Online Shoppers Purchasing Intention

Istraživanje podataka 1

Zoran Vujičić

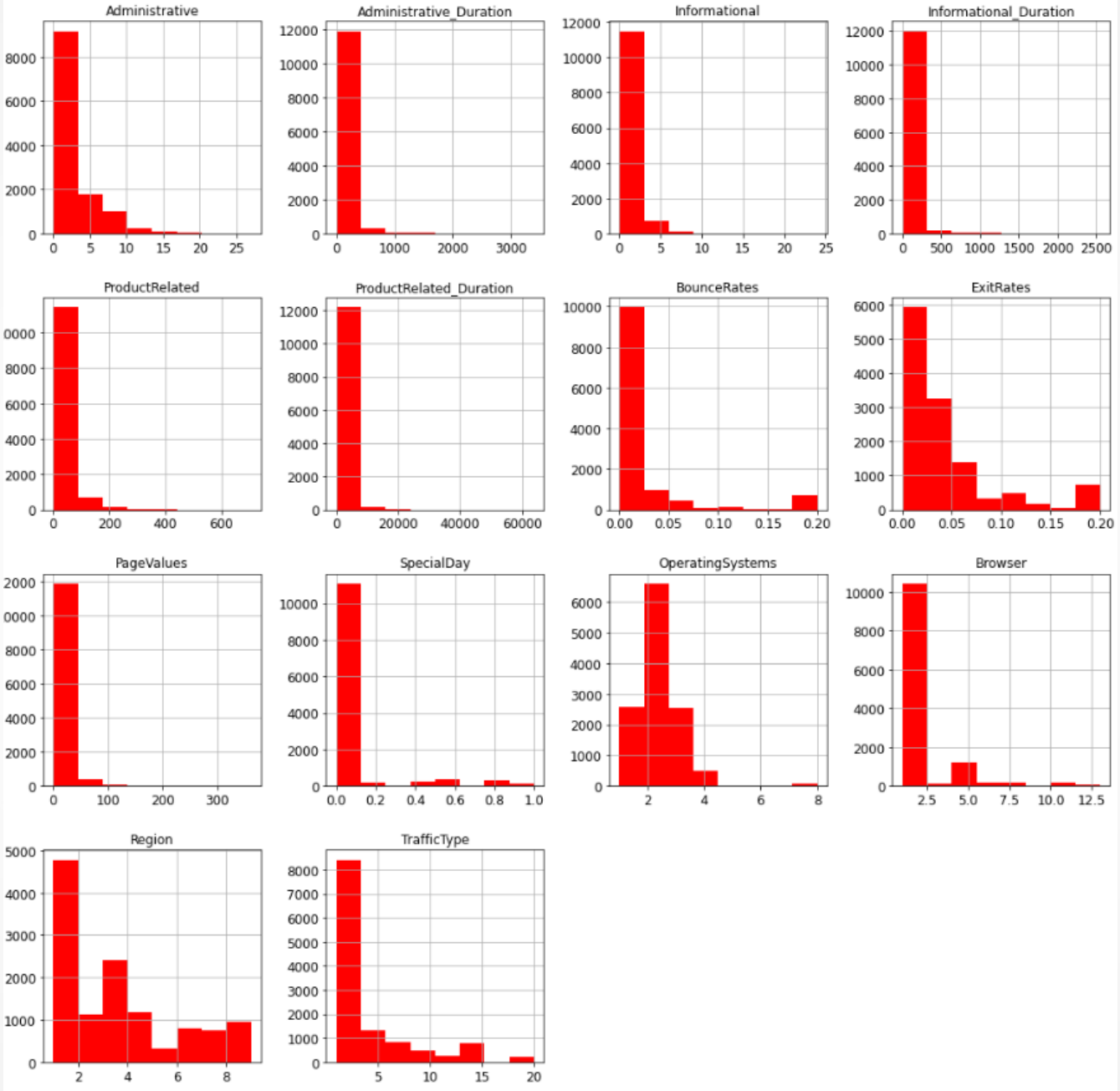
Uvod

- **Google Analytics**
- **Period 1 godina**
- **Sajt za maloprodaju**

- **12330 pristupa korisnika**
- **18 atributa**
- **10 numeričkih**
- **8 kategoričkih**

