Analiza skupa podataka *Online Shoppers Purchasing Intention*

Seminarski rad iz kursa Istraživanje podataka

Olivera Popović

25.6.2019.

Analiza skupa podataka *Online Shoppers Purchasing Intention*

Uvod

Uvod

- **Online Shoppers Purchasing Intention** skup podataka sadrži informacije o aktivnostima korisnika na internetu, a podaci su prikupljeni sa sajta za maloprodaju.
- Sastoji se iz 12330 pristupa korisnika.

Opis i analiza skupa podataka

Opis skupa podataka

- Skup podataka *Online Shoppers Purchasing Intention* sadrži 10 numeričkih i 8 kategoričkih atributa.

Numerički atributi

Naziv atributa	Opis atributa	Min	Max	Std. dev
Administrative	Broj posećenih veb strana o upravljanju profilom	0	27	3.32
Administrative Duration	Vreme (u sekundama) provedeno na veb stranama o upravljanju profilom	0	3398	176.70
Informational	Broj posećenih veb strana sa informacijama o veb sajtu, komunikaciji i adresi sajta za kupovinu	0	24	1.26
Informational Duration	Vreme (u sekundama) provedeno na veb stranama za informacije	0	2549	140.64
Product Related	Broj posećenih veb strana vezanih za proizvode	0	705	44.45
Product Related Duration	Vreme (u sekundama) provedeno na veb stranama vezanim za proizvode	0	63973	1912.25
Bounce Rate	Procenat korisnika koji nakon ulaska na veb sajt izađu bez pokretanja drugih zahteva ka serveru	0	0.2	0.04
Exit Rate	Procenat koliko je puta veb strana bila poslednja u jednom pristupu korisnika internetu, u odnosu na ukupan broj pregleda	0	0.2	0.05
Special Day	Pokazuje koliko je vreme posete veb sajtu blizu nekog specijalnog dana u godini, u kojima je veća verovatnoća da se uspešno izvrši transakcija	0	1.0	0.19

Slika 1: Numerički atributi

Kategorički atributi

Naziv atributa	Opis atributa	Broj kategoričkih vrednosti
Operating Systems	Operativni sistem korisnika	8
Browser	Veb pregledač korisnika	13
Region	Geografski region iz kog se korisnik prijavio	9
Traffic Type	Izvor, odakle je korisnik pristupio veb sajtu	20
Visitor Type	Tip korisnika koji može biti: “New Visitor”, “Returning Visitor” i “Other”	3
Weekend	Indikator da li je datum posete vikend ili nije	2
Month	Mesec u kom je korisnik pristupio veb sajtu	12
Revenue	Oznaka klase koja pokazuje da li je poseta bila završena transakcijom ili ne	2

Slika 2: Kategorički atributi

Analiza skupa (korišćena tehnika)

- Za analiziranje skupa podataka biće korišćen programski jezik Python, kao i biblioteke: Numpy, Pandas, Seaborn, Matplotlib i Sklearn.

Rezultat izvršavanja (opis skupa)

	Administrative	Administrative_Duration	...	Weekend	Revenue
10842	0	0.000	...	False	False
7804	6	132.025	...	False	False
8540	4	28.500	...	True	False
7346	0	0.000	...	False	False
2014	2	36.000	...	False	False
5731	0	0.000	...	False	True
1230	0	0.000	...	False	False
12025	0	0.000	...	True	False
9468	0	0.000	...	False	False
1387	3	63.000	...	True	False

[10 rows x 18 columns]

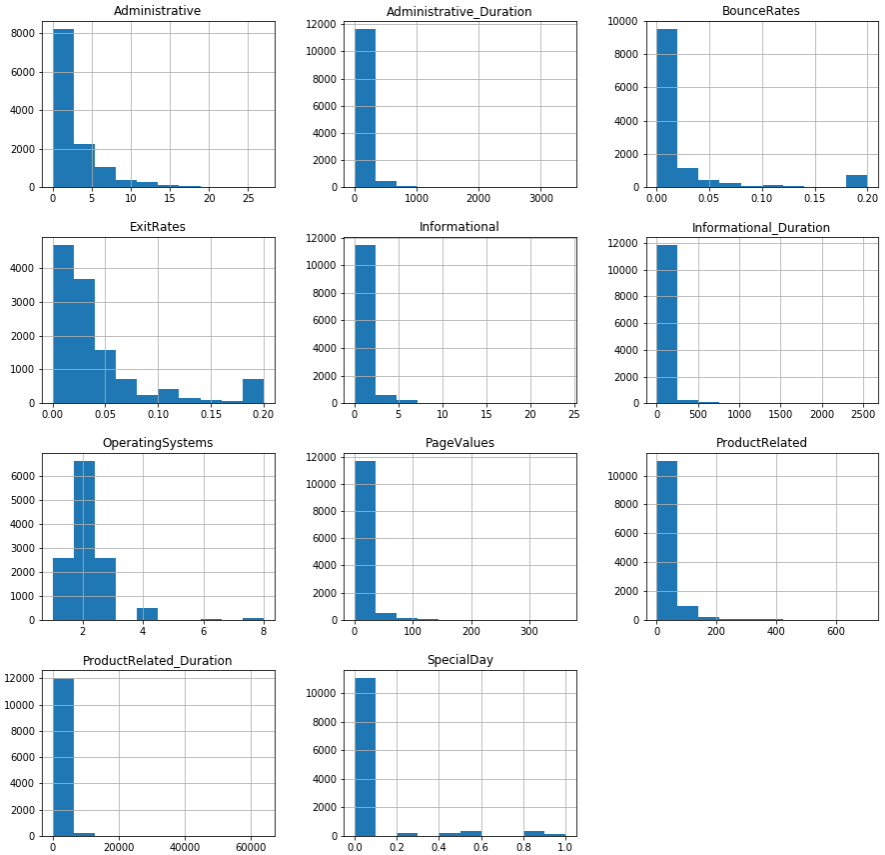
	Administrative	Administrative_Duration	...	Region	TrafficType
count	12330.000000	12330.000000	...	12330.000000	12330.000000
mean	2.315166	80.818611	...	3.147364	4.069586
std	3.321784	176.779107	...	2.401591	4.025169
min	0.000000	0.000000	...	1.000000	1.000000
25%	0.000000	0.000000	...	1.000000	2.000000
50%	1.000000	7.500000	...	3.000000	2.000000
75%	4.000000	93.256250	...	4.000000	4.000000
max	27.000000	3398.750000	...	9.000000	20.000000

Slika 3: Opis skupa

Rezultat izvršavanja (opis skupa)

- Iz rezultata izvršavanja vidi se da su podaci retki i da su atributi različito skalirani, što će biti rešeno u fazi preprocesiranja podataka.

