Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Krištof Zubricky
Norbert Varga

Predikcia mŕtvice

Správa o riešení semestrálneho projektu z predmetu OZNAL

Predmet:	Objavovanie znalostí - OZNAL
Zodpovedný za predmet:	doc. Mgr. Michal Kováč, PhD.

Obsah

Úvod

Tento dokument je prezentáciou výsledkov práce počas semestra na predmete OZNAL. Celková práca je zdokumentovaná v priložených notebookoch. Dokument začína informáciami k spôsobu práce na projekte, ako aj informáciami k odovzdaným súborom. V nasledujúcich častiach sú popísané výstupy z prieskumnej analýzy, riešenie nedostatkov v dátach a pokusy o predikciu mŕtvice.

Informácie k projektu

Táto časť obsahuje základné informácie k projektu, ako podiel práce oboch študentov, použitý jazyk a prostredie a popis odovzdaných súborov. Tiež sú stručne popísané dva datasety, na ktorých sa počas semestra pracovalo.

Podiel práce

Na všetkých častiach práce študenti pracovali spoločne a do spoločného repozitára posielali výsledky striedavo.

Viac informácií je možné zistiť na samotnom GitHub repozitári: Network Graph · xvargan/oznal-projekt (github.com)

Použité technológie

Projekt bol riešený v jazyku R a použité bolo prostredie Jupyter Notebook.

Dataset 1 - Free Wifi

Pôvodne boli vybraté dáta z Open Data Bratislava. Plánom bolo vyskladať dataset z dostupných dát k Free Wifi a následne nad ním riešiť projekt. Po vytvorení datasetu sa však ukázalo, že dáta majú často nízky prekryv, čo spôsobovalo množstvo chýbajúcich hodnôt. Na základe konzultácií po prvej prezentácii bolo teda prijaté rozhodnutie prejsť na iný, predpripravený dataset. Proces vytvárania Free Wifi datasetu je zdokumentovaný v samostatnom Jupyter Notebooku.

Dataset 2 - Stroke Prediction Dataset

Po prvom neúspešnom pokuse bol vybratý dataset Stroke Prediction Dataset | Kaggle, nad ktorým bol nakoniec vypracovaný projekt.

Prieskumná analýza

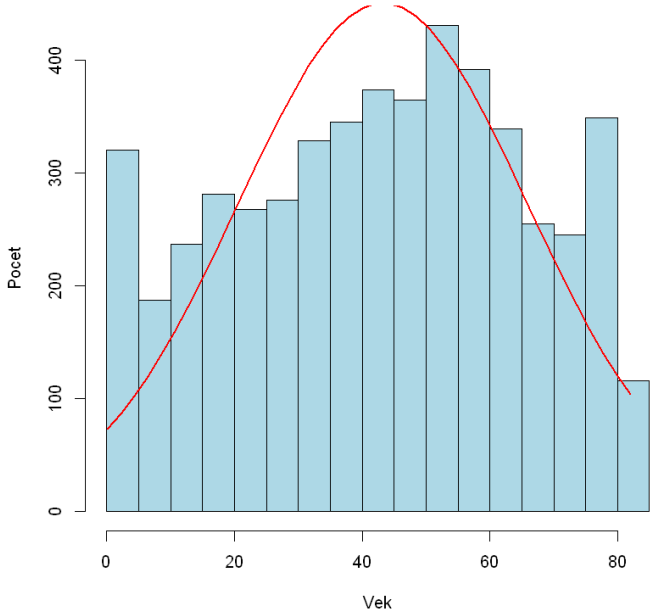
Táto časť popisuje výsledky prieskumnej analýzy. Postupne sú opísané dáta, identifikované vzťahy v dátach a hypotézy, ktoré boli na základe identifikovaných vzťahov stanovené.

Informácie o dátach

Dataset obsahuje 3 numerické atribúty, 8 kategorických. Ich vysvetlivky sú na [Stroke Prediction Dataset | Kaggle](). Počet riadkov je 5110.

