Otkrivanje prijevara kreditnim karticama korištenjem varijacijskog autoenkodera

Grubelić, Damjan

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 539

OTKRIVANJE PRIJEVARA KREDITNIM KARTICAMA KORIŠTENJEM VARIJACIJSKOG AUTOENKODERA

Damjan Grubelić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 539

OTKRIVANJE PRIJEVARA KREDITNIM KARTICAMA KORIŠTENJEM VARIJACIJSKOG AUTOENKODERA

Damjan Grubelić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 539

Pristupnik:	**Damjan Grubelić (0036514575)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	prof. dr. sc. Zvonko Kostanjčar

Zadatak: **Otkrivanje prijevara kreditnim karticama korištenjem varijacijskog autoenkodera**

Opis zadatka:

Otkrivanje prijevara kreditnim karticama ključno je za financijsku sigurnost. Ovo područje proučava algoritme čiji je zadatak odrediti radi li se o legitimnoj transakciji ili pokušaju prijevare. Varijacijski autoenkoder, kao predstavnik probabilističkih modela dubokog učenja, opisuje distribuciju nekog skupa podataka. Jedna od primjena tog modela je otkrivanje anomalija, gdje se anomalijama smatraju primjeri podataka koji po svojim značajkama bitno odudaraju od ostalih primjera. Cilj ovog diplomskog rada je iskoristiti varijacijski autoenkoder kako bi se naučila distribucija transakcijskih skupova podataka, te na temelju nje prepoznale prijevare kao anomalije među regularnim transakcijama. Model treba naučiti i ispitati nad javno dostupnim skupovima transakcijskih podataka.

Rok za predaju rada: 28. lipnja 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 539

OTKRIVANJE PRIJEVARA KREDITNIM KARTICAMA KORIŠTENJEM VARIJACIJSKOG AUTOENKODERA

Damjan Grubelić

Zagreb, kolovoz, 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 539

Pristupnik:	**Damjan Grubelić (0036514575)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	prof. dr. sc. Zvonko Kostanjčar

Zadatak: **Otkrivanje prijevara kreditnim karticama korištenjem varijacijskog autoenkodera**

Opis zadatka:

Otkrivanje prijevara kreditnim karticama ključno je za financijsku sigurnost. Ovo područje proučava algoritme čiji je zadatak odrediti radi li se o legitimnoj transakciji ili pokušaju prijevare. Varijacijski autoenkoder, kao predstavnik probabilističkih modela dubokog učenja, opisuje distribuciju nekog skupa podataka. Jedna od primjena tog modela je otkrivanje anomalija, gdje se anomalijama smatraju primjeri podataka koji po svojim značajkama bitno odudaraju od ostalih primjera. Cilj ovog diplomskog rada je iskoristiti varijacijski autoenkoder kako bi se naučila distribucija transakcijskih skupova podataka, te na temelju nje prepoznale prijevare kao anomalije među regularnim transakcijama. Model treba naučiti i ispitati nad javno dostupnim skupovima transakcijskih podataka.

Rok za predaju rada: 28. lipnja 2024.

Sadržaj

1. Uvod

1.1. Opis problema

Plaćanje kreditnom karticom oblik je elektroničkog prijenosa novca pri čemu kartica služi kao sučelje prema računu plaćatelja. Cilj je prepoznati kada je plaćanje kreditnom karticom učinjeno bez dozvole vlasnika računa, tj. otkrivanje prijevara kreditnim karticama (OPKK).

1.2. Izazovi

Svojstveni izazovi za ovaj problem su (Mekterović et al., 2018; Cherif et al., 2023; Hilal et al., 2022):

1.2.1. Nedostatak stvarnih podataka

Značajke transakcije su povjerljivi podatci vlasnika kartice. Javna objava predstavlja rizik za privatnost tih podataka zbog čega ne postoji puno javno dostupnih skupova „stvarnih" podataka, odnosno onih nastalih korištenjem kreditnih kartica stvarnih kupaca. Postoje skupovi umjetno generiranih podataka, ali rijetko se koriste (Marazqah Btoush et al., 2023).

1.2.2. Obrada značajki

Postoji suglasje da podatci o transakciji samoj nisu dovoljni za model koji bi otkrivao prijevare. Zato se ti podatci trebaju obogaćivati dodatnim značajkama, uglavnom onima koje opisuju navike kupca. Takve značajke općenito ne budu objavljene u literaturi, a i javno dostupni skupovi podataka se objavljuju bez objašnjenja o značenju dodatnih

značajki, što otežava usporedbu metoda, reprodukciju rezultata te prijenos i nadogradnju znanja o najboljim metodama.

1.2.3. Skalabilnost

Potrebno je obraditi velik broj transakcija u sekundi, pogotovo ako je cilj detektirati je li transakcija prijevara prije nego se obradi ili bar prije nego se zareda veći broj prijevara.

1.2.4. Nesrazmjer ciljnih klasa

U ukupnom broju transakcija udio prijevara je malen. Zato standardni klasifikacijski algoritmi strojnog učenja (na primjer logistička regresija, stroj potpornih vektora (SVM) i stabla odluke (Haixiang et al., 2017)) te mjere za evaluaciju tih algoritama često nisu optimalni.

1.2.5. Vrednovanje metoda

Ovaj izazov djelomično proizlazi iz prethodnog: Mjere za vrednovanje rješenja trebaju uzeti u obzir neravnomjernost klasa. Na primjer, za skup podataka s 1.5% prijevara mjera „98.5% transakcija ispravno je klasificirano“ nije naročito korisna jer može značiti da model nikada ne otkriva prijevaru, kao i da otkriva sve prijevare, samo i 1.5% legitimnih transakcija pogrešno označi kao prijevare.

Osim toga, lažno pozitivna predviđanja općenito nisu jednako štetna kao lažno negativna, a ni sva lažno negativna međusobno nisu jednako štetna: neke prijevare nanose više štete. U nekim su slučajevima resursi za obradu transakcija označenih kao prijevare ograničeni, pa je i to potrebno uzeti u obzir prilikom vrednovanja metoda.

1.2.6. Pomak koncepta

Pomak koncepta najčešća je podvrsta pomaka u podatcima kod OPKK. Događa se kad dođe do promjene u distribuciji oznaka uvjetovanoj transakcijama, čak i kad distribucija podataka ostaje ista (Lucas et al., 14.10.2020.). Ovakav problem izražen je jer prevaranti prilagođavaju metode prijevare te jer se i navike potrošača mijenjaju s vremenom.

1.2.7. Preklapanje podataka

Kada su podatci dviju različitih ciljnih klasa jako slični, klasifikatoru je teže razlikovati ih. Kod prijevara kreditnim karticama taj je problem izražen jer prevaranti na taj način aktivno nastoje postići cilj suprotan cilju klasifikatora.

1.3. Dosadašnja istraživanja

Glavna područja istraživanja kod otkrivanja prijevara krednitnim karticama mogu se svrstati u sljedeće tipove zadataka strojnog učenja (Goodfellow et al., 2016):

1.3.1. Otkrivanje anomalija

Otkrivanje anomalija tip je zadatka u kojem je cilj u skupu transakcija pronaći one neobične. Algoritmi karakteristični za otkrivanje anomalija trebaju biti osmišljeni su s nesrazmjerom ciljnih klasa i heterogenošću anomalija na umu (Pang et al., 2021). Prednost takvog pristupa je lakše otkrivanje novih, još neviđenih načina prijevare. Klasični primjeri takvih algoritama su jednoklasni stroj potpornih vektora (OCSVM), izolacijska šuma (IF) (Ounacer et al., 2018) i samoorganizirajuća mapa (Zaslavsky i Strizhak, 2006). Primjeri algoritama *dubokog učenja* korištenih za otkrivanje anomalija u sklopu OPKK su autoenkoder (AE) i ograničeni Boltzmannov stroj (Kazemi i Zarrabi, 2017; Pumsirirat i Yan, 2018).

1.3.2. Klasifikacija

Kod ovog tipa zadatka cilj je primjerima pridružiti ispravnu kategoriju. Za razliku od otkrivanja anomalija, kod binarne je klasifikacije cilj podatku koji odudara od do tad viđenih pridružiti onu kategoriju čiji su elementi sličniji tom podatku.

Od klasičnih klasifikatora najčešći su naivni Bayesov klasifikator, logistička regresija, algoritam k najbližih susjeda, stabla odluke te strojevi potpornih vektora (Mekterović et al., 2018). Ansambli više modela pokazali su se korisnima u savladavanju neravnomjernosti klasa, kao i klasični algoritmi posebno prilagođeni za taj problem, najčešće uzimanjem podzastupljenosti ciljne klase u obzir tijekom treniranja (Sahin et al., 2013; Correa Bahnsen et al., 2015). Slučajne šume primjer su često korištenih ansambala kod

OPKK (Mekterović et al., 2021). Kod dubokih modela dobre rezultate daju inačice rekurzivnih neuronskih mreža (Jurgovsky et al., 2018; Wiese i Omlin, 2009) i konvolucijske mreže (Fu et al., 2016; Heryadi i Warnars, 2017; Zhang et al., 2018).

1.3.3. Sinteza i uzorkovanje

Istraživanja u ovom području uglavnom se koriste kako bi se olakšao zadatak klasifikacije pronalaskom najbolje metode za poduzorkovanje ili stvaranje novih transakcija radi otklanjanja problema nedostatka podataka ili nesrazmjera ciljnih klasa. Tradicionalno je najkorišteniji pristup poduzorkovanje (Bhattacharyya et al., 2011), ali koristi se i naduzorkovanje (Liu i Liu, 2016), pogotovo nakon popularizacije generativnih modela (Tanaka i Aranha, 2019; Fiore et al., 2019; Ba, 2019).

2. Glavni dio

2.1. Skupovi podataka i pretprocesiranje

Korišteni su javni skupovi stvarnih podataka za provođenje eksperimenata kako bi se izbjeglo pitanje znače li rezultati rada da je odabrana metoda dobra u modeliranju stvarnih procesa ili da je dobra u modeliranju algoritma pomoću kojeg su podatci sintetizirani (npr. (Harris, 2022)). Uz to, birani su veći skupovi jer nije jasno koliko bi na primjer skup od oko 3000 transakcija (Joshi, 2018) bio dobar za donošenje općenitih zaključaka.

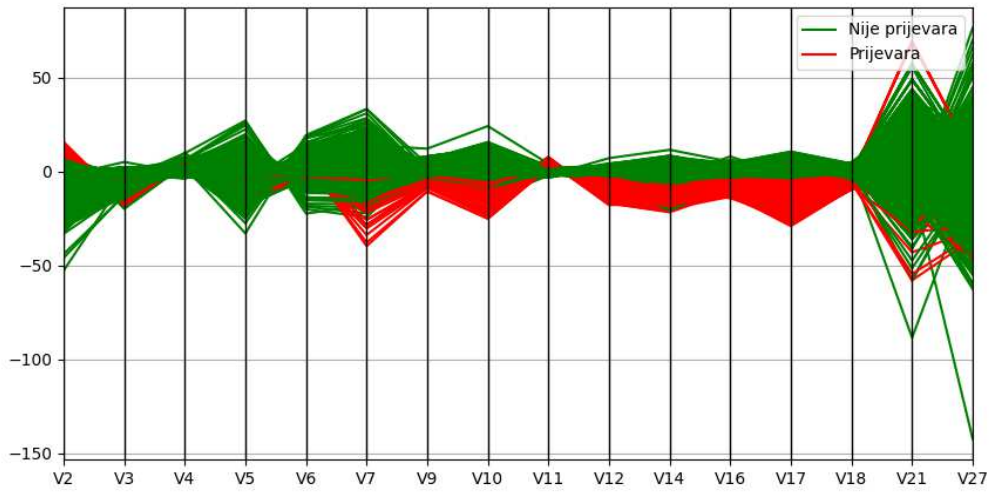
Postoje dva takva skupa podataka i korištena su oba kako bi se stekao dojam koliko su rezultati metoda ovisni o skupu podataka nad kojim su metode provedene. Njihov najveći nedostatak je to što ne sadrže nikakav identifikator korisnika pomoću kojeg bi bilo moguće grupirati vremenske nizove. Značenja značajki su općenito prikrivena zbog povjerljivosti podataka, ali moguće je da sadrže kumulativne statistike računa platitelja, kao i para (platitelj, primatelj), pa bi odabrana metoda mogla način na koji ponašanje prevaranata općenito odudara od tih statistika prepoznati kao anomaliju. Osim toga, skupovi su stari najmanje pet godina, što zbog očekivanog pomaka koncepta dovodi u pitanje relevantnost donesenih zaključaka.

Oba skupa podataka svedena su na 16 značajki s najvećom transinformacijom $I(\mathbf{X}, \mathbf{y})$ (Ross, 2014) u odnosu na ciljnu varijablu. Prednost ovog pristupa je bolje prepoznavanje prijevara zanemarivanjem anomalija onih značajki među kojima su anomalije češće kod legitimnih transakcija, odnosno značajki među kojima su prijevare slične legitimnim transakcijama, pa s obzirom na te značajke prijevare nisu anomalije. Nedostatak je potencijalno zanemarivanje anomalija u međuodnosima odbačenih značajki s ostalim značajkama *kod prijevara* te potencijalna veća osjetljivost na pomak koncepta: ako neka od odbačenih značajki postane dobar indikator ciljne varijable, to neće biti odmah

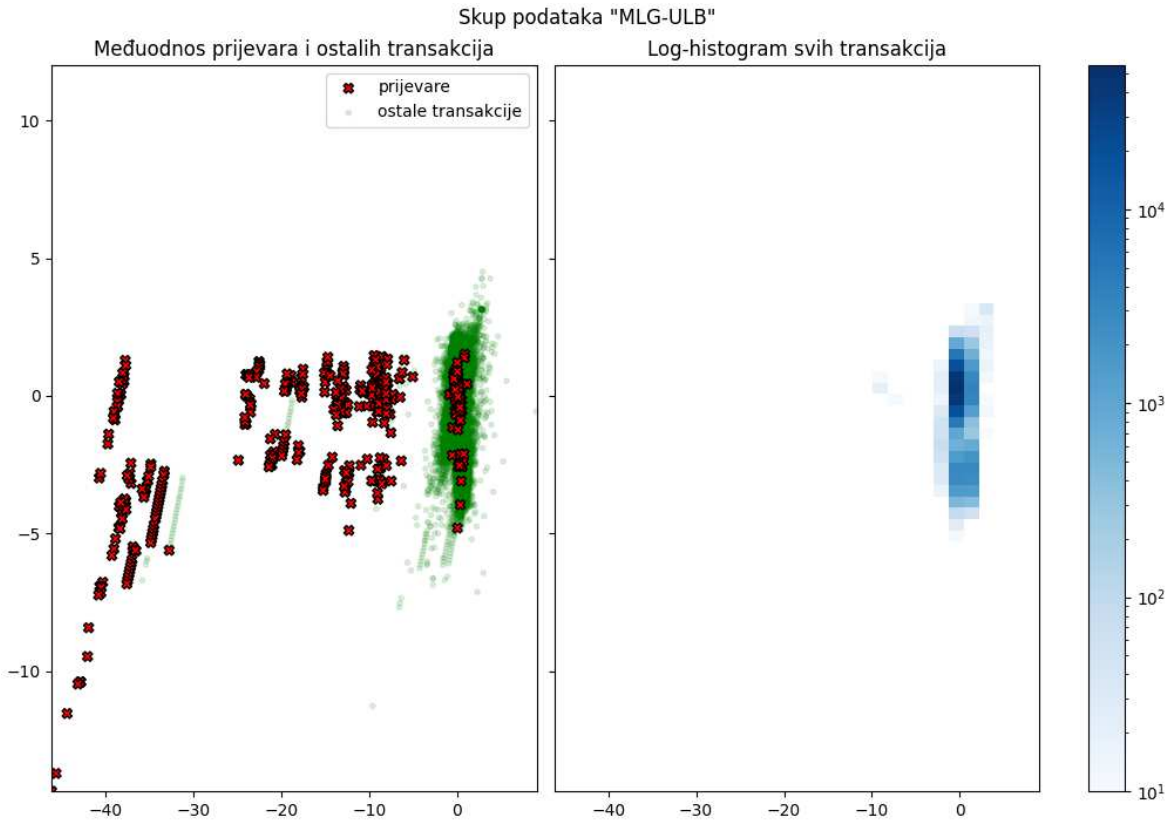
jasno, a kad bude, u tom će slučaju biti potrebno ponovno pretprocesirati podatke i trenirati model. Takvi nedostatci nisu potvrđeni na ovim skupovima podataka, pa je pristup odabran na temelju vidljivo boljih rezultata. Pretprocesiranje podataka ispostavilo se ključnim za postizanje boljih rezultata, što je i očekivano s obzirom na prirodu problema (veliko preklapanje ciljnih klasa), cilj (otkrivanje anomalija, a ne klasifikacija) i odabrane metode (nenadzirano učenje).

2.1.1. MLG-ULB

Ovaj skup podataka objavila je *Machine Learning Group* sveučilišta u Bruxellesu na Kaggleu (Dal Pozzolo). Sastoji se od 31 značajke: oznaka klase, vrijeme, iznos i 28 glavnih komponenata ostalih značajki. Od 284,807 transakcija 492 su prijevare (0.172%), a sve su se te transakcije dogodile tijekom dva dana u rujnu 2013. Odbačena je vremenska dimenzija, a ostale su skalirane na temelju njihovog interkvartilnog ranga, što manje utječe na procjenu sredine i varijance te na anomalije koje je cilj otkriti, nego na primjer „z-standardizacija“ ili „minmax standardizacija“, koje u obzir uzimaju sve podatke, uključujući stršeće vrijednosti. Na kraju je odabrano 16 značajki s najvećom transinformacijom u odnosu na ciljnu klasu, a ostale je moguće vidjeti u dodatku A1.



