Primerjava klasifikacijskih modelov na neuravnoteženih podatkih

Ž. Nagelj, L. Loncaric

September 11, 2019

1 Uvod

Cilj naloge je predstaviti in primerjati štiri razlicnih metod za klasifikacijo na primeru neuravnovešenih podatkov. Gre se za vec kot 500 tistoc transakcijskih podatkov razlicnih tipov na podlagi katerih želimo zaznati prevaro. Za konec bomo v primerjavo dodali še enasamble modelov in sicer v primeru vecine glasov ter konsenza.

2 Podatki

Podatke je na spletni strani Kaggle (https://www.kaggle.com/c/ieee-fraud-detection/data) zagotovilo podjetje Vesta. Podatkovni set ima vec kot 350 številcnih in kategoricni neodvisnih spremenljik. Pomen posameznih spremenljivk ni pojasnjen, so pa definirani naslednji sklopi:

- TransactionDT : timedelta from a given reference datetime
- TransactionAMT : transaction payment amount in USD
- ProductCD : product code, the product for each transaction
- card1 - card6 : payment card information, such as card type, card category, issue bank, country, etc.
- addr : address
- dist : distance
- P and R emaildomain : purchaser and recipient email domain
- C1-C14 : counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked
- D1-D15 : timedelta, such as days between previous transaction, etc.
- M1-M9 : match, such as names on card and address, etc.
- Vxxx: : Vesta engineered rich features, including ranking, counting, and other entity relations

2.1 Nove spremenljivke

Posebno naravo ima spremenljivka *TransactionDT*, ki je periodicna. Pomensko nas zanima tedenska perioda, ki jo bomo zajeli tako, da bomo glede na vrednost spremenljivke transakcijo uvrstili v enega izmed sedmih razredov (dni).

2.2 Priprava podatkov

Prvi korak obdelave podatkov je obsegal izbiro kakovostnih spremenljivk glede na delež manjkajocih vrednosti. Odstranili smo tiste spremenljivke, katere dežel manjkajocih vrednosti je presegal 20%. Vecina odstranejenih spremenljivk je imela oznako V (umetno ustvarjene spremenljivke) in D (informacije o casih med transakcijami).

Saj smo opazili, da pri nekaterih spremenljivkah nastopajo vedno enake vrednosti smo nato odstranili tudi tiste spremenljivke, katere varianca je bila prakticno nic. Pri tem smo zaradi velikega števila transakcij brez prever s testom ANOVA preverili, da nizka varianca ni posledica neuravnoteženega podatkovnega seta. Testirali smo torej statisticno znacilnost razlik med povprecij ob in brez prevare.

Analizo bomo izvajali na popolnih podatkih, torej tistih brez manjkajocih vrednosti. Po obdelavi na ostane 346873 transakcij in 96 neodvisnih spremenljivk. Izkaže se, da je kolicina podatkov prevelika za procesiranje na osebnih racunalnikih, zato bomo s pomocjo stratificiranega vzorcenja vzeli le polovico podatkov, ki jih bomo razdelili na dva dela z namenom nepristranske validacije modela.

2.3 Izbira spremenljivk

Kljub manjšemu številu transakcij imamo še vedno preveliko število neodvisnih spremenljivk. Saj cilj analize ni iskanje cim boljšega modela temvec primerjava razlicnih metod se odlocimo, da bomo na podlagi random foresta izbrali 20 spremenljivk, ki pripelje do najvecje klasifikacijske tocnosti. Rezultat v primeru 96 spremenljivk poda tocnost modela 0.9644 ± 0.0008 ter v primeru 20 spremenljivk 0.9686 ± 0.0012. Izbrane so bile naslednje spremenljivke:

- TransactionAmt
- card1, card2, card5, card6
- emaildomain
- C1, C2, C6, C9, C11, C13, C14
- V76, V78, V83, V283, V285, V294, V296

