Predikcija i usporedba statusa kredita korištenjem random forest modela i logističke regresije unutar Power BI Okruženja

BELMA ĐELILOVIĆ[1]
Softversko inženjerstvo, Politehnički fakultet, Univerziteta u Zenici
Autor: Belma Đelilović (e-mail: belma.djelilovic.20@size.ba).

APSTRAKT Ovaj rad analizira tačnost predikcije i usporedbu statusa kredita koristeći Random Forest model i logističku regresiju, primijenjene na podatke iz Relational Financial Dataset-a unutar Power BI okruženja. Primarni cilj istraživanja je evaluacija performansi oba modela u kontekstu predikcije i analize statusa kredita, koristeći različite finansijske parametre kao ulazne varijable. Rad uključuje detaljan pregled literature, objašnjenje Random Forest algoritma i logističke regresije, opis dataset-a i metodologije korištene za treniranje i evaluaciju modela. Random Forest je moćna metoda koja koristi kombinaciju više odluka za predviđanje kontinuiranih varijabli, dok logistička regresija nudi robusnost u klasifikaciji binarnih varijabli koje su prilagođene za predikciju iznosa kredita. Podaci su prethodno obrađeni i podijeljeni na trening i test setove kako bi se osigurala validnost modela. Evaluacija modela obuhvata metrike kao što su srednja kvadratna greška (MSE) i koeficijent determinacije (R^2) za Random Forest, te ROC AUC i preciznost za logističku regresiju, čime se dobija sveobuhvatna slika uspješnosti predikcija. Rezultati pokazuju visoku tačnost oba modela u predikciji statusa kredita, uz identifikaciju prostora za dalja poboljšanja i optimizacije. Power BI je korišten za vizualizaciju rezultata, omogućavajući preglednost i razumijevanje predikcija u različitim rasponima statusa kredita. Diskusija se fokusira na analizu rezultata, ograničenja istraživanja i prijedloge za buduća istraživanja, uključujući dalju primjenu naprednijih metoda i drugih algoritama.

INDEX RIJEČI predikcija statusa kredita, random forest, logistička regresija, Power BI, matrica konfuzije, finansijski podaci, mašinsko učenje

I. UVOD

OPTIMIZACIJA procesa donošenja odluka u finansijskom sektoru postala je imperativ zbog sve veće složenosti i obima podataka. Precizna predikcija statusa kredita igra ključnu ulogu u smanjenju rizika i povećanju profitabilnosti finansijskih institucija. Tradicionalne metode procjene kreditne sposobnosti često nisu dovoljne da obuhvate sve varijable koje utiču na donošenje odluka, što stvara potrebu za naprednijim analitičkim tehnikama.

Random Forest, kao napredna metoda mašinskog učenja, pruža moćan način za analizu i predikciju kontinuiranih varijabli koristeći kombinaciju više odlučujućih stabala. S druge strane, logistička regresija nudi robustnost u klasifikaciji binarnih varijabli. Ovaj rad istražuje primjenu oba modela za predikciju statusa kredita koristeći podatke iz Relational Financial Dataset-a u Power BI okruženju. Primarni cilj istraživanja je evaluacija performansi oba modela u kontekstu predikcije statusa kredita, kao i identifikacija ključnih faktora koji utiču na njihovu tačnost.

Kroz detaljnu analizu rezultata, korištenjem matrice konfuzije vizualizaciju predikcija u Power BI-u, ovaj rad nastoji doprinijeti boljem razumijevanju potencijala mašinskog učenja u finansijskom sektoru. Diskusija se fokusira na rezultate predikcija, njihove implikacije na praktične primjene i mogućnosti za buduća istraživanja.

II. PREGLED LITERATURE

U ovom radu istraživali smo različite alate i algoritme mašinskog učenja vezane za predikciju kreditnog statusa. Predikcija kreditnog statusa često se povezuje sa različitim tehnikama mašinskog učenja, regresionim modelima i analizom podataka. Fokus se stavlja na evaluaciju performansi modela, njihovu sposobnost generalizacije i identifikaciju ključnih faktora koji utiču na tačnost predikcija.

A. DATASET

Dataset korišten u ovom istraživanju preuzet je sa Relational Financial Dataset. PKDD'99 Financial[4] dataset je popularan dataset koji se koristi za analizu finansijskih podataka i predikciju kreditnog rizika. Ovaj dataset sadrži nekoliko tablica koje predstavljaju različite aspekte finansijskih transakcija i informacija o klijentima. Tabele ovog dataseta: loan, order, trans, account, card, disp, client, district.

III. ALATI

A. POWERBI

Power BI je zbirka softverskih usluga, aplikacija i konektora koji rade zajedno kako bi vaši nepovezani izvori podataka postali koherentni, vizualno impresivni i interaktivni uvidi. Vaši podaci mogu biti Excel tabela ili zbirka hibridnih skladišta podataka koja se nalaze u oblaku i na lokalnim serverima. Power BI vam omogućava da lako povežete svoje izvore podataka, vizualizujete i otkrijete šta je važno, te to podijelite s bilo kim ili svima koje želite.[1]

Power BI se sastoji od nekoliko elemenata koji rade zajedno, počevši od ovih osnovnih:

- Aplikacija za Windows desktop pod nazivom Power BI Desktop.
- Online usluga kao softver kao usluga (SaaS) pod nazivom Power BI servis.
- Power BI mobilne aplikacije za Windows, iOS i Android uređaje.[1]

Ova tri elementa—Power BI Desktop, servis i mobilne aplikacije—dizajnirana su kako bi vam omogućili kreiranje, dijeljenje i konzumiranje poslovnih uvida na način koji najefikasnije odgovara vašoj ulozi.[1]