Slika 2.1. Skup za treniranje „MLG-ULB“: Odnosi značajki



Slika 2.2. Skup za treniranje „MLG-ULB“: Raspodjela transakcija i prijevara u podatkovnom prostoru (dvije glavne komponente)

2.1.2. Vesta

Tvrtka Vesta je ovaj skup podataka ustupila za Kaggle natjecanje *IEEE-CIS Fraud Detection* (Howard et al., 2019). Ima 590,540 transakcija u rasponu od 6 mjeseci od kojih su 20,663 prijevare (3.5%). Ima velik broj značajki i velik broj nedostajećih vrijednosti. Osim značajki transakcija, za podskup zapisa postoje „značajke identiteta“ u posebnoj tablici. Zato je najprije provedeno pretprocesiranje:

1. Transakcijama su pridružene značajke identiteta gdje je to bilo moguće, ostale su uklonjene — ostaje 144,233 transakcije od kojih je 11,318 prijevara (7.8%)

2. Uklonjene su sve značajke bez ijedne vrijednosti u skupu za treniranje

3. Odabrane su kategoričke varijable (sve koje nisu brojevi ili u skupu za treniranje imaju k različitih cjelobrojnih vrijednosti iz intervala $[1, k]$ ili je $k < 5$

4. Kategoričke značajke enkodirane su ordinalnim enkoderom i odabrano ih je jedna trećina koje su prema χ^2 testu najinteresantnije s obzirom na ciljnu klasu

5. Kategoričke značajke enkodirane su one-hot enkoderom

6. Pomoću stabla odluke odabrane su 32 najvažnije *numeričke* značajke s obzirom na ciljnu klasu

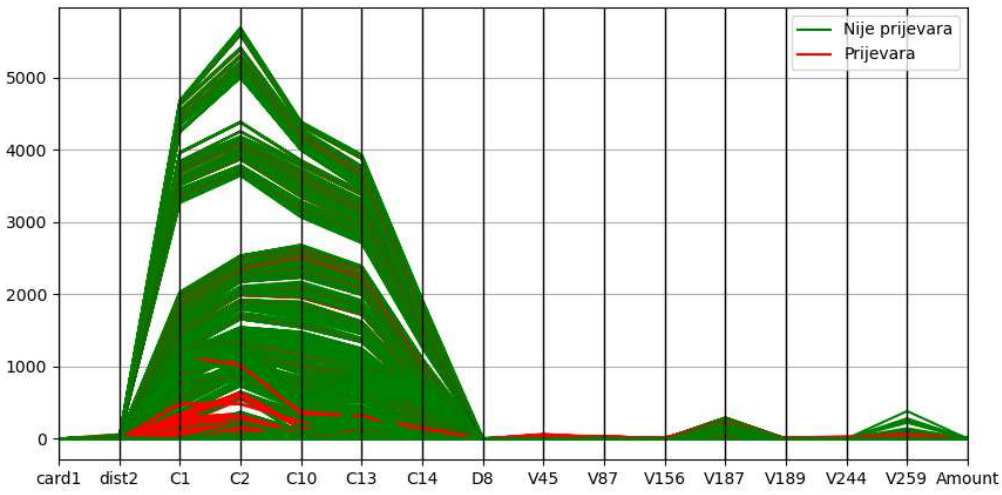
7. Sve značajke skalirane su na temelju njihova interkvartilnog ranga

8. Dodane su nedostajeće vrijednosti na temelju vrijednosti značajki najbližeg susjeda, a zbog visoke složenosti algoritma to je učinjeno u grupama od 8192 primjera (prethodnih desetak dana u prosjeku)

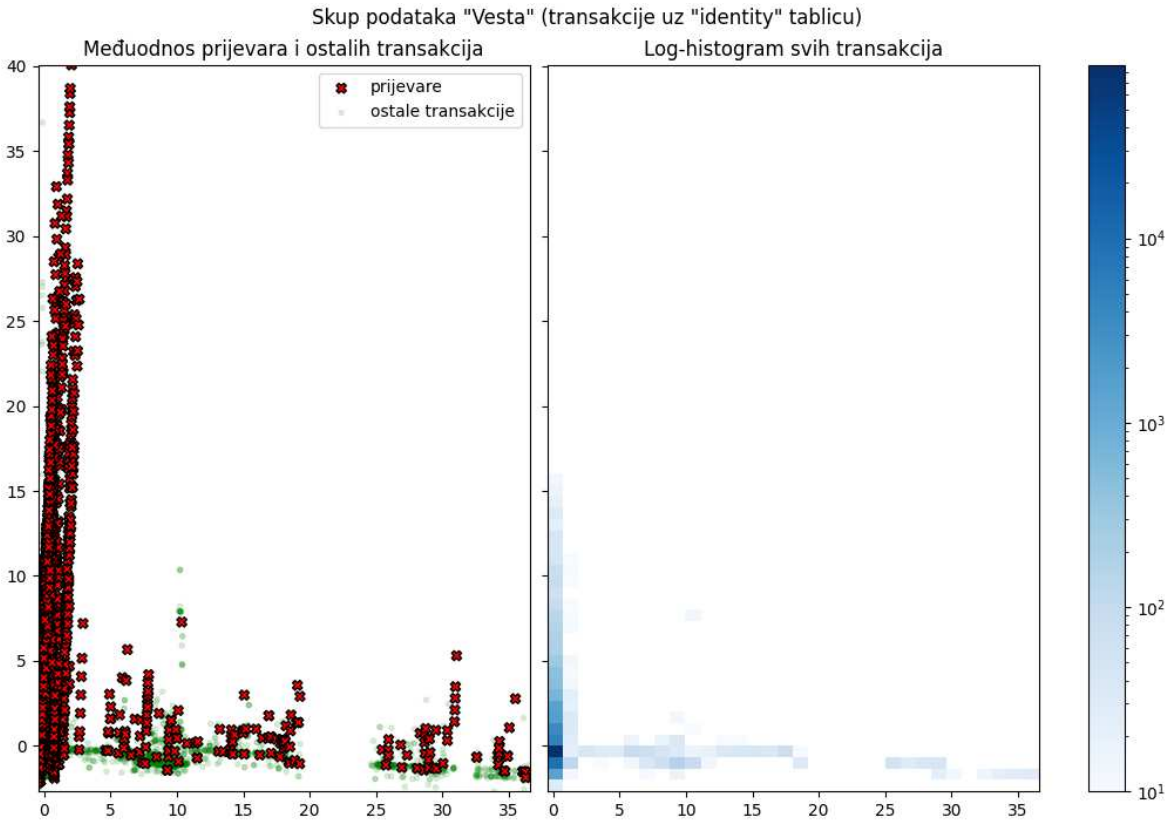
9. Odabrano je 28 značajki s najvećom transinformacijom $I(X, y)$ u odnosu na ciljnu varijablu te vrijeme i iznos transakcije

10. Kao i kod „MLG-ULB“ skupa podataka, provedeno je skaliranje i dodatno filtriranje na temelju transinformacije

Usporedba značajki u odnosu na ciljnu klasu prije zadnjeg filtriranja prikazana je u dodatku A2.



Slika 2.3. Skup za treniranje „Vesta“: Odnosi značajki



Slika 2.4. Skup za treniranje „Vesta“: Raspodjela transakcija i prijevara u podatkovnom prostoru (dvije glavne komponente)

2.2. Varijacijski autoenkoder

2.2.1. Detaljna definicija problema i motivacija

Potrebno je modelirati slučajan proces kojim su generirani podatci $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ pomoću neosmotrene slučajne varijable $\mathbf{z}$ koja se ravna prema apriornoj distribuciji $p_{\theta^*}(\mathbf{z})$ te kojom se uvjetuje izglednost $p_{\theta^*}(\mathbf{x}|\mathbf{z})$. Parametri θ^* tih distribucija su nepoznati, a gustoće su im skoro svugdje diferencijabilne s obzirom na $\mathbf{z}$ i θ.

Model bi trebao efikasno raditi i kad je $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})$ netraktabilna, pa nije moguće evaluirati ili diferencirati marginalnu izglednost te kad je netraktabilna i $p_\theta(\mathbf{z}|\mathbf{x})$, pa nije moguće koristiti algoritam maksimizacije očekivanja. Ta netraktabilnost čest je slučaj kod modeliranja distribucija dubokim neuronskim mrežama.

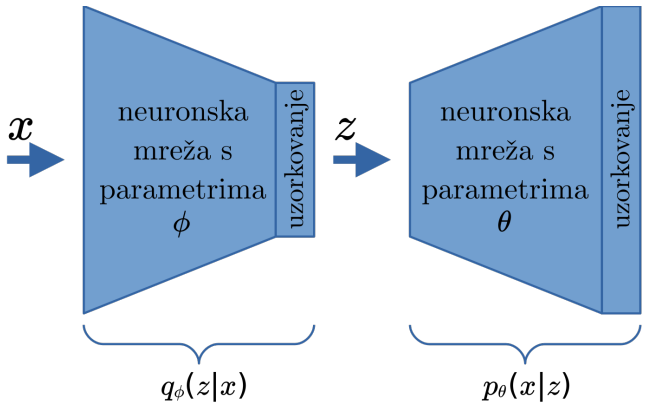
Uz to, model bi trebao efikasno raditi i u slučaju velikih skupova podataka koje bi bilo presporo rješavati pristupima temeljenima na uzorkovanju, poput Monte Carlo maksimizacije očekivanja (Kingma i Welling, 2022).

Krajnji ciljevi takvog pristupa modeliranju su ostvariti efikasnu procjenu:

- **parametara** θ npr. u svrhu generiranja podataka sličnih stvarnima
- **aposteriorne distribucije** $p_\theta(\mathbf{z}|\mathbf{x})$ npr. za transformiranje podataka u oblik korisniji za daljnju obradu, poput klasifikacije
- **marginalne inferencije ulazne varijable x** npr. za izoštravanje slike

2.2.2. Opis modela VAE

U skladu s motivacijom, VAE je generativan model strojnog učenja definiran parametriziranom apriornom distribucijom $p_\theta(\mathbf{z})$, izglednošću $p_\theta(\mathbf{x}|\mathbf{z})$ te procjenom posteriorne distribucije $p_\theta(\mathbf{z}|\mathbf{x})$ za koju se koristi zapis $q_\phi(\mathbf{x}|\mathbf{z})$. Ta se procjena često naziva „koder“ jer efektivno preslikavanjem „kodira“ osmotrenu varijablu **x** u neosmotrenu (latentnu) varijablu **z** (kôd), a izglednost „dekoder“ jer preslikavanjem „dekodira“ iz **z** u **x**.



Slika 2.5. Shema VAE

2.2.3. Procjena parametara VAE

Cilj je modelirati slučajan proces $p_\theta(\mathbf{x})$ s neosmotrenom slučajnom varijablom **z** koja uvjetuje **x** (Kingma i Welling, 2019):

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) \tag{2.1}$$

Uz skup podataka X i pretpostavku da je nastao iz distribucije $p_\theta(\mathbf{x})$ s upravo onim parametrima θ za koje je vjerojatnost najveća da generira takav uzorak, kako bi se što bolje modelirao željeni proces, cilj postaje naći parametre θ maksimizacijom njihove iz-

glednosti:

$$\underset{\theta}{\operatorname{argmax}}\, p_\theta(\mathbf{x}) \tag{2.2}$$

Budući da integral u izrazu marginalne razdiobe (2.1) općenito nije traktabilan, nije ga moguće optimirati diferenciranjem s obzirom na parametre.

No, ako se uvede $q_\phi(\mathbf{z}|\mathbf{x})$ kao aproksimacija aposteriorne distribucije $p_\theta(\mathbf{z}|\mathbf{x})$ te se umjesto izglednosti optimira log-izglednost, koja zbog monotonosti logaritamske funkcije postiže maksimum za iste vrijednosti parametara kao izglednost, cilj optimizacije može se raspisati na sljedeći način:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] && (2.3)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] && (2.4)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] && (2.5)\\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right]}_{\substack{=\mathcal{L}_{\theta,\phi}(\mathbf{x})\\(\text{ELBO})}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right]}_{=D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))} && (2.6)
\end{aligned}$$

Zahvaljujući tome što $\log p_\theta(\mathbf{x})$ ne ovisi o $q_\phi(\mathbf{z}|\mathbf{x})$, može se samo ubaciti u očekivanje u prvom izrazu (2.3).

U zadnjem redu (2.6) drugi pribrojnik naziva se Kullback-Leibler (KL) divergencija te je mjera odstupanja distribucije $q_\phi(\mathbf{z}|\mathbf{x})$ od $p_\theta(\mathbf{z}|\mathbf{x})$. Njeno je svojstvo da je uvijek nenegativna, a nula akko $q_\phi(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z}|\mathbf{x})$.

Budući da je KL divergencija nenegativna, prvi pribrojnik u tom izrazu (2.6) označen s $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ uvijek je manji od izraza s lijeve strane jednakosti, tj. od $\log p_\theta(\mathbf{x})$. Kako se $p_\theta(\mathbf{x})$ često naziva dokazom (engl. *evidence*), prikladan naziv za $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ je „donja granica dokaza“ (engl. *Evidence Lower BOund*) (ELBO).

Dakle maksimizacija donje granice dokaza ima dvojaku ulogu: minimizira odstupanje procijenjenog posteriora od stvarnoga te maksimizira izglednost $p_\theta(\mathbf{x})$, bez da u izrazu za optimiranje ostaju netraktabilni $p_\theta(\mathbf{x})$ i $p_\theta(\mathbf{z}|\mathbf{x})$.

Za maksimizaciju $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ potrebno je naći gradijent s obzirom na parametre θ i ϕ. Što se tiče θ, postupak je sljedeći:

$$\nabla_\theta \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_\theta \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\big] \tag{2.7}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\nabla_\theta \left(\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right)\big] \tag{2.8}$$

$$\simeq \nabla_\theta \left(\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right) \tag{2.9}$$

$$= \nabla_\theta \log p_\theta(\mathbf{x},\mathbf{z}) \tag{2.10}$$

Ukratko, jednostavnom Monte Carlo procjenom dobiva se nepristrani procjenitelj gradijenta ELBO s obzirom na θ.

Problem nastaje kod procjene gradijenta ELBO s obzirom na parametre ϕ:

$$\nabla_\phi \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\big] \tag{2.11}$$

$$\neq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\nabla_\phi \left(\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right)\big] \tag{2.12}$$

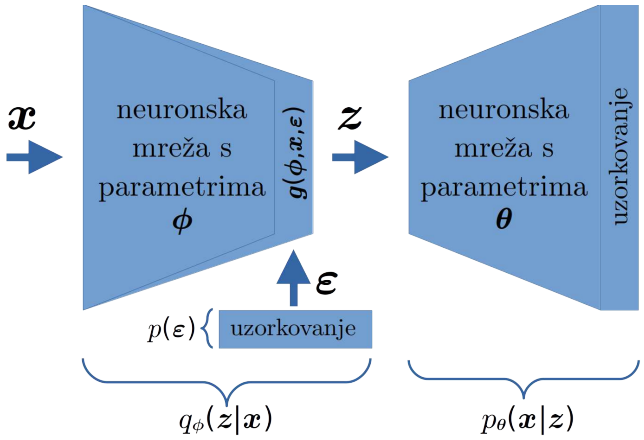
Cilj je izbjeći da uzorkovanje slučajne varijable ovisi o parametrima modela. Za to se koristi **reparametrizacijski trik**. Ukratko, slučajna varijabla $\mathbf{z}$ raspiše se kao diferencijabilna i invertibilna funkcija koja drugu slučajnu varijablu ϵ (čije realizacije ne ovise ni o $\mathbf{x}$ ni ϕ) transformira pomoću $\mathbf{x}$ i ϕ:

$$\mathbf{z} = \mathbf{g}(\phi, \mathbf{x}, \epsilon) \tag{2.13}$$

Zato se izraz za ELBO zapisuje kao:

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{p(\epsilon)}\big[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\big] \tag{2.14}$$

Za Monte Carlo procjenitelj donje granice dokaza izvorno je predložena jedna realizacija slučajne varijable ϵ. Uz to, korisno je u izrazu potpuno zamijeniti $\mathbf{z}$ s ϵ da bude



Slika 2.6. Reparametrizacijski trik

jasnije što se stvarno događa. Tako je funkcija koja se maksimizira:

$$\tilde{\mathcal{L}}_{\theta,\boldsymbol{\phi}}(\mathbf{x}) = \log p_\theta(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \tag{2.15}$$

$$= \log p_\theta(\mathbf{x},\mathbf{g}(\boldsymbol{\phi},\mathbf{x},\epsilon)) - \log p(\epsilon) + \log \left|\det\left(\frac{\partial \mathbf{g}(\boldsymbol{\phi},\mathbf{x},\epsilon)}{\partial \epsilon}\right)\right| \tag{2.16}$$

Pritom je za zamjenu $\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ korišteno svojstvo invertibilnosti funkcije $\mathbf{g}(.)$.

Pseudokodom 1 opisan je izvorni algoritam optimizacije parametara VAE, tj. gornjeg izraza (2.16).

Algoritam 1 *Auto-Encoding Variational Bayes* (AEVB)

Ulaz:
Skup podataka $\mathbf{X} = \{x^{(i)}\}_{i=1}^N$
Generativni model $p_\theta(\mathbf{x},\mathbf{z})$
Model zaključivanja $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$
$(\theta,\boldsymbol{\phi}) \leftarrow$ inicijalizacija parametara
while optimizator nije konvergirao **do**
$\mathcal{M} \sim \mathcal{D}$ ▷ Nasumični podskup podataka
for $i = 1,\ldots,|\mathcal{M}|$ **do**
$\epsilon \sim p(\epsilon)$
$\text{ELBO}^{(i)} \leftarrow \log p_\theta(\mathbf{x}^{(i)},\mathbf{g}(\boldsymbol{\phi},\mathbf{x}^{(i)},\epsilon)) - \log p(\epsilon) + \log \left|\det\left(\frac{\partial \mathbf{g}(\boldsymbol{\phi},\mathbf{x}^{(i)},\epsilon)}{\partial \epsilon}\right)\right|$
$\Delta\theta^{(i)}, \Delta\boldsymbol{\phi}^{(i)} \leftarrow \nabla_{\theta,\boldsymbol{\phi}}\left(\log p_\theta(\mathbf{x}^{(i)},\mathbf{g}(\boldsymbol{\phi},\mathbf{x}^{(i)},\epsilon)) + \log \left|\det\left(\frac{\partial \mathbf{g}(\boldsymbol{\phi},\mathbf{x}^{(i)},\epsilon)}{\partial \epsilon}\right)\right|\right)$
end for
OPTIMIZATOR$(\theta,\boldsymbol{\phi},\Delta\theta,\Delta\boldsymbol{\phi})$ ▷ Ažuriranje parametara pomoću gradijenata
end while
return $\theta,\boldsymbol{\phi}$

2.2.4. Općenite nadogradnje VAE

Iz izvornog VAE izvedeno je više inačica kako bi ga se (učinkovitije) moglo koristiti za razne zadaće. Najčešće se radi o modeliranju distribucija slika. Već 2014., manje od godinu dana nakon objave rada u kojem je predstavljen VAE, izvorni autori prilagodili su VAE za klasifikaciju uz polunadzirano učenje (Kingma et al., 2014), a jedan od modela iz tog rada može se koristiti i za stvaranje podataka koji pripadaju proizvoljnoj klasi.

Neke od utjecajnijih prilagodbi VAE drugih autora su *importance weighted AE* (IWAE) (Burda et al., 2016) čija je glavna razlika od VAE promjena ciljne funkcije iz procjene ELBO iz jednog uzorka u procjenu ELBO iz više uzoraka uz pomoć uzorkovanja prema važnosti, što daje za rezultat precizniji procjenitelj, *categorical VAE* (Jang et al., 2017) koji koristi Gumbel-Softmax distribuciju za *kategoričke* latentne varijable i *Wasserstein AE* (WAE) (Tolstikhin et al., 2019) u kojem se odstupanje prave apriorne distribucije od procijenjene računa za sve primjere odjednom pomoću Wasserstein udaljenosti, umjesto primjer po primjer pomoću KL divergencije — što je slučaj kod VAE. Uz to, rekonstrukcijska pogreška može biti proizvoljna udaljenost u prostoru primjera, što omogućuje bolje odvajanje primjera u latentnom prostoru i rezultira boljim rekonstrukcijama. Poticanje razdvajanja u latentnom prostoru postiže i β-*VAE* (Burgess et al., 2018) koji hiperparametrom β množi KL divergenciju između uzorkovane aproksimacije aposteriorne distribucije i apriorne distribucije u sklopu ELBO. Poboljšanje rezultata daje i korištenje mješavine aposteriornih distribucija uvjetovanih pseudo-ulaznim podatcima (posebna vrsta parametara modela) kao procjene apriorne distribucije (Tomczak i Welling, 2018). Opsežniji pregled dan je u sklopu rada uz biblioteku za „benchmarking“ inačica varijacijskih autoenkodera (Chadebec et al., 2022).

2.2.5. VAE za otkrivanje anomalija i vremenske nizove

Opisana svojstva VAE omogućavaju njihovo učinkovito korištenje za otkrivanje anomalija (An i Cho, 2015). Za razliku od uobičajenih dubokih generativnih modela čije se otkrivanje anomalija često temelji na kvadratnom odstupanju danog podatka od njegove rekonstrukcije, VAE pruža konkretniju probabilističku podlogu na temelju koje je moguće uzeti u obzir i varijancu — ne samo u prostoru podataka, nego i u latentnom prostoru, uzorkovanjem aproksimacije posteriorne distribucije.

Osim originalnih VAE, za otkrivanje anomalija osmišljene su i prilagodbe, poput *Donut* (Xu et al., 2018) algoritma vrlo uspješnog u otkrivanju anomalija u sezonalnosti KPI-jeva (ključnih indikatora performansi), a *variational LSTM* (Zhou et al., 2021) generalizira bolje od svih testiranih alternativa. Osim navedenih, za obradu vremenskih nizova razvijeno je još inačica VAE, za koje se koristi zajednički naziv *dynamical VAE* (Girin et al., 2021).

Kod otkrivanja anomalija u sustavu okidača akceleratora čestica (LHC) u CERN-u, uvjetovanje VAE kategoričkom značajkom (Pol et al., 2019) pokazalo se kao još jedan koristan trik za poboljšanje učinkovitosti: značajka manje za enkodirati latentnom varijablom korištena je kao značajka više za sigurnije usmjeravanje dekodera.

Ipak, korištenje VAE ne daje uvijek dobar detektor anomalija. U slučaju otkrivanja anomalija na slikama pluća pomoću VAE i GAN-a, ni nakon iscrpnog ugađanja hiperparametara nisu postignute zavidne mjere površine pod ROC krivuljom (mjera opisana u dijelu 2.5.) (Buitrago et al., 2018). Autori tvrde da je razlog to što VAE i GAN ne mogu obuhvatiti složenost takvih podataka.