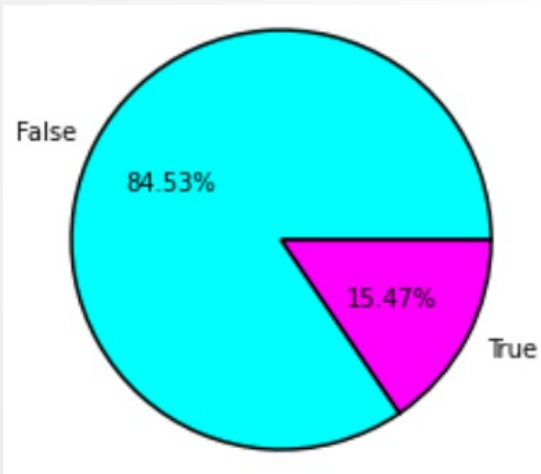
Atributi	Opis
Administrative	Broj posećenih veb strana vezanih za upravljanje profilom
Administrative_Duration	Vreme provedeno na veb stranama o upravljanju profilom u sekundama
Informational	Broj posećenih veb strana vezanih za informacije o sajtu
Informational_Duration	Vreme provedeno na stranama za informacije u sekundama
ProductRelated	Broj posećenih veb strana vezanih za proizvode
ProductRelated_Duration	Vreme provedeno na veb stranama vezanim za proizvode u sekundama
BounceRates	Procenat korisnika koji nakon ulaska na veb sajt izađu bez pokretanja drugih zahteva ka serveru
ExitRates	Koliko je puta u procentima veb strana bila poslednja u jednom pristupu korisnika internetu, u odnosu na ukupan broj pregleda
PageValues	Predstvalja prosečnu vrednost veb stranica koje je korisnik posetio pre nego što je izvršio transakciju
SpecialDay	Pokazuje koliko je vreme posete veb sajtu blizu nekog specijalnog dana u godini (npr. 8. Mart), u kojima je veća verovatnoća da se uspešno izvrši transakcija
Month	Mesec u godini u kome je korisnik pristupio veb sajtu
OperatingSystems	Operativni sistem koji je koristio korisnik
Browser	Internet pregledač koji je koristio korisnik
Region	Geografski region iz kog se prijavio korisnik
TrafficType	Izvor, odakle je korisnik pristupio veb sajtu
VisitorType	Tip korisnika koji može biti *Novi Korisnik*, *Povratnik* i *Ostali*
Weekend	Pokazauje da li je datum posete vikend ili ne
Revenue	Pokazuje da li je korisnik pri poseti veb sajtu ivršio transakciju ili nije