Pored ova tri elementa, Power BI također sadrži još dva:

- Power BI Report Builder, za kreiranje paginiranih izvještaja koji se mogu dijeliti u Power BI servisu. Više o paginiranim izvještajima možete pročitati kasnije u ovom članku.
- Power BI Report Server, on-premises server za izvještaje gdje možete objaviti svoje Power BI izvještaje nakon što ih kreirate u Power BI Desktop-u. Više o Power BI Report Server-u možete pročitati kasnije u ovom članku.[1]

Jedan čest radni tok u Power BI-u počinje povezivanjem s izvorima podataka u Power BI Desktop-u i izradom izvještaja. Zatim taj izvještaj objavite iz Power BI Desktop-a na Power BI servis i podijelite ga kako bi poslovni korisnici u Power BI servisu i na mobilnim uređajima mogli pregledati i interaktirati s izvještajem.[1]

Ovaj radni tok je uobičajen i pokazuje kako tri glavna Power BI elementa međusobno dopunjuju jedan drugog.[1]

1) Matrice in PowerBi

Matrica vizualizacija u Power BI-u je moćan alat koji se koristi za prikazivanje i analizu podataka kroz više dimenzija. Slično kao pivot tabela u Excelu, matrica vizualizacija omogućava organizovanje podataka u redove i kolone, čineći razumijevanje složenih skupova podataka lakšim. Svaka ćelija u matrici predstavlja specifičnu tačku podataka, s mogućnošću dubljeg istraživanja u detaljnije nivoe informacija. Ovaj dizajn omogućava sveobuhvatan pregled podataka, omogućujući korisnicima da identifikuju obrasce i trendove kroz različite dimenzije.[2]

Sposobnost matrice vizualizacije da agregira podatke i podržava hijerarhijske strukture čini je ključnim alatom za dubinsku analizu podataka i izveštavanje. Matrinske vizualizacije se često koriste u poslovanju i financijama za analizu performansi, kao što su prodaja po regionima ili kategorijama proizvoda. Također, od velike su vrijednosti u marketingu za ispitivanje ponašanja potrošača, efikasnosti kampanja i segmentacije tržišta. [2]

B. PYTHON I MAŠINSKO UČENJE

Python je izuzetno popularan izbor za razvoj modela mašinskog učenja zbog svoje jednostavnosti, fleksibilnosti i moćne ekosisteme biblioteka. Zbog jednostavne sintakse, Python olakšava validaciju podataka, kao i procese poput sakupljanja, obrade, čišćenja i analize podataka, što omogućava lakšu suradnju među programerima. Python također nudi veliki broj biblioteka, poput Scikit-Learn, koje olakšavaju implementaciju osnovnih algoritama mašinskog učenja.

Python je izuzetno fleksibilan, podržava rad s drugim programskim jezicima i može se koristiti na svim operativnim sistemima, uključujući Windows, macOS, Linux i Unix. Zbog svoje čitljivosti i podrške široke zajednice programera, Python je vrlo poželjan jezik, a zajednica stalno doprinosi razvoju novih paketa koji olakšavaju rad u mašinskom učenju. Ukratko, Python omogućava lakše rješenje problema i brže razvojne cikluse u mašinskom učenju.

IV. PRIPREMA PODATAKA

Proces pripreme i čišćenja podataka započeo je preuzimanjem PKDD'99 Financial dataset-a, koji sadrži širok spektar informacija relevantnih za analizu finansijskih podataka i predikciju kreditnog rizika. Kako bismo osigurali tačnost i pouzdanost naših analiza, odabrali smo pet ključnih tablica iz dataset-a: loan, trans, account, client i district. Svaka od ovih tablica pružila je specifične podatke koji su bili ključni za našu studiju. U prvom koraku, fokusirali smo se na čišćenje podataka kako bismo uklonili sve potencijalne pristranosti i nesigurnosti koje bi mogle uticati na rezultate analize. Ovo je uključivalo identifikaciju i zamjenu nedostajućih vrijednosti, koje bi u suprotnom mogle dovesti do netačnih predikcija. Korištene su različite metode za rješavanje nedostajućih vrijednosti, uključujući imputaciju srednjom vrijednošću ili modalnim vrijednostima gdje je to bilo prikladno, kao i potpuno uklanjanje zapisa koji su imali previše nedostajućih podataka. Normalizacija podataka bila je sljedeći ključni korak. Kako bismo osigurali da svi podaci budu konzistentni i usporedivi, skalirali smo numeričke vrijednosti u odgovarajuće opsege. Ovo je omogućilo da naš model efikasnije obrađuje podatke i poboljša tačnost predikcija. Transformacija varijabli također je bila potrebna za pripremu podataka

za Random Forest model. Kategorijalne varijable, kao što je status kredita, pretvorene su u numeričke formate koje model može lako obraditi.

Nakon čišćenja i normalizacije, pristupili smo fazi spajanja odabranih tablica. Kako bismo spojili podatke iz različitih izvora, prvo smo spojili tabelu klijenata s tabelom distrikta koristeći zajednički ključ district-id. Zatim smo spojili računovodstvene podatke s distriktima i klijentima, također koristeći district-id. Konačno, spojili smo podatke o kreditima s računovodstvenim podacima koristeći account-id. Ova faza spajanja omogućila nam je da objedinjavamo sve relevantne informacije iz različitih tablica u jedinstven skup podataka.

Nakon spajanja podataka, pristupili smo fazi inženjeringa karakteristika. Izračunali smo godine klijenata na osnovu njihovog datuma rođenja kako bismo dobili preciznu informaciju o dobi svakog klijenta. Također smo pretvorili status kredita u binarnu vrijednost, gdje je 1 označavalo odobreni kredit, a 0 odbijeni kredit.

