Predikcia nákladov zdravotnej starostlivosti pomocou vybraných metód strojového učenia

Analýza a návrh riešenia, výsledky

Autor: Bc. Pavel Lukačik
Vedúci práce: RNDr. Ľubomír Antoni, PhD.

Úvod

Rastúce náklady na zdravotnú starostlivosť predstavujú jednu z aktuálnych výziev vo svetovom meradle. Jedným z možných nástrojov na kontrolu nákladov na zdravotnú starostlivosť je presná predikcia pravdepodobných budúcich nákladov jednotlivcov na zdravotnú starostlivosť. Prediktívne modelovanie s využitím metód umelej inteligencie môže zvýšiť účinnosť systému zdravotnej starostlivosti a umožniť produktívnejšiu alokáciu zdrojov. V práci predstavujeme oblasť umelej inteligencie, strojového učenia a dátovej vedy, popisujeme výpočtové metódy umelej inteligencie v predikcii nákladov zdravotnej starostlivosti. Prezentujeme metódy regresných stromov, umelých neurónových sietí, konvolučných neurónových sietí a metódy vylepšovania presnosti pomocou metód bagging a boosting. Ďalej pokračujeme popisom ukazovateľov hodnotenia úspešnosti metód umelej inteligencie pri predikcii nákladov zdravotnej starostlivosti a K-zložkovej krížovej validácie. Prácu uzatvárame popisom trénovania, našimi výsledkami a porovnaním s výsledkami v článkoch.

1. Umelá inteligencia a dátová veda

Umelá inteligencia je vedná disciplína, ktorej cieľom je automatizácia intelektuálnych úloh pomocou inteligentných strojov. Inteligentné stroje pomocou algoritmu vytvoreného človekom napodobňujú ľudské kognitívne funkcie a správanie, ako je napríklad učenie alebo riešenie problémov. Typickým príkladom je riadenie robotov, autonómna navigácia automobilov, rozpoznávanie reči alebo obrazu alebo súťaženie s človekom v strategických hrách. V roku 2019 vzniklo Slovenské centrum pre výskum umelej inteligencie ako platforma pre excelenciu v umelej inteligencii, ktorá prepája študentov, výskumníkov, podnikateľov, učiteľov, investorov a všetkých ďalších, ktorí sa zaujímajú o umelú inteligenciu [1].

Umelú inteligenciu môžeme vnímať aj širšie. Môžeme hovoriť o digitálnej inteligencii, ktorá integruje do jedného spoločného celku umelú inteligenciu, cloudové počítanie, internet všetkého, virtuálnu realitu a niektoré ďalšie oblasti [2].

Kým v počiatkoch sa počítače používali na spracovanie číselných údajov, v posledných desaťročiach sa presadila aj reprezentácia vo forme relačných údajov, pretože umožňuje skúmať vzťahy medzi inštanciami objektov v oveľa širšom rozmere. Navyše znalosti objavené v údajoch pomáhajú riešiť mnohé problémy týkajúce sa nielen vedy, ale aj podnikania či špecifických problémov spoločnosti. Neustála modifikácia vstupných údajov do formy vedomostí a znalostí je preto prospešná a nevyhnutná. Pomerne často sa údaje vyskytujú v tabuľkovej podobe, pričom takúto formu údajov môžeme získavať zaznamenávaním výsledkov vyšetrení u lekára, správania sa zákazníkov a ich spotrebiteľských návykov, či zaznamenávaním výsledkov študentov [3].

Dátová veda je interdisciplinárna oblasť, ktorá využíva vedecké metódy, procesy, algoritmy a systémy na získanie poznatkov a poznatkov z dát v rôznych formách, štruktúrovaných i neštruktúrovaných [4]. Predstavuje koncepciu zjednotenia štatistiky, analýzy údajov, strojového učenia a súvisiacich metód s cieľom analyzovať aktuálne procesy s údajmi. Dátová veda využíva techniky a postupy čerpané z oblastí matematiky, štatistiky, či informatiky [5].

Oblasť analýzy údajov je vysoko aktuálna, riešenia dátovej analýzy sa už využívajú v mnohých oblastiach technických, prírodných, humanitných a ekonomických vied. Pri riešení úloh v tejto oblasti je potrebné zaoberať sa prepojením množstva dostupných údajov, ktoré sú pravidelne generované mnohými komerčnými zariadeniami a vedeckými prístrojmi a dostupných metód ich analýzy, pomocou ktorých je možné v tých údajoch objavovať nové znalosti. Jedným z cieľov analýzy údajov je extrahovať nové, platné a potenciálne užitočné znalosti z dostupných údajov v rôznych oblastiach akademického a firemného života. K splneniu cieľov a k samotnému spoznávaniu znalostí vedie niekoľko fáz, v rámci ktorých je potrebné rozlišovať fázu pochopenia dát, prípravy dát, modelovania, vyhodnotenia výsledkov a nasadenia výsledkov do praxe.

2. Predikcia nákladov zdravotnej starostlivosti

Jedným z možných nástrojov na kontrolu nákladov na zdravotnú starostlivosť je presná predikcia pravdepodobných budúcich nákladov jednotlivcov na zdravotnú starostlivosť. Lepšie prediktívne modelovanie s využitím metód umelej inteligencie môže zlepšiť účinnosť systému zdravotnej starostlivosti a umožniť produktívnejšiu alokáciu zdrojov. Z viacerých hľadísk je prospešná aj identifikácia úrovní rizika pre rôzne skupiny obyvateľstva [6].

Medzi kľúčové organizácie v úsilí o efektívne riadenie nákladov na zdravotnú starostlivosť patria zdravotné poisťovne, poskytovatelia zdravotnej starostlivosti a tiež ďalšie organizácie. Zdravotné poisťovne disponujú komplexnými informáciami o nákladoch pacientov v predchádzajúcich obdobiach, keďže zdravotnú starostlivosť hradia jej poskytovateľom.

Pri hľadaní riešenia vzhľadom na trvalo neudržateľné zvyšovanie nákladov na zdravotnú starostlivosť by mohlo byť prospešné, ak by jednotlivé zdravotnícke organizácie mali k nahliadnutiu pravdepodobné budúce náklady jednotlivcov. Tieto informácie by tak mohli pomôcť efektívne riadiť starostlivosť o osoby s najvyšším rizikom vzniku významných nákladov.

Predpovedanie nákladov na zdravotnú starostlivosť pre jednotlivcov pomocou presných predikčných modelov je dôležité z rôznych dôvodov. Pre zdravotné poisťovne a poskytovateľov zdravotnej starostlivosti môžu presné predpovede pravdepodobných nákladov pomôcť pri všeobecnom plánovaní a tiež pri prideľovaní obmedzených zdrojov starostlivosti. Vopred známe pravdepodobné výdavky na budúci rok môžu potenciálne pomôcť pri vytvorení vhodného poistného plánu aj samotným pacientom.

Rastúce náklady na zdravotnú starostlivosť predstavujú jednu z aktuálnych výziev vo svetovom meradle. Predikcia týchto nákladov a jej presnosť poskytujú prvý krok k hľadaniu nových riešení. Od 80. rokov 20. storočia bol realizovaný výskum prediktívnych metód na modelovanie nákladov zdravotnej starostlivosti na základe údajov zdravotného poistenia s využitím heuristických pravidiel a rôznych regresných metód. Nevýhodou tohto prvotného výskumu v tejto oblasti bola absencia validácie výsledkov na rôznych skupinách obyvateľov. Z týchto dôvodov je skúmanie výpočtových metód na predikciu nákladov zdravotnej starostlivosti naďalej aktuálne a prospešné [7].

3. Výpočtové metódy na predikciu nákladov zdravotnej starostlivosti

Medicínske údaje zvyčajne vykazujú vysoký stupeň odchýlok v rámci jednotlivých pacientov a sú zvyčajne extrémne zošikmené vzhľadom na pravdepodobnosť rozdelenia. Je teda veľmi ťažké predvídať náklady zdravotnej starostlivosti na individuálnej úrovni. Situácia je na sumárnej úrovni o niečo jednoduchšia, ale výsledok ovplyvňuje množstvo zložito kvantifikovateľných faktorov, napríklad dostupnosť určitých liekov, platná legislatíva a iné faktory [8]. Na riešenie týchto problémov bolo navrhnutých viacero triviálnych metód, štatistických metód, ale aj metód umelej inteligencie, ktoré si predstavíme v nasledujúcej podkapitole.

Jednou z triviálnych metód a postupov, ako určiť náklady zdravotnej starostlivosti na nasledujúci rok, je použitie priemernej hodnoty nákladov z niekoľko predchádzajúcich rokov. Podobnou možnosťou je použiť pre triviálnu predikciu náklady posledného roka. Takýto postup je síce úspešnejší ako náhodná predikcia, ale metódy umelej inteligencie dokážu takéto triviálne predikcie veľmi ľahko prekonať.

Literárne zdroje uvádzajú tri typy metód, ktoré boli publikované pri riešení predikcie nákladov zdravotnej starostlivosti. Patria k nim rôzne metódy založené na pravidlách, štatistické metódy a kontrolované učenie v rámci metód umelej inteligencie. Medzi nevýhody metód založených na pravidlách patrí skutočnosť, že vyžadujú veľké množstvo doménových znalostí. Tieto znalosti často nie sú ľahko dostupné a ich získanie je navyše nákladné. Medzi štatistické metódy patria najmä viacnásobné regresné metódy [9]. Aj keď predstavujú výkonné nástroje na zachytenie vzťahov medzi vysvetľujúci premennými a závislou premennou, čelia dvom významným výzvam. Jednou z nich je skutočnosť, že práca s niekoľkými vysvetľujúci premennými často spôsobuje multi-kolinearitu, ktorá je charakterizovaná prítomnosťou významných korelácií medzi vysvetľujúcimi vstupnými premennými. Okrem toho je výkonnosť štatistických metód pri predikcii nákladov zdravotnej starostlivosti obmedzená z dôvodu špeciálnej povahy údajov o zdravotnej starostlivosti. Údaje o nákladoch zdravotnej starostlivosti zvyčajne vykazujú svoj hrot v nulovej hodnote, distribúcie sú silne ovplyvnené ťažkým pravým chvostom a v údajoch môžu byť prítomné extrémne hodnoty. Ak príslušné pravdepodobnostné rozdelenie údajov nie je normálne, uvedené faktory môžu spôsobiť neúčinnosť štatistických metód pri malých až stredných veľkostiach vzorky. Aj keď bolo navrhnutých niekoľko pokročilých štatistických metód na vyrovnanie šikmosti, ktorú môžeme pozorovať v údajoch o zdravotnej starostlivosti, tento typ predikčných metód nie je schopný prekonať kontrolované učenie s využitím metód umelej inteligencie, ktoré si popíšeme detailnejšie [6].