Raspodela nekih atributa

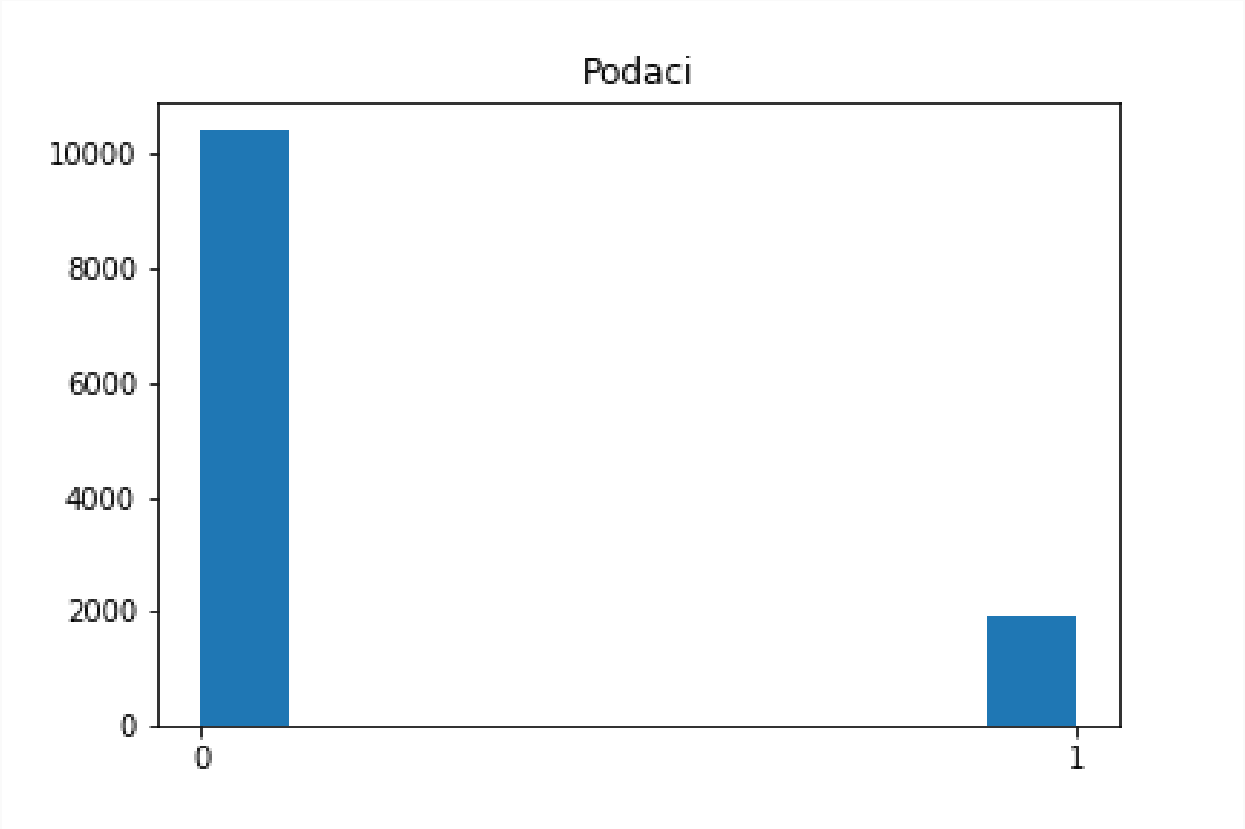


Slika 4: Histogrami raspodela nekih atributa

Ciljna promenljiva

- Ciljna promenljiva u procesu klasifikacije biće *Revenue*, odnosno da li je korisnik kupio neki proizvod ili nije, jer je to informacija koju želimo da dobijemo nakon istraživanja.
- Postoje dve klase, klasa *True* i klasa *False*. Klasi False pripada 10422 instance, dok klasi True pripada preostalih 1908 instanci.
- Može se primetiti da klase nisu balansirane, što će predstavljati problem u procesu klasifikacije.

Ciljna promenljiva



Slika 5: Histogram zastupljenosti klasa

Preprocesiranje podataka

Rad sa nedostajućim vrednostima

- Potrebno je proveriti da li u podacima postoje nedostajuće vrednosti i ako postoje treba ih na odgovarajući način ukloniti, odnosno zameniti.
- Iz rezultata izvršavanja može se primetiti da među podacima nema nedostajućih vrednosti.

Rad sa nedostajućim vrednostima

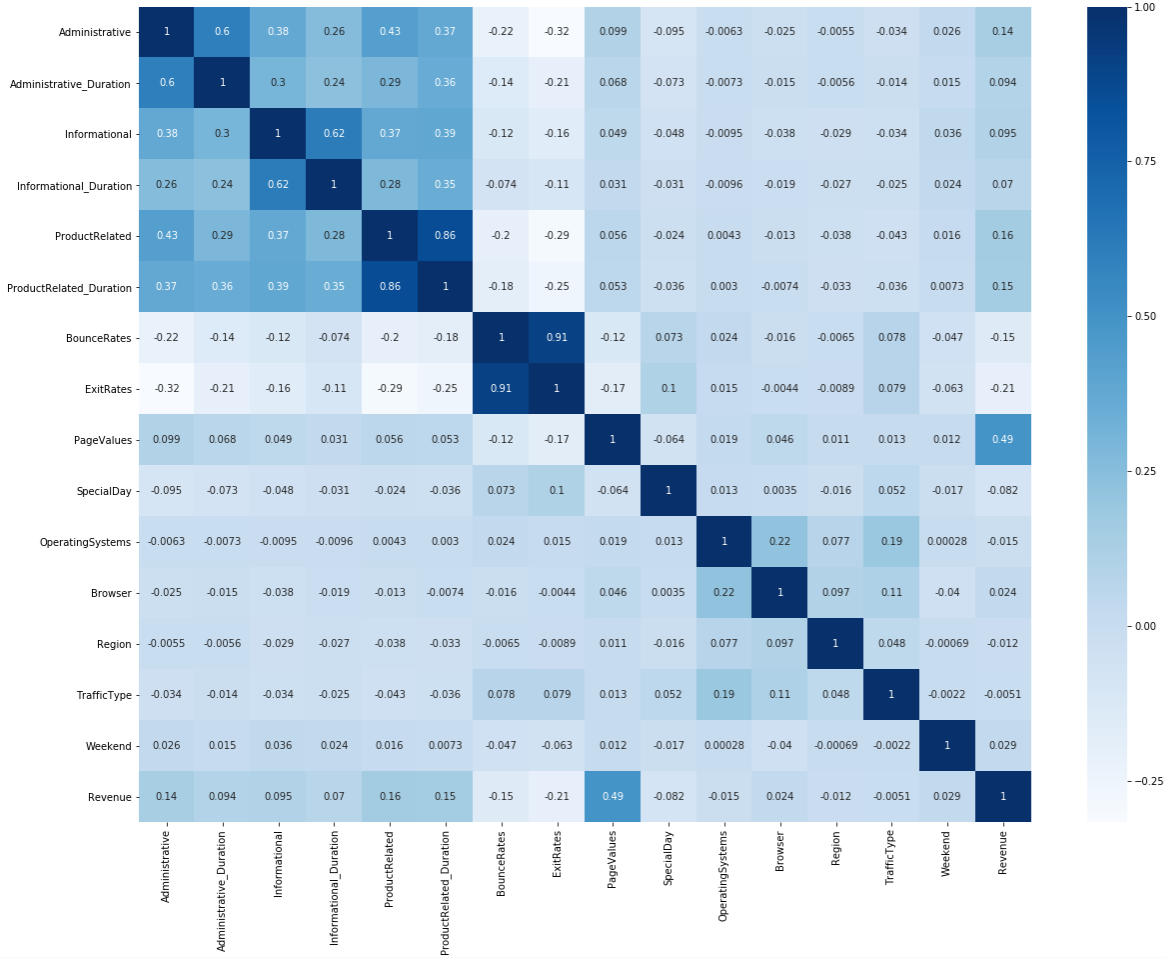
```
Nedostajuce vrednosti u podacima:

Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0
dtype: int64
```

Slika 6: Rezultat izvršavanja programskog koda

Početni izbor atributa

- Da bismo odabrali atribute potrebna nam je matrica korelacije. Matricom korelacije dobićemo informacije o tome koliko atributi utiču jedni na druge, odnosno koliko su korelirani.
- Ovo će biti samo početni odabir jer će kasnije atributi biti ponovo birani, kao pokušaj da se unapredi rezultat algoritama.

