Mozgový infarkt
I64	Cievna mozgová príhoda bližšie neurčená ako krvácanie alebo infarkt
I67	Iné cievne choroby mozgu
I69	Neskoré následky cievnych mozgových chorôb
L89	Dekubitálny vred - preležanina
N31	Neuromuskulárna dysfunkcia močového mechúra, nezatriedená inde
R26	Poruchy chôdze a mobility (pohyblivosti)
R32	Inkontinencia moču, bližšie neurčená
R33	Retencia moču
R54	Senilita
U99[6]	Inkontinencia

Tabuľka 12: Vybrané diagnózy pre skupinu HYP

[6]Pri tejto diagnóze sme zlúčili konkrétne diagnózy U99.01 - U99.03. Ostatné kódy začínajúce na U99 tam zarátané neboli, lebo sa v dátach nevyskytovali alebo slúžili na iné účely.

Príloha C

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-0.10294	0.013018	-7.90794	2.62E-15
vek_kat_25-29	0.018049	0.015694	1.150057	0.250120789
vek_kat_30-34	0.032279	0.015127	2.133915	0.032850092
vek_kat_35-39	0.086342	0.015063	5.732047	9.93E-09
vek_kat_40-44	0.201748	0.014935	13.50884	1.40E-41
vek_kat_45-49	0.392249	0.015355	25.54532	7.20E-144
vek_kat_50-54	0.684303	0.015479	44.20886	0
vek_kat_55-59	1.079408	0.015658	68.93745	0
vek_kat_60-64	1.47968	0.01581	93.59325	0
vek_kat_65-69	1.87217	0.016774	111.6089	0
vek_kat_70-74	2.303044	0.018935	121.6312	0
vek_kat_75-79	2.498843	0.021573	115.8335	0
vek_kat_80-84	2.587758	0.026478	97.73185	0
vek_kat_85+	2.578929	0.031386	82.16786	0
pohlavie_zena	-0.0729	0.005655	-12.8912	5.09E-38
plati_zanho_stat	0.123585	0.006825	18.10655	2.94E-73
umrtie	-0.04828	0.027473	-1.7573	0.078866611
ADOS	0.220557	0.031298	7.047051	1.83E-12
ANGIO	0.209216	0.019246	10.87052	1.60E-27
ARYT	1.004939	0.06379	15.75394	6.60E-56
CIECHI	0.75225	0.024779	30.3578	2.68E-202
DIABET	1.478976	0.010567	139.962	0
DZS	0.083461	0.01525	5.472992	4.43E-08
GER	-0.1048	0.038152	-2.74682	0.006017704
KARDIO	0.641546	0.011099	57.8013	0
VNUTRO	0.061137	0.008078	7.568709	3.77E-14

Tabuľka 13: Výstup z lineárnej regresie pre skupinu HYP

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
E78	0.043308	0.008683	4.987643	6.11E-07
H40	0.057378	0.011817	4.855569	1.20E-06
I10	1.091471	0.008701	125.4454	0
I11	1.038091	0.011	94.36912	0
I12	0.739865	0.055675	13.28912	2.71E-40
I25	0.554599	0.012375	44.81629	0
I42	1.530373	0.049982	30.61818	9.58E-206
I48	0.473208	0.019796	23.90374	3.14E-126
I50	1.386365	0.02562	54.11177	0
I63	0.483758	0.026935	17.95999	4.16E-72
I70	0.766921	0.017947	42.73234	0
I83	0.103525	0.015252	6.787737	1.14E-11
M17	0.270235	0.013289	20.33557	6.62E-92
M51	0.128749	0.013751	9.362933	7.78E-21
M81	-0.43203	0.014403	-29.9946	1.54E-197
N08	0.855697	0.042591	20.09092	9.39E-90
N18	0.649596	0.031368	20.70905	3.06E-95
R07	-0.14695	0.014012	-10.487	9.96E-26
Z95	0.92823	0.037128	25.00051	6.94E-138

Tabuľka 14: Výstup z lineárnej regresie pre skupinu HYP - pokračovanie

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-0.06209	0.021211	-2.92753	0.003416806
vek_kat_25-29	-0.03788	0.025562	-1.48171	0.138418518
vek_kat_30-34	-0.05871	0.024639	-2.38285	0.017179356
vek_kat_35-39	-0.07485	0.024536	-3.05057	0.002284175
vek_kat_40-44	-0.07875	0.024316	-3.2387	0.001200808
vek_kat_45-49	-0.11453	0.02497	-4.58682	4.50E-06
vek_kat_50-54	-0.14254	0.025113	-5.67615	1.38E-08
vek_kat_55-59	-0.17212	0.025279	-6.80884	9.85E-12
vek_kat_60-64	-0.16466	0.025339	-6.49827	8.13E-11
vek_kat_65-69	-0.14256	0.0267	-5.33938	9.33E-08
vek_kat_70-74	-0.00889	0.03003	-0.2959	0.767309714
vek_kat_75-79	0.31149	0.034327	9.074159	1.15E-19
vek_kat_80-84	1.587235	0.042693	37.17774	3.08E-302
vek_kat_85+	5.48925	0.051149	107.3197	0
pohlavie_zena	0.153034	0.009222	16.59471	7.82E-62
plati_zanho_stat	0.021729	0.011189	1.942107	0.052124598
umrtie	2.133081	0.044875	47.5339	0
URO	0.380082	0.016759	22.67902	7.93E-114
NEU	0.093521	0.013288	7.03817	1.95E-12
GER	5.700409	0.063859	89.26568	0
ADOS	2.446645	0.053533	45.70344	0
DZS	1.173602	0.025014	46.91745	0
PSY	0.412281	0.019706	20.92154	3.65E-97

Tabuľka 15: Výstup z lineárnej regresie pre skupinu B

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
E86	1.790501	0.096607	18.53386	1.15E-76
F00	6.854423	0.08021	85.45606	0
F01	5.087769	0.090303	56.3414	0
F03	1.915534	0.132133	14.49697	1.29E-47
F06	1.760945	0.048288	36.46739	6.83E-291
F72	16.67277	0.203999	81.72951	0
G80	11.95973	0.137936	86.70505	0
G82	8.079368	0.141715	57.01145	0
G93	-0.15679	0.073861	-2.12282	0.033769143
I50	-0.38523	0.0399	-9.65474	4.71E-22
I63	0.988378	0.047697	20.72189	2.35E-95
I64	-0.25946	0.063061	-4.11438	3.88E-05
I69	3.052045	0.048538	62.87946	0
L89	9.893314	0.127067	77.85899	0
N31	1.759514	0.048144	36.54681	3.77E-292
R26	1.267863	0.037318	33.97448	8.51E-253
R32	2.339064	0.126476	18.49416	2.40E-76
R33	1.716213	0.093492	18.35681	3.03E-75
R54	4.197683	0.396929	10.57541	3.89E-26
U99	5.327088	0.035155	151.5293	0

Tabuľka 16: Výstup z lineárnej regresie pre skupinu B - pokračovanie