SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Otkrivanje prijevare s kreditnom karticom

Antonio Ilinović
Voditelj: *Goran Delač*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod

U današnje vrijeme, korištenje kartica kao oblika plaćanja postaje sve učestalije, dok se uporaba gotovine smanjuje. Stoga je iznimno važno prepoznati slučajeve zlouporabe kreditnih kartica. Ovaj rad fokusira se na detekciju transakcija kreditnim karticama koje predstavljaju zlouporabu. Rješavanje ovog problema od interesa je bankama i svima koji koriste kreditne kartice.

U prvom dijelu ovog rada izvršena je analiza podataka, prikazane su distribucije ciljne varijable 'Class' i ostalih varijabli. Provedeno je skaliranje varijabli koje prethodno nisu bile skalirane, kao što su 'Amount' i 'Time'. Implementirane su tehnike slučajnog poduzorkovanja i naduzorkovanja pomoću metode SMOTE [8]. Kako bi se bolje razumjela struktura podatkovnog skupa, prikazane su distribucije varijabli i matrice korelacije te njihova interpretacija u odnosu na originalni, neuravnoteženi skup podataka i poduzorkovane skupove podataka. Također su prikazane vizualizacije grupiranja podataka pomoću metoda PCA i t-SNE. Na kraju, trenirani su klasifikacijski modeli koji su evaluirani upotrebom ugniježđene unakrsne provjere, te su prikazani rezultati tih modela i kako oni ovise o tome jesu li modeli trenirani na originalnom, nebalansiranom skupu ili na poduzorkovanom ili naduzorkovanom skupu.

2. Analiza podatkovnog skupa

2.1. Skup podataka

Jedan od ključnih koraka pri primjeni algoritama strojnog učenja na zadani problem je upoznavanje s podacima. Skup podataka [12], [6], [5], [7], [4], [1], [2], [10], [3], [9], [11] sastoji se od transakcija kreditnim karticama Europskih građana u rujnu 2013. godine. Sve transakcije u skupu podataka napravljene su unutar dva dana, a od ukupno 284807 transakcija, njih 492 je lažnih. Imajući u vidu da pozitivna klasa, odnosno lažne transakcije, čine samo 0.172% svih transakcija, skup podataka je neuravnotežen, što možemo vidjeti na grafikonu 2.2. Stoga, potrebno je obraditi podatke kako bi algoritmi strojnog učenja pravilno funkcionirali na takvom skupu.

Skup podataka koji se koristi sastoji se od 30 brojčanih varijabli. Kako bi se zaštitila privatnost korisnika kreditnih kartica, originalne značajke nisu dostupne u skupu podataka. Umjesto toga, 28 značajki dobiveno je pomoću analize glavnih komponenti (PCA). Značajke dobivene PCA-om su označene kao V1, V2, V3, ..., V28, dok su dvije značajke koje nisu transformirane PCA-om 'Time' i 'Amount'. Značajka 'Time' predstavlja vrijeme proteklo između svake transakcije i prve transakcije u skupu podataka, dok značajka 'Amount' predstavlja iznos koji je terećen s kartice. Konačno, binarna ciljna značajka 'Class' označava transakcije koje su prijevare (1) i transakcije koje nisu prijevare (0).

Važno je napomenuti da u skupu podataka nema nedostajućih podataka, tako da se ne treba baviti tehnikama nadopunjavanja takvih podataka.

Najveći problem ovog podatkovnog skupa je nebalansiranost. U takvim slučajevima, ciljna varijabla ima puno više uzoraka jedne klase, zbog čega algoritmi strojnog učenja mogu biti pristrani prema većinskoj klasi, te imati problema s detekcijom manjinske klase. Kako bismo riješili ovaj problem, primijenit ćemo metode poduzorkovanja i naduzorkovanja te usporediti točnost modela pri korištenju ovih metoda. Primjenom ovih metoda, dobivamo balansiraniji skup podataka za koji očekujemo da će biti reprezentativniji i pridonijeti povećanju performansi algoritama strojnog učenja.

2.1.1. Skaliranje značajki

Značajke ćemo skalirati kako bi imale sličnu skalu podataka. To može pomoći performansama algoritama strojnog učenja, kao i smanjiti utjecaj stršećih vrijednosti. Značajke V1-V28 su unaprijed skalirane, dok ćemo značajke 'Amount' i 'Time' skalirati. Za skaliranje koristimo RobustScaler transformaciju iz paketa Sklearn. Nove značajke su naziva 'Amount_scaled' i 'Time_scaled'. RobustScaler skalira podatke koristeći interkvartalni raspon IQR koji se dobiva kao razlika 75-tog i 25-tog percentila podataka. Medijan se oduzima od svih podataka i rezultat se dijeli sa izračunatim interkvartalnim rasponom. Podaci će imati medijan 0 i raspršenje slično kao originalni podaci. Ova metoda je robusna na stršeće vrijednosti.