Početni izbor atributa



Slika 7: Matrica korelacije

Početni izbor atributa

- Prateći tamno plavu boju može se uočiti da atributi *Exit Rates* i *Bounce Rates* veoma utiču jedan na drugi, te da jedan možemo isključiti. Slično je i sa *Product Related* i *Product Related Duration* atributima.

Transformacija kategoričkih atributa

- Da bi se mogla vršiti izračunavanja u algoritmima klasifikacije potrebno je da atributi budu numeričkog tipa.
- Biće izvršena **binarizacija**. Ako kategorički atribut ima n različitih vrednosti formira se n novih, različitih binarnih atributa. Svaki binarni atribut odgovaraće jednoj mogućoj vrednosti kategoričkog atributa. U jednom redu tačno jedan od n atributa imaće vrednost 1, dok će ostali imati vrednost 0.
- Binarni kategorički atributi koji uzimaju logičke vrednosti *True* ili *False* biće zamenjeni sa 0 i 1, jer su binarni podaci specijalni slučaj i kategoričkih i numeričkih atributa.

Novi izgled skupa

Primer atributa koji su dobijeni od atributa *Operating Systems*:

```
Novi atributi:
['OperatingSystems_1', 'OperatingSystems_2', 'OperatingSystems_3', 'OperatingSystems_4',
'OperatingSystems_5', 'OperatingSystems_6', 'OperatingSystems_7', 'OperatingSystems_8']

Novi izgled skupa:
   Administrative  Administrative_Duration  ...  Weekend_False  Weekend_True
0               0                     0.0  ...              1             0
1               0                     0.0  ...              1             0
2               0                     0.0  ...              1             0
3               0                     0.0  ...              1             0
4               0                     0.0  ...              0             1

[5 rows x 74 columns]
```

Slika 8: Rezultat izvršavanja programskog koda

Standardizacija

- Kao što je prethodno napomenuto, atributi su različito skalirani, što znači da ih je nemoguće međusobno upoređivati. Kako bi se rešio ovaj problem biće korišćena **standardizacija**.
- To se postiže tako što se od atributa oduzme njegova srednja vrednost i to se podeli njegovom standardnom devijacijom, odnosno: $X_s = \frac{X-\mu}{\delta}$.
- Pre nego što se izvrši standardizacija potrebno je podeliti skup podataka na trening i test skup, koji će biti korišćeni u procesu klasifikacije, kako bi se izbegao uticaj podatak iz test skupa na statistike trening skupa.

Klasifikacija

Logistička regresija

- Prvi algoritam koji će biti primenjen na skup podataka je logistička regresija, upravo zato što je jedna od najkorišćenijih metoda. Jednostavna je i pruža efikasno treniranje.
- Ova metoda primenljiva je samo na binarnu klasifikaciju , što jeste slučaj u opisanom skupu podataka.
- Alternativa bi bila koršćenje pristupa jedan protiv svih.

Prvi pokušaj

- Za odabir regularizacionog parametra korišćen je Grid Search algoritam iscrpne pretrage, koji se koristi za pronalaženje optimalnih parametara za prosleđeni metod. Algoritam za evaluaciju koristi unakrsnu validaciju.

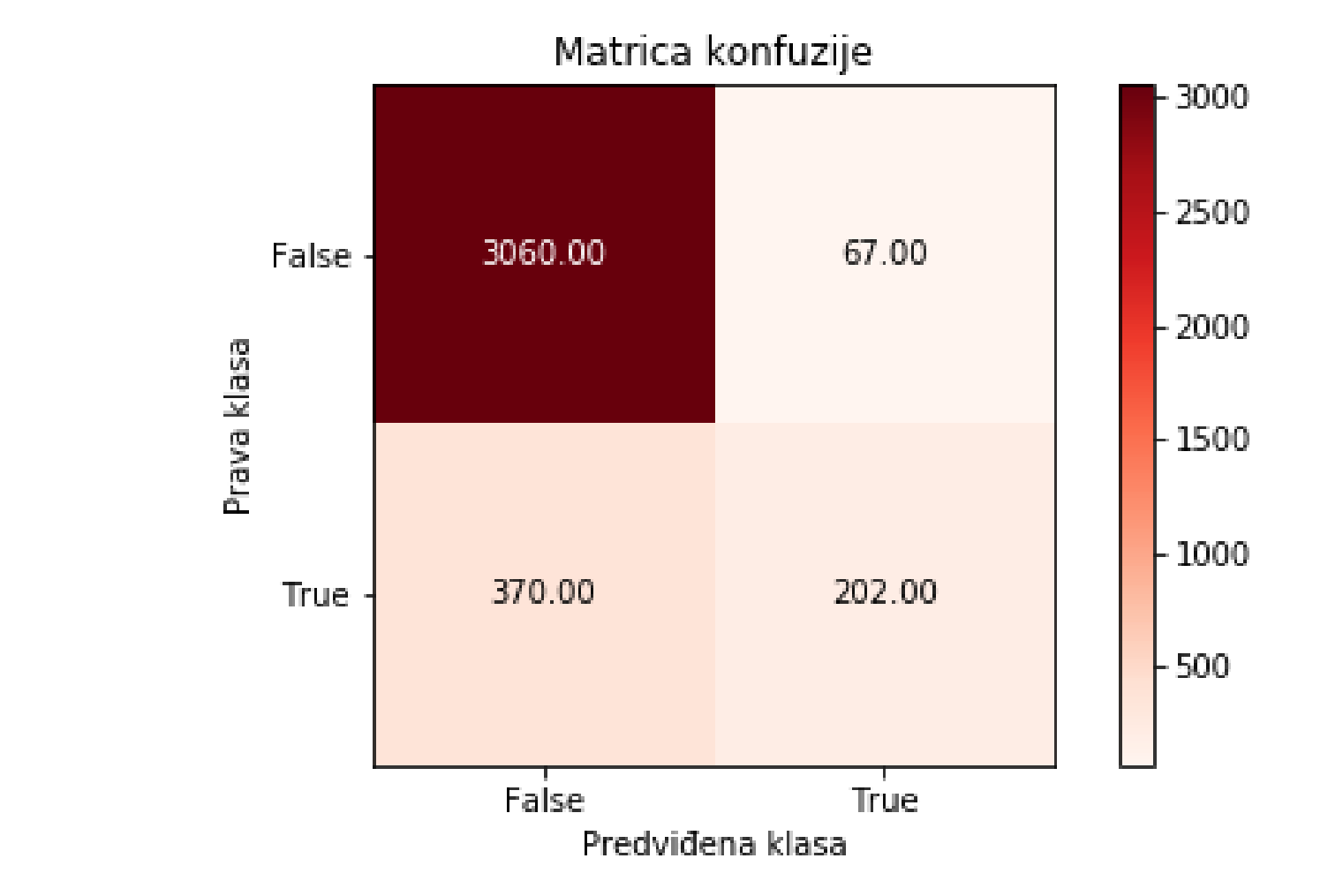
Evaluacija dobijenog modela

- Tačnost na trening skupu: 0.8870351060132082
- Tačnost na test skupu: 0.881859962151933

	precision	recall	f1-score	support
0	0.89	0.98	0.93	3127
1	0.75	0.35	0.48	572
accuracy			0.88	3699
macro avg	0.82	0.67	0.71	3699
weighted avg	0.87	0.88	0.86	3699

Slika 9: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 10: Matrica konfuzije

Popravka performansi

- Iz prethodnih rezultata može se zaključiti da se dobijaju znatno lošiji rezultati za klasifikaciju podataka koji pripadaju klasi tačno, jer u toj klasi postoji manje podataka.
- Balansiranjem klasa trebalo bi da se popravi taj rezultat.