2.4 Vizualizacija izbiranih spremenljivk

TransactionAMT

***

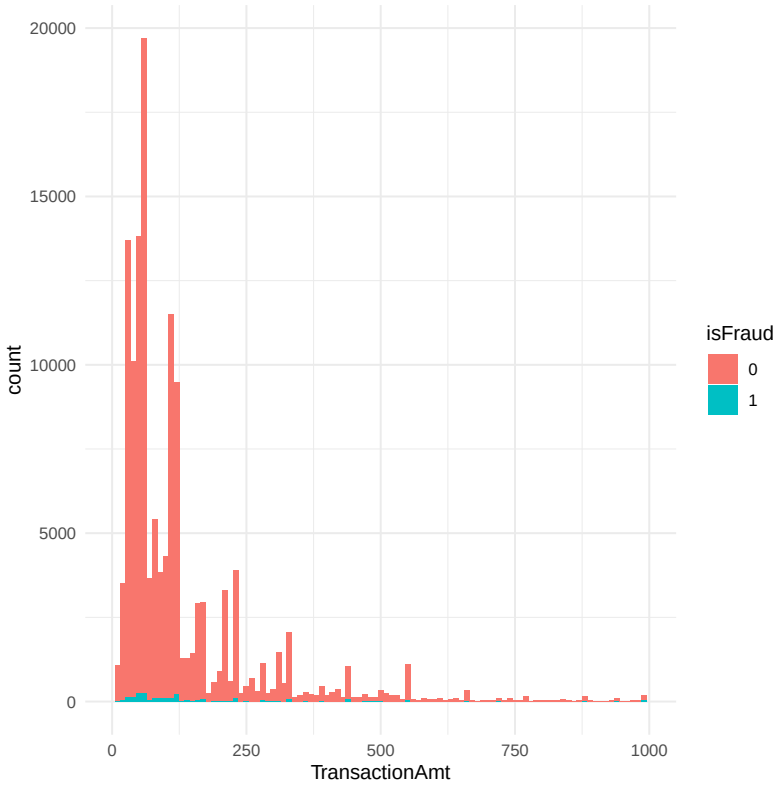


Figure 1: Histogram zneskov transakcij

Card

***

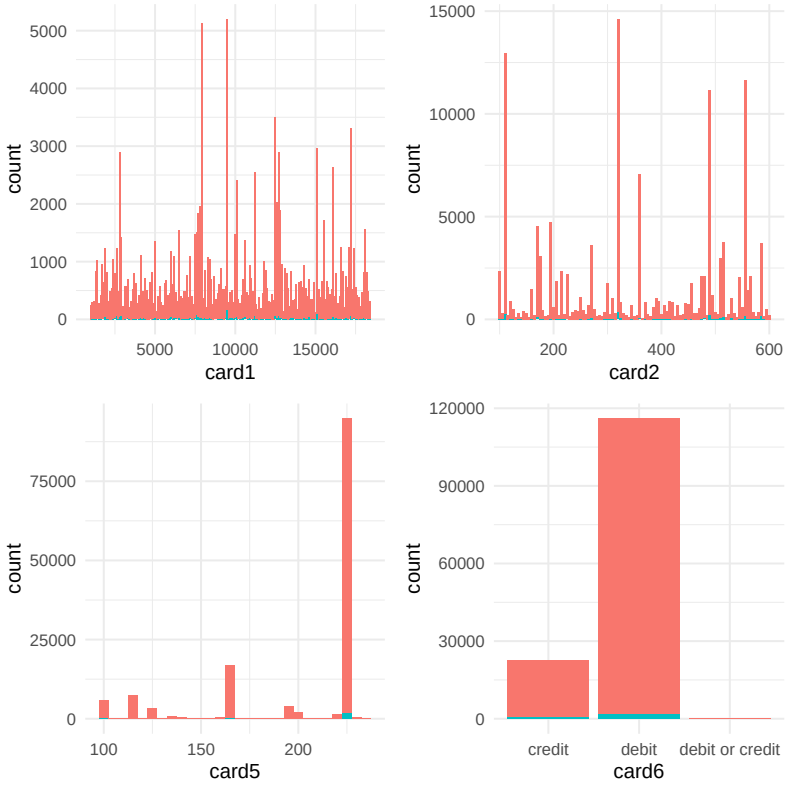


Figure 2: Histogrami spremenljivk kategorije card

Email domain

***

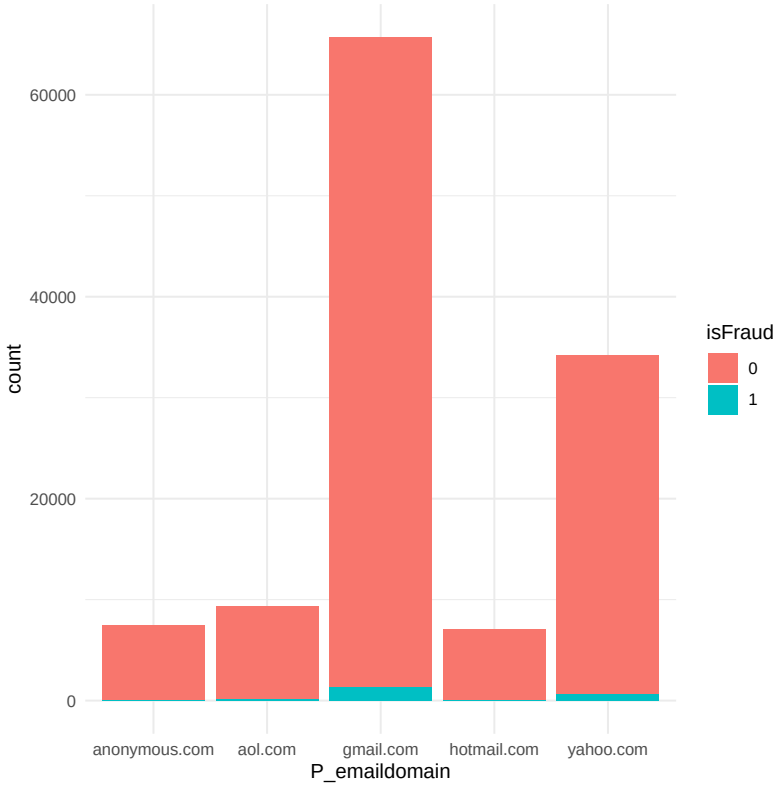


Figure 3: Stolpicni diagram email domen placnika

C

***

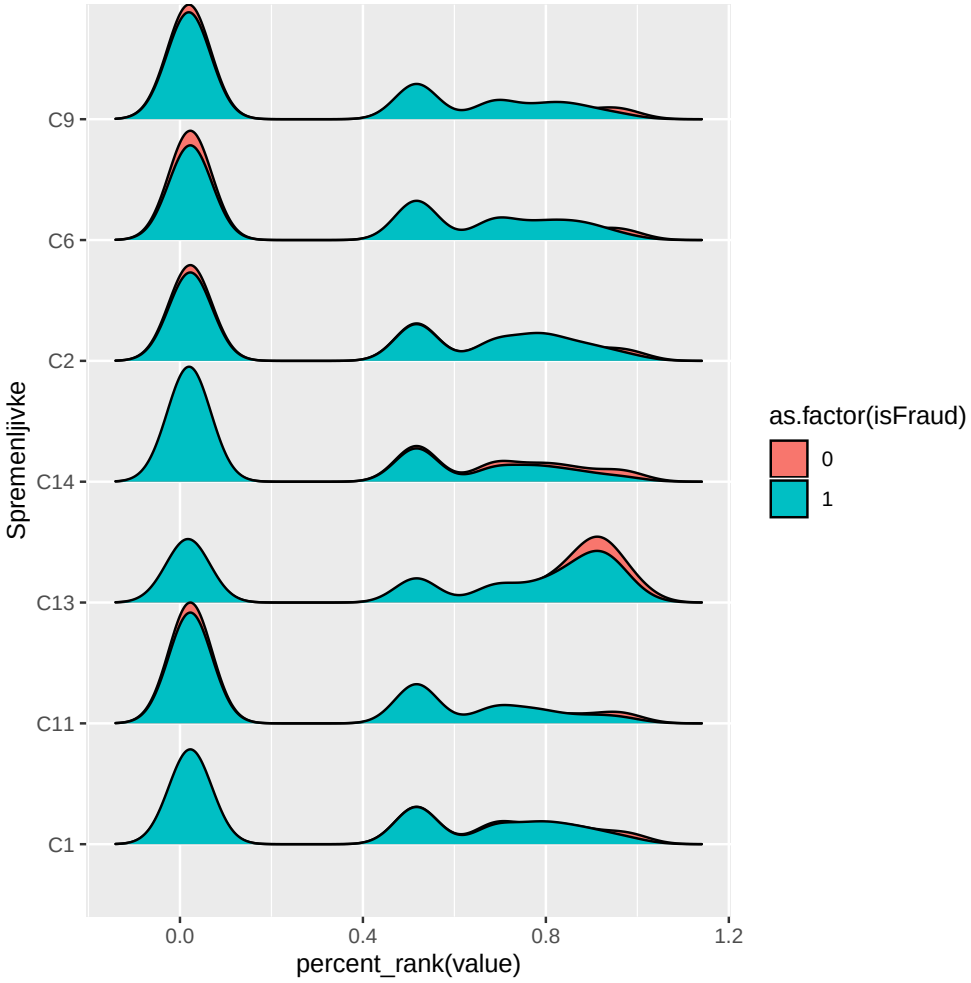


Figure 4: Empiricina porazdelitev spremenljivk kategorije C