2.2. Uzorkovanje

U radu provodimo testiranje utjecaja uzorkovanja na točnost modela. Za tu svrhu pripremljen je skup podataka koji se sastoji od jednakog broja transakcija koje su prijevare i transakcija koje nisu prijevare. Naš originalni skup podataka je nebalansiran, te ćemo napraviti i naduzorkovani i poduzorkovani skup koji su balansirani. Jedan od razloga za balansiranje je mogućnost prenaučenosti modela na nebalansiranom skupu, što može dovesti do pretpostavke da većina primjera ne predstavlja prijevaru. Drugi razlog je nemogućnost preciznog utvrđivanja korelacija između ciljne značajke i ostalih značajki na nebalansiranom skupu podataka. Iako nam nije poznato što pojedine značajke V1-V28 točno predstavljaju, važno je vidjeti kako te značajke utječu na ciljnu značajku.

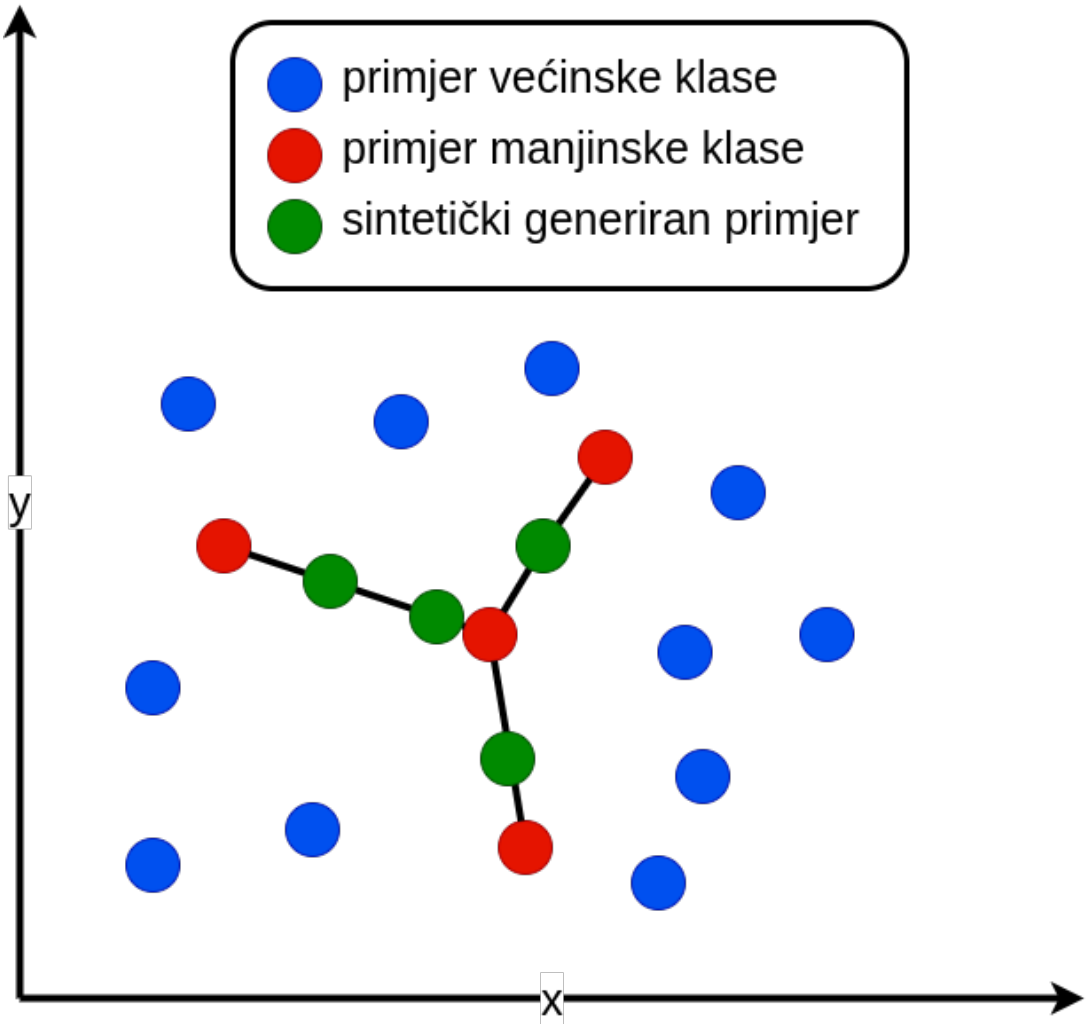
2.2.1. Slučajno poduzorkovanje

Kao metoda balansiranja podataka korištena je metoda slučajnog poduzorkovanja. Ova metoda uključuje slučajni odabir primjera iz većinske klase kako bi se stvorio poduzorak jednak veličini manjinske klase, čime se postiže balansiranje podatkovnog skupa. Međutim, mogući problem kod ovog pristupa je da izbacivanje primjera većinske klase može rezultirati gubitkom bitnih značajki te klase, što može utjecati na sposobnost klasifikatora da nauči i donese ispravne odluke za tu klasu [8].