Zatim smo selektovali relevantne kolone za modeliranje, uključujući trajanje kredita (duration), mjesečne isplate (payments) i godine klijenata (age). Ove kolone su služile kao ulazne varijable (X), dok je kolona iznos kredita (amount) predstavljala izlaznu varijablu (y).

Spajanjem podataka iz različitih izvora, kreirali smo sveobuhvatan i kvalitetan dataset koji je sadržavao ključne informacije potrebne za našu studiju. Ovaj pripremljeni dataset poslužio je kao osnova za treniranje Random Forest modela, omogućavajući preciznu predikciju iznosa kredita.

V. RANDOM FOREST

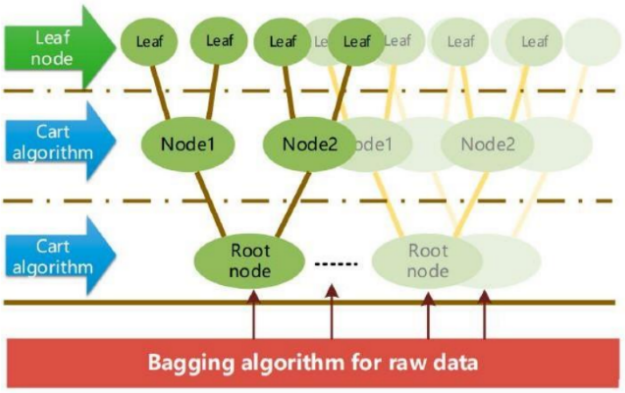
Random forest su kombinacija stabla prediktora, gdje svako stablo zavisi od vrijednosti nasumičnog vektora uzorkovanog nezavisno i s istom distribucijom za sva stabla u šumi. Generalizacijska greška za šume konvergira a.s. do limita kako broj stabala u šumi postaje veliki. Generalizacijska greška šume stabala klasifikatora zavisi od snage pojedinačnih stabala u šumi i korelacije među njima. Korištenje nasumične selekcije značajki za podjelu svakog čvora daje stope grešaka koje su povoljne u poređenju s Adaboost-om[5], ali su robusnije u odnosu na šum. Interni procjenitelji prate grešku, snagu i korelaciju, a ovi podaci se koriste za prikaz odgovora na povećanje broja značajki korištenih u podjeli. Interni procjenitelji se također koriste za mjerenje važnosti varijabli. Ove ideje su također primjenjive na regresiju. Značajna poboljšanja u tačnosti klasifikacije postignuta su rastom skupa stabala i omogućavanjem da glasaju za najpopularniju klasu. Da bi se uzgajali ovi skupovi, često se generišu nasumični vektori koji upravljaju rastom svakog stabla u skupu. Rani primjer je bagging [6], gdje se za rast svakog stabla vrši nasumična selekcija (bez ponavljanja) iz primjera u skupu za treniranje. Još jedan primjer je nasumična selekcija podjele [7], gdje se na svakom čvoru podjela bira nasumično među K najboljih podjela. Breiman [8] generiše nove skupove za treniranje randomizacijom izlaza u originalnom skupu za treniranje. Drugi pristup je odabir skupa za treniranje iz nasumičnog skupa težina na primjerima u skupu za treniranje.

A. OBJAŠNJENJE RADA RANDOM FOREST ALGORITMA

Slika 1. pruža intuitivno razumijevanje radio frekvencija. Uočeno je da dnevna količina punjenja različitih punionica pokazuje specifičnu karakteristiku, naime, količina punjenja je značajno raspršena. Zbog toga je proces dijeljenja originalnih podataka u korake neophodan. Intervali pokrivenosti opsega odabrani su na temelju raspona vrijednosti količine isporuke podataka. Ovaj pristup poboljšava efikasnost i preciznost RF algoritma predikcije eliminisanjem manjih smetnji. Postoje dva osnovna principa za dijeljenje perioda[9]:

Raspon količine isporuke se dijeli na jednake vremenske periode kako bi se osigurala jednaka raspodjela vremena. Ovo osigurava da su podaci organizovani na konzistentan način, što olakšava identifikaciju trendova i obrazaca.

Periodi se dijele prema jačini količine punjenja. U ovom scenariju, vremenski intervali variraju i zavise od koncentracije podataka u specifičnim regijama unutar raspona. Ova vrsta segmentacije omogućava fokusiranje na najvažnije dijelove podataka koji sadrže složenije i statistički značajne informacije.



Slika 1. Shema dijagrama algoritma Random forest

Segmentacija podataka ima nekoliko prednosti. Jedna od ključnih prednosti je smanjenje smetnji. Dijeljenjem podataka u specifične periode, smanjuje se interferencija između podataka. Ovo smanjenje smetnji na kraju vodi do poboljšane tačnosti predikcija. Segmentacija podataka poboljšava efikasnost algoritama korištenih u analizi podataka i RF predikcijama, čime se povećava njihova ukupna performansa. Analiza je jednostavnija kada su podaci adekvatno segmentirani, jer to omogućava lakše prepoznavanje obrazaca i trendova. Primjenom ovih koncepta za dijeljenje vremenskih intervala, algoritmi se mogu unaprijediti, što vodi do preciznijih i učinkovitijih predikcija. Ovo, zauzvrat, omogućava tačnije razumijevanje i predikciju RF frekvencija. [9]

B. IMPLEMENTACIJA RANDOM FOREST ALGORITMA ZA PREDIKCUJU STATUSA

Koristeći Random Forest algoritam, izvršili smo nekoliko koraka kako bismo osigurali tačnu i pouzdanu predikciju statusa kredita. Prvo smo podijelili podatke na trening i test skupove korištenjem funkcije train-test-split() iz biblioteke sklearn. Postavljena veličina test seta na 30 procenata znači da je 70 procenata podataka išlo na trening set, dok je 30 procenata podataka išlo na test set. Preciznije, ukoliko imamo ukupno 1000 podataka, 700 podataka bi bilo korišteno za trening (X-train, y-train), a preostalih 300 za testiranje (X-test, y-test). Ova podjela je ključna jer omogućava da model trenira na većem dijelu podataka i da se kasnije evaluira na manjem, neviđenom dijelu podataka, čime se dobija realna procjena njegove performanse.