Iako nadzirano učenje nije naročito zastupljeno u otkrivanju anomalija (Ruff et al., 2021), iskušan je i taj pristup prilagođenim VAE (Kawachi et al., 2018, 2019), što je posebno zanimljivo u kontekstu OPKK gdje ima dovoljno označenih podataka, no ispostavilo se da takvi pristupi ne daju naročito bolje rezultate nego običan VAE.

2.2.6. Primjena na izazove kod OPKK

Zahvaljujući tome što je generativan model, VAE je korišten za dodatno uzorkovanje (engl. *oversampling*) transakcija-prijevara, na temelju kojega je klasifikator (troslojna neuronska mreža) ostvarivao bolje rezultate nego uz SMOTE ili generativnu suparničku mrežu (GAN) (Tingfei et al., 2020), a *Dual Sequential VAE* (Alazizi et al., 2020) dao je bolje rezultate nego običan VAE, *autoencoding binary classifier* (ABC) (Yamanaka et al., 2019) i *learning reconstruction capability* (LRC) (Munawar et al., 2017) pristupi, s površinom ispod PRC krivulje (2.5.) ~ 0.53 prilikom klasifikacije *CatBoost* klasifikatorom.

Primjena varijacijskog autoenkodera baš kao modela za otkrivanje anomalija kod OPKK može se naći u jednom završnom radu gdje je manje od 4% prijevara među svim

transakcijama označenima kao prijevare (Sweers, 2018).

2.3. Metodologija

Budući da nije pronađen naročito uspješan nadziran pristup otkrivanju anomalija pomoću VAE, kao referentni odabrani su modeli za otkrivanje anomalija također zasnovani na nenadziranom učenju, koji se često pojavljuju u literaturi u sklopu OPKK.

2.3.1. VAE, IWAE i WAE

U sklopu ovog rada implementiran je originalni VAE, a kako bi se ustanovilo pomaže li bolja procjena donje granice dokaza pri postizanju boljih rezultata, implementiran je i IWAE. Dodatno, korišten je i WAE kako bi se omogućilo lakše odvajanje u latentnom prostoru, što bi trebalo dovesti do manje pogreške prilikom mapiranja legitimnih transakcija, pa bi ih trebalo biti lakše odvojiti od prijevara ako njihove pogreške ostanu veće.

VAE

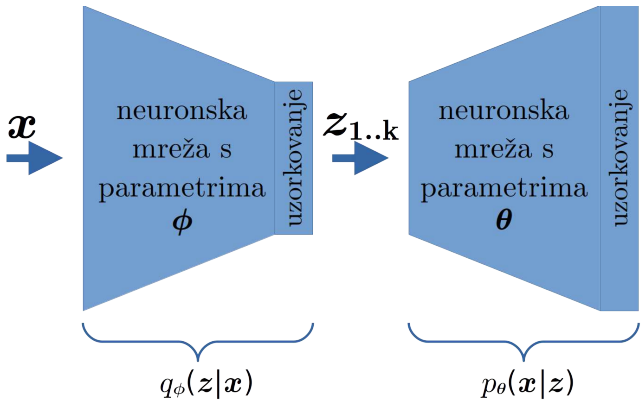
Opisani običan varijacijski autoenkoder prilikom modeliranja distribucije legitimnih transakcija optimira vjerojatnost da ulazna transakcija dolazi iz uvjetne distribucije $p_\theta(\mathbf{x}|\mathbf{z})$ tako što ju enkoderom preslikava u latentni prostor $\mathcal{Z}$, a zatim iz toga dekoderom u parametre distribucije $p_\theta(\mathbf{x}|\mathbf{z})$.

Ako podatak nije viđen tijekom treniranja, teže će biti preslikan u ono područje $\mathcal{Z}$ koje bi dekoder preslikao u one parametre $p_\theta(\mathbf{x}|\mathbf{z})$ za koje je vjerojatno da su dali originalni $\mathbf{x}$, nego što bi to bilo za onakve podatke nad kakvima se upravo to optimiralo. Zato su taj postupak i vjerojatnost da je $\mathbf{x}$ dobiven iz $p_\theta(\mathbf{x}|\mathbf{z})$ korišteni u otkrivanju anomalija (An i Cho, 2015) te se ta vjerojatnost naziva „rekonstrukcijska vjerojatnost“. Kako bi se uzela u obzir varijabilnost u $\mathcal{Z}$, moguće je koristiti Monte Carlo uzorkovanje za dobivanje uzorka iz $q_\phi(\mathbf{z}|\mathbf{x})$, na temelju kojeg se računaju parametri $p_\theta(\mathbf{x}|\mathbf{z})$ i za dane parametre vjerojatnost $p_\theta(\mathbf{x}|\mathbf{z})$. Uprosječivanjem te vjerojatnosti dolazi se do indikatora je li podatak anomalija.

IWAE

Kod VAE ELBO se računa na temelju jednog uzorka, što je dalo dovoljno dobre rezultate, ali pokazalo se da uzorkovanje po važnosti uz više uzoraka aposteriorne distribucije $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ daje procjenu bližu $p_\theta(x)$, pa je ciljna funkcija pretvorena u:

$$\mathcal{L}_{\theta,\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_k \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p_\theta(\mathbf{x}|\mathbf{z}_i)p(\mathbf{z}_i)}{q_\phi(\mathbf{z}_i|\mathbf{x})}\right] \tag{2.17}$$



Slika 2.7. IWAE: VAE čiji je i-ti od k uzoraka otežan s $p_\theta(\mathbf{x}, \mathbf{z}_i)/q_\phi(\mathbf{z}_i|\mathbf{x})$

Za $k = 1$ ovaj je izraz jednak izrazu ELBO kod VAE. Cilj je ovim pristupom povećati stabilnost treniranja i po mogućnosti postići bolje rezultate.

Još jedna razlika IWAE i originalnog VAE je što IWAE izričito dopušta korištenje više stohastičkih slojeva, no ta karakteristika nije korištena u ovom radu.

WAE

Ciljna funkcija običnog VAE funkcionira na temelju sprege rekonstrukcijske pogreške $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ i KL divergencije između aproksimacije aposteriorne distribucije i apriorne dis-

tribucije $D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}))$.

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right]\right] \tag{2.18}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log\left[p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right] - \log\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]\big] \tag{2.19}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log\left[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] + \log\left[p_{\boldsymbol{\theta}}(\mathbf{z})\right] - \log\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]\big] \tag{2.20}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log\left[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]] - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})) \tag{2.21}$$

Zahvaljujući minimizaciji KL divergencije navedenih distribucija svaki primjer dobiven iz aposteriorne distribucije $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ koji se ne ravna prema apriornoj distribuciji $p_{\boldsymbol{\theta}}(\mathbf{z})$ utječe na gradijent VAE kako bi ubuduće i taj primjer odgovarao aprioronoj distribuciji. Jedino što spriječava parametre $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (tj. enkoder) da ne vraća uvijek apriornu distribuciju, a ne neku uvjetovanu s $\mathbf{x}$ — je rekonstrukcijska pogreška. Dekoder ne bi znao koji podatak generirati da je KL divergencija 0. Dakle, očito je njena minimizacija kontraproduktivna vjernoj rekonstrukciji. Zato je cilj olabaviti taj regularizacijski dio ciljne funkcije.

Kako bi se to postiglo, umjesto KL divergencije uvodi se Wassersteinova udaljenost, poznata i kao „udaljenost premještanja zemlje“ jer kad bi gustoće distribucija (P_x i P_y) bile hrpe zemlje, Wassersteinova udaljenost je općenito najmanja „cijena“ (c) potrebna da se premještanjem zemlje s jedne hrpe dobije druga hrpa.

$$W_c(P_x, P_y) = \min_{\Gamma \in \mathcal{P}(P_x,P_y)} \int_{X\times Y} c(x,y)\, d\Gamma(x,y) \tag{2.22}$$

gdje je $\mathcal{P}(P_x, P_y)$ skup svih zajedničkih distribucija čije su marginalne distribucije P_x i P_y

U ovom kontekstu ta je „cijena“ metrička udaljenost i tako zadana omogućava korištenje Kantorovich-Rubinstein dualnosti čiji je rezultat:

$$W_c(P_x, P_y) = \max_{f\in\mathcal{F}_L} \mathbb{E}_{X\sim P_X}[f(X)] - \mathbb{E}_{Y\sim P_Y}[f(Y)] \tag{2.23}$$

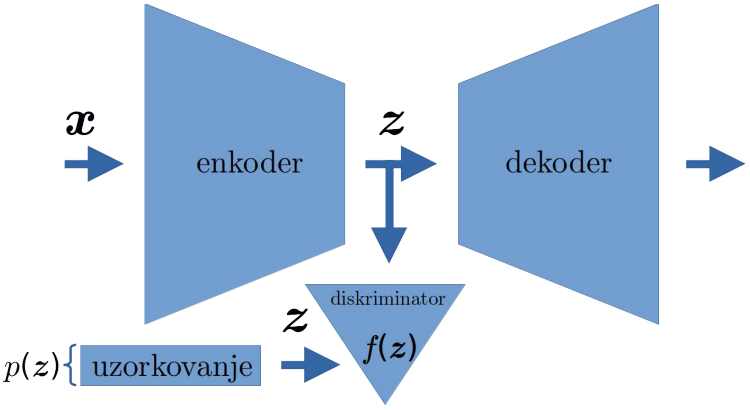
uz ograničenje da je $f(.)$ funkcija čiji je gradijent ograničen s 1.

Navedena funkcija f je u ovoj implementaciji dobivena učenjem parametara neuron-

ske mreže te ima sličnu ulogu kao diskriminator u generativnoj suparničkoj mreži (GAN), samo u latentnom prostoru. U svakom se koraku optimizacije WAE prvo provede korak optimizacije f za uzorak apriorne i aposteriorne distribucije, a zatim se takva funkcija f koristi u ciljnoj funkciji za optimiranje cjelokupnog WAE:

$$\min_{\phi,\theta} \mathbb{E}_{\mathbf{x}\sim p_{\theta^*}(\mathbf{x})} \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[c(\mathbf{x}, \text{dekoder}(\mathbf{z}))] + \lambda \cdot W_c(q_\phi(\mathbf{z}|\mathbf{x}), p(\mathbf{z})) \tag{2.24}$$

gdje je c proizvoljna funkcija pogreške rekonstrukcije. Prilikom otkrivanja anomalija upravo se iznos te pogreške koristi kao indikator je li podatak anomalija.



Slika 2.8. WAE: VAE čije se realizacije $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ *pomoću diskriminatora* uspoređuju s apriornom distribucijom $p(\mathbf{z})$

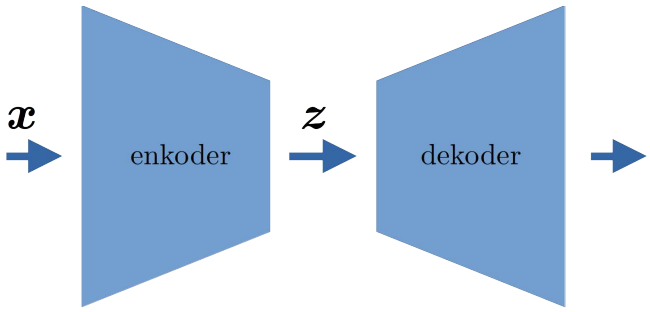
Na ovaj se način ni regularizacijom više ne nastoji postići da se svaki element uzorka mora ravnati prema apriornoj distribuciji, nego ih se temeljem $f(.)$ rasporedi na ona mjesta koja će sveukupno rezultirati apriornom distribucijom uz najmanje premještanja.

U originalnom su radu enkoder i dekoder determinističke funkcije, pa je moguće koristiti običan AE (opisan u 2.3.2.) za implementaciju, a slučajno uzorkovanje koristi se za treniranje diskriminatora kako bi se dobio uzorak apriorne distribucije $p(\mathbf{z})$.

2.3.2. Autoenkoder

Kao primjer referentnog dubokog modela strojnog učenja odabran je autoenkoder zbog svoje sličnosti s varijacijskim autoenkoderom. Arhitektura ove neuronske mreže također ima „usko grlo“, koje efektivno služi za sažimanje ulaznog prostora u manji, latentni prostor, samo je kod AE cijela arhitektura deterministička. Nakon uskog grla postoji niz slojeva čija je zadaća rekonstrukcija ulaznih podataka. Oni podatci čije se karakteristike

prerijetko javljaju da budu rezultat enkodiranja u usko grlo (sažimanje s gubitcima) neće se tako dobro rekonstruirati kao podatci sa zastupljenijim karakteristikama te će razlika u kvaliteti rekonstrukcije (najčešće kvadratna pogreška) biti indikator za otkrivanje anomalija.



Slika 2.9. Autoenkoder

2.3.3. f-AnoGAN

Kako bi se kvalitetnije opisao f-AnoGAN, prvo je u kratkim crtama dan opis običnog AnoGAN modela, a zatim nadogradnje koje čine f-AnoGAN.

AnoGAN

Obični AnoGAN (Schlegl et al., 2017) je model temeljen na generativnoj suparničkoj mreži, proširen kako bi se mogao koristiti za otkrivanje anomalija. Kao i GAN, sastoji se od generativne i diskriminativne neuronske mreže sa suparničkim zadaćama: generativna mreža (generator G) nastoji iz svakog uzorka latentne varijable $\mathbf{z}$ stvoriti podatke / transakcije koje diskriminativna mreža (diskriminator D) ne može razlikovati od podataka iz skupa za treniranje, a cilj diskriminatora je prepoznati koji podatci su „stvarni“, to jest iz skupa za treniranje, a koji ne. Danom podatku diskriminator pridružuje procjenu vjerojatnosti da je „stvaran“. U skladu s time, GAN optimira mjeru unakrsne entropije:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{\theta^*}(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{2.25}$$

gdje p_{θ^*} označava distribuciju čiji je uzorak skup podataka za treniranje, a $p(\mathbf{z})$ je distribucija latentne varijable (često uniformna ili normalna), čiji su uzorci transformirani generatorom $G(\mathbf{z})$.

Par G i D se kod AnoGAN-a koristi za detekciju anomalija tako što se za svaki testni

podatak $\mathbf{x}$ procjenjuje mjera $A(\mathbf{x})$ (engl. *anomaly score*) dana formulom:

$$A(\mathbf{x}) = (1 - \lambda) \cdot R(\mathbf{x}) + \lambda \cdot D(\mathbf{x}) \tag{2.26}$$

u kojoj se hiperparametrom λ zadaje omjer rezidualnog odstupanja $R(\mathbf{x})$ i diskriminativnog odstupanja $D(\mathbf{x})$ u ukupnoj mjeri $A(\mathbf{x})$.

Rezidualno odstupanje označava koliko je dani podatak $\mathbf{x}$ različit od najsličnijeg podatka $\mathbf{x}'$ kojeg je generator naučio generirati, a taj najsličniji podatak dobije se u Γ koraka gradijentnog spusta po latentnom prostoru varijable $\mathbf{z}$ u potrazi za onom koja daje minimalno odstupanje $R(\mathbf{x})$:

$$R(\mathbf{x}) = \sum |\mathbf{x} - G(\mathbf{z}_\Gamma)| \tag{2.27}$$

Diskriminativno odstupanje označava koliko je diskriminator uvjeren da je dani podatak „stvaran“, tj. sličan podatcima iz skupa za treniranje. No, umjesto da se ažuriranje temelji na mjeri vjerojatnosti $D(\mathbf{x})$ koju za taj podatak vrati diskriminator, korisnije je temeljiti procjenu na podudaranju značajki (engl. *feature matching*) iz jednog od slojeva diskriminatora na temelju kojih diskriminator donosi odluku. Drugim riječima, traži se podudaranje značajki dobivenih iz danog podatka $\mathbf{x}$ s najsličnijim značajkama dobivenim generiranjem primjerka $\mathbf{x}' = G(\mathbf{z}_\Gamma)$ iz latentnog prostora $\mathcal{Z}$. Dakle, i u ovom se slučaju traži najsličniji podatak kojeg generator može stvoriti, pa se i diskriminativno odstupanje dodaje u ciljnu funkciju gradijentnog spusta po prostoru latentne varijable $\mathbf{z}$. Izraz koji opisuje diskriminativno odstupanje zato izgleda ovako:

$$D(\mathbf{x}) = \sum |\mathbf{f}(\mathbf{x}) - \mathbf{f}(G(\mathbf{z}_\Gamma))| \tag{2.28}$$

gdje $\mathbf{f}(\mathbf{x})$ označava prvih nekoliko slojeva neuronske mreže diskriminatora.

Korist uporabe i ovakvog diskriminativnog odstupanja (uz rezidualno) posebno je jasna kod slika, koje i za mali pomak imaju veliku pogrešku kad se uspoređuju pikseli na istim koordinatama. Cilj je da se na taj način bitnije značajke izdvoje i kod transakcija u kartičnom plaćanju.

Očit problem ovog pristupa je što za svaki primjer treba provesti Γ koraka gradijentnog spusta kako bi se procijenilo je li anomalija (što posebno nije prikladno za provjeru

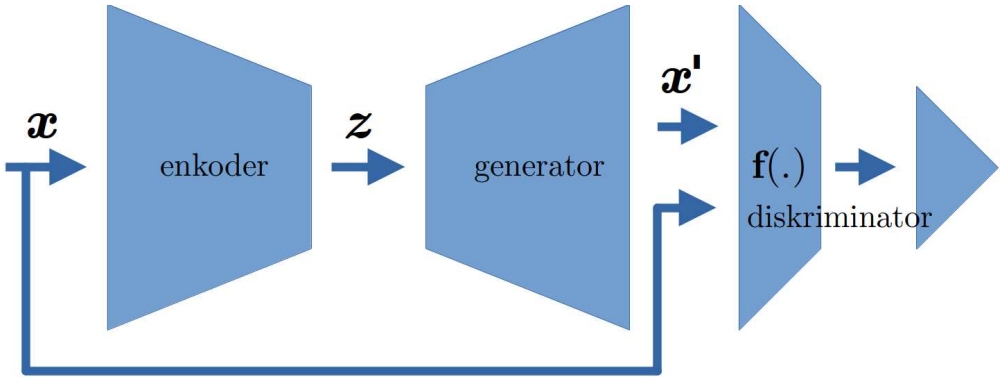
velikog broja transakcija). Zato su isti autori dvije godine kasnije predložili poboljšanje (Schlegl et al., 2019).

f-AnoGAN

Glavna prednost koju f-AnoGAN ima nad AnoGAN-om je skraćenje vremena inferencije korištenjem enkodera kojeg se prethodno nauči preslikavati ulazne podatke u pripadne vrijednosti latentne varijable z, što se potom koristi za izračun mjere anomalije na sličan način kao kod AnoGAN-a. Jedine razlike su korištenje kvadratne umjesto apsolutne razlike značajki te korištenje drugačije linearne kombinacije rezidualnog i diskriminativnog odstupanja:

$$A(\mathbf{x}) = \frac{1}{n} \cdot ||\mathbf{x} - G(E(\mathbf{x}))||^2 + \frac{\kappa}{n_d} ||\mathbf{f}(\mathbf{x}) - \mathbf{f}(G(E(\mathbf{x})))||^2 \tag{2.29}$$

Ovdje hiperparametar κ mijenja hiperparametar λ, a n i n_d redom označavaju broj značajki ulazne i latentne varijable. Budući da elementi skupa podataka za treniranje ne bi smjeli biti anomalije, ovaj se izraz minimizira kao ciljna funkcija prilikom treniranja enkodera. Dodatno, umjesto GAN-a koristi se Wassersteinova generativna suparnička mreža (WGAN) što omogućuje robusniji postupak učenja: izraz za unakrsnu entropiju u ciljnoj funkciji mijenja se u izraz za Wassersteinovu udaljenost uz korištenje Kantorovich-Rubinstein dualnosti:



Slika 2.10. f-AnoGAN

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{g^*}(\mathbf{x})}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[D(G(\mathbf{z}))] \tag{2.30}$$

Kako bi vrijedio taj izraz dobiven iz Kantorovich-Rubinstein dualnosti, potrebno je pretpostaviti da je gradijent diskriminatora u svakoj točki najviše 1. Zato se prilikom treniranja diskriminatora korisnim ispostavilo kažnjavanje magnitude gradijenta različite od

1 korištenjem kvadratnog odstupanja (Gulrajani et al., 2017).