3.1. Metódy umelej inteligencie na predikciu nákladov zdravotnej starostlivosti

Strojové učenie je podoblasťou umelej inteligencie, ktorá sa zaoberá metódami a algoritmami učenia sa stroja z údajov. Cieľom strojového učenia je modelovanie algoritmov učenia sa pomocou stroja na základe vstupných dát v definovanom priestore riešení. Vo všeobecnosti, rozoznávame pri strojovom učení tri typy úloh, a teda kontrolované učenie (učenie s dozorom, učenie pod dohľadom), nekontrolované učenie (učenie bez dozoru, samostatné učenie) a učenie posilňovaním [10].

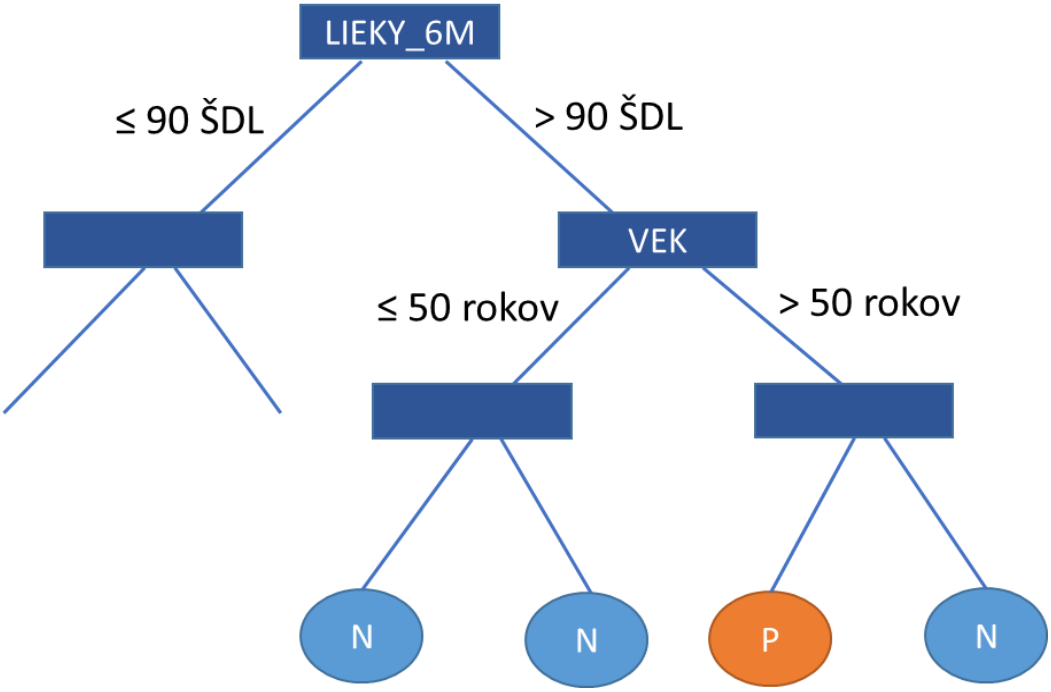
Kontrolované učenie prebieha v dvoch fázach, ktoré zahŕňa trénovanie a testovanie. Tréningovú údajovú sadu tvoria príklady, ktoré obsahujú hodnoty vysvetľujúcich premenných, ale aj hodnotu cieľovej predpovedanej premennej. V testovacej údajovej sade sú prítomné len príklady, ktoré obsahujú hodnoty vysvetľujúcich premenných. Pomocou tréningových príkladov je vybraným algoritmom vytvorený model, ktorý je použitý pre generovanie výstupov pre príklady testovacej údajovej sady. Patria tu rôzne algoritmy predikcie, ktoré sa používajú na predikciu zdravotnej starostlivosti, napríklad regresné stromy, regresná metóda k- najbližších susedov, regresná metóda podporných vektorov, neurónové siete a iné metódy [11].

Nekontrolované učenie prebieha v jednej fáze. Údajovú sadu tvoria príklady s hodnotami premenných a algoritmus hľadá vzťahy medzi vstupnými príkladmi vzhľadom na podobnosť hodnôt jednotlivých atribútov. Patria tu napríklad algoritmy zhlukovania, či asociačné pravidlá, ktoré je tiež možné použiť pri predikcii nákladov zdravotnej starostlivosti.

V poslednej dobe sa rozvíjajú aj riešenia úloh s využitím učenia sa posilňovaním, ktorá využíva pravidlá pre odmeňovanie správnych rozhodnutí. Rozvíja sa aj oblasť samo-kontrolovaného učenia, ktoré prebieha v dvoch fázach rovnako ako kontrolované učenie, ale už aj samotný tréning prebieha bez použitia hodnôt cieľovej premennej.

Regresné stromy sú populárnou metódou strojového učenia a umelej inteligencie, ktorá je založená na rekurzívnom delení priestoru príkladov na menšie časti [6]. Cieľom regresných stromov je predpovedať hodnotu cieľovej numerickej premennej na základe vysvetľujúcich vstupných premenných. Pri rekurzívnom delení priestoru príkladov pomocou najviac relevantných atribútov vzniká strom, v ktorom z každého uzla vychádzajú dve alebo viac hrán. V každom uzle tohto stromu sa príslušný príklad testuje vzhľadom na prahovú hodnotu príslušnej numerickej premennej alebo vzhľadom na triedu nominálnej premennej. V závislosti od výsledku tohto testu je pre tento príklad vybraná jedna z hrán tohto stromu. V prípade, že sa príklad nachádza v niektorom z listov vygenerovaného stromu, určí sa predpovedaná hodnota cieľovej premennej tohto

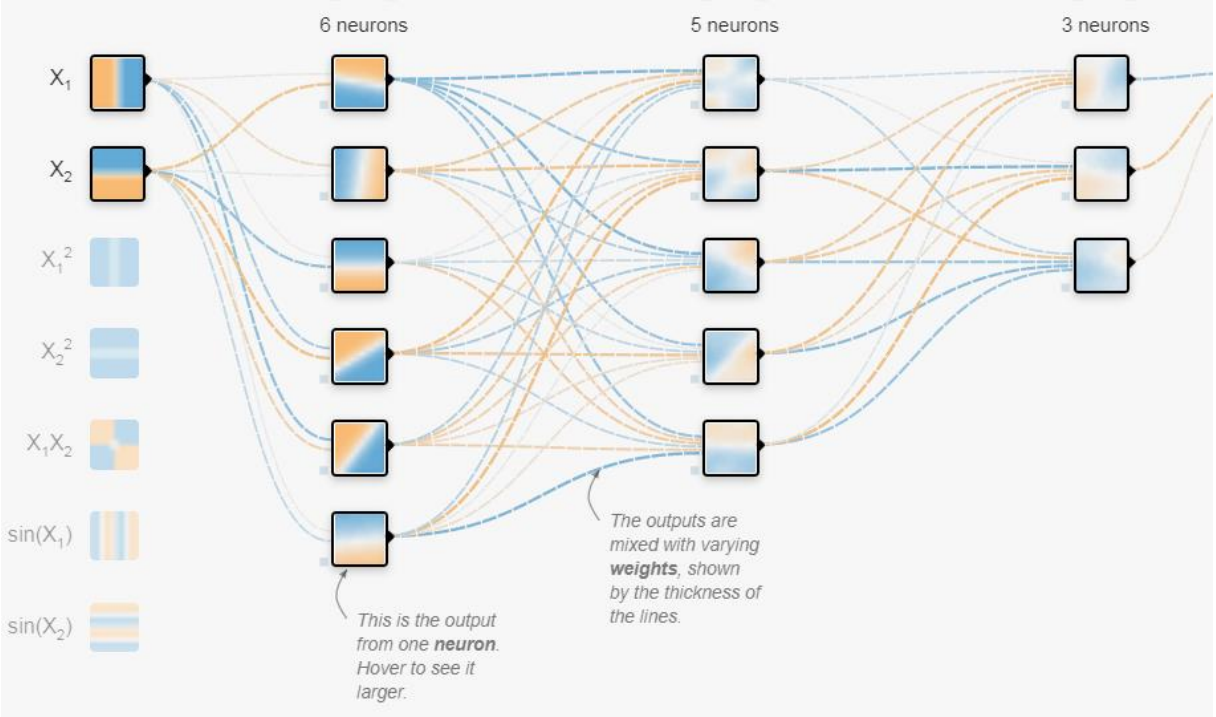
príkladu podľa vytvoreného modelu. V našej úlohe sú predpovedanou hodnotou náklady na zdravotnú starostlivosť v cieľovom období, napríklad v nasledujúcom roku, pomocou nákladov poistenca v predchádzajúcom období, napríklad v posledných troch rokoch. Metóda náhodných regresných lesov je založená na konštrukcii veľkého počtu regresných stromov, pričom každý regresný strom sa generuje nad náhodnou podmnožinou príkladov a náhodnou podmnožinou premenných. Predpovedanú hodnotu cieľovej premennej dostaneme ako priemernú hodnotu alebo medián hodnôt predpovedaných jednotlivými regresnými stromami. V špeciálnom prípade môžu vznikať len binárne stromy, teda stromy, v ktorých z každého uzla vychádzajú dve hrany. Takáto situácia nastáva, ak medzi vysvetľujúcimi premennými sú len numerické atribúty alebo nominálne atribúty s dvoma triedami.



