

Detekcija lažnih blockchain računa primjenom algoritama strojnog učenja

Mentor: prof.dr.sc Zlatan Car
Komentor: v. asist. dr. sc. Nikola Anđelić

22.1.2024

David Mustač

Sadržaj

1. Uvod
2. Rješenje
3. Prikupljanje podataka
4. Čišćenje podataka
5. Pred procesiranje podataka
6. Treniranje
7. Algoritmi
8. Stacking ansambl
9. Usporedba preformansi modela
10. Zaključak

Uvod

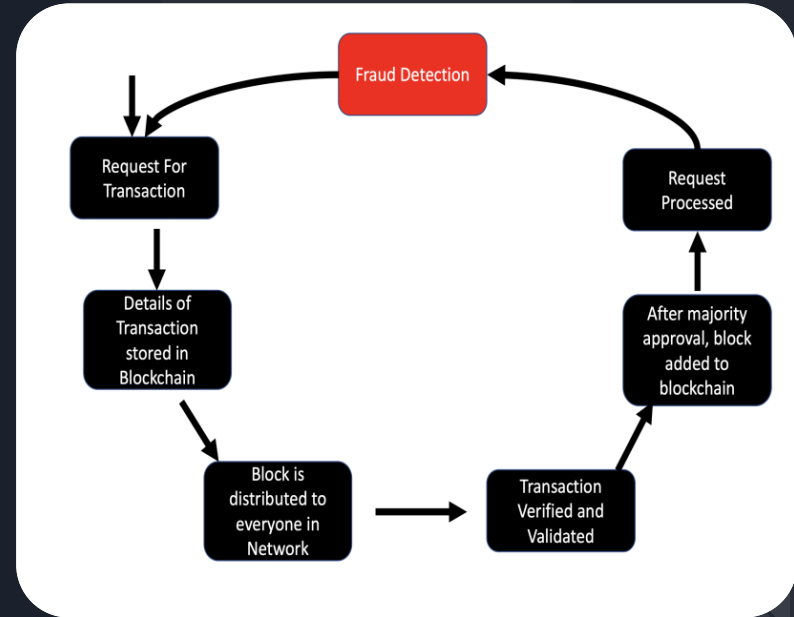


Ethereum je distribuirana računalna platforma otvorenog koda koja se temelji na blockchain tehnologiji i omogućava stvaranje ugovora čiji su uvjeti izvršenja zapisani u linijama koda

Iako Ethereum ima veliki potencijal, pati od transakcijskih prijevara koje bi mogle predstavljati rizik za njegove korisnike. Otkrivanje prijevara ključno je u svakom financijskom sustavu.

Rješenje

Algoritmi strojnog učenja imaju potencijal riješiti ovaj problem brzim i točnim otkrivanjem lažnih transakcija proizašlih iz blockchain računa





Prikupljanje podataka

Prvi korak u otkrivanju prijevare je prikupljanje podataka iz različitih izvora kao što su blockchain, SEC-jeva baza podataka financijskih prijevara, javne baze na Kaggle-u i prema potrebi ako nisu pretvoriti u csv. format

Čišćenje podataka

- Izostavili su se podaci koji nisu relevantni za analizu
- Podaci sa varijancom 0

FLAG	1.724110e-01
Avg min between sent txn	4.616718e+08
Avg min between received txn	5.327656e+08
Time Diff between first and last (Mins)	1.042889e+11
Sent txn	5.733918e+05
Received Txn	8.851734e+05
Number of Created Contracts	2.000685e+04
Unique Received From Addresses	8.917457e+04
Unique Sent To Addresses	6.960121e+04
min value received	1.062298e+05
max value received	1.692294e+08
avg val received	8.323238e+06
min val sent	1.921264e+04
max val sent	4.394646e+07
avg val sent	5.715935e+04
min value sent to contract	5.080371e-08
max val sent to contract	2.660652e-07
avg value sent to contract	1.046096e-07
total transactions (including txn to create contract)	1.828997e+06
total Ether sent	1.283952e+11
total ether received	1.326451e+11
total ether sent contracts	2.660625e-07
total ether balance	5.877009e+10
Total ERC20 txns	2.002821e+05
ERC20 total Ether received	1.110618e+20
ERC20 total ether sent	1.393321e+18
ERC20 total Ether sent contract	3.756017e+07
ERC20 uniq sent addr	1.107809e+04
ERC20 uniq rec addr	6.694262e+03
ERC20 uniq sent addr.1	4.316210e-03
ERC20 uniq rec contract addr	2.974444e+02
ERC20 avg time between sent txn	0.000000e+00
ERC20 avg time between rec txn	0.000000e+00
ERC20 avg time between rec 2 txn	0.000000e+00
ERC20 avg time between contract txn	0.000000e+00
ERC20 min val rec	2.850451e+08
ERC20 max val rec	1.110370e+20
ERC20 avg val rec	4.584705e+16
ERC20 min val sent	1.110004e+12
ERC20 max val sent	1.392176e+18
ERC20 avg val sent	3.498443e+17
ERC20 min val sent contract	0.000000e+00
ERC20 max val sent contract	0.000000e+00
ERC20 avg val sent contract	0.000000e+00
ERC20 uniq sent token name	4.536185e+01
ERC20 uniq rec token name	2.781759e+02
dtype: float64	