Kada je podjela izvršena, koristili smo RandomForestClassifier za treniranje modela. Random Forest algoritam funkcioniše tako što kreira mnoštvo stabala odlučivanja tokom treniranja, gdje svako stablo koristi različite uzorke podataka i karakteristike. Svako stablo donosi svoje predikcije, a konačna odluka algoritma je bazirana na većini predikcija stabala. Ovaj pristup, koji uključuje nasumično biranje podskupova podataka i karakteristika, smanjuje mogućnost prekomjernog učenja (overfitting) i poboljšava robusnost modela.

Nakon što smo trenirali model, izvršili smo predikciju statusa kredita za podatke iz test seta pomoću metode predict(). Rezultate predikcija evaluirali smo koristeći metrike kao što su tačnost (accuracy), matrica konfuzije (confusion matrix) i klasifikacijski izvještaj (classification report). Tačnost nam pokazuje koliko su predikcije modela tačne u odnosu na stvarne vrijednosti, dok matrica konfuzije pruža dublji uvid u tačnost i netačnost predikcija po klasama. Klasifikacijski izvještaj pruža dodatne metrike kao što su preciznost, odziv i F1-score za svaku klasu.

Pored toga, koristili smo križnu validaciju (cross-val-score) kako bismo dodatno evaluirali stabilnost i performanse modela. Križna validacija podrazumijeva podjelu podataka na više dijelova i treniranje modela više puta, pri čemu se svaki put koristi različiti dio podataka za testiranje, a ostali za treniranje. Ovo osigurava da procjena performansi modela bude preciznija i pozdanija.

VI. LOGISTIČKA REGRESIJA

Logistička regresija je jedan od najvažnijih analitičkih alata u društvenim i prirodnim naukama. U obradi prirodnog jezika, logistička regresija je osnovni nadzirani algoritam mašinskog učenja za klasifikaciju i takođe ima veoma blisku vezu sa neuronskim mrežama. [6] Logistička regresija radi vrlo slično linearnoj regresiji, ali sa binomnom zavisnom varijablom. Najveća prednost u poređenju sa Mantel-Haenszel OR je činjenica da možete koristiti kontinuirane objašnjavajuće varijable i lakše je rukovati sa više od dvije objašnjavajuće varijable istovremeno. Iako se čini trivijalnim, ova posljednja karakteristika je suštinska kada smo zainteresirani za utjecaj različitih objašnjavajućih varijabli na zavisnu varijablu. Ako promatramo više objašnjavajućih varijabli neovisno, ignoriramo kovarijaciju među varijablama i izloženi smo učincima miješanja, kao što je demonstrirano u gore navedenom primjeru kada je učinak tretmana na vjerojatnost smrti bio djelomično skriven učinkom starosti.[11]

Logistička regresija će modelirati šansu za ishod na osnovu individualnih karakteristika. Budući da je šansa omjer, ono što će zapravo biti modelirano je logaritam šanse dat sljedećom formulom:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (1)$$

gdje označava vjerojatnost događaja (npr., smrt u prethodnom primjeru), a i su regresijski koeficijenti povezani sa referentnom grupom i xi objašnjavajućim varijablama. U ovom trenutku, važno je istaknuti jedan važan koncept. Referentna grupa, predstavljena sa 0, čine oni pojedinci koji prezentiraju referentni nivo svake varijable x1...m.[11]

A. OBJAŠNJENJE RADA LOGISTIČKE REGRESIJE

Logistička regresija je vjerovatnosni klasifikator koji koristi nadzirano mašinsko učenje. Klasifikatori mašinskog učenja zahtijevaju trening korpus od m ulaznih/izlaznih parova $(x^{(i)}, y^{(i)})$. Sistem mašinskog učenja za klasifikaciju ima četiri komponente[10]:

1. **Reprezentacija karakteristika ulaza** - za svaku ulaznu opservaciju $x^{(i)}$, ovo će biti vektor karakteristika $[x_1, x_2, \ldots, x_n]$. Generalno ćemo se referencirati na karakteristiku i za ulaz $x^{(j)}$ kao $x_i^{(j)}$, ponekad pojednostavljeno kao x_i, ali ćemo također koristiti oznake f_i, $f_i(x)$, ili za multiklasa klasifikaciju $f_i(c, x)$.

2. **Funkcija klasifikacijen** koja računa ^y, procijenjenu klasu, putem p(y|x). U sljedećem odjeljku ćemo uvesti sigmoid i softmax alate za klasifikaciju.

3. **Ciljna funkcija** koju želimo optimizirati za učenje, obično uključujući minimiziranje funkcije gubitka koja odgovara grešci na trening primjerima. Uvest ćemo cross-entropy funkciju gubitka.

4. **Algoritam za optimizaciju ciljne funkcije**.

Logistička regresija ima dvije faze:

1. Trening: Treniramo sistem (posebno težine w i b, uvedene ispod) koristeći stohastičko gradijentno spuštanje i cross-entropy gubitak.

