Primena Mašinskog Učenja u Predikciji Kreditnog Rizika za Optimizaciju Auto Kredita

Miloš Pavlović 2021/1057, Miron Petrik Popović 2020/0219, Nikola Marinković 2019/0336

Mašinsko učenje
Fakultet organizacionih nauka

1. Uvod

Kada finansijske institucije odobravaju auto kredite, one teže da maksimizuju broj odobrenih kredita dok minimizuju rizik od neplaćanja. Međutim, postoji značajan rizik da klijent neće biti u mogućnosti da otplati kredit. Ovo može dovesti do značajnih gubitaka za banke i druge finansijske institucije. Problem sa kojim se suočavaju ove institucije je kako pravilno proceniti kreditni rizik, kako bi se smanjila stopa neplaćanja bez suvišnog odbijanja zahteva za kredit.

Mnogi faktori mogu uticati na neplaćanje kredita, a predviđanje ovih faktora može biti izazovno zbog složenosti finansijskih podataka i promenljivih ekonomskih uslova. Iz tog razloga, finansijske institucije mogu značajno profitirati od primene mašinskog učenja kako bi unapred identifikovale klijente koji su u većem riziku od neplaćanja. Ako bi postojala mogućnost da se unapred utvrdi verovatnoća da će klijent prestati sa otplatom kredita, banke bi ne samo mogle da smanje svoje gubitke, već i da optimizuju proces odobravanja kredita, pružajući bolje usluge svojim klijentima.

2. Opis podataka

Opis dataset-a

Podaci korišćeni u ovoj analizi preuzeti su sa *Kaggle.com* iz dataset-a *L&T Vehicle Loan Default Prediction*, autora *Mamta Dhaker*, za predikciju neplaćanja prve rate kredita za vozila (*loan_default*). Ovaj dataset sadrži informacije o korisnicima, kreditima, podacima i istoriji iz kreditnog biroa, sa ukupno 41 atributom, uključujući i izlaznu varijablu *loan_default*, koja označava da li je korisnik izmirio svoju prvu ratu kredita na vreme (0 - nema default-a, 1 - default).

Opis atributa

Atributi u ovom datasetu pokrivaju tri ključne kategorije: informacije o korisniku, informacije o kreditu, i podaci iz kreditnog biroa i istorija. Prva kategorija obuhvata osnovne demografske podatke korisnika, uključujući jedinstvene identifikatore, trenutnu adresu, datum rođenja, tip zaposlenja, kao i indikatore dostupnosti različitih identifikacionih dokumenata. Ovi podaci omogućavaju praćenje i verifikaciju korisničkog profila.

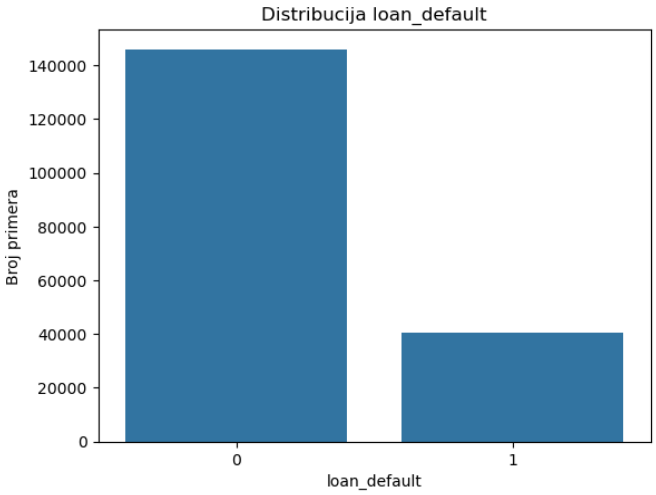
Druga kategorija se fokusira na detalje vezane za odobrene kredite, uključujući iznos kredita, cenu imovine koja se finansira, odnos vrednosti kredita prema vrednosti imovine, kao i informacije o filijali, prodavcu i proizvođaču. Ovi atributi omogućavaju dubinsku analizu finansijskih obaveza korisnika, uslova kredita, i odnosa između različitih učesnika u procesu odobrenja kredita.

Treća kategorija obuhvata kreditni skor korisnika, broj aktivnih i kašnjenja u otplati kredita, ukupne iznose odobrenih i isplaćenih kredita, kao i dužinu kreditne istorije. Ovi podaci su ključni za procenu kreditne sposobnosti korisnika, omogućavajući analizu njegovih finansijskih navika i rizika povezanih sa davanjem novih kredita.