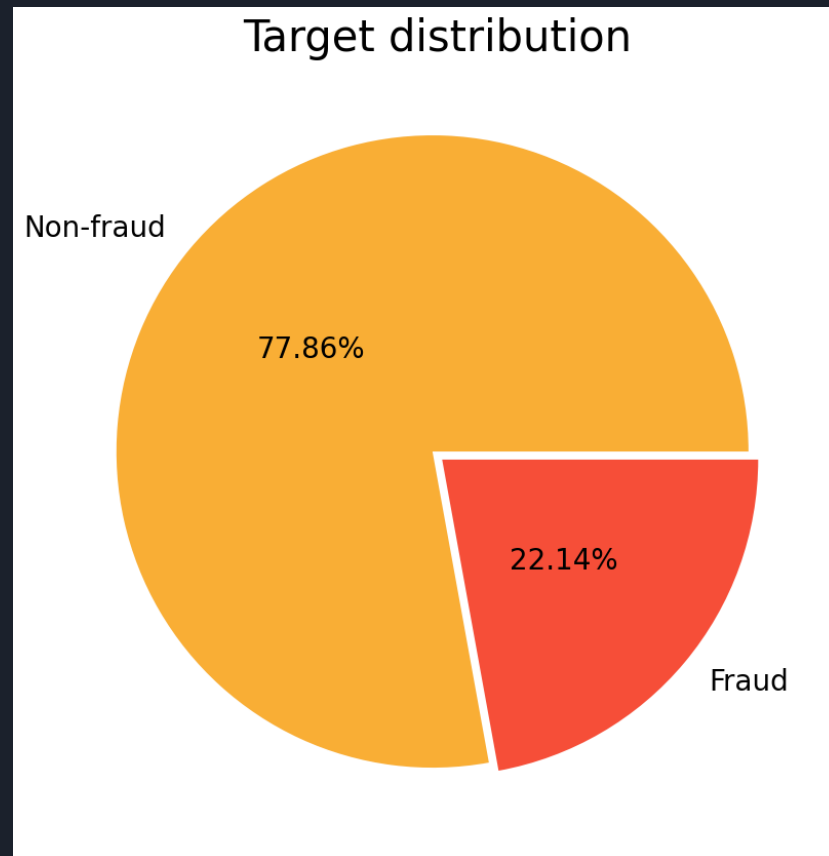
Predprocesiranje podataka

- Izvukli smo informacije kao što su prosječna vrijednost, standardna odstupanja, minimum i maksimum,
- Objektne varijable su pretvorene u kategoričke varijable dtype radi učinkovitijeg procesa računanja.