V

***

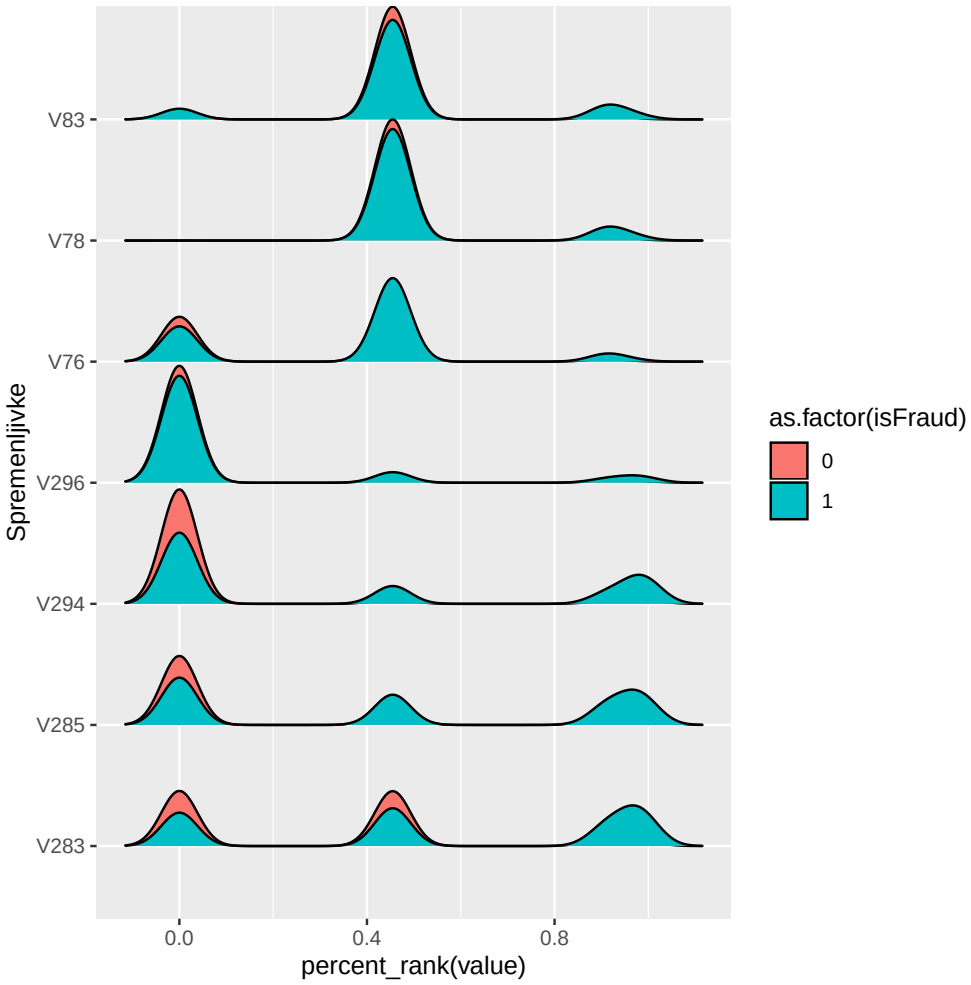


Figure 5: Empiricina porazdelitev spremenljivk kategorije V

Correlation

***

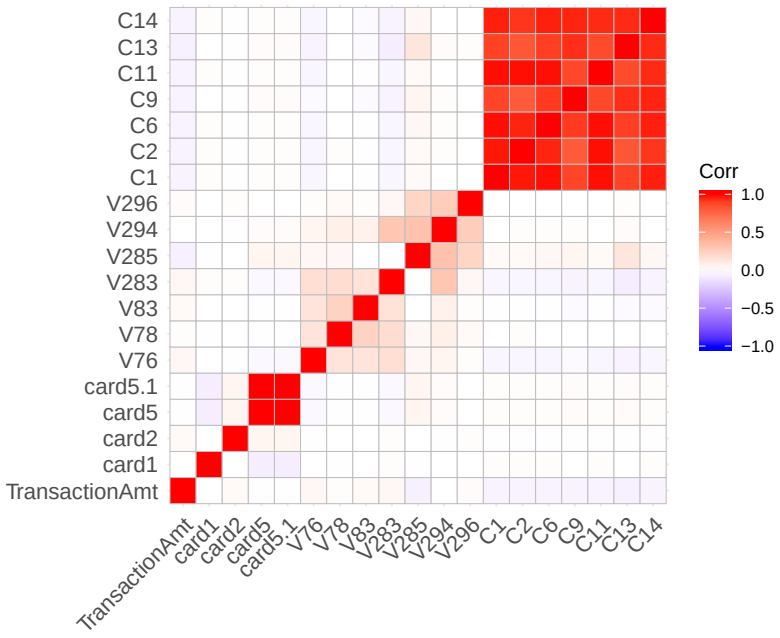


Figure 6: Korelacijska matrika med vsemi izbranimi spremenljivkami

3 Rezultati

Ogledali si bomo rezultate štirih razlicnih modelov in dveh na podlagi glasovanja. Pri delu z neuravnoteženimi podatki se moramo zavedati zavedljivosti klasifikacijske tocnosti. Zato je potrebno, da poznamo dve številki in sicer delež pravilno uvršcenih enot v primeru, da vse enote razvrstimo v skupino 0 (ni prevara) in delež pravilno razvršcenih enot v primeru nakljucnega razvršcanja ob predpostavki, da so populacijski deleži enaki vzorcnim. Naši referencni vrednosti sta torej 0.98 in 0.96. V takšnem primeru je pomemben tudi tip napake oziroma ali se gre za False Positive ali False Negative. Zato bomo poleg klasifikacijske tocnosti primerjali tudi Precision (manjhna vrednost nakazuje na veliko število False Positive napak) in Recall (manjhna vrednost nakazuje na veliko število False Negative napak). Ogledali si bomo F1 Score, ki združi prej omenjeni metriki. Vsi modeli so bili ocenjeni na standariziranih podatkih, kakovost modela pa je bila vrednotena na loceni, testni množici podatkov, ki ni bila del eksploratorne analize in ocenjevanja modela.