2.2.2. SMOTE naduzorkovanje

Kao metoda naduzorkovanja koristi se metoda SMOTE (engl. Synthetic Minority Oversampling Technique). SMOTE je tehnika naduzorkovanja, prikazana na slici 2.1, koja umjetno generira dodatne primjere manjinske klase kako bi se postigao balans između manjinske i većinske klase. Primjeri se generiraju na sljedeći način:

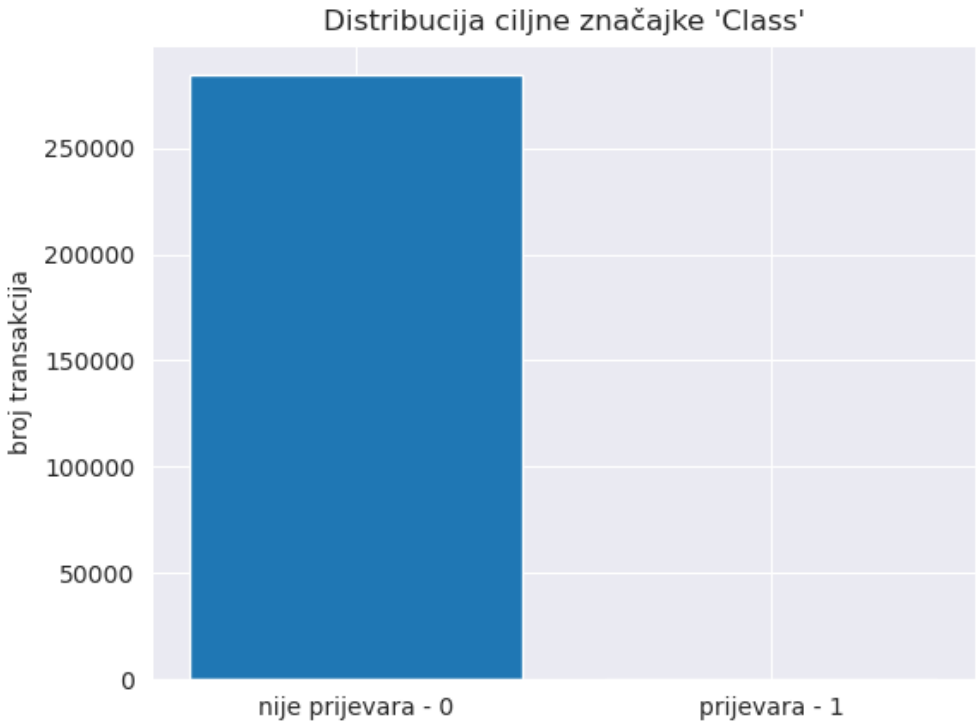
1. Odabire se slučajan primjerak manjinske klase.
2. Pronalaze se njegovih k najbližih susjeda unutar manjinske klase.
3. Na linijama koje povezuju odabrani primjerak s njegovim k najbližih susjeda, slučajno se odabiru točke koje će postati novi umjetno generirani primjerci manjinske klase.



Slika 2.1: Ilustracija procesa generiranja sintetičkih primjera SMOTE metodom.

Iako SMOTE ima mnogo obećavajućih prednosti, također ima svoje nedostatke [8]. Naime, ovakvo naduzorkovanje može dovesti do prenaučenosti modela, jer će model imati mnogo sličnih primjera koji su generirani na temelju postojećih primjera manjinske klase. Ova sličnost među primjerima može rezultirati manjom sposobnošću modela da se generalizira na nove, neviđene primjere.

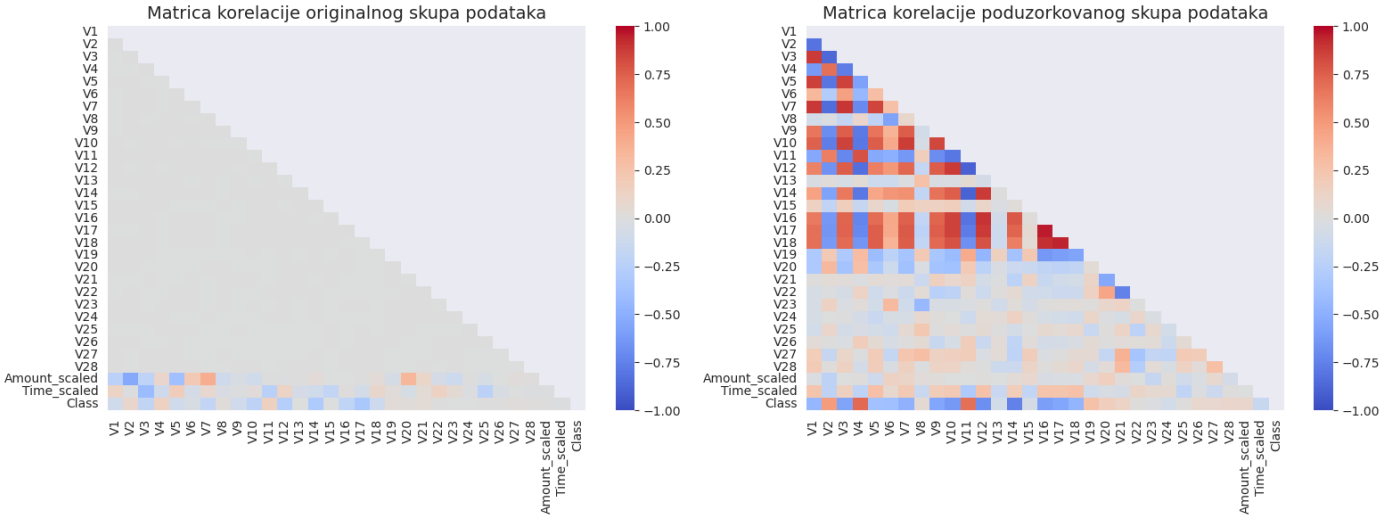
2.3. Distribucija značajki



Slika 2.2: Distribucija ciljne značajke 'Class'. Samo mali broj transakcija su prijevare, odnosno vrijednost značajke 1.