Analiza ciljne distribucije

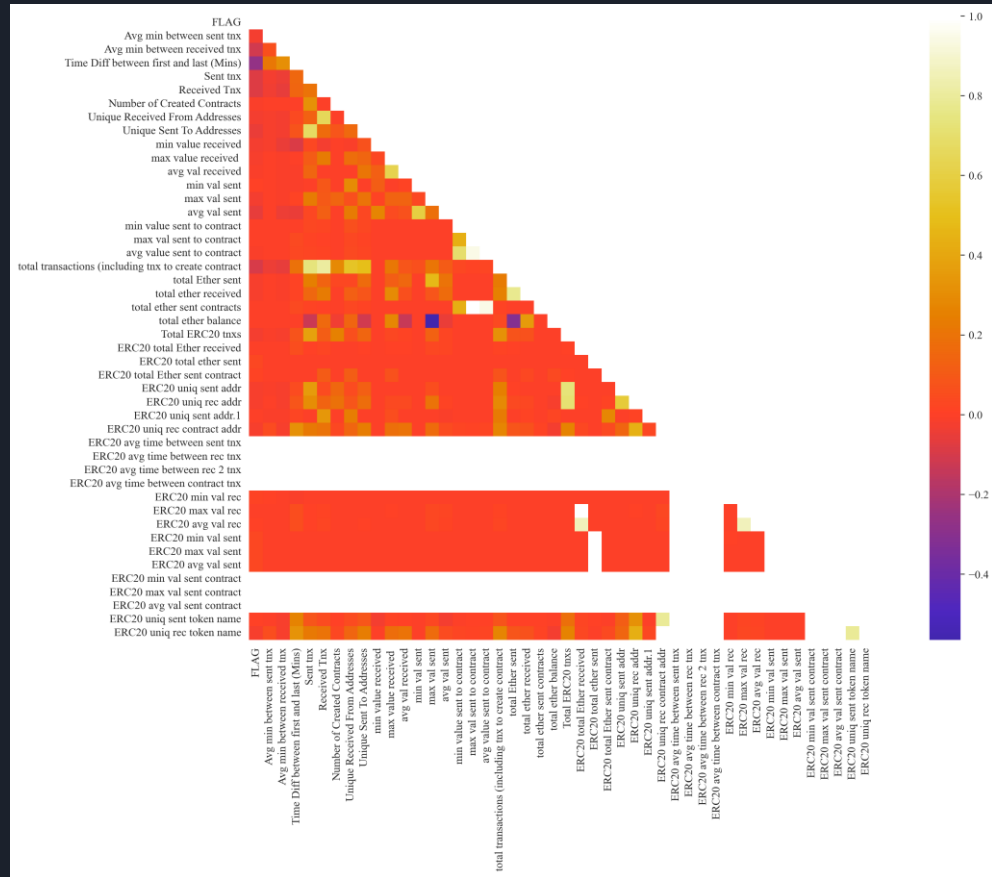
- uvid o ciljnoj varijabli
- učestalost različitih vrijednosti
- procjena ravnoteže skupa podataka, dominantne klase i utjecaj neravnoteže

Prilikom izbora modela ključno je rješavanje neravnoteže u distribuciji kako bi se izbjegla njegova pristranost



Korelacijska matrica

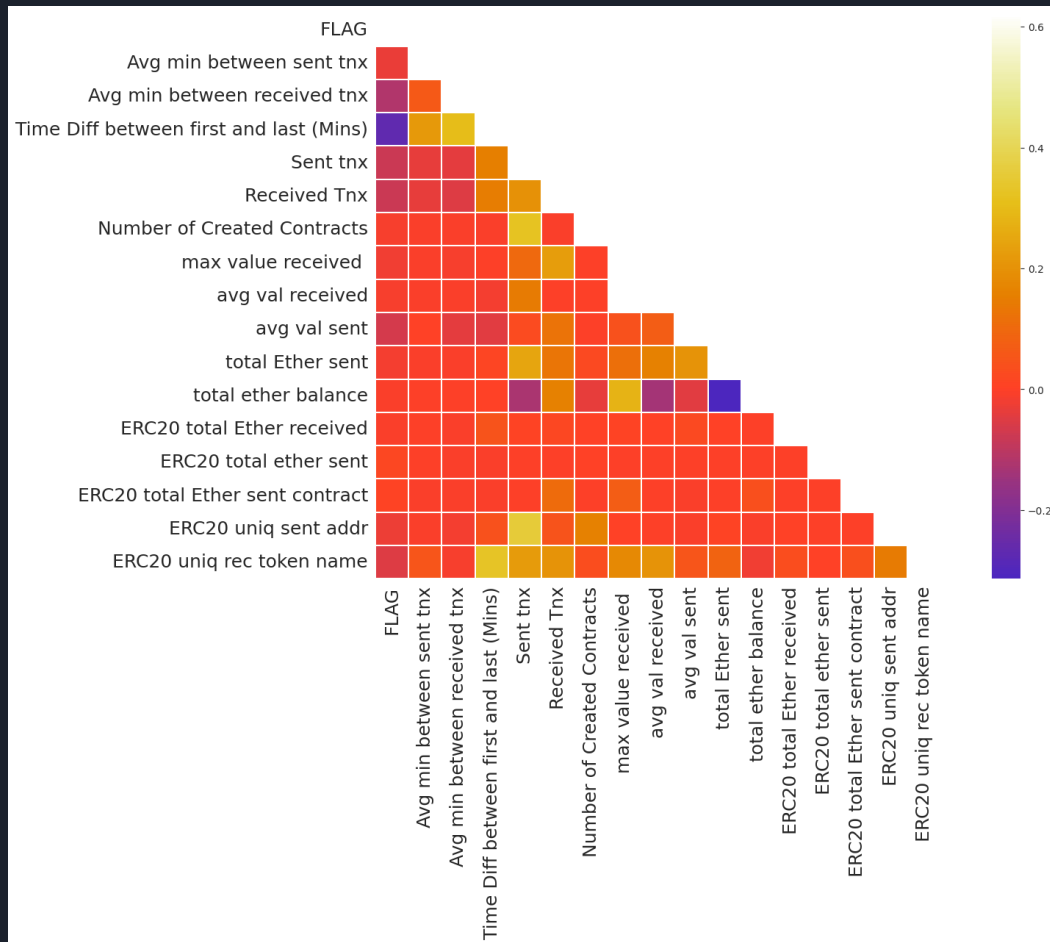
- pomaže u identificiranju suvišnih značajki i otkrivanju multikolinearnosti
- mjeri samo linearne odnose između varijabli