Dizbalans klasa i opis cene različitih grešaka

U našem datasetu, izlazne klase su izrazito neravnomerno raspoređene, sa većinskom klasom koja obuhvata 78.29% instanci i manjinskom klasom koja čini 21.71% ukupnog broja. Ovaj disbalans klasa može uzrokovati pristrasnost modela ka predikciji većinske klase, što rezultira visokim procentom tačnih predikcija za nju, ali lošom tačnošću za manjinsku klasu, koja je od suštinskog značaja jer predikcija manjinske klase (neplaćanje kredita) nosi veći rizik.

Greške koje model može napraviti u ovom kontekstu imaju različite implikacije. Lažno pozitivne greške, kada model pogrešno predvidi da korisnik neće platiti kredit, mogu dovesti do neopravdanog odbijanja kredita, gubitka potencijalnih prihoda, i nezadovoljstva klijenata. S druge strane, lažno negativne greške, gde model ne prepoznaje korisnika koji ne plaća kredit, mogu dovesti do odobravanja kredita korisnicima koji kasnije neće plaćati kredit, što rezultira direktnim finansijskim gubicima.



3. Priprema podataka

Nedostajuće vrednosti

Za atribut *Employment.Type*, koji je jedini atribut sa nedostajućim vrednostima (6,069), odlučili smo da dodamo vrednost "*Unknown*" umesto da brišemo redove sa nedostajućim podacima. Ova odluka omogućava očuvanje što više informacija, budući da brisanje redova može dovesti do gubitka važnih podataka koji su relevantni za analizu ili modeliranje. S obzirom na to da je broj nedostajućih vrednosti relativno mali u odnosu na ukupni skup podataka, ovaj pristup minimizira uticaj nedostajućih podataka na rezultate analize, dok istovremeno omogućava bolje očuvanje integriteta skupa podataka.

Transformacija tipova podataka

U procesu pripreme podataka, izvršena je transformacija tipova podataka kako bi se poboljšala efikasnost obrade i analize. Numerički (*int64*) i tekstualni (*object*) tipovi podataka su promenjeni u category gde je to bilo odgovarajuće.

Identifikatori poput *branch_id*, *supplier_id*, *i manufacturer_id* su konvertovani u *category* tip jer predstavljaju diskretne kategorije i ne zahtevaju numeričke operacije. Takođe, atributi kao što su *Employment_Type* i različiti flagovi (*MobileNo_Avl_Flag*, *Aadhar_flag*, *PAN_flag*, *VoterID_flag*, *Driving_flag*, *Passport_flag*) su pretvoreni u category zbog svoje prirode binarnih ili nominalnih vrednosti. Ova promena omogućava efikasnije skladištenje i analizu, smanjujući potrošnju memorije i poboljšavajući performanse.

Pored toga, *loan_default* takođe je promenjen u category kako bi se unapredila analiza u klasifikacionim modelima. Ove transformacije optimizuju obradu podataka, čineći analize bržim i preciznijim.

Izvodjenje novih atributa

Kako bi se pojednostavila struktura podataka i smanjila redundansa, originalni atributi *AVERAGE_ACCT_AGE*, *CREDIT_HISTORY_LENGTH*, *Date_of_Birth*, i *DisbursalDate* su uklonjeni nakon što su njihovi korisni podaci transformisani u nove kolone.

Date_of_Birth je pretvoren u atribut *Age*, koji jednostavno prikazuje starost korisnika u godinama. *DisbursalDate* je korišćen za izračunavanje *Days_from_Ref*, koji pokazuje broj dana od najstarijeg datuma u skupu podataka. Ove promene smanjuju kompleksnost obrade i olakšavaju praćenje vremenskih intervala između transakcija, jer koriste čiste brojeve godina i dana umesto složenih datumskih formata.

AVERAGE_ACCT_AGE i *CREDIT_HISTORY_LENGTH*, pretvoreni su u mesece, što je omogućilo stvaranje novih atributa *AVERAGE_ACCT_AGE_MONTHS* i *CREDIT_HISTORY_LENGTH_MONTHS*. Ova konverzija osigurava uniformnost i olakšava upotrebu, jer su prethodni atributi bili izraženi u kombinaciji godina i meseci, što je moglo otežati preciznu analizu i poređenje.