Korištenjem korelacijske matrice možemo prikazati linearnu povezanost između parova varijabli u skupu podataka. Na slici 2.3 možemo vidjeti dvije korelacijske matrice. Lijeva matrica prikazuje korelacije originalnog skupa podataka, dok desna matrica prikazuje korelacije poduzorkovanog skupa podataka koji sadrži jednak broj primjera prijevara i primjera koji nisu prijevare. U originalnom skupu, vrijednosti korelacije su većinom blizu nule, što je rezultat neravnoteže između klasa primjera. Korelacijska matrica dobivena iz poduzorkovanog skupa podataka pruža jasniji uvid u

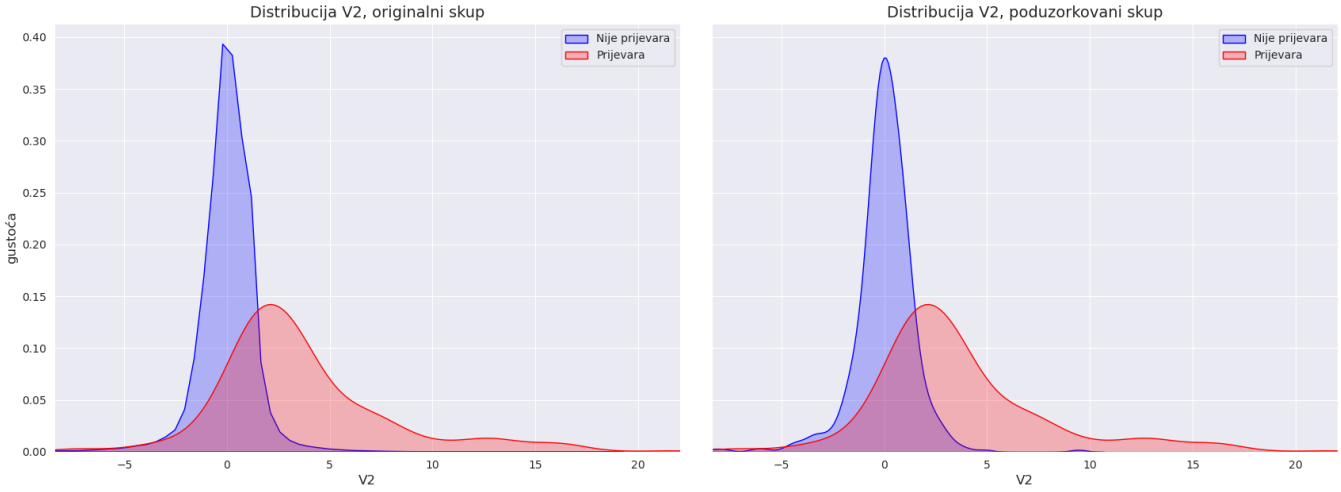
međuovisnost varijabli. Najzanimljivije su nam vrijednosti korelacije između ciljne varijable 'Class' i ostalih varijabli, jer one pokazuju koliko ciljna varijabla 'Class' ovisi o pojedinim varijablama. Varijable V10, V12 i V14 imaju najvišu negativnu korelaciju, dok varijable V2, V4 i V11 imaju najvišu pozitivnu korelaciju s ciljnom varijablom 'Class'. Što su niže vrijednosti varijabli V10, V12, V14 i što su više vrijednosti varijabli V2, V4 i V11, veća je vjerojatnost da će primjer biti lažna transakcija.



Slika 2.3: Matrica korelacije originalnog i poduzorkovanog skupa podataka. U matrici originalnog skupa, zbog neravnoteže u klasama, međuovisnosti varijabli nisu jasno vidljive. Međutim, u matrici poduzorkovanog skupa jasnije su vidljive međuovisnosti među varijablama.

Grafovi 2.4 i 2.5 prikazuju distribucije varijabli V2 i V10 u originalnom i poduzorkovanom skupu podataka. Tijekom poduzorkovanja, ne uklanjamo primjere koji predstavljaju prijevaru, pa distribucija pozitivne klase ostaje jednaka u oba skupa. Međutim, uklanjamo primjere negativne klase. Distribucija primjera pozitivne klase nije znatno različita u oba slučaja. Utvrdili smo da varijabla V2 ima najveću pozitivnu, a V10 najveću negativnu korelaciju s ciljnom varijablom. To se može vidjeti na grafikonima distribucija. Kod varijable V2, središte distribucije pozitivnih primjera smješteno je desno od središta distribucije negativnih primjera, što ukazuje na pozitivnu korelaciju. S druge strane, kod varijable V10 situacija je obrnuta. Središte distribucije pozitivnih primjera nalazi se lijevo od središta distribucije negativnih primjera, što odgovara negativnoj korelaciji.