Obr. 1 Rozhodovací strom - klasifikácia

Umelé neurónové siete sú ďalšími populárnymi metódami strojového učenia a umelej inteligencie, ktoré môžeme použiť na predikciu nákladov zdravotnej starostlivosti [6]. Model umelých neurónových sietí je tvorený súborom umelých neurónov, základných výpočtových jednotiek tohto modelu. Každý umelý neurón je spojený s jedným alebo viacerými umelými neurónmi, ktoré sú usporiadané do viacerých vrstiev. Počet vrstiev a počet neurónov v jednotlivých vrstvách tvoria architektúru umelej neurónovej siete. V každej vrstve môže byť rozličný počet neurónov, pričom počet neurónov na vstupnej vrstve zodpovedá väčšinou počtu vysvetľujúcich vstupných

premenných. V prípade našej úlohy je na výstupnej vrstve už len jeden neurón, ktorý obsahuje hodnotu predpovedaných nákladov zdravotnej starostlivosti v nasledujúcom kalendárnom roku. Takýto výstup neurónu dostaneme postupným šírením vstupných hodnôt v umelej neurónovej sieti pomocou výpočtov v jednotlivých neurónoch, ktoré sú v tejto sieti navzájom prepojené. Model vzniká na základe optimalizácie hodnôt váh, ktoré reprezentujú prepojenia medzi jednotlivými neurónmi.



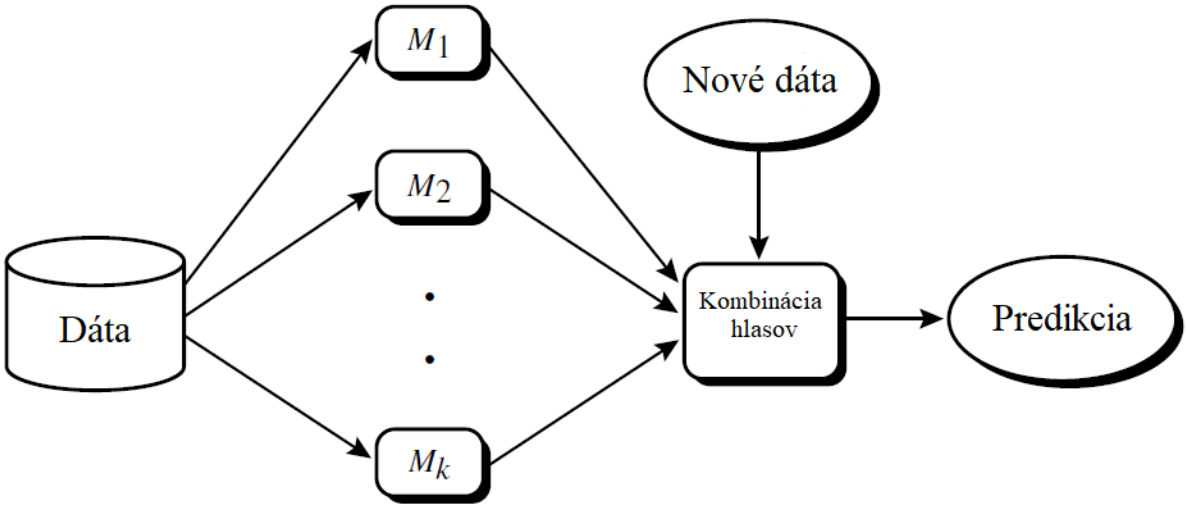
Obr. 2 Model umelej neurónovej siete

Konvolučná neurónová sieť patrí do oblasti hĺbkového učenia. V podstate je to viacvrstvový perceptrón, ktorého hlavnými výhodami sú lokálne spojenia a zdieľanie váh. Model konvolučnej siete nielen znižuje počet váh ale aj uľahčuje optimalizáciu siete. To pomáha redukovať zložitosť modelu a riziko preučenia. Model vo všeobecnosti obsahuje konvolučnú vrstvu, pooling vrstvu a plne prepojenú vrstvu, z ktorých každá sa skladá z mnohých neurónov. Spoločné váhy sa používajú na výpočet neurónov a aktivačná funkcia je použitá ako vstup do ďalšej vrstvy na nelineárny výpočet výstupu. Medzi najpoužívanejšie aktivačné funkcie patrí ReLU, ktorá zmení záporné hodnoty na nulu a kladné hodnoty ponechá bezo zmeny. Váhy sa získavajú trénovaním na dátach. Konvolučná vrstva extrahuje príznaky, vylepšuje príznaky z pôvodného signálu pomocou operácie konvolúcie a redukuje šum. Pooling vrstva komprimuje mapy príznakov, extrahuje hlavné príznaky a znižuje dimenzionalitu a preučenie.

Obr. 3 Model konvolučnej neurónovej siete

3.2. Vylepšovanie presnosti skladaním modelov

Na vylepšovanie presnosti rozhodovacích stromov, sa okrem prerezávania dajú použiť aj všeobecné stratégie na vylepšovanie presnosti modelov. Bagging a boosting sú dva príklady takzvaných ensemble metód, teda metód, ktoré používajú kombináciu modelov na predikciu alebo klasifikáciu. Každá kombinuje sériu k modelov M_1, ..., M_k, s cieľom vytvoriť vylepšený zložený model M* [11].



Obr. 4 Zložený model kombinujúci hlasy množiny vygenerovaných modelov na výslednú predikciu

Zoberme si najprv metódu bagging. Pre jednoduchosť predpokladajme, že modelom je klasifikátor. Zoberme si pacienta, ktorý chce určiť svoju diagnózu na základe svojich symptómov. No namiesto navštívenia jedného doktora môže navštíviť viacero doktorov a nechať si diagnózu od každého z nich. Ak sa konkrétna diagnóza objaví viackrát ako ostatné, môže ju určiť ako finálnu a najlepšiu diagnózu. Teda výsledná diagnóza je určená

na základe väčšinového hlasu, kde každý z doktorov má rovnocenný hlas. Ak vymeníme doktora za klasifikátor tak máme myšlienku za metódou bagging. Intuitívne väčšinový hlas väčšej skupiny doktorov môže byť spoľahlivejší ako väčšinový hlas menšej skupiny Bagging môže byť aplikovaný aj pri predikcii spojitých hodnôt, a to určením priemernej hodnoty z každej predikcie, ktorú prezentujeme ako výslednú predikciu. Klasifikátor vytvorený takýmto zložením viacerých klasifikátorov, má často výrazne vyššiu presnosť ako jediný jednoduchý klasifikátor. Zložený klasifikátor nebude ani nikdy výrazne horší a je robustnejší voči negatívnemu vplyvu zašumených dát. K zvýšeniu presnosti dochádza vďaka tomu, že zložený model znižuje rozptyl jednotlivých klasifikátorov. Pre predikciu bolo teoreticky dokázané, že takto zložený prediktor bude mať vždy lepšiu presnosť v porovnaní s jedným prediktorom.

Podobne ako pri baggingu, zoberme si pacianta, ktorý má určité symptómy a namiesto navštívenia iba jedného doktora si vyberie navštíviť a prekonzultovať svoje symptómy s viacerými doktormi. Každému doktorovi, ale priradí váhu, alebo cenu doktorovej diagnózy, založenú na presnosti predchádzajúcich diagnóz, ktoré určil. Výsledná diagnóza je potom kombinácia vážených diagnóz, čo je myšlienkou za boostingom.

Pri boostingu sa každému príkladu z tréningovej množiny priradí váha. Následne sa iteratívne buduje k klasifikátorov. Po tom čo je vytvorený klasifikátor M_i, váhy príkladov sú upravené tak, aby sa nasledujúci klasifikátor M_{i+1} sústredil viac na tie príklady, ktoré boli pomocou M_i nesprávne klasifikované. Výsledný klasifikátor M* kombinuje hlasy individuálnych operátorov, kde váha hlasu každého klasifikátora je funkciou jeho presnosti. Medzi populárne boostovacie algoritmy patrí Adaboost, Gradient boosting, Extreme gradient boosting alebo Light gradient boosting. Asi očividnou nevýhodou je výpočtová zložitosť tejto metódy. Vykonávanie mnohých iterácií, kde v každej sa vygeneruje nový model si vyžaduje veľa výpočtového času a priestoru. Fakt, že modely musia byť vytvárané postupne, zložitosti nepomáha, každopádne vytváranie jednoduchších modelov namiesto veľkých a zložitých to trochu zmierňuje. Ďalšou veľkou nevýhodou je citlivosť na zašumené dáta. Keďže algoritmus sa snaží vylepšiť presnosť výstupu pre príklady, ktoré neboli doteraz dobre klasifikované, tak keď dané dáta majú nejaké ťažko klasifikovateľné príklady, takzvaných outlierov, tak algoritmus sa bude veľmi snažiť vytvárať ďalšie modely tak, aby správne klasifikovali tento šum. Ako výsledok takejto snahy sa veľmi pravdepodobne objaví preučenie alebo overfitting. Teda narozdiel od baggingu, pri boostingu môže nastať situácia, že vytvorený zložený model bude menej presný ako samostatný jednoduchý model. Avšak pri boostingu sa zvyčajne dosahujú vyššie presnosti ako pri baggingu.

3.3. Nová implementácia skupinových modelov stromov

Gradient boosting rozhodovacích stromov je populárna metóda strojového učenia a má niekoľko efektívnych implementácií ako XGBoost a pGBRT. Bolo využitých mnoho inžinierskych optimalizácií no efektivita a škálovateľnosť je stále neuspokojivá aj je dimenzia príznakov ale atribútov vysoká, a ak je dátová sada obrovská. Hlavným dôvodom je, že pre každý príznak (stĺpec v tabuľke) je nutné preskenovať všetky inštancie (riadky tabuľky) aby sa odhadol informačný zisk všetkých možných bodov rozdelenia a to je časovo náročné.