Uklanjanjem originalnih kolona, smanjena je složenost skupa podataka i poboljšana čitljivost, dok su novim atributima pružene jasnije i direktnije informacije za analizu i modelovanje.

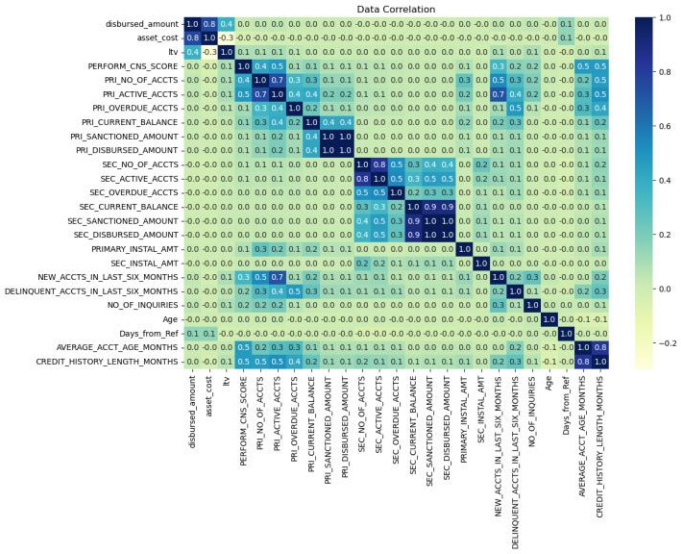
Izbacivanje nepotrebnih atributa

Neki atributi su uklonjeni zbog njihove ograničene korisnosti u analizi. *UniqueID* je izbačen jer svaka instanca ima jedinstvenu vrednost (186,523 instanci), što ga čini nepotrebnim za analitičke svrhe. Atribut *MobileNo_Avl_Flag* je takođe uklonjen jer sadrži samo jednu vrednost za sve instance, što ne doprinosi razlikovanju podataka. Identifikatori kao što su *supplier_id* (2,915 instanci), *Current_pincode_ID* (6,513 instanci), *branch_id* (82 instanci), i *Employee_code_ID* (3,258 instanci) su izbačeni zbog velikog broja jedinstvenih vrednosti koje služe samo za identifikaciju i nisu korisni za analizu. Iako su *manufacturer_id* (11 instanci) i *State_ID* (22 instanci) identifikatori, zadržani su zbog manjeg broja različitih klasa, što može pružiti korisne informacije.

Korelacije i selekcija atributa

Analizom korelacija među atributima, primetili smo vrlo snažne veze izmedju nekih atributa. Ove visoke korelacije sugerišu redundansu, pa smo odlučili da izbacimo kolone *SEC_DISBURSED_AMOUNT*,

PRI_DISBURSED_AMOUNT, i *SEC_CURRENT_BALANCE* kako bismo smanjili suvišnost u podacima. Na osnovu grafičkih prikaza odnosa između atributa i izlaznog atributa, teško je precizno odrediti koji atributi su najkorisniji za predikciju.



Stoga ćemo se odlučiti za selekciju atributa koristeći Principal Component Analysis (PCA) kako bismo identifikovali najvažnije karakteristike za modeliranje.

Normalizacija Atributa

U cilju poboljšanja efikasnosti, izvršili smo *one-hot enkodiranje* kategorijalnih atributa, dok smo numeričke atribute *normalizovali* na raspon od 0 do 1. Ove promene su omogućile dosledno poređenje svih atributa i ravnomernu zastupljenost tokom modeliranja, što je dovelo do boljih rezultata u klasifikaciji.

4. Treniranje više algoritama i interpretacija dobijenih rezultata

Validacija Rešenja i Evaluacija Modela

Za validaciju rešenja odlučili smo se za kombinaciju kros validacije i klasičnog deljenja podataka na trening i test skupove. Kros validaciju koristimo tokom optimizacije modela kako bismo osigurali da model generalizuje dobro na različitim podacima, izbegavajući pretreniranost. Ovaj pristup omogućava modelu da bude testiran na više podskupova podataka, što daje stabilniju procenu performansi. Međutim, za konačnu evaluaciju modela koristimo tradicionalni pristup sa jednim testnim skupom, jer je naš testni skup dovoljno velik, a kros validacija bi zahtevala previše vremena za izvođenje na celokupnom skupu podataka.