Korelacijska matrica

Uklonili smo:

- značajke sa varijancom jednakom 0
- visoko korelirane značajke što može ukazivati na redundanciju





Treniranje


- 80% treniranje, 20% testiranje
- skup za testiranje služi za procjenu izvedbe modela usporedbom njegovog predviđanja sa stvarnim oznakama

Normalizacija

- skaliranje značajki na zajednički raspon
- kako bi se olakšalo učenje algoritma

	Avg min between sent txn	Avg min between received txn	Time Diff between first and last (Mins)	Sent txn	Received Txn	Number of Created Contracts	max value received	avg val received	avg val sent	total Ether sent	total ether balance	ERC20 total Ether received	ERC20 total ether sent	ERC20 total Ether sent contract	ERC20 uniq sent addr	ERC20 uniq rec token name
0	1.294061	1.151313	1.393751	1.591951	1.017881	-0.401539	1.170696	0.988757	0.651223	1.416308	-0.007274	1.815951	2.508169	-0.038483	2.398649	1.831406
1	-1.096066	-1.184221	-1.638410	-1.391726	-1.785005	-0.401539	-1.407378	-1.283886	-1.138468	-1.252291	-0.006836	-0.746114	-0.410600	-0.038483	-0.437145	0.226082
2	-0.006354	0.213137	1.103220	1.970707	1.876994	-0.401539	0.613575	-0.995919	-0.869934	1.171479	-0.006819	-0.746114	-0.410600	-0.038483	-0.437145	-0.994019
3	-1.096066	1.220438	0.195684	-1.391726	-0.618856	2.490307	-0.871196	-0.822389	-1.138468	-1.252291	-0.006832	-0.746114	-0.410600	-0.038483	-0.437145	-0.994019
4	0.628503	-1.121221	-0.904665	-0.206187	-0.618856	-0.401539	0.889875	1.277099	1.332663	0.787323	-0.006836	-0.746114	-0.410600	-0.038483	-0.437145	-0.994019
...
7867	-1.096066	0.249518	-0.705997	-0.590325	-0.618856	-0.401539	-0.535113	-0.269750	0.183532	-0.463079	-0.006836	-0.746114	-0.410600	-0.038483	-0.437145	0.226082
7868	1.128623	0.633621	0.895636	0.707356	1.432439	-0.401539	0.339924	-0.354586	1.043745	1.010645	-0.006835	1.611838	-0.410600	-0.038483	-0.437145	1.238838
7869	-1.096066	-1.184221	-1.154101	-0.590325	-1.004903	-0.401539	-1.275895	-1.091970	-0.960939	-1.159157	-0.006835	-0.746114	-0.410600	-0.038483	-0.437145	-0.994019
7870	1.385084	0.988959	0.715180	0.707356	0.556810	-0.401539	1.858790	1.531808	1.671553	1.592460	-0.006835	1.514388	-0.410600	-0.038483	-0.437145	1.122262
7871	-1.096066	0.565250	0.761464	-1.391726	1.419117	2.490307	-0.956946	-0.971335	-1.138468	-1.252291	-0.006675	0.014594	-0.410600	-0.038483	-0.437145	0.956416

