UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MODELOVANIA POISTNÝCH PODVODOV V ZDRAVOTNOM POISTENÍ

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Vedúci práce:	Mgr. Lenka Petrů

Bratislava 2019 **Bc. Daniel RIZMAN**



19655768

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Daniel Rizman
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: Modelovania poistných podvodov v zdravotnom poistení
Modeling of fraud in health insurance

Anotácia: V prvej časti bude práca obsahovať teoretické informácie o existujúcich prístupoch na odhaľovanie podvodov v zdravotníctve. Druhá časť bude kľúčová pre celú prácu a bude obsahovať návrh a tvorbu vlastného modelu, resp. konštrukciu rôznych modelov a hľadanie optimálneho. Následne sa bude model testovať v praxi a výsledky reálnych kontrol u poskytovateľov poukážu na potenciálne slabé miesta modelu, ktorý sa následne bude dolaďovať.

Vedúci: Mgr. Lenka Petrů
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 24.01.2018

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

.....
študent

.....
vedúci práce

Poďakovanie Rád by som sa poďakoval zdravotnej poisťovni Union za skvelú možnosť pracovať s ich dátami a vzdelávať sa v oblasti zdravotníctva. Kamarátom a kolegom z controllingu za vysvetľovanie všetkých mojich dotazov, revíznym lekárom a revíznym farmaceutom za pomoc pri tvorbe modelov a za spoluprácu pri posudzovaní mojich výsledkov. Špeciálne by som sa rád poďakoval svojej školiteľke Mgr. Lenke Petrů za vedenie, pomoc a rady pri vypracovávaní tejto diplomovej práce.

Abstrakt

RIZMAN, Daniel: MODELOVANIA POISTNÝCH PODVODOV V ZDRAVOTNOM POISTENÍ [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Lenka Petrů, Bratislava, 2019, 60 s.

V tejto práci sa zaoberáme problematikou podvodov v zdravotnom poistení. Prvá časť práce zoznamuje čitateľa s oblasťou zdravotného poistenia, podvodov v zdravotnom poistení a dátami, s ktorými pracuje zdravotná poisťovňa. Uvádzame vo svete najčastejšie používané analytické metódy na identifikáciu týchto podvodov. Druhá časť práce opisuje nami vytvorené modely lineárnej regresie a zhlukovací model DBSCAN. Tieto modely aplikujeme na dáta zdravotnej poisťovne v jazyku R. Ich výsledky verifikujeme pomocou dát o externých revíznych kontrolách a taktiež pomocou revíznych lekárov. Model lineárnej regresie sa napriek svojim obmedzeniam a problémom pri aplikácii ukázal byť ako dobrá pomôcka na identifikáciu podvodníkov. Na aplikáciu zhlukovacieho modelu do rôznych oblastí zdravotníctva je potrebná spolupráca s expertmi z danej oblasti. Po časovo náročnej príprave modelu sa však ukazuje, že to môže byť kvalitná pomôcka na identifikáciu podvodníkov.

Kľúčové slová: Zdravotníctvo, Podvody, Lineárna regresia, Zhlukovanie, DBSCAN

Abstract

RIZMAN, Daniel: Modeling of fraud in health insurance [Master's Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: RNDr. Mgr. Lenka Petrů, Bratislava, 2019, 60 p.

In this work we deal with the issue of health insurance fraud. The first part of the work introduces the reader to the area of health insurance, health insurance fraud and data that the health insurance company works with. We present the most commonly used analytical methods in the world to identify these frauds. The second part of the work describes our linear regression model and our DBSCAN clustering model. We apply these models to data from the health insurance company in programming language R. We verify the results using external revision data as well as by the help of the revision doctors. Despite its limitations, the linear regression model has proven to be a good tool for identifying fraudulent health care providers. In order to apply a clustering model to different areas of health care, collaboration with field experts is needed. However, after time-consuming model preparation, it turns out that clustering model can also be a good tool to identify fraudulent providers.

Keywords: Health care, Fraud, Linear regression, Clustering, DBSCAN

Obsah

Úvod **9**

1 Zdravotníctvo, dáta a podvody **11**

1.1 Odborné pojmy 11

1.2 Dáta 12

1.3 Podvody 13

2 Odhaľovanie podvodov vo svete **17**

2.1 Základné delenie metód 17

2.2 Postupy 18

3 Teória používaných metód **19**

3.1 Stepwise metóda 19

3.2 Outlieri v lineárnej regresii 20

3.3 DBSCAN 21

3.3.1 Definovanie algoritmu 22

3.3.2 Hľadanie zhlukovania 22

3.4 Analýza hlavných komponentov 23

3.4.1 Definícia 23

3.4.2 Hľadanie hlavných komponentov 24

3.4.3 Práca s hlavnými komponentmi 25

4 Praktická časť **27**

4.1 Model lineárnej regresie 27

4.1.1 Prediktory a postup modelovania 29

4.1.2 Výstupy a výsledky 31

4.2 Zhlukovací model 37

4.2.1 Myšlienka a postup modelovania 38

4.2.2 Aplikácia na lekárne 39

4.2.3 Aplikácia na všeobecných lekárov 50

4.3 Zhrnutie modelovania a porovnanie modelov 54

Záver **57**

Zoznam použitej literatúry **59**

Úvod

Náklady na zdravotnú starostlivosť podľa [6] každoročne stúpajú, pričom prognóza uvádza, že oproti roku 2015, kedy náklady vo vyspelých krajinách predstavovali približne 6% HDP, by mali v roku 2060 predstavovať až 14% HDP. Nie všetky peniaze však putujú tam, kam by mali. Podľa [2] a [4] je odhadnutý podiel peňazí, ktoré sú v zdravotníctve minuté z dôvodu podvodov a zneužitia systému od 3 až do 10 percent.

To je signifikantné množstvo peňazí, ktoré zhoršuje úroveň zdravotnej starostlivosti. Detekcia podvodov je dôležitá nielen z hľadiska poisťovne, ktorej sa zvyšujú náklady, ale aj z hľadiska poistenca, ktorý kvôli vyšším nákladom na zdravotnú starostlivosť platí vyššie poistné alebo dostáva menej benefitov. Poistenec môže taktiež byť vystavený procedúram, ktoré v skutočnosti nepotrebuje, alebo záznam o jeho zdravotnom stave môže byť poškodený fiktívnymi výkonmi. To všetko môže v budúcnosti vytvoriť nežiadúce komplikácie.

V súčasnosti zdravotné poisťovne venujú detekcii podvodov čoraz viac a viac pozornosti. Okrem automatických kontrol dát zavedených v systéme poisťovne sa odhaľovaniu podvodov venujú aj revízni lekári. S odhaľovaním podvodov môžu pomôcť aj samotní poistenci, ktorí si môžu ľahko skontrolovať na nich vykázanú zdravotnú starostlivosť.

Položkovitá kontrola revíznym lekárom je časovo veľmi náročná a zdravotná poisťovňa má k dispozícii obrovské množstvo dát, ktoré nie je takýmto spôsobom možné skontrolovať. Automatické kontroly sú založené na situáciách preddefinovaných odborníkom z danej oblasti zdravotníctva. Tým pádom sú však veľmi limitované. Podnety poistencov sú pri odhaľovaní podvodov veľmi užitočné, avšak často sa stáva, že sú nesprávne, napríklad kvôli komplikovanosti zdravotníctva.

Oproti uvedeným spôsobom odhaľovania podvodov majú metódy strojového učenia a dolovania dát mnohé výhody. Dokážu spracovať obrovské množstvo dát v krátkom čase, zároveň dokážu odhaliť aj nové typy podvodov. Preto sa v praxi tento prístup objavuje čoraz častejšie. Prehľad o používaných metódach môžeme nájsť napríklad v článkoch [2], [3].

Podvodmi v zdravotnej starostlivosti sa zaoberá aj diplomová práca [15] z Fakulty matematiky, fyziky a informatiky na Univerzite Komenského. Táto práca sa venuje vy-

lepšovaniu algoritmu na hľadanie poskytovateľov zdravotnej starostlivosti, ktorí medzi sebou zdieľajú kódy poistencov.

Naším cieľom bolo vytvoriť dostatočne všeobecný model, ktorý by vedel identifikovať podozrivých poskytovateľov zdravotnej starostlivosti a overiť správnosť predikcie pomocou revíznych lekárov a kontrol u poskytovateľov. V prvej kapitole sa nachádzajú základné pojmy zo zdravotníctva a prehľad podvodov v zdravotníctve. V druhej kapitole sa nachádza prehľad o používaných metódach na identifikáciu podvodov, ktorý budeme čerpať z [2], [3], [5] a [8]. V tretej kapitole uvedieme matematickú teóriu metód používaných v praktickej časti, ktorú sme získali z [7], [9], [13] a [11]. V štvrtej kapitole použijeme dané metódy na dáta zdravotnej poisťovne Union a popisujeme výsledky jednotlivých modelov. Na analýzu dát používame jazyk R.

1 Zdravotníctvo, dáta a podvody

V tejto kapitole vysvetľujeme niektoré pojmy zo zdravotníctva a poisťovníctva a taktiež opisujeme v akej forme dostáva zdravotná poisťovňa dáta. Tieto informácie sme pritom čerpali priamo od pracovníkov zdravotnej poisťovne, ale aj zo stránky Úradu pre dohľad nad zdravotnou starostlivosťou. Nakoniec vymenovávame niekoľko druhov známych podvodov.

1.1 Odborné pojmy

Pre lepšie pochopenie textu uvádzame niekoľko pojmov zo zdravotníckej terminológie, ale aj pojmy s ktorými pracuje zdravotná poisťovňa:

- **Poskytovateľ zdravotnej starostlivosti (PZS)** - Je to firma alebo živnostník, ktorý má povolenie poskytovať zdravotnú starostlivosť. Poskytovateľ zdravotnej starostlivosti môže vlastniť napríklad dve ambulancie, v ktorých zamestnáva troch lekárov. Okrem firiem prevádzkujúcich ambulancie to však môže byť aj firma prevádzkujúca lekáreň alebo laboratórium.

- **Odbornosť** - Je to špecializácia lekára/PZS. Ako príklad môžeme použiť urológiu, gynekológiu, kardiológiu alebo všeobecné lekárstvo.

- **Vykázaná zdravotná starostlivosť** - Ide o zoznam vykonanej zdravotnej starostlivosti, ktorú PZS poslal poisťovni na preplatenie.

- **Špecializovaná ambulantná starostlivosť (ŠAS)** - Poskytuje ju lekár so špecializáciou (výnimkami sú špecializácie všeobecné lekárstvo a zubné lekárstvo). Až na niekoľko výnimiek sa ŠAS poskytuje na základe odporúčania všeobecného lekára. V ŠAS vo väčšine prípadov platí, že PZS je lekár, ktorý pracuje len v jednej ambulancii. Nie je to však pravidlo. V tejto diplomovej práci budeme, okrem iných dát, analyzovať dáta práve zo ŠAS a VAS - všeobecnej ambulantnej starostlivosti.

- **PCG (Pharmacy-based Cost Group)** - Je to klasifikácia poistencov na základe liekov, ktoré im boli predpísané. Používa sa na roztriedenie poistencov na základe ich nákladov na lieky. Ako príklad môžeme uviesť astmatikov, ktorí by sa nachádzali v jednej skupine a diabetikov, ktorí by sa nachádzali v druhej skupine.

- **Uznaná suma** - Nazývaná aj medicínsky uznaná suma, je suma, ktorú by mal mať PZS vyplatenú, ak by neexistovali limity pre poskytovateľov. Limit môže byť napríklad typu, že PZS môže za isté časové obdobie vykonať len 10 istých druhov výkonov. Ak ich vykonal viac, bude mu preplatených len prvých 10.
- **Účtovná suma** - Ide o reálne preplatenú sumu zdravotnou poisťovňou.
- **Interná revízia** - Ide o kontrolu dát poisťovne odborným revíznym lekárom alebo revíznym farmaceutom. Je to kontrola za počítačom.
- **Externá revízia** - Ide o kontrolu v konkrétnej ambulancii poskytovateľa alebo v konkrétnej lekárni, kde sa kontroluje zdravotná dokumentácia.
- **Finančný efekt** - Výsledok po externej revízii. Množstvo peňazí, ktoré bolo vyplatené neprávom plus zmluvná pokuta.
- **Lieky na výnimku** - Ide o drahé lieky, ktorých úhradu musí schváliť revízny lekár.

1.2 Dáta

Zdravotná poisťovňa dostáva dáta od poskytovateľov zdravotnej starostlivosti, ktorí aspoň raz za mesiac posielajú zoznam všetkých zdravotných výkonov, ktoré za dané časové obdobie vykonali, prípadne zoznam všetkých vydaných liekov a pomôcok, ktoré vydali. Takýto zoznam sa nazýva dávka. Jeden riadok dát, ktoré má zdravotná poisťovňa k dispozícii, teda zodpovedá jednému výkonu/lieku/pomôcke.

Informácie, ktoré musia poskytovatelia ku každej položke v dávke pravdivo dodať, sú presne dané Úradom pre dohľad nad zdravotnou starostlivosťou a dajú sa nájsť na stránke Úradu pod metodickými usmerneniami. Dátové rozhranie pre dávky je však veľmi komplikované pre čitateľa, ktorý nemá predošlé skúsenosti s odbornou zdravotníckou terminológiou. Taktiež neobsahuje všetky informácie, ktoré poisťovňa môže dohľadať na základe informácií, ktoré sa tam nachádzajú. Preto uvádzame prehľad o tom, aké informácie obsahuje jeden riadok dát zdravotnej poisťovne, ale aj o informáciách, ktoré sa k nemu dajú dohľadať:

- **Ambulancia a PZS** - PZS, ktorý poslal dávku. Dá sa dohľadať odbornosť, adresa ambulancie, ale aj či ambulancia má k dispozícii nejaký špeciálny, nadštandardný liečebný prístroj. V prípade, ak sa konkrétny riadok dát týka vydaného lieku alebo pomôcky, tak ide o lekáreň alebo výdajňu zdravotníckych pomôcok.
- **Informácie o výkone/lieku/pomôcke** - Okrem údaju o tom, o aký výkon, liek či pomôcku ide je možné zistiť aj uznanú sumu, reálne hradenú sumu, množstvo vydaného lieku, pomôcky alebo vykonaných výkonov a iné informácie o type liekov, pomôcok a výkonov.
- **Diagnóza** - Diagnózy sa dajú ďalej zoskupovať podľa druhu choroby.
- **Dátum výkonu alebo dátum výdaja a dátum predpisu**
- **Informácie o poistencovi** - Vek, pohlavie, bydlisko, PCG v minulom roku.
- **Informácie o objednávajúcom PZS a ambulancii** - Uvedené v prípade, ak ide o objednaný liek/výkon, pričom niektoré lieky alebo výkony môže predpisovať či objednávať len lekár s istou špecializáciou.
- **Informácie o predpisujúcom PZS a ambulancii** - Uvedené v prípade, ak ide o vydaný liek alebo pomôcku.
- **Informácie o vykonávajúcom, predpisujúcom a objednávajúcom lekárovi**

Okrem dát z jednotlivých dávok má poisťovňa k dispozícii aj dáta z externých revízií. Tieto dáta sa dajú použiť na učenie modelov alebo na validáciu výstupov modelu. V našom prípade však máme k dispozícii veľmi málo takýchto dát, a preto ich budeme používať iba na verifikáciu.

1.3 Podvody

Existuje mnoho druhov podvodov, podľa článkov [1], [2], [3] sa podvody delia napríklad aj podľa toho, kto vykonáva podvod. My sa budeme zaoberať podvodmi na úrovni poskytovateľov zdravotnej starostlivosti.

V prípade zdravotných výkonov existuje niekoľko známych druhov podvodov, ktoré podľa [1], [2], [3] môžeme rozdeliť nasledovne:

- **Fiktívne výkony** - Ide o výkony, ktoré poskytovateľ vykáže, ale v skutočnosti neboli vykonané.
- **Nadhodnocovanie** - Vykazovanie drahšej procedúry namiesto lacnejšej (napr. dlhšia namiesto kratšej procedúry).
- **Duplicitné výkony** - Vykazovanie toho istého výkonu niekoľkokrát, napríklad len s pozmenením dátumu.
- **Zbytočné výkony** - Vykazovanie výkonov, ktoré nie sú potrebné pri danom stave pacienta (diagnóze).
- **Falšovanie diagnózy** - Za účelom odôvodnenia niektorých výkonov.
- **Zamieňanie nehradených výkonov za hradené**
- **Vykazovanie výkonu ako skupina jednoduchších výkonov** - Niektoré výkony sa skladajú z mnohých menších výkonov, ktoré sa dajú vykázať samostatne. Príkladom môže byť skupinová terapia vykázaná ako niekoľko individuálnych terapií.