2. Test: Dajući test primjer x, računamo p(y|x) i vraćamo oznaku sa većom vjerovatnoćom y = 1 ili y = 0. [10]

Na primjer, imamo dvije klase: Klasa 0 i Klasa 1. Ako je vrijednost logističke funkcije za neki ulaz veća od 0.5 (prag vrijednost), tada pripada Klasi 1, inače pripada Klasi 0. Naziva se regresijom jer je produžetak linearne regresije, ali se uglavnom koristi za probleme klasifikacije.

B. IMPLEMENTACIJA LOGISTIČKE REGRESIJE ZA PREDIKCUJU STATUSA

Koristeći logističku regresiju, poduzeli smo nekoliko ključnih koraka kako bismo predvidjeli status kredita. Prvo

smo podijelili podatke na trening i test skupove koristeći funkciju train-test-split() iz biblioteke sklearn. Test skup je činilo 30 procenata podataka, dok je preostalih 70 procenata korišteno za trening modela. Ova podjela omogućava modelu da uči iz većeg dijela podataka, a testiranje na neviđenim podacima daje realnu procjenu performansi modela. Na primjer, ako imamo ukupno 1000 podataka, 700 podataka bi bilo za trening (X-train, y-train), a 300 za testiranje (X-test, y-test).

Nakon podjele, trenirali smo model logističke regresije koristeći klasu LogisticRegression iz sklearn. Ovaj model koristi metodologiju optimizacije kako bi se odredili koeficijenti za različite značajke koje predviđaju status kredita. Korištenje multinomijalne logističke regresije omogućava modelu da predviđa četiri moguće kategorije statusa kredita: A, B, C, i D. Svaka od tih kategorija predstavlja različite razine rizika u vraćanju kredita, a model koristi značajke kao što su iznos kredita, broj uplata i starost klijenta za određivanje vjerojatnosti svake od tih kategorija. Trening modela omogućava mu da pronađe optimalne parametre koji minimiziraju razliku između predviđenih i stvarnih vrijednosti.

Nakon što smo trenirali model, izvršili smo predikciju statusa kredita za testne podatke pomoću metode predict(). Rezultati predikcija su evaluirani koristeći metrike kao što su tačnost, matrica konfuzije i klasifikacijski izvještaj. Tačnost nam pokazuje postotak točnih predikcija u odnosu na stvarne vrijednosti, dok matrica konfuzije omogućava dublje razumijevanje grešaka modela po klasama. Klasifikacijski izvještaj daje dodatne metrike poput preciznosti, odziva i F1-skorova za svaku od kategorija statusa kredita. Na kraju, kako bismo procijenili stabilnost modela, koristili smo kros-validaciju, koja uključuje treniranje i testiranje modela na različitim dijelovima podataka više puta, čime se dobija pouzdanija procjena performansi modela.

VII. ANALIZA REZULTATA

A. ANALIZA REZULTATA SCATTER PLOTA

Graf prikazuje poređenje stvarnih i predviđenih statusa kredita u odnosu na iznos kredita. Korišteni su Random Forest i Logistička regresija kako bi se predvidio status kredita, pri čemu je svaka tačka na grafu predstavljala jedan kredit. Os X prikazuje iznos kredita, dok os Y označava status kredita koji može biti A, B, C ili D. Različiti oblici i boje korišteni su za vizualizaciju stvarnih vrijednosti i predikcija modela, što omogućava direktno poređenje performansi algoritama.

Analizom grafičkog prikaza može se primijetiti da su predikcije modela u mnogim slučajevima tačne, ali postoje i određena neslaganja između stvarnih i predviđenih vrijednosti. Statusi A i C su relativno konzistentno predviđeni, pri čemu većina predikcija odgovara stvarnim vrijednostima. S druge strane, kod statusa B primjećuje se veća odstupanja, što ukazuje na to da modeli imaju poteškoće u razlikovanju ove kategorije. Također, kod statusa D može se vidjeti da Logistička regresija pokazuje veću disperziju predikcija, dok Random Forest izgleda preciznije predviđa ovu kategoriju.

Jedan od mogućih razloga za neslaganja u predikcijama može biti neravnomjerna raspodjela podataka među kategorijama. Ako su određene klase manje zastupljene u skupu podataka, modeli mogu imati problema s učenjem njihovih obrazaca. To bi se moglo poboljšati balansiranjem podataka putem tehnika poput oversamplinga ili undersamplinga. Osim toga, predikcije se trenutno zasnivaju samo na tri značajke: iznosu kredita, rati otplate i starosti klijenta. Dodavanje dodatnih informacija, kao što su podaci o prethodnim kreditima ili finansijskoj stabilnosti klijenta, moglo bi značajno poboljšati preciznost modela.