Balansiranje klasa

- Za balansiranje klasa biće korišćena *SMOTE (Synthetic Minority Oversampling TEchnique)* tehnika, koja počiva na sintezi elemenata manje klase od onih koji već postoje u skupu.
- Takođe, biće primenjen i algoritam PCA, kako bi se smanjila dimenzionalnost i time povećala efikasnost algoritma.

Rezultat izvršavanja (balansiranje klasa)

- Balansiranje izvršeno na test skupu:

```
Broj instanci u klasi "1" pre balansiranja: 1336
Broj instanci u klasi "0" pre balansiranja: 7295

Broj instanci u klasi "1" nakon balansiranja: 7295
Broj instanci u klasi "0" pre balansiranja: 7295
```

Slika 11: Rezultat balansiranja

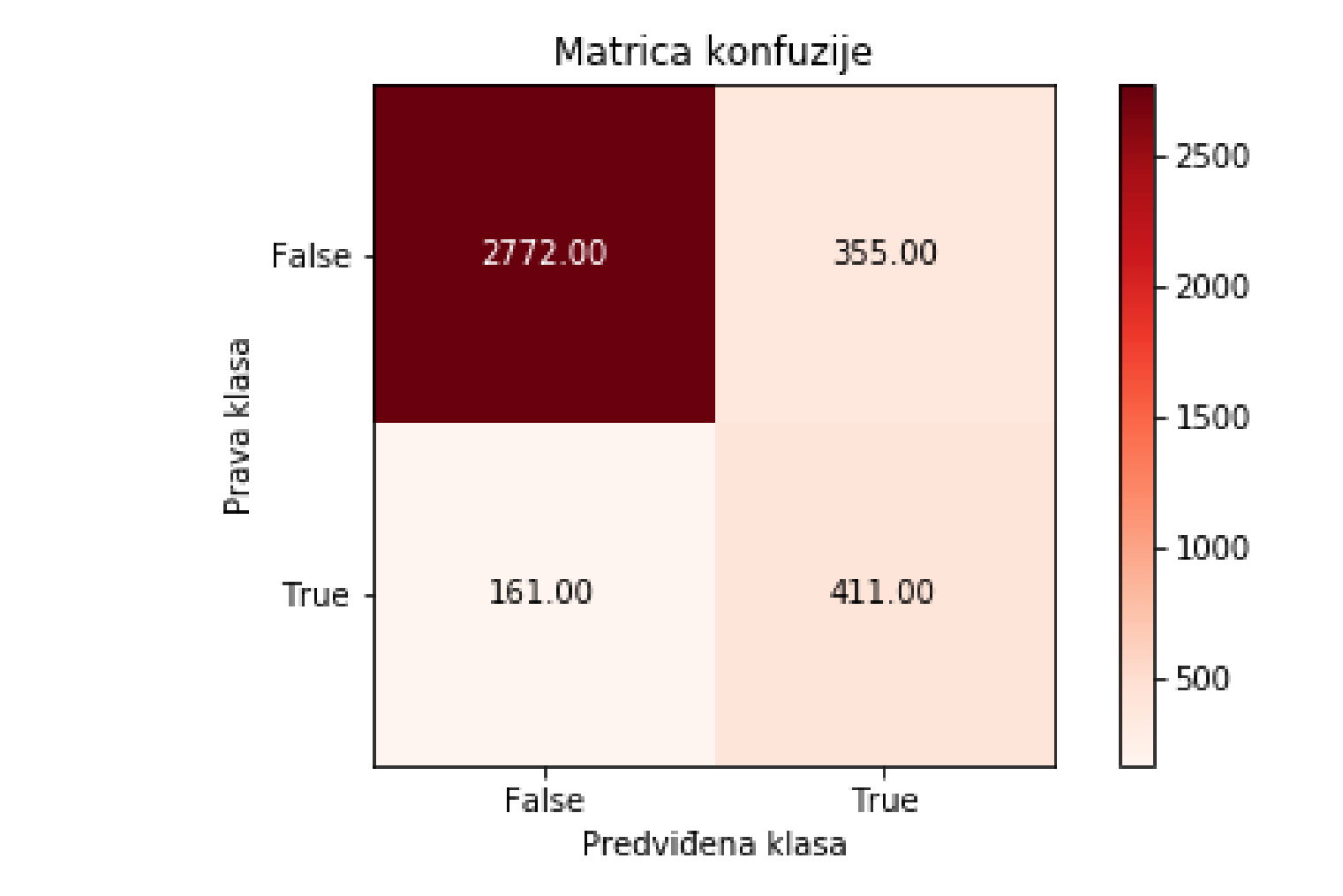
Evaluacija dobijenog modela

- Train score: 0.8503084304318026
- Test score: 0.8605028386050284

	precision	recall	f1-score	support
0	0.95	0.89	0.91	3127
1	0.54	0.72	0.61	572
accuracy			0.86	3699
macro avg	0.74	0.80	0.76	3699
weighted avg	0.88	0.86	0.87	3699

Slika 12: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 13: Odgovarajuća matrica konfuzije