Okrem vyššie spomínaných druhov podvodov určite existujú aj ďalšie druhy podvodov, ktoré sa v článkoch nespomínajú. Dobrou vlastnosťou niektorých analytických metód na rozdiel od preddefinovaných systémových kontrol je, že sa nesústreďujú na jeden typ podvodu a môžu odhaliť aj nový, doteraz neznámy druh podvodu.

Odlišný prípad nastáva pri predpisovaní a výdaji liekov/zdravotných pomôcok. V prípade podvodného predpisu je nutná istá dohoda medzi predpisovateľom a lekárňou, pretože peniaze z poisťovne neputujú ku predpisovateľovi, ale práve do lekárne. Napriek tomu sa však niektoré druhy podvodov pri výkonoch dajú preniesť aj na tento prípad. Ako príklad môžeme zobrať nadhodnocovanie - lekár predpíše pacientovi drahšiu pomôcku, ale vydaná mu bude lacnejšia. Taktiež môžu nastať aj prípady, kedy podvodníkom je iba lekáreň. Lekáreň môže napríklad namiesto predpísaného lieku alebo pomôcky vydať niečo iné. V tejto práci sa budeme zaoberať podvodným predpisom lekárov, ale aj podvodným výdajom liekov a pomôcok v lekárňach samostatne.

Niekedy môže lekár na základe svojej chyby alebo nevedomosti vykazovať výkon nesprávne alebo inak, ako by sa mal vykazovať. Na chyby pri vykazovaní majú zdravotné

poisťovne automatické kontroly systémom, ktoré môžu vykázaný riadok posúdiť ako chybný a úplne ho odstrániť, môžu riadok odoslať na kontrolu revíznym lekárom alebo môžu posúdiť riadok ako bez chyby. Týmto systémom sa odstránia niektoré z chybne vykázaných riadkov. Niektoré nesprávne vykázané riadky však prejdú systémom bez problémov. V takýchto prípadoch síce nejde o podvod, ale napriek tomu má takáto chyba podobný dopad na poisťovňu a poistenca ako podvod. V tejto práci preto nebudeme chybné prípady a podvody rozlišovať.

Dáta o kontrolách jednotlivých riadkov interným revíznym lekárom má poisťovňa taktiež k dispozícii. V tejto práci sa však nebudeme zaoberať podvodmi na úrovni riadkov, ale celkovými podvodmi na úrovni poskytovateľov, a preto s nimi nebudeme ďalej pracovať. Napriek tomu však ide o dáta, ktoré by mohli byť vhodné na tvorbu rôznych predikčných modelov a zaoberá sa nimi množstvo odborných prác, o čom viac píšeme v ďalšej kapitole.

2 Odhaľovanie podvodov vo svete

V tejto kapitole budeme vychádzať hlavne z článkov [2], [3], [5], [8], ktoré recenzujú množstvo literatúry na tému odhaľovania podvodov v zdravotníctve. Naším cieľom je získať stručný prehľad o metódach spolu s ich výhodami a nevýhodami.

2.1 Základné delenie metód

Uvedieme základné rozdelenie používaných analytických metód:

- metódy bez učiteľa,
- metódy s učiteľom,
- zmiešané metódy.

Učiteľom sa nazýva vzorka dát, ktorá ma preddefinované, či daný riadok je podvod, alebo nie je. Cieľom metód s učiteľom je na preddefinovanej vzorke odhaliť závislosti medzi cieľovou premennou a vysvetľujúcimi premennými.

Zjavnou nevýhodou metód s učiteľom je potreba mať preddefinovanú vzorku dát, ktorá pri mnohých metódach musí mať dostatočnú veľkosť. Tieto metódy sú naopak dobré na hľadanie prípadov, ktoré sú podobné ako podvodné prípady, ktoré poznáme. Taktiež výstupy z týchto metód zvyknú mať dobrú interpretáciu. Príkladom takýchto metód použitých v zdravotníctve sú všetky možné klasifikačné a regresné algoritmy. Podľa článku [3] zo všetkých skúmaných prípadov, kedy boli použité metódy s učiteľom boli neurónové siete použité v 40% a klasifikačné stromy tiež v 40% prípadov.

V článku [3], ktorý vyšiel v roku 2008, sa metódy bez učiteľa moc nepoužívali, viac sa používali zmiešané metódy, o ktorých sa dá dočítať napr. v článku [1]. V neskoršie vydaných článkoch [5], [8] sa však objavujú tieto metódy vo väčšom počte. Ich výhodou je, že nepotrebujú vzorku dát s danou cieľovou premennou. Ďalšou výhodou je, že dokážu objaviť aj nové druhy podvodov. Príkladom takýchto metód môže byť zhlukovanie alebo analýza anomálií.

Zmiešané metódy fungujú väčšinou na princípe, že sa na dáta aplikuje metóda bez učiteľa, ktorá roztriedi dáta do niekoľkých skupín a následne sa použije metóda s učiteľom. Týmto spôsobom sa dokážu odstrániť niektoré nevýhody, ktoré modely

majú, keď sa použijú samostatne. Pre viac informácií o týchto metódach odporúčame článok [1].

2.2 Postupy

Zdravotné poisťovne majú k dispozícii milióny riadkov dát. Počet poistencov sa pritom môže pohybovať tiež rádovo v miliónoch. Takéto obrovské dáta sa často spájajú s pojmom big data. Big data prístup zažíva veľkú popularitu a dostáva sa aj do povedomia poisťovní. Viac o tomto prístupe sa dá dočítať v článku [2].

Jednou z úplne prvých vecí potrebných pred analýzou dát je stanoviť si cieľ, ktorým je v našom prípade identifikácia podozrivých poskytovateľov zdravotnej starostlivosti. Veľká časť literatúry sa však zaoberá identifikáciou konkrétneho podozrivého riadku v zdravotníckych dátach. V takomto prípade existuje veľké množstvo dát z internej revízie, ktorá sa zaoberá kontrolou jednotlivých riadkov. To je dobré pre metódy s učiteľom. Môžu však existovať poskytovatelia, ktorých jednotlivé riadky dát sú úplne v poriadku, ale pri pohľade na poskytovateľa ako celok sa môže objaviť podozrivá vlastnosť, ako je napríklad frekvencia výkonov. Preto v tejto diplomovej práci budeme skúmať PZS ako celok. Jedným z článkov, ktoré využívajú takýto postup je [4].

Väčšina štatistických metód používa dáta vo forme, kde jeden riadok zodpovedá jednému skúmanému objektu, v našom prípade to je poskytovateľ zdravotnej starostlivosti, preto potrebujeme dáta transformovať. O transformácii zdravotníckych dát a ich predspracovaní sa dá dočítať viac v článku [3].

Počet rôznych PZS v špecializovanej ambulantnej starostlivosti a všeobecnej ambulantnej starostlivosti na Slovensku o veľa nepresahuje desaťtisíc. Tento počet nie je veľký, preto budeme môcť využívať klasické štatistické metódy bez problémov.

Na transformovanie potrebujeme najskôr zvoliť časovú jednotku, za ktorú budeme dáta analyzovať. Ďalšou kľúčovou časťou prípravy dát je výber vhodných premenných a aglomeračných funkcií. Výber premenných úzko súvisí s kvalitou výsledného modelu a aj podľa [3] je na ňom potrebné si dať záležať.

Predspracovanie dát nebudeme v tejto práci podrobne opisovať. Zameriame sa hlavne na postupy a výsledky. Musíme však spomenúť, že predspracovanie dát bolo robené v jazyku SQL a pracovali sme priamo s databázou zdravotnej poisťovne Union.

3 Teória používaných metód

V tejto kapitole si vysvetľujeme teóriu niektorých metód použitých v praktickej časti. Nezaoberáme sa pritom základnou teóriou lineárnej regresie a zhlukovacích metód. Čitateľ neoboznámený s touto teóriou, si ju môže doštudovať napríklad v knihe [7] a v článku [9].

3.1 Stepwise metóda

Stepwise metóda je užitočná metóda na výber ideálneho modelu, respektíve na výber správnych premenných pomocou informačných kritérií. Opis tejto metódy preberáme z práce [10], pričom informačné kritériá sú definované ako v knihe [7] a opis algoritmu je prebratý z knihy [11].

Najznámejšie informačné kritérium je AIC (Akaike information criterion), ktorého definícia je

$$AIC = -2l(\hat{\beta}) + 2kd,$$

kde $l(\hat{\beta})$ je maximum vierohodnostnej funkcie modelu a d je dimenzia vektoru parametrov modelu β. k je pritom parameter, ktorý má pre AIC hodnotu 1. Je ním však možné ovplyvňovať citlivosť kritéria na počet parametrov modelu. Čím je hodnota tohto kritéria nižšia, tým je model lepší.

Stepwise metóda má tri nasledovné varianty:

- dopredná,
- spätná,
- obojsmerná.

Dopredná metóda začína s modelom, ktorý má iba absolútny člen. Pre tento základný model sa vypočíta hodnota informačného kritéria. Metóda hľadá premennú, ktorej pridanie zaznamená najväčší pokles informačného kritéria. Pokračuje sa pridávaním ďalšej premennej. Metóda skončí vtedy, keď pridanie hociktorej premennej už nie je štatisticky významné.

Spätná metóda začína s modelom obsahujúcim všetky premenné a odoberá ich analogicky s dopredným variantom, až kým zlepšenie kritéria pri odobratí hociktorej premennej nie je štatisticky významné.

Obojsmerná metóda začína s modelom obsahujúcim všetky premenné, ale v každom kroku testuje aj odobratie, aj pridanie premennej do modelu.

3.2 Outlieri v lineárnej regresii

V tejto podkapitole preberáme teóriu z knihy [13] a z kurzu [14]. Časť teórie preberáme aj z práce [10].

Outlier v lineárnej regresii je pozorovanie, ktoré má vzhľadom na hodnoty svojich prediktorov veľmi nezvyčajnú hodnotu cieľovej premennej. Pokiaľ má pozorovanie veľké rezíduum, je veľká pravdepodobnosť, že je to outlier. Môže však existovať pozorovanie, ktoré má veľký vplyv na odhadnuté parametre lineárneho modelu a vďaka tomu má nízke rezíduum. Preto sa na odhaľovanie outlierov používajú studentizované rezíduá, ktoré takéto vlastnosti odstraňujú.

Nech je daná rovnica:

$$\hat{Y_j} = \sum_{i=1}^{n} h_{ij} Y_i,$$

kde $\hat{Y_j}$ je j-ta modelom predpovedaná hodnota a Y_i je cieľová premenná prislúchajúca i-temu pozorovaniu. Potom h_{ij} určuje prínos pozorovania Y_i ku predpovedanej hodnote $\hat{Y_j}$ a hodnoty so strieškou sú definované ako

$$h_i = \sum_{i=1}^{n} h_{ij}^2.$$

Studentizované rezíduá E_i^* potom môžeme definovať nasledovne:

$$E_i^* = \frac{E_i}{S_E \sqrt{1 - h_i}}.$$

Pritom E_i je rezíduum a S_E^2 je odhad variancie rezíduí z modelu zostrojeného bez i-teho pozorovania. Rozdelenie studentizovaných rezíduí je podľa [13] studentovo s $n - p - 2$ stupňami voľnosti. n je pritom počet pozorovaní v modeli a p je počet prediktorov (bez absolútneho člena).

Na testovanie outlierov sa dá použiť takzvaný model s posunutým priemerom (mean shift model), ktorý vyzerá nasledovne:

$$Y = X\beta + D_i\Theta + \epsilon,$$
$$E(\epsilon) = 0,$$
$$Var(\epsilon) = \sigma^2 I.$$

D_i je premenná, ktorá má jednotku pri i-tom pozorovaní a všade inde má nuly. Nenulové Θ následne implikuje, že i-te pozorovanie je outlier. Dá sa ukázať [13], že testovacia štatistika pre test signifikantnosti parametra Θ je rovná studentizovanému rezíduu.

V štandardnom prípade nepoznáme dopredu kandidátov na outlierov, a preto je potrebné použiť tento test na každé pozorovanie. V takomto prípade pri použití veľkého množstva testov rastie pravdepodobnosť chyby prvého druhu a preto je vhodné použiť Bonferroniho korekciu. Bonferroniho korekcia používa namiesto klasickej hladiny testu α upravenú hladinu testu α/n.

3.3 DBSCAN

V tejto podkapitole preberáme teóriu z článku [9]. DBSCAN (Density based spatial clustering of applications with noise), alebo priestorové zhlukovanie založené na hustote v aplikáciách so šumom, je zhlukovací algoritmus. Jeho výhodami oproti bežným zhlukovacím algoritmom je okrem jeho efektívnosti na databázach väčších rozmerov aj jeho schopnosť odhaliť zhluky nekonvexného tvaru. Okrem toho berie do úvahy, že v dátach sa môžu nachádzať outlieri, inak povedané objekty, ktoré nepatria do žiadneho zhluku.

Myšlienkou tohto algoritmu je, že v zhlukoch je hustota bodov väčšia ako mimo zhluku. Vďaka tomu, že tento algoritmus identifikuje zhluky na základe hustoty bodov, vie odhaliť aj zhluky nekonvexného tvaru. Navyše vie označiť body, ktoré nemajú vo svojom okolí dostatočný počet iných bodov ako outlierov. Pri vhodných dátach môže byť takáto identifikácia outlierov užitočná aj pri odhaľovaní podvodov, keď outlierov označíme za podozrivých.

3.3.1 Definovanie algoritmu

Teraz formálne opíšeme metódu DBSCAN, podobne ako je to v článku [9].

Eps okolie bodu p je definované nasledovne: Nech platí $N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$. Potom $N_{Eps}(p)$ je okolie bodu p. Pritom D je databáza (množina) bodov, ktoré je potrebné zhlukovať.

Nech $Minpts$ je prirodzené číslo. Bod, ktorý má vo svojom *Eps* okolí aspoň $Minpts$ bodov sa nazýva jadrový bod.

Nech je dané $Minpts \in N$ a $Eps \in R^+$. Potom môžeme definovať:

- Bod p je priamo dosiahnuteľný z bodu q, ak $p \in N_{Eps}(q)$ a zároveň q je jadrový bod.
- Bod p je dosiahnuteľný z bodu q, ak existuje postupnosť bodov $p_1, p_2, ..., p_n$ taká, že $p_1 = p$ a $p_n = q$, pričom platí, že p_i a p_{i+1} sú navzájom priamo dosiahnuteľné.
- Body p a q sú prepojené, ak existuje bod o taký, že p aj q sú dosiahnuteľné z bodu o.

Zhluk C je potom vzhľadom na $Minpts$ a *Eps* taká neprázdna množina bodov z D, že pre všetky body p, q platí:

- Ak $p \in C$ a q je dosiahnuteľné z p, tak potom aj $q \in C$.
- Pre všetky body $p, q \in C$ platí, že p, q sú prepojené.

Ak bod nepatrí do žiadneho zhluku, tak je označený za šum.

3.3.2 Hľadanie zhlukovania

Hľadanie zhlukovania sa vykonáva v dvoch krokoch. V prvom kroku je potrebné nájsť nejaký bod, ktorý je jadrový. V druhom kroku sa zoberú všetky body, ktoré sú dosiahnuteľné z daného bodu. Všetky tieto body sú zaradené do jedného zhluku. Čitateľa, ktorého viac zaujíma tento algoritmus odkazujeme na článok [9].