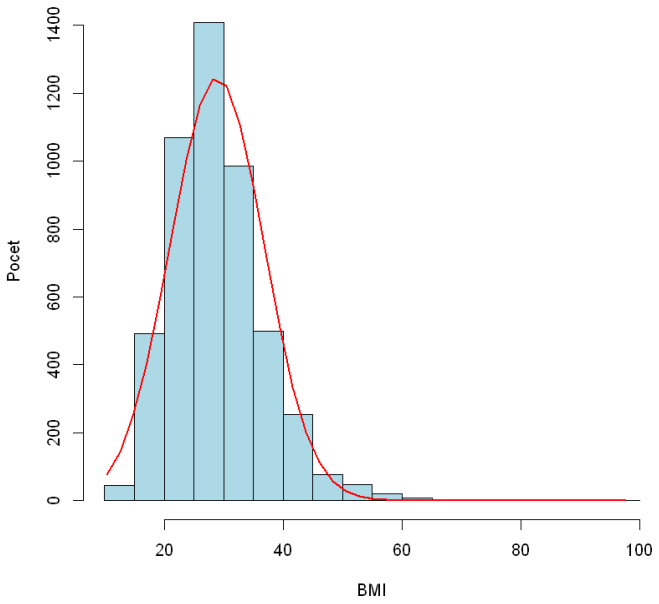
Numerické atribúty predstavujú vek, BMI a úroveň glukózy v krvi. Žiadny z týchto atribútov nie je normálne rozdelený.
V atribúte Pohlavie bol jeden riadok s hodnotou unknown - bol odstránený.

Vek
Najmladší respondent má len 0.08 rokov, teda je ešte novorodenec. Najstarší človek v datasete má 82 rokov, priemerný vek je 43 rokov, stredná hodnota je 45, najčastejšie sa vyskytujúca hodnota je 78. Smerodajná odchýlka je 22.6, koeficient asymetrie je -0.13, teda naľavo od priemeru sa vyskytujú vzdialenejšie hodnoty ako napravo a väčšina hodnôt sa nachádza viac vpravo od priemeru. Špicatosť je 2, teda je špicatejší ako krivka normálneho rozdelenia. Na Obr. 1 je histogram pre Vek v porovnaní s normálnym rozdelením. Na grafe môžeme vidieť, že veľa hodnôt je zhromaždených na začiatku, v strede a na konci. Z toho môže predpokladať, že vek nebude normálne rozdelený.



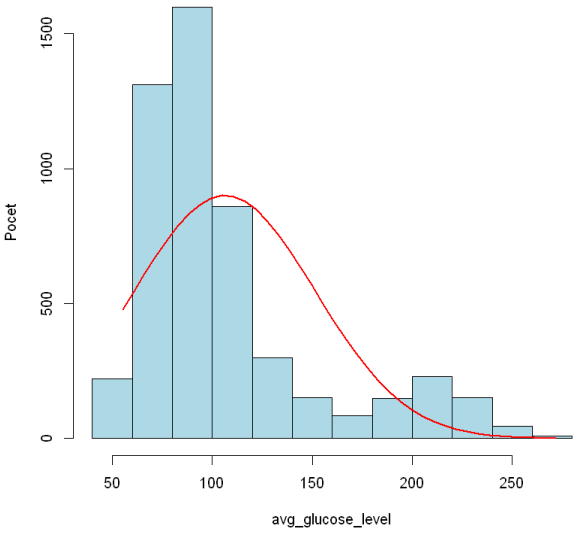
BMI
Počet riadkov, ktoré neobsahujú NA hodnoty je 4908, minimálna hodnota je 10.3, maximálna hodnota je 97.6. Priemerný BMI index je 28.8, stredná hodnota je 28.9, najčastejšie sa vyskytujúca hodnota je 28.7. Smerodajná odchýlka je 7.85, koeficient asymetrie je 1.05, teda napravo od priemeru sa vyskytujú vzdialenejšie hodnoty ako naľavo a väčšina hodnôt sa

nachádza viac vľavo od priemeru. Špicatosť je 6.36, teda je špicatejší ako krivka normálneho rozdelenia. Na Obr. 2 je histogram pre BMI v porovnaní s normálnym rozdelením. Z grafu je vidno, že v BMI budú vychýlené vysoké hodnoty. Takéto hodnoty sú však reálne a tak ich nepovažujeme za chybu v dátach.