7872 rows × 16 columns



Transformacija distribucije s logaritamskom funkcijom

- rješavanje asimetrije u distribuciji podataka
- transformacija uzimajući logaritam svake značajke, sažima veće vrijednosti i širi manje vrijednosti čineći podatke slične normalnoj distribuciji jer algoritam pretpostavlja takvu



SMOTE

- Kako ne bismo dobili pristranost modela jer postoji neravnoteža u broju instanci između većinske (6115) klase i manjinske (1757) klase, morali smo uravnotežiti klase
- SMOTE metoda je tehnika sintetskog preuzorkovanja manjinskih uzoraka,
- Nakon preuzorkovanja broj uzoraka izjednačen je na 6115,
- Sada je vjerojatnije uhvatiti temeljne obrasce u podacima i točno identificirati slučajeve prijevare i slučajeve koji nisu prijevare.



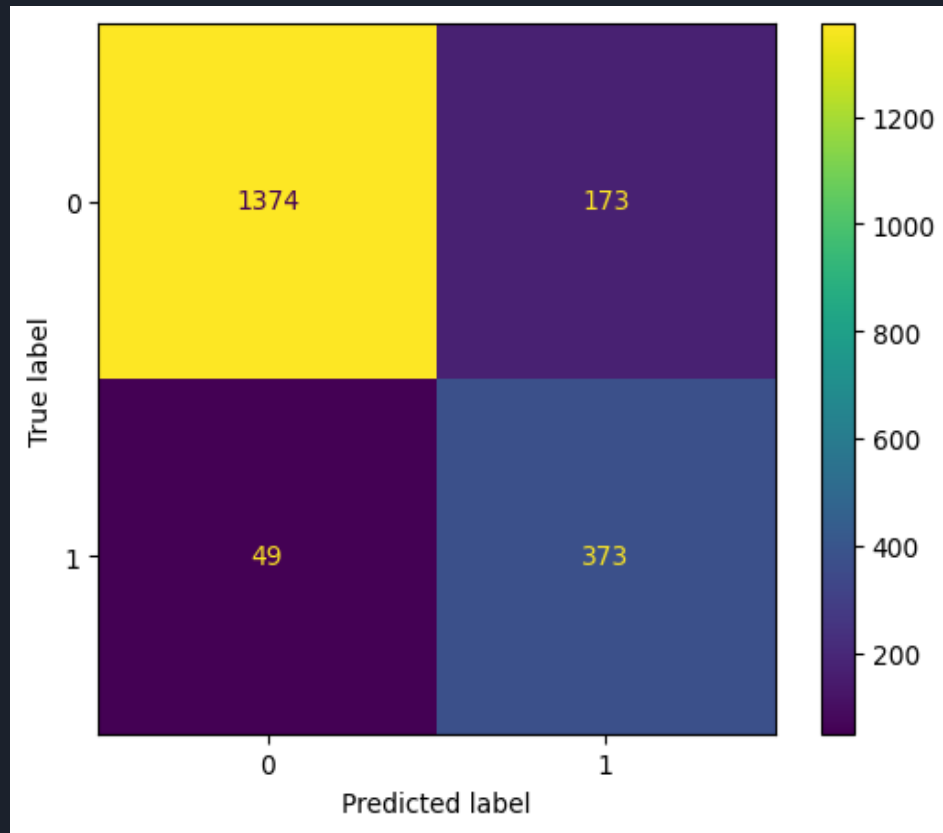
Algoritmi

- omogućavaju nam da podatkovne točke kategoriziramo u unaprijed definirane skupine na temelju njihovih karakteristika.
- Koristili su se logistička regresija, random forest i xgboost algoritmi