Analiza podataka

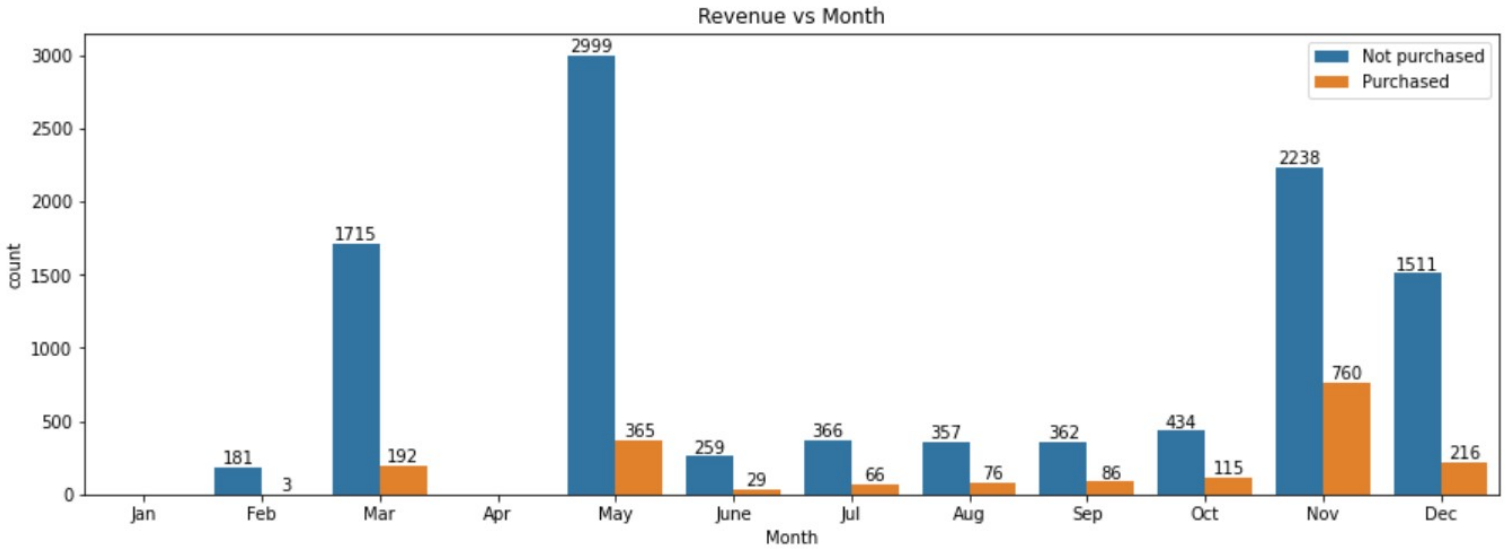


Analiza podataka

Administrative	Administrative_Dura	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValue	SpecialDay	Month	OperatingS	Browser	Region	TrafficTyp	VisitorTyp	Weekend	Revenue
0	0	0	0	1	0	0.2	0.2	0	0	Feb	1	1	1	1	Returning	FALSE	FALSE
0	0	0	0	2	64	0	0.1	0	0	Feb	2	2	1	2	Returning	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	Feb	4	1	9	3	Returning	FALSE	FALSE
0	0	0	0	2	2.666666667	0.05	0.14	0	0	Feb	3	2	2	4	Returning	FALSE	FALSE
0	0	0	0	10	627.5	0.02	0.05	0	0	Feb	3	3	1	4	Returning	TRUE	FALSE
0	0	0	0	19	154.2166667	0.015789474	0.024561	0	0	Feb	2	2	1	3	Returning	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0.4	Feb	2	4	3	3	Returning	FALSE	FALSE
1	0	0	0	0	0	0.2	0.2	0	0	Feb	1	2	1	5	Returning	TRUE	FALSE
0	0	0	0	2	37	0	0.1	0	0.8	Feb	2	2	2	3	Returning	FALSE	FALSE
0	0	0	0	3	738	0	0.022222	0	0.4	Feb	2	4	1	2	Returning	FALSE	FALSE
0	0	0	0	3	395	0	0.066667	0	0	Feb	1	1	3	3	Returning	FALSE	FALSE
0	0	0	0	16	407.75	0.01875	0.025833	0	0.4	Feb	1	1	4	3	Returning	FALSE	FALSE
0	0	0	0	7	280.5	0	0.028571	0	0	Feb	1	1	1	3	Returning	FALSE	FALSE
0	0	0	0	6	98	0	0.066667	0	0	Feb	2	5	1	3	Returning	FALSE	FALSE
0	0	0	0	2	68	0	0.1	0	0	Feb	3	2	3	3	Returning	FALSE	FALSE
2	53	0	0	23	1668.285119	0.008333333	0.016313	0	0	Feb	1	1	9	3	Returning	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	Feb	1	1	4	3	Returning	FALSE	FALSE
0	0	0	0	13	334.9666667	0	0.007692	0	0	Feb	1	1	1	4	Returning	TRUE	FALSE
0	0	0	0	2	32	0	0.1	0	0	Feb	2	2	1	3	Returning	FALSE	FALSE
0	0	0	0	20	2981.166667	0	0.01	0	0	Feb	2	4	4	4	Returning	FALSE	FALSE
0	0	0	0	8	136.1666667	0	0.008333	0	1	Feb	2	2	5	1	Returning	TRUE	FALSE
0	0	0	0	2	0	0.2	0.2	0	0	Feb	3	3	1	3	Returning	FALSE	FALSE
0	0	0	0	3	105	0	0.033333	0	0	Feb	3	2	1	5	Returning	FALSE	FALSE
0	0	0	0	2	15	0	0.1	0	0.8	Feb	2	4	1	3	Returning	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	Feb	2	2	4	1	Returning	TRUE	FALSE
0	0	0	0	5	156	0	0.04	0	0	Feb	1	1	9	3	Returning	FALSE	FALSE
4	64.6	0	0	32	1135.444444	0.002857143	0.009524	0	0	Feb	2	2	1	3	Returning	FALSE	FALSE
0	0	0	0	4	76	0.05	0.1	0	0	Feb	1	1	1	3	Returning	FALSE	FALSE