Problémom zostáva nájsť ideálne parametre $Minpts$ a *Eps*. Ideálne parametre môžu byť rozdielne pre každý zhluk, preto by sme potrebovali nájsť ideálne parametre pre zhluk, ktorý má najmenšiu hustotu bodov. Najlepší prípad nastáva, keď parametre môžeme

stanoviť priamo z kontextu danej aplikácie. Keď sa však takýmto spôsobom parametre nepodarí stanoviť, potrebujeme mať aj iné možnosti. Preto existujú jednoduché a používané heuristiky:

Označme d vzdialenosť bodu p od k-teho najbližšieho suseda bodu p (vzdialenosť od k-teho najbližšieho bodu k bodu p). Potom d okolie bodu p obsahuje práve $k + 1$ bodov. Hraničné prípady pritom neberieme do úvahy. Ak zmeníme k, tak vzdialenosť d pre bod patriaci do nejakého zhluku sa veľmi nezmení. Veľký nárast by bol v prípade, ak by zhluk bol podobný rovnej čiare, čo je nepravdepodobný prípad.

Nech je dané k, pre všetky body z D sa vypočíta vzdialenosť d. Tieto vzdialenosti sa zoradia a uložia sa do grafu. Alternatívou môže byť okrem týchto vzdialeností počítať aj všetky vzdialenosti ku k najbližším susedom. Takýto graf sa nazýva graf vzdialeností k najbližších susedov (KNN distance plot).

Heuristika sa snaží nájsť bod p, ktorý zodpovedá „lakťu“ grafu. Za Eps sa potom zoberie $d(p)$ a za parameter $Minpts$ sa zoberie k. Body s vyššou vzdialenosťou d sú potom považované za šum.

Táto heuristika nám pomôže nájsť ideálny parameter Eps pre zvolené k alebo $Minpts$. Parameter $Minpts$ sa väčšinou určuje ďalšími heuristikami, ktoré určujú tento parameter v závislosti od dimenzie. Podľa [9] pre dimenziu 2 stačí za tento parameter zobrať číslo 4. Vo všeobecnosti táto heuristika hovorí, že je vhodné za parameter $Minpts$ zobrať dvojnásobok dimenzie.

3.4 Analýza hlavných komponentov

Analýza hlavných komponentov (PCA - Principal component analysis) je metóda na redukciu dimenzie, ktorá z niekoľkých korelovaných premenných vyrába novú sadu takých nekorelovaných premenných, že aj ich malé množstvo vysvetľuje veľkú časť variability v dátach. V tejto práci budeme pomocou tejto metódy vizualizovať mnohorozmerné dáta. Teóriu ohľadom analýzy hlavných komponentov sme prebrali z knihy [12].

3.4.1 Definícia

Nech x je vektor p náhodných premenných a α_1 je vektor p konštánt. V prvom kroku analýzy hlavných komponentov hľadáme lineárnu funkciu $\alpha_1^T x$ závislú od x s ma-

ximálnou varianciou. V druhom kroku hľadáme lineárnu funkciu $\alpha_2^T x$ nekorelovanú s $\alpha_1^T x$ a s čo najväčšou varianciou. V kroku k je potrebné nájsť funkciu $\alpha_k^T x$ nekorelovanú s $\alpha_1^T x, ..., \alpha_{k-1}^T x$ a s čo najväčšou varianciou. Postupuje sa analogicky, až kým sa nenájde p hlavných komponentov $\alpha_1^T x, ..., \alpha_p^T x$.

3.4.2 Hľadanie hlavných komponentov

Nech Σ predstavuje kovariančnú maticu vektora náhodných premenných x. Potom k-ty hlavný komponent nájdeme ako $z_k = \alpha_k^T x$, kde α_k je vlastný vektor matice Σ zodpovedajúci k-tej najväčšej vlastnej hodnote λ_k. Naviac ak α_k má jednotkovú dĺžku, tak potom variancia z_k je rovná λ_k.

Hľadajme najskôr prvý hlavný komponent $\alpha_1^T x$, teda hľadajme α_1, ktoré maximalizuje $var(\alpha_1^T x) = \alpha_1^T \Sigma \alpha_1$. Je vidno, že riešenie tejto úlohy nie je konečné, preto sa aplikuje normalizačná podmienka $\alpha_1^T \alpha_1 = 1$. Túto úlohu je možné riešiť pomocou Lagrangeových multiplikátorov. Lagrangeova funkcia $\mathscr{L}$ vyzerá nasledovne:

$$\mathscr{L} = \alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1).$$

Pritom λ je Lagrangeov multiplikátor. Derivácia Lagrangeovej funkcie podľa α_1 sa musí rovnať nule, a preto platí:

$$\Sigma \alpha_1 = \lambda \alpha_1.$$

Z toho vyplýva, že λ je vlastná hodnota a α_1 je vlastný vektor matice Σ. Maximalizujeme pritom $\alpha_1^T \Sigma \alpha_1 = \lambda \alpha_1^T \alpha_1$, a preto musí byť λ najväčšia vlastná hodnota a α_1 jej zodpovedajúci vlastný vektor.

Teraz uvedieme dôkaz aj pre druhý hlavný komponent, pre každý ďalší hlavný komponent je dôkaz podobný. Maximalizujeme $\alpha_2^T \Sigma \alpha_2$, pričom α_1 a α_2 sú nekorelované, teda $cov(\alpha_1^T x, \alpha_2^T x) = 0$. Platí:

$$cov(\alpha_1^T x, \alpha_2^T x) = \alpha_1^T \Sigma \alpha_2 = \lambda \alpha_1^T \alpha_2 = 0.$$

Namiesto podmienky na nekorelovanosť je teda možné zobrať rovnicu $\alpha_1^T \alpha_2 = 0$. Opäť je potrebné pridať aj normalizačnú podmienku. Lagrangeova funkcia potom vyzerá nasledovne:

$$\mathscr{L} = \alpha_2^T \Sigma \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi \alpha_2^T \alpha_1.$$

Pritom λ a ϕ sú Lagrangeove multiplikátory. Derivovaním Lagrangeovej funkcie podľa α_2 a jej vynásobením α_1^T sprava dostaneme nasledujúcu rovnicu:

$$\alpha_1^T \Sigma \alpha_2 - \lambda \alpha_1^T \alpha_2 - \phi \alpha_1^T \alpha_1 = 0.$$

Preto $\phi = 0$. Dosadením naspäť do derivácie Lagrangeovej funkcie dostávame $\Sigma\alpha_2 = \lambda\alpha_2$, a teda λ je vlastná hodnota a α_2 je jej zodpovedajúci vlastný vektor. Takisto ako v prvom kroku maximalizujeme $\alpha_2^T \Sigma \alpha_2$, a preto musí byť λ čo najväčšia. Pokiaľ Σ nemá rovnaké vlastné hodnoty, čo je výnimočný prípad, ktorý v realite nikdy nenastáva, tak ide o druhú najväčšiu vlastnú hodnotu.

3.4.3 Práca s hlavnými komponentmi

Po vytvorení hlavných komponentov je dobré poznať nejakú mieru ich kvality. Najpoužívanejšou mierou kvality je podiel vysvetlenej variancie P. Ide o podiel variancie hlavných komponentov, ktoré používame a celkovej variancie v dátach. Ak používame k hlavných komponentov, tak podiel vysvetlenej variancie je:

$$P_k = \frac{\lambda_1 + \lambda_2 +, ..., +\lambda_k}{\lambda_1 + \lambda_2 +, ..., +\lambda_p}$$

4 Praktická časť

V tejto kapitole analyzujeme dáta zdravotnej poisťovne Union v jazyku R. Príprava dát bola urobená pomocou jazyka SQL.

Najskôr opisujeme model lineárnej regresie, ktorý sme vytvorili pomocou stepwise algoritmu. Jeho výstupy sme kontrolovali spolu s revíznymi lekármi. Po potvrdení podozrivého vykazovania boli jednotliví PZS poslaní na externú revíziu, ktorá definitívne potvrdila alebo vyvrátila podozrenie.

Ďalej opisujeme zhlukovací model DBSCAN založený na identifikátoroch podvodu. Jeho výsledky vykresľujeme pomocou analýzy hlavných komponentov a verifikujeme pomocou dát o externých revíziách.

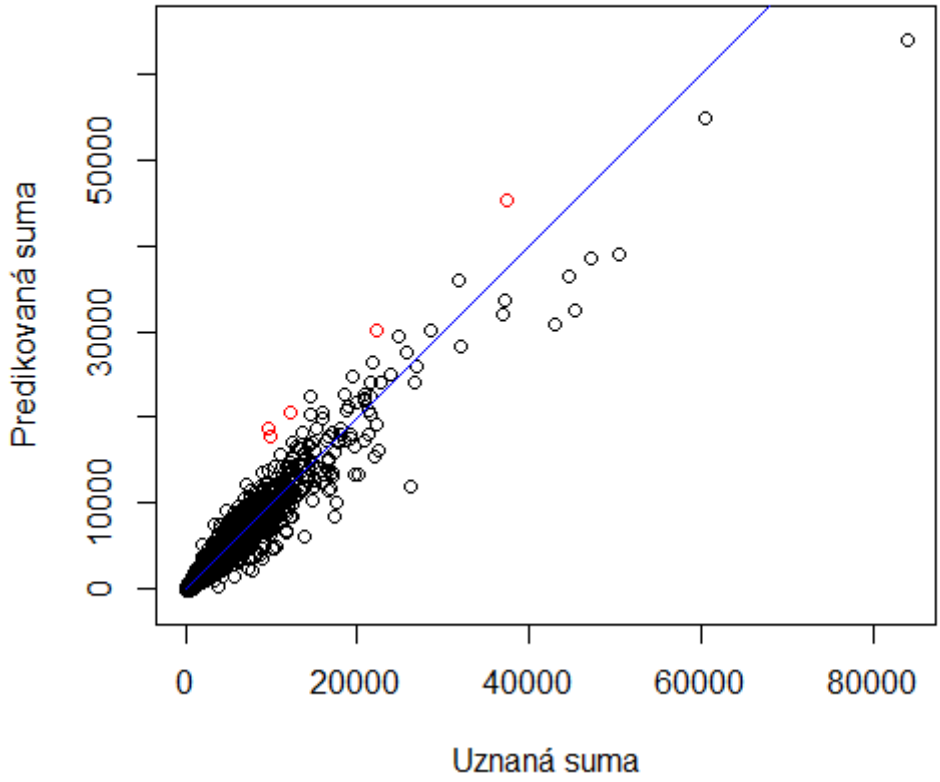
4.1 Model lineárnej regresie

Myšlienka tohto modelu je jednoduchá, ako cieľ sme si zvolili vytvoriť jeden model, ktorý by vedel pokryť výdavky čo najväčšej časti špecializovanej a všeobecnej zdravotnej starostlivosti. Jedným modelom sme chceli pokryť aj podvodné vykazovanie výkonov a aj nadmerný predpis poskytovateľov. Pri výkonoch sme predpovedali medicínsky uznanú sumu poskytovateľov za časové obdobie jedného roka. Medicínsky uznanú sumu sme použili preto, že výkony, ktoré neboli preplatené kvôli limitom sa napriek tomu objavujú vo vykázanej zdravotnej starostlivosti a navyšujú tým pádom hodnoty prediktorov. Pri predpísaných liekoch a pomôckach sme predpovedali reálne hradenú sumu.

Časové obdobie jedného roka sme zvolili preto, lebo je to dostatočne dlhé obdobie na to, aby sa podvodné správanie štatisticky prejavilo a zároveň je to dostatočne krátke obdobie na to, aby analýzu príliš neovplyvňovali zmeny súvisiace s PZS (napríklad zrušenie poskytovateľa, zakúpenie nových prístrojov, otvorenie novej ambulancie atď.). Pri výkonoch pritom ide približne o 9 miliónov a pri predpise o 3,5 milióna riadkov zdravotnej starostlivosti.

Po odhadnutí parametrov modelu sme sa pozreli na rezíduá, ktoré sú pri lineárnej regresii rovné reálnej uznanej sume mínus sume odhadnutej modelom. Tento rozdiel je v teórii odhad sumy, ktorú má poskytovateľ navyše od priemerného poskytovateľa

(poskytovateľa na regresnej priamke), teda v podstate ide aj o istý odhad veľkosti podvodu. Pre ilustráciu postupu sa môžeme pozrieť na obrázok 4.1, na ktorom je vykreslená uznaná suma oproti predikovanej sume modelu z jednej odbornosti. Červenou farbou je znázornených prvých 5 PZS s najväčším rezíduom. Rezíduá nám síce neurčili presne podozrivých poskytovateľov, ale ich zoradením sme dostali zoznam poskytovateľov od najpodozrivejších po najmenej podozrivých, čo je ešte viac užitočné.



Obr. 4.1: Predikovaná vs. uznaná suma

Lineárna regresia má aj rôzne spôsoby a testy, ako odhaľovať outlierov. Na jeden z týchto testov sme sa taktiež stručne pozreli, ale v prvom rade sme testovali metódu s rezíduami. Dôvodmi sú hlavne odhad veľkosti podvodu a jednoduchosť modelu. Jednoduchosť modelu môže v zdravotnej poisťovni mať veľkú výhodu kvôli interpretácii výstupov a výsledkov.

Podstatnú časť prediktorov v lineárnej regresii tvorili počty pacientov, ktorí daného

poskytovateľa navštívili v rôznych skupinách, ako sú demografické skupiny alebo skupiny diagnóz. Myšlienkou fungovania tohto modelu teda je, že na liečenie ľudí v rovnakých vekových skupinách s rovnakými diagnózami by mali mať lekári v rovnakej odbornosti podobné náklady. Lekár, ktorý má vyššie náklady na liečenie istého druhu diagnóz by mal mať vo výsledku veľké rezíduum.

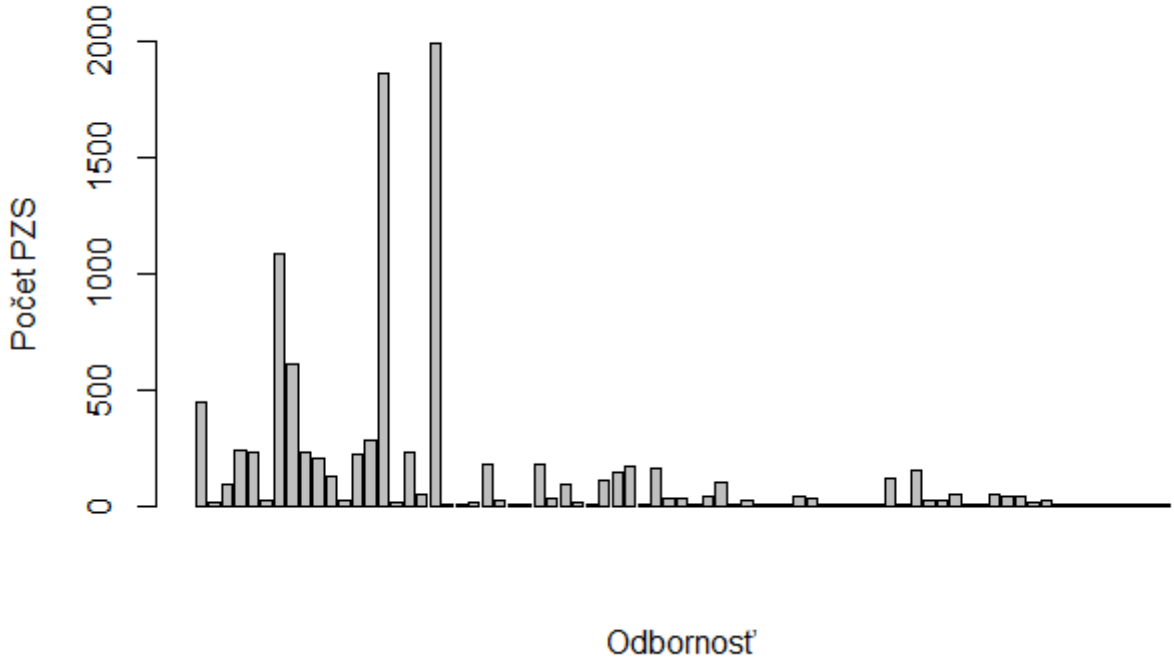
4.1.1 Prediktory a postup modelovania

Ako je spomenuté vyššie, poskytovatelia by mali mať v tej istej odbornosti približne rovnaké náklady na pacienta s istou diagnózou, vekom atď. To však neplatí medzi rôznymi odbornosťami. Preto máme pre každú odbornosť rovnaký základ modelu (prediktory), ale s rôznymi odhadnutými parametrami (a aj rôznymi vybratými premennými). Existujú však PZS, ktorí majú dve a viac odborností. Jeden riadok v našich dátach preto zodpovedá jednej konkrétnej odbornosti daného poskytovateľa, vo väčšine prípadov to zodpovedá jednej ambulancii, v menej častých prípadoch to môže zodpovedať aj viacerým ambulanciám.