FLAG	Broj transakcija	OPIS
0	1547	Valjanje transakcije
1	422	Prijevare

Logistička regresija

- Istinski pozitivno (TP): Model je točno identificirao 373 slučaja PRIJEVARE od 422 (tj. 88%).
- Lažno pozitivno (FP): Model je netočno označio 173 transakciju kao PREVAREU od 1547, dok one zapravo NISU PREVARE
- Lažno negativan (FN): Model nije uspio identificirati 49 slučajeva PRIJEVARE, tretirajući ih kao da nisu prijevare. Neke lažne transakcije nisu otkrivene, što potencijalno dovodi do financijskih gubitaka za organizaciju.
- Više nam je stalo do transakcija koje su zapravo bile PREVARE, ali koje je naš model tretirao kao NISU PRIJEVARE (FN - 49) GREŠKA VRSTE II. Takve greške je potrebno maksimalno smanjiti, a to ćemo pokušati sa sljedećim modelima.



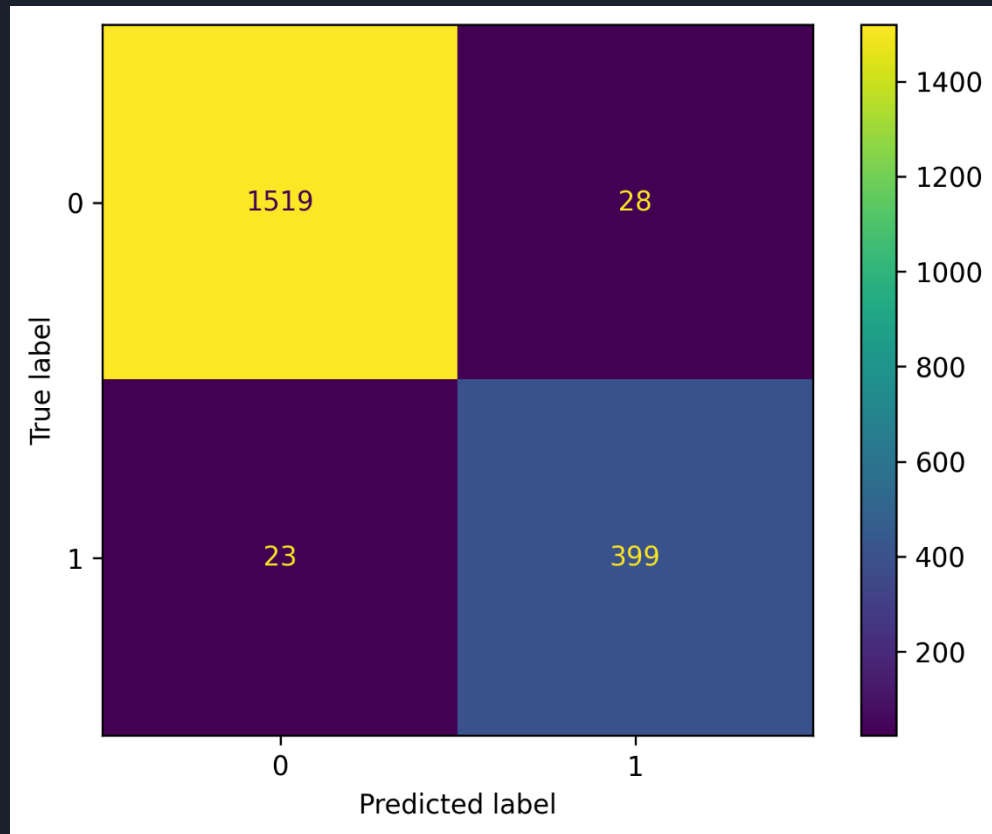
Preformanse modela

Best Hyperparameters	'C'	'Penalty'	'Solver'
	1	L2	liblinear

Classification report	Precision	Recall	F1- score	Support
0	0.97	0.89	0.92	1547
1	0.68	0.88	0.77	422
Accuracy			0.89	1969
Macro avg	0.82	0.89	0.85	1969
Weighted avg	0.91	0.89	0.89	1969
Accuracy			0.8842	
ROC AUC Score			0.8841	
Precision			0.6757	
Recall			0.8839	
F1 Score			0.7659	

Random forest

- Istinski pozitivno (TP): Model je točno identificirao 399 slučaja PRIJEVARE od 422. To ukazuje da je model učinkovit u identificiranju stvarnih lažnih transakcija.
- Lažno pozitivno (FP): Model je netočno označio 28 transakciju kao PREVARU od 1547, što pokazuje napredak u odnosu na prethodni model
- Lažno negativan (FN): Model nije uspio identificirati 23 slučajeva PRIJEVARE, tretirajući ih kao da nisu prijevare, što je isto poboljšanje na prethodni model