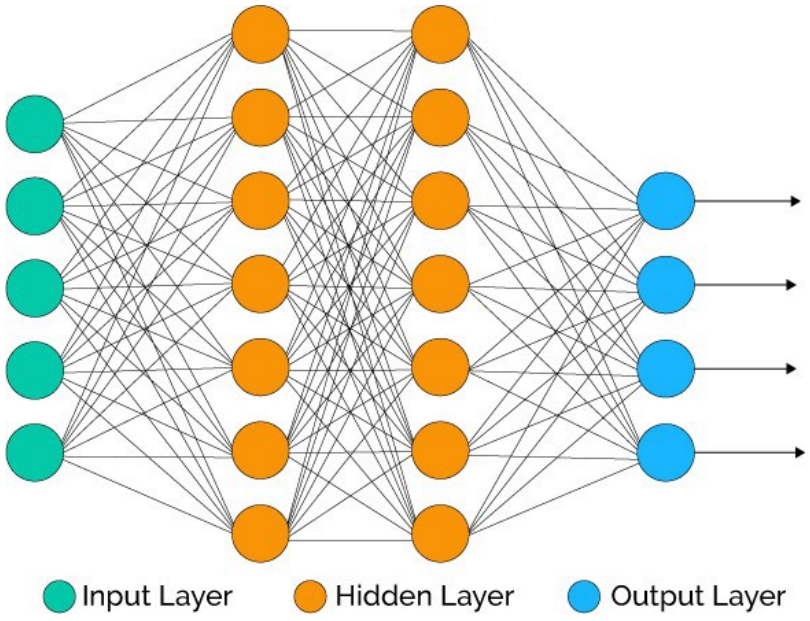
3.1 Logistic regression

3.2 kNN

3.3 Globoke nevronske mreže

Globoke nevronske mreže so vecnivojske mreže z vec skritimi nivoji, ki omogocajo vecjo kompleksnost modela. Nevronske mreže so priljubljen alogritem klasifikacije, katera ideja izhaja iz delovanja možganov. Vsaka posamezna enota nevronske mreže ima v prvi fazi dodeljene poljubne uteži (bodisi nakljucne ali s pridobljene s pomocjo specificne inicializacije). Vhodni podatki so nato propagirani skozi posamezne nivoje nevronske mreže kjer se na vsakem nivoju ponovljena operacija matricnega množenja z utežim trenutnega nivja ter apliciranje aktivacijske funkcije. Vloga aktivacijske funkcije je, da v model vnaša nelinearno preslikavo prejšnega nivja in s tem omogoca kompleksnejši model. Hkrati zalogo vrednosti preslika na omejen interval, najpogosteje med 0 in 1. Ko enkrat izracunamo izhod vseh nivojev nevronske mreže sledi proces optimizacije kriterijske funkcije. V tem procesu s pomocjo optimizacijskih metod prilagjamo modelske uteži in s tem minimiziramo napako. Minimizacijo napake izvajamo v obratnem vrstnem redu glede na strukturo modela kot je pretok podatkov, torej iz zadnje (izhodne) plasti se pomikamo proti zacetni (vhodni) plasti. Pomemben del ucenja modela je tudi proces regularizacije s katerim poskrbimo, da pri modelu ne pride do preprileganja. To naredimo tako, da pri kriterijski funkciji dodamo dodaten regularizacijski clen (L1, L2) ali pa z metodo dropout, kjer v postopku ucenja ob vsaki iteraciji uteži nakljucnih nevronom postavimo na nic.

Naš model je sestavljen iz štirih nivojev, vhodni, izhodni ter dva skrita (15, 10, 5, 1). Pri tem se pri vseh nivojih uporabili aktivacijsko funkcijo hiperpolicni tangens razen pri izhodni, kjer sem uporabil sigmoidno aktivacijsko funkcijo. Za optimizacijo sem uporabil algoritem **adam**, ki opzimiriza binarno precno entropijo. Za regularizacijo sem uporabil postopek dropout pri prvem in drugem nivoju, kjer ob vsaki iteraciji "ugasne" 20% nevronov. Zaradi velikega števila podatkov sem pri procesu ucenja uporabil treniranje na manjših delih podatkov (minibatch) velikost 128 transakcij. Število iteracij ucenja (epoch) sem nastavil na 100, a sem omogocil možnost zgodnjega ustavljanja v primeru, ce se vrednost kriterijske funkcije ne niža vec.



