Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/14BE12Stanic

AUTOMATIZACIJA PROCESA ANALIZE PODATAKA I OBUČAVANJA MODELA ZA PREDIKCIJU STOPE SAMOUBISTAVA

AUTOMATING THE DATA ANALYSIS AND MODEL TRAINING PROCESS FOR SUICIDE RATE PREDICTION

Aleksandar Stanić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U okviru ovog rada opisani su softverski moduli kreirani TDD procesom razvoja, poštujući SOLID principe. Oni su namenjeni automatizaciji određenih koraka ka kreiranju kvalitetnog prediktivnog modela kako bi se rešio problem predikcije stope samoubistava po državama.*

Ključne reči: *TDD, SOLID principi, mašinsko učenje, prediktivni model,*

Abstract – *This paper describes software modules created following the TDD implementation process, respecting the SOLID principles. These modules are intended to automate certain steps towards creating a machine learning model that predicts suicide rates by countries.*

Keywords: *TDD, SOLID principles, machine learning, predictive model*

1. UVOD

Velika većina problema iz oblasti nauke sa podacima koji su vezani za kreiranje prediktivnih modela, rešava se na skoro približan način. On započinje samim prikupljanjem skupova podataka, njihovom obradom i analizom, pa zatim treniranjem prediktivnih modela na osnovu predhodno preprocesiranih podataka. Isti slučaj je bio i u radu [1] gde su Boris Bibić, Marko Katić i autor teksta rešavali problem predikcije stope samoubistava na državnom nivou. Koraci koji su tamo bili primenjivani predstavljaće vodilje ka kreaciji softverskih modula i samog problemsko-specifičnog softverskog rešenja, a sam problem predikcije stope samoubistava bice iskorišćen kao test slučaj kako bi se ispitali rezultati softverskog rešenja.

Tema ovog rada jeste kreiranje prediktivnog modela na osnovu predhodno analiziranih i procesovanih obeležja, odnosno kreiranje softverskih modula uz pomoć koji se dolazi do odgovarajućeg prediktivnog modela.

Kreiranje softverskog rešenja se vrši preko Test Driven Development [2] pristupa razvoja softvera poštujući tzv. SOLID principe [3]. Sa ovim pristupom razvoja aplikacije, njena arhitektura postaje razumnija, lakša za modifikaciju i lakša za proširivanje.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr. prof.

Uz pomoć spomenutih softverskih modula, biće implementirano specifično softversko rešenje koje koristi spomenute module i automatizuje i generalizuje proces analize i predikcije problema sličnih problemu vezanog za predikciju stope samoubistava (gde su kao obavezne kolone svakog skupa podataka kolone koje reprezentuju godinu i državu). A samim drugačijim korišćenjem tih modula i dodavanjem novih implementacija apstrakcija unutar njih, moguće je napraviti novo softversko rešenje u svrhe rešavanja neke druge grupe problema. Ovakvo softversko rešenje će kao rezultat korisniku olakšati proces rešavanja problema i dati mu informacije potrebne za kreiranje najboljeg mogućeg modela ili pak dobru smernicu ukoliko je potrebno to rešiti ručno.

Neobrađeni skupovi podataka prikupljeni u radu [1] iskorišćeni su i u ovom radu. Takvi skupovi podataka su dodatno analizirani i obrađeni pred njihovo spajanje u zajednički skup podataka. Analizu predstavlja ispitivanje kvaliteta samih podataka skupa i proveru pripadnosti kolona koje predstavljaju godinu i državu koje su bile potrebne za mogućnost spajanja skupova podataka. Obradu podataka predstavlja isecanje skupa u izabranom opsegu godina, dopunu nedostajućih redova skupa, kako bi svi skupovi imali jednaki broj redova za svoje parove godina-država, kao i dopunu nedostajućih vrednosti sa jednom od metoda koji će kasnije biti opisani.

Spojeni skup podataka se dodatno procesira kako bi mogao biti korišćen od strane prediktivnih algoritama. Nakon čega se nad njim primenjuju algoritmi za redukciju dimenzionalnosti kao što su Low Variance Filter [4], Random Forest [5], Principal Component Analysis [6] i Factor Analysis [7].

Sa tako redukovanim skupovima podataka, kao i sa originalom trenirani su prediktivni regresioni algoritmi: Decision Tree [8], Random Forest, Multiple Linear Regression [9], Partial Least Squares Regression [10], Elastic Net Regression [11] i XGB Regression [12]. Tačnost modela ispitana je sa metrikom pronalaska koeficijenta determinacije [13] i srednjim kvadrantom greške. Rezultati o najznačajnijim varijablama modela su ispitani preko Feature Importance informacije Random Forest i XGB regresionih algoritama.