Za evaluaciju modela koristili smo kombinaciju više mera kako bismo dobili potpuniju sliku njegovih performansi. *Accuracy* nam pokazuje osnovnu tačnost, ali sam po sebi nije dovoljan, posebno kod neuravnoteženih podataka. *Precision* nam govori koliko tačno model predviđa pozitivne klase, dok *recall* pokazuje koliko dobro prepoznaje sve pozitivne primere. *F1 score* kombinuje ova dva aspekta, dajući uravnotežen uvid u performanse. *AUC (Area Under the Curve)* dodatno procenjuje sposobnost modela da razlikuje između klasa, nezavisno od praga. Ova kombinacija nam omogućava sveobuhvatnu i balansiranu procenu modela.

Balansiranje podataka

Kako bismo rešili problem neuravnoteženosti klasa, balansirali smo podatke tako da sada imamo jednak broj primera iz obe klase. Prethodno je skup podataka imao značajnu dominaciju jedne klase, što bi moglo negativno uticati na performanse modela. Nakon balansiranja, klasa 0 i klasa 1 imaju po 40,457 instanci. Od ovog novog skupa, 10% podataka ćemo koristiti za optimizaciju modela. U nastavku će biti prikazani rezultati algoritama na balansiranim i nebalansiranim podacima kako bismo uporedili njihove performanse.

KNN

Analiza performansi *KNN* modela pokazuje značajne promene pre i posle balansiranja podataka. Pre balansiranja, model je imao *tačnost* od 74.17%, ali su *preciznost* i *odziv* za manjinsku klasu bili niski (30.05% *preciznost* i 14.02% *odziv*), što znači da nije uspevao da identifikuje korisnike koji ne plaćaju kredit.

Nakon balansiranja, *tačnost* je opala na 53.24%, kao i *preciznost* 25.06%, dok je *odziv* porastao na 57.67%. Ovo pokazuje da model bolje prepoznaje korisnike sa većim rizikom neplaćanja kredita, iako ukupna tačnost opada. *F1 skor* se poboljšao sa 0.19 na 0.35, a *AUC* se blago povećao sa 0.52 na 0.55, što ukazuje na bolju ravnotežu između preciznosti i odziva za manjinsku klasu.

Random Forest

Analiza performansi *Random Forest* modela pre balansiranja podataka otkriva visoku *tačnost* od 77.68%, dok su *preciznost* (32.15%) i *odziv* za manjinsku klasu (2.24%) bili veoma niski. Ovo ukazuje da model uspešno identifikuje većinsku klasu, ali ne uspeva da prepozna korisnike sa rizikom od neplaćanja kredita.

Nakon balansiranja, *tačnost* modela opada na 56.99%, kao i *preciznost* 25.97%, dok je *odziv* za manjinsku klasu poboljšan na 52.67%. *F1 skor* je porastao sa 0.04 na 0.35, a *AUC* se povećao sa 0.50 na 0.55, što pokazuje poboljšanu

ravnotežu između *preciznosti* i *odziva* za manjinsku klasu. Ovi rezultati sugerišu da model sada bolje prepoznaje korisnike sa rizikom od neplaćanja kredita.

Gradient Boosting

Performanse *Gradient Boosting* modela pre balansiranja podataka pokazuju visoku *tačnost* od 78.22%, uz značajnu *preciznost* (49.43%) ali vrlo nizak *odziv* za manjinsku klasu (0.85%). Model uspeva da prepozna većinsku klasu, ali je njegova sposobnost da identifikuje korisnike sa rizikom od neplaćanja kredita veoma slaba.

Nakon balansiranja, *tačnost* modela opada na 55.01 kao i *preciznost* 27.69%, dok je *odziv* za manjinsku klasu poboljšan na 66.18%. *F1 skor* je povećan na 0.39, dok je *AUC* porastao na 0.59, što ukazuje na bolju ravnotežu između preciznosti i odziva za manjinsku klasu. Ovi rezultati pokazuju da model bolje prepoznaje korisnike sa rizikom neplaćanja kredita.