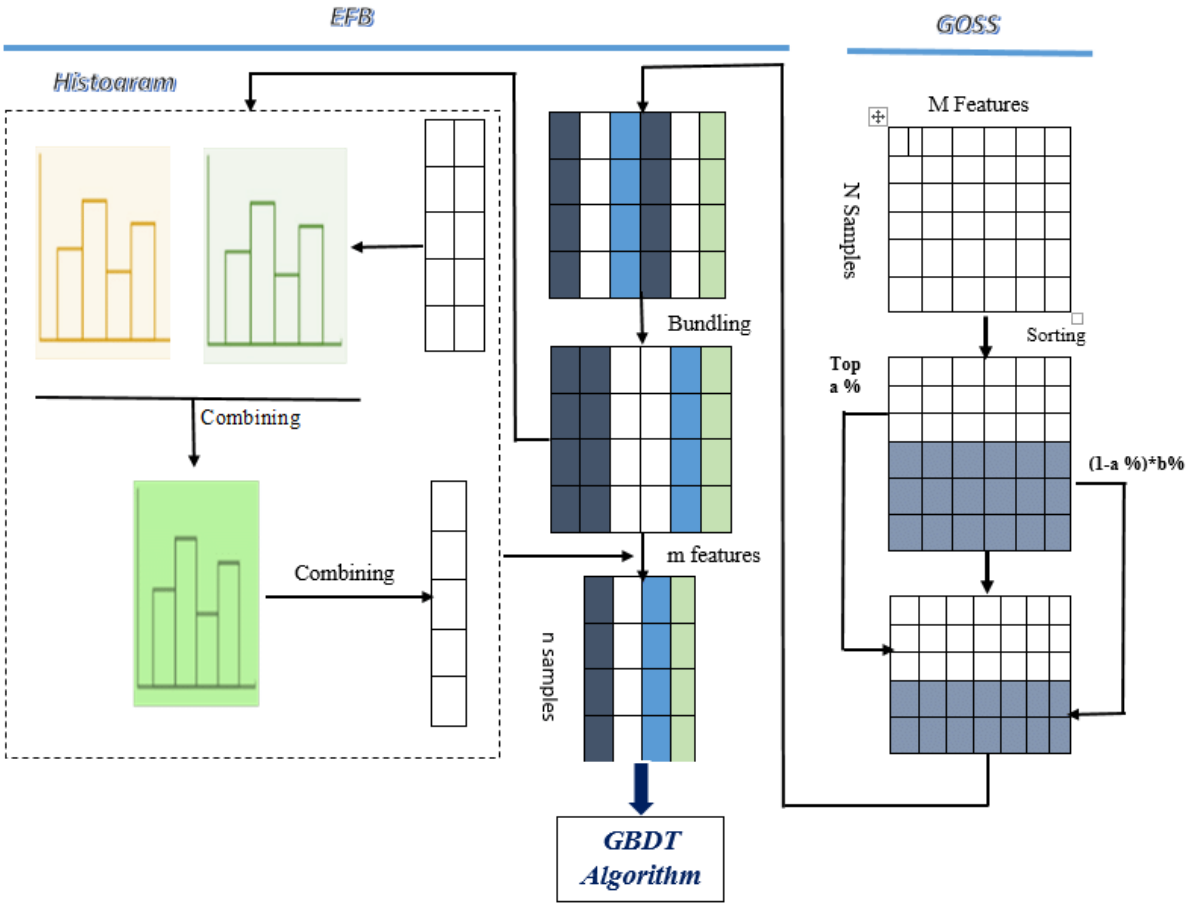
Na riešenie tohto problému autori článku [12] navrhujú dve nové techniky a to Gradient-based one-side sampling (GOSS) a Exclusive feature bundling (EFB).

Pri GOSS sa vylúči významná časť dát (riadkov) s malým gradientom a iba to čo nám zostane použijeme na odhad informačného zisku. V tomto článku je dokázané, že keďže inštancie dát s vyššími gradientmi hrajú významnejšiu úlohu vo výpočte informačného zisku, teda že GOSS dokáže získať celkom presný odhad informačného zisku s omnoho menším objemom dát.

Pomocou EFB sa zabalia vzájomne vylučujúce sa príznaky, teda tie príznaky, ktoré majú málokedy súčasne nenulové hodnoty znížili počet príznakov. Dokazujú, že nájdenie optimálneho balíka vylučujúcich sa príznakov je NP-ťažké, ale greedy, alebo pažravý algoritmus dokáže dosiahnuť celkom dobrý aproximačný pomer, teda dokáže efektívne redukovať počet príznakov bez výrazného zhoršenia presnosti deliacich bodov.

Táto implementácia gradient boostingu s GOSS a EFB sa nazýva Light Gradient Boosting Machine (LGBM). Experimenty, ktoré predviedli autori na viacerých verejných datasetoch ukazujú, že LGBM zrýchľuje trénovací čas konvenčného gradient boostingu rozhodovacích stromov až do 20-krát pri zachovaní takmer rovnakej presnosti.

Znázornenie novej implementácie skupinového modelu rozhodovacích stromov, ktorá sa zvykne označovať v literatúre ako LGBM, obsahuje obrázok 5.



Obr. 5 Znázornenie fungovania LGBM

4. Dáta

Ako dátovú sadu máme k dispozícii reálne anonymizované záznamy zo súkromnej zdravotnej poisťovne so vzorkou okolo 17 000 poistencov. Jedná sa o záznamy pacientov, ktorí sa liečia na astmu, čo je záchvatovo sa vyskytujúca, spravidla intenzívna dychová nedostatočnosť (dýchavičnosť), ktorá môže mať rôzne príčiny. Záznamy predstavujú sumárne údaje o cene liekov v konkrétnych mesiacoch obdobia január 2010 až december 2013. Lieky sú z Anatomicko-terapeuticko-chemického klasifikačného systému, kde sú lieky klasifikované do skupín podľa toho, na aký orgán alebo sústavu sú zamerané.

Pre daných pacientov máme kompletné údaje za dané obdobie, takže boli celý čas poistení, a teda pacienti, ktorí v danom období prešli do inej zdravotnej poisťovne alebo boli poistení v zahraničí boli vyradení. Zároveň platí, že každý pacient má aspoň jednoročnú históriu liečby, respektíve užívania liekov z daných ATC skupín liekov. Teda začali sa liečiť najneskôr v januári roku 2012.

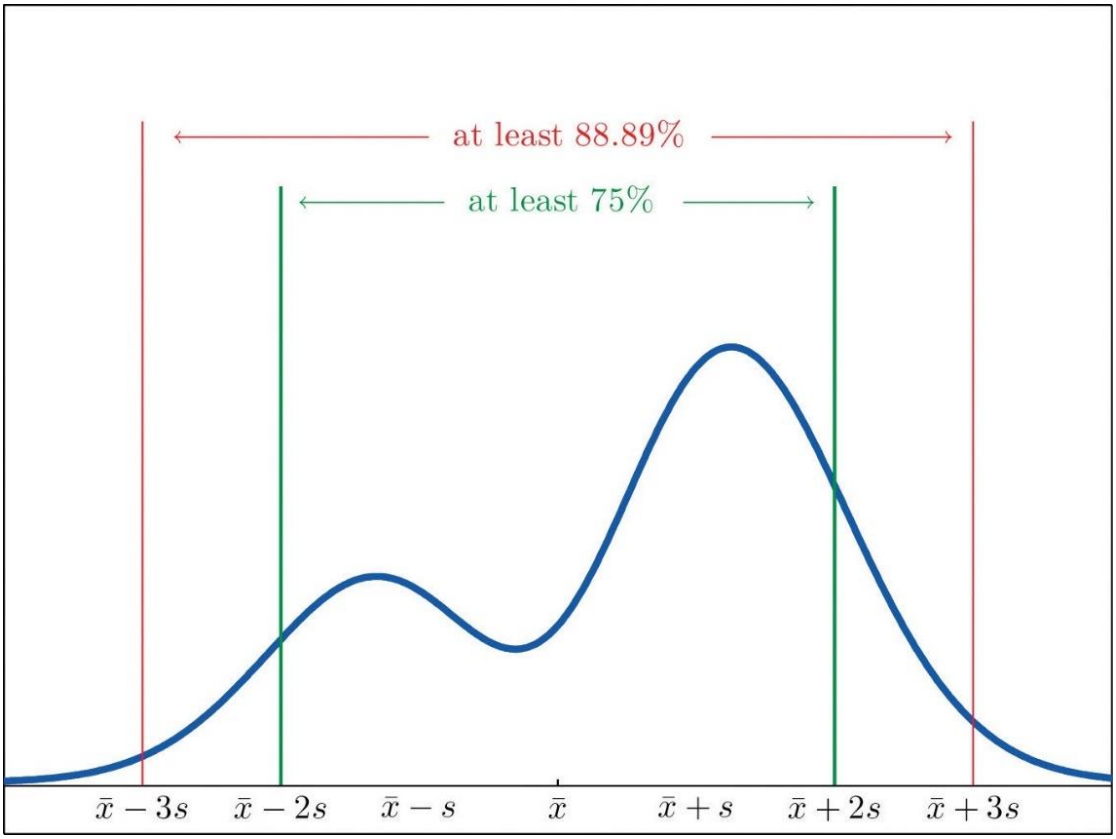
Keďže máme k dispozícii 4 plné roky záznamov, tak trénovanie bude prebiehať na prvých troch rokoch a na poslednom štvrtom roku budeme vyhodnocovať presnosť predikcie našich modelov.