2. METODOLOGIJA

U ovom poglavlju biće prikazani koraci koji za rezultat imaju kreaciju prediktivnog modela. Takodje, uz svaki korak biće ukratko opisan i softverski modul implementiran za rešavanje njegovog problema.

2.1. Unos podataka u aplikaciju

Zbog raznolikosti struktura skupova podataka sa kojima se radilo, nemoguće je napraviti softversko rešenje koje će pokriti sve slučajeve nepravilnosti u strukturi skupa podataka, ili jednostavno problema nekonzistetnosti između više skupova podataka. Stoga, odlučeno je da će prvi modul čija je odgovornost unos skupova podataka imati preduslov sa kojim se mogu navoditi koje su obavezne kolone koji svaki skup podataka mora sadržati. Dok, sve probleme kao što su na primer semantička tačnost podataka unutar skupa ili način na koji su predstavljene nedostajuće vrednosti biti dužnost korisnika da validira pred unos skupa podataka u softver. Za obavezne kolone su izabrane kolone koje predstavljaju godinu i ISO3 kod države.

Prikaz trenutno unetih skupova podataka unutar softverskog rešenja je zadatak posebnog softverskog modula. Ono za ulaz dobija učitane podatke sa svojim nazivima, a kao izlad daje izgenerisanu reportažu korisniku u određenom formatu, u ovom slučaju je izabrano kao tekstualni prikaz unutar terminala.

2.2. Prikaz statistike skupova podataka

Još jedan deo koji se ne može automatizovati jeste proces eksplorativne analize. Ali, ono što je moguće uraditi jeste generisati određenu statistiku skupova podataka korisniku kako bi mu se olakšao sam taj proces analize. To predstavlja ulogu narednog softverskog modula.

Rezultat njegovog izvršavanja predstavlja Excel datoteku (takođe i ovde može biti drugi format prikaza rezultata) sa podacima za svaki učitani skup podataka.

Od informacija korisnik dobija na uvid nazive kolona skupa podataka, opseg godina obuhvaćenih podacima, broj država obuhvaćenih podacima, srednju vrednost za kolonu godina, njenu standardnu devijaciju i varijansu, ukupan broj podataka za svaku neobaveznu kolonu (konkretno kolone koje nisu kolona godina i kolona ISO3 kod države), kao i broj nedostajućih podataka i sam procenat nedostajućih podataka u odnosu na ukupan broj podataka tih kolona. Naravno, svaka od ovih statistika predstavlja implementaciju određene apstrakcije unutar spomenutog modula, što znači moguće je u budućnosti dodati implementacije novih statistika. Važno je napomenuti da se preko softverskog rešenja koji koristi ovaj modul definisalo za koju kolonu se na primer generisala statistika opsega podataka (kolona godina u ovom specifičnom slučaju).

2.3. Brisanje skupova podataka / brisanje njihovih kolona po određenom kriterijumu

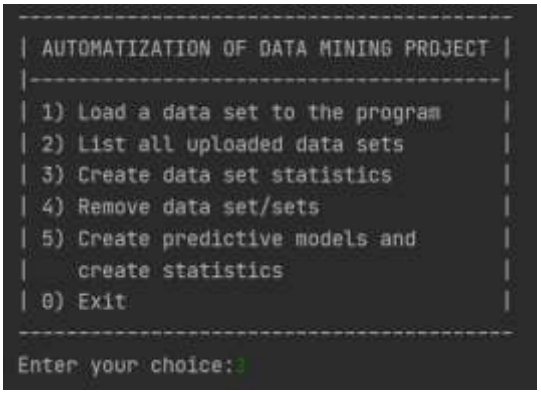
Kako postoji mogućnost unosa skupova podataka u aplikaciju, potrebno je korisniku omogućiti i brisanje istih iz aplikacije. Taj zadatak kao i zadatak filtera kolona svih skupova podataka po određenom kriterijumu su dodeljeni novom softverskom modulu.

Drugi zadatak daje mogućnost korisniku da obriše kolone skupova podataka ukoliko nisu zadovoljile jedan od kriterijuma: ukoliko je broj unikatnih vrednosti kolone manji od određenog broja ili ukoliko je procenat nedostajućih podataka u koloni veći od neke granične vrednosti. Jedini specijalan slučaj predstavlja situaciju kada je obrisana kolona jedna od obaveznih kolona ili jedina od neobaveznih kolona jednog skupa podataka, onda se ceo skup podataka briše iz aplikacije.