Na slici 2.6 prikazana je distribucija značajke 'Time_scaled' u originalnom, poduzorkovanom i naduzorkovanom skupu. Primjećujemo značajne razlike u distribucijama između poduzorkovanja i naduzorkovanja. Posebno je zanimljiva distribucija pozitivnih primjera u naduzorkovanom skupu, što je rezultat sintetičkog dodavanja

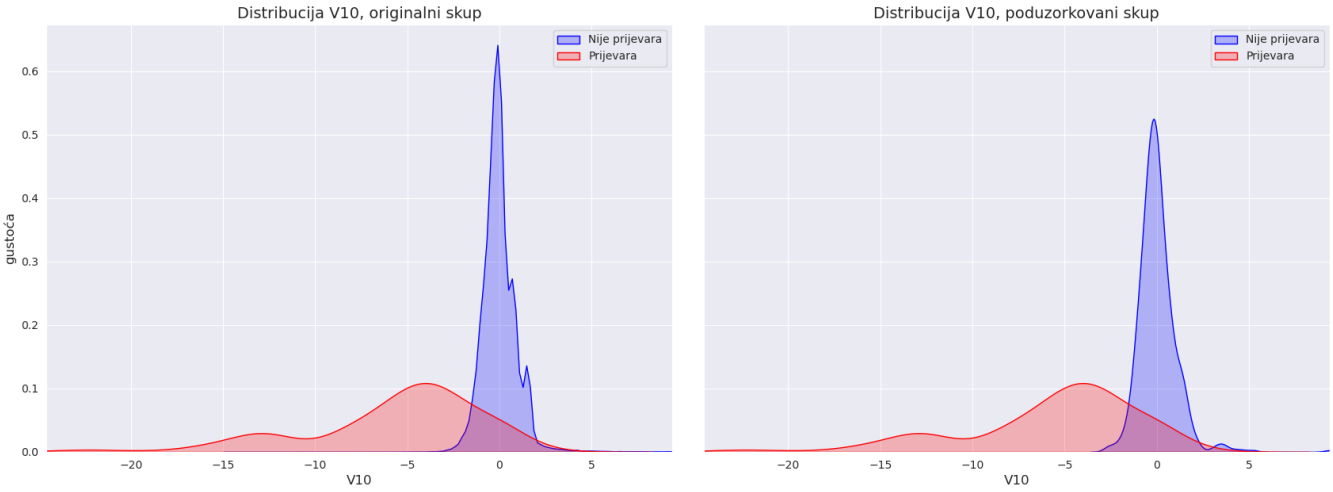


Slika 2.4: Distribucija varijable 'V2' u originalnom i poduzorkovanom skupu. Distribucije pozitivne klase su slične, a pozitivna korelacija je vidljiva jer je središte pozitivnih primjera pomaknuto desno u odnosu na središte negativnih primjera.

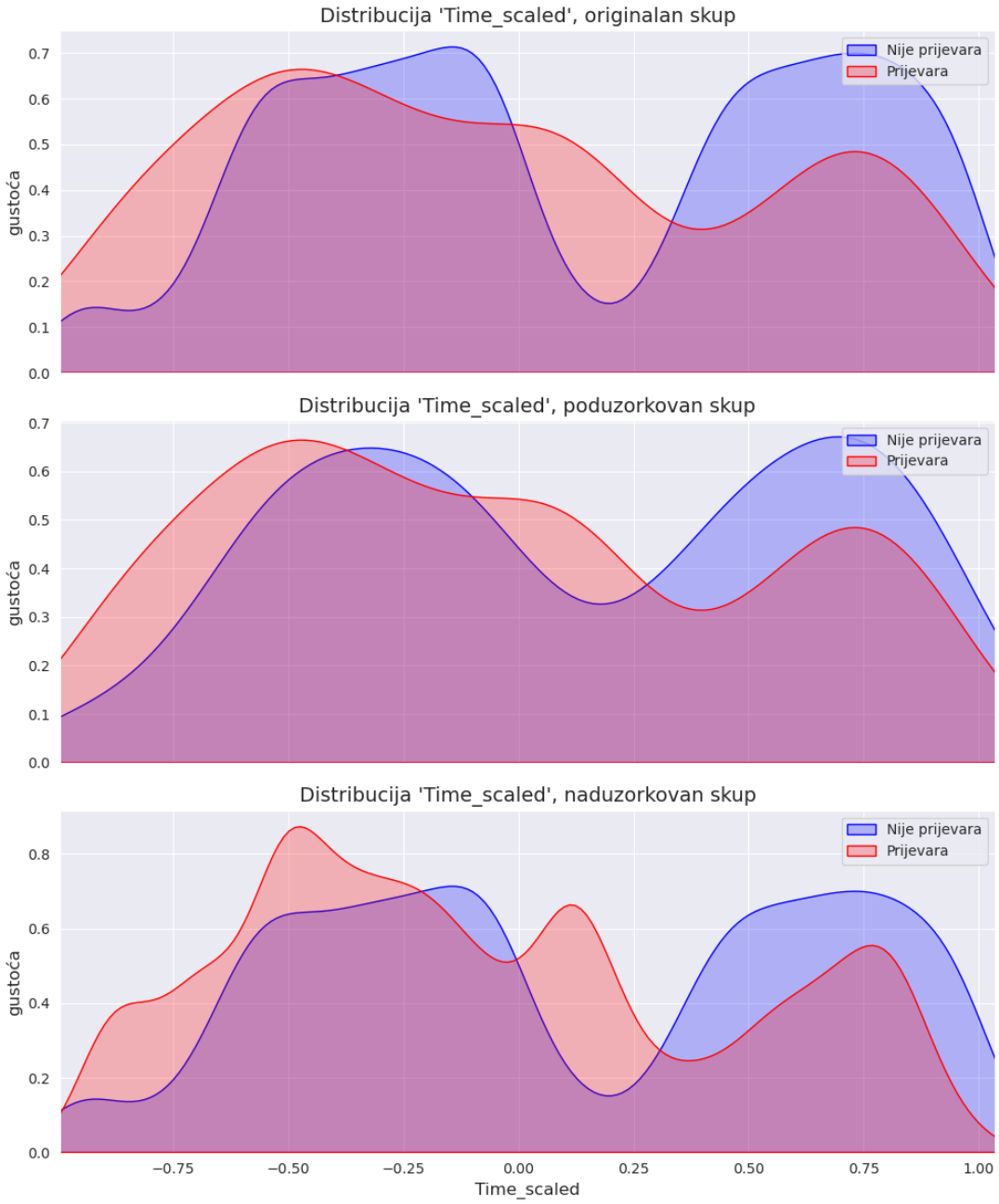
manjinskih primjera.