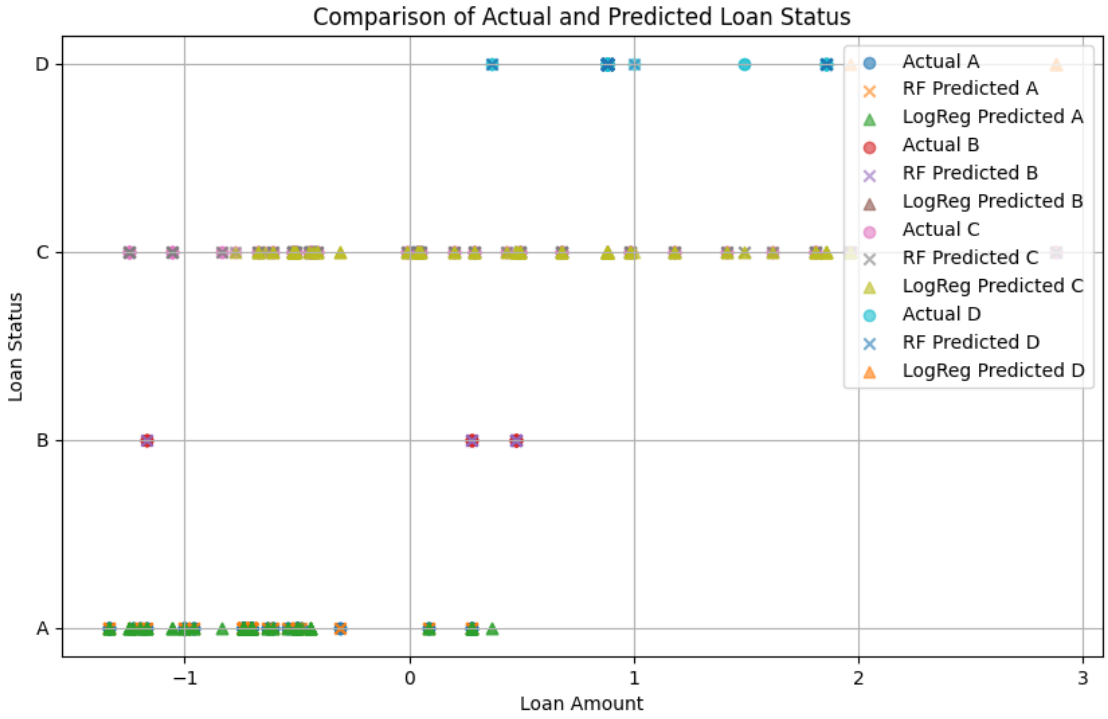
Još jedan aspekt koji bi mogao poboljšati performanse modela jeste optimizacija hiperparametara. Random Forest model bi se mogao poboljšati prilagođavanjem broja stabala u šumi ili dubine svakog stabla, dok bi Logistička regresija mogla imati bolju tačnost uz korištenje različitih tehnika regularizacije. Također, korištenje drugih algoritama poput neuronskih mreža ili gradijentnog boostinga moglo bi se ispitati kako bi se vidjelo da li bi donijeli bolje rezultate.

Možemo reći da Random Forest model generalno postiže veću tačnost u predikcijama u poređenju s Logističkom regresijom, ali i dalje postoje slučajevi kada su predikcije pogrešne. Određene klase, poput B, predstavljaju izazov za oba modela, što sugeriše potrebu za dodatnim poboljšanjima. Optimizacijom modela, dodavanjem novih značajki i boljim balansiranjem podataka moguće je postići preciznije predikcije statusa kredita i time poboljšati upotrebljivost modela u realnim scenarijima.

B. EVALUACIJA LOGISTIČKE REGRESIJE I RANDOM FORESTA

Random Forest model je pokazao izuzetno visoku tačnost od 98,5 procenata, što sugerira da je model vrlo efikasan u predikciji statusa kredita. Preciznost (precision) i odziv (recall) su također na visokom nivou za većinu klasa, posebno za klase A i C, koje su najbrojnije u skupu podataka. Model postiže 100procenata recall za klasu C i 99 procenata za klasu A, što znači da gotovo svi pozitivni slučajevi za ove klase bivaju ispravno klasificirani. Međutim, klasa B pokazuje nešto niži recall (88 procenata) i precision (94 procenata), što ukazuje na blagi nesklad u predikciji ove kategorije. Cross-validation score od 74 procenta sa standardnom devijacijom od 30 procenata sugerira određenu nestabilnost modela prilikom generalizacije na neviđene podatke.

S druge strane, logistička regresija postiže značajno nižu tačnost od 70,2 procenata, što ukazuje da se ovaj model teže nosi sa zadatkom klasifikacije kreditnog statusa. Dok se klasa A klasificira relativno dobro sa recall vrijednošću od 98procenata, preciznost modela za ostale klase je niska. Posebno problematične su klase B i D, gdje recall iznosi 0procenata što znači da model uopće ne uspijeva ispravno klasificirati ove statuse kredita. Ovo sugerira da je model logističke regresije neadekvatan za ovaj problem, vjerovatno zbog linearnih ograničenja koja ne mogu dobro modelirati kompleksne obrasce u podacima. Cross-validation tačnost od 64 procenta uz standardnu devijaciju od 21 procenat pokazuje



Slika 2. Poređenje stvarnog i predviđenog statusa kredita korištenjem algoritama Random Forest i Logistic Regression na scatter plotu

```
Random Forest Accuracy: 0.9853801169590644
Random Forest Classification Report:
              precision    recall  f1-score   support

           A       0.99      0.99      0.99       139
           B       0.94      0.88      0.91        17
           C       0.99      1.00      0.99       135
           D       1.00      0.96      0.98        51

    accuracy                           0.99       342
   macro avg       0.98      0.96      0.97       342
weighted avg       0.99      0.99      0.99       342

Random Forest Cross-Validation Accuracy: 0.74 (+/- 0.30)
```

Slika 3. Konzolni ispis evaluacije random forest algoritma za klasifikaciju statusa kredita

```
Logistic Regression Accuracy: 0.7017543859649122
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           A       0.79      0.98      0.87       139
           B       0.00      0.00      0.00        17
           C       0.63      0.77      0.70       135
           D       0.00      0.00      0.00        51

    accuracy                           0.70       342
   macro avg       0.36      0.44      0.39       342
weighted avg       0.57      0.70      0.63       342

Logistic Regression Cross-Validation Accuracy: 0.64 (+/- 0.21)
```

Slika 4. Konzolni ispis evaluacije logističke regresije za klasifikaciju statusa kredita

nešto stabilniju, ali i slabiju prediktivnu sposobnost u odnosu na Random Forest.