Izbor kolona koji ulaze u proveru korisnik odabira tokom rada na aplikaciji.

2.4. Obrada skupova podataka kao priprema za proces spajanja u zajednicki skup

Ulaskom u petu stavku menija softverskog rešenja (slika 1) korisnik započinje proces kreiranja prediktivnog modela. Prvi korak ovog procesa predstavlja obradu skupova podataka, kako bi oni međusobno bili konzistetni sa aspekta pokrivenosti jednakog broja država i opsega godina. Ovo je neophodno iz razloga što će se u budućem koraku svi skupovi podataka morati svesti na jedan skup (spojeni skup), a u ovom specifičnom slučaju to će se vršiti na osnovu kolona: godina i ISO3 kod države.



Slika 1. *Prikaz početnog menija specifičnog softverskog rešenja*

Stoga, prvi korak u tom procesu jeste izbor najoptimalnijeg opsega godina koji će se dalje uzeti u obzir. Opseg godina korisnik sam zaključuje i to na osnovu statistike kreirane u trećoj stavci menija (opisano u 2.2), ali isto tako i statistike o kojoj će sada biti reči. Nakon što je korisnik uneo opseg godina, započinje proces obrade skupova podataka. Prvo se svi skupovi podataka odsecaju tako da svi sadrže redove sa zajedničkim skupom ISO3 koda države, pa se zatim odsecaju redovi sa godinama koje ne spadaju u traženi opseg. Ovaj posao je odrađen od strane softverskog modula implementiranog samo u te svrhe. Njegov rezultat daje skupove podataka sa određenim brojem nedostajućih redova za par ISO3 kod države-godina.

Kako bi svi skupovi podataka imali isti broj redova u sebi (kako bi obuhvatali sve parove ISO3 kod države-godina), izgenerisani su redovi za svaki skup podataka ponaosob. Takav red u sebi sadrži traženi par obaveznih kolona i prazne vrednosti za ostale kolone. Za ovo je takođe napravljen poseban softverski modul.

Uz pomoć novonastalih skupova podataka i modula za generisanje statistika, nova statistika biva generisana. Sa tom statistikom korisnik sistema imaće još bolji uvid koliko je dobar njegov izbor. Zatim, korisniku se daje upit da li je zadovoljan nakon ponovne analize da nastavi dalje ili da se vrati na prvi korak i izabere novi opseg godina i ponovi proces kreacije skupova podataka.

2.5. Dopuna nedostajućih vrednosti obrađenim skupovima podataka

Pre procesa spajanja skupova podataka ostaje još jedan korak, ne toliko bitan za samo spajanje podataka, ali bitan za buduću kreaciju prediktivnih modela i korišćenja algoritama za redukciju dimenzionalnosti, a to je dopuna nedostajućih vrednosti sa ne Null vrednostima.

Ovaj korak započinje sa korisnikovim unosom maksimalnog procenta nedostajućih podataka unutar obrađenog skupa koje on smatra zadovoljavajućim za dalji rad. Uz pomoć modula za brisanje kolona skupa podataka, sve kolone koje sadrže veći procenat od unetog biće obrisane. Što je veći procenat, veći deo podataka mora biti dopunjen kao nedostajuća vrednost, što može ugroziti originalnost podataka.

Svaki skup podataka ima drugačiju semantiku, stoga nije moguće automatizovati proces punjenja nedostajućih vrednosti bez da se obezbedi korisniku da za svaku kolonu skupova podataka sam izabere način na koji će da joj se popune te vrednosti. Dostupni načini koje korisnik može izabrati su: punjenje ćelija sa 0 brojnom vrednosti, minimalnom brojnom vrednosti te kolone, maksimalnom brojnom vrednosti te kolone, srednjom brojnom vrednosti te kolone, tekstualnom konstantom, punjenje uz pomoc višestruke linearne regresije (gde se model trenira sa podacima koje nemaju nedostajuce vrednosti) i punjenje uz pomoc logistic regresije.

2.6. Spajanje skupova podataka

Spajanje obrađenih skupova podataka se vrši na osnovu dve kolone: kolona godina i kolona ISO3 kod države. Naravno, implementacija nam omogucava da navedemo bilo koji skup kolona po kojim se žele spojiti skupovi podataka. Zbog test problema za koji se pravi softversko rešenje su iskorišćene spomenute kolone.