5. Optimizacija parametara i interpretacija dobijenih rezultata

KNN

Optimizacija parametara *KNN* modela sprovedena je korišćenjem *Grid Search-a* sa različitim vrednostima za broj najbližih suseda *k*. Pre optimizacije, model sa balansiranim podacima imao je *tačnost* od 53.24%, *preciznost* od 25.06%, *odziv* od 57.67%, *F1 skor* od 0.35 i *AUC* od 0.55. Nakon optimizacije, gde je najbolji broj suseda *k* bio 3, performanse su se blago promenile: *tačnost* je opala na 52.83%, *preciznost* na 24.59%, odziv na 56.42%, *F1 skor* na 0.34, i *AUC* na 0.54. Ove promene ukazuju da optimizacija nije dovela do značajnog poboljšanja modela.

Random Forest

Optimizacija parametara *Random Forest* modela sprovedena je korišćenjem *Grid Search-a*, gde su istraživane različite kombinacije vrednosti za max_depth, min_samples_leaf, i max_features. Pre optimizacije, model sa balansiranim podacima imao je tačnost od 56.99%, preciznost od 25.97%, odziv od 52.67%, F1 skor od 0.35 i AUC od 0.55.

Nakon optimizacije, gde su najbolji parametri bili *max_depth=8, max_features=15*, i *min_samples_leaf=1*, performanse modela su se značajno promenile. *Tačnost* je opala na 45.37%, dok su *preciznost* i *odziv* za manjinsku klasu promenjeni na 24.69% i 73.61%, respektivno. *F1 skor* je porastao na 0.37, a *AUC* je blago povećan na 0.56. Ove promene sugerišu da optimizacija nije dovela do poboljšanja ukupne tačnosti, ali je poboljšala sposobnost modela da prepozna korisnike sa visokim rizikom neplaćanja kredita.

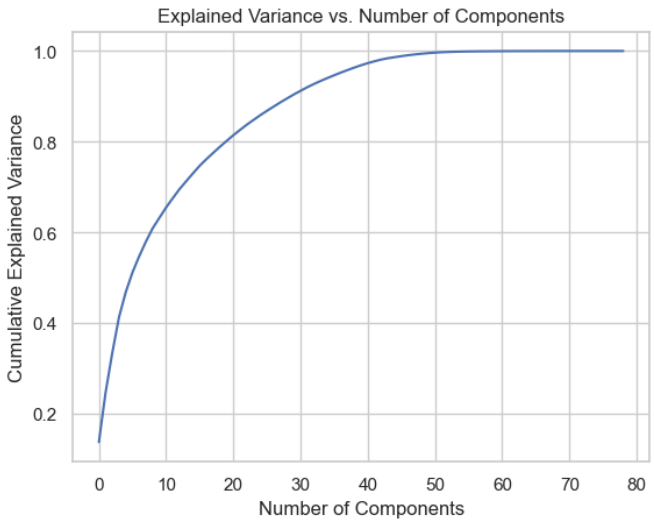
Gradient Boosting

Optimizacija parametara *Gradient Boosting* modela izvršena je korišćenjem *Grid Search-a* sa različitim vrednostima za *learning_rate* i *max_depth*. Pre optimizacije, model sa balansiranim podacima imao je *tačnost* od 55.01%, *preciznost* od 27.69%, *odziv* od 66.18%, *F1 skor* od 0.39, i *AUC* od 0.59.

Nakon optimizacije, gde su najbolji parametri bili *learning_rate=0.1* i *max_depth=3*, performanse modela su ostale nepromenjene u poređenju sa prethodnim rezultatima: *tačnost* je ostala na 55.01%, *preciznost* na 27.69%, *odziv* na 66.18%, *F1 skor* na 0.39, i *AUC* na 0.59. Ove rezultate sugerišu da optimizacija nije uticala na poboljšanje modela, već su performanse ostale iste.

6. Selekcija atributa i interpretacija dobijenih rezultata

Na osnovu grafičkih prikaza odnosa između atributa i izlaznog atributa, nismo mogli precizno utvrditi koji atributi su najvažniji za predikciju. Zbog toga smo primenili *Principal Component Analysis (PCA)* kako bismo smanjili broj atributa. *PCA* nam omogućava da redukujemo dimenzionalnost sa početnih 79 atributa na 25 glavnih komponenti, koje objašnjavaju 80% - 90% varijabilnosti podataka. Rezultati modela na skupu podataka sa smanjenim brojem atributa biće prikazani u nastavku.