U poređenju ova dva modela, jasno je da Random Forest nadmašuje logističku regresiju u skoro svim metrikama – preciznosti, odzivu, F1-score-u i ukupnoj tačnosti. Glavna prednost Random Forest-a je njegova sposobnost da modelira nelinearne odnose i da efikasno rukuje velikim brojem varijabli, dok logistička regresija pokazuje ograničenja u složenijim problemima klasifikacije. Međutim, visok cross-validation standardni otklon kod Random Forest modela može ukazivati na osjetljivost modela na različite podskupove podataka, što može biti izazov za generalizaciju u realnim scenarijima.

C. MATRICE KONFUZIJE IZ POWERBI

Prva matrica konfuzije prikazuje poređenje stvarnog statusa kredita (actual-status) s predikcijama modela logističke regresije (logreg-predicted-status). Vidimo da je model izuzetno dobar u predviđanju klase A, s 138 tačno klasificiranih primjera od ukupno 139. Također, klasa C je savršeno klasificirana, sa svih 135 primjera pravilno predviđenih. Međutim, model ima ozbiljne probleme s klasama B i D. Za klasu B, model je tačno klasificirao samo 15 od 17 primjera, dok su 2 slučaja pogrešno svrstana u klasu A. Najveći problem je s klasom D, gdje je svih 51 primjer pogrešno

actual_status	A	B	C	D	Total
A	138	1			139
B	2	15			17
C			135		135
D			2	49	51
Total	**140**	**16**	**137**	**49**	**342**

Slika 5. Matrice konfuzije za model logističke regresije

klasificiran – većina je svrstana u klasu C. Ovo ukazuje na to da logistička regresija ima problema s razdvajanjem klasa B i D, vjerojatno zbog njihove male zastupljenosti i sličnosti s drugim klasama.

actual_status	A	C	D	Total
A	136	3		139
B	9	8		17
C	25	104	6	135
D	2	49		51
Total	**172**	**164**	**6**	**342**

Slika 6. Matrice konfuzije za model Random Forest-a

Druga matrica konfuzije prikazuje poređenje stvarnog statusa kredita s predikcijama modela Random Forest (rf-predicted-status). Ovaj model također pokazuje odlične rezultate za klasu A, tačno klasificirajući 136 od 139 primjera. Međutim, primjetno je da model ima više grešaka kod klase C, gdje je samo 104 od 135 primjera tačno klasificirano, dok je čak 25 primjera greškom svrstano u klasu A. S druge strane, za klasu B, model također pokazuje slabosti – samo 8 primjera su tačno klasificirana, dok je 9 greškom svrstano u klasu A. Najveća razlika u odnosu na logističku regresiju je kod klase D, gdje su svi primjeri i dalje loše klasificirani, ali su ovdje raspoređeni između klasa C i A.

Kada usporedimo obje matrice konfuzije, primjećujemo da logistička regresija bolje klasificira klasu C, dok Random Forest bolje raspoređuje klasu B, ali lošije predviđa klasu C. Klasa A je dobro klasificirana u oba modela, dok je klasa D problematična u oba slučaja – nijedan model ne uspijeva da je pravilno predvidi. Random Forest izgleda ima veću tendenciju ka pogrešnom klasificiranju klase C u A, dok logistička regresija griješi svrstavajući klasu D u C. Ovo pokazuje da različiti modeli imaju različite prednosti i slabosti, što može ukazivati na potrebu za dodatnim inženjeringom značajki ili upotrebom drugih tehnika, kao što su balansiranje podataka ili kombinacija više modela.

VIII. ZAKLJUČAK

Evo proširenog zaključka i diskusije, uključujući natuknice za daljnje istraživanje, te uporedbe s radovima:

Zaključak

U ovom radu analizirali smo problem predikcije statusa kredita kroz vizualizaciju podataka, evaluaciju metrika tačnosti i analizu matrica konfuzije. Također, analizirana je važnost balansiranja podataka i odabira relevantnih značajki, što su ključni faktori za poboljšanje tačnosti modela.

Rezultati istraživanja pokazuju da Random Forest značajno nadmašuje logističku regresiju, sa tačnošću od 98,5 procenata u poređenju sa 70,2 procenata kod logističke regresije. Dok je logistička regresija pokazala poteškoće u predikciji složenih obrazaca, Random Forest je bolje generalizovao podatke, ali je istovremeno pokazao određenu nestabilnost pri cross-validaciji. Posebno su problematične klase B i D, gdje modeli nisu postigli zadovoljavajuće rezultate, što ukazuje na potrebu za dodatnim unaprijeđenjima.

Zaključno, naš rad pruža čvrstu osnovu za daljnje istraživanje u oblasti predikcije kreditnog statusa. Rezultati sugerišu da upotreba naprednijih metoda mašinskog učenja i detaljna obrada podataka mogu značajno poboljšati preciznost i pouzdanost modela u realnim scenarijima. Daljnje istraživanje može se fokusirati na integraciju dodatnih podataka i primjenu naprednijih tehnika kako bi se dodatno unaprijedili modeli predikcije.

IX. DISKUSIJA

Pored analize performansi modela, rad sam uporedila s relevantnim istraživanjima u oblasti mašinskog učenja za kreditne analize, što je omogućilo validaciju naših nalaza i metodoloških pristupa.