2.3.4. Jednoklasni stroj potpornih vektora (OCSVM)

Za razliku od originalnog stroja potpornih vektora čiji je cilj u prostoru značajki (često zadan jezgrenom funkcijom) naći hiperravninu najudaljeniju od dvije klase, kod jednoklasnog stroja potpornih vektora (Schölkopf et al., 2001) cilj je naći hiperravninu između ishodišta i podataka, što udaljeniju od ishodišta. Naravno, za to mora vrijediti da su ishodište i podatci linearno odvojivi u prostoru značajki. To primjerice vrijedi za Gaussovu jezgru jer je za svaki par primjera iznos jezgrene funkcije pozitivan, odnosno pozitivan je njihov skalarni umnožak u prostoru značajki, dakle između svaka dva primjera kosinus kuta je pozitivan, što znači da su u istom „kvadrantu“ ((engl. *orthant*) u n-dimenzionalnom prostoru). Kod malih skupova podataka ovaj se problem optimira kvadratnim programiranjem, no za veće se koristi gradijentni spust. Osim odabira jezgre kod OCSVM bitan je hiperparametar ν koji predstavlja gornju granicu udjela anomalija u skupu podataka i donju granicu udjela potpornih vektora.

2.3.5. Isolation Forest (IF)

IF (Liu et al., 2012) je ansambl od t stabala odluke generiranih slučajnim particioniranjem skupa podataka. Stabla se grade do predodređene dubine ili dok u particiji ne preostane jedan od ψ podataka korištenih za izgradnju tog stabla. Prilikom detekcije anomalija prosječna dubina stabala $\mathbb{E}[h(\mathbf{x})]$ kod pretrage stablom odluke određuje je li primjer anomalija ili nije: kraći put označava da je primjer anomalija jer nije bilo puno sličnih primjera u skupu za treniranje koji bi na tom mjestu produbili stablo da bude potrebno donijeti više odluka.

$$s(\mathbf{x}, \psi) = 2^{-\frac{\mathbb{E}[h(\mathbf{x})]}{c(\psi)}} \tag{2.31}$$

s prosječnom duljinom puta neuspješne pretrage binarnog stabla pretraživanja:

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi & ;\psi > 2, \\ 1 & ;\psi = 2 \\ 0 & ;\psi = 1 \end{cases} \tag{2.32}$$

gdje je $H(i)$ i-ti harmonijski broj.

2.4. Provedba eksperimenta

Radi što bolje usporedbe gore opisanih metoda na raspolaganju su im bile slične količine resursa. Svi enkoderi, dekoderi ili generatori te diskriminatori imaju jednaku arhitekturu do na aktivacijsku funkciju zadnjeg sloja. Pokretane su nad jednako pretprocesiranim podatcima, s jednakim brojem latentnih varijabli (kod metoda koje ih koriste) i jednakim maksimalnim brojem epoha (1024). Iznos i strategija ažuriranja stope učenja variraju između pristupa radi postizanja što boljih rezultata svakim od njih.

Hiperparametri dubokih modela (poput arhitektura mreža) podešavani su ručno, a za IF i OCSVM korištena je iscrpna pretraga po rešetci (engl. *grid search*).

Konačna arhitektura mreže je potpuno povezana mreža sa slojevima veličine 128, 64 i 32 za enkoder, a obrnuto za dekoder, odnosno generator. Latentna varijabla ima 8 dimenzija. Diskriminator je također potpuno povezana mreža, sa slojevima veličine 16, 8 i 1. Značajke za f-AnoGAN dobivaju se iz predzadnjeg sloja. Između svih slojeva je aktivacijska funkcija „LeakyReLU“, koja vrijednosti manje od nule množi s 0.2, a veće ostavlja kakve jesu.

Distribucije kod VAE i IWAE modela su: multivarijatna normalna na izlazu enkodera, normalna uz dijagonalnu matricu varijance na izlazu dekodera te standardna normalna kao apriorna distribucija latentne varijable.

IF koristi 4 značajke i 2048 transakcija za izradu 64 stabla kod „MLG-ULB“ skupa podataka, a na „Vesta“ skupu podataka korišteno je 512 stabala koja su izgrađena na temelju 4 značajke i 256 transakcija.

OCSVM je treniran uz toleranciju 10^{-3} uz $\nu = 0.47$ na „MLG-ULB“ skupu podataka. Tolerancija na „Vesta“ skupu podataka je 10^{-8} i $\nu = 0.15$.

2.5. Evaluacija

U literaturi postoje različite mjere kvalitete rezultata, većina počiva na vrijednostima iz „matrice zabune“ 2.1.:

Tablica 2.1. Matrica zabune

		stvarno	
predviđeno		stvarno pozitivni (**TP**)	lažno pozitivni (**FP**)
		lažno negativni (**FN**)	stvarno negativni (**TN**)

Točnost

Metoda opisana u uvodu, koja se računa dijeljenjem broja točno (ne)označenih primjera s ukupnim brojem primjera neovisno o veličinama ciljnih klasa ipak je često korištena u literaturi vezanoj za OPKK. Zove se točnost i pomoću elemenata matrice zabune može se zapisati na sljedeći način:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2.33}$$

Odziv

Korisnija mjera kod neuravnoteženih ciljnih klasa je odziv(2.34). U sklopu OPKK odgovara na pitanje „od svih prijevara, koliko ih je prepoznato kao prijevare?“, no kod odziva je problem što u njega nije uračunata količina pogrešno označenih transakcija koje nisu prijevare. Algoritam koji sve transakcije označava kao prijevare nije naročito koristan, jer ako služi kao filtar nakon kojeg slijedi ručna provjera, neće biti dovoljno ljudi da tu provjeru provedu, a i njima će biti teže zaključivati.

$$\frac{TP}{TP + FN} \tag{2.34}$$

Kod algoritama koji poredaju transakcije prema nekoj mjeri nerijetko se koristi *odziv@k*, mjera koja vraća odziv u slučaju označavanja k najviše rangiranih primjera kao prijevare. Primjerice, kad su ograničeni resursi za daljnju obradu na k transakcija, ova mjera daje udio prijevara među k „najkritičnijih“ transakcija koje će biti označene i dodatno obrađene.

Preciznost

Udio prijevara među svim transakcijama koje su tako označene naziva se preciznost. U nju nije uračunat broj prijevara koje nisu označene kao prijevare. Algoritam koji samo jednu transakciju ispravno označi kao prijevaru (na primjer jer je to po nekom kriteriju trivijalno), a sve ostale ne označi nije koristan — iako je u tom slučaju ova mjera maksimalna.

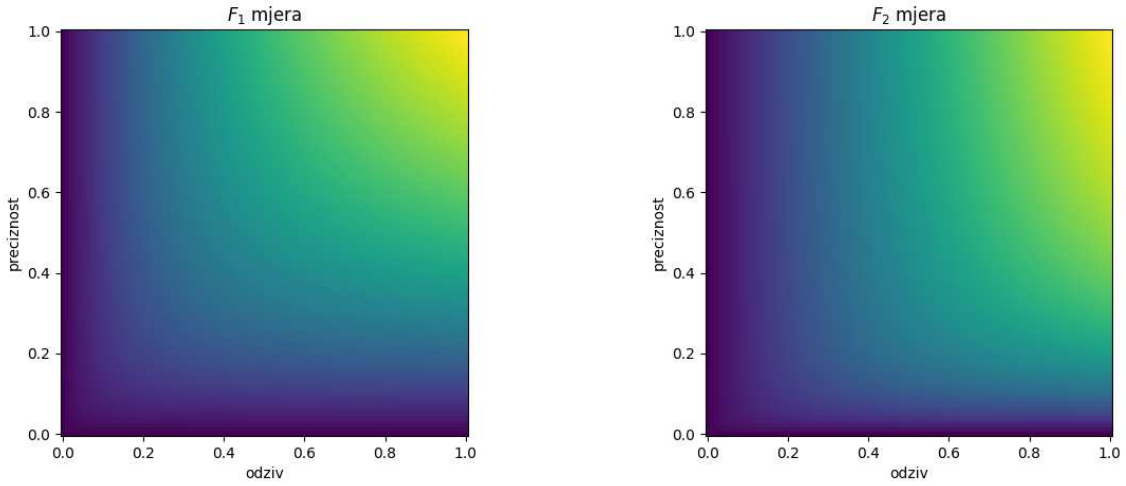
$$\frac{TP}{TP + FP} \tag{2.35}$$

Kao kod odziva, kod preciznosti postoji *preciznost@k*.

F_1 i F_β mjera

Odziv i preciznost očito su obje korisne, ali same za sebe imaju manjkavosti. Iz navedenih primjera daje se zaključiti da korist jedne odgovara manjkavosti druge i obratno. Zato je još korisnija mjera koja bi bila niska u slučaju kad je bar jedna od te dvije mjere niska, a kad su obje visoke da bude i ta mjera visoka, odnosno harmonijska sredina preciznosti i odziva. Ta mjera naziva se F_1 mjera. U slučaju da je odziv β puta važniji od preciznosti ili obrnuto, jednu od tih mjera moguće je otežati te se to zove F_β mjera:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{preciznost} \cdot \text{odziv}}{(\beta^2 \cdot \text{preciznost}) + \text{odziv}} \tag{2.36}$$



Slika 2.11. F_1 i F_2 mjera: vizualna usporedba doprinosa preciznosti i odziva

Matthew's correlation coefficient (MCC)

Jedna od prednosti (Chicco i Jurman, 2020) MCC nad F_1 mjerom je uračunavanje broja stvarno negativnih primjera (TN).

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.37}$$

Površina pod ROC krivuljom (AUC ROC)

Za algoritme koji mogu poredati transakcije prema tome koliko su vjerojatno prijevare (ili koliko odudaraju od ostalih ili koliko su kritične...), može se definirati krivulja rasta odziva (prepoznatih prijevara u svim prijevarama) u odnosu na broj legitmnih transakcija označenih kao prijevare ($FPR = FP/(FP + TN)$) s obzirom na poredak kojim se transakcije odabiru iz skupa svih transakcija. Krivulja raste kad su poredane transakcije zaista prijevare, a stagnira kad nisu. Što je više prijevara prije legitimnih transakcija, krivulja prije raste, pa je površina pod njom veća. Maksimalnu površinu pod ROC krivuljom imaju algoritmi koji mogu odvojiti bitnije transakcije od ostalih, čime prvo maksimalno povećaju odziv, odnosno visinu krivulje, a tek onda krivulja počne stagnirati. Nasumičnim označavanjem (bacanje novčića) dobiva se AUC ROC $= 0.5$.

Površina ispod krivulje preciznosti i odziva (AUC PRC), prosječna preciznost (AP)

Iako je površina pod ROC krivuljom česta u literaturi, za podatke s nesrazmjerom ciljnih klasa bolja je površina ispod PRC krivulje (Saito i Rehmsmeier, 2015). Ta krivulja pada kad su transakcije označene kao prijevare zapravo legitimne, a raste kad raste odziv (udio prijevara u svim transakcijama označenima kao prijevare). Što dulje krivulja stagnira bez padanja, površina pod njom je veća. Mana ove mjere je što nema jasne referentne vrijednosti, kao što je kod AUC ROC očekivana vrijednost za nasumični klasifikator 0.5 (Ruff et al., 2021).

Čest robustan način izračuna površine ispod krivulje preciznosti i odziva je korištenjem prosječne preciznosti (Boyd et al., 2013). Za preciznost i odziv kod označavanja

prvih i najbolje rangiranih transakcija kao prijevare formula je (Mekterović et al., 2018):

$$\text{AP} = \sum_i (\text{odziv}_i - \text{odziv}_{i-1})\text{preciznost}_i \tag{2.38}$$

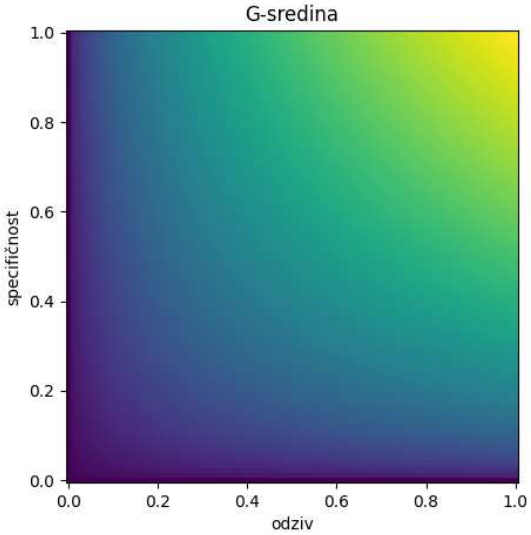
Napomena: PRC krivulja leži na istim osima kao gore prikazani dijagrami mjera F_1 i F_2 2.11. Dakle, prilikom usporedbe metoda korisno je znati da točka PRC krivulja najbliža gornjem desnom kutu označava najbolju F_1 mjeru.

G-sredina

Analogno odzivu, mjera kojom se računa udio transakcija koje ispravno *nisu* označene kao prijevare među svim legitimnim transakcijama naziva se specifičnost. Geometrijska sredina specifičnosti i odziva također se pokazala kao korisna mjera kod nesrazmjera ciljnih klasa za izbjegavanje prenaučenosti na primjere veće klase i podnaučenosti kod primjera manje klase (Akosa, 2017).

$$\text{G-sredina} = \sqrt{\frac{TP \cdot TN}{(TP + FN)(TN + FP)}} \tag{2.39}$$

U literaturi postoje i druge mjere, ali rjeđe se koriste, pa će se u sklopu ovog rada koristiti



Slika 2.12. Geometrijska sredina specifičnosti i odziva

samo gore navedene. Kao temeljna mjera odabrana je AP jer uzima u obzir sve podjele podataka, gdje god da se podvuče granica. Radi lakše usporedbe s drugim istraživanjima koriste se ROC AUC i F_1 mjera uz granicu odabranu na temelju maksimuma na skupu

podataka za treniranje.

3. Rezultati i rasprava

3.1. Rezultati

3.1.1. Rezultati za „MLG-ULB“ skup podataka

VAE

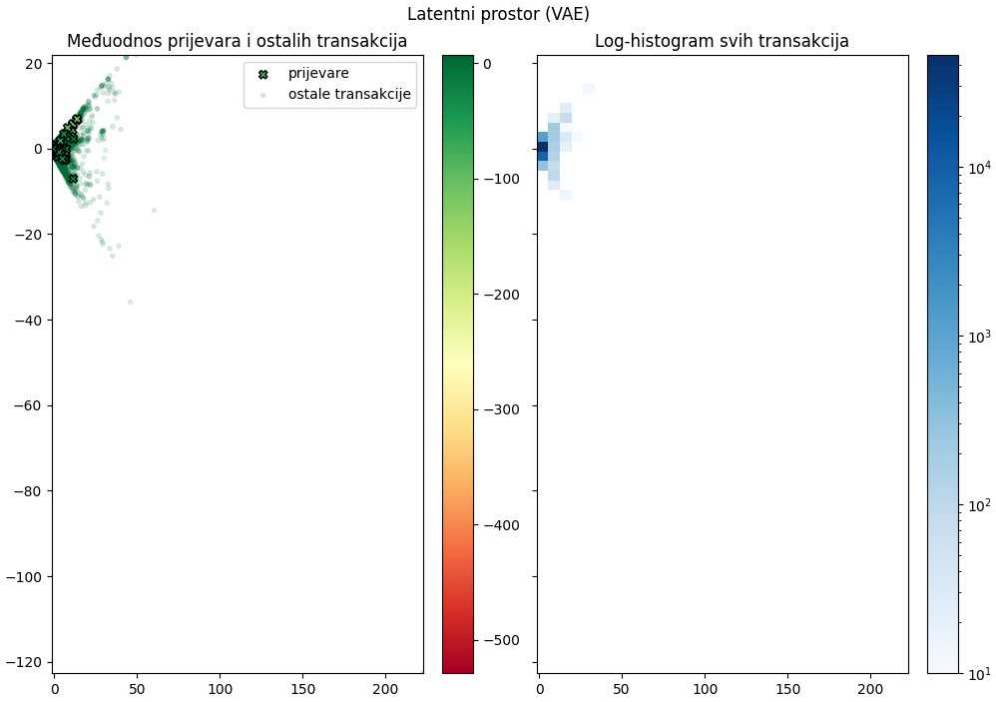
Na dijagramu dobivenom PCA transformacijom latentnog prostora VAE (Slika 3.1.) pojedine transakcije puno više odskaču nego u podatkovnom prostoru, a prijevare nisu jasno odvojene od ostalih transakcija iako to je bio slučaj u podatkovnom prostoru (Slika 2.2.).

Te stršeće vrijednosti mogu se vidjeti na violinskom dijagramu gustoće raspodjele log-vjerojatnosti rekonstrukcije (Slika 3.2.) kao dugački repovi raspodjele legitimnih transakcija, što objašnjava i zašto su teže uočljive: legitimne transakcije prozirnije su kako bi se bolje stekao dojam gustoće raspodjele i u latentnom prostoru te kako bi prijevare, kojih ima manje i bitnije su, bile uočljivije. Zbog spomenutog dugačkog repa, koji se pruža ispod svih rekonstrukcijskih vjerojatnosti prijevara, očekuje se drastičan pad na samom početku PRC krivulje.

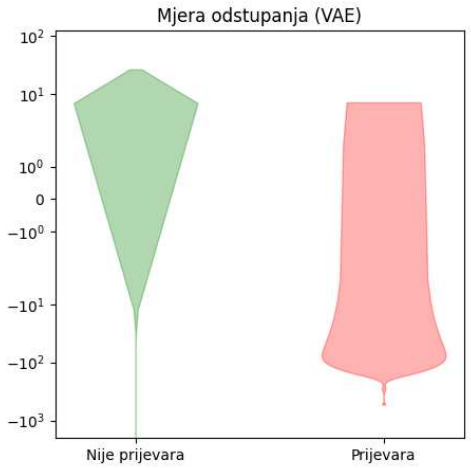
Na log-histogramu (Slika 3.1.) vidljivo je da je velik dio prijevara u području najveće gustoće, što odgovara relativno širokoj bazi oko nule violinskog dijagrama. Drugim riječima, rekonstrukcijska vjerojatnost dijela prijevara je visoka, što je slučaj i za velik broj legitimnih transakcija, pa se očekuje niska preciznost prije nego odziv dođe do 1. Zato će oni koji žele označiti sve prijevare pritom označiti i velik broj legitimnih transakcija zbog čega će PRC krivulja biti nisko i uz desni rub.

Ipak, iz toga što je većina raspodjele *prijevara* svojom rekonstrukcijskom vjerojatnošću *ispod* najvećeg dijela raspodjele *legitimnih transakcija*, većinu prijevara moguće je odvojiti od najvećeg broja legitimnih transakcija, dakle za velik broj prijevara preciznost

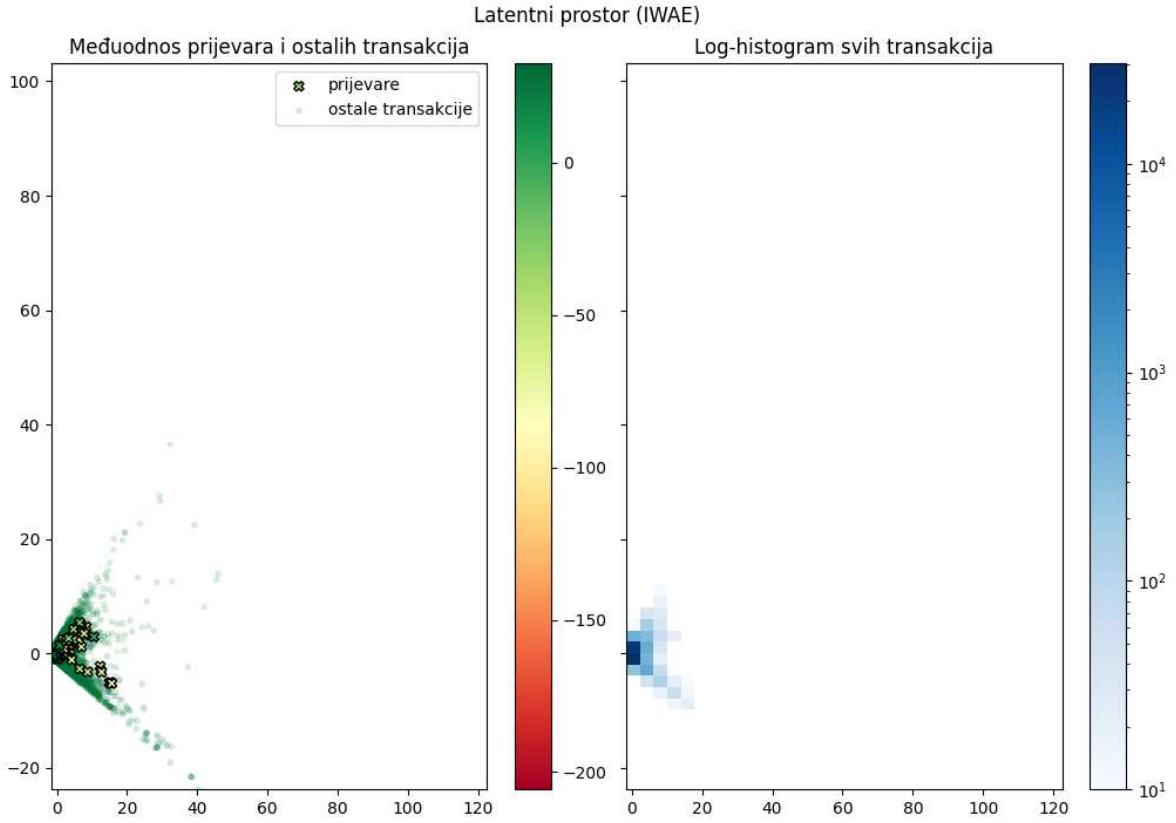
je na razini puno, puno boljoj nego da su u igri i desetci tisuća legitimnih transakcija iz baze raspodjele legitimnih transakcija. Drugim riječima, pozitivno je što crvena površina ima zadebljanje dolje, a zelena suženje te što crvena površina cijela ima veći odmak prema dolje. Takav oblik i odnos raspodjela je indikator velike površine pod ROC krivuljom.