4.1. Pridanie ďalších atribútov

K atribútom dátovej sady, ktorú máme k dispozícii, sme na základe článku [7] vytvorili a pridali ďalšie atribúty odvodené s použitím pôvodných atribútov pozorovacieho obdobia.

Pridané atribúty:

- overall_costs – celková suma nákladov za pozorovacie obdobie
- six_costs – suma za posledných 6 mesiacov v pozorovacom období
- three_costs – suma za posledné 3 mesiace v pozorovacom období
- trend – tento atribút sme získali tak, že sme na sumách nákladov za jednotlivé mesiace v pozorovacom období nafitovali model lineárnej regresie, teda získali sme priamku, z ktorej sme extrahovali koeficienty, čiže sklon priamky
- acute – atribút nadobúda hodnotu 0 alebo 1 a je založený na porovnaní najdrahšieho mesiaca v pozorovacom období s priemernou hodnotou všetkých mesiacov v pozorovacom období a ak významne odlišné tak tam dáme hodnotu 1, inak 0. Aby sme zistili, či sú tieto hodnoty významne odlišné sme použili Čebyševovu nerovnosť, ktorá hovorí, že nie viac ako určitý zlomok hodnôt môže mať viac ako určitú vzdialenosť od priemeru. Presnejšie teda nie viac ako $1/k^2$ hodnôt distibúcie môže byť vzdialených k alebo viac štandardných odchýlok od priemeru, alebo ekvivalentne viac ako $1-1/k^2$ hodnôt distribúcie je vzdialených menej ako k štandardných odchýlok od priemeru. Táto nerovnosť je veľmi užitočná, keďže môže byť aplikovaná pri akomkoľvek pravdepodobnostnom rozdelení ak poznáme priemer a štandardnú odchýlku. Teda podľa tejto nerovnosti sa vo vzdialenosti 5 a viac štandardných odchýlok môže nachádzať maximálne 4% hodnôt. Ak je hodnota najdrahšieho mesiaca v pozorovacom období v takejto vzdialenosti od priemeru tak bude pre nás dostatočne významná.
- highest_cost – hodnota nákladov za najdrahší mesiac
- num_above_average – počet mesiacov v pozorovacom období s hodnotou nákladov nad priemerom mesiacov v pozorovacom období. Ak sú náklady relatívne konštantné v pozorovacom období tak hodnota je okolo polovice počtu mesiacov v pozorovacom období, čo indikuje, že pacient je postihnutý chronickou chorobou.



Obr. 6 Znázornenie Čebyševovej nerovnosti

4.2. Rozdelenie do skupín

Pre účely redukcie šumu v dátach a zároveň redukcie efektu extrémne nákladných pacientov sme podľa článku [7] vytvorili rozdelenie pacientov do piatich skupín podľa výšky nákladov v poslednom roku pozorovacieho obdobia. Rozdeľovali sme takým spôsobom, aby suma v každej skupine bola približne rovnaká. Celková suma nákladov všetkých pacientov v poslednom roku pozorovacieho obdobia je 4 390 212 € a rozsahy súm jednotlivých pacientov sa pohybovali od 0 po 13 029 €. Takže suma nákladov, ktorú chceme aby mala každá skupina je okolo 878 042,4 €. Teda zoradili sme si pacientov podľa sumy ich nákladov za posledný rok pozorovacieho obdobia od najlacnejšieho po najdrahšieho a zaradom sme pacientov pridávali do skupiny pokým suma danej skupiny nepresahovala hodnotu, ktorú sme si určili, inak bola daná skupina kompletná a tak sme vytvorili ďalšiu skupinu, kde sme pridávali ďalších pacientov, kým sme znova neprekročili stanovenú hodnotu alebo sme neprešli všetkých pacientov.

Skupiny 1 až 5 môžu byť interpretované ako reprezentácia nízkeho až veľmi vysokého rizika zdravotných komplikácií. Pomocou takéhoto rozdelenia dokážeme použiť dva klasifikačné ukazatele a to Hit Ratio a Penalty Error, a taktiež dokážeme všetky ukazatele vyhodnocovať aj na jednotlivých skupinách pacientov samostatne a tak sledovať úspešnosť rôznych metód na rôznych skupinách pacientov.

Tabuľka Tab. 1 znázorňuje rozsahy každej skupiny, percentuálnu časť a počet pacientov, ktorí sú v skupine pre náš dataset. Pre porovnanie tabuľka Tab. 2 rovnako znázorňuje skupiny, ktoré boli vytvorené v článku [7] nad ich dátovou sadou z USA. Pri našej dátovej sade nemáme až také veľké ceny a rozsahy a ani percentuálne rozdelenie pacientov. Pri tabuľke Tab. 2 si môžeme všimnúť, že 80% všetkých nákladov pochádza od menej ako 20% pacientov, no pri našej tabuľke Tab. 1 je to menej ako 50%.

Skupina	Rozsah (€)	Percentuálna časť	Počet pacientov
1	< 216.30	54.27%	9 374
2	216.30 – 331.75	18.97%	3 278
3	331.75 – 471.71	12.85%	2 221
4	471.71 – 722.75	8.91%	1 540
5	> 722.75	4.97%	859

Tab. 1 Popis nášho rozdelenia do skupín

Skupina	Rozsah ($)	Percentuálna časť	Počet pacientov
1	< 3 200	83.9%	204 420
2	3 200 – 8 000	9.7%	23 606
3	8 000 – 18 000	4.2%	10 261
4	18 000 – 50 000	1.7%	4 179
5	> 50 000	0.5%	1 175

Tab. 2 Popis rozdelenia do skupín z článku [7]

5. Hodnotenie úspešnosti metód umelej inteligencie pri predikcii nákladov zdravotnej starostlivosti

Základnou myšlienkou predikcie nákladov zdravotnej starostlivosti je použitie nákladov poistenca počas prechádzajúceho obdobia, napríklad v posledných troch rokoch, na predikciu nákladov zdravotnej starostlivosti v cieľovom období, napríklad v nasledujúcom kalendárnom roku.

Na vyhodnotenie úspešnosti metód môžeme použiť ukazovatele, medzi ktoré patrí napríklad priemerná absolútna chyba, priemerná absolútna percentuálna chyba, priemerná kvadratická chyba alebo R- kvadrát [13]. Priemerná absolútna chyba vyjadruje priemernú chybu modelu medzi predpovedanou hodnotou nákladov p_i a skutočnou hodnotou nákladov a_i vypočítanú vzhľadom na všetkých pacientov v údajovej sade.

$$MAE = \frac{\sum_i |a_i + p_i|}{n}$$

Rovnica 1 Priemerná absolútna chyba (Mean absolute error)

Táto modifikovaná verzia priemernej absolútnej percentuálnej chyby vyjadruje podiel medzi priemernou absolútnou chybou a priemernou hodnotou nákladov všetkých poistencov. V pôvodnej verzii by sa totiž delilo skutočnou hodnotou, čo pri našom type dát prinášalo viacero delení nulou. Narozdiel od MAE, ktorá je závislá od konkrétnych dát, MAPE je vhodnejšia pre porovnávanie úspešnosti rôznych modelov z rôznych štúdií.

$$MAPE = \frac{\frac{\sum_i |a_i + p_i|}{n}}{\bar{a}}$$

Rovnica 2 Priemerná absolútna percentuálna chyba (Mean absolute percentage error)

Priemerná kvadratická chyba je založená na výpočte druhej mocniny medzi predpovedanou hodnotou nákladov a skutočnou hodnotou nákladov pre všetkých pacientov v údajovej sade. Často sa využíva aj odvodený ukazateľ, ktorý je druhou odmocninou z priemernej kvadratickej chyby.

$$MSE = \frac{\sum_i (a_i + p_i)^2}{n}$$

Rovnica 3 Priemerná kvadratická chyba (Mean squared error)

Ukazovateľ R-kvadrátu slúži na vyjadrenie korelácie medzi aktuálnymi a predpovedanými hodnotami nákladov, pričom v prípade ideálneho modelu umelej inteligencie je jeho hodnota rovná 1.

$$R^2 = 1 - \frac{\sum_i (a_i + p_i)^2}{\sum_i (a_i + \bar{a})^2}$$

Rovnica 4 R - kvadrát (R^2)

Ak hodnotíme priemernú absolútnu chybu, priemernú absolútnu percentuálnu chybu alebo priemernú kvadratickú chybu, v ideálnom prípade je ich hodnota rovná 0. Táto situácia zodpovedá modelu, ktorého predpovedané hodnoty nákladov sú identické ako skutočné hodnoty nákladov.

Naša úloha dá riešiť aj pomocou klasifikácie, ak ľudí rozdelíme do skupín podľa výšky ich minulých nákladov a následne sa budeme snažiť predpovedať či pacient zostane vo svojej skupine alebo preskočí do niektorej skupiny s vyššími či nižšími nákladmi. V takomto prípade je možné použiť aj ukazatele používané pri klasifikácii. Napríklad „pomer zásahov“ alebo Hit Ratio, čo je vlastne podiel počtu pacientov správne zaradených do skupiny a počtu všetkých pacientov. Ďalším vhodným ukazateľom je Penalty error [6], ktorý trestá podcenenie vyšších nákladov viac ako precenenie nižších nákladov. Teda ak má pacient patriť do skupiny s najnižšími nákladmi a model ho zaradí do skupiny s najvyššími nákladmi tak penalizácia je polovičná oproti penalizácii ak nastane situácia, že pacient má patriť do skupiny s najvyššími nákladmi a je zaradený do skupiny s najnižšími nákladmi.