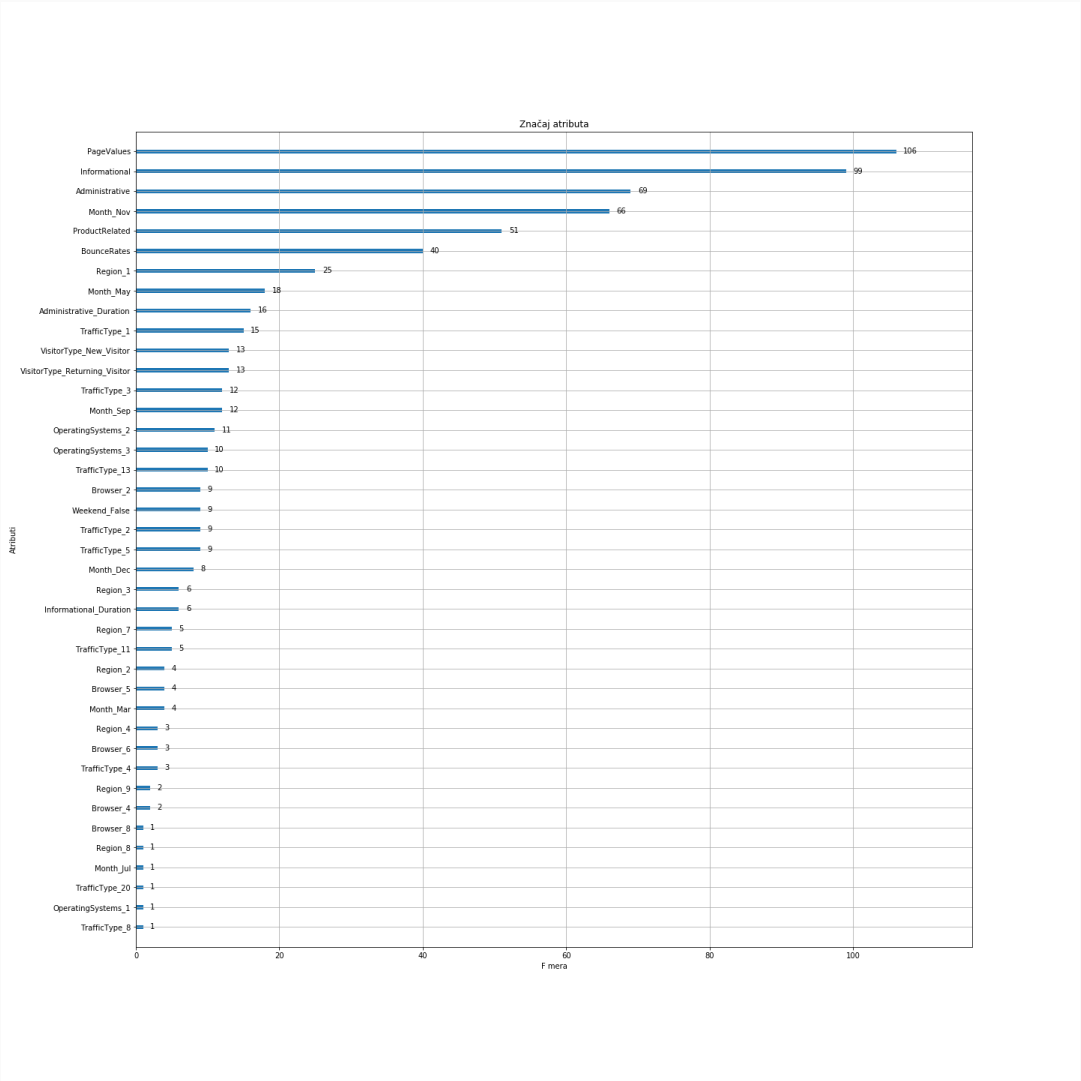
Analiza

- Na prvi pogled drugi model deluje lošiji, jer je dobijena manja ukupna tačnost. Međutim, odziv za klasu ‘1’ je sada čak duplo bolji, što znači da će model manje grešiti kada je ta klasa u pitanju.

Promena atributa

- Sledeće što će biti pokušano, zajedno sa svim prethodnim, jeste promena atributa koji se koriste. Za određivanje najbitnijih atributa biće korišćen algoritam XGBoost.
- XGBoost je jedan od najpopularnijih algoritama danas za regresiju i klasifikaciju, a baziran je na stablima odlučivanja. U ovom slučaju biće korišćen za selekciju važnih atributa.

Rezultatl algoritma XGBoost



Slika 14: Grafik značajnosti atributa

Izbor atributa

- Na osnovu grafika može se zaključiti da atributi *Informational Duration*, *Browser* i *Weekend* nisu mnogo značajni, pa ćemo ih isključiti.
- Atribut *Special Day* ima iznenađujuće malu značajnost, jer intuitivno veoma utiče na kupovinu, ali ćemo ga ipak isključiti.

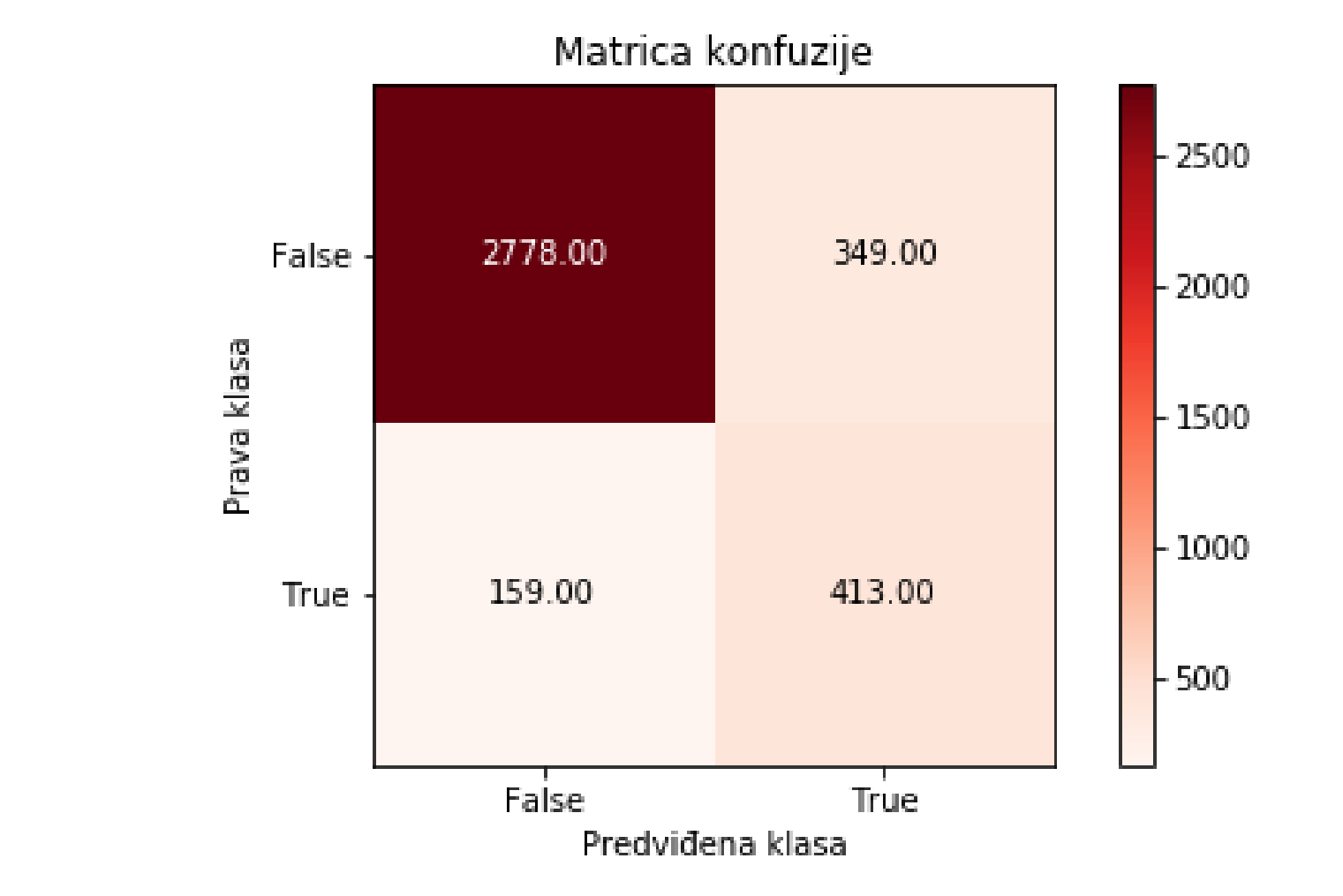
Evaluacija dobijenog modela

- Train score: 0.8465387251542152
- Test score: 0.8626655852933225

	precision	recall	f1-score	support
0	0.95	0.89	0.92	3127
1	0.54	0.72	0.62	572
accuracy			0.86	3699
macro avg	0.74	0.81	0.77	3699
weighted avg	0.88	0.86	0.87	3699

Slika 15: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 16: Matrica konfuzije sa novim izborom atributa

Analiza

- Može se primetiti da se dobijaju neznatno bolji rezultati, tako da će nadalje biti korišćen stari skup, jer je bogatiji, odnosno nosi više informacija.