2.4. Redukcija dimenzionalnosti i grupiranje

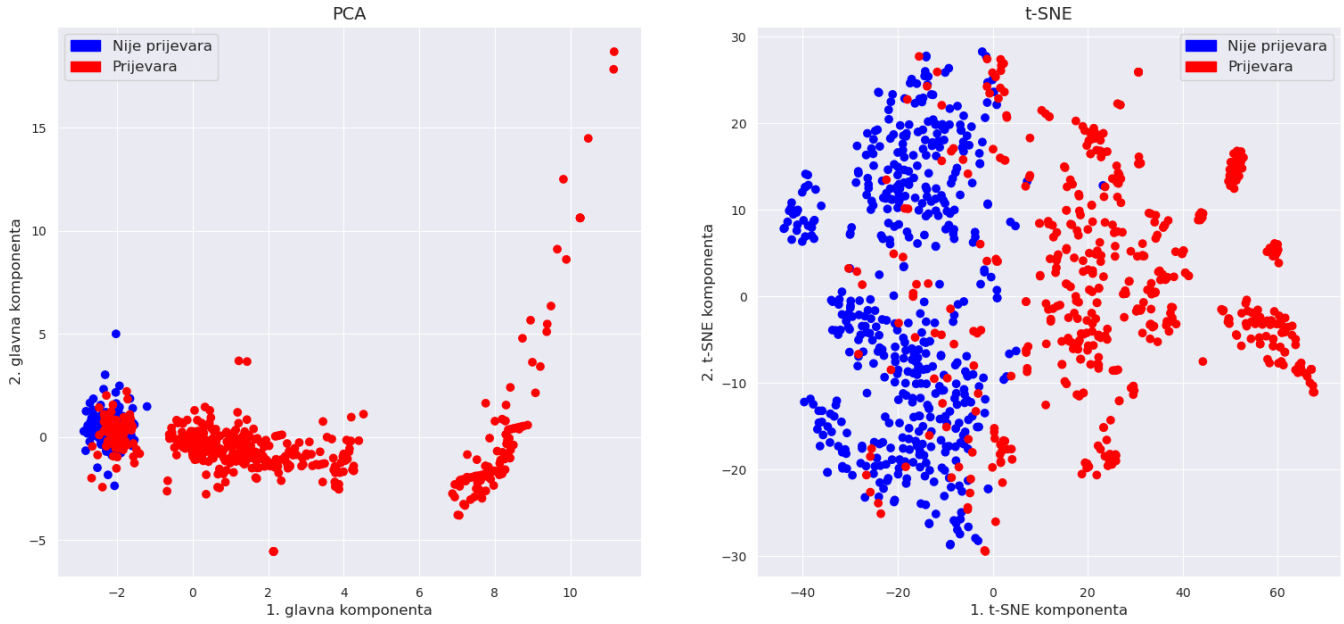
Za vizualizaciju transakcija koje predstavljaju prijevare i onih koje to nisu, korištene su analiza glavnih komponenti (PCA) i algoritam t-SNE (t-distributed stochastic neighbor embedding) [13]. Broj komponenti postavljen je na 2 kako bi se podaci mogli prikazati pomoću grafa raspršenja. PCA je algoritam koji traži komponente tako da prva komponenta redom objašnjava što je moguće više varijance, dok druga komponenta objašnjava što više preostale varijance. T-SNE je nelinearni algoritam za smanjenje dimenzionalnosti i vizualizaciju podataka visokih dimenzija, koji je posebno koristan za prikaz kompleksnih podataka koji se ne mogu jednostavno prikazati pomoću linearnih metoda kao što je PCA. Na grafu 2.7 prikazane su grupacije pomoću algoritama PCA i t-SNE. U obje vizualizacije možemo vidjeti grupe transakcija koje su prijevare i onih koje nisu. Međutim, t-SNE se pokazuje boljim jer bolje razdvaja podatke, dok kod PCA vizualizacije, u grupi transakcija koje nisu prijevare, možemo pronaći i primjere koji su prijevare.



Slika 2.5: Distribucija varijable 'V10' u originalnom i poduzorkovanom skupu. Distribucije pozitivne klase su slične, a negativna korelacija je vidljiva jer je središte pozitivnih primjera pomaknuto lijevo u odnosu na središte negativnih primjera.