Pri takomto postupe sme narazili na problém, pretože niektoré odbornosti nemajú dostatok poskytovateľov na odhadnutie modelu. Počet rôznych poskytovateľov za roky 2016 a 2017 je 10154 a 10160, pričom celkový počet prediktorov je 354. Na obrázku 4.2 sa nachádza stĺpcový graf počtu poskytovateľov v jednotlivých odbornostiach za rok 2017.

Nízky počet PZS v niektorých odbornostiach nám znemožňuje použiť lineárnu regresiu v prípadoch, keď je počet prediktorov väčší ako počet pozorovaní. Niektoré prediktory sú však nulové, pretože napríklad u detského lekára sa neobjaví 60-ročný pacient. Takéto premenné môžeme pre jednotlivé odbornosti odstrániť.

Odstránili sme aj premenné, pri ktorých platí, že suma dát v danom stĺpci je menšia ako počet poskytovateľov v danej odbornosti vydelených 25 (väčšina premenných je počet poistencov s istou vlastnosťou, ktorí navštívili daného poskytovateľa, pri ostatných premenných táto podmienka nikdy neplatila). Takéto premenné by takmer úplne odstránili niektorých poskytovateľov z vysokej priečky v zoradenom zozname poskytovateľov podľa veľkosti rezídua. Pridaním takejto premennej by model získal parameter, ktorý by sa mohol optimálne nastaviť tak, aby minimalizoval štvorce rezíduí



Obr. 4.2: Počet PZS v odbornostiach

len pre niekoľkých poskytovateľov. Ak by to bol napríklad len jeden poskytovateľ, tak by mal automaticky rezíduum nula, pretože parameter prislúchajúci danej premennej by sa mohol ľubovoľne nastaviť. Príčinou by pritom bolo len malé množstvo pacientov so špecifickou diagnózou na odbornosť, ktorej liečba dokonca nemusí byť ani veľmi nákladná.

Aj napriek odstráneniu niektorých prediktorov však zostal vo viacerých prípadoch (odbornostiach) problém s príliš veľkým počtom premenných. Potrebovali sme preto spôsob, ako vybrať najdôležitejšie premenné do modelu. Z rôznych metód na výber premenných sme si vybrali metódu stepwise. Táto metóda nám umožnila meniť parameter k, ktorý mení toleranciu k počtu premenných v modeli. Tým pádom sme dokázali vytvoriť model aj pre nízky počet pozorovaní. Vždy je však potrebné myslieť na to, že zvyšovaním parametra k sa do modelu nemuseli dostať premenné, ktoré výrazne ovplyvňujú výsledok. Tým pádom čím je parameter k vyšší, tým môžu byť výsledky menej presné a mohli byť odhalení nesprávni PZS. Taktiež čím máme väčší počet pozorovaní, tým môžeme modelu viac veriť a pravdepodobne by mal mať lepšie

výsledky.

Stepwise metóda sa v jazyku R aplikuje jednoducho príkazom `step()`. Vstupmi sú variant metódy, inicializačný model a parameter k. Ako variant sme v našom prípade zvolili doprednú metódu. Ostatné metódy by sme v niektorých prípadoch nemohli použiť, pretože by sme sa snažili odhadnúť model s viac premennými ako pozorovaniami. Dopredná metóda má taktiež výhodu v rýchlosti algoritmu, pretože jej stačí odhadovať modely s menším počtom premenných. Za inicializačný model sme tým pádom zobrali jednoduchý model, ktorý predikuje iba na základe absolútneho člena. V jazyku R sme ho vytvorili pomocou príkazu `lm()`.

Ďalšou príčinou, prečo sme potrebovali dobrý a rýchly spôsob na výber vhodných premenných je, že pre jednu odbornosť sme predpovedali na základe rovnakých premenných až 2 modely, jeden pre zdravotné výkony a druhý pre predpis. Niektoré premenné sú pritom dôležitejšie pri predpise ako pri zdravotných výkonoch a naopak.

Alternatívou k výberu vhodných premenných by mohla byť niektorá z metód na redukciu dimenzie, ako je napríklad analýza hlavných komponentov. V takomto prípade by sa však výrazne zhoršila interpretovateľnosť modelu, a preto sme sa rozhodli pracovať len s pôvodnými premennými.

V tabuľke 4.1 opisujeme, na základe akých premenných sme predikovali:

Z uvedeného postupu je zrejmé, že na predikciu by bolo možné použiť aj iný model, ako je lineárna regresia. Na dáta sme preto vyskúšali použiť jednovrstvové neurónové siete. Na výber premenných do neurónových sietí sme použili boruta algoritmus, ktorý vyberá vhodné premenné na základe dôležitosti premenných získaných z klasifikačných stromov. Čitateľa, ktorého by tento postup viac zaujímal odkazujeme na článok [16]. Neurónové siete však dokázali aj na základe menšieho počtu premenných vysvetliť takmer všetku varianciu v dátach. Modely teda boli pretrénované a na našu aplikáciu boli tým pádom nepoužiteľné. Neurónové siete taktiež majú mnohé nevýhody, ako je napríklad dĺžka trénovania modelov a taktiež to, že ide o čiernu skrinku.

4.1.2 Výstupy a výsledky

V tejto podkapitole štatisticky zhŕňame niektoré výstupy a výsledky ohľadom modelu lineárnej regresie.

Tabuľka 4.1: Prehľad premenných v modeli lineárnej regresie

Premenné	Typ premennej	Počet premenných	Popis
Prístroj	faktor	1	1-PZS nemá prístroj, 2-PZS má prístroj, 3-PZS má prístroj lepšej úrovne
Skupina diagnóz	číselná	265	Počet poistencov, ktorí navštívili PZS v danej skupine diagnóz
PCG	číselná	24	Počet poistencov, ktorí navštívili PZS v danej PCG skupine
Geografické	číselné	2	Informácie o rozložení a počte poistencov v okrese PZS
Demografické	číselná	62	Pohlavie, veková skupina a typ platiteľa poistného (platí štát alebo osoba)

Model bol aplikovaný na dáta za roky 2016 a 2017. Získali sme tým pádom dve sady výsledkov. Modely za jednotlivé roky sa v štatistikách veľmi nelíšili, pre jednoduchosť bližšie opisujeme hlavne štatistiky modelov na novších dátach. Model sa nám za rok 2017 podarilo natrénovať pri výkonoch pre 28 odborností a pri predpise pre 27 odborností.

Výsledky z jedného roku boli pritom veľmi podobné výsledkom z druhého roku (modely z rôznych rokov často predpovedali rovnakých PZS na prvých priečkach podľa veľkosti rezíduí). Zo zoznamu poskytovateľov zoradených podľa veľkosti rezíduí z modelu pre výkony z roku 2016 sme zobrali prvých 10 percent poskytovateľov pre každú odbornosť a porovnali sme ich s rovnako vybratými poskytovateľmi z roku 2017. Týmto spôsobom sme dostali zhodu 40%. Pri predpise bola táto zhoda 39%. Keď zoberieme do úvahy, že kvalita modelov bola rôzna (niektoré neboli vôbec použité), tak to je celkom veľké číslo. To nám naznačuje, že model naozaj poukazuje na nejaké štatistické výchylky a nevyberá poskytovateľov náhodne.

V tabuľke 4.2 uvádzame štatistiky o koeficiente determinácie R^2 a upravenom ko-

eficiente determinácie $\bar{R}^2$ modelov pre rok 2017.

Tabuľka 4.2: Prehľad o koeficientoch determinácie

Model	Koeficient	Min.	Priemer	Max.
Výkony	R^2	0,8076	0,9369	0,9987
Výkony	$\bar{R}^2$	0,7928	0,9189	0,9977
Predpis	R^2	0,6728	0,9009	0,9972
Predpis	$\bar{R}^2$	0,6436	0,8688	0,9950

Z tabuľky je možné vidieť, že koeficienty determinácie sú vcelku vysoké, to znamená, že modely dokážu vysvetliť veľkú časť variancie v dátach. To, že sú modely dobré nám potvrdzuje aj test signifikancie modelu. Nulovú hypotézu, že vektor koeficientov modelu $\beta = 0$, zamietame pri každom z uvádzaných modelov. Môžeme si všimnúť, že pri predpise sú priemery koeficientov determinácie o niečo nižšie ako pri výkonoch. V tomto prípade sa v dátach nachádzajú mnohé záznamy, ktoré dokážu výrazne ovplyvniť výsledok. Napríklad pacient, ktorý poberá špeciálne potraviny, lieky alebo potrebuje isté pomôcky môže mať ročné náklady vyššie, ako 2000 eur. Priemerné náklady sú pritom omnoho nižšie. Preto je v tomto prípade ťažšie vytvoriť dobrý model. Taktiež si pri predpise a aj pri výkonoch môžeme všimnúť až podozrivo vysokú maximálnu hodnotu R^2. V prípade predpisu išlo o odbornosť, kde je predpisovanie liekov a pomôcok zriedkavé a preto väčšina hodnôt v cieľovej premennej bola nula. Podobne sa dá vysvetliť veľké R^2 pri výkonoch. V tomto prípade ide o odbornosť, kde je hradených len málo výkonov. Posledná vec, ktorú si môžeme všimnúť, sú rozdiely medzi R^2 a $\bar{R}^2$. Tie nám naznačujú, že niektoré z modelov majú veľký počet prediktorov, rozdiely však nie sú príliš veľké a preto sa nimi nemusíme ďalej zaoberať.

Upozorniť by sme chceli na to, že už po kontrole internou revíziou by sme mali byť schopní povedať, či model predpovedal dobre, alebo zle. Ak sa po kontrole odborníkom nájde v dátach niečo podozrivé, tak to znamená, že model našiel nejakú štatistickú odchýlku nevysvetliteľnú z dát, čo je dobré. Na druhú stranu táto odchýlka môže byť v reáli vysvetliteľná, takáto vec už však zrejme z dát nie je bez reálnej kontroly predpovedateľná. Ak sa po kontrole odborníkom nenájde nič podozrivé, znamená to, že v dátach nám chýba nejaký prediktor.

Pri kontrole internou revíziou sme kontrolovali taktiež početnosť jednotlivých výkonov alebo predpísaných liekov na pacienta v porovnaní s ostatnými PZS. Takéto štatistiky by mali mať vysokú váhu v určovaní správnosti predpovedania modelu. Môžeme povedať, že vo veľkej väčšine kontrolovaných prípadov sa PZS nachádzali v týchto štatistikách na prvých pozíciách. To však nemusí ešte znamenať, že PZS je podvodník, môže to znamenať aj presný opak, že si svoju prácu robí poctivo.

Výber podozrivých PZS prebiehal v troch sadách a s nasledovnými zásadami:

- Vybraní boli poskytovatelia, ktorí sa nachádzajú v prvých desiatich percentách v zozname poskytovateľov zoradenom podľa veľkosti rezídua. Prednostne boli pritom vyberaní poskytovatelia na prvých priečkach.
- Vybraní PZS museli mať veľký potenciál šetrenia peňazí, čo znamená, že mali dostatočne veľké rezíduum a zároveň dostatočne veľkú uhradenú sumu (aspoň 5000 eur za rok). Podmienka na veľkosť rezídua pritom závisí aj od uhradenej sumy. Ak máme napríklad PZS so sumou uhradenou za rok 80000 eur a má rezíduum 2000 eur, tak to nie je tak zaujímavé, ako PZS s uhradenou sumou 5000 eur a rovnakým rezíduom.
- Vybraní PZS neboli za posledných 5 rokov kontrolovaní externou revíziou.
- Pri PZS v druhej a hlavne tretej sade sa brali do úvahy poznatky, ktoré sa zistili predtým. Tieto poznatky podrobnejšie opisujeme neskôr v tejto práci.

Výsledky kontrol internou revíziou pre model na výkony sa nachádza v tabuľke 4.3.

Tabuľka 4.3: Kontrola výkonov internou revíziou

Sada	Počet kontrolovaných PZS	Počet podozrivých PZS	Úspešnosť
1	12	5	42%
2	16	7	44%
3	11	6	55%

Môžeme si všimnúť, že úspešnosť predikcie sa postupne so zvyšujúcou sa sadou zlepšovala. Hodnotíme to na dosť malej vzorke dát, čo je potrebné brať do úvahy. Vzorka dát je taká malá kvôli časovej náročnosti kontroly poskytovateľov. V tretej

sade dát sa úspešnosť vyšplhala až na 55%. Ak zoberieme do úvahy, že vybraní PZS neboli predtým kontrolovaní externou revíziou (niektorí predtým kontrolovaní PZS sa objavovali na prvých priečkach modelov) a ide o model bez učiteľa, ktorý dokáže odhaliť aj úplne nové druhy podvodov, tak je to dobrá úspešnosť. Určite k nej prispeli aj poznatky zistené z predošlých kontrol, preto ich tu podrobnejšie opisujeme. Tieto poznatky by sa do budúcnosti mohli dať buď zakomponovať do modelu alebo aplikovať na výstupy modelu.

Prvým poznatkom bolo, že v niektorých odbornostiach existujú výkony, ktoré vykonáva len malé percento poskytovateľov. V takýchto prípadoch väčšinou išlo o nadštandardnú liečbu alebo liečbu, ktorá sa dá vykonávať iba nejakým špeciálnym prístrojom. Takýto poskytovateľ má oproti klasickému poskytovateľovi na pacienta s rovnakou diagnózou náklady navyše, a preto sa bude nachádzať vysoko nad regresnou priamkou (hodnoty prediktorov budú rovnaké, ale náklady na liečbu vyššie).

Tento poznatok poukázal aj na problém nášho modelu a ten je, že faktorová premenná určujúca, či poskytovateľ má nadštandardný prístroj nie je dostatočne rozčlenená - pod rovnakou úrovňou prístrojovej vybavenosti sa môžu stále nachádzať rôzne prístroje.

Ďalším problémom s týmto prediktorom je to, že suma, ktorú vyplatíme navyše poskytovateľovi s prístrojom závisí aj od počtu poistencov, na ktorých prístroj použije, nie len na tom, či ho má. Môžeme si to demonštrovať na príklade: Majme štyroch poskytovateľov, o ktorých vieme dáta v tabuľke 4.4. Platí, že bežné náklady na pacienta sú 1 a náklady na pacienta u PZS s prístrojom sú 2. PZS s prístrojom pritom používajú daný prístroj na všetkých pacientoch.

Tabuľka 4.4: Príklad k prístrojovej vybavenosti

Uznaná suma	Prístroj	Počet pacientov
10	0	10
20	1	10
40	1	20

Keby sme na takýchto dátach natrénovali lineárnu regresiu bez absolútneho člena, odhadol by $\beta_1 = 10$ a $\beta_1 = 1,33$. V skutočnosti je situácia ešte komplikovanejšia, pretože každý poskytovateľ môže použiť prístroj na jednom pacientovi niekoľkokrát a

na iných pacientoch zase nemusí prístroj použiť vôbec.

Ďalšie poznatky sú príklady podvodného správania, ale aj príklady bežného správania, ktoré sa zistili pri kontrolovaní minulej sady poskytovateľov. Ak narazíme na PZS, ktorý je správaním sa veľmi podobný predošlému skontrolovanému prípadu, tak bude vybraný na kontrolu internou revíziou iba v prípade, ak bol jemu predtým kontrolovaný PZS po internej revízii označený za podozrivého.

Tým, že sme brali do úvahy aj takéto poznatky sme model vylepšovali a vytvorili sme v podstate rozhodovací strom, ktorý preveril podozrivého poskytovateľa určeného modelom lineárnej regresie pomocou získaných poznatkov.

Na základe internou revíziou identifikovaných podozrivých PZS bolo zatiaľ vykonaných 9 externých revízií, z ktorých 6 odhalilo podvodné správanie daných poskytovateľov. Celková úspešnosť spočítaná ako úspešnosť po internej revízii a vynásobená úspešnosťou po externej revízii je 31%.