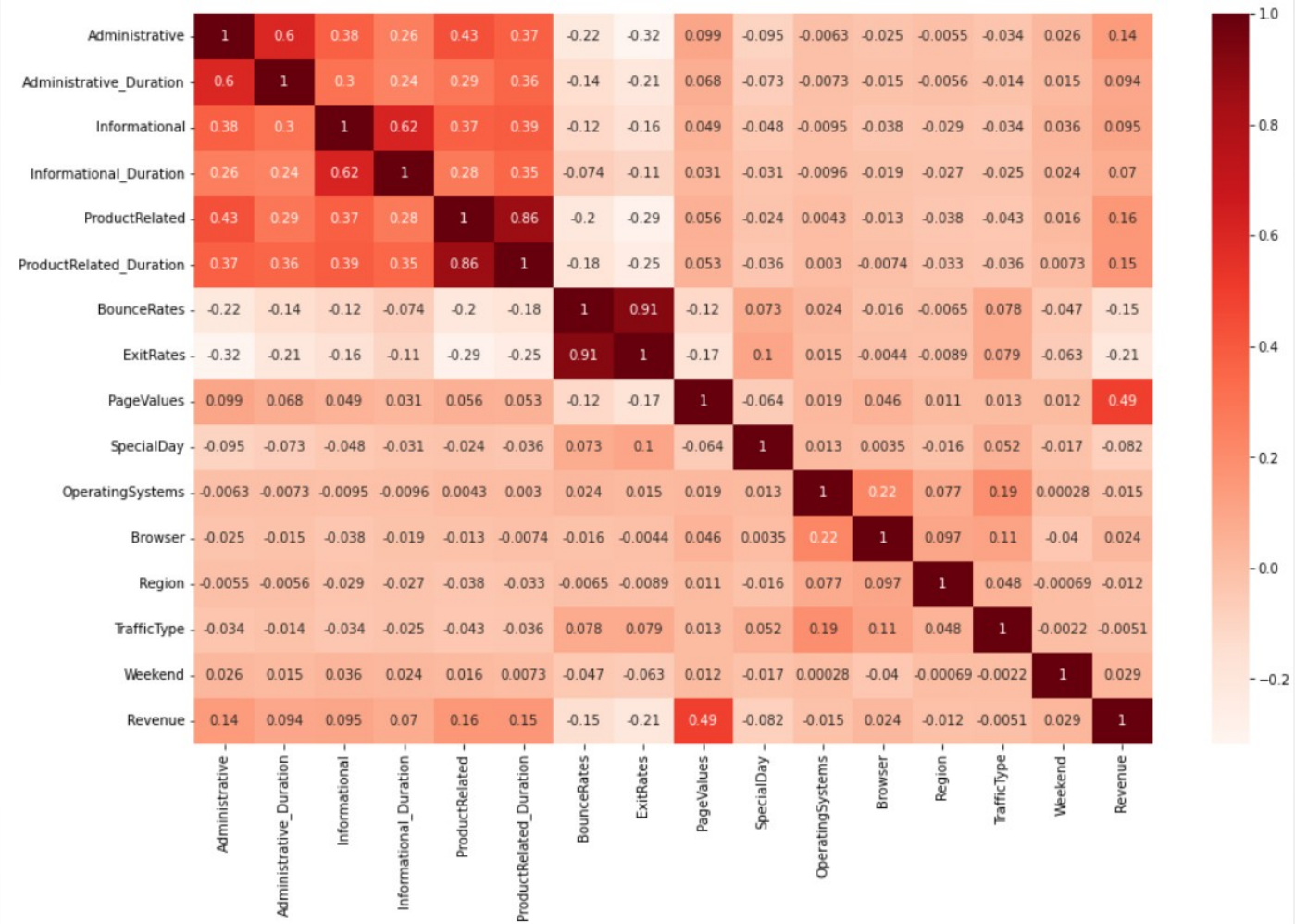
Revenue



Preprocesiranje

```
dataset.isna().sum()
```

```
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
```



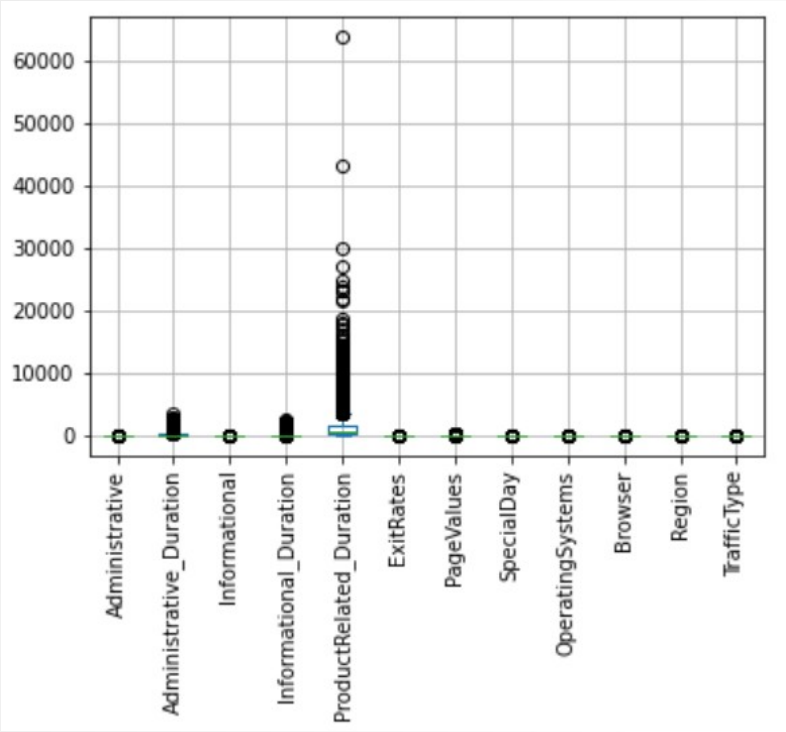
- **Nema nedostajućih vrednosti**
- **Matrica korelacije, izbačeni visokorelirani >= 0.85**