Slika 2.6: Na grafu je prikazana usporedba distribucija značajke 'Time_scaled' u originalnom, poduzorkovanom i naduzorkovanom skupu. U poduzorkovanom skupu su izbačene transakcije koje nisu prijevare, što je rezultiralo blagom promjenom distribucije negativne klase. Zanimljiva je i distribucija prijevara u naduzorkovanom skupu, gdje su sintetički dodani pozitivni primjeri, što se može primijetiti po promjeni distribucije pozitivne klase.



Slika 2.7: Na grafu su prikazane dvije metode grupiranja podataka: PCA i t-SNE. Na lijevom grafu prikazano je grupiranje poduzorkovanih podataka korištenjem PCA metode, dok je na desnom grafu prikazano grupiranje dobiveno primjenom t-SNE metode. Primjećuje se da t-SNE metoda bolje razdvaja podatke, dok je na PCA grafu vidljiva grupa koju čine transakcije koje su prijevare i one koje nisu.

3. Primjena algoritama strojnog učenja

3.1. Vrednovanje modela

U formulama za točnost, preciznost i odziv koristimo sljedeće oznake, gdje pozitivan primjer označava transakciju koja je prijevara, a negativan transakciju koja nije prijevara:

- TP (engl. True Positive) predstavlja broj ispravno klasificiranih pozitivnih primjera.
- TN (engl. True Negative) predstavlja broj ispravno klasificiranih negativnih primjera.
- FP (engl. False Positive) predstavlja broj pogrešno klasificiranih negativnih primjera.
- FN (engl. False Negative) predstavlja broj pogrešno klasificiranih pozitivnih primjera.

$$\text{Točnost} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.1}$$

Točnost (engl. accuracy), koja se definira formulom 3.1, predstavlja metriku koja mjeri omjer točno klasificiranih primjera u odnosu na ukupan broj primjera [14]. Međutim, kod nebalansiranih skupova podataka, točnost nije prikladna mjera za evaluaciju modela [8]. Uzmimo za primjer naš podatkovni skup, koji sadrži samo 0,17% pozitivnih primjera. Ako kreiramo model koji će sve primjere klasificirati kao negativne, točnost takvog modela iznosila bi 99,83%, što predstavlja preoptimističan rezultat za tako naivan model.

Stoga je za vrednovanje modela korištena F1 'Makro' varijanta, definirana formulom 3.4. Oznaka P označava preciznost (engl. precision), koja je definirana formulom

3.2, dok oznaka R predstavlja odziv (engl. recall), definiran formulom 3.3. Makro varijanta podrazumijeva zasebno izračunavanje preciznosti i odziva za pozitivne i negativne primjere, te se računa njihov prosjek. Korištenjem takve makro varijante osigurava se jednaka težina za svaku klasu prilikom evaluacije modela, čime se omogućuje "pravo glasa" manjinskoj klasi. Na taj način izbjegavamo probleme koji se javljaju pri korištenju metrike točnosti.

$$\text{Preciznost} = P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

$$\text{Odziv} = R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

$$F1 = 2 \cdot \frac{\text{P} \cdot \text{R}}{\text{P} + \text{R}} \tag{3.4}$$

3.2. Optimiziranje hiperparametara

Za evaluaciju i odabir optimalnih hiperparametara modela, korištena je ugniježđena stratificirana unakrsna validacija [14]. Ugniježđena stratificirana unakrsna validacija se sastoji od vanjske i unutarnje petlje. U vanjskoj petlji podatkovni skup se podijeli na pet disjunktnih skupova. Važno je naglasiti da su skupovi stratificirani, što znači da omjer primjera iz manjinske i većinske klase ostaje konzistentan u svakom skupu. Jedan od pet skupova koristi se za konačnu provjeru modela, dok ostala četiri skupa ulaze u unutarnju petlju, gdje se vrši treniranje i validacija modela s različitim kombinacijama hiperparametara. U unutarnjoj petlji, skup podataka se dalje dijeli na tri dijela, od kojih se jedan koristi kao validacijski set za određivanje optimalnih parametara. Preostala dva dijela koriste se kao skup za treniranje. Ovaj skup za treniranje može se naduzorkovati, poduzorkovati ili ostaviti nepromijenjen, ovisno o odabranoj strategiji.

Prednost ovakve evaluacije je što pruža preciznu procjenu konačne pogreške modela. Međutim, nedostatak je što ovakav pristup zahtijeva značajan broj treninga modela što može biti vremenski zahtjevno.

3.3. Eksperimentalni rezultati

U svrhu predikcije korištena su tri modela: logistička regresija (engl. logistic regression), model slučajne šume (engl. random forest classifier) i stroj potpornih vektora

(engl. support vector machine - SVM). Rezultati, izraženi prosječnom F1 mjerom, prikazani su u tablici 3.1.

Model	**Poduzorkovan**	**Originalan skup**	**Naduzorkovan**
Logistička Regresija	54.1%	**89.1%**	55.0%
Slučajna Šuma	59.1%	**90.0%**	76.1%
SVM	56.2%	**81.1%**	58.5%

Tablica 3.1: Prosječna F1 mjera za tri modela na različitim skupovima podataka.