V tabuľke 4.5 sa nachádzajú výsledky po kontrole internou revíziou pre model na predpis liekov a pomôcok.

Tabuľka 4.5: Kontrola predpisu internou revíziou

Sada	Počet PZS	Počet podozrivých	Úspešnosť
1	8	4	50%
2	9	5	56%

Vzhľadom na to, že sme nemali veľa možností konzultovať výstupy modelu na predpis s odborníkom, počet dát je malý a kontroly prebiehali len v dvoch sadách. Napriek tomu úspešnosť prevyšuje 50%, čo je veľmi dobrý výsledok. Obzvlášť keď zoberieme do úvahy, že išlo o model na predpis, pri ktorom sa v dátach nachádza veľa riadkov, ktoré môžu ľahko ovplyvniť výsledok modelu. Po potvrdení podozrivosti internou revíznou kontrolou boli vybraté lekárne, do ktorých smeroval predpis podozrivých PZS. Externou revíziou boli skontrolované zatiaľ tri z týchto lekární, pričom vo všetkých sa potvrdil podvodný výdaj.

Na model na výkony za rok 2017 sme použili aj test na hľadanie outlierov využívajúci Bonferroniho korekciu. Tento test je možné aplikovať na model v jazyku R funkciou `outlierTest()`. Namiesto výberu pomocou klasických rezíduí sme využili tým

pádom studentizované rezíduá. Testovanie nám pomohlo určiť outlierov a nemuseli sme preto vymýšľať žiadne ďalšie pravidlá na výber poskytovateľov. Pokiaľ sa v modeli nenachádzal podľa testu žiaden outlier, vybrali sme aspoň poskytovateľa s najväčším studentizovaným rezíduom.

Dokopy sa nám týmto testom podarilo nájsť 49 outlierov s kladným studentizovaným rezíduom. Z toho 18 poskytovateľov bolo skontrolovaných na základe modelu s klasickými rezíduami a 11 z nich bolo označených za podozrivých a odporučených na externú revíziu. To nám dáva úspešnosť 61%, čo je vyššie číslo ako úspešnosti pri výbere poskytovateľov na základe štandardných rezíduí. Taktiež 9 poskytovateľov bolo v posledných piatich rokoch externe revidovaných, pričom 6 z nich malo finančný efekt väčší ako nula. To nám dáva ďalšiu štatistiku úspešnosti, ktorá sa v tomto prípade pohybuje na úrovni 67%. Studentizované rezíduá nám teda zrejme pomohli lepšie odhaliť outlierov v lineárnom modeli a mohli by byť do budúcnosti používané na zlepšenie predikcie modelu.

4.2 Zhlukovací model

V tejto podkapitole postupujeme podobne ako v článku [4]. Okrem výkonov u všeobecných lekárov sa však zameriavame hlavne na inú oblasť zdravotníctva, konkrétne na podvodný výdaj liekov a pomôcok v lekárňach. Taktiež používame zhlukovací algoritmus, ktorý berie do úvahy aj šum - DBSCAN. Výsledky znázorňujeme graficky pomocou prvých dvoch hlavných komponentov.

Pomocou dát o externých revíznych kontrolách lekární sa pokúšame identifikovať podozrivo sa správajúce lekárne.

Táto práca sa zaoberá odhaľovaním podvodných PZS. Teraz sa však zaoberáme aj odhaľovaním podvodných lekární. V mnohých prípadoch platí, že jedna lekáreň zodpovedá práve jednému PZS. Môže však existovať PZS, ktorý vlastní aj viac lekární. Dáta o externých revíziách sa viažu na konkrétnu lekáreň a preto hľadáme podvody na úrovní lekární.

4.2.1 Myšlienka a postup modelovania

Základnou myšlienkou tohto modelu je na základe premenných popisujúcich správanie sa jednotlivých lekární ich roztriediť na také, ktoré sa správajú štandardne a také, ktoré sa správajú podozrivo. Za premenné sme zobrali identifikátory podvodu vytvorené pomocou skúseností revíznych lekárov a okrem toho aj ďalšie štatistické premenné popisujúce dáta o lekárňach. Premenné sme pritom zobrali dostatočne všeobecné, aby sme dokázali okrem známych druhov podvodov odhaliť aj nové, doteraz neznáme podvody.

Parameter zhlukovacieho algoritmu DBSCAN *Eps* sme našli pomocou grafu vzdialeností k najbližších susedov. Parameter $Minpts$ sme zvolili na základe kontextu aplikácie. DBSCAN zaraďuje jednotlivé lekárne buď do zhlukov, alebo ich označí ako šum. Identifikácia šumu je veľmi užitočná, pretože lekárne, ktoré sú zaradené do šumu majú významne odlišné premenné od ostatných lekární. Keďže sme za premenné zobrali práve identifikátory podvodu, tak šum je pre nás veľmi zaujímavý a môže sa v ňom nachádzať veľa podvodníkov. DBSCAN nám teda umožnil okrem nájdenia potencionálnych zhlukov obsahujúcich podvodníkov nájsť aj podvodníkov na základe štatistických výchyliek a to je dôvod, prečo sme si zvolili tento zhlukovací algoritmus.

Postup, ktorý najprv aplikujeme na dáta o výdaji liekov a pomôcok sa dá použiť aj na iné oblasti zdravotníctva, ako je napríklad aj ŠAS. V prípade ŠAS by sme však pre dobré výsledky potrebovali pre každú odbornosť vytvoriť rôzne identifikátory podvodu, čo je časovo veľmi náročné. Na demonštráciu sme si však zvolili odbornosť všeobecné lekárstvo pre dospelých. V článku [4] sa nachádza aplikácia podobného postupu na všeobecných lekároch. Tento článok však z veľkej časti zameriava na podvodnú preskripciu (teda inú oblasť zdravotníctva) a pracuje s odlišným zdravotným systémom. Indikátory podvodu sme v tomto prípade vytvorili na základe našich skúseností pri verifikácii lineárneho modelu.

Po vytvorení zhlukovania ho vykresľujeme pomocou prvých dvoch hlavných komponentov. Na základe dát o externých revíziách, ktoré taktiež zakresľujeme do obrázku, daný model a postup hodnotíme, prípadne vylepšujeme.

4.2.2 Aplikácia na lekárne

Model sme aplikovali na dáta o výdaji liekov a zdravotníckych pomôcok za rok 2017. Ide pritom o 4 milióny riadkov dát, ktoré obsahujú výdaje z lekární a výdajní zdravotníckych pomôcok. Iba 70000 z týchto riadkov sa zaoberá výdajňami zdravotníckych pomôcok. Preto sme sa rozhodli aplikovať postup iba na lekárne. S pomocou revíznych farmaceutiek a skúseností získaných pri verifikácii lineárneho modelu sme vytvorili 42 premenných, z ktorých veľká väčšina sú identifikátory podvodu.

Uvádzame opis premenných, ktoré sme použili pri aplikácii na lekárne:

- Uhradená suma na poistenca, množstvo liekov a pomôcok na poistenca, štandardná odchýlka sumy na poistenca a priemerný počet návštev lekárne jedného poistenca. Okrem týchto premenných sme vytvorili také isté premenné pre pravidelných poistencov a pre nepravidelných poistencov (poistencov, ktorí sa objavili v lekárni len niekoľkokrát). Dokopy ide o 12 premenných.

- Sumy na poistenca pre lieky a pomôcky s vysokými cenami v dvoch skupinách a ich podiel oproti celkovej uhradenej sume danej lekárni. 4 premenné.

- Sumy a množstvá na poistenca pre rôzny podozrivý materiál (materiál, s ktorým sa často podvádza). Podiely uhradenej sumy na tento materiál oproti celkovej uhradenej sume. Celkovo ide o 18 premenných.

- Podiel najväčšieho lekára (lekára, ktorí predpísal zdravotnícky materiál za najväčšiu sumu) a podiel najväčšieho PZS na celkovej uhradenej sume. Dokopy 2 premenné.

- Vzdialenosť na 1 euro. Ide pritom o najkratšiu vzdialenosť (medzi obcami), ktorú musel poistenec prejsť buď z trvalého bydliska, alebo od prevádzky predpisovateľa do lekárne. Tieto hodnoty sú vypočítané pre 3 rôzne obmedzenia na veľkosť vzdialenosti (dôvodom je, že výbery poistencov z východu Slovenska na západe Slovenska výrazne ovplyvňujú túto štatistiku). 3 premenné.

- Suma na poistenca na lieky na výnimku a podiel uhradenej sumy na tieto lieky oproti celkovej uhradenej sume danej lekárne. 2 premenné.

- Suma na poistenca na diagnózu spojenú s rakovinou. 1 premenná.

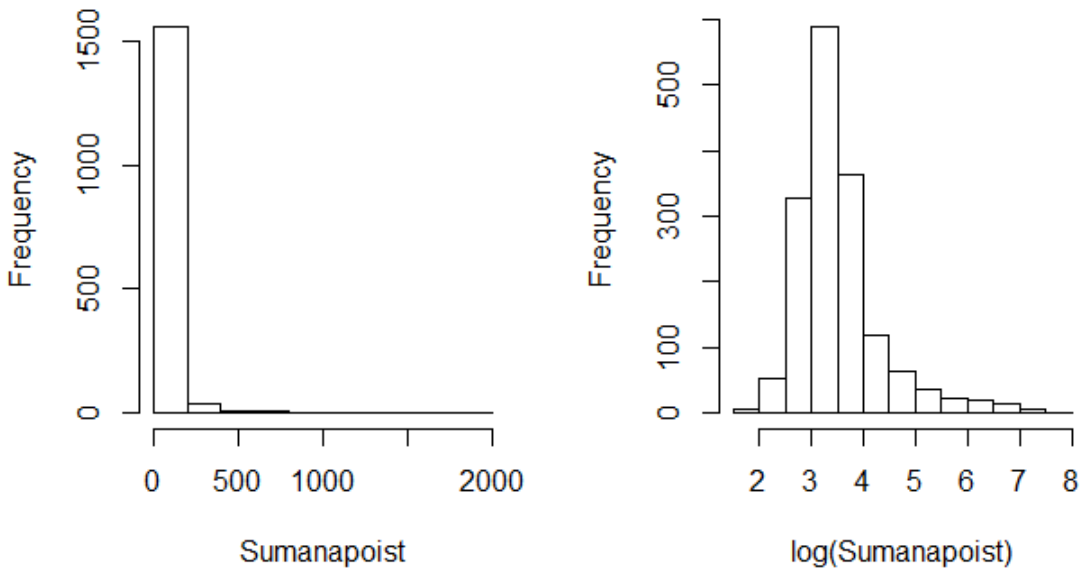
Ako si môžeme všimnúť, ide síce o pomerne veľký počet premenných, ale sú vytvorené podobnou logikou a veľa z nich sa dokonca zameriava na ten istý problém. Z pohľadu na tieto premenné je zrejmé, že niektoré sú vysoko korelované. Premenné typu suma na poistenca a množstvo na poistenca by pri zameraní sa na jeden liek mali koreláciu dokonca jedna. Avšak v našej analýze sa zameriavame na skupiny liekov alebo pomôcok, v ktorých niektoré sú drahé a niektoré lacné. Suma na poistenca tým pádom bude odhaľovať skôr drahé lieky (samozrejme ale aj veľké množstvo lacných liekov), ale množstvo na poistenca bude odhaľovať práve veľké množstvo lacnejších liekov. Pritom veľmi drahý liek nebude mať na túto premennú silný vplyv.

Ďalšia vec, ktorú by sme si mohli pri týchto premenných všimnúť je, že existuje veľmi veľa možností, ako ich obmieňať a aj veľa možností, ako vymyslieť nové premenné na základe podobného princípu. V tejto práci pracujeme už iba s týmito premennými, ale musíme poukázať na dôležitosť výberu premenných pri tejto metóde. Premenné musia byť dobre premyslené a musia reálne poukazovať na nejaký podvod. Pokiaľ pri analýze uvidíme, že niektoré premenné poukazujú na nejaký podvod, tak ich vždy môžeme ešte vylepšovať, obmieňať a testovať, aby model fungoval ešte lepšie. Ak však zo začiatku pracujeme s úplne zlými premennými, tak len veľmi ťažko dospejeme k použiteľným výsledkom.

Najprv sme vyskúšali vytvoriť zhlukovanie na základe všetkých 42 premenných. Do analýzy sme zahrnuli iba lekárne, ktoré mali ročnú uhradenú sumu aspoň 5000, aby sme sa vyhli štatistike malého čísla (táto suma je však stále dosť malá, no napriek tomu je to už dostatočne veľká suma na to, aby bola do terénu vyslaná externá revízia pri takmer istom podvode). Počet takýchto lekární je 1617.

Keď sa pozrieme na histogramy jednotlivých premenných, tak zistíme, že sú veľmi vychýlené a nachádzajú sa tam outlieri, ktorí sú veľmi vzdialení väčšine ostatných lekární. Vychýlenosť premenných môže veľmi ovplyvniť kvalitu zhlukovania. Jedným zo spôsobov, ako opraviť takúto vychýlenosť je logaritmická transformácia. Na obrázku 4.3 môžeme vidieť histogram jednej z vychýlených premenných a histogram jej logaritmickej transformácie. Môžeme pozorovať, že logaritmická transformácia opravila vychýlenosť premennej. V našom prípade sa však snažíme práve o identifikáciu outlie-

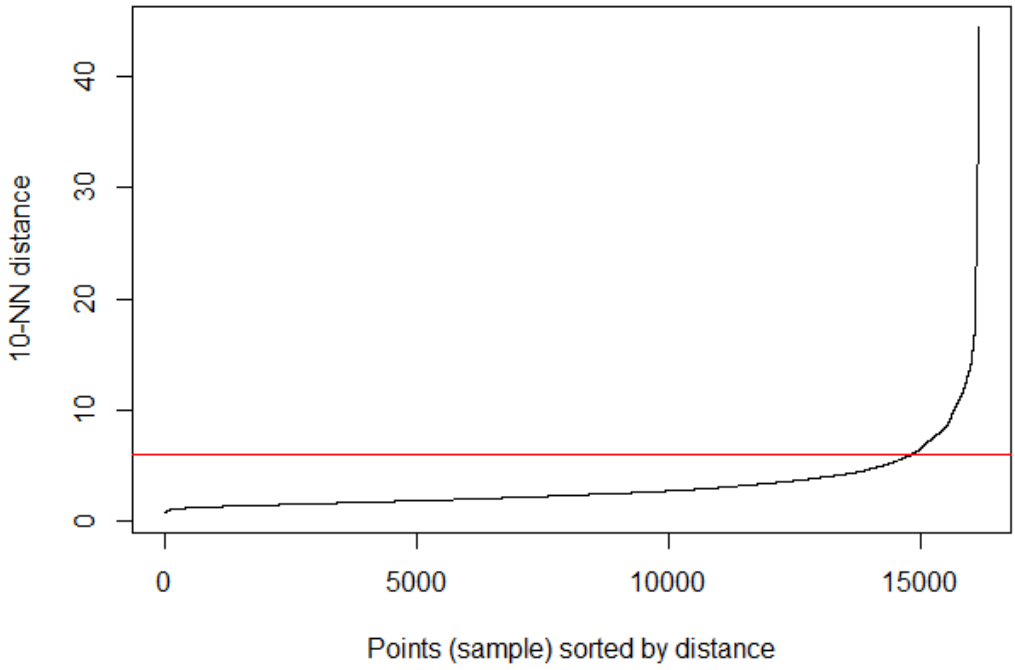
rov. Logaritmická transformácia nezachováva pôvodné vzdialenosti a outlierov posúva veľmi blízko všetkým ostatným pozorovaniam. Preto sme ju na naše dáta nepoužili.



Obr. 4.3: Logaritmická transformácia vychýlenej premennej

Ešte pred použitím zhlukovacieho algoritmu sme všetky premenné preškálovali tak, aby mali výberovú disperziu rovnú 1 a priemer rovný 0. Vďaka tomu majú odchýlky v rôznych premenných rovnakú dôležitosť. Jedným zo spôsobov, ako sa dá v tomto kroku ovplyvňovať výsledok je dať jednotlivým premenným rôzne váhy a priradiť im tak rôznu dôležitosť. V našom prípade sme tento krok preskočili. Je dobré však myslieť na to, že to je jeden zo spôsobov, ako vylepšiť výsledok.