	Skutočná skupina				
Predikovaná skupina	0	2	4	6	8
	1	0	2	4	6
	2	1	0	2	4
	3	2	1	0	2
	4	3	2	1	0

Tab. 3 Hodnoty penalizácie na základe skutočnej a predikovanej skupiny pri rozdelení do piatich skupín

Porovnaním výkonnosti rôznych modelov umelej inteligencie vzhľadom na rôzne ukazovatele úspešnosti a rôzne skupiny obyvateľov môžeme získať informáciu o metódach, ktoré sú pri riešení našej úlohy najúspešnejšie. Výsledky rôznych štúdií ukazujú, že pri predpovední hodnôt nákladov zdravotnej starostlivosti v nasledujúcom období preukazujú najvyššiu úspešnosť metódy umelej inteligencie založené na kontrolovanom učení, konkrétne náhodné regresné lesy a umelé neurónové siete, ktoré sme si stručne predstavili v predchádzajúcej kapitole. Umelé neurónové siete sú mimoriadne úspešné pri pacientoch s vyššími nákladmi. O žiadnej metóde však nemôžeme tvrdiť, že je univerzálna. Vo viacerých prípadoch sa v praxi ako najúčinnejšie javí použitie kombinácie viacerých metód.

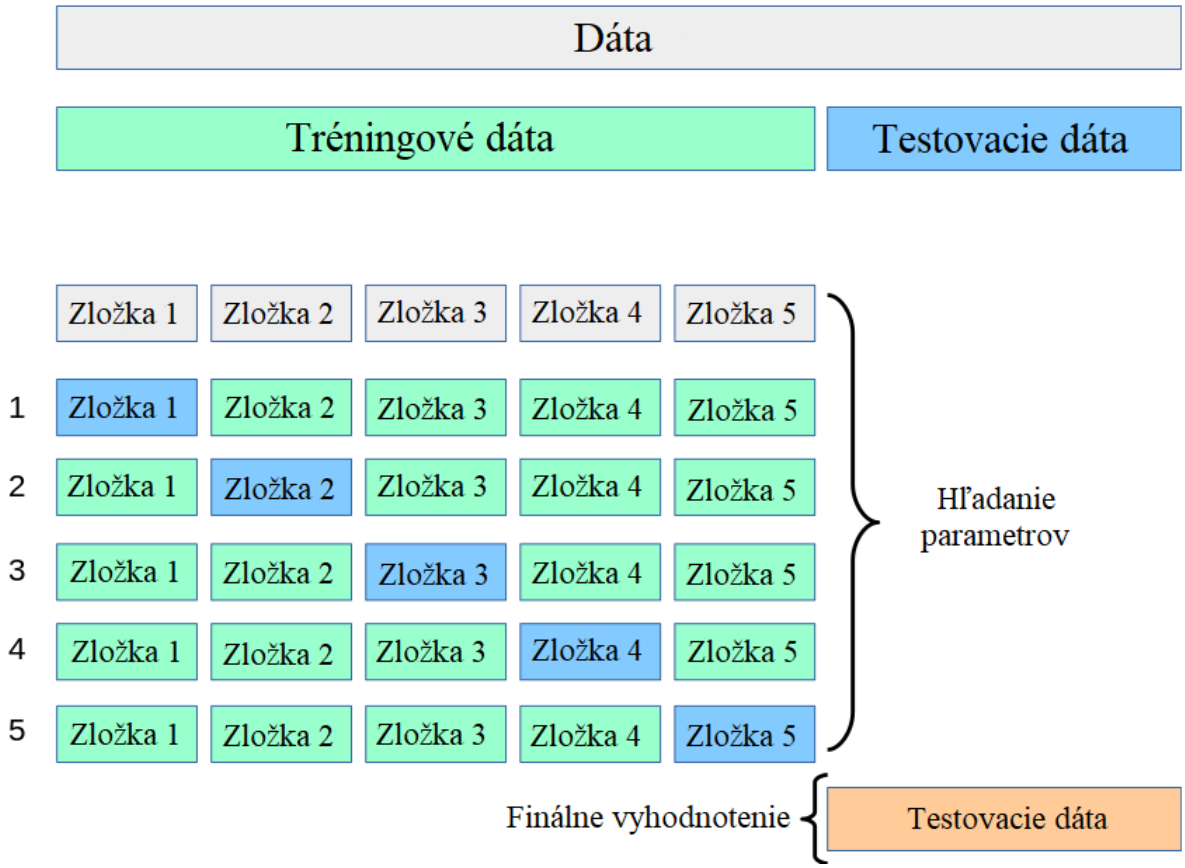
5.1. K-zložková krížová validácia

Krížová validácia je populárna štatistická metóda používaná na odhad presnosti vytvorených modelov. Pri k-zložkovej krížovej validácii sa počiatočné dáta náhodne rozdelia do k vzájomne disjunktných podmnožín alebo zložiek D_1, D_2, ... , D_k, pričom všetky sú približne rovnako veľké. Trénovanie a testovanie je vykonávané K-krát. V iterácii i je zložka D_i rezervovaná ako testovacia množina a zostávajúce zložky sú spolu použité na trénovanie modelu. každá zložka je teda presne (k-1)-krát použitá pri trénovaní a jedenkrát na testovanie. Pri klasifikácii je odhadom presnosti celkový počet správne klasifikovaných príkladov zo všetkých k iterácií delený celkovým počtom kladov v počiatočných dátach.

Leave-one-out je špeciálny prípad k-zložkovej krížovej validácie, kde k je zvolené ako počet príkladov v počiatočných dátach. Teda v jednej iterácii je vynechaný iba jeden príklad pre testovanie.

Pri stratifikovanej krížovej validácii sú jednotlivé zložky vytvorené tak, aby pomer príkladov z rôznych tried bol približne rovnaký, ako v počiatočných dátach.

Vo všeobecnosti je 10-zložková krížová validácia odporúčaná pre odhad presnosti, aj keď výpočtový výkon dovoľuje použitie viacerých zložiek, kvôli relatívne nízkemu bias a variancii.



Obr. 7 5-zložková krížová validácia

6. Trénovanie metód

Trénovanie, ale aj analýza dátovej sady prebiehala na platforme Colaboratory (skrátene Colab) od spoločnosti Google. Colab umožňuje písať a spúšťať kód v jazyku Python cez prehliadač a je špeciálne prispôsobený pre strojové učenie, dátovú analýzu a vzdelávanie. Je to vlastne Jupyter notebook služba, ktorá nevyžaduje žiadne úvodné nastavovania a poskytuje voľný prístup k výpočtovým zdrojom vrátane grafických kariet.

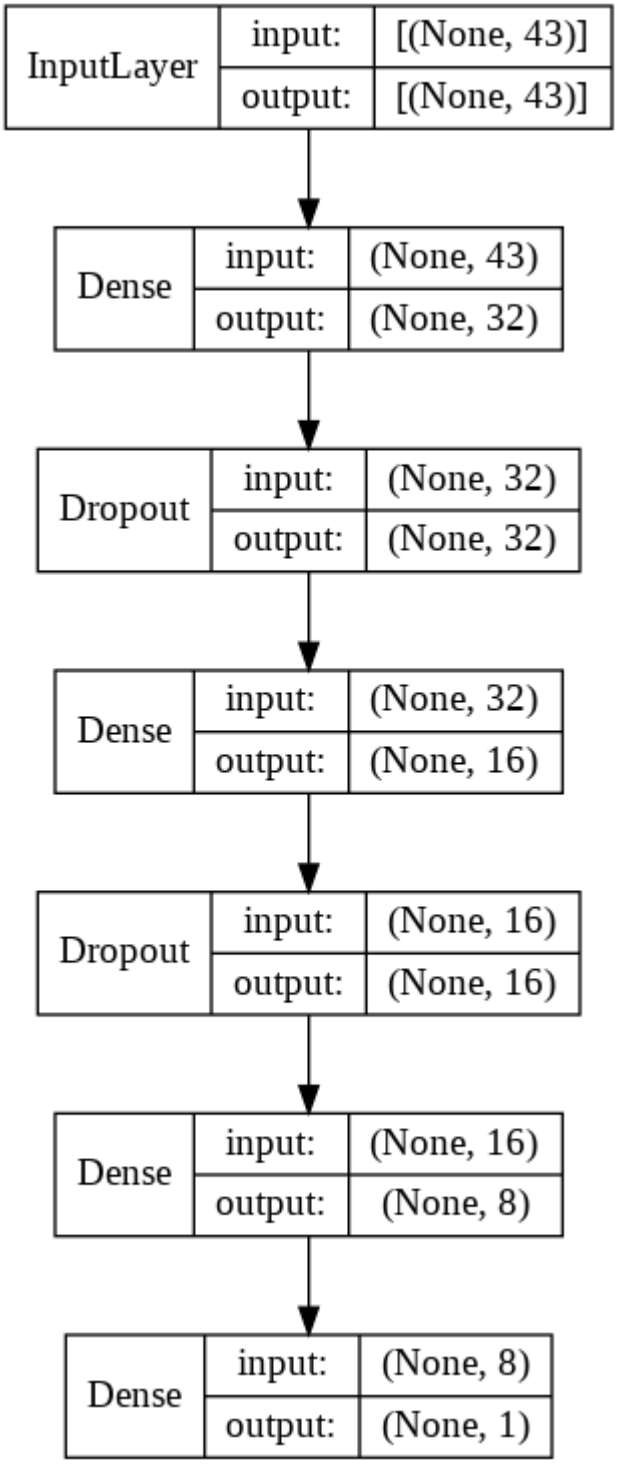
6.1. Dopredná neurónová sieť

Pre prácu s doprednou neurónovou sieťou sme využívali Keras, ktorý je výkonnou a jednoduchou open source knižnicou pre Python. Obaľuje totiž výpočtovo efektívne knižnice Theano a Tensorflow, a tak dovoľuje definovať a trénovať neurónové siete pomocou pár riadkov kódu.