Glukóza

Minimálna priemerná hodnota glukózy je 55.12, maximálna hodnota je 271.74. Priemerný hodnota je 106.14, medián je 91.88, najčastejšie sa vyskytujúca hodnota je 93.88. Smerodajná odchýlka je 45.29, koeficient asymetrie je 1.57, teda napravo od priemeru sa vyskytujú vzdialenejšie hodnoty ako naľavo a väčšina hodnôt sa nachádza viac vľavo od priemeru. Špicatosť je 4.68, teda je špicatejší ako krivka normálneho rozdelenia. Na Obr. 3 je histogram pre Glukózu v porovnaní s normálnym rozdelením.



Transformácia na normálne rozdelenie
Na základe absolútnej hodnoty koeficientu asymetrie a znamienka pred ním sme rôznymi technikami skúsili transformovať numerické atribúty do normálneho rozdelenia. BMI bolo transformované pomocou log10(x) a Glukóza bola transformovaná pomocou 1/x. Vek sa transformovať nepodarilo.

Vzťahy v dátach

Jednotlivé stĺpce boli analyzované párovo, každý s každým, s cieľom zistiť vzájomné korelácie, ale hlavne súvislosti s výskytom mŕtvice. Zistené boli hlavne samozrejmé skutočnosti, ako napríklad to, že pacienti v manželskom zväzku sú starší ako slobodní, prípadne že mladí ľudia (deti) sú skoro výlučne nefajčiari.

Neboli zistené žiadne spoločné znaky pacientov s mŕtvicou.

Pri hľadaní súvislostí s mŕtvicou však boli zistené nasledovné skutočnosti:

1) Mŕtvicu malo viac starších pacientov
2) Mŕtvicu malo viac pacientov s nižšou hladinou glukózy
3) Mŕtvicu malo viac pacientov s vysokým krvným tlakom
4) Mŕtvicu malo viac bývalých fajčiarov
5) Mŕtvicu malo viac self-employed ľudí
6) Mŕtvicu malo viac pacientov, ktorí majú aj ochorenie srdca
7) Mŕtvicu malo viac pacientov, ktorí sú v manželskom zväzku

Hypotézy

Na základe identifikovaných súvislostí v dátach boli sformulované nasledovné hypotézy:

1) Vek ovplyvňuje mozgovú mŕtvicu.
2) BMI ovplyvňuje mozgovú mŕtvicu.
3) Manželstvo ovplyvňuje mozgovú mŕtvicu.
4) Vysoký krvný tlak ovplyvňuje mozgovú mŕtvicu.
5) Hladina glukózy ovplyvňuje mozgovú mŕtvicu.

Problémy v dátach

V tejto časti sú popísané identifikované problémy v dátach a spôsob ich riešenia.

Chýbajúce hodnoty

Stĺpce smoking status a BMI obsahovali chýbajúce hodnoty. Keďže veľa chýbajúcich hodnôt pre smoking status patrilo deťom, nahradili sme ich hodnotou never_smoked. Chýbajúce BMI bolo nahradené priemerom.

Nevyvážený dataset

Pozorovania v datasete sú nevyvážené. Dataset obsahuje podstatne viac pozorovaní, ktoré nemali mŕtvicu ako pozorovaní, ktoré ju mali. Konkrétne je 249 pozorovaní s mŕtvicou a 4860 bez. Toto je problém pre trénovanie, pretože výsledná predikcia bude odrážať toto rozdelenie dát. Toto sme si uvedomili až počas trénovania modelov.

Tento problém riešime tromi prístupmi - oversampling, undersampling a kombináciou oboch. Následne testujeme natrénovaný model na pôvodnom datasete.

Kategorické dáta

Keďže ML algoritmy rozumejú hlavne číslam, kategorické dáta (gender, ever_married, work_type, Residence_type, smoking_status) sme prekonvertovali na čísla, aby tomu rozumel aj stroj. Vzhľadom na to, že majú tieto kategorické atribúty málo množín hodnôt, ktoré môžu nadobudnúť, manuálne sme ich prekonvertovali. Napr. v prípade atribútu „work_type“ hodnota „children“ je zakódovaná ako 0, hodnota „Govt_job“ ako 1 atď. Okrem manuálneho zakódovania jednotlivých hodnôt, sme vyskúšali aj „one hot encoding“ pomocou knižnice „caret“.