2.7. Preprocesiranje spojenog skupa podataka

Prediktivni algoritmi za vrednosti svakog obeležja zahtevaju numeričku vrednost. Pošto u spojenom skupu podataka postoje obeležja sa tekstualnim vrednostima potrebno je izvršiti enkodovanje podataka. Enkodovanje je izvršeno uz pomoc One Hot enkodera.

Takođe, modul zadužen za preprocesiranje omogućava podelu skupova podataka (X-prediktorski skup, Y-skup predikcione varijable) na trening i test skupove.

2.8. Redukcija dimenzionalnosti

Zbog novonastalog preprocesiranog skupa podataka koji sadrži enkodovane tekstualne kolone sa One Hot enkoderom, rodila se potreba da se prediktivni algoritmi, pored redovnog skupa treniraju i sa skupovima podataka redukovanim od strane algoritama redukcije dimenzionalnosti.

Implementacija modula namenjenog u te svrhe je vršena na drugačiji način u odnosu na predhodno spomenute module. TDD pristup razvoja nije se mogao primeniti. Razlog tome je priroda algoritama redukcije dimenzionalnosti, odnosno njihova nepredividivost. Stoga, na osnovu iskustva prilikom implementacije predhodnih modula, kreiran je modul namenjen za redukciju dimenzionalnosti. Na osnovu ulaznog preprocesiranog skupa, softverski modul generiše novi skup podataka za svaki redukcioni algoritam primenjen na originalnom skupu. Redukcioni algoritmi koji su korišćeni su Low variance filter, PCA, Random Forest i Factor Analysis. Pored redukovanih skupova podataka, modul generiše i reportažu u vidu Excel dokumenta koji sadrži za svaki redukcioni algoritam informaciju o broju varijabli pre i posle njegove primene. Kao kod svakog drugog modula, moguće je proširiti broj algoritama koji će se koristiti kao algoritmi za redukciju dimenzionalnosti.

2.9. Treniranje prediktivnih modela

Softverski modul koji radi sa prediktivnim modelima je implementiran na isti način kao i predhodno spomenuti modul za redukciju dimenzionalnosti, znači ne koristeći TDD pristup razvoja. Razlog je isti, nepredvidivost rezultata treniranih modela nije dalo za mogućnost da se može unapred znati rezultat, te napisati i test za isti.

Algoritmi koji su korisćeni kao implementacija apstrakcije modula zaduženog za kreaciju prediktivnih modela su: Decision Tree, Random Forest, Multiple Linear Regression, Partial Least Squares Regression, Elastic Net Regression i XGB Regression. A sam proces treniranja modela kod većine algoritama prepušten je klasi GridSearchCV iz sklearn.model_selection biblioteke. GridSearchCV je implementacija koja nam uz pomoć unakrsne validacije prilikom treniranja modela omogućava automatizovanje procesa traženja optimalnog skupa parametara potrebnih treniranom modelu.

2.10. Validacija modela

Za potrebe validacije modela kreiran je poseban softverski modul koji u svoj ulaz metode validacije dobija dve liste podataka: stvarne vrednosti i prediktovane vrednosti, a kao rezultat vraća informaciju koliki su rezultati ostvareni za svaku metriku greške. Pošto je test problem stope samoubistava regresione prirode, za metrike greške su korišćeni koeficijent determinacije i srednji kvadrant greške. Kao i kod svakog modula, moguće je dodati nove implementacije metrika greške ukoliko za to bude bilo potrebe.

2.11. Pronalazak najznačajnije promenljive prediktivnog modela

U svrhe pronalaska najznačajnijih promenljivih modela isprobana su četiri načina. Prvi od njih je da se značajnost neke promenljive pronađe preko koeficijenata linearne regresije, drugi preko PCA Loading rezultata, a preostala dva načina pokrivaju feature_importance funkcionalnost Random Forest i XGBoost regresije.

3. REZULTATI

3.1. Rezultati prediktivnih modela

Rezultati koji će biti prikazani predstavljaju rezultat koji je dobijen na osnovu odluke korisnika da nakon obrade podataka (sa koraka predstavljenim u poglavlju 2.5) želi da nastavi sa maksimalnom tolerancijom na 20% nedostajućih podataka.

Najbolji rezultati su se pokazali kod modela istreniranih sa originalnim spojenim skupom podataka koji je u sebi sadržao 153 varijable. Prosečna tačnost tih modela iznosi za R^2 86,5% tačnosti, a MSE 0,084. Najbolje se pokazao

Random Forest algoritam sa 98,8% tačnosti za R^2 meru i 0,007 za MSE. Dok, XGBoost završio sa najlošijim rezultazima od svega 43,6% tačnosti za R^2 i 0,352 za MSE. Ostali modeli za neredukovani skup podataka u proseku su imali 94,1% za R^2 i 0,036 za MSE.