Slika 3.1. Skup podataka „MLG-ULB“ u latentnom prostoru VAE



Slika 3.2. Logaritam rekonstrukcijske vjerojatnosti VAE



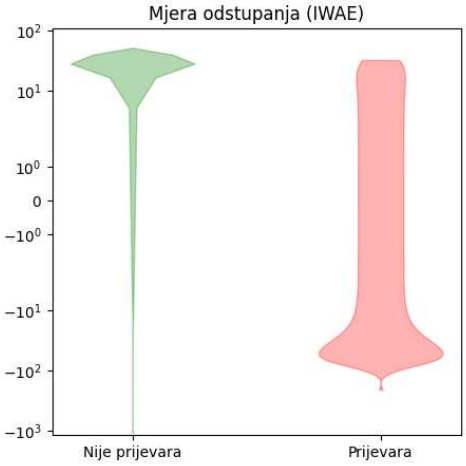
Slika 3.3. Skup podataka „MLG-ULB“ u latentnom prostoru IWAE

IWAE

Raspodjela u latentnom prostoru (Slika 3.3.) slična je za IWAE kao što je bio slučaj kod VAE.

Violinski dijagram isprva izgleda bolje nego kod VAE, ali valja primjetiti da raspodjela legitimnih transakcija ima širu bazu, dakle proširenje u repu koje izgleda usko u odnosu na deseke tisuća transakcija iz baze ima velik apsolutan broj transakcija u odnosu na broj prijevara s istom rekonstrukcijskom vjerojatnošću. To znači da preciznost@k neće biti skroz oko 1 — ne samo za mali broj označenih transakcija k, tj. kod niske rekonstrukcijske vjerojatnosti (zahvaljujući dužem repu raspodjele legitimnih transakcija) — nego ni za bilo koji razuman k. Iz toga se da zaključiti da AP mjera neće biti velika, no za to ustanoviti bolje je pogledati PRC dijagram (Slika 3.15.) jer iz ovog je dijagrama to teže procijeniti.

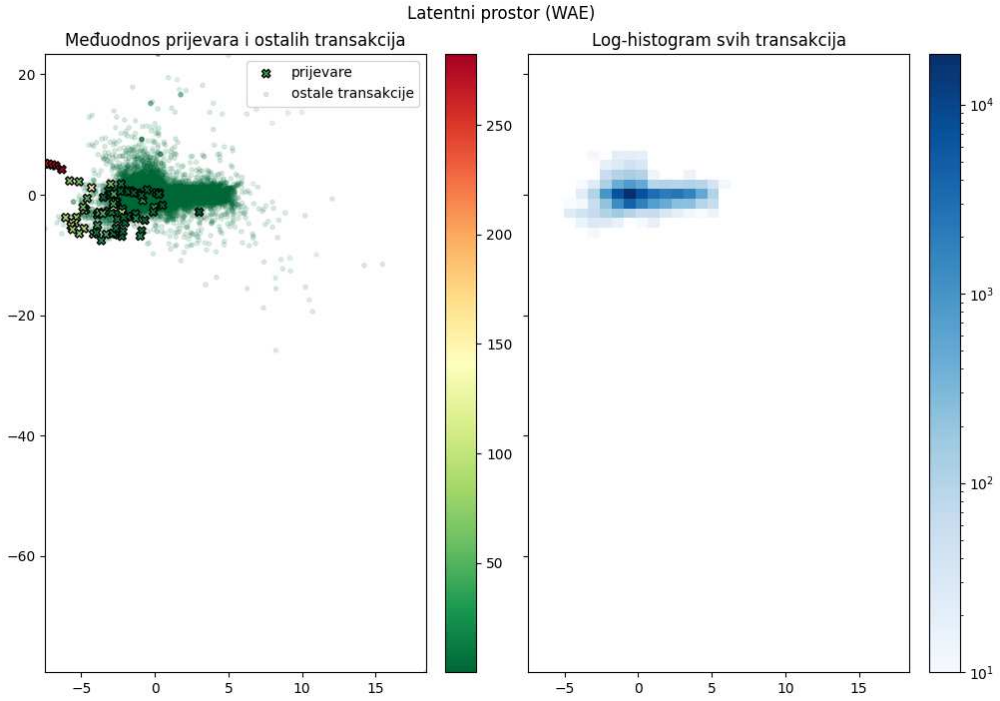
Zanimljiva je konzistentno uska baza raspodjele rekonstrukcijske vjerojatnosti prijevara. Ona označava bolju preciznost pred kraj (ali ne nužno i na samom kraju) PRC krivulje, kad je velik broj *prijevara* (ne svih transakcija) označen. To je slučaj kad je F_2



Slika 3.4. Logaritam rekonstrukcijske vjerojatnosti IWAE

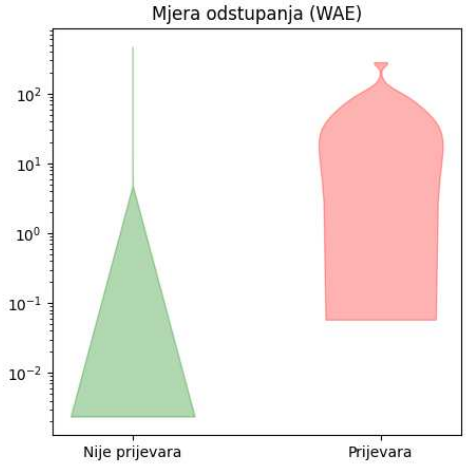
mjera visoka, a primjerice AUC ROC i AP su niže zbog pogrešnog poretka svih označenih transakcija.

WAE



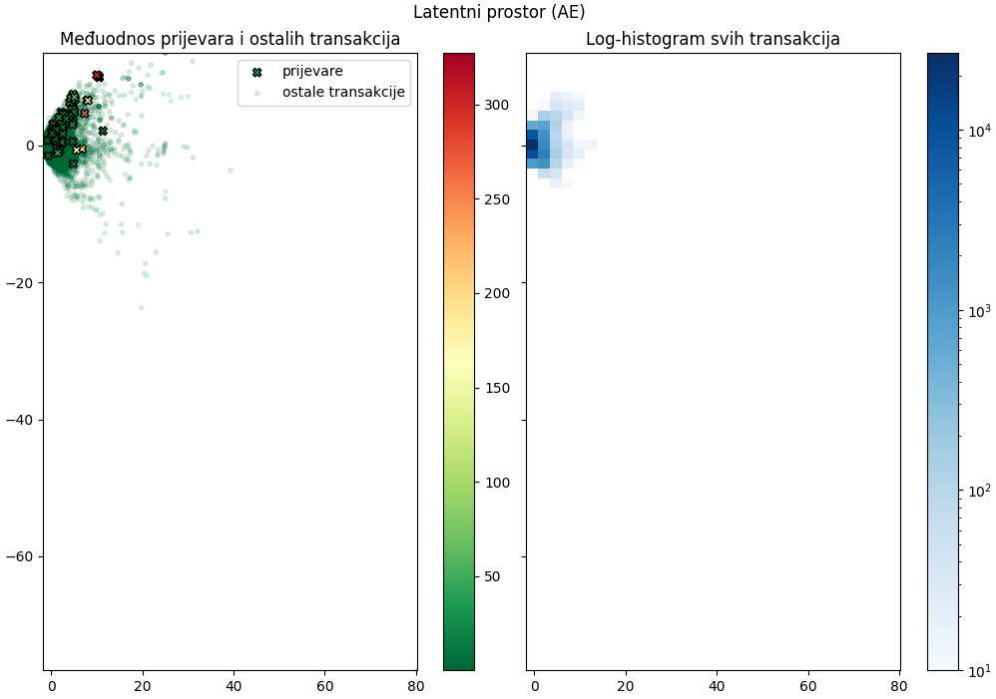
Slika 3.5. Skup podataka „MLG-ULB“ u latentnom prostoru WAE

Kod WAE se bolje vidi podjela u latentnom prostoru nego kod prethodnih modela, što je bio i cilj. Stršeće su vrijednosti manje udaljene. Pomak među raspodjelama rekonstrukcijske pogreške puno je veći nego kod prethodnih modela, što znači da će i krivulje PRC i ROC imati veće površine.



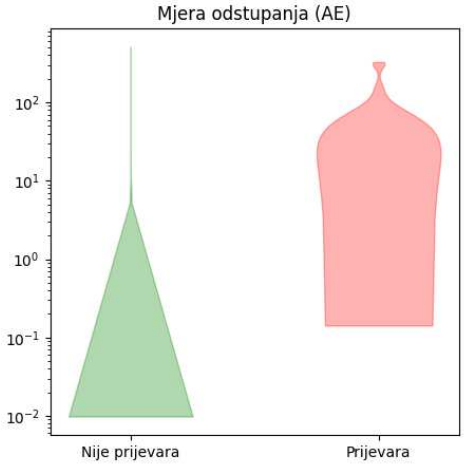
Slika 3.6. Rekonstrukcijska greška WAE

AE



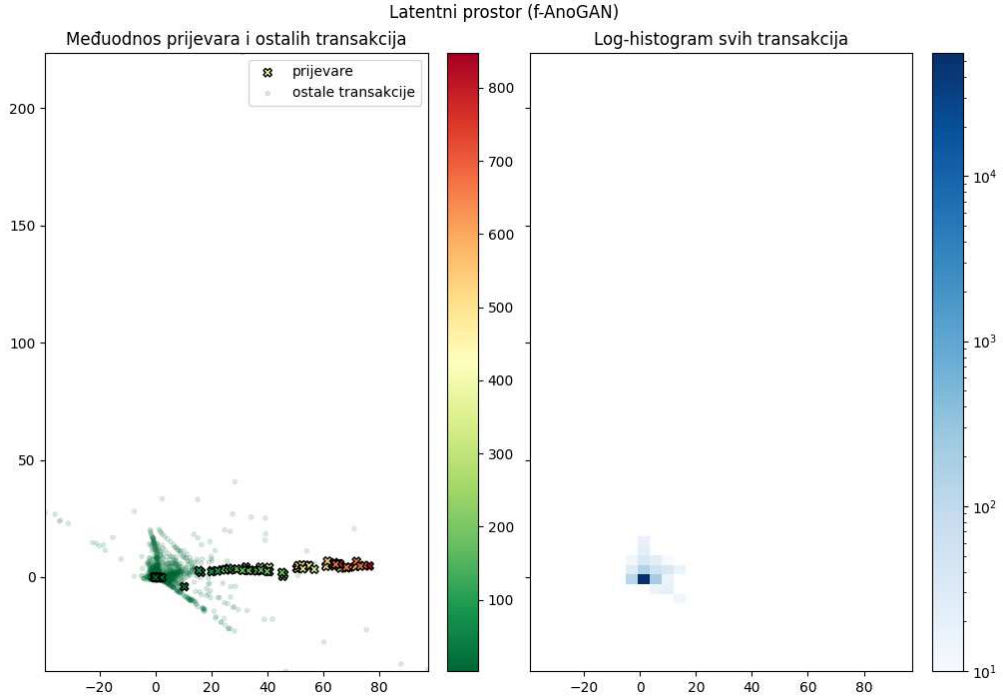
Slika 3.7. Skup podataka „MLG-ULB“ u latentnom prostoru AE

Raspodjela u latentnom prostoru ovog modela slična je raspodjeli VAE, a volinske krivulje sličnije su WAE, ali pogledom na skalu rekonstrukcijske pogreške vidljivo je da WAE bolje generira transakcije. Za usporedbu s VAE i IWAE potrebno je pogledati dijagrame ROC i PRC krivulja na slici 3.15.



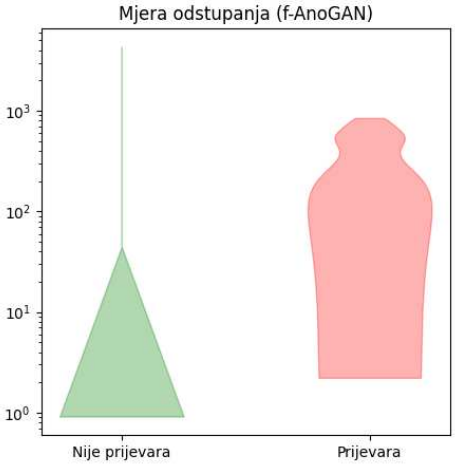
Slika 3.8. Rekonstrukcijska greška AE

f-AnoGAN



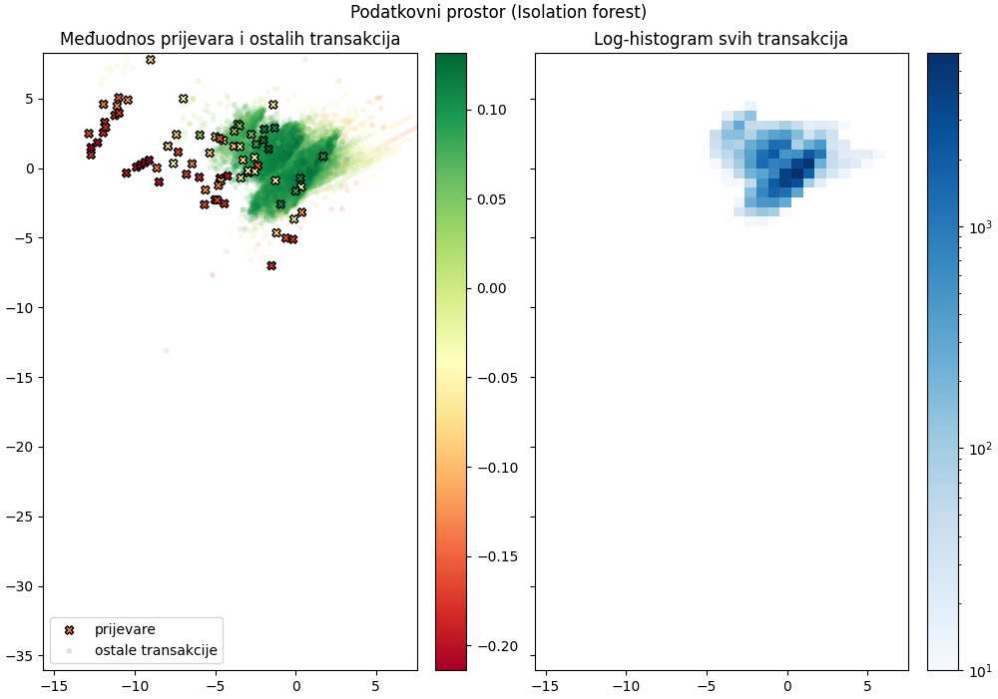
Slika 3.9. Skup podataka „MLG-ULB“ u latentnom prostoru fAnoGAN-a

Za f-AnoGAN violinski dijagram (Slika 3.10.) ne pokazuje naročitu razliku u odnosu na neke prethodne modele, ali zato je usporedbom log-histograma s dijagramom raspršenja (Slika 3.9. jasno). Os najvećeg raspršenja (glavna komponenta skupa podataka) u latentnom prostoru upravo je ona koja dijeli velik dio prijevara od ostalih transakcija. Ipak, iz violinskog se dijagrama daje zaključiti da je u području najveće gustoće i dalje relativno velik broj prijevara koje se na dijagramu raspršenja preklapaju, tako da AP ipak neće biti blizu 1.



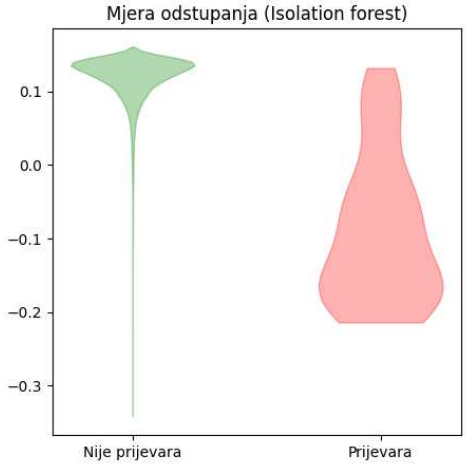
Slika 3.10. Rekonstrukcijska greška f-AnoGAN-a

IF



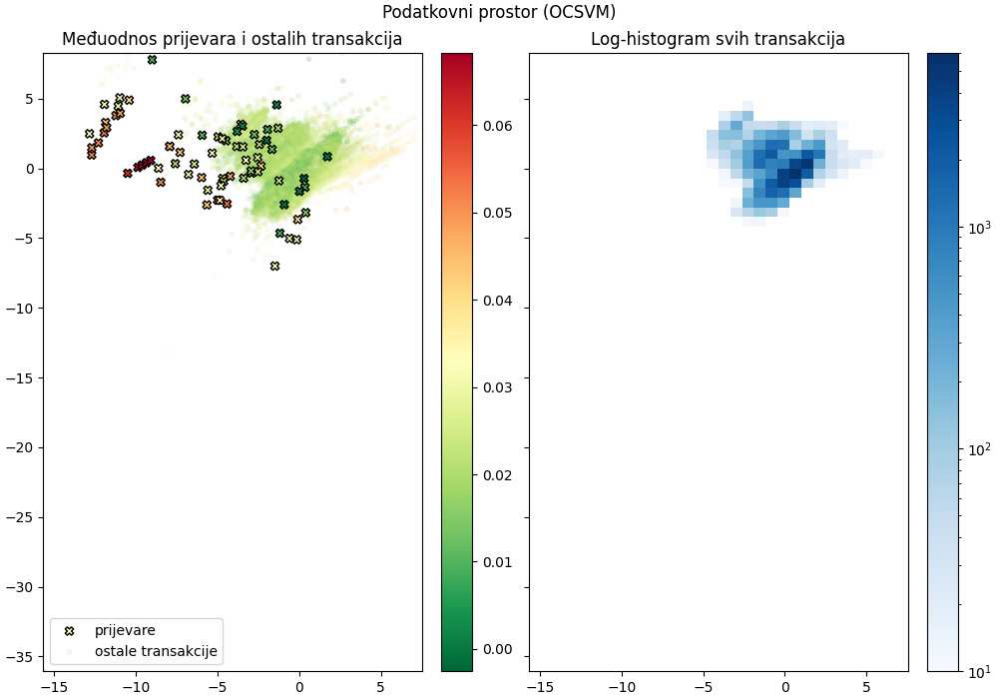
Slika 3.11. Skup podataka „MLG-ULB“ s dodijeljenim mjerama odstupanja IF

Za IF prikazan je podatkovni prostor umjesto latentnog na slici 3.11. Kao i kod IWAE, širina baze raspodjele mjere odstupanja prikriva širinu repa, pa zato i izgled PRC krivulje. Ipak, ovaj je dijagram dobar indikator izgleda ROC krivulje, a ona je za ovakav odnos i oblik raspodjela jako visoka, kao što je opisano i kod rezultata VAE (zadnji odlomak dijela 3.1.1.).