Stabla odlučivanja

- Sledeći algoritam koji će biti primenjen na opisani skup jesu stabla odlučivanja. To je metoda u kojoj se proces klasifikacije modeluje pomoću skupa hijerarhijskih odluka donetih na osnovu atributa trening podataka, uređenih u obliku drvolike strukture.
- Kao kriterijum podele biće korišćena Entropija, a čvorovi će se nalaziti na dubini najviše 7.
- Parametri su, kao i kod Logističke regresije, odabrani korišćenjem Grid Search algoritma.

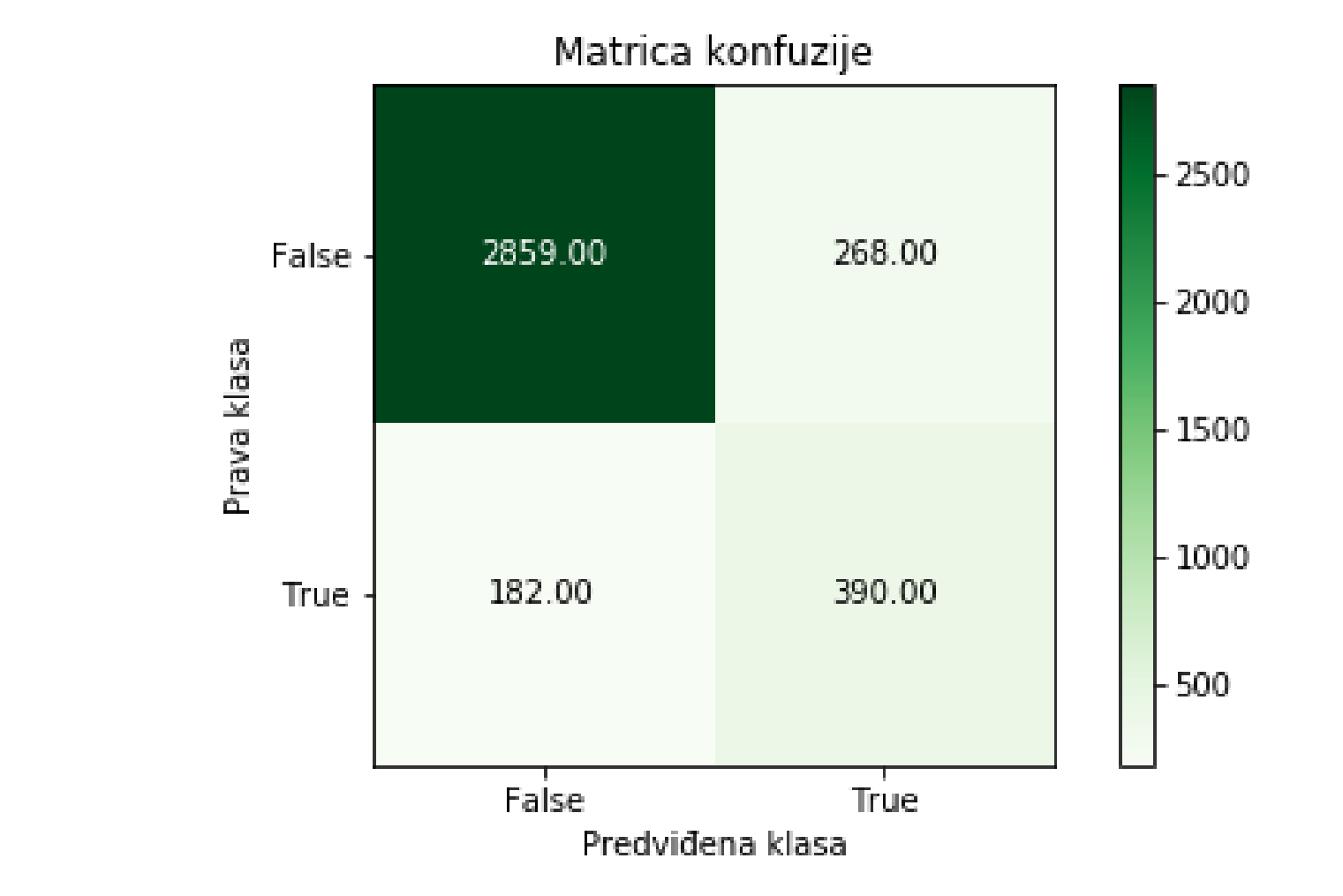
Evaluacija dobijenog modela

- Train score: 0.922960932145305 1
- Test score: 0.878345498783455

	precision	recall	f1-score	support
0	0.94	0.91	0.93	3127
1	0.59	0.68	0.63	572
accuracy			0.88	3699
macro avg	0.77	0.80	0.78	3699
weighted avg	0.89	0.88	0.88	3699

Slika 17: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 18: Matrica konfuzije za stablo odlučivanja

Metod potpornih vektora (SVM)

- Naredna tehnika koja će biti primenjena je *SVM (Support Vector Machine)*, odnosno tehnika potpornih vektora. To je tehnika za klasifikaciju zasnovana na ideji vektorskih prostora.
- Prvo će biti prikazan linearni SVM, a zatim SVM sa kernel funkcijom.

Linearni SVM

- Za regularizacioni parametar biće uzeta vrednost 1.0, što je dobijeno Grid Search algoritmom.