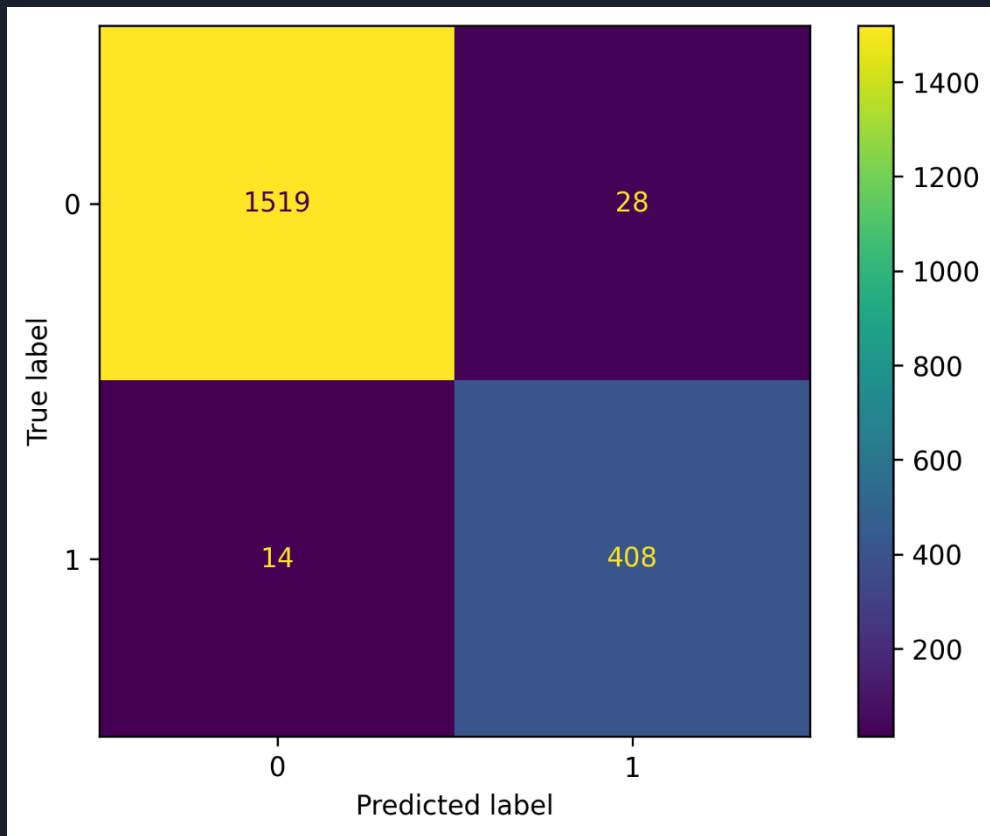
Preformance modela

Best Hyperparameters	'max_depth'	'max_features'	'min_samples_leaf'	'min_samples_split'	n_estimators'
	'None'	'Auto'	1	2	100

Classification report	Precision	Recall	F1- score	Support	
0	0.99	0.98	0.98	1547	
1	0.92	0.95	0.94	422	
Accuracy			0.97	1969	
Macro avg	0.95	0.97	0.96	1969	
Weighted avg	0.97	0.97	0.97	1969	
Accuracy				0.974	
ROC AUC Score				0.967	
Precision				0.926	
Recall				0.955	
F1 Score				0.941	

XGBoost

- Istinski pozitivno (TP): Model je točno identificirao 408 slučajeva PRIJEVARE od 422. To ukazuje da je model učinkovit u identificiranju stvarnih lažnih transakcija.
- Lažno pozitivno (FP): Model je netočno označio 28 transakciju kao PREVARU od 1547, međutim one to nisu bile
- Lažno negativan (FN): Model nije uspio identificirati 14 slučajeva PRIJEVARE, tretirajući ih kao da nisu prijevare, ali su bile, što je isto poboljšanje na prethodni model