Predikcie

Našim hlavným cieľom bolo predikovať mŕtvicu na základe dát, ktoré máme k dispozícií. Snažili sme sa pri tom dosiahnuť čo najvyššiu presnosť.

Prvým krokom bolo rozdeliť náš dataset na 2 časti:

- trénovací dataset
- testovací dataset.

Na prvom (trénovacom) datasete sme natrénovali náš model a na testovacom sme model otestovali a výsledky sme vyhodnotili. Pomer rozdelenia celkového datasetu je 75:25. Trénovací dataset obsahuje 182 záznamov, kde ľudia mali mŕtvicu, kým testovací dataset 67.

Modely sme vyhodnocovali podľa metrík recall a precision.

Vďaka nevyváženosti datasetu však predikcie dosahovali extrémne nízke hodnoty metrík, hoci veľká väčšina predikcií bola správna.

Preto boli dáta upravené pomocou oversamplingu, undersamplingu a kombináciou. Oversampling bol jednoduchý - duplikoval existujúce riadky. Vznikli tak tri trénovacie datasety a modely boli testované na pôvodných dátach.

Lineárny model

Na predikciu mŕtvice sme sa rozhodli použiť lineárny model. Využili sme na to knižnicu „caret“. Pomocou funkcie glm() sme vytvorili lineárny model, kde sme ako argument určili dataset a atribút, ktorý sledujeme (v našom prípade je to atribút „stroke“). Následne sme model natrénovali a otestovali.

Naivný Bayes

Ako druhý klasifikátor sme použili Naive Bayes Classifier, ktorý bol testovaný a trénovaný rovnakými spôsobmi.

Výsledky

Výsledné klasifikátory dosahovali v oboch metrikách takmer 80% na trénovacích dátach, no na testovacích dátach to bolo podstatne horšie. Hoci metrika recall stále dosahovala aspoň 70%, presnosť klesla medzi 10% až 15%. Metriky sa nepodarilo zvýšiť ani vypustením rôznych stĺpcov, či rôznymi nastaveniami parametrov. Možné dôvody, prečo bola presnosť tak nízka, sú rozobraté v nasledujúcej kapitole.

Záver

Nepodarilo sa natrénovať dostatočne efektívny model pomocou metód, ktoré sme vyskúšali. Predpokladáme, že jedným z dôvodov je skutočnosť, že pacienti v datasete, ktorí mali mŕtvicu, nemajú žiadnu výraznú spoločnú črtu, alebo sa nám ju aspoň nepodarilo objaviť.

Porovnali sme tiež vlastnú prácu s prácou skúsenejšieho analytika na SMOTE + Voting Classifier + Moving Threshold = 96% | Kaggle, ktorý dosahuje presnosť okolo 96%. Predspracovanie dát z práce je podobné nášmu, ale je precíznejšie - napríklad atribút BMI je transformovaný na kategórie a vek s glukózou sú naškálované na podobný rozsah, atď. Tiež rôzne parametre pri tvorbe modelov sú starostlivo ponastavované, v čom sa odrážajú spomínané vyššie skúsenosti. Hlavný rozdiel oproti nášmu prístupu však vidíme v použití metódy SMOTE pri oversamplovaní. Táto metóda vytvára syntetické dáta namiesto replikácie existujúcich, čím prináša do datasetu novú informáciu. Za druhý významný rozdiel považujeme použitie komplexného klasifikátora (Voting Classifier), ktorý využíva hneď niekoľko rôznych modelov.

Hoci sa v rámci práce na tomto projekte nepodarilo prísť s použiteľným klasifikátorom, dovoľujeme si konštatovať, že projekt splnil svoj účel v zmysle, že sme sa mnohé naučili a získali užitočnú skúsenosť. V tomto smere hlavne pozitívne hodnotíme možnosť porovnať sa so skúsenejšími riešiteľmi z Kaggle.