U poređenju sa radom "BANK LOAN PREDICTION USING MACHINE LEARNING TECHNIQUES "[12] jasno je da oba rada prepoznaju visoku tačnost Random Forest algoritma, ali s različitim rezultatima (99,98 procenata naspram 98,5 procenata). Prvi rad naglašava superiornost ensemble metoda u smanjenju prekomjernog prilagođavanja i poboljšanju generalizacije, dok vaš rad dodatno ukazuje na izazove s klasama B i D te potrebu za stabilnijim performansama modela pri cross-validaciji. Integracija nalaza iz oba rada može pomoći u boljem razumijevanju performansi Random Forest Classifier-a i identifikaciji područja za unapređenja, posebno u kontekstu stabilnosti i prilagodljivosti modela različitim klasama podataka.

Moj rad se usmjerava na praktične izazove u implementaciji modela, dok drugi rad nudi širu perspektivu o potencijalu ensemble metoda u prevazilaženju problema s prekomjernim prilagođavanjem. Kombinovanjem ovih uvida, buduće istraživanje može se fokusirati na razvijanje metoda za stabilizaciju performansi, posebice u slučajevima kada su podaci neuravnoteženi ili kada se suočavamo s izazovima u klasama koje model teže prepoznaje.

U poređenju sa radom "Loan Status Prediction" [13] vidimo da oba rada prepoznaju visoku tačnost Random Forest algoritma, ali s različitim rezultatima (99,98 procenata naspram 98,5 procenata). Ovaj rad naglašava superiornost ensemble metoda i potrebe za poboljšanjima u oblasti sigurnosti i pouzdanosti sistema, dok vaš rad dodatno ističe tehničke izazove s određenim klasama podataka i potrebu za daljnjim unapređenjima. Logistička regresija se pokazala manje efikasnom u poređenju s Random Forest algoritmom u oba rada, što potvrđuje potrebu za korištenjem naprednijih algoritama u predikciji kreditnog statusa.

Integracija nalaza iz oba rada može pružiti sveobuhvatan uvid u performanse različitih algoritama i identifikaciju područja za unapređenja, posebno u kontekstu stabilnosti i prilagodljivosti modela različitim klasama podataka.

Smjernice za dalje istraživanje:

- **Proširenje dataset-a** – Dodavanje demografskih i finansijskih podataka, poput prihoda i kreditne istorije.
- **Napredne tehnike obrade podataka** – Korištenje metoda za balansiranje podataka i transformaciju varijabli.
- **Optimizacija hiperparametara** – Primjena Grid Search-a ili Bayesian Optimization-a za bolje prilagođavanje modela.
- **Istraživanje naprednijih modela** – Korištenje XGBoost-a, LightGBM-a ili neuronskih mreža za bolju tačnost.
- **Analiza interpretabilnosti modela** – Primjena SHAP i LIME tehnika za bolje razumijevanje odluka modela.
- **Kombinovanje modela (Ensemble Learning)** – Korištenje stacking-a i boosting-a za poboljšanje tačnosti.
- **Istraživanje vremenskih serija** – Analiza kreditnog statusa kroz duži period za identifikaciju trendova.

Literatura

[1] Microsoft. (2025). Power BI Overview. [Online]. Dostupno: https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview. [Pristupano:23.01.2025].

[2] Microsoft. (2025). Power BI Visualization Matrix Visual. [Online]. Dostupno: https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-matrix-visual?tabs=powerbi-desktop. [Pristupano: 23.01.2025].

[3] Builtin. (2025). Python for Machine Learning. [Online]. Dostupno: https://builtin.com/machine-learning/python-machine-learning.[Pristupano: 23.01.2025].

[4] CTU Relational. (2025). Financial Dataset. [Online]. Dostupno: https://relational.fel.cvut.cz/dataset/Financial.[Pristupano: 15.01.2025].

[5] Yoav Freund and Robert E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, 19.12.1996, [Online]. Dostupno: https://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf. [Pristupano: 25.01.2025].

[6] Leo Breiman, Bagging Predictors, Septembar 1994, [Online]. Dostupno: https://www.stat.berkeley.edu/~breiman/bagging.pdf. [Pristupano: 25.01.2025].

[7] Thomas G, Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, 30.12.1997., [Online]. Dostupno: https://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf. [Pristupano: 25.01.2025].

[8] Leo Breiman, RANDOM FORESTS -RANDOM FEATURES, Septembar 1999, [Online]. Dostupno: https://www.stat.berkeley.edu/~breiman/bagging.pdf. [Pristupano: 26.01.2025].

[9] ResearchGate. (2025). Random Forest Algorithm Overview. [Online]. Dostupno: https://www.researchgate.net/publication/382419308_Random_Forest_Algorithm_Overview,[Pristupano: 27.01.2025].

[10] Daniel Jurafsky i James H. Martin., Logistic Regression, January 12, 2025 [Online]. Dostupno: https://web.stanford.edu/~jurafsky/slp3/5.pdf?form=MG0AV3,[Pristupano: 27.01.2025].

[11] Sandro Sperandei, Understanding logistic regression analysis, Februar, 2014 [Online]. Dostupno: https://www.researchgate.net/publication/260810482_Understanding_logistic_regression_analysis,[Pristupano: 27.01.2025].

[12] Md. Mahedi Hassan, BANK LOAN PREDICTION USING MACHINE LEARNING TECHNIQUES, bez datuma,[Online]. Dostupno: https://arxiv.org/pdf/2410.08886,[Pristupano: 28.01.2025].

[13] Nalla Sai Ram Reddy1, D. Ram Babu2, Shaik Sajid, Ch Nihal Reddy, Loan Status Prediction, bez datuma,[Online]. Dostupno: https://ijarsct.co.in/Paper9790.pdf,[Pristupano: 28.01.2025].