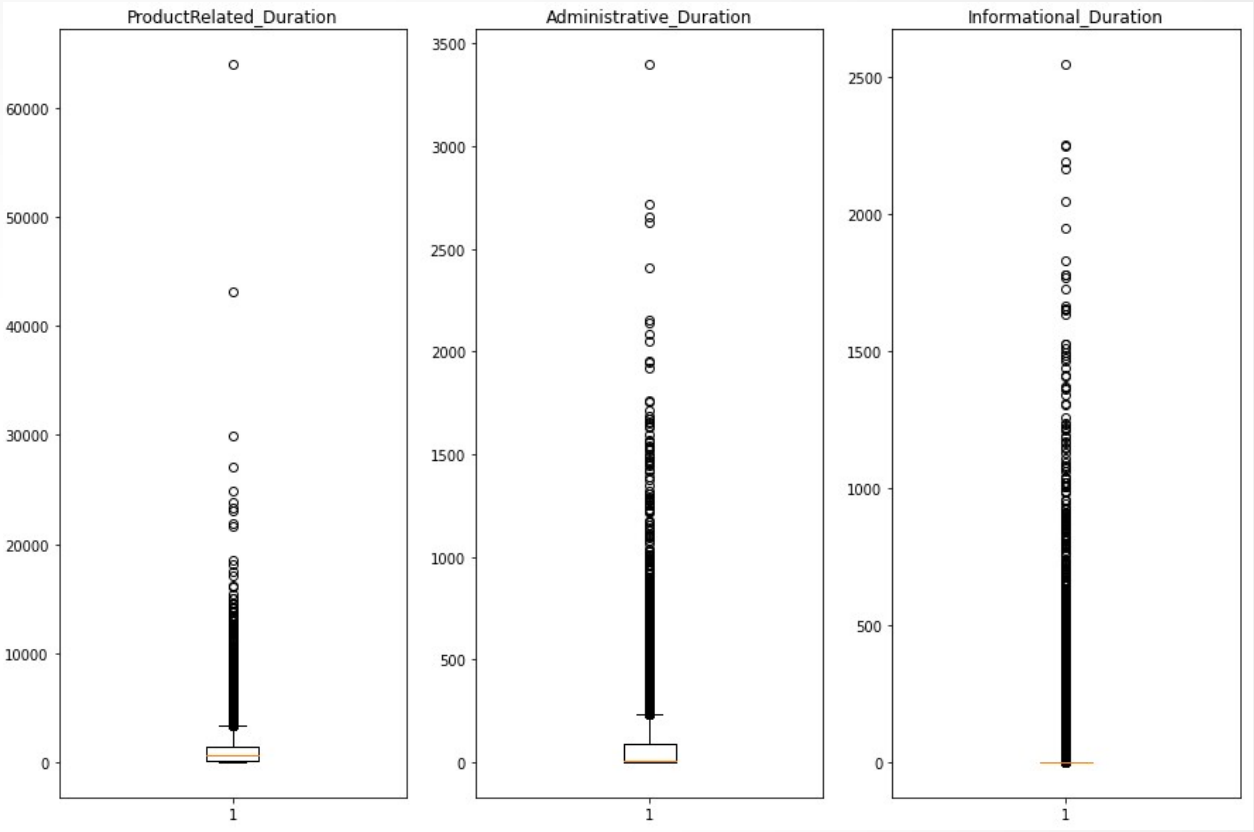
Transformisanje atributa

- **Izvršena binarizacija**
- **Binarni kategorički TRUE, FALSE zamenjeni sa 0, 1**

Administrative	OperatingSystems_1	Browser_12	TrafficType_9
Administrative_Duration	OperatingSystems_2	Browser_13	TrafficType_10
Informational	OperatingSystems_3	Region_1	TrafficType_11
Informational_Duration	OperatingSystems_4	Region_2	TrafficType_12
ProductRelated_Duration	OperatingSystems_5	Region_3	TrafficType_13
ExitRates	OperatingSystems_6	Region_4	TrafficType_14
PageValues	OperatingSystems_7	Region_5	TrafficType_15
SpecialDay	OperatingSystems_8	Region_6	TrafficType_16
Revenue	Browser_1	Region_7	TrafficType_17
Month_Aug	Browser_2	Region_8	TrafficType_18
Month_Dec	Browser_3	Region_9	TrafficType_19
Month_Feb	Browser_4	TrafficType_1	TrafficType_20
Month_Jul	Browser_5	TrafficType_2	VisitorType_New_Visitor
Month_June	Browser_6	TrafficType_3	VisitorType_Other
Month_Mar	Browser_7	TrafficType_4	VisitorType_Returning_Visitor
Month_May	Browser_8	TrafficType_5	Weekend_False
Month_Nov	Browser_9	TrafficType_6	Weekend_True
Month_Oct	Browser_10	TrafficType_7	
Month_Sep	Browser_11	TrafficType_8	