Najprv je potrebné zadefinovať architektúru neurónovej siete. Na to sme použili sekvenčný model, pomocou ktorého sme pridali vrstvy aké sme potrebovali. Dopredná neurónová sieť sa zvyčajne skladá z dvoch typov vrstiev:

- Dense – husto prepojená vrstva, nastavuje sa pri nej počet neurónov vo vrstve a aktivačná funkcia, ktorá sa aplikuje na jej výstup
- Dropout – regularizačná vrstva, ktorá zabraňuje preučeniu siete tým, že náhodne nahradzuje vstupné hodnoty hodnotou 0 a tým ako keby zabúdala časť z toho čo sa už naučila, nastavuje sa pomer, koľko hodnôt má nahradiť

V našej sieti majú všetky Dense vrstvy aktuvačnú funkciu ReLU, okrem poslednej, ktorá je výstupná a má aktivačnú funkciu Linear. Všetky Dropout vrstvy majú pomer nastavený na 0,1. Architektúra siete je na obrázku Obr. 8.



Obr. 8 Architektúra doprednej neurónovej siete

Ďalej je potrebné tento sekvenčný model skompilovať a nastavili sme parametre:

- loss = mse – nastavenie funkcie Stredná kvadratická chyba (Mean Squared Error) ako stratovú funkciu
- optimizer = Adam – nastavenie algoritmu Adam ako optimalizátor, ktorý sa snaží efektívne znižovať hodnotu stratovej funkcie
- learning_rate = 0,01 – pomer použitý pri trénovaní, aby sa predišlo preučeniu

Pri trénovaní sme použili tieto nastavenia:

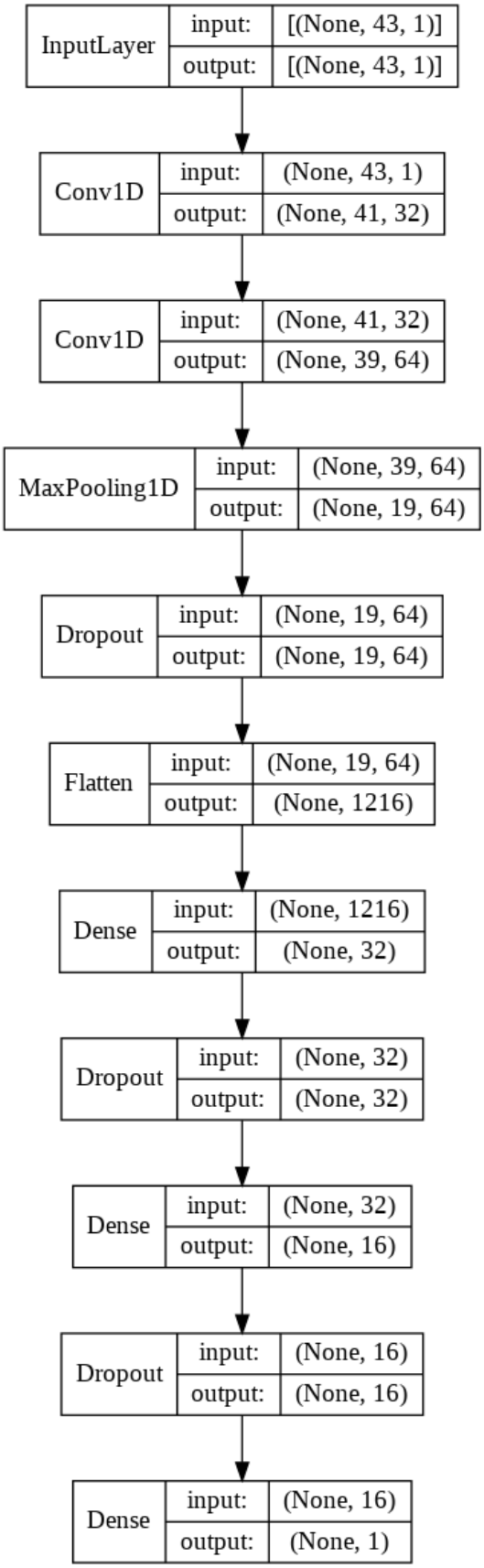
- epoch = 100 – epocha je jeden cyklus prezentovania všetkých príkladov sieti
- batch_size = 32 – počet príkladov prezentovaných sieti v epoche bez toho aby sa aktualizovali váhy

6.2. Konvolučná neurónová sieť

Pre prácu s konvoluč neurónovou sieťou sme používali taktiež knižnicu keras. Najprv je potrebné vytvoriť sekvenčný model pomocou rôznych typov vrstiev ako sú:

- Conv1D – jednorozmerná konvolučná vrstva, nastavuje sa počet neurónov a veľkosť konvolučného filtra
- MaxPool1D – vrstva, ktorá zmenšuje dimenziu vstupu tým, že vyberie len najväčšiu hodnotu z okna veľkosti akú nastavíme
- Dropout - regularizačná vrstva, ktorá zabraňuje preučeniu siete tým, že náhodne nahradzuje vstupné hodnoty hodnotou 0 a tým ako keby zabúdala časť z toho čo sa už naučila, nastavuje sa pomer, koľko hodnôt má nahradiť
- Flatten – vrstva, ktorá transformuje vstup z matice na vektor, aby bolo možné použiť Dense vrstvu
- Dense - husto prepojená vrstva, nastavuje sa pri nej počet neurónov vo vrstve a aktivačná funkcia, ktorá sa aplikuje na jej výstup

V tejto sieti majú všetky Conv1D aj Dense vrstvy aktivačnú funkciu ReLU, okrem poslednej, ktorá je výstupná a ma aktivačnú funkciu Linear a všetky Dropout vrstvy majú pomer 0,1. Znázornenie architektúry siete je na obrázku Obr. 9.



Obr. 9 Architektúra konvolučnej neurónovej siete

Parametre pri kompilácii modelu:

- loss = mse - nastavenie funkcie Stredná kvadratická chyba (Mean Squared Error) ako stratovú funkciu
- optimizer = Adam – nastavenie algoritmu Adam ako optimalizátor, ktorý sa snaží efektívne znižovať hodnotu stratovej funkcie
- learning_rate = 0,0005 – pomer použitý pri trénovaní, aby sa predišlo preučeniu

Nastavenia pri trénovaní:

- epoch = 100 – epocha je jeden cyklus prezentovania všetkých príkladov sieti
- batch_size = 32 – počet príkladov prezentovaných sieti v epoche bez toho aby sa aktualizovali váhy

6.3. Extreme Gradient Boosting

XGBoost je optimalizovaná distribuovaná knižnica pre gradient boosting dizajnovaná na to aby bola efektívna a flexibilná. Pri trénovaní sme použili XGBRegressor, kde je možné nastaviť množstvo parametrov a my sme nastavili tieto:

- objective = reg:squarederror – špecifikuje úlohu, a teda regresnú úlohu s použitím kvadratickej chyby ako stratovej funkcie
- learning_rate = 0,09 – pomer použitý pri trénovaní, aby sa predišlo preučeniu
- n_estimators = 250 – maximálny počet stromov, ktorý sa môže počas učenia vytvoriť
- reg_alpha = 0,4 – L1 regularizácia, aby sa zabránilo preučeniu
- subsample = 0,7 – pomer, akú veľkú časť príkladov náhodne vyberie a použije pri učení, zabráňuje preučeniu
- colsample_bytree = 0,7 – pomer, akú veľkú časť atribútov náhodne vyberie pre vytváranie stromu

6.4. Light Gradient Boosting Machine

LGBM je framework pre gradient boosting a je dizajnovaný, aby bol distribuovaný a efektívny s danými výhodami oproti ostatným:

- rýchlejšie trénovanie a výššia efektivita
- nižšie použitie pamäte
- lepšia presnosť
- podpora paralelného, distribuovaného a GPU učenia
- schopný zvládnuť obrovské dáta

Pre učenie sme použili LGBMRegessor s týmito parametrami:

- learning_rate = 0,05 – pomer použitý pri trénovaní, aby sa predišlo preučeniu
- n_estimators = 90 – maximálny počet stromov, ktorý sa môže počas učenia vytvoriť
- num_leaves = 40 – maximálny počet listov v strome
- reg_lambda = 0,5 – L2 regularizácia, aby sa zabránilo preučeniu
- subsample = 0,8 – pomer, akú veľkú časť príkladov náhodne vyberie a použije pri učení, zabráňuje preučeniu
- subsample_freq = 20 – ako často sa má použiť subsampling, daná hodnota *k* znamená, že na každú *k-tu* iteráciu sa použije subsampling
- min_split_gain = 0.2 – minimálny zisk potrebný, aby sa uzol v strome mohol ďalej deliť
- colsample_bytree = 0.7 – pomer, akú veľkú časť atribútov náhodne vyberie pre vytváranie stromu

7. Výsledky a porovnania

Všetky výsledky sú priemernou hodnotou z hodnôt získaných použitím 10-zložkovej krížovej validácie.

7.1. Výsledky s použitím základných metód

Najprv sme pomocou vybraných ukazateľov vyhodnotili naše triviálne baseline metódy, ktorých výsledky sú v tabuľke Tab. 4 a Tab. 5. Z týchto výsledkov je vidieť, že metóda Baseline1 predpovedá náklady na budúci rok ako sumu nákladov za posledný rok v pozorovacom období je úspešnejšia ako metóda Baseline2, ktorá zasa predpovedá náklady vo výške priemernej hodnoty súm posledných 3 rokov v pozorovacom období.