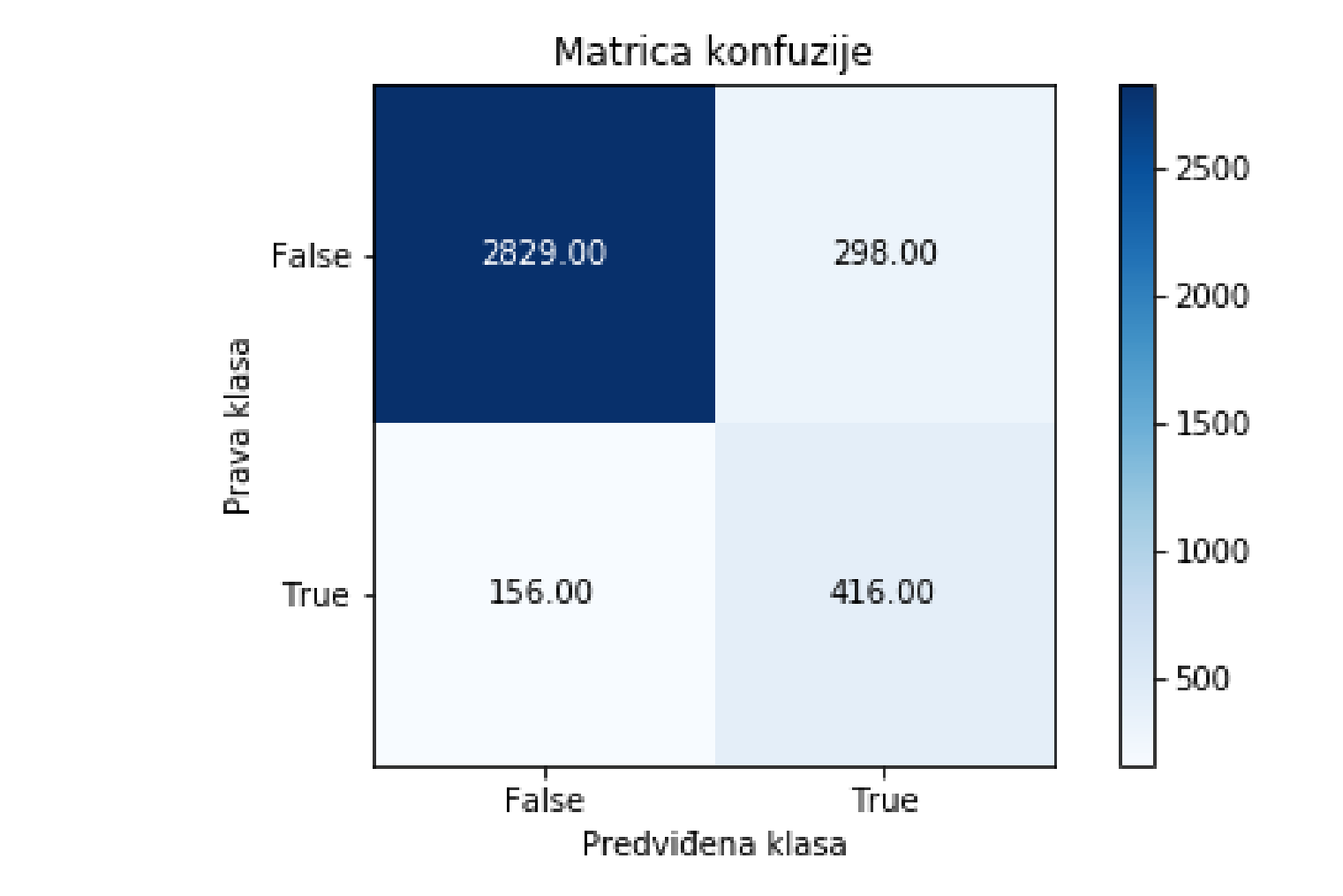
Evaluacija dobijenog modela

- Train score: 0.8613433858807402
- Test score: 0.8772641254393079

	precision	recall	f1-score	support
0	0.95	0.90	0.93	3127
1	0.58	0.73	0.65	572
accuracy			0.88	3699
macro avg	0.77	0.82	0.79	3699
weighted avg	0.89	0.88	0.88	3699

Slika 19: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 20: Matrica konfuzije za linearni SVM

SVM sa kernel funkcijom

- Za regularizacioni parametar biće uzeta vrednost 1.0, a za kernel je odabran rbf kernel.

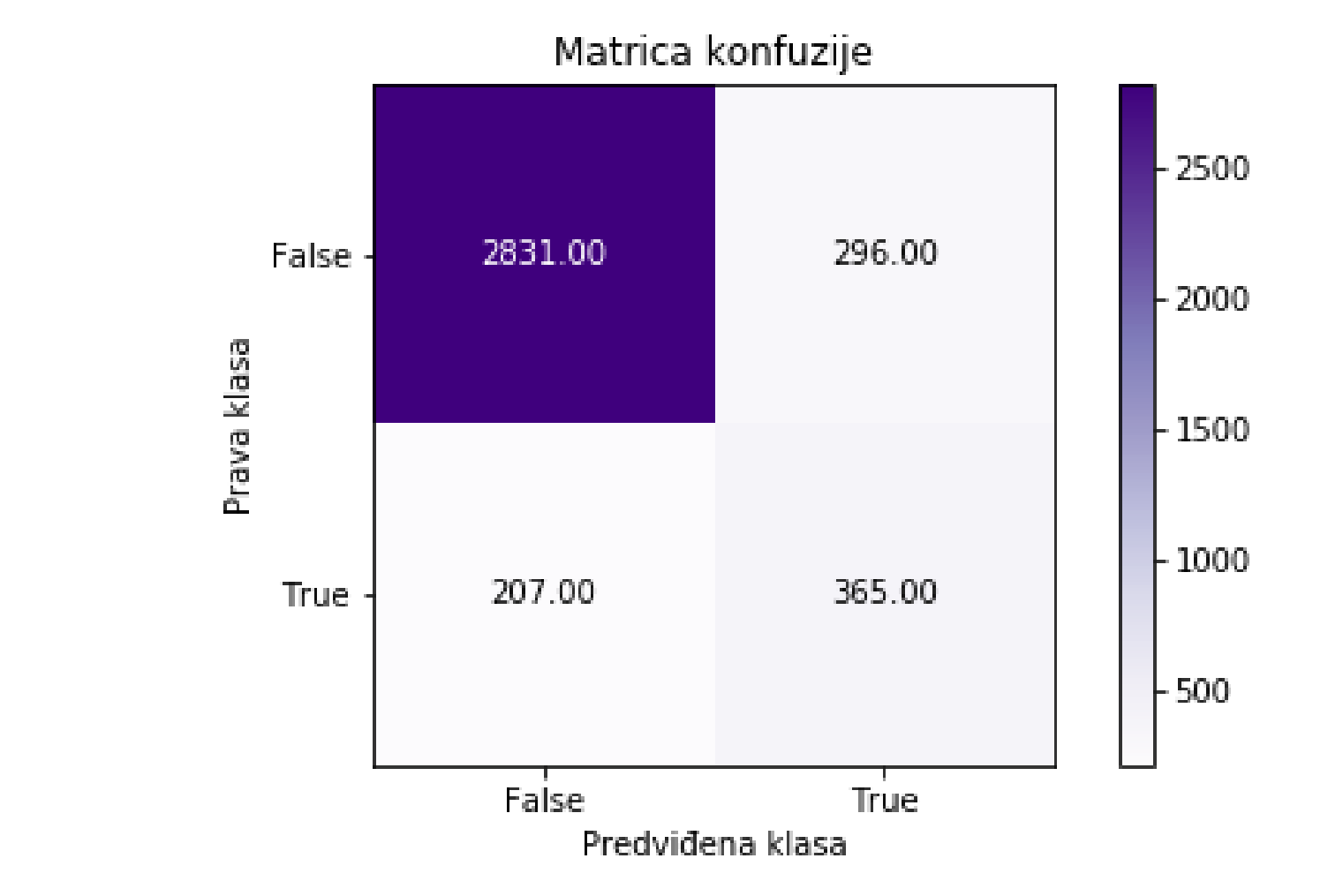
Evaluacija dobijenog modela

- Train score: 0.9372172721041809
- Test score: 0.8640173019735063

	precision	recall	f1-score	support
0	0.93	0.91	0.92	3127
1	0.55	0.64	0.59	572
accuracy			0.86	3699
macro avg	0.74	0.77	0.76	3699
weighted avg	0.86	0.87	0.86	3699

Slika 21: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 22: Matrica konfuzije za SVM sa kernel funkcijom

Random Decision Forests

- Poslednji metod koji će biti primenjen na skup je *Random Forest* algoritam. Ovaj metod spada u grupu ansambala.
- Ansambli koriste više algoritama za učenje kako bi postigli bolje rezultate u predikciji klasa. Konkretno, Random Forest metod funkcioniše tako što izgrađuje mnoštvo stabala odlučivanja pri treniranju i dodeljuje instanci onu klasu koja se najčešće pojavljivala.
- Kao kriterijum podele biće korišćena Entropija, a broj stabala u šumi biće jednak 15. Ti parametri su još jednom dobijeni iz Grid Search algoritma.