Elementi izvan granica





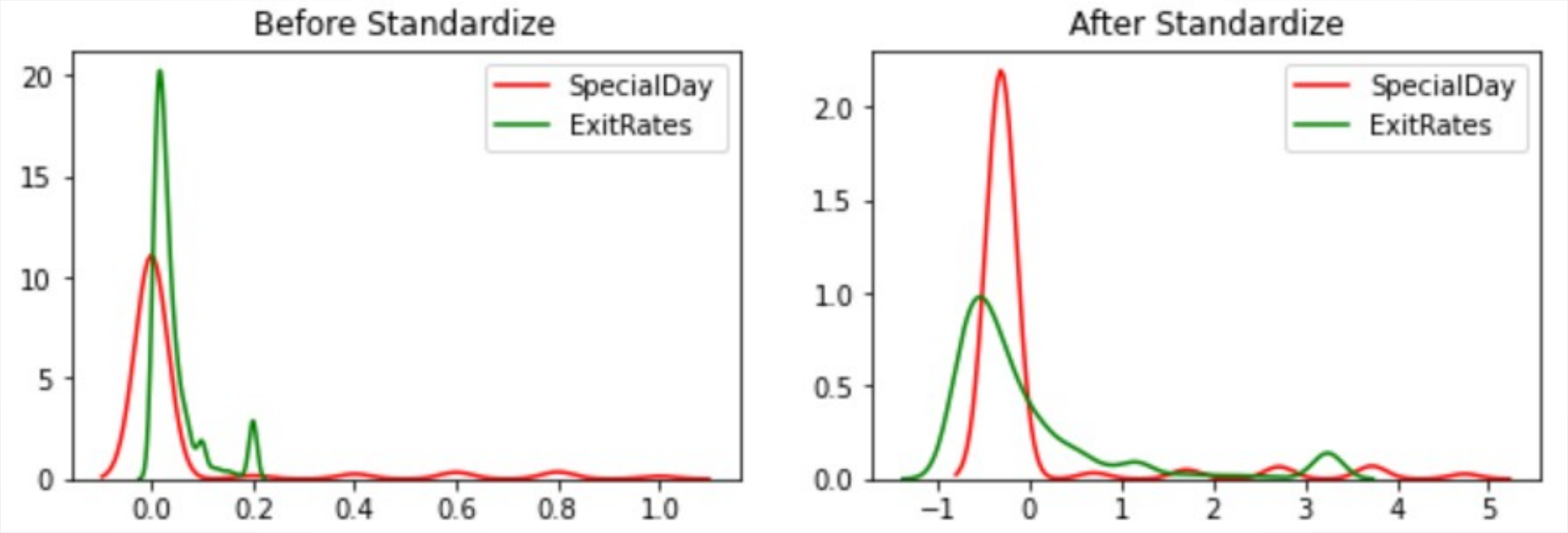
> 20000 > 2400 > 1900

IQR

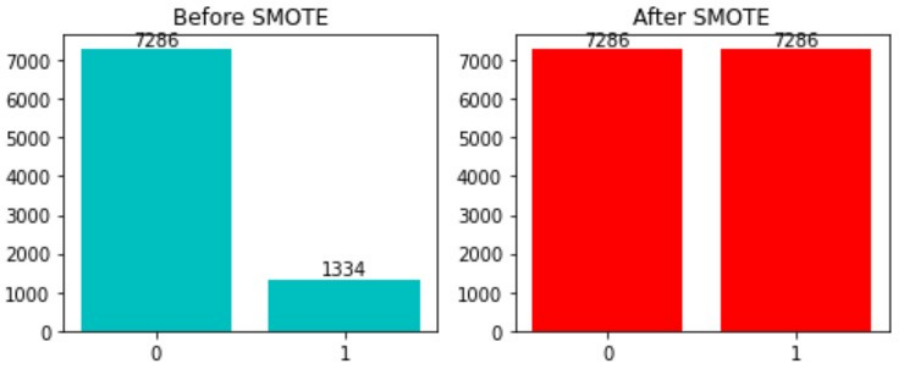
```
ProductRelated_Duration outliers: 961 in percent: 7.793
Administrative_Duration outliers: 1172 in percent: 9.50
Informational_Duration outliers: 2405 in percent: 19.50
```

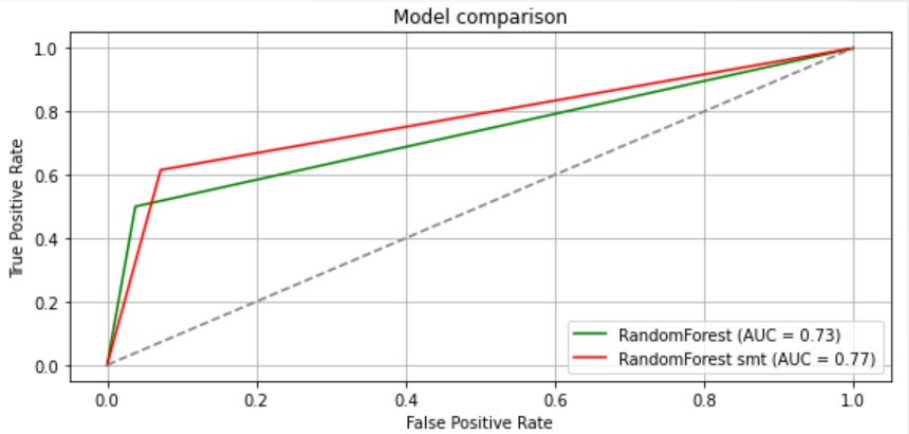
Standardizacija

Pre standardizacije, skup se **deli na skup atributa i na specijalni atribut** koji će biti korišćen kao oznaka klase. Nakon toga se oba skupa dele na **trening i test** skup koji će biti korišćeni u procesu klasifikacije. Pošto su atributi ***različito skalirani***, to znači da ih ***ne možemo međusobno upoređivati.*** Zbog toga se vrši standardizacija koja funkcioniše tako što se od atributa oduzme njegova srednja vrednost i to se podeli njegovom standardnom devijacijom.



Klasifikacija – Nasumične šume

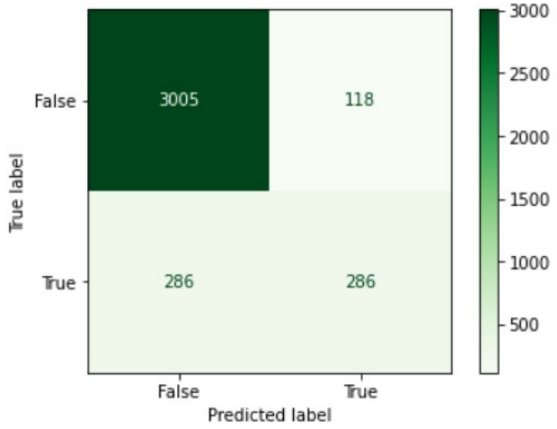




- **GridSearchCV**
- **Br. Stabla: 15**
- **Kriterijumi podele:**
  Entropija

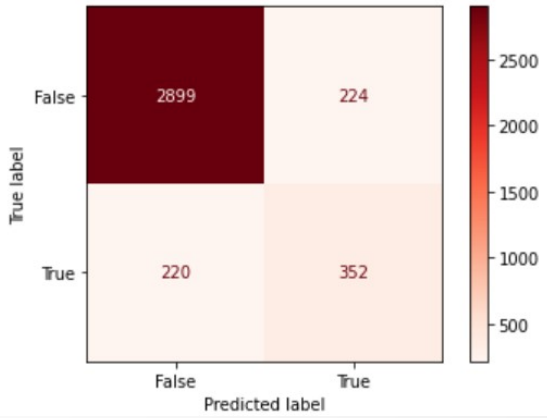
```
Train result: 0.9970997679814385
Test result: 0.8906630581867389
```

	precision	recall	f1-score	support
0	0.91	0.96	0.94	3123
1	0.71	0.50	0.59	572
accuracy			0.89	3695
macro avg	0.81	0.73	0.76	3695
weighted avg	0.88	0.89	0.88	3695