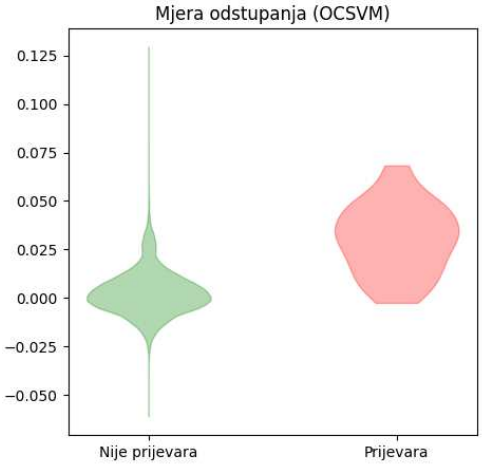
Slika 3.12. Raspodjela mjera odstupanja IF prema ciljnoj klasi

OCSVM



Slika 3.13. Skup podataka „MLG-ULB“ s dodijeljenim mjerama odstupanja OCSVM

Kod OCSVM više od pola raspodjele mjere odstupanja kod prijevara je u razini s proširenjem u raspodjeli mjere odstupanja legitimnih transakcija, što znači da je ovom metodom i velik broj legitimnih transakcija nužno označiti kao prijevare za korisnu razinu odziva. Uz to, opet postoje podatci s većom mjerom odstupanja nego što ju ima bilo koja prijevara (stršeće vrijednosti). Zato će se pokazati da, iako postoji pomak među raspodjelama, nije dovoljan da preciznost dosegne razinu veću od 0.18 ni za koju granicu mjere odstupanja (engl. *threshold*) prilikom označavanja transakcija.



Slika 3.14. Raspodjela mjera odstupanja OCSVM prema ciljnoj klasi

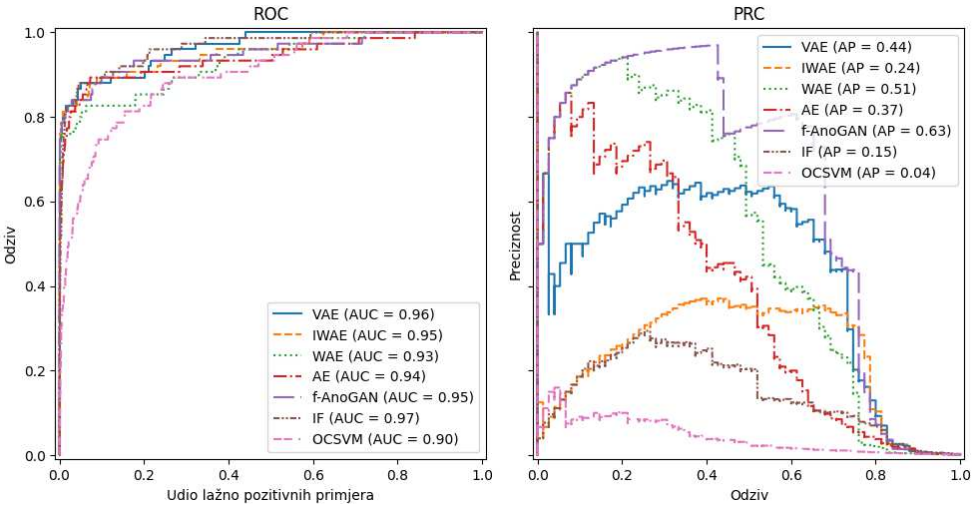
Ukupni rezultati

Tablica 3.1. Iznosi mjera za „MLG-ULB“ skup podataka

	AP	AUC ROC	F_1	MCC	G-sredina	Vrijeme inf.
f-AnoGAN	**0.629917**	0.951389	**0.656250**	**0.665805**	0.806584	0.123525
WAE	0.510689	0.931347	0.482412	0.496688	0.735004	0.086737
VAE	0.438157	0.964113	0.595745	0.596465	0.807528	9.760391
AE	0.371717	0.939052	0.422764	0.455934	0.793433	0.130203
IWAE	0.236367	0.952133	0.410526	0.485957	0.795643	9.048160
IF	0.147820	**0.965376**	0.071429	0.057710	**0.829326**	0.290513
OCSVM	0.043662	0.899512	0.109290	0.074084	0.576764	**0.003349**

Iz PRC krivulje (Slika 3.15.) vidi se da u je u ovom (i ovako pretprocesiranom) skupu podataka otprilike 20% prijevara jako slično velikom broju legitimnih transakcija. Može se reći da za 20% prijevara postoji bar još toliko legitimnih transakcija koje u jednakoj mjeri ili više odudaraju od većine preostalih transakcija na temelju rezultata svih iskušanih metoda (ni jedna PRC krivulja nije na više od 50% preciznosti pri odzivu od 80%).

Ipak, to je i dalje mali udio u ukupnom broju legitimnih transakcija, pa je zato AUC ROC za sve metode vrlo visoka. Taj nesrazmjer najbolje se vidi kod IF: ta metoda bolje poreda onih zadnjih 20% prijevara u odnosu na legitimne transakcije, što u odnosu na ostale metode najviše poveća površinu pod ROC krivuljom, no je li bitno hoće li se zadnjih 10% prijevara naći među 200 000 legitimnih transakcija ili među 100 000 legitimnih transakcija? Taj broj legitimnih transakcija, odnosno udaljenost od desnog ruba dijagrama, drastično utječe na rast površine pod ROC krivuljom za svaku dobro označenu prijevaru. U usporedivoj mjeri utječe na površinu pod ROC krivuljom kao i razlika je li preciznost



Slika 3.15. Krivulje svih metoda

10% ili 90%, što je u ovom kontekstu puno korisnije. Zato je AP bolja mjera nego AUC ROC kod vrednovanja metoda OPKK.

Općenito se f-AnoGAN pokazao kao najbolja metoda za OPKK na ovakvom skupu podataka. Iako je u jednom od nekoliko pokretanja VAE postigao vrlo dobar rezultat, najčešće se pokazalo da KL divergencija kao način regularizacije ne daje jednako dobre rezultate kao ostale metode. Za razliku od VAE koji često daje i lošiji rezultat nego IWAE, IWAE konzistentno daje prikazani rezultat.

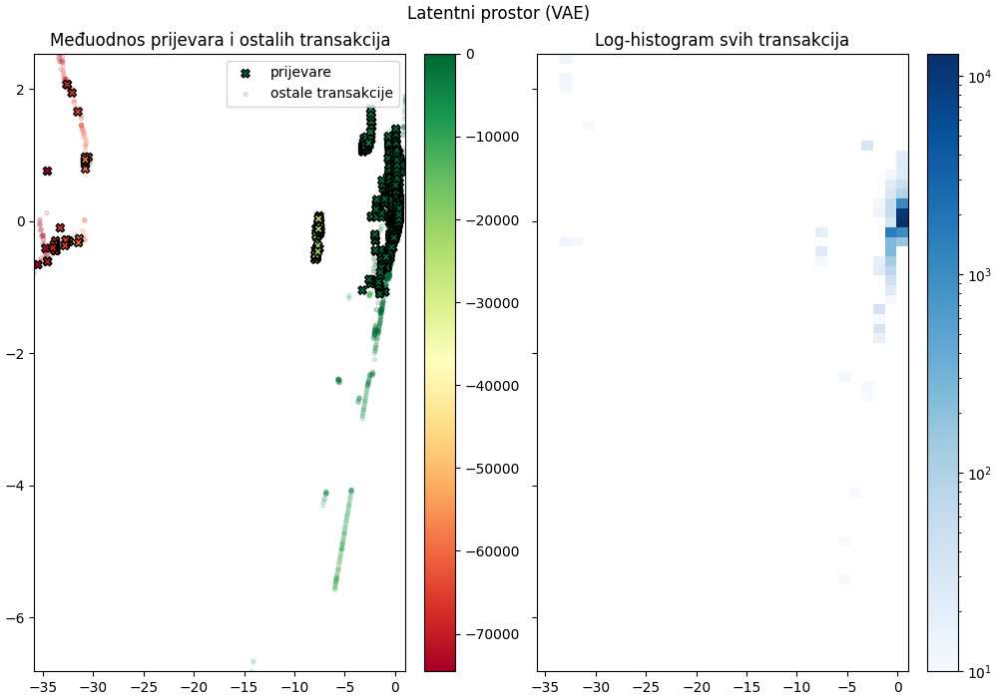
Na ovom skupu podataka inferencija je pokretana na CPU za sve metode i očito su metode koje svoje mjere temelje na uzorkovanju slučajnih brojeva puno sporije. Uz manji broj uzoraka (rezultati su uz $k = 100$) procjena bi bila neprecizinija, ali i vrijeme kraće. Kraće bi bilo i uz pokretanje na GPU, a takva usporedba prikazana je kod drugog skupa podataka.

3.1.2. Rezultati za „Vesta“ skup podataka

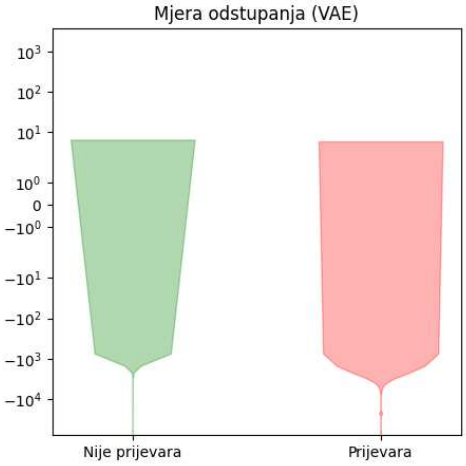
VAE

Za razliku od „MLG-ULB“ skupa podataka, „Vesta“ skup podataka zbog heterogenosti značajki nema samo jednu približno sferičnu grupu transakcija. Stršeće vrijednosti pojedinih značajki nalaze se na pravcu koji prolazi kroz područje najveće gustoće podatka te postoji nekoliko manjih grupa stršećih vrijednosti udaljenih od glavne grupe podataka. To se odražava i u latentnom prostoru, u kojem je transakcijama u središtu glavne grupe i na spomenutom pravcu dodijeljena najveća rekonstrukcijska vjerojatnost te se

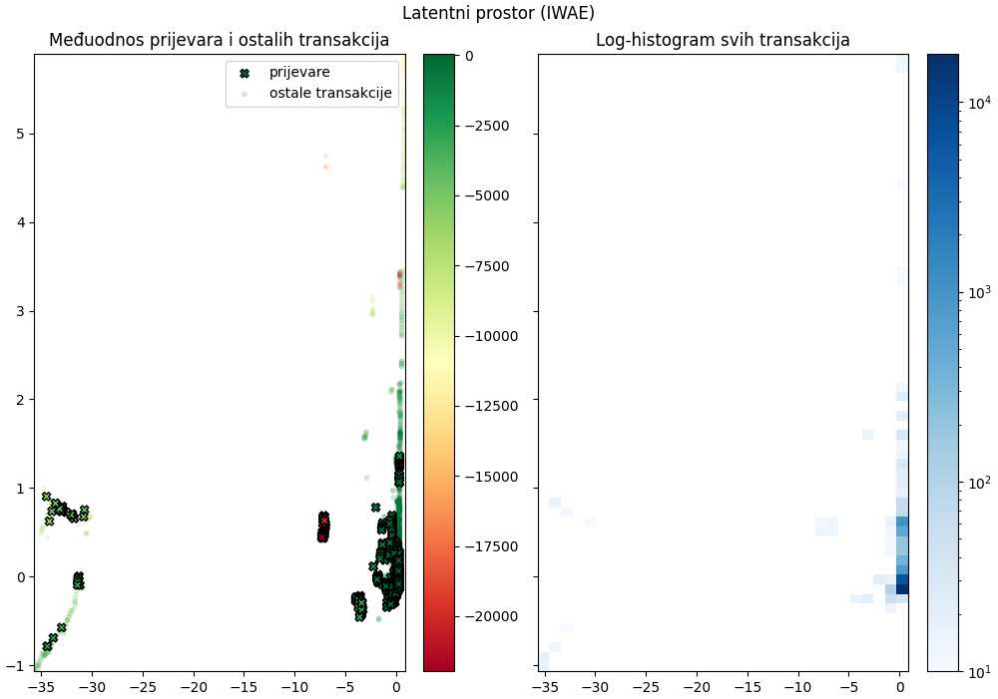
ona smanjuje s udaljenošću od te glavne grupe i od pravca. Budući da i prijevara i legitimnih transakcija ima u spomenutim udaljenim grupama, veliko je preklapanje njihovih rekonstrukcijskih vjerojatnosti, pa iako je u prosjeku rekonstrukcijska vjerojatnost prijevara manja, ovaj model nije koristan u otkrivanju anomalija kod ovakvih podataka.



Slika 3.16. Skup podataka „Vesta“ u latentnom prostoru VAE



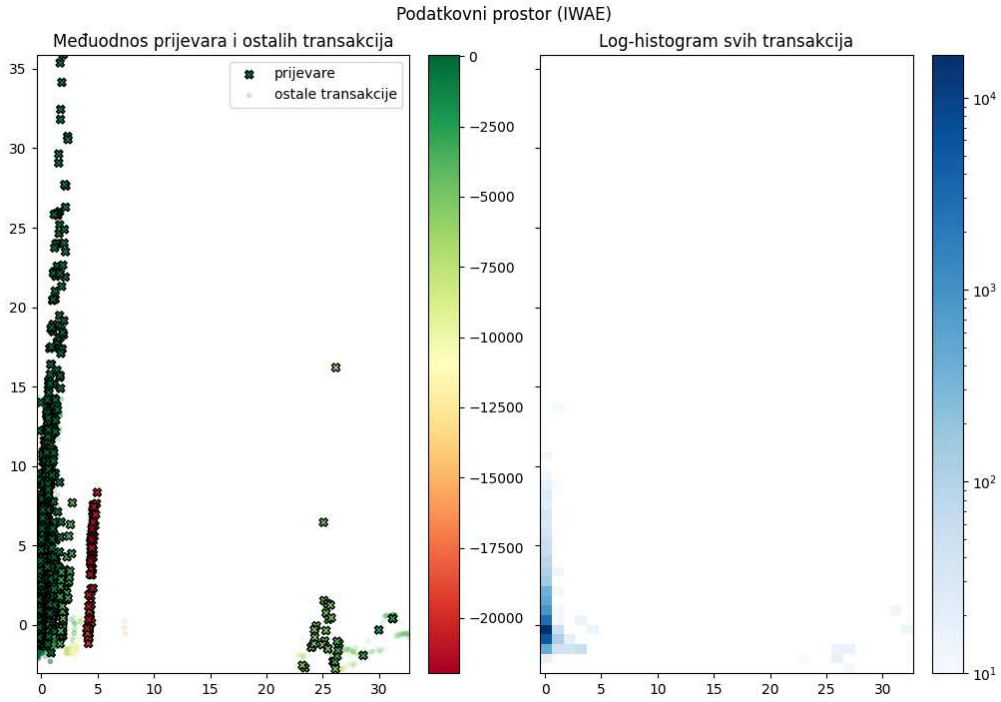
Slika 3.17. Logaritam rekonstrukcijske vjerojatnosti VAE



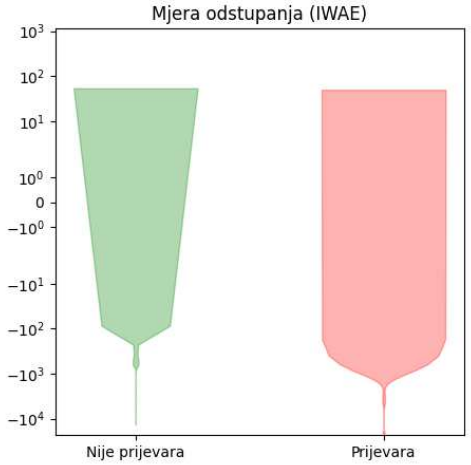
Slika 3.18. Skup podataka „Vesta“ u latentnom prostoru IWAE

IWAE

U latentnom prostoru IWAE je sličan raspored kao u latentnom prostoru VAE (Slika 3.18.), no što se tiče rekonstrukcijske vjerojatnosti, od dvije skupine transakcija odvojene od glavne grupe, ona bliža ima niže rekonstrukcijske vjerojatnosti. Razlog tome je to što ta grupa nije zastupljena u skupu za učenje, što znači da IWAE bolje uspijeva naučiti generirati i jako različite stršeće vrijednosti, a ne samo one bliže glavnom skupu ali nezastupljene u skupu za učenje (usporediti slike 3.19. i 2.4., crveni križići iznad 5 na x-osi izbačeni su iz skupa za učenje, a zelenih točaka baš i nema na tom mjestu). Violinske krivulje pokazuju preklapanje u rekonstrukcijskim vjerojatnostima, što je i za očekivati uz brojnost prijevara u najvećoj grupi..

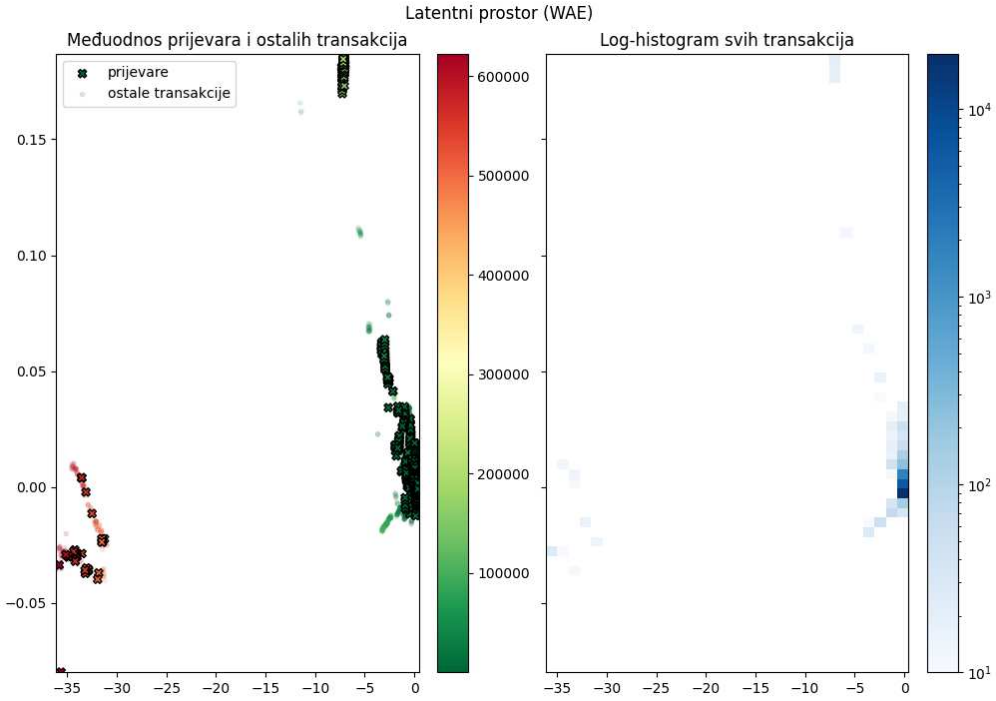


Slika 3.19. Skup podataka „Vesta“ u podatkovnom prostoru s dodijeljenim rekonstrukcijskim vjerojatnostima IWAE

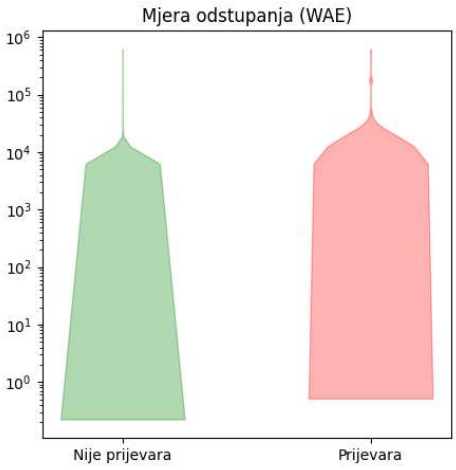


Slika 3.20. Logaritam rekonstrukcijske vjerojatnosti IWAE

WAE



Slika 3.21. Skup podataka „Vesta“ u latentnom prostoru WAE

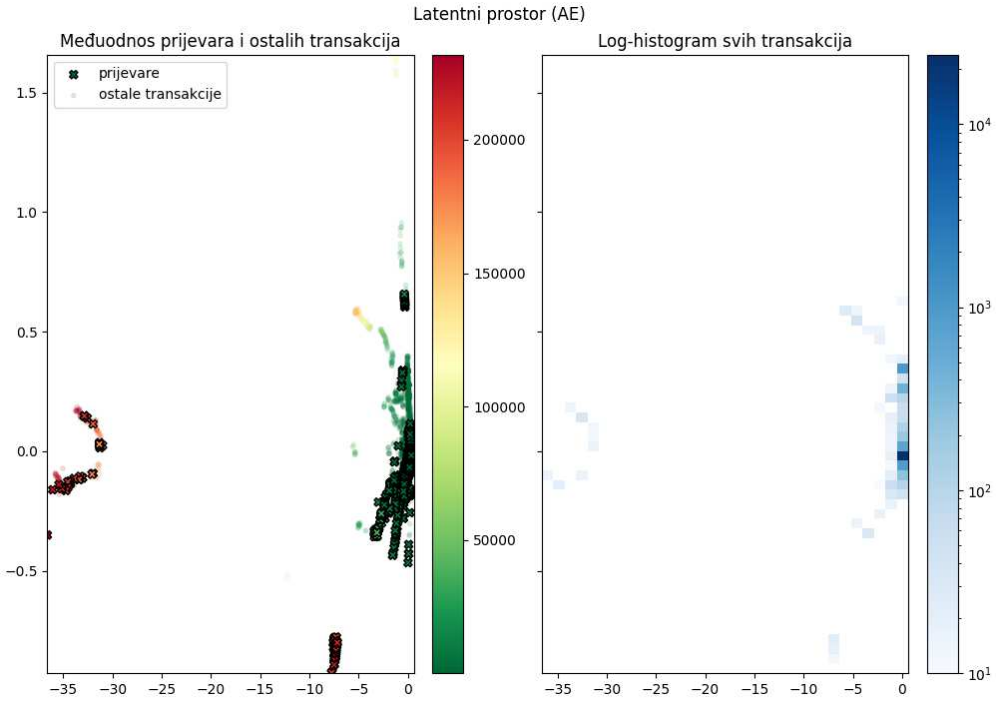


Slika 3.22. Rekonstrukcijska greška WAE

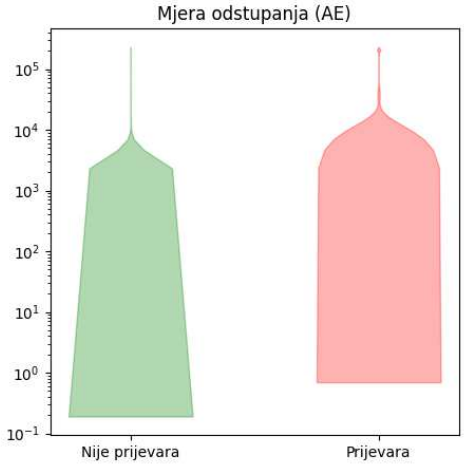
Za WAE vrijedi isto što i za VAE, čak im i latentni prostori izgledaju slično.