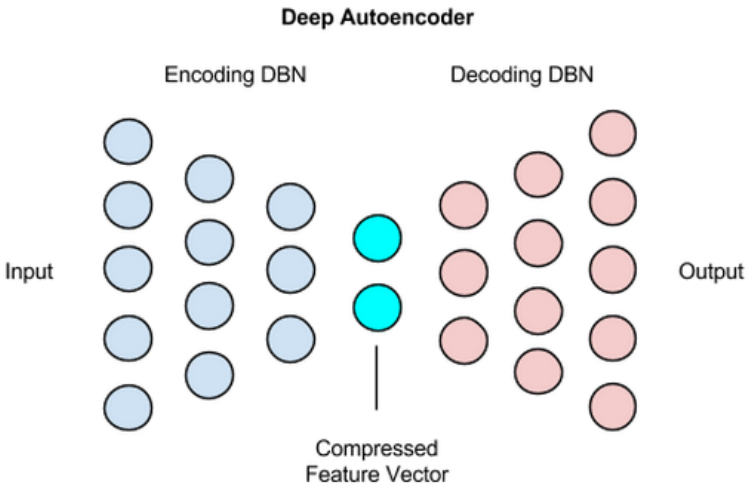
V spodnji tabeli vidimo, da v splošnem klasifikacijska tocnost ni boljša kot, ce vse enote klasificiramo kot normalne transakcije, je pa boljša v primeru slucajnega razvrščanja. F1 score je zelo majhen predvsem zaradi majhne vrednosti recalla, torej bo vecina napak tipa False Negative.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 33988   683
         1     3    11

               Accuracy : 0.9802
                 95% CI : (0.9787, 0.9817)
    No Information Rate : 0.98
    P-Value [Acc > NIR] : 0.389

                  Kappa : 0.0303
 Mcnemar's Test P-Value : <2e-16

              Precision : 0.7857143
                 Recall : 0.0158501
                     F1 : 0.0310734
             Prevalence : 0.0200086
         Detection Rate : 0.0003171
   Detection Prevalence : 0.0004036
      Balanced Accuracy : 0.5078809

       'Positive' Class : 1
```

3.4 Autoencoder

Autoencoderji so nevronske mreže s katerimi se lahko naucimo latentno reprezentacijo (encoding) poljubnega podatkovnega seta. Tradicionalno so bili vecinoma uporabljeni z namenom zmanjševanja dimenzij podatkov, trenutno pa so aktulani tudi na podrocju generativnih modelov. V principu delujejo tako, da skozi plasti nevronske mreže zmanjšamo dimenzijo podatkov (kodirnik), ter nato na podlagi te latentne reprezenzacije vhodne podatke rekonstruiramo s cim manjšo napako (dekodirnik). Da model deluje je potrebna predpostavka, da so porazdelitve spremenljivk transakcij pri katerih je prisotna prevara drugacne od normalnih. Ideja pri uporabi za klasifikacijo pri neuratnoteženih podatkih je naslednja: saj imamo veliko število normalnih transakcij se naucimo latentno reprezentacijo teh. Ko bomo v z modelom napovedovali transakcije, ob upoštevanju predpostavke pricakujemo, da bo napaka pri rekonstrukciji normalnih transakcij manjša kot, ko je prisotna prevara. Dolociti moramo še mejno vrednost napake, na podlagi katere bomo klasificirali transakcije. Vrednost napake dolocimo glede na izbrano metriko, v našem primeru bo to F1.

Naš model je sestavljen iz petih nivojev, vhodni, izhodni ter trije skriti (15, 10, 5, 10, 15). Pri tem se pri vseh nivojih uporabili aktivacijsko funkcijo hiperpolicni tangens. Za optimizacijo sem uporabil algoritem **adam**, ki opzimiriza povprecen kvadrat napake (MSE). Za regularizacijo sem uporabil postopek dropout za nivojema z desetimi nevroni, kjer ob vsaki iteraciji "ugasne" 20% nevronov. Zaradi velikega števila podatkov sem pri procesu ucenja uporabil treniranje na manjših delih podatkov (minibatch) velikost 128 transakcij. Število iteracij ucenja (epoch) sem nastavil na 100, a sem omogocil možnost zgodnjega ustavljanja v primeru, ce se vrednost kriterijske funkcije ne niža vec.



V spodnji tabeli vidimo, da je klasifikacijska tocnost zelo slaba (0.7049) in ni boljša niti, ce vse enote klasificiramo kot normalne transakcije, niti v primeru slucajnega razvrščanja. F1 score je zelo majhen predvsem zaradi majne vrednosti preciznosti, torej bo vecina napak tipa False Positive.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 24126   370
         1  9865   324

               Accuracy : 0.7049
                 95% CI : (0.7001, 0.7097)
    No Information Rate : 0.98
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0229
 Mcnemar's Test P-Value : <2e-16

              Precision : 0.031799
                 Recall : 0.466859
                     F1 : 0.059542
             Prevalence : 0.020009
         Detection Rate : 0.009341
   Detection Prevalence : 0.293758
      Balanced Accuracy : 0.588317

       'Positive' Class : 1
```