Ovaj rad je pokazao da se sva tri ispitivana modela, logistička regresija, slučajna šuma i SVM najbolje ponašaju na originalnom, nebalansiranom skupu podataka. Najbolje rezultate postiže model slučajne šume, neovisno o skupu podataka na kojem se trenira. Uspoređujući rezultate dobivene na poduzorkovanom i naduzorkovanom skupu podataka, uočavamo da su modeli postigli bolje rezultate na naduzorkovanom skupu.

Iako se uzorkovanje često koristi kao metoda za poboljšanje rezultata modela treniranih na nebalansiranim skupovima podataka [8], u ovom radu nije došlo do očekivanog poboljšanja. Mogući razlog za to mogu biti velika nebalansiranost skupa podataka i gubitak podataka većinske klase pri poduzorkovanju, kao i dodavanje velikog broja sličnih podataka manjinske klase pri naduzorkovanju.

Zaključno, za daljnja istraživanja predlaže se eksperimentiranje s različitim metodama uzorkovanja koje ne uključuju savršeni balans između primjera većinske i manjinske klase. Također, moglo bi se istražiti kako kombinacija metoda uzorkovanja, konkretno istovremeno poduzorkovanje većinske klase i naduzorkovanje manjinske klase, utječe na performanse modela i može li se na taj način postići bolji rezultat od onoga dobivenog treniranjem na originalnom skupu podataka.

4. Zaključak

U ovom radu obrađen je problem nebalansiranih podatkovnih skupova, s naglaskom na metode uzorkovanja koje mogu poboljšati učenje klasifikacijskih modela na takvim skupovima. Izvedeni eksperimenti pokazali su kako su se testirani modeli najbolje ponašali na originalnom, neizmijenjenom skupu. Ovakav rezultat mogao bi biti posljedica izrazite nebalansiranosti skupa podataka gdje poduzorkovanje eliminira prevelik broj primjera iz većinske klase, dok naduzorkovanje stvara prevelik broj sličnih primjera unutar manjinske klase.

Daljnji rad mogao bi uključivati eksperimentiranje s različitim omjerima uzorkovanja, umjesto da se uvijek teži savršenoj ravnoteži. Također, mogla bi se istražiti mogućnost kombinacije poduzorkovanja i naduzorkovanja, gdje bi se smanjivao broj primjera većinske klase, a povećavao broj primjera manjinske klase. Dodatno, istraživanje bi se moglo proširiti na veći broj klasifikacijskih modela.

Metode uzorkovanja su korisne tehnike koje se mogu primijeniti na različite nebalansirane skupove podataka. One su posebno korisne jer većina podataka pokazuje određeni stupanj nebalansiranosti.

5. Literatura

[1] Fabrizio Carcillo, Andrea Dal Pozzolo, Yann-Aël Le Borgne, Olivier Caelen, Yannis Mazzer, i Gianluca Bontempi. Scarff: a scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41:182–194, 2018.

[2] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, i Gianluca Bontempi. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. *International Journal of Data Science and Analytics*, 5 (4):285–300, 2018.

[3] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, i Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection information sciences. *Information Sciences*, 2019.

[4] Andrea Dal Pozzolo. *Adaptive machine learning for credit card fraud detection*. Doktorska disertacija, Université Libre de Bruxelles (ULB), 2015.

[5] Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, i Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, 41(10):4915–4928, 2014.

[6] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, i Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. U *Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 2015.

[7] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, i Gianluca Bontempi. Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*, 29(8):3784–3797, 2018.

[8] Haibo He i Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[9] Yann-Aël Le Borgne i Gianluca Bontempi. Reproducible machine learning for credit card fraud detection - practical handbook. *SpringerBriefs in Computer Science*, 2021.

[10] Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, i Gianluca Bontempi. Deep-learning domain adaptation techniques for credit cards fraud detection. U *INNSBDDL 2019: Recent Advances in Big Data and Deep Learning*, stranice 78–88, 2019.

[11] Bertrand Lebichot, Gianmarco Paldino, Wissam Siblini, Liyun He, Frederic Oblé, i Gianluca Bontempi. Incremental learning strategies for credit cards fraud detection. *International Journal of Data Science and Analytics*, 2021.

[12] Machine Learning Group - ULB. Kaggle Competition: Credit Card Fraud Detection. `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`.

[13] Laurens van der Maaten i Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[14] Jan Šnajder i Bojana Dalbelo Bašić. *Strojno učenje*. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2014.

Otkrivanje prijevare s kreditnom karticom

Sažetak

U današnjem svijetu, upotreba kartica kao način plaćanja postaje sve učestalija. Zbog toga je od velike važnosti prepoznavanje zlouporabe kreditnih kartica. Ovaj rad se fokusira na detekciju transakcija koje nisu izvršene od strane korisnika, već predstavljaju prijevare.

U okviru rada izvršena je analiza podatkovnog skupa, prikazane su distribucije podataka. Primijenjene su tehnike poduzorkovanja i naduzorkovanja. Podaci su interpretirani pomoću matrice korelacije i algoritama grupiranja. Na kraju, trenirani su klasifikacijski modeli te su prikazani rezultati tih modela nad nebalansiranim skupom i balansiranim uzorkovanim skupovima.

Ključne riječi: kreditne kartice; lažne transakcije; nebalansiran skup podataka; uzorkovanje podataka