Na vytvorenie zhlukovania pomocou algoritmu DBSCAN sme potrebovali stanoviť parametre $Minpts$ a Eps. $Minpts$ je v našej aplikácii počet lekární, ktoré sa musia nachádzať v okruhu Eps, aby bol daný bod považovaný za jadrový. Ak sa v okolí nejakej lekárne nachádza aspoň 10 rôznych lekární, tak predpokladáme, že ich správanie je podobné z nejakej príčiny a nejde len o šum. Preto sme stanovili parameter $Minpts = 10$. Parameter Eps sme našli pomocou grafu vzdialeností k najbližších susedov, kde $k = 10$. Graf vzdialeností k najbližších susedov sme vytvorili v R pomocou príkazu `KNNdistplot()`. Tento graf môžeme vidieť na obrázku 4.4.



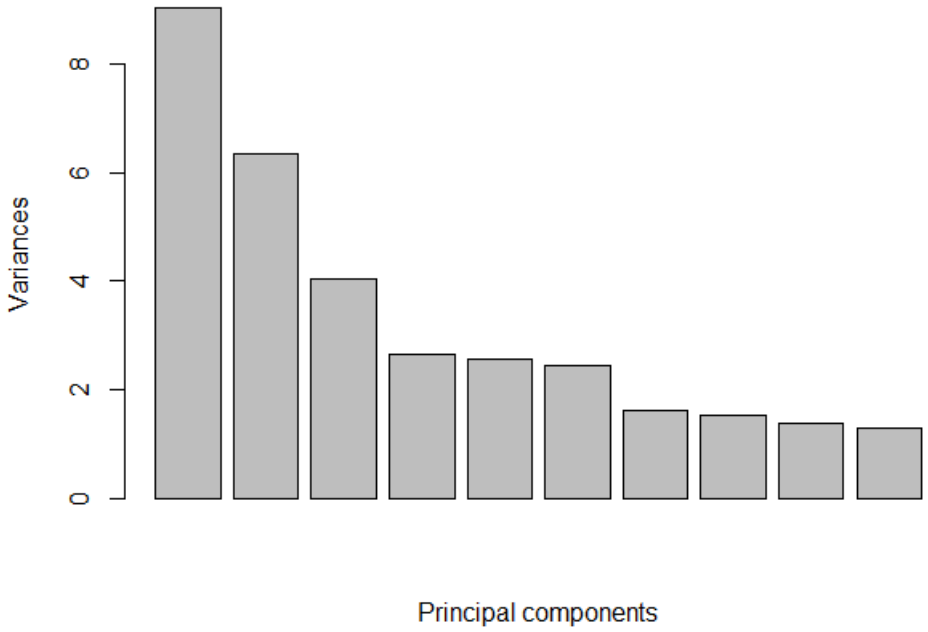
Obr. 4.4: KNN distance plot pre lekárne

Na obrázku môžeme vidieť, že lakeť grafu sa nachádza približne v hodnote 6. Preto sme stanovili $Eps = 6$. Táto hodnota je v grafe zakreslená červenou čiarou.

Pred vykreslením zhlukovania je potrebné spraviť analýzu hlavných komponentov. Analýzu hlavných komponentov sme vytvorili v R príkazom `prcomp()`. Na obrázku 4.5 vidíme variancie prvých 10 hlavných komponentov.

Podiel vysvetlenej variancie v dátach pre prvé dva hlavné komponenty je iba 37 percent. Vzhľadom na to, že sme sa pokúšali vykresliť 42 premenných do dvojrozmerného priestoru sme však väčšie číslo nemohli očakávať. Pri opisovaní obrázka je však potrebné myslieť na to, že mu nemožno úplne dôverovať. Vzhľadom na veľký počet premenných je interpretácia hlavných komponentov pomocou ich prislúchajúcich koeficientov veľmi náročná, preto sme sa o ňu nepokúšali.

Okrem vykreslenia zhlukovania sme do obrázku špeciálne zaznačili aj lekárne, v ktorých externá revízia kontrolovala výdaj za rok 2017 a potvrdila podvodnú schému a lekárne, v ktorých bola externá revízia a vyvrátila podvodnú schému. Za podvodné lekárne sme zobrali také, pri ktorých finančný efekt externej revízie bol aspoň 2000



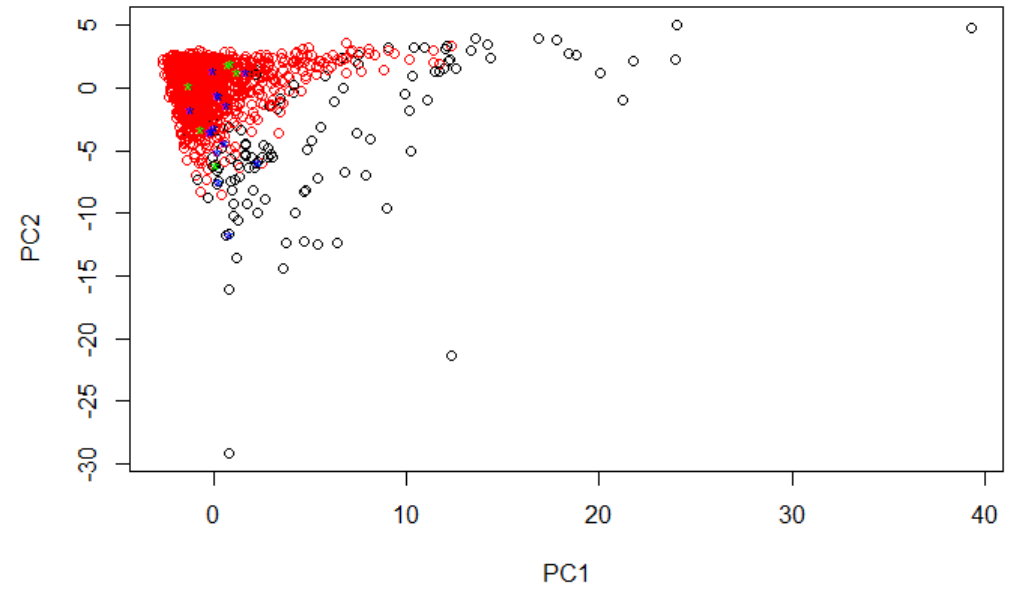
Obr. 4.5: Variancie hlavných komponentov pre lekárne

eur. Takýchto lekární je dokopy 14. Za lekárne, ktoré sú v poriadku sme označili také, kde finančný efekt externej revízie bol rovný nule. Ich počet je pritom 6.

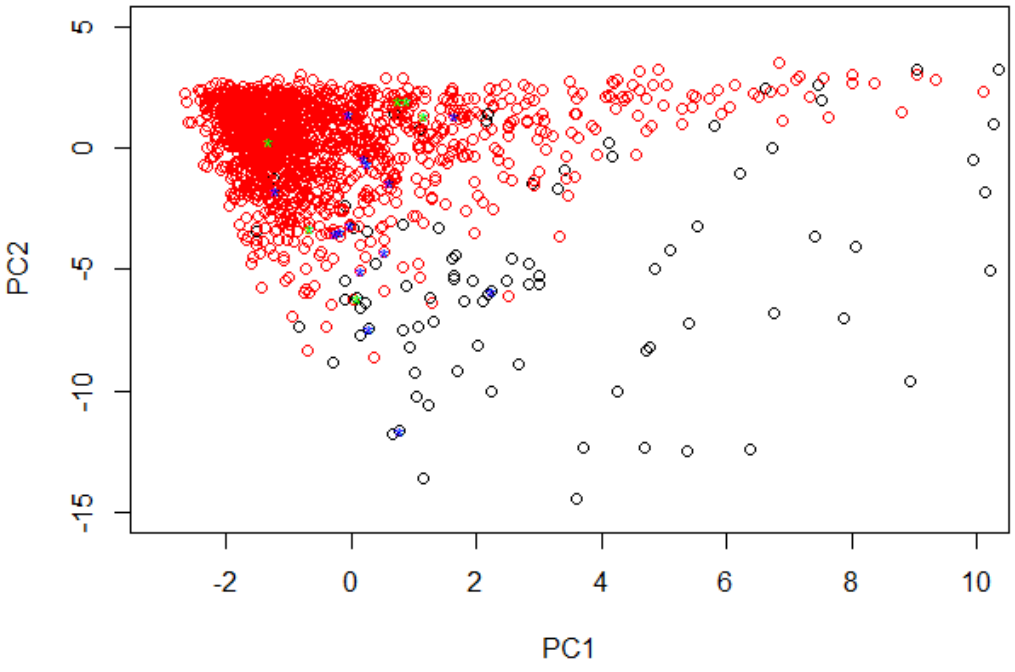
Zhlukovanie algoritmom DBSCAN sa v jazyku R dá vytvoriť pomocou príkazu `dbscan()`. Zhlukovanie našlo jeden zhluk, do ktorého patrí 1503 lekární a šum, do ktorého patrí 114 lekární. Na obrázku 4.6 môžeme vidieť vykreslenie tohto zhlukovania pomocou prvých dvoch hlavných komponentov.

Červené krúžky pritom znázorňujú lekárne patriace do zhluku, čierne krúžky znázorňujú lekárne patriace do šumu, modré hviezdičky označujú podvodné lekárne a zelené hviezdičky označujú lekárne, u ktorých externá revízia nič nenašla. Tento obrázok je veľmi skreslený kvôli outlierom vzdialeným ďaleko od zhluku, preto prikladáme priblížený obrázok 4.7.

Tento obrázok už má väčšiu výpovednú hodnotu. Môžeme si všimnúť, že iba 3 z podvodných lekární sa nachádzajú v šume a ostatné sa nachádzajú v zhluku. Keďže naše dáta obsahujú iba jeden zhluk, je jasné, že podvodné lekárne by sa mali nachádzať v šume. Vzhľadom na tento fakt môžeme vypočítať istú úspešnosť modelu - koľko z podvodných lekární sa reálne nachádza v šume. To nám pre tento model dáva úspešnosť



Obr. 4.6: Zhlukovanie pre lekárne



Obr. 4.7: Priblížené zhlukovanie pre lekárne

21%. To je síce dosť málo, ale tento model má veľký priestor na zlepšenie. Takéto počítanie úspešnosti však veľa vecí neberie do úvahy, ako je napríklad poloha podvodných lekární v šume. Kvalita modelu sa však dá dosť dobre odhadnúť aj z obrázka, a preto sa nebudeme zaoberať vymýšľaním nejakého lepšieho hodnotiaceho kritéria.

Jednou z vecí, čo by sme však mohli posúdiť je koľkokrát je pravdepodobnejšie, že podvodná lekáreň spadne do šumu, ako do zhluku. Spočítali sme teda hodnotu $(3/114)/(11/1503)$, čo je rovné približne $3,6$. To síce nie je obrovské číslo, ale dáva nám to predstavu o tom, že tento model má istú predpovedaciu silu.

Keď sa na obrázok pozrieme bližšie vidíme, že žiadna z podvodných lekární sa nenachádza úplne v centre zhluku, kde sa nachádza najviac lekární a veľké množstvo podvodných lekární sa nachádza na úplnom okraji zhluku. To je dobrý znak a znamená to, že keby sa nám podarilo zhlukovanie vylepšiť, mohli by sa do šumu ľahko dostať. Vylepšiť zhlukovanie by sa mohlo dať rôznymi spôsobmi, ako sú výber najlepších premenných, tvorba nových premenných, váženie premenných atď. Všetky z týchto spôsobov sú však časovo veľmi náročné a vyžadujú veľa pokusov. Navyše na ďalšom príklade si ukážeme, že to nemusí byť úplne najlepší postup.

Na obrázku 4.7 si ešte môžeme všimnúť polohu lekární, ktoré nie sú podvodné. Vidíme, že aj keď sú priemerne bližšie ku stredu zhluku, sú tam aj také, ktoré sa nachádzajú na okraji zhluku. Dôvodom je, že lekárne sa na externú revíziu nevyberajú úplne náhodne. Vyberajú sa buď na podnety poistencov, alebo na základe skúseností revíznych farmaceutiek, ktoré nám pomáhali vytvoriť niektoré z identifikátorov podvodu. Ukazuje nám to, že aj keď sa lekárne štatisticky vychyľujú od priemeru, stále to nemusí znamenať, že ide o podvod. Ešte raz by sme chceli pripomenúť, že napriek tomu, že sme z obrázku zistili zaujímavé informácie, tak prvé dva komponenty vysvetľujú len 37% variancie a preto treba získané informácie brať s rezervou.

Po prezretí obrázka znázorňujúceho zhlukovanie by nám mohlo napadnúť, či má význam robiť zhlukovanie na dátach, ktoré majú iba jeden zhluk. DBSCAN je však algoritmus, ktorý nám okrem hľadania zhlukovaní pomáha nájsť aj outlierov. Outlierov si vieme definovať dvoma parametrami: $Minpts$ a Eps. Pri stanovení menšieho $Minpts$ by sme v dátach vedeli nájsť aj viac zhlukov, ktoré by sme mohli samostatne skúmať. Ďalším argumentom v prospech tohto algoritmu je to, že teraz pracujeme s dátami

o lekárňach, ktoré majú zrejme len jeden zhluk. Ak by sme pracovali s inými dátami, alebo ak by sme do týchto dát pridali ďalšie premenné, mohlo by sa nám podariť nájsť aj viac zhlukov. Napriek týmto argumentom sa nám vďaka tvaru dát vytvárajú nové možnosti, ako s nimi pracovať. Jednou z nich by mohlo byť vytváranie obálky dát a následne hľadanie dát, ktoré sú najbližšie k obálke. Použitím logaritmickej transformácie by sme vedeli získať dáta, ktoré majú elipsoidný tvar, čo by nám ponúkalo ďalšie možnosti. V tejto práci sa však budeme zaoberať len zhlukovacím algoritmom DBSCAN.

V tejto aplikácii sme použili na zhlukovanie 42 premenných, ktoré sú identifikátormi podvodu. Tieto identifikátory sa pritom zameriavajú na rôzne druhy podvodu, ktoré sa v dátach môžu prejavovať veľmi odlišne. Lekárne na okraji zhluku v našom príklade mohli byť v niektorých identifikátoroch outlieri, ale v ostatných mohli byť veľmi priemerné. To mohlo zapríčiniť, že sa nedostali do šumu.

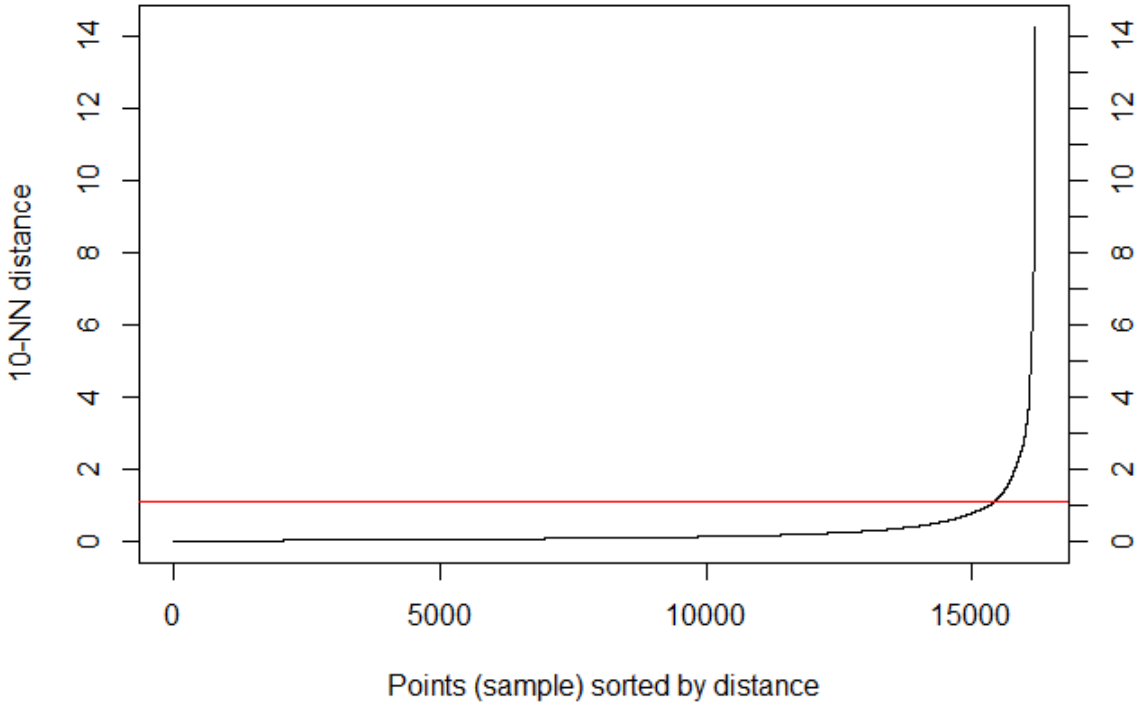
V ďalšom príklade sme sa zamerali na podvod, kedy lekárne vydávajú príliš veľké množstvá podozrivého materiálu (veľké množstvo lacnejších výrobkov). Vybrali sme teda identifikátory, ktoré sa zameriavajú na takýto podvod. Zobrali sme pritom 4 identifikátory, pri ktorých je najväčší rozdiel medzi priemerom daného identifikátora u podvodných lekární a priemerom identifikátora u všetkých ostatných lekární. V tabuľke 4.6 uvádzame popis výsledných štyroch premenných.