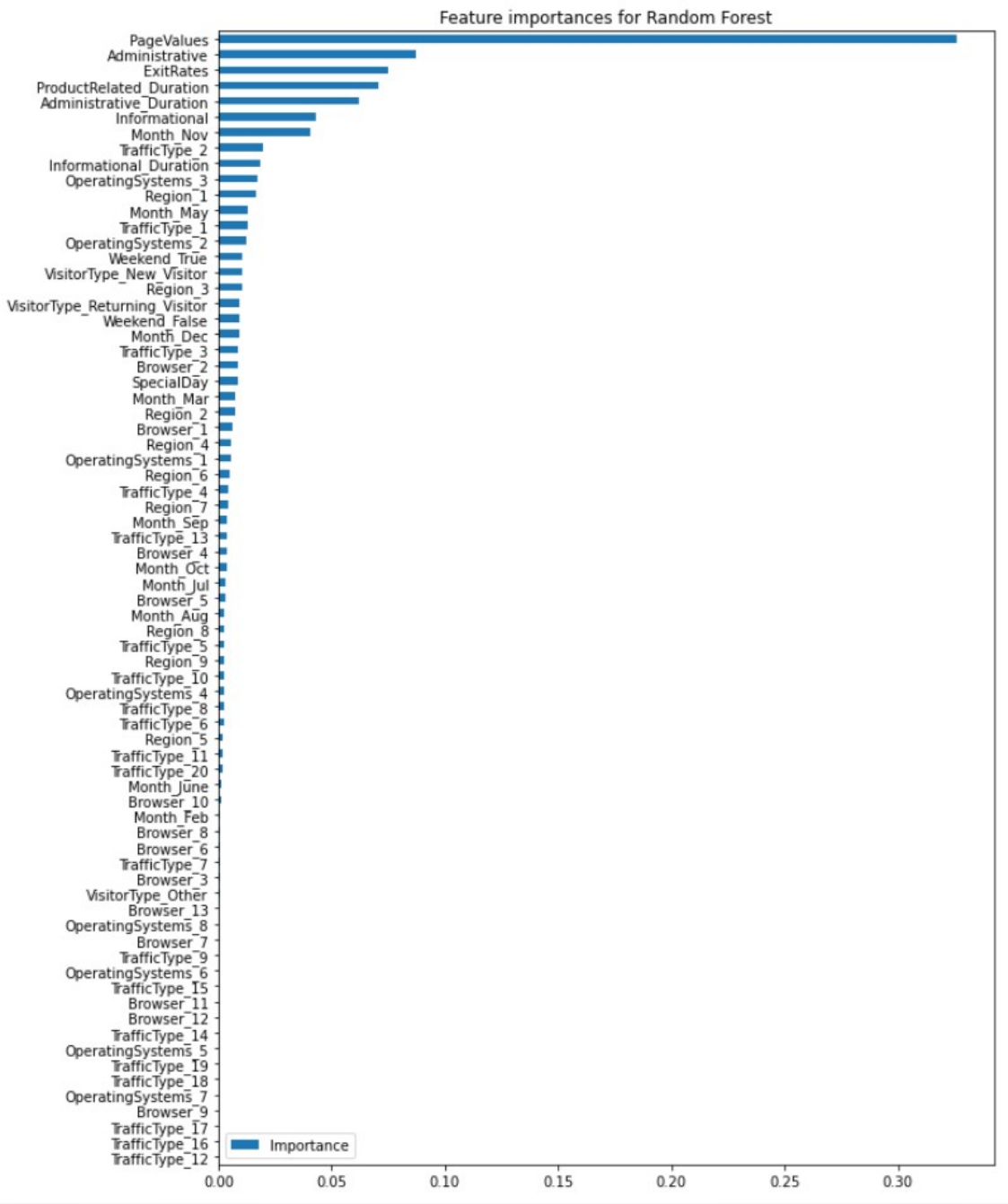



```
Train result: 0.9984216305242931
Test result: 0.8798376184032476
```

	precision	recall	f1-score	support
0	0.93	0.93	0.93	3123
1	0.61	0.62	0.61	572
accuracy			0.88	3695
macro avg	0.77	0.77	0.77	3695
weighted avg	0.88	0.88	0.88	3695