Baseline1 vlastne predpokladá, že sa u všetkých pacientov v nasledujúcom roku zopakujú náklady z minulého roka, a teda budú v rovnakej skupine. Takže pomocou ukazateľa Hit Ratio môžeme zistiť, koľko pacientov zostalo v rovnakej skupine aj na ďalší rok a teda ich stav sa nezlepšil ani nezhoršil. Z tabuľky TAB teda môžeme vidieť, že 60,1% pacientov zostalo vo svojej skupine, a že dve najstabilnejšie skupiny sú najlacnejšia s 70,7% a najdrahšia s 72,8%.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.173	0.373	60.1%	0.826
1	-1.217	0.575	70.7%	0.812
2	0.030	0.286	43.7%	0.924
3	0.051	0.277	41.1%	0.972
4	0.104	0.234	50.9%	0.751
5	-3.025	0.246	72.8%	0.352

Tab. 4 Naše výsledky pre metódu Baseline1

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.046	0.425	55.9%	0.998
1	-1.415	0.619	66.8%	0.801
2	-1.175	0.371	42.7%	1.212
3	-0.463	0.332	35.8%	1.422
4	-0.460	0.288	42.4%	1.282
5	-2.768	0.275	63.9%	0.721

Tab. 5 Naše výsledky pre metódu Baseline2

V článku [7] použili ako baseline metódu rovnakú ako je Baseline1 a výsledky ukazateľov okrem MAPE na ich dátovej sade sú v tabuľke . Tu si môžeme všimnúť, že až 80% pacientov zostalo v rovnakej skupine ako minulý rok. Avšak ukazateľ Hit Ratio ovplyvňuje aj veľkosť rozsahu pre dané skupiny, ktorý je pri našej dátovej sade celkom úzky, a je teda prísnejší.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	-0.001	-	80.0%	0.431
1	-0.033	-	90.1%	0.287
2	-0.056	-	52.3%	0.992
3	-0.087	-	41.7%	1.358
4	-0.057	-	30.5%	1.669
5	0.500	-	19.3%	1.825

Tab. 6 Výsledky pre baseline metódu z článku [7]

7.2 Výsledky s použitím doprednej neurónovej siete

Ďalej sme vyhodnocovali doprednú neurónovú sieť, ktorej výsledky sú v tabuľke Tab. 7. Tento model má výrazne lepšiu hodnotu ukazateľa R^2 a mierne lepšie MAPE oproti Baseline1 no horšie Hit Ratio a Penalty Error.

Na porovnanie máme aj výsledky z článku [6], ktoré sú v tabuľke Tab. 8. Ukazatele R^2 a MAPE sú v našom prípade lepšie no ukazatele Hit Ratio a Penalty Error sú miestami až výrazne horšie, čo však znova môže byť zapríčinené rôznymi veľkosťami rozsahov pre skupiny. Teda aj keď robíme menšie chyby čo sa týka rozdielu medzi reálnou a predpovedanou hodnotou, tak stále táto predpovedaná hodnota spadá do inej skupiny ako reálna podľa určených hraníc.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.451	0.328	52.4%	0.999
1	-0.095	0.461	62.3%	0.902
2	-0.170	0.384	69.8%	0.621
3	-0.183	0.377	64.9%	0.715
4	-0.075	0.344	62.7%	0.719
5	0.404	0.304	44.5%	1.158

Tab. 7 Naše výsledky pre metódu Dopredej neurónovej siete

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.440	0.400	89.2%	0.220
1	0.020	0.840	94.0%	0.140
2	0.070	0.690	69.5%	0.700
3	0.110	0.660	53.9%	0.970
4	0.250	0.520	54.9%	0.980
5	0.440	0.450	49.6%	0.970

Tab. 8 Výsledky pre ANN metódu z článku [6]

7.3 Výsledky s použitím konvolučnej neurónovej siete

Jednorozmerná konvolučná neurónová sieť mala výrazne lepšie výsledky v rámci všetkých ukazateľov oproti doprednej neurónovej sieti. Výsledky sú v tabuľke Tab. 9. Čo sa týka porovnania s neurónovou sieťou z tabuľky Tab. 8, tak na niektorých skupinách má aj lepšie Hit Ratio alebo Penalty Error, ale celkovo stále horšie.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.573	0.280	58.7%	0.846
1	0.009	0.415	69.0%	0.872
2	0.032	0.330	76.7%	0.574
3	0.043	0.327	71.8%	0.660
4	0.109	0.299	68.9%	0.675
5	0.550	0.255	50.6%	0.927

Tab. 9 Naše výsledky pre metódu Konvolučnej neurónovej siete

7.4 Výsledky s použitím skupinového modelu stromov (Extreme Gradient Boosting)

Metóda gradient boostingu XGBoost si poradila najlepšie a jej výsledky sú v tabuľke Tab. 10. Výrazne lepšie R^2 a mierne lepšie MAPE oproti konvolučnej neurónovej sieti.

Oproti obyčajnému gradient boostingu použitého v článku [6], ktorého výsledky sú v tabuľke Tab. 11 bol výrazne lepší v R^2 a MAPE a znova horší v Hit Ratio a Penalty Error, kvôli rôznym veľkostiam rozsahov skupín.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.677	0.277	57.6%	0.784
1	0.016	0.429	60.0%	0.901
2	0.048	0.346	76.4%	0.571
3	0.066	0.338	70.8%	0.671
4	0.146	0.308	66.9%	0.662
5	0.694	0.247	51.5%	0.817

Tab. 10 Naše výsledky pre metódu XGBoost

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.460	0.650	92.9%	0.200
1	0.040	0.760	96.4%	0.120
2	0.110	0.630	72.3%	0.670
3	0.150	0.600	61.2%	0.920
4	0.130	0.590	50.8%	1.080
5	0.320	0.540	35.2%	1.200

Tab. 11 Výsledky pre Gradient boosting metódu z článku [6]

7.5 Výsledky s použitím skupinového modelu stromov (Light Gradient Boosting Machine)

LGBM má veľmi podobné výsledky v rámci ukazateľov ako XGBoost, no zato proces učenia tohto modelu je výrazne rýchlejší, aj 2 a viac násobne. Naviac je vhodnejší pre prácu s obrovskými datasetmi. Výsledky modelu LGBM sú v tabuľke Tab. 12.

Skupina	R^2	MAPE	Hit Ratio	Penalty Error
Všetci	0.673	0.279	58.7%	0.784
1	0.013	0.430	69.3%	0.873
2	0.036	0.350	43.7%	0.571
3	0.058	0.343	71.1%	0.675
4	0.136	0.312	66.9%	0.659
5	0.689	0.248	50.9%	0.824

Tab. 12 Naše výsledky pre metódu LGBM

Záver

V našej práci sme predstavili možné použitie umelej inteligencie v oblasti predikcie nákladov zdravotnej starostlivosti. Porovnaním výkonnosti rôznych modelov umelej inteligencie vzhľadom na rôzne ukazovatele úspešnosti a rôzne skupiny obyvateľov poukazujú rôzne štúdie na možnosť použitia regresných stromov, náhodných regresných lesov, umelých neurónových sietí, konvolučných neurónových sietí a metódy boostingu.

Popísali sme našu dátovú sadu a aj atribúty, ktoré sme k nej pridali. Ďalej sme rozdeľovali pacientov do skupín podľa výšky ich nákladov, aby sme mohli použiť aj klasifikačné ukazatele, ale aj aby sme mohli všetky ukazatele vyhodnotiť aj na jednotlivých skupinách samostatne. Následne sme popísali naše modely s ich parametrami, ktoré sme použili pri trénovaní a nakoniec sme všetky modely vyhodnotili pomocou ukazateľov a porovnali s výsledkami z článkov. Zistili sme, že najlepšie výsledky dosahovali metódy gradient boostingu XGBoost a LGBM. Pričom obe mali porovnateľné výsledky, ale LGBM je rýchlejší pri fáze trénovania.

Pridanie nenákladových medicínskych alebo demografických atribútov do údajovej sady a využitie ich prediktívnej schopnosti môže byť ďalším smerom budúceho výskumu v tejto oblasti.

Zoznam použitej literatúry

1. ANTONI, Ľ., PARALIČ, J., BUTKA, P. et al. *Dátová veda a jej aplikácie*, 1. vydanie. Košice, Univerzita Pavla Jozefa Šafárika v Košiciach, 2020. ISBN 978-80-8152-917-7.
2. ANTONI, Ľ., GALČÍK, F., GUNIŠ, J. et al. Case Studies in Data Science and Internet of Things. In: *17th IEEE International conference on emerging elearning technologies and applications: Information and communication technologies in learning*, Denver: Institute of Electrical and Electronics Engineers, 2019. pp. 23-28.
3. ALPAYDIN, E. *Introduction to Machine Learning*, 3rd edition. Boston, MIT Press, 2014. ISBN: 978-0262-02818-9.
4. DHAR, V.: Data science and prediction. In *Communications of the ACM*, Vol. 56, no. 12, 2013, pp. 64-73.
5. NEIL, C.O., SCHUTT, R. *Doing Data Science: Straight Talk from the frontline*, 1st edition. O'Reilly Media, 2014. ISBN 978-1449358655.
6. MORID, M. A., KAWAMOTO, K., AULT, T. et al. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. In *American Medical Informatics Association Annual Symposium Proceedings*, Vol. 2017, pp. 1312-1321.
7. BERTSIMAS, D., BJARNADÓTTIR, M. V., KANE, M. A. et al. Algorithmic Prediction of Health-Care Costs. In *Operation Research*, Vol. 56, no. 6, 2008, pp. 1382-1392.
8. GUO, X., GANDY, W., COBERLEY, C. et al. Predicting health care cost transitions using a multidimensional adaptive prediction process. In *Population health management*, Vol. 18, no. 4, 2015, pp. 290-299.
9. MIHAYLOVA, B., BRIGGS, A., O'HAGAN, A. et al. Review of statistical methods for analysing healthcare resources and costs. In *Health economics*, Vol. 20, no. 8, 2011, pp. 897-916.
10. BISHOP, C. M. Pattern Recognition and Machine Learning, 1st edition. New York, Springer, 2007. ISBN 978-0387-31073-2.
11. HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques*, 3rd edition. Waltham, Elsevier, Morgan Kaufmann Publishers, 2012. ISBN 978-0-12-381479-1.
12. KE, G. et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, pp.3146–3154..
13. DUNCAN, I., LOGINOV, M., LUDKOVSKI, M. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. In *North American Actuarial Journal*, Vol. 20, no. 1, 2016, pp. 65-87.