AE

Običan AE visoke iznose srednje kvadratne pogreške daje i bližoj i daljoj grupi u latentnom prostoru, a ne samo daljoj, no rezultati su slični VAE i WAE.

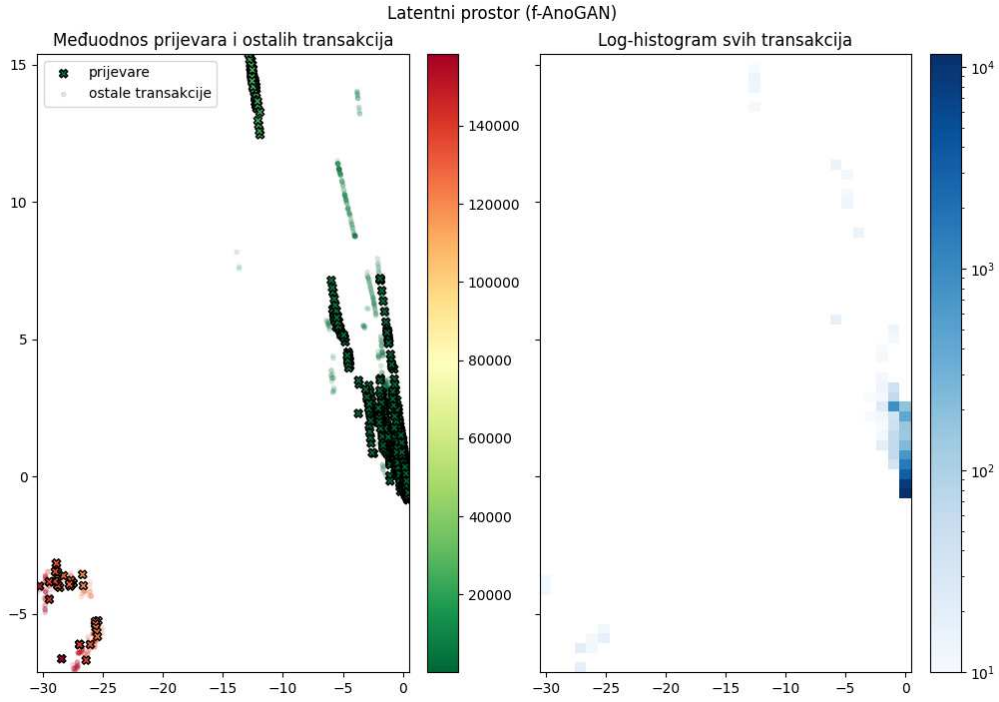


Slika 3.23. Skup podataka „Vesta“ u latentnom prostoru AE

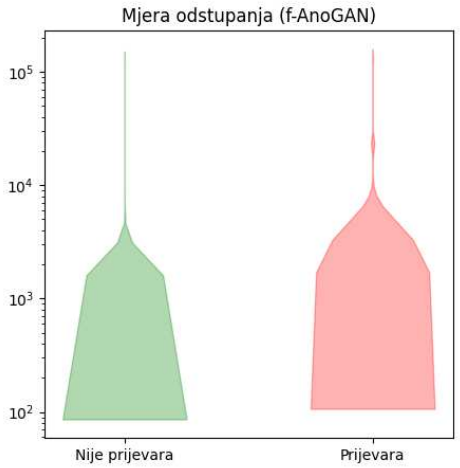


Slika 3.24. Rekonstrukcijska greška AE

f-AnoGAN



Slika 3.25. Skup podataka „Vesta“ u latentnom prostoru fAnoGAN-a

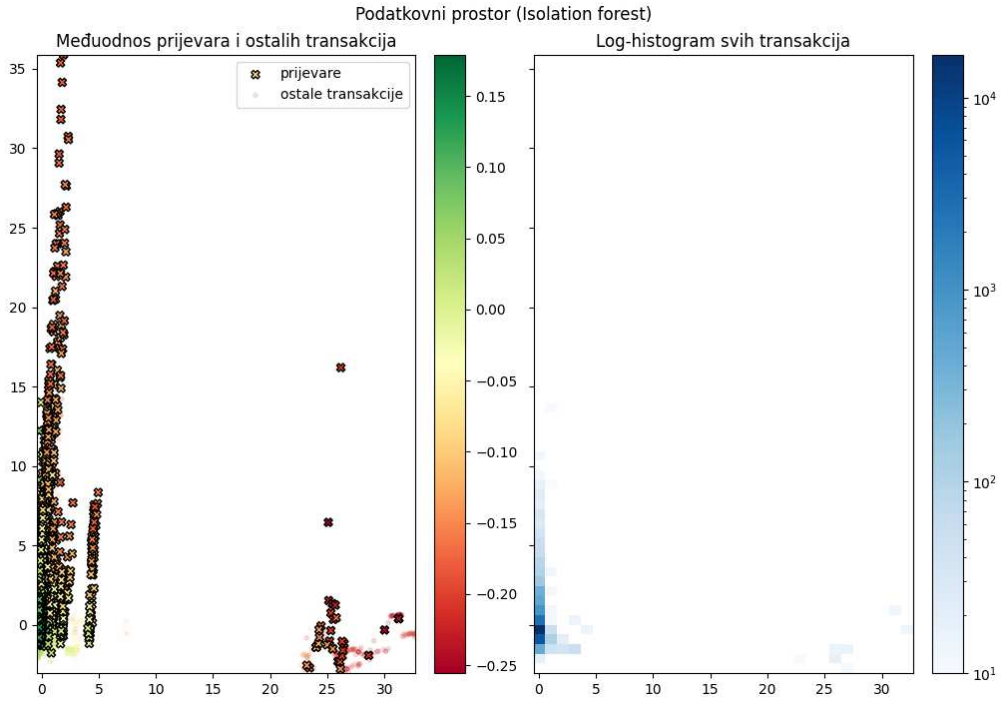


Slika 3.26. Rekonstrukcijska greška f-AnoGAN-a

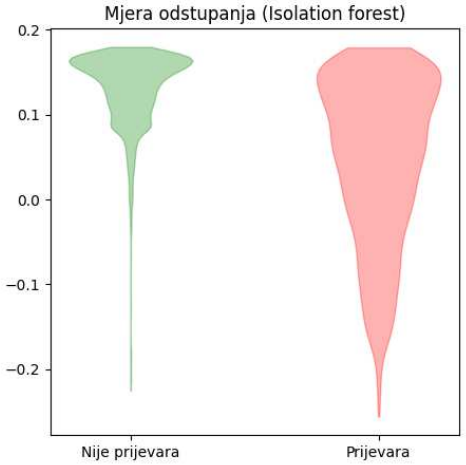
Kod f-AnoGAN-a se opet može vidjeti dodjela najveće greške najudaljenijoj skupini.

IF

Za razliku od prethodnih metoda kod kojih sve transakcije u najvećoj grupi imaju nisku mjeru odstupanja čak i uz veliko raspršenje unutar grupe, kod IF mjera odstupanja izgleda slično mjeri udaljenosti transakcije od nule, tj. od mjesta najveće gustoće podataka. Na violinskom dijagramu vidi se da je velik udio prijevara s većim odstupanjem od većine legitimnih transakcija, a repovi su kratki.

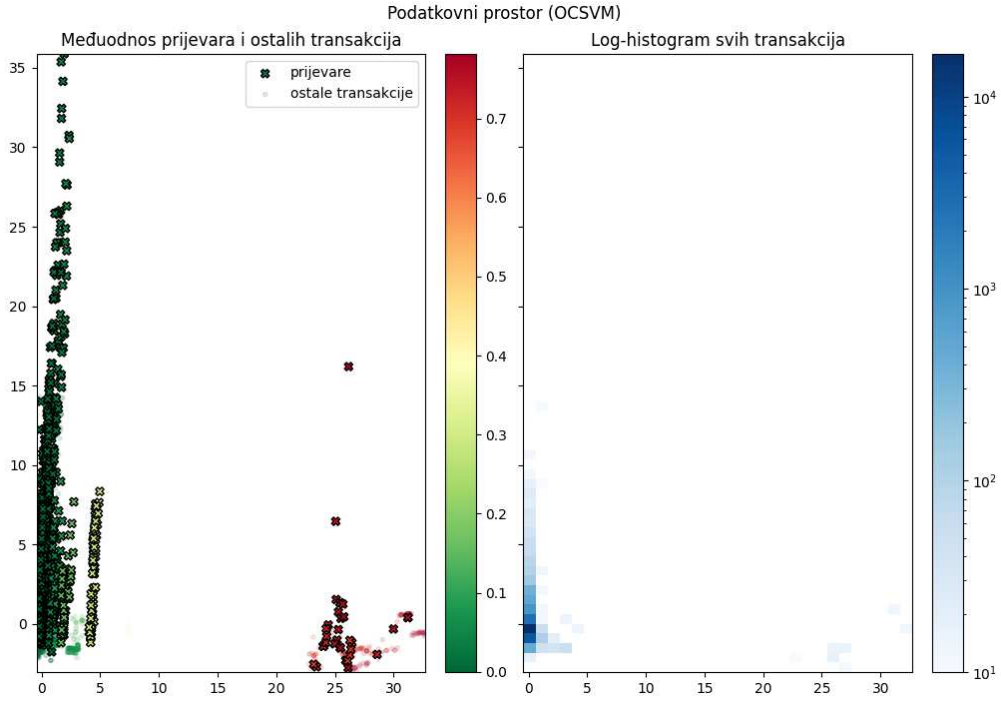


Slika 3.27. Skup podataka „Vesta“ s dodijeljenim mjerama odstupanja IF

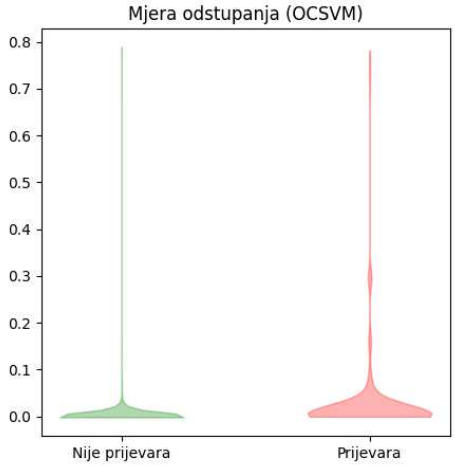


Slika 3.28. Raspodjela mjera odstupanja IF prema ciljnoj klasi

OCSVM



Slika 3.29. Skup podataka „Vesta“ s dodijeljenim mjerama odstupanja OCSVM



Slika 3.30. Raspodjela mjera odstupanja OCSVM prema ciljnoj klasi

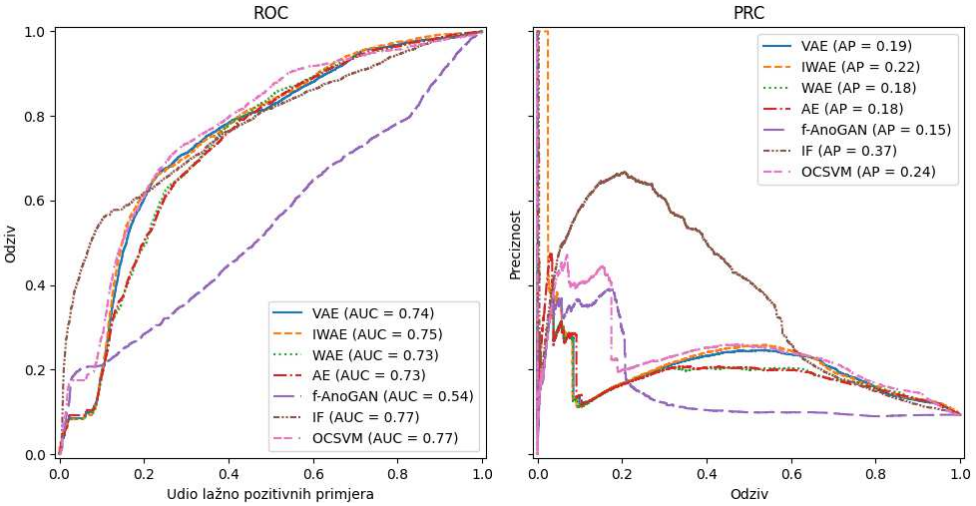
Kod OCSVM raspodjela mjera odstupanja slična je onoj VAE, WAE, i AE.

Ukupni rezultati

Klasične metode otkrivanja anomalija koje su na prošlom skupu podataka imale izrazito lošu AP, na ovakvom su skupu podataka imale puno bolji rezultat. Zbog manjeg nesrazmjera ciljnih klasa, stršeće vrijednosti legitimnih klasifikacija drastično su smanjile površinu ispod ROC krivulje. U odnosu na ostale metode dubokog učenja IWAE opet ima visoke rezultate zahvaljujući robusnijem učenju. I VAE i WAE su zahtijevali

Tablica 3.2. Iznosi mjera za „Vesta“ skup podataka

	AP	AUC ROC	F_1	MCC	G-sredina	Vrijeme inf.
IF	**0.369867**	**0.774683**	**0.450296**	**0.395486**	0.483415	0.735562
OCSVM	0.237001	0.766360	0.344713	0.276172	0.515039	**0.002713**
IWAE	0.195889	0.737583	0.357324	0.292710	**0.517617**	7.216910
VAE	0.194069	0.744944	0.339734	0.272145	0.503629	7.797144
AE	0.184485	0.727670	0.271720	0.229361	0.465305	0.074726
WAE	0.179574	0.729527	0.288586	0.238874	0.468115	0.030784
f-AnoGAN	0.111194	0.454961	0.159956	0.132920	0.212416	0.083436



Slika 3.31. Krivulje svih metoda

snižavanje stope učenja i otežanje norme težina (engl. *weight decay*) kako ne bi divergirali – i svejedno upadnu u lokalni minimum, što se ispostavilo da je generiranje samo onakvih transakcija kakve su u najvećoj grupi. Rezultati općenito nisu zadovoljavajući i zahvaljujući preklapanju podataka i zahvaljujući heterogenosti značajki, čemu ovakvi modeli izgleda teže mogu doskočiti.

3.2. Rasprava

S obzirom na okolnosti: cilj je otkrivanje anomalija, tj. nije klasifikacija i nema nadziranog učenja, velika su preklapanja s obzirom na ciljnu klasu te je ograničeno pretprocesiranje sakrivanjem značenja značajki i nemogućnošću analize s obzirom na povijest jednog računa – dobiveni rezultati su očekivani. Ipak, IWAE se pokazao kao robustan model koji može zadržati relativno visoku preciznost i uz visok odziv te može biti korištan kao filter pred dodatnu obradu: ne samo da bolje ograničava broj legitimnih transakcija označenih kao prijevare, nego i prepoznaje nove uzorke u podatcima, kao što se moglo vidjeti na primjeru „Vesta“ skupa podataka.

4. Zaključak

Ovim je radom dan kratak pregled problema OPKK i pristup tom problemu otkrivanjem anomalija. Opisan je VAE i dan je pregled modela zasnovanih na VAE, uz fokus na tri (VAE, IWAE i WAE) koja su primjenjena na problem OPKK i uspoređena s drugim poznatim metodama (AE, f-AnoGAN, IF, OCSVM) prikladnim za takav tip zadatka.

Potvrđeno je da odabrane metode nenadziranog strojnog učenja nisu prikladne za samostalnu primjenu na ovakav tip problema. Odnosno, na cijelom skupu podataka veća su odstupanja kod legitimnih korisnika koji za odstupanja ne mare, nego kod prevaranata koji nastoje što bolje oponašati većinu – što lakše mogu nego oponašati pojedinu žrtvu. Nadogradnje na VAE svojim prednostima donekle poboljšavaju rezultate, no potrebno je pribjeći konkretnijoj prilagodbi skupa podataka (poput eksplicitnog praćenja ponašanja računa kroz vrijeme) ili je potrebno prilagoditi algoritam odmakom od metoda nenadziranog učenja za otkrivanje anomalija prema metodama klasifikacije kako bi se mogle samostalno koristiti za OPKK.

Među inačicama VAE, IWAE se istaknuo kao robustan model za učenje dane distribucije podataka i može biti koristan za filtriranje prije ručne obrade ako je prihvatljiva preciznost u rasponu $[0.2, 0.33]$, no kod skupova podataka s jednom skupinom i manjim pomakom između većine prijevara i legitimnih transakcija f-AnoGAN se pokazao kao bolje rješenje, a na skupu podataka s raznolikim značajkama i više grupa najbolje rezultate daju plitki modeli.

Daljnji rad može uključivati testiranje nad informativnijim skupom podataka uz prilagodbe za obradu vremenskih nizova (Donut, LSTM, dinamički VAE) ili korištenje drugih generativnih modela, kao što je polunadzirano učenje uz Deep SAD (Ruff et al., 2019).

Literatura

Josephine Akosa. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. 2017.

Ayman Alazizi, Amaury Habrard, François Jacquenet, Liyun He-Guelton, i Frédéric Oblé. Dual Sequential Variational Autoencoders for Fraud Detection. U Michael R. Berthold, Ad Feelders, i Georg Krempl, ur., *Advances in Intelligent Data Analysis XVIII*, Lecture Notes in Computer Science, str. 14–26, Cham, 2020. Springer International Publishing. ISBN 978-3-030-44584-3. https://doi.org/10.1007/978-3-030-44584-3_2.

Jinwoon An i Sungzoon Cho. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. *Special Lecture on IE*, 2(1):1–18, prosinac 2015.

Hung Ba. Improving Detection of Credit Card Fraudulent Transactions using Generative Adversarial Networks, srpanj 2019.

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, i J. Christopher Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, veljača 2011. ISSN 0167-9236. https://doi.org/10.1016/j.dss.2010.08.008.

Kendrick Boyd, Kevin H. Eng, i C. David Page. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. U Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, i Filip Železný, ur., *Machine Learning and Knowledge Discovery in Databases*, str. 451–466, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-40994-3. https://doi.org/10.1007/978-3-642-40994-3_29.