Rezultati modela na osnovu dimenziono redukovanih skupova podataka dala je dosta lošije rezultate. Jedini izuzetak predstavlja Low Variance filter koji je sa parametrom granične vrednosti od 20% dao najbolji rezultat od 93,7% za R^2 i 0,039 za MSE. Tačnost kod ostalih modela nije prelazila 45% za R^2 i nije bila manja od 0,35 za MSE.

3.1. Najznačanije varijable prediktivnog modela

Za otkrivanje najznačajnijih varijabli problema stope samoubistava jedino su metode od strane Random Forest i XGBoost algoritma dale konkretne rezultate.

Za najznačajniju varijablu modela Random Forest je izabrao stopu plodnosti sa nešto više od 20% uticaja na odluku o izboru prediktivne vrednosti. Između 10% i 15% uticaja imaju variajble koje označavaju depresivne poremećaje, pitanje da li je država Šri Langa i poremećaj upotrebe alkohola i supstanci. Što su rezultati veoma slični kao u radu [1].

4. ZAKLJUČAK

Najbolji rezultati dobijeni su sa originalnim skupom podataka i Random Forest Algoritmom. Njegove predikcije imaju tačnost od 98,9% za R^2 i 0,007 za MSE. Takođe, ostatak algoritama, sem XGBoost-a su pokazali dobre rezultate na originalnom skupu podataka. Modeli trenirani na osnovu redukovanih skupova podataka nisu dobili dobre rezultate. Ovo u neku ruku i ima smisla, jer samom redukcijom podataka gube se potencijalno bitne informacije za trenirani model. Takođe, sama primena algoritama za redukciju dimenzionalnosti je više namenjena za skupove podataka koji imaju i po par miliona varijabli, u cilju izbegavanja pretreniranosti modela ili radi poboljšanja performansi samog trenižnog procesa imajući u vidu da će to rezultirati malo lošijim predikcijama. U posmatranom test slučaju kreiranja modela za predikciju stope samoubistava, spojeni skup podataka sadržao je „svega'' 153 varijable. Trenižni proces je takođe bio brz, stoga, benefiti redukcije dimenzionalnosti nisu pogledali svetlost dana kod ovog problema. Za analizu najznačajnijih varijabi uzeta je u obzir reportaža Random Forest algoritma.

Modularnost sistema softverskog rešenja i način na koji je on razvijen omogućava proširenja ili izmene na bilo kom njegovom delu. Stoga, mogla bi se izmeniti funkcionalnost od načina kako se učitavaju skupovi podataka (na primer, koje su obavezne kolone), koje će statistike da se prikazuju i na koji način, način dopune nedostajućih vrednosti, pa sve do dodavanja ili izmene algoritama za redukciju dimenzionalnosti, algoritama za predikciju ili novih metrika koje validiraju predikcije modela.

5. LITERATURA

[1] https://github.com/CodeEatSleepRepeat/Data-Mining/blob/master/izvestaj.pdf (pristupljeno u maju 2021.)

[2] https://www.digite.com/agile/test-driven-development-tdd/ (pristupljeno u maju 2021.)

[3] https://en.wikipedia.org/wiki/SOLID (pristupljeno u maju 2021.)

[4] https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-low-variance-filter-and-its-implementation/ (pristupljeno u maju 2021.)

[5] https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 (pristupljeno u maju 2021.)

[6] https://heartbeat.fritz.ai/understanding-the-mathematics-behind-principal-component-analysis-efd7c9ff0bb3 (pristupljeno u maju 2021.)

[7] Factor Analysis - Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in quantitative methods for psychology, 9(2), 79-94.

[8] https://medium.com/swlh/decision-tree-regression-and-its-mathematical-implementation-58c6e9c5e88e (pristupljeno u maju 2021.)

[9] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley & Sons.

[10] http://www.imm.dtu.dk/~perbb/MAS/ST116/module03/module.pdf (pristupljeno u maju 2021.)

[11] https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6 (pristupljeno u maju 2021.)

[12] https://www.youtube.com/watch?v=OtD8wVaFm6E (pristupljeno u maju 2021.)

[13] https://medium.com/@erika.dauria/looking-at-r-squared-721252709098 (pristupljeno u maju 2021.)

Kratka biografija:

Aleksandar Stanić rođen je u Vrbasu 1996. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike –Sistemi za istraživanje i analizu podataka odbranio je 2021.god.
kontakt: sale96@protonmail.com