KNN

Prvi rezultati modela na balansiranim podacima pokazali su *tačnost* od 53.24%, *preciznost* od 25.06%, *odziv* od 57.67%, i *F1 skor* od 34.94%. Ovi rezultati ukazuju na to da model prepoznaje manjinsku klasu (neplaćanje kredita), ali još uvek ima prostora za poboljšanje u preciznosti i sveukupnoj tačnosti.

Nakon primene *PCA*, gde smo smanjili broj atributa sa 79 na 25, rezultati su se blago promenili. *Tačnost* je opala na 52.65%, dok su *preciznost* i *F1 skor* takođe pokazali blagi pad na 23.25% i 31.95%, respektivno. Ova promena ukazuje na to da je redukcija dimenzionalnosti, iako korisna za smanjenje redundancije podataka, donela određene gubitke u informativnosti, što je uticalo na sposobnost modela da precizno prepozna ciljne klase.

U poređenju sa početnim rezultatima, *PCA* je smanjio performanse modela, što sugeriše da iako *PCA* pomaže u smanjenju složenosti podataka, može dovesti do gubitka korisnih informacija koje su bile prisutne u originalnim atributima. Iz tog razloga, i s obzirom na ograničene resurse, odlučili smo da obustavimo ovaj deo eksperimenta.

1. Zaključak i dalji rad

Na osnovu priloženih rezultata, može se konstruisati tabela (*tabela 1*) koja prikazuje performanse modela pre i posle balansiranja, kao i nakon optimizacije i primene *PCA*.

Analiza performansi modela *KNN*, *Random Forest* i *Gradient Boosting* pokazuje da su svi modeli poboljšali prepoznavanje korisnika sa rizikom od neplaćanja kredita kada su trenirani na balansiranom skupu podataka, uz poboljšanje *F1 statistike*. *Gradient Boosting* je postigao najbolje rezultate sa *F1 statistikom* od 0.390, što sugeriše da je ovaj model najefikasniji u prepoznavanju rizičnih korisnika. Ipak, očekivano bi bilo da bi kombinacija *Gradient Boosting*-a i *Random Forest*-a mogla dodatno unaprediti performanse modela, jer bi kombinovani pristup mogao iskoristiti prednosti oba algoritma i pružiti bolje rezultate u praktičnoj primeni.

Optimizacija parametara i *Principal Component Analysis (PCA)* nisu doneli značajna poboljšanja, što ukazuje da su trenutni parametri i dimenzionalnost podataka već blizu optimalnih za ove modele. U budućnosti, istraživanje dodatnih modela može doneti dodatne uvide i unaprediti rezultate, dok bi detaljna selekcija i inženjering atributa mogli doprineti boljoj efikasnosti modela. Iako je trenutni pristup koristan, dalja istraživanja i unapređenja su neophodna za postizanje boljih rezultata.

U praktičnoj primeni, ovi modeli mogu pružiti određenu vrednost u prepoznavanju rizičnih korisnika, ali nisu dovoljno robusni za samostalnu upotrebu u stvarnim situacijama. Iako najbolji model pokazuje da može prepoznati većinu korisnika sa rizikom, visoka stopa lažno pozitivnih i relativno niska preciznost mogu dovesti do problema u realnom svetu, gde je potrebno da se minimizira broj lažno pozitivnih odluka. Testiranje u stvarnim uslovima, koristeći konkretne podatke, ključno je za procenu efikasnosti modela i prilagodjavanje strategija prema specifičnim potrebama finansijskih institucija.

Model	Status	Tačnost	Preciznost	Odziv	F1 Skor	AUC
KNN	Pre Balansiranja	0.741	0.300	0.140	0.191	0.524
	Posle Balansiranja	0.532	0.250	0.576	0.349	0.548
	Posle Optimizacije	0.528	0.245	0.564	0.342	0.541
	Posle PCA	0.526	0.232	0.510	0.319	0.520
Random Forest	Pre Balansiranja	0.776	0.321	0.022	0.041	0.504
	Posle Balansiranja	0.569	0.259	0.526	0.347	0.554
	Posle Optimizacije	0.453	0.246	0.736	0.369	0.555
Gradient Boosting	Pre Balansiranja	0.782	0.494	0.008	0.016	0.503
	Posle Balansiranja	0.550	0.276	0.661	0.390	0.590
	Posle Optimizacije	0.550	0.276	0.661	0.390	0.590

Tabela 1 – rezultati modela