Nazly Rocio Santos Buitrago, Loek Tonnaer, Vlado Menkovski, i Dimitrios Mavroeidis.

Anomaly Detection for imbalanced datasets with Deep Generative Models, studeni 2018.

Yuri Burda, Roger Grosse, i Ruslan Salakhutdinov. Importance Weighted Autoencoders, studeni 2016.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, i Alexander Lerchner. Understanding disentangling in β-VAE, travanj 2018.

Clément Chadebec, Louis J. Vincent, i Stéphanie Allassonnière. Pythae: Unifying Generative Autoencoders in Python – A Benchmarking Use Case. https://arxiv.org/abs/2206.08309v2, lipanj 2022.

Asma Cherif, Arwa Badhib, Heyfa Ammar, Suhair Alshehri, Manal Kalkatawi, i Abdessamad Imine. Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, 35(1):145–174, 2023. https://doi.org/10.1016/j.jksuci.2022.11.008.

Davide Chicco i Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, siječanj 2020. ISSN 1471-2164. https://doi.org/10.1186/s12864-019-6413-7.

Alejandro Correa Bahnsen, Djamila Aouada, i Björn Ottersten. Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19):6609–6619, studeni 2015. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2015.04.042.

Andrea Dal Pozzolo. Credit Card Fraud Detection.

Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, i Francesco Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479:448–455, travanj 2019. ISSN 0020-0255. https://doi.org/10.1016/j.ins.2017.12.030.

Kang Fu, Dawei Cheng, Yi Tu, i Liqing Zhang. Credit Card Fraud Detection Using Convolutional Neural Networks. U Akira Hirose, Seiichi Ozawa, Kenji Doya, Ka-

zushi Ikeda, Minho Lee, i Derong Liu, ur., *Neural Information Processing*, str. 483–490, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46675-0. https://doi.org/10.1007/978-3-319-46675-0_53.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, i Xavier Alameda-Pineda. Dynamical Variational Autoencoders: A Comprehensive Review. *Foundations and Trends® in Machine Learning*, 15(1-2):1–175, 2021. ISSN 1935-8237, 1935-8245. https://doi.org/10.1561/2200000089.

Ian Goodfellow, Yoshua Bengio, i Aaron Courville. *Deep Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 978-0-262-03561-3.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, i Aaron Courville. Improved Training of Wasserstein GANs, prosinac 2017.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, i Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, svibanj 2017. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2016.12.035.

Brandon Harris. Namebrandon/Sparkov_Data_Generation, lipanj 2022.

Yaya Heryadi i Harco Leslie Hendric Spits Warnars. Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM. U *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, str. 84–89, studeni 2017. https://doi.org/10.1109/CYBERNETICSCOM.2017.8311689.

Waleed Hilal, S. Andrew Gadsden, i John Yawney. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, 193:116429, svibanj 2022. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2021.116429.

Addison Howard, Bernadette Bouchon-Meunier, IEEE CIS, inversion, John Lei, Lynn@Vesta, Marcus2010, i Hussein Abbas. IEEE-CIS Fraud Detection, 2019.

Eric Jang, Shixiang Gu, i Ben Poole. Categorical Reparameterization with Gumbel-Softmax, kolovoz 2017.

Shubham Joshi. Abstract data set for Credit card fraud detection, 2018.

Johannes Jurgovsky, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, i Olivier Caelen. Sequence classification for credit-card fraud detection. *Expert systems with applications*, 100:234–245, 2018. https://doi.org/10.1016/j.eswa.2018.01.037.

Yuta Kawachi, Yuma Koizumi, i Noboru Harada. Complementary Set Variational Autoencoder for Supervised Anomaly Detection. U *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, str. 2366–2370, travanj 2018. https://doi.org/10.1109/ICASSP.2018.8462181.

Yuta Kawachi, Yuma Koizumi, Shin Murata, i Noboru Harada. A Two-class Hyperspherical Autoencoder for Supervised Anomaly Detection. U *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, str. 3047–3051, svibanj 2019. https://doi.org/10.1109/ICASSP.2019.8683790.

Zahra Kazemi i Houman Zarrabi. Using deep networks for fraud detection in the credit card transactions. U *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, str. 0630–0633, prosinac 2017. https://doi.org/10.1109/KBEI.2017.8324876.

Diederik P. Kingma i Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237, 1935-8245. https://doi.org/10.1561/2200000056.

Diederik P. Kingma i Max Welling. Auto-Encoding Variational Bayes, prosinac 2022.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, i Max Welling. Semi-supervised learning with deep generative models. U *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, str. 3581–3589, Cambridge, MA, USA, prosinac 2014. MIT Press.

Fei Tony Liu, Kai Ming Ting, i Zhi-Hua Zhou. Isolation-Based Anomaly Detection. *ACM*

Transactions on Knowledge Discovery from Data, 6(1):1–39, ožujak 2012. ISSN 1556-4681, 1556-472X. https://doi.org/10.1145/2133360.2133363.

Tiebin Liu i Shiping Liu. Fraud detection model & application for credit card acquiring business based on data mining technology. U *2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)*, str. 963–967. Atlantis Press, prosinac 2016. ISBN 978-94-6252-265-7. https://doi.org/10.2991/iceeecs-16.2016.185.

Dr Yvan Lucas, INSA Lyon, i Dr Johannes Jurgovsky. Credit card fraud detection using machine learning: A survey. 14.10.2020.

Eyad Abdel Latif Marazqah Btoush, Xujuan Zhou, Raj Gururajan, Ka Ching Chan, Rohan Genrich, i Prema Sankaran. A systematic review of literature on credit card cyber fraud detection using machine and deep learning. 2023. https://doi.org/10.7717/peerj-cs.1278.

Igor Mekterović, Ljiljana Brkić, i Mirta Baranović. A systematic review of data mining approaches to credit card fraud detection. *WSEAS Trans. Bus. Econ*, 15:437, 2018.

Igor Mekterović, Mladen Karan, Damir Pintar, i Ljiljana Brkić. Credit Card Fraud Detection in Card-Not-Present Transactions: Where to Invest? *Applied Sciences*, 11(15): 6766, siječanj 2021. ISSN 2076-3417. https://doi.org/10.3390/app11156766.

Asim Munawar, Phongtharin Vinayavekhin, i Giovanni De Magistris. Limiting the Reconstruction Capability of Generative Neural Network using Negative Learning. https://arxiv.org/abs/1708.08985v1, kolovoz 2017.

Soumaya Ounacer, Hicham Ait El Bour, Younes Oubrahim, Mohamed Yassine Ghoumari, i Mohamed Azzouazi. Using isolation forest in anomaly detection: The case of credit card transactions. *Periodicals of Engineering and Natural Sciences (PEN)*, 6(2): 394, studeni 2018. ISSN 23034521. https://doi.org/10.21533/pen.v6i2.533.

Guansong Pang, Chunhua Shen, Longbing Cao, i Anton Van Den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):38:1–38:38, ožujak 2021. ISSN 0360-0300. https://doi.org/10.1145/3439950.

Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, i Maurizio Pierini. Anomaly Detection with Conditional Variational Autoencoders. U *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, str. 1651–1657, prosinac 2019. https://doi.org/10.1109/ICMLA.2019.00270.

Apapan Pumsirirat i Liu Yan. Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. *International Journal of Advanced Computer Science and Applications*, 9(1), 2018. ISSN 21565570, 2158107X. https://doi.org/10.14569/IJACSA.2018.090103.

Brian C. Ross. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE*, 9(2):e87357, veljača 2014. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0087357.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, i Marius Kloft. Deep Semi-Supervised Anomaly Detection, 2019.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, i Klaus-Robert Müller. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5): 756–795, svibanj 2021. ISSN 1558-2256. https://doi.org/10.1109/JPROC.2021.3052449.

Yusuf Sahin, Serol Bulkan, i Ekrem Duman. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 40(15):5916–5923, studeni 2013. ISSN 09574174. https://doi.org/10.1016/j.eswa.2013.05.021.

Takaya Saito i Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10 (3):e0118432, ožujak 2015. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0118432.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, i Georg Langs. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. U Marc Niethammer, Martin Styner, Stephen Aylward,

Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, i Dinggang Shen, ur., *Information Processing in Medical Imaging*, sv. 10265, str. 146–157. Springer International Publishing, Cham, 2017. ISBN 978-3-319-59049-3 978-3-319-59050-9. https://doi.org/10.1007/978-3-319-59050-9_12.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, i Ursula Schmidt-Erfurth. F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, svibanj 2019. ISSN 1361-8415. https://doi.org/10.1016/j.media.2019.01.010.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, i Robert C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471, srpanj 2001. ISSN 0899-7667. https://doi.org/10.1162/089976601750264965.

Tom Sweers. *Autoencoding Credit Card Fraud*. Doktorska disertacija, Radboud University, lipanj 2018.

Fabio Henrique Kiyoiti dos Santos Tanaka i Claus Aranha. Data Augmentation Using GANs, travanj 2019.

Huang Tingfei, Cheng Guangquan, i Huang Kuihua. Using Variational Auto Encoding in Credit Card Fraud Detection. *IEEE Access*, 8:149841–149853, 2020. ISSN 2169-3536. https://doi.org/10.1109/ACCESS.2020.3015600.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, i Bernhard Schoelkopf. Wasserstein Auto-Encoders, prosinac 2019.

Jakub M. Tomczak i Max Welling. VAE with a VampPrior, veljača 2018.

Bénard Wiese i Christian Omlin. Credit Card Transactions, Fraud Detection, and Machine Learning: Modelling Time with LSTM Recurrent Neural Networks. U Monica Bianchini, Marco Maggini, Franco Scarselli, i Lakhmi C. Jain, ur., *Innovations in Neural Information Paradigms and Applications*, str. 231–268. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-642-04003-0. https://doi.org/10.1007/978-3-642-04003-0_10.

Haowen Xu, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, i Dan Pei. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. U *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, str. 187–196, Lyon, France, 2018. ACM Press. ISBN 978-1-4503-5639-8. https://doi.org/10.1145/3178876.3185996.

Yuki Yamanaka, Tomoharu Iwata, Hiroshi Takahashi, Masanori Yamada, i Sekitoshi Kanai. Autoencoding Binary Classifiers for Supervised Anomaly Detection. https://arxiv.org/abs/1903.10709v1, ožujak 2019.

Vladimir Zaslavsky i Anna Strizhak. Credit Card Fraud Detection Using Self-Organizing Maps. *Information & Security: An International Journal*, 18:48–63, 2006. ISSN 08615160. https://doi.org/201308010750.

Zhaohui Zhang, Xinxin Zhou, Xiaobo Zhang, Lizhi Wang, i Pengwei Wang. A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection. *Security and Communication Networks*, 2018:1–9, kolovoz 2018. ISSN 1939-0114, 1939-0122. https://doi.org/10.1155/2018/5680264.

Xiaokang Zhou, Yiyong Hu, Wei Liang, Jianhua Ma, i Qun Jin. Variational LSTM Enhanced Anomaly Detection for Industrial Big Data. *IEEE Transactions on Industrial Informatics*, 17(5):3469–3477, svibanj 2021. ISSN 1941-0050. https://doi.org/10.1109/TII.2020.3022432.

Sažetak

Otkrivanje prijevara kreditnim karticama korištenjem varijacijskog autoenkodera

Damjan Grubelić

Otkrivanje prijevara jedan je od ključnih problema kod kartičnog plaćanja. Naći skalabilan način za rješavanje tog problema nije jednostavan zadatak jer broj legitimnih transakcija daleko nadmašuje broj prijevara. Osim toga, prevaranti neprestano prilagođavaju pristupe pa detekcija mora biti robusna na promjene kroz vrijeme. Varijacijski autoenkoder može enkodirati distribucije iz kojih takvi podatci proizlaze što omogućava robusnost na nesrazmjer u podatcima te generiranje novih, sličnih podataka za smanjenje nesrazmjera. U ovom radu uspoređujem nekoliko inačica varijacijskog autoenkodera te pristupa problemu detekcije prijevara pomoću njih.

Ključne riječi: VAE; generativni modeli; otkrivanje anomalija

Abstract

Credit card fraud detection using variational autoencoders

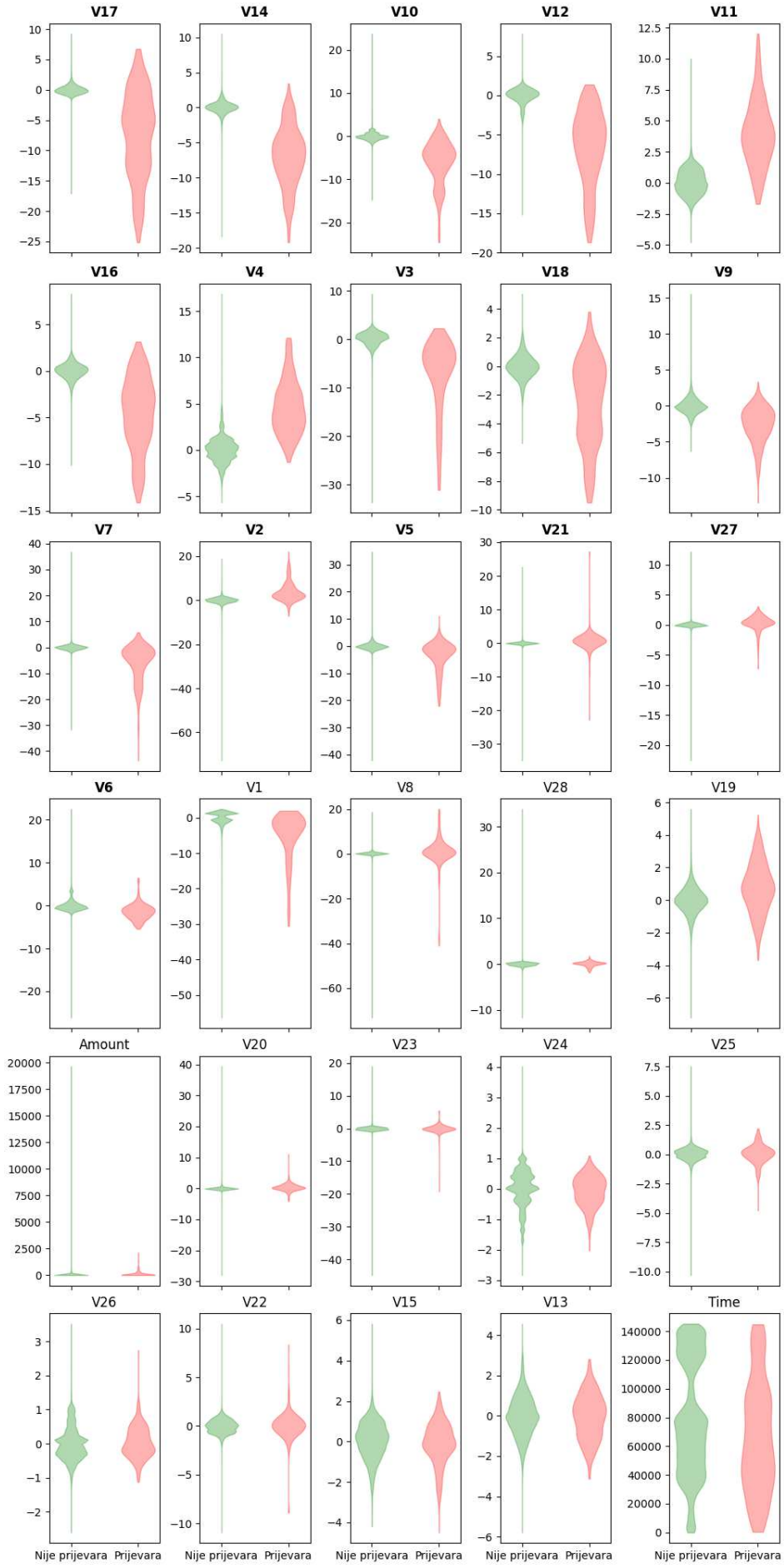
Damjan Grubelić

Detecting fraud is one of the key issues in card payments. Finding a scalable way to address this problem is not an easy task because the number of legitimate transactions far exceeds the number of frauds. Moreover, fraudsters constantly adapt their approaches, so detection must be robust to changes over time.

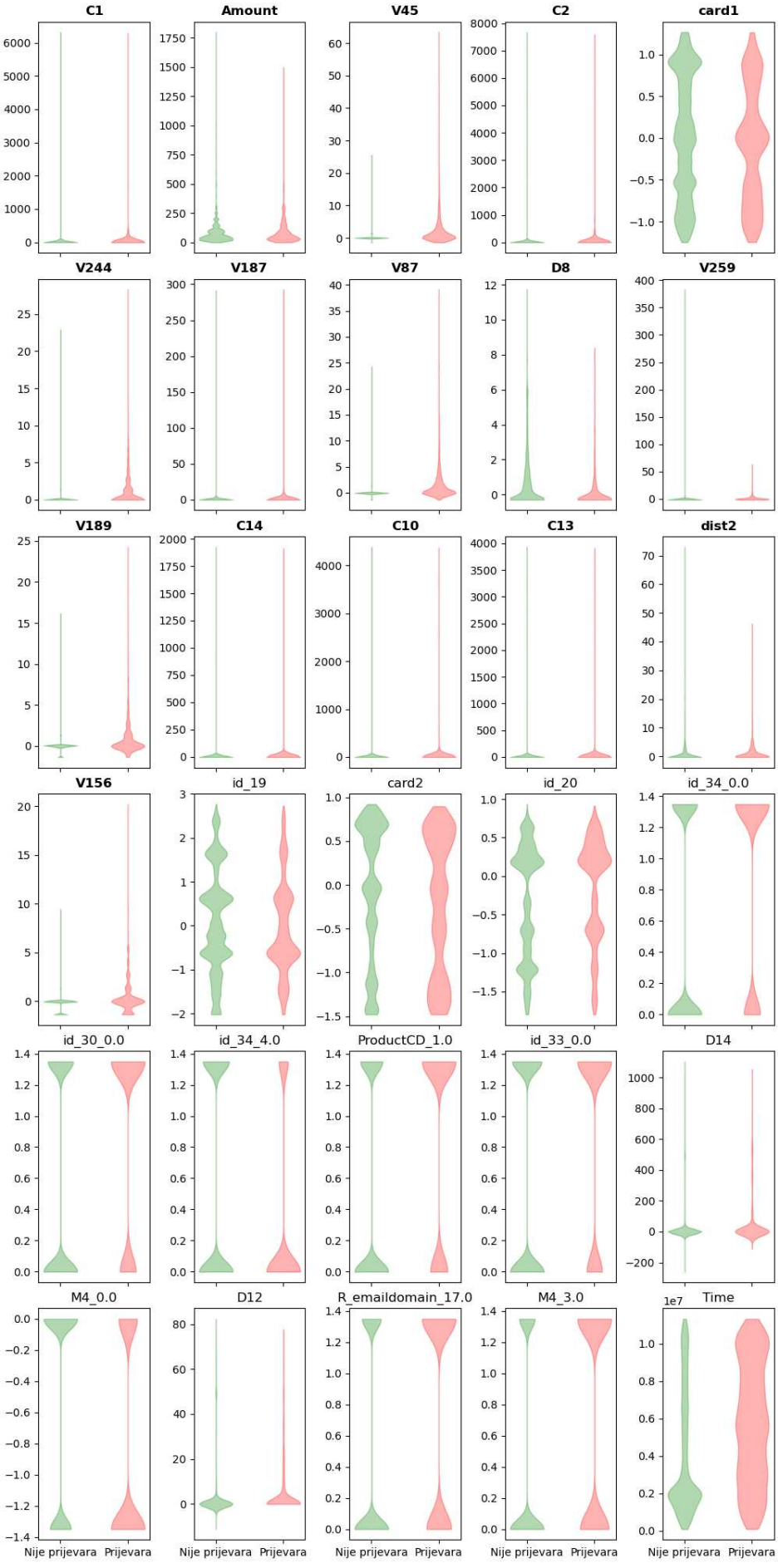
A variational autoencoder can encode the distributions from which such data originate, enabling robustness to the imbalance in the data and the generation of new, similar data to reduce this imbalance. In this paper, I compare several variants of the variational autoencoder and their approaches to the problem of fraud detection.

Keywords: VAE; generative models; anomaly detection

Privitak A: Odnos varijabli i ciljne klase



Slika A1. Skup podataka „MLG-ULB“: varijable o kojima najviše ovisi ciljna klasa



Slika A2. Skup podataka „Vesta“: varijable o kojima najviše ovisi ciljna klasa