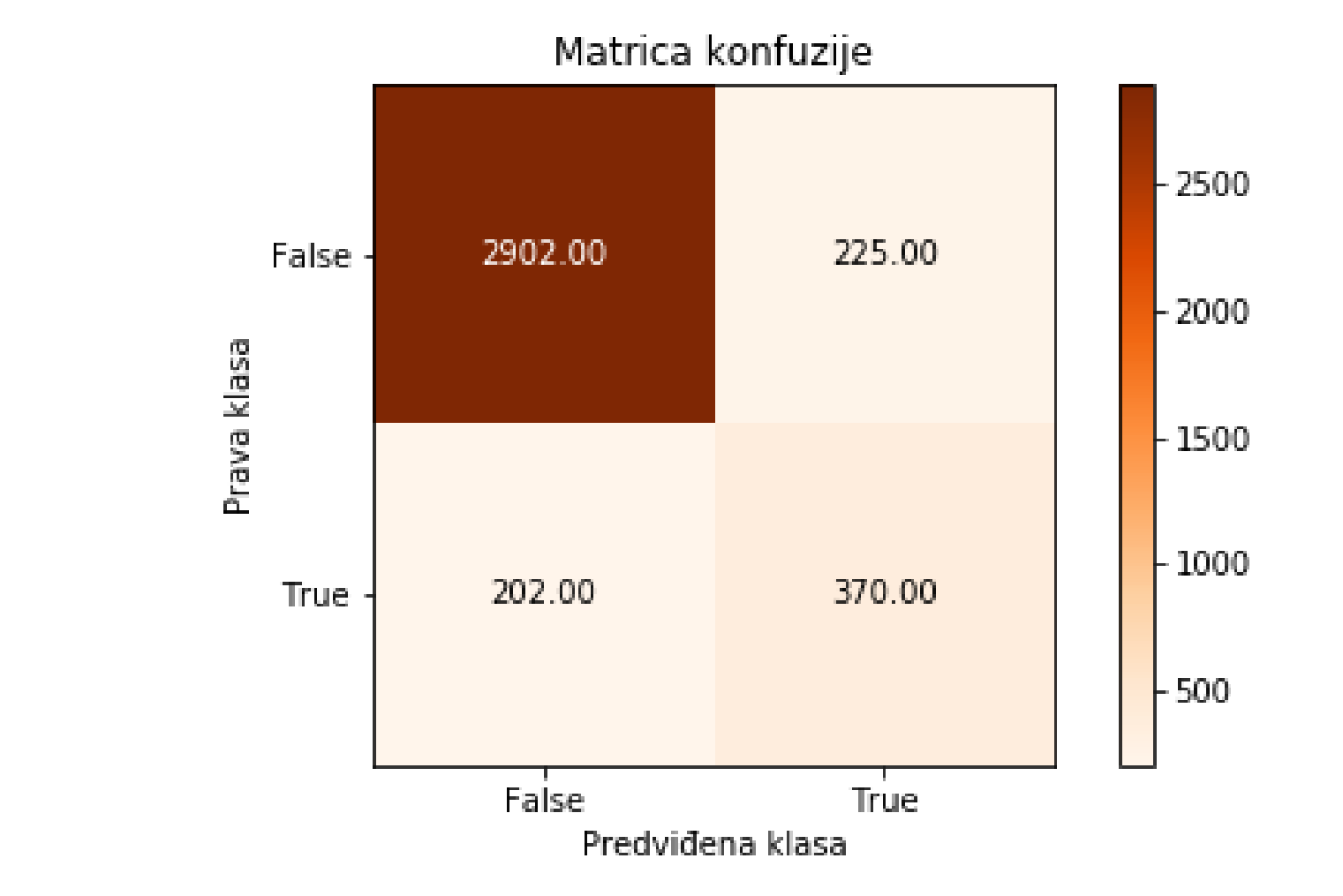
Evaluacija dobijenog modela

- Train score: 0.997943797121316
- Test score: 0.8821303054879697

	precision	recall	f1-score	support
0	0.93	0.93	0.93	3127
1	0.61	0.64	0.63	572
accuracy			0.88	3699
macro avg	0.77	0.78	0.78	3699
weighted avg	0.88	0.88	0.88	3699

Slika 23: Izveštaj klasifikacije

Evaluacija dobijenog modela



Slika 24: Matrica konfuzije za Random Forest metod

Zaključak

Rezultati svih primenjenih metoda bez balansiranja

Naziv metode	Tačnost na trening skupu	Tačnost na test skupu	Preciznosz klase '0'	Preciznost klase '1'	Odziv klase '0'	Odziv klase '1'
Logistička regresija	0.85	0.86	0.95	0.54	0.89	0.72
Stabla odlučivanja	0.90	0.88	0.92	0.66	0.95	0.57
Linearni SVM	0.88	0.88	0.89	0.73	0.98	0.36
SVM sa kernel funkcijom	0.91	0.88	0.90	0.74	0.98	0.38
Random Forest	0.99	0.88	0.91	0.68	0.96	0.49

Slika 25: Rezultat klasifikacije

Rezultati svih primenjenih metoda sa balansiranjem klasa

Naziv metode	Tačnost na trening skupu	Tačnost na test skupu	Preciznosz klase '0'	Preciznost klase '1'	Odziv klase '0'	Odziv klase '1'
Logistička regresija	0.85	0.86	0.95	0.54	0.89	0.72
Stabla odlučivanja	0.92	0.87	0.94	0.59	0.91	0.68
Linearni SVM	0.86	0.88	0.95	0.58	0.90	0.73
SVM sa kernel funkcijom	0.94	0.86	0.93	0.55	0.91	0.64
Random Forest	0.99	0.88	0.93	0.61	0.93	0.64

Slika 26: Rezultat klasifikacije

Zaključak

- Svi prikazani metodi se veoma slično ponašaju na opisanom skupu.
- Takođe, kao što se vidi iz prethodnih tabela, na svaki metod je balansiranje klasa uticalo na isti način: tako što je opala tačnost na test skupu, ali se povećao odziv klase ‘1’, što svakako znači da je model dobijen balansiranjem klasa bolji.

Zaključak

- Iz tabela se takođe vidi da najbolje mere ima metod Random Forest, što je i bilo očekivano zbog načina na koji ovaj metod radi.
- Dakle, krajnji rezultat istraživanja bio bi zaključak da iako su samo nijanse u pitanju u rezulatatima metoda, za klasifikaciju ovog skupa, ukoliko se uzimaju u obzir mere klasifikacije a ne i druge osobine poput jednostavnosti i brzine, trebalo bi koristiti algoritam Random Forest.