Preformanse modela

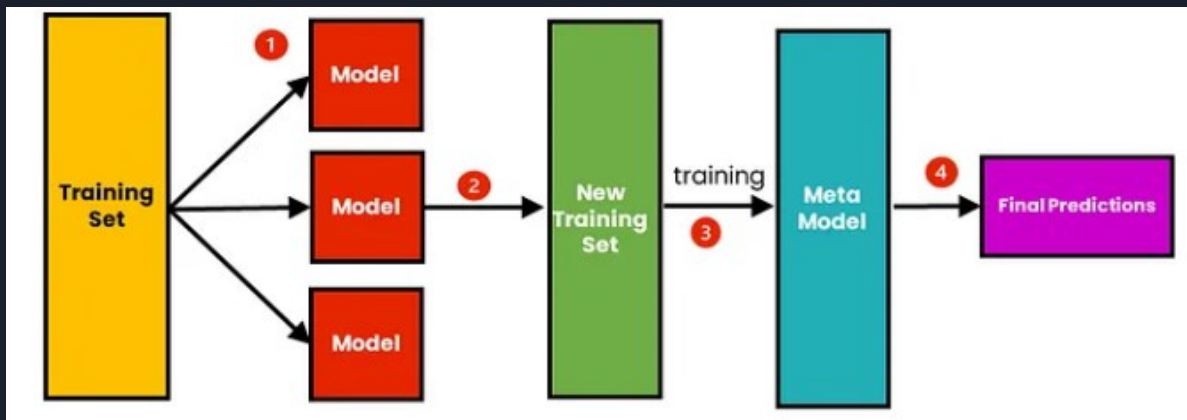
Best Hyperparameters	'colsample_bytree'	'learning_rate'	'max_depth'	'n_estimators'	'subsample'
	0.5	0.5	4	200	0.5

Classification report	Precision	Recall	F1- score	Support
0	0.99	0.99	0.99	1547
1	0.95	0.96	0.95	422
Accuracy			0.98	1969
Macro avg	0.97	0.97	0.97	1969
Weighted avg	0.98	0.98	0.98	1969
Accuracy			0.9802	
ROC AUC Score			0.9727	
Precision			0.9484	
Recall			0.9597	
F1 Score			0.9541	

Usporedba preformansi algoritama

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
0	Random Forest	0.972067	0.917995	0.954976	0.936121	0.965853
1	XGBoost	0.980193	0.948478	0.959716	0.954064	0.972747
2	Logistic Regression	0.885729	0.679417	0.883886	0.768280	0.885059

Stacking ansamble



- strategija strojnog učenja koja kombinira predviđanja više osnovnih modela kako bi se dobila bolja prediktivna izvedba,
- osnovni modeli koji su korišteni logistička regresija, random forest i XGBoost algoritmi.



Preformanse stacking ansambla sa različitim meta modelima

Meta model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Logistic regression	0.974099	0.926437	0.954976	0.940490	0.967146
Random forest	0.980193	0.954869	0.950607	0.943737	0.970162
XGBoost	0.980193	0.954869	0.952607	0.953737	0.970162



Usporedba preformanse stacking ansambla i najboljeg pojedinačnog modela

Vrste algoritma/metode	Meta model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Stacking ansambl	XGBoost	0.980193	0.954869	0.952607	0.953737	0.970162
XGBoost	-	0.980193	0.948478	0.959716	0.954064	0.972747

Zaključak

Pojedinačni model XGBoost ima više vrijednosti za metriku prisjećanja, F1 scorea i ROC AUC. To sugerira da je pojedinačni model XGBoost nešto bolji u predviđanju pozitivnih primjera i razlikovanju između pozitivnih i negativnih primjera.

Međutim, model Stacking ensemble ima više vrijednosti za metriku preciznosti. To sugerira da je model Stacking ensemble nešto bolji u ukupnoj prediktivna izvedbi.

U konačnici, izbor modela koji se koristi ovisi o specifičnim zahtjevima aplikacije. Ako su prisjećanje i ROC AUC osobito važni, tada je pojedinačni model XGBoost bolji izbor. Međutim, ako je ukupna prediktivna izvedba važnija, tada je model Stacking ensemble bolji izbor.

Vrste algoritma/metode	Meta model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Stacking ansambl	XGBoost	0.980193	0.954869	0.952607	0.953737	0.970162
XGBoost	-	0.980193	0.948478	0.959716	0.954064	0.972747



Hvala na pažnji !

David Mustač

Linked in profil