Tabuľka 4.6: Indikátory na zhlukovanie pre lekárne

Názov	Popis premennej
MPP	Množstvo na poistenca u pravidelných poistencov
Podiel_B	Podiel pomôcok na inkontinenciu
MPF	Množstvo na poistenca na stomické pomôcky
Podiel_celk	Celkový podiel podozrivého zdravotníckeho materiálu

Na základe vybratých identifikátorov sme urobili zhlukovanie. Za parameter $Minpts$ sme tentokrát zobrali podľa heuristiky dvojnásobok dimenzie, teda 8. *Eps* sme našli ako v predošlom prípade pomocou grafu vzdialeností k najbližších susedov, ktorý sa nachádza na obrázku 4.8.

Ideálne *Eps* nám v tomto prípade vychádza 1,1.



Obr. 4.8: KNN distance plot pre lekárne

V tomto prípade prvé dva komponenty vysvetľujú až 86% variancie v dátach. To je už dosť veľké množstvo a dvojrozmernému obrázku sa preto dá dobre veriť. Keďže nepracujeme s veľkým počtom premenných, pokúsili sme sa o interpretáciu hlavných komponentov. V tabuľke 4.7 zobrazujeme koeficienty α_1 a α_2 prvých dvoch hlavných komponentov.

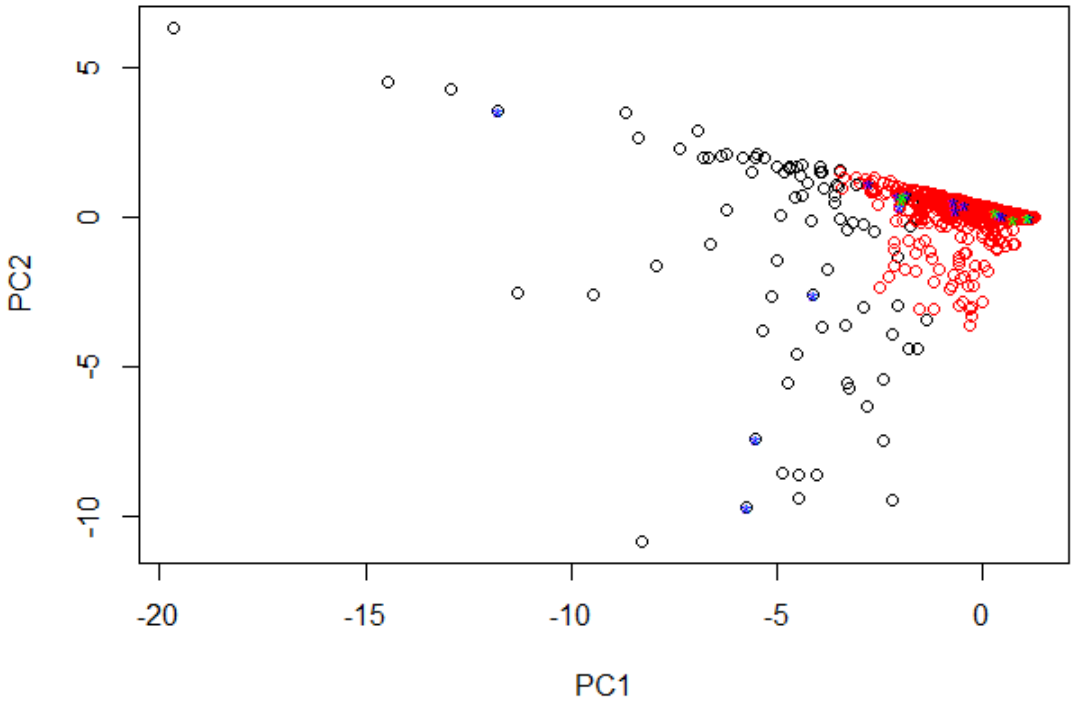
Tabuľka 4.7: Koeficienty PCA pre lekárne

	PC1	PC2
MPP	-0,58	0,11
Podiel_B	-0,55	0,37
MPF	-0,24	-0,92
Podiel_celk	-0,55	-0,06

Vidíme, že čím je nižší prvý komponent, tým sú vyššie hodnoty všetkých identi-

fikátorov. Najmenšia váha však je na premennej MPF. Druhý komponent má výrazne najväčšiu váhu (v absolútnej hodnote) práve pri premennej MPF. V prvom komponente by sa teda mali zobrazovať hlavne podvody na inkontinencii, ale s menšou váhou aj na stomických pomôckach. V druhom komponente by sa mali zobrazovať podvody iba na stomických pomôckach a nie na pomôckach na inkontinenciu, pretože pri premennej Podiel_B je v druhom komponente výrazne kladná váha.

Zhlukovanie lekárne rozdelilo na 1537 lekární v zhluku a 81 lekární v šume. Podobne ako v predchádzajúcom príklade sme okrem zhlukovania do obrázka 4.9 zakreslili aj podvodné a nepodvodné lekárne.



Obr. 4.9: Zhlukovanie pre lekárne

Z obrázku vidíme, že tentokrát sme identifikovali v šume 4 podvodné lekárne, čo nám dáva úspešnosť 29%. Správne identifikované lekárne sú zároveň jednými z najväčších outlierov, teda sa nachádzajú veľmi blízko k obálke dát, čo je dobré. Taktiež si môžeme všimnúť jednu podvodnú lekáreň, ktorá sa nachádza na úplnom okraji zhluku. Ostatné lekárne (aj podvodné, aj nepodvodné) sa nachádzajú dosť blízko k centru zhluku.

Pravdepodobnosť, že podvodná lekáreň sa bude nachádzať v šume je tentokrát 7, 6-krát väčšia ako pravdepodobnosť, že podvodná lekáreň sa bude nachádzať v zhluku. Dostali sme teda lepší model.

Výsledok tohto zhlukovania vyzerá veľmi dobre, okrem toho, že sme odhalili v šume 4 lekárne, ktoré sú zároveň jedny z najvzdialenejších od centra zhluku, tak sa nám podarilo objaviť aj nové lekárne, ktoré sa v týchto identifikátoroch vychyľujú od priemeru. Ostatné podvodné lekárne, ktoré sa nachádzajú blízko k centru zhluku sa zrejme vychyľujú od priemeru v iných identifikátoroch. Keďže sme sa zamerali len na jeden druh podvodu, tak lekárne, ktoré tento podvod vykonávali sa jasne odlíšili od zhluku. Zároveň podvodné lekárne, ktoré tento druh podvodu nevykonávali sa pohybovali veľmi blízko centru zhluku.

Ak by sme sa zamerali na iný druh podvodu, napríklad výdaj veľmi drahých liekov, tak by sme pravdepodobne mohli odhaliť iných pár podvodných lekární a ostatné lekárne by sa zase pohybovali okolo centra zhluku.

Jednou z vecí, ktoré by nám mohli napadnúť na základe tohto výsledku je, či by nebolo lepšie sledovať identifikátory podvodu jednotlivo a skúmať tak outlierov jednorozmerne. To síce môže byť dobrý spôsob, ale zhlukovanie nám dáva oveľa viac informácií o postavení voči ostatným lekárňam a takisto vie dobre spracovať vzťahy medzi identifikátormi podvodu.

Podvodov v lekárňach môže byť veľmi veľa druhov a každý druh podvodu sa môže v dátach prejavovať inak. Aj preto keď sme zobrali všetky identifikátory podvodu, výsledné zhlukovanie na obrázku nevyzeralo veľmi dobre. Podvodné lekárne sa pohybovali väčšinou na okrajoch zhluku. V niektorých identifikátoroch mohli byť dané podvodné lekárne outlieri, ale v iných mohli byť zase veľmi priemerné.

Ďalším z dôvodov, prečo sme nemali až takú dobrú úspešnosť mohlo byť aj nie úplné prepojenie našich dát s dátami o externých revíziách. Externé revízie kontrolujú v lekárni väčšinou posledné dva roky výdajov. Pritom my sme brali do úvahy externé revízie, ktoré kontrolovali dáta z roku 2017, ale mohli kontrolovať aj dáta z iného roku. Podvod pritom mohol nastať buď v inom roku, alebo mohla v inom roku nastať jeho veľká časť.

Keď zoberieme do úvahy, že sme analyzovali dáta o výdaji liekov a pomôcok,

v ktorých sa nachádza veľmi veľa rôznych neštandardných dát, dosiahli sme dobré výsledky. V nasledujúcej podkapitole aplikujeme zhlukovací algoritmus na dáta o výkonoch všeobecných lekárov. V týchto dátach sa nenachádza moc neštandardných výkonov a podvodné praktiky sú nám aj vďaka verifikácii lineárneho modelu veľmi dobre známe.

4.2.3 Aplikácia na všeobecných lekárov

Na analýzu dát o výkonoch všeobecných lekárov pre dospelých sme použili rovnaký postup ako pri aplikácii na lekárne. Analyzovali sme dáta za rok 2017, čo predstavuje približne $1,7$ milióna riadkov zdravotnej starostlivosti. V tejto odbornosti sa nachádza 2012 rôznych poskytovateľov zdravotnej starostlivosti.

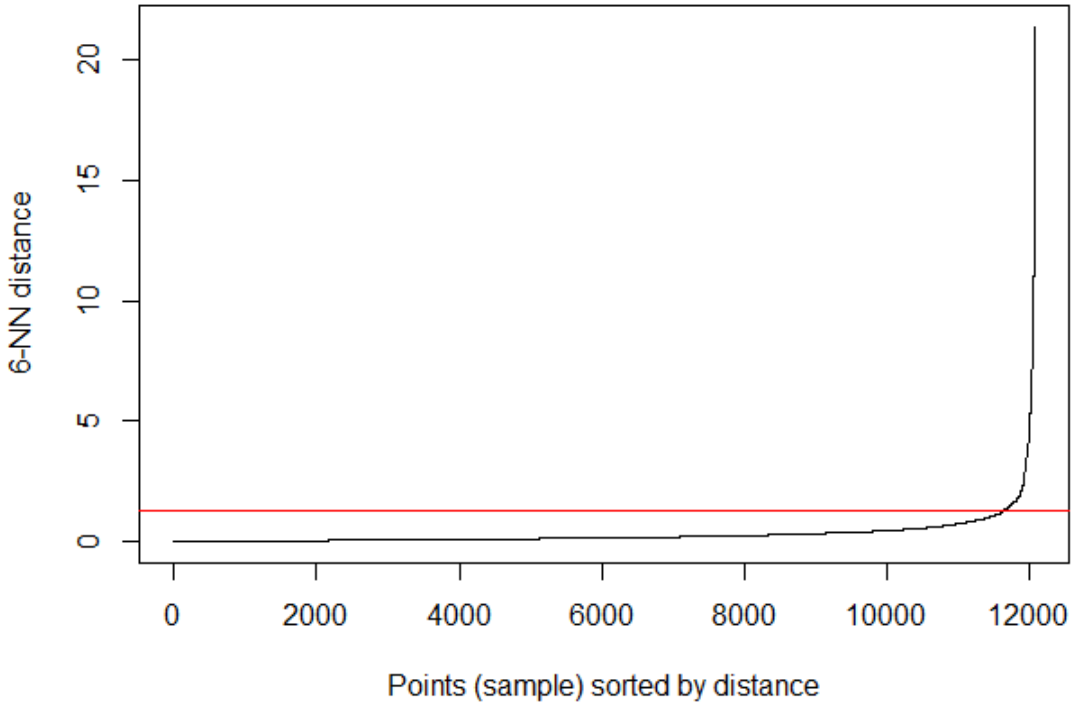
Všeobecní lekári sú špecifickí tým, že sú z veľkej časti hradení kapitáciami. Kapitácie sú mesačné platby za prihláseného pacienta. Pokiaľ suma, ktorá by mala byť hradená všeobecnému lekárovi za výkony za dané časové obdobie neprekročí sumu kapitácií, nebude mu za dané výkony vyplatené nič.

Na identifikáciu podvodných poskytovateľov sme vytvorili 6 identifikátorov podvodu, ktoré opisujeme v tabuľke 4.8. V tomto prípade sme vytvorili špecifické identifikátory, ktoré sa sústreďujú hlavne na známe rizikové oblasti zdravotnej starostlivosti a nedávajú tak veľký priestor na objavenie nových druhov podvodov. Výsledky by však vďaka tomu mali byť oveľa presnejšie než pri aplikácii na lekárne.

Tabuľka 4.8: Indikátory na zhlukovanie pre VAS

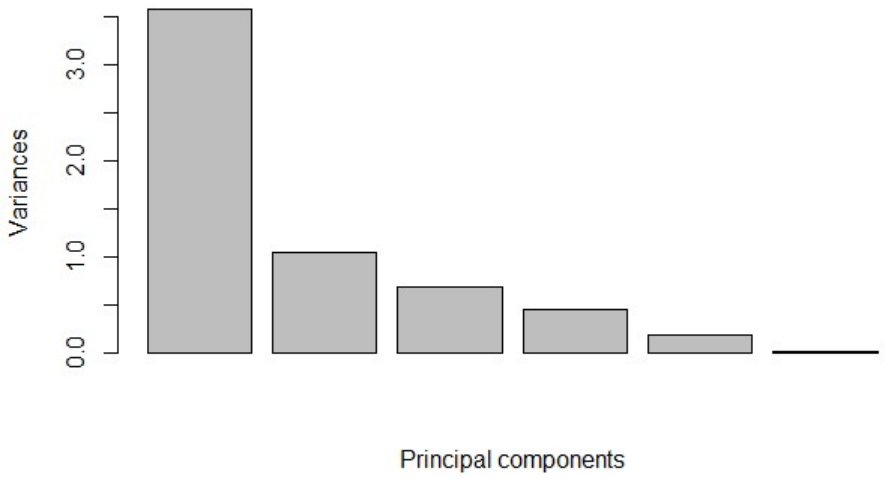
Názov	Popis premennej
Sumanapoist	Suma nad rámec kapitácie na ošetreného poistenca
Navstevy_B	Suma na poistenca na návštevné výkony
INJ	Suma na poistenca na podané injekcie
CRP_celk	Suma na poistenca na stanovovanie C-reaktívneho proteínu
NEKAP_celk	Suma na poistenca na nekapitovaných poistencoch
Podiel_KAP_celk	Podiel sumy za kapitácie ku celkovej uhradenej sume

Na základe týchto identifikátorov sme vytvorili zhlukovanie. Parameter $Minpts$ sme tentokrát zvolili 6. Podľa grafu vzdialeností k najbližších susedov, ktorý sa nachádza na obrázku 4.10, sme stanovili parameter $Eps = 1,3$



Obr. 4.10: KNN distance plot pre VAS

Variancie hlavných komponentov sa nachádzajú na obrázku 4.11. Variancie prvých dvoch hlavných komponentov predstavujú 77% celkovej variancie v dátach, čo je dostatočne veľké množstvo na dobrú interpretáciu z obrázku.



Obr. 4.11: Variancie hlavných komponentov pre VAS

Opäť sme sa pokúsili interpretovať prvé dva hlavné komponenty. Do tabuľky 4.9 sme vložili koeficienty prvých dvoch hlavných komponentov.