3.5 Ensamble

Tak tip klasifikatorja združuje rezultate vecih modelov. Zdržili jih bomo na dva nacina, gleda na vecinski delež glasov (zaradi sodega števila modelov bomo v primeru deleža 0.5 transakcijo klasificirali kot prevaro) ali z konsenzom vseh glasov. V primeru konsenza bomo transakcijo kot prevaro klasificirali le v primeru, ce jo za prevaro oznacijo vsi štirje algoritmi.

Vidimo, da v primeru vecinskega glasu je klasifikacijska tocnost statisticno znacilno boljša od referencne. Vecina napak je False Negative. V primeru konsenza klasifikacijska tocnost ni statisticno znacilno razlicna kot tista, ce bi vse enote klasificirali kot normalne. Opazimo, da False Positive napaka ni vec prisotna vendar pravilno klasificiramo le eno prevaro.

Vecinsko glasovanje

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 33990   473
         1     1   221

               Accuracy : 0.9863
                 95% CI : (0.9851, 0.9875)
    No Information Rate : 0.98
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4775
 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.995495
                 Recall : 0.318444
                     F1 : 0.482533
             Prevalence : 0.020009
         Detection Rate : 0.006372
   Detection Prevalence : 0.006400
      Balanced Accuracy : 0.659207

       'Positive' Class : 1
```

Konsenz

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 33991   693
         1     0     1

               Accuracy : 0.98
                 95% CI : (0.9785, 0.9815)
    No Information Rate : 0.98
    P-Value [Acc > NIR] : 0.4948

                  Kappa : 0.0028
 Mcnemar's Test P-Value : <2e-16

              Precision : 1.000e+00
                 Recall : 1.441e-03
                     F1 : 2.878e-03
             Prevalence : 2.001e-02
         Detection Rate : 2.883e-05
   Detection Prevalence : 2.883e-05
      Balanced Accuracy : 5.007e-01

       'Positive' Class : 1
```

4 Zakljucek

Glede na klasifikacijsko tocnost je najslabši model autoencoder. Najbolj podobna sta si modela logisticne regresije in nevronskih mrež, zelo blizu pa jima je tudi model, ko napovedi dolocamo s konsenzom, le da je njegov recall precej manjši od prej navedenih modelov. KNN ima najboljše razmerje med preciznostjo in recallom in posledicno najboljši F1 score. Zelo blizu mu je tudi, model, ko napovedi dolocamo na podlagi vecinskega glasu. Ta model bi oznacil tudi kot najboljši, saj edini presega referecno klasifikacijsko tocnost, ter ima dobro razmerje med preciznostjo in recallom.

	logistic	knn	neuralnetwork	autoencoder	majority	consensus
Accuracy	0.980	0.982	0.980	0.705	0.986	0.980
Precision	0.652	0.555	0.786	0.032	0.995	1.000
Recall	0.022	0.510	0.016	0.467	0.318	0.001
F1	0.042	0.532	0.031	0.060	0.483	0.003

Table 1: Metrike koncnih modelov