Rozšírené zadanie diplomovej práce

Názov: Predikcia nákladov zdravotnej starostlivosti pomocou vybraných metód strojového učenia

Autor: Bc. Pavel Lukačik

Vedúci práce: RNDr. Ľubomír Antoni, PhD.

Ciele:

1. Vytvoriť prehľad aktuálne existujúcich metód strojového učenia na predikciu nákladov zdravotnej starostlivosti.
2. Navrhnúť a implementovať kombináciu metód strojového učenia vhodnú na riešenie predikcie nákladov zdravotnej starostlivosti.
3. Porovnať presnosť predikcie navrhnutého riešenia s inými dostupnými štúdiami.

V USA národné výdavky za zdravotníctvo stúpli v roku 2019 o 4.8% na 3.2 biliónov dolárov (11 582 dolárov na osobu), čo predstavovalo 17.7% hrubého domáceho produktu [1]. Pri hľadaní spôsobu ako kontrolovať takéto neudržateľné nárasty nákladov na zdravotnú starostlivosť je nevyhnutné, aby zdravotnícke organizácie vedeli predikovať možné budúce náklady jednotlivých pacientov, aby sa mohli efektívne rozdeliť finančné prostriedky pacientom s pravdepodobnými vysokými nákladmi. Dostatočne presné predikovanie nákladov na zdravotnú starostlivosť je teda dôležité pre poisťovne, ale aj ďalšie zúčastnené strany, napríklad pacienti, ktorí budú poznať ich možné zdravotné náklady na ďalší rok si dokážu zvoliť vhodnejšie zdravotné poistenie.

Budeme vychádzať a porovnávať sa hlavne s dvoma článkami, a to článkom [2], ktorý popisuje experimenty na MIT, kde problém predikcie zdravotných nákladov klasifikačné rozhodovacie stromy a klasterizáciu. Zistili, že pre dostatočne presné naučenie modelov stačia nákladové atribúty. Medicínske (údaje o vyšetreniach) a ostatné (vek, pohlavie, ...) zásadne nepomohli presnosti predikcie, dokázali pomôcť iba pri pacientoch s veľmi vysokými nákladmi aj to iba pri metóde klasterizácie. Ďalší článok [3] pochádza z univerzity v Utahu, kde vytvorili prehľad metód vhodných pre daný problém. Tento článok sa odkazuje a porovnáva s článkom [2], ale aj inými a prezentuje aj ďalšie modely. Ako najpresnejšie boli vyhodnotené metódy Gradient boosting, Neurónová sieť a Ridge. Článok tvrdí, že supervised learning, teda učenie s učiteľom sa najviac hodí na predikciu nákladov zdravotnej starostlivosti, a podobne ako článok [2] tvrdí, že úplne stačia nákladové atribúty, no v závere uvažuje, že pri použití novších modelov ako deep learning alebo štrukturálna analýza by mohli medicínske atribúty pomôcť.

Najprv plánujeme implementovať dva typy Baseline metód. Jedna ako predikciu nákladov pre nasledujúci rok zoberie sumu nákladov za posledný rok, túto metódu použil aj článok [2]. Druhá zasa zoberie priemer predchádzajúcich troch rokov a ten bude brať ako

predikciu pre nasledujúci rok. Ďalej plánujeme vyskúšať rôzne modely neurónovej siete, pričom začneme doprednou sieťou, a ďalej využijeme metódu Gradient boosting.

Pri porovnávaní našich výsledkov s inými štúdiami použijeme metriky, ktoré používajú aj články [2][3], a teda strednú absolútnu percentuálnu chybu odhadu, R^2, mieru úspešnosti a absolútnu chybu predikcie.

Ako dátovú sadu máme k dispozícii reálne anonymizované údaje z nemenovanej súkromnej zdravotnej poisťovne so vzorkou 17 000 poistencov. Dáta sú z obdobia január 2010 až december 2013 a obsahujú len pacientov s plnou históriou v danom období. Ku každému pacientovi teda máme údaj o sumárnej cene za lieky v danom mesiaci. Lieky sú z ATC (Anatomicko-terapeuticko-chemického) klasifikačného systému, kde lieky sú klasifikované do skupín podľa toho, na aký orgán alebo sústavu sú zamerané. Keďže máme k dispozícii 4 plné roky záznamov, tak trénovať budeme na prvých troch rokoch (2010, 2011, 2012) a na štvrtom roku (2013) budeme testovať presnosť našej predikcie. Zatiaľ máme k dispozícii len nákladové atribúty no neskôr sa pokúsime pridať aj ďalšie, ako vek, pohlavie, atď.

Pri práci budeme používať programovací jazyk Python s knižnicami, a to Keras pre prácu s neurónovými sieťami a XGBoost pre prácu s Gradient boosting metódou.

Zdroje:

1. The Centers for Medicare & Medicaid Services (CMS) DoHaHS, United States. National Health Expenditure Data 2019. Dostupné na: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet
2. BERTSIMAS, D., BJARNADOTTIR, M. V., KANE, M. A. a kol. (2008). Algorithmic prediction of health care costs. Operations Research, 56(6), 1382 1392.
3. MORID, M.A, KAWAMATO, K. a kol. (2017). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation.AMIA Annu Symp Proc. 2017; 2017: 1312 1321.