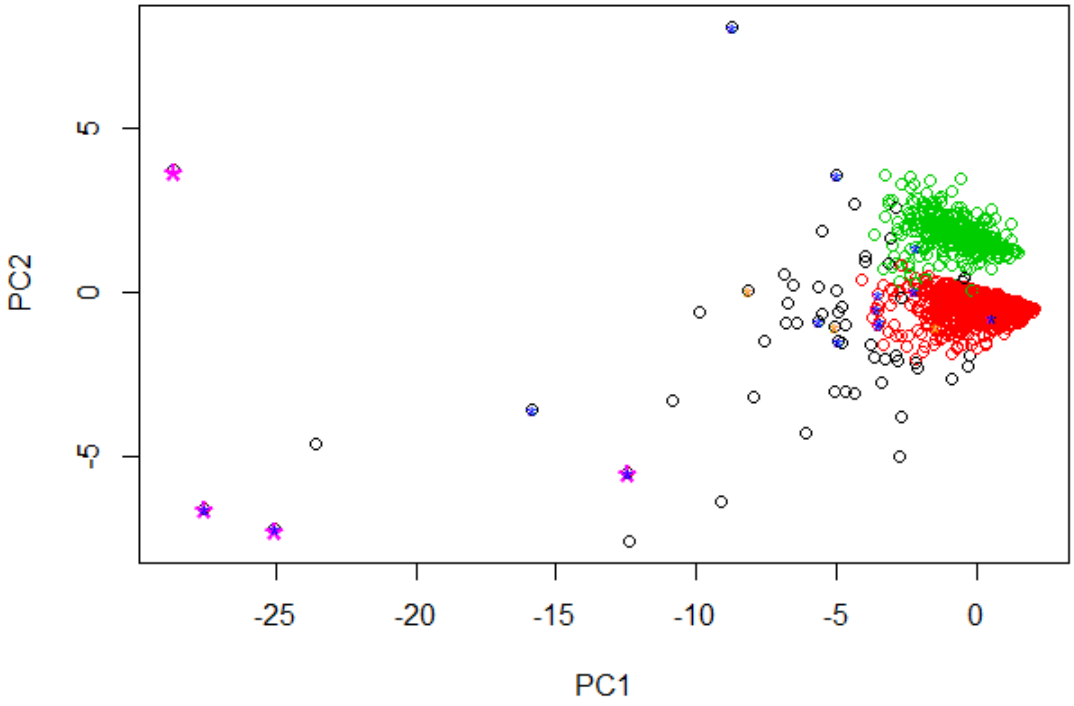
Tabuľka 4.9: Koeficienty hlavných komponentov pre VAS

Názov premennej	PC1	PC2
Sumanapoist	-0,51	-0,06
Navstevy_B	-0,35	-0,34
INJ	-0,38	-0,42
CRP_celk	-0,16	0,82
NEKAP_celk	-0,5	0,06
Podiel_KAP_celk	-0,45	-0,19

Vidíme, že prvý hlavný komponent je mínus vážená suma všetkých identifikátorov, pričom podstatne menší koeficient zodpovedá identifikátoru CRP. Druhý hlavný komponent má najväčšiu váhu práve pri identifikátore CRP. Signifikantne záporné váhy má pri identifikátoroch Navstevy a INJ. PZS s veľkým druhým hlavným komponentom by mal byť taký, že podvádza na CRP, ale nepodvádza na injekciách a návštevách.

Zhlukovací algoritmus našiel dva zhluky, do prvého zhluku zaradil 1579 PZS, do druhého 375 PZS a do šumu zaradil 58 PZS. Porovnaním premenných pre PZS patriacich do prvého a druhého zhluku sme zistili, že v prvom zhluku sa nachádzajú poskytovatelia, ktorí nevlastnia alebo nepoužívajú prístroj na identifikáciu C-reaktívneho proteínu v krvi. V druhom zhluku sa nachádzajú PZS, ktorí tento prístroj používajú. Zhlukovanie sme zakreslili do obrázku 4.12. Prvý zhluk je pritom zakreslený červenými kruhmi a druhý zhluk je zakreslený zelenými kruhmi. Okrem toho je modrými hviezdička-mi na základe dát o externých revíziách označených 15 podvodných PZS. Za podvodných PZS sme pritom brali takých PZS, kde finančný efekt externej revízie bol väčší ako nula. Fialovými hviezdičkami sme zakreslili štyroch PZS, ktorí boli označení modelom lineárnej regresie a po verifikácii internou revíziou boli označení za podozrivých. Nakoniec oranžovou farbou sú zakreslení traja PZS, ktorí boli skontrolovaní externou revíziou, ale finančný efekt bol nulový.

Model správne identifikoval 8 PZS, čo nám dáva úspešnosť 53%. Úspešnosť by sme mohli ešte zvýšiť, ak by sme zobrali do úvahy aj poskytovateľa, ktorý bol odhalený



Obr. 4.12: Zhlukovanie pre VAS

modelom lineárnej regresie, ale nebola na ňom ešte vykonaná externá revízia. Vidíme, že týmto modelom sa nám podarilo odhaliť takmer všetky doteraz známe podvodné schémy. V úplnom centre zhluku sa nachádza iba jeden podvodný PZS a ostatní sa nachádzajú buď na úplnom okraji zhluku, alebo mimo zhluku. Zároveň najväčší outlieri sú odhalení ako podvodníci. Taktiež pravdepodobnosť, že sa podvodník bude nachádzať v šume je $38,5$-krát väčšia ako pravdepodobnosť, že sa podvodník bude nachádzať v jednom zo zhlukov. Podarilo sa nám teda nájsť veľmi kvalitný model na odhaľovanie podvodov. Upravením parametrov *Minpts* a *Eps* by sme mohli docieliť, aby sa niektorí z PZS na okraji zhluku dostali do šumu a zvýšiť tak úspešnosť ešte viac.

Ďalším prínosom tohto modelu je, že sa nám podarilo istým spôsobom verifikovať model lineárnej regresie pre všeobecné lekárstvo pre dospelých. Poloha PZS odhalených lineárnym regresným modelom svedčí o jeho kvalite. Kvalitu vidno však hlave na prvom komponente. Povedať to, že model nepredpovedá dobre aj v druhom komponente sa však zo štyroch dát povedať nedá.

4.3 Zhrnutie modelovania a porovnanie modelov

Ukázali sme, že obidva modely sa dajú použiť na veľmi veľkú časť zdravotníctva a majú dobrú schopnosť identifikovať podozrivých poskytovateľov.

Jedným z veľkých prekážok pri analýze dát zdravotnej poisťovne Union sa ukázal byť jej podiel na slovenskom trhu, ktorý sa pohybuje približne okolo desiatich percent. To znamená, že v dátach sa priemerne nachádza len desať percent všetkej preplatenej zdravotnej starostlivosti, ktoré poskytovatelia reálne vykonali. Kvalita výstupov zo štatistických metód je preto veľmi náchylná na také druhy liečby, ktoré sú zriedkavé, ale veľmi drahé. Takýchto rôznych druhov liečby sa v zdravotníctve objavuje veľké množstvo. V prípade lineárneho modelu nám výsledky zhoršovala nadštandardná liečba v rámci jednej odbornosti. V prípade zhlukovacieho modelu je tento problém vzhľadom na veľké množstvo rôznych druhov liekov a pomôcok ešte väčší.

Regresným modelom sme sa snažili hľadať podvodníkov na základe premenných, ktoré spočítavali poistencov do skupín s rovnakou diagnózou, vekom atď. Na výrobu takýchto premenných nie je potrebné mať veľké znalosti zo zdravotníctva. Nevýhodou je vysoký počet premenných, ktoré sú potrebné na dobré fungovanie modelov. Zhlukovací model využíva oveľa menší počet premenných, avšak na ich výrobu sú potrebné skúsenosti s oblasťou zdravotníctva. Taktiež na ich testovanie je potrebné veľké množstvo času. Kvalita zhlukovacieho modelu je pritom od premenných vysoko závislá. Zhlukovací model sa dá použiť aj na oblasti zdravotníctva, v ktorých sa nenachádza veľké množstvo poskytovateľov, teda napríklad aj na odbornosti v ŠAS, na ktoré sme nemohli použiť regresný model.

Ďalšou výhodou zhlukovacieho modelu je to, že jeho výstupy sú priamo podložené identifikátormi podvodu a preto po identifikácii podozrivých PZS ihneď vieme, prečo boli identifikovaní a teda aj čo je potrebné u nich skontrolovať.

Výsledky regresného modelu sú síce tiež podložené počtami poistencov, ktorí navštívili daného PZS, ale v realite môže byť veľa vecí, ktoré môžu vysvetliť, prečo bol daný PZS modelom identifikovaný. Preto je po identifikácii PZS týmto modelom potrebná dodatočná kontrola odborníkmi. Interpretovateľnosť lineárnej regresie je vo všeobecnosti veľmi dobrá. Pri takom veľkom počte prediktorov, aký používame je však interpretovateľnosť dosť komplikovaná. Rozdiel uvidíme hlavne keď porovnáme interpretovateľnosť

so zhlukovacím modelom, ktorý je interpretovateľný okamžite bez žiadnych ďalších medzikrokov. Obidva modely majú rôzne výhody a nevýhody. V tejto práci sme však ukázali, že pri správnom použití má každý z nich svoje miesto v hľadaní podvodných poskytovateľov zdravotnej starostlivosti.

Záver

Cieľom tejto diplomovej práce bolo vytvoriť vlastný model na odhaľovanie podvodov u poskytovateľov zdravotnej starostlivosti v špecializovanej a všeobecnej ambulantnej starostlivosti, vytvorený model otestovať v praxi a postupne ho dolaďovať.

V prvej kapitole sme čitateľa uviedli do problematiky a v druhej kapitole sme opísali na základe článkov [2], [3], [5] a [8], ktoré poskytujú prehľad literatúry o podvodoch v zdravotníctve, niektoré z používaných metód a postupov. Táto problematika je riešená v mnohých publikáciách. Väčšina z publikácií sa však zaoberá odhaľovaním podvodov iným spôsobom, v inej oblasti alebo len v jednej špecializácii. V tejto práci sme sa pokúšali vytvoriť model na identifikáciu podvodných poskytovateľov zdravotnej starostlivosti, ktorý by mohol pokryť čo najväčšiu časť zdravotnej starostlivosti.

Teóriu používaných metód sme popísali v tretej kapitole. Okrem zhlukovacieho algoritmu DBSCAN sme uviedli aj algoritmus na hľadanie ideálneho regresného modelu pomocou informačných kritérií - stepwise algoritmus.

Náš model lineárnej regresie, ktorý popisujeme vo štvrtej kapitole, sme vytvorili tak, aby pokrýval čo najviac odborností špecializovanej a všeobecnej zdravotnej starostlivosti a aby sa dal aplikovať na podvody v zdravotných výkonoch a aj na predpis liekov a pomôcok. Na výber premenných, ktorý je veľmi dôležitou časťou modelovania, sme použili stepwise algoritmus.

Náš lineárny regresný model mal na začiatku mnohé slabé stránky. Po kontrole niekoľkých identifikovaných poskytovateľov s pomocou revíznych lekárov sme získali množstvo informácií a indikátorov, ktoré pomohli odstrániť alebo zmierniť niektoré slabé stránky a tým pádom predikciu tohto modelu vylepšiť. Taktiež sme poukázali na možnosť použitia testu na outlierov a studentizovaných rezíduí, ktoré majú potenciál úspešnosť predikcie modelu posunúť z 50% až na 60% alebo aj vyššie. To je vzhľadom na všeobecnosť modelu a na to, že ide o model bez učiteľa veľmi dobrá úspešnosť.

V praktickej časti sme opísali okrem lineárneho regresného modelu aj zhlukovací model vytvorený pomocou algoritmu DBSCAN. Tento model sme aplikovali na dáta o výdajoch liekov a pomôcok, ale aj na dáta o výkonoch všeobecných lekárov. Tento model je postavený na základe identifikátorov podvodu, čo mu poskytuje veľmi dobrú interpretáciu. Jeho výstupy sú vysoko závislé od vhodného použitia a od kvality identi-

fikátorov podvodu. Na základe dát o externých revíziách sme však ukázali, že pri vhodnom použití môže tento model slúžiť ako veľmi dobrá pomôcka na odhaľovanie podozrivých poskytovateľov.

V tejto diplomovej práci sme sa zaoberali témou, ktorá je v súčasnosti veľmi populárna. Existuje mnoho prístupov a metód, ktoré by sme mohli ďalej otestovať a prácu tak rozšíriť. Zameraním sa na špecifickú odbornosť v lineárnom regresnom modeli by sme mohli vymyslieť nové prediktory a vylepšiť tak úspešnosť predikcie. Zhlukovací model taktiež poskytuje obrovské množstvo priestoru na testovanie starých a vymýšľanie nových premenných. Priestor by sa našiel aj na dolaďovanie parametrov modelu a váhovanie premenných. Vykreslenie dát pri zhlukovacom algoritme nám pomohlo prísť aj na ďalšie možné postupy. Jedným z nich bola obálková analýza dát.

V tejto práci sme skúmali poskytovateľov zdravotnej starostlivosti ako celok. Zameraním sa na jednotlivé riadky by sa nám otvorili možnosti k vytváraniu rôznych predikčných modelov, ako sú napríklad neurónové siete či klasifikačné stromy. Taktiež by sme sa namiesto riadkov mohli zamerať aj na poistencov, alebo na iné zložky zdravotníctva. Možností ako postupovať ďalej je teda obrovské množstvo a každá z nich by k odhaľovaniu podvodov mohla priniesť niečo nové.

Zoznam použitej literatúry

[1] Rawte, V. & Anuradha, G.: *Fraud detection in health insurance using data mining techniques*, Communication Information & Computing Technology (ICCICT 2015), pp. 1-5, 2015

[2] Ebru Aydoğan Duman & Şeref Sağıroğlu: *Health care fraud detection and new approaches*, Computer Science and Engineering (UBMK) 2017 International Conference on, pp. 839-844, 2017

[3] Li, J. & Huang, K.-Y. & Jin, J. & Shi, J.: *A survey on statistical methods for health care fraud detection*, Health Care Manag Sci, 11 (3), pp. 275-287, 2008

[4] Joudaki, H.& Rashidian, A.& Minaei-Bidgoli, B.: *Improving fraud and abuse detection in general physician claims: a data mining study*, International Journal of Health Policy and Management, 5(3), pp.165–172,2016

[5] Waghade, S. S. & Karandikar, A. M. .: *A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning*, International Journal of Applied Engineering Research, 13 , pp. 4175-4178 ,2018

[6] OECD : *Fiscal Sustainability of Health Systems*, OECD Publishing, Paríž, 2015

[7] Wasserman, L.: *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York, 2004

[8] Joudaki, H.& Rashidian, A.& Minaei-Bidgoli, B.& Mahmoodi, M.& Geraili, B.& Nasiri, M.: *Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature*, Glob J Health Sci., 2015

[9] Ester, M. & Kriegel, H. P.& Sander, J.& Xiaowei Xu: *A density-based algorithm for discovering clusters in large spatial databases with noise*, KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,pp. 226-231, 1996

[10] Rizman, D.: *Základné metódy Datamininguu*, Bakalárska práca, FMFI UK, Bratislava, 2017

[11] Rud, O. P.: *Data Mining Cookbook*, John Wiley & Sons Inc., New York, 2000

[12] Jolliffe, I.T.: *Principal Component Analysis*, Springer-Verlag, New York, 2002

[13] Cook, D. R. & Weisberg, S.: *Residuals and Influence in Regression* , Chapman and Hall , New York, 1982

[14] Fox, J.: *Course on regression diagnostics*, 2009, dostupné na internete (21.4.2017), http://socserv.socsci.mcmaster.ca/jfox/Courses/Brazil-2009/index.html

[15] Santrová, A.: *Návrh a optimálizácia algoritmov na odhaľovanie fraudov v dátach Zdravotnej poisťovne* , Diplomová práca, FMFI UK, Bratislava, 2017

[16] Kursa, B.M.& Rudnicki, R. W.: *Feature Selection with the Boruta Package*, Journal of Statistical Software, vol. 36, issue 11, 2010