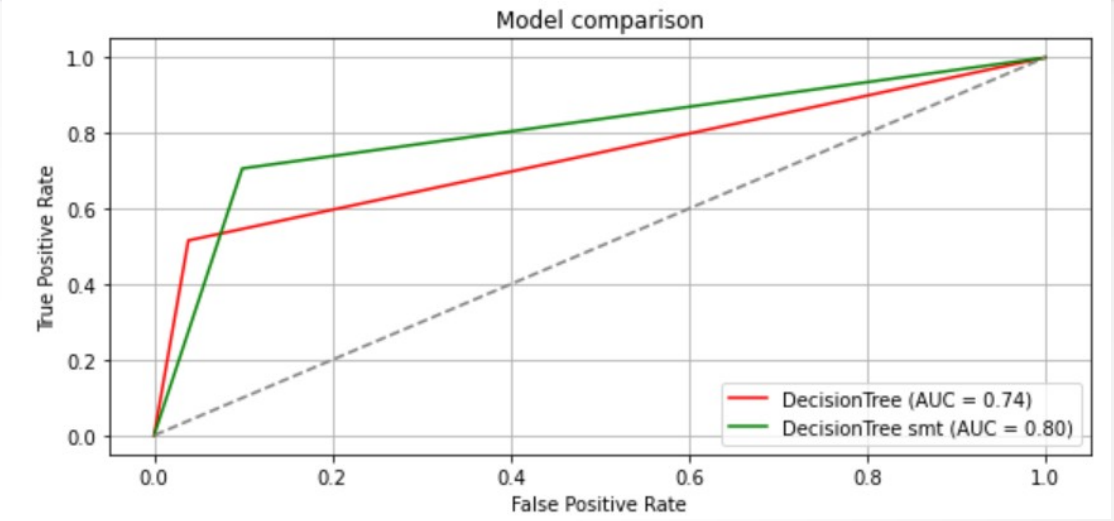


Značajnost atributa



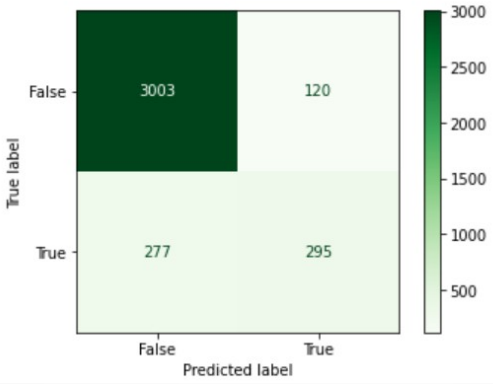
Drveta odlučivanja

- **GridSearchCV**
- **Max dubina čvorova 5 pre 10 posle balansiranja**
- **Kriterijumi podele: Entropija**



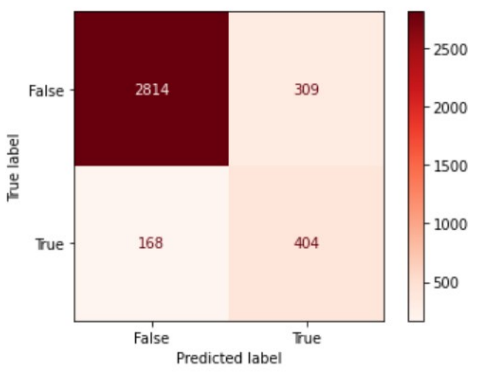

```
Train result: 0.9046403712296984
Taest result: 0.8925575101488498
```

	precision	recall	f1-score	support
0	0.92	0.96	0.94	3123
1	0.71	0.52	0.60	572
accuracy			0.89	3695
macro avg	0.81	0.74	0.77	3695
weighted avg	0.88	0.89	0.89	3695



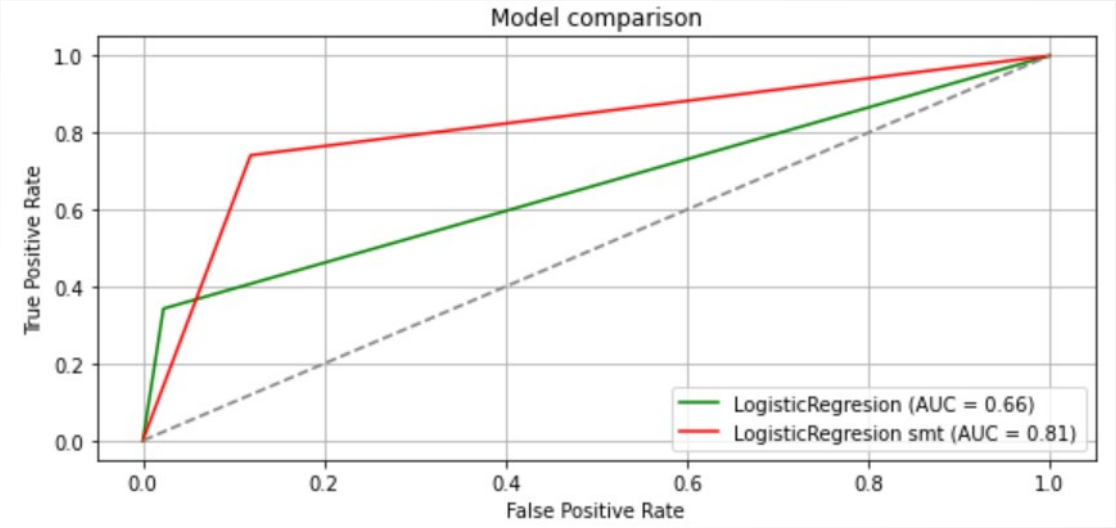

```
Train result: 0.9380318418885534
Test result: 0.8709066305818673
```

	precision	recall	f1-score	support
0	0.94	0.90	0.92	3123
1	0.57	0.71	0.63	572
accuracy			0.87	3695
macro avg	0.76	0.80	0.78	3695
weighted avg	0.89	0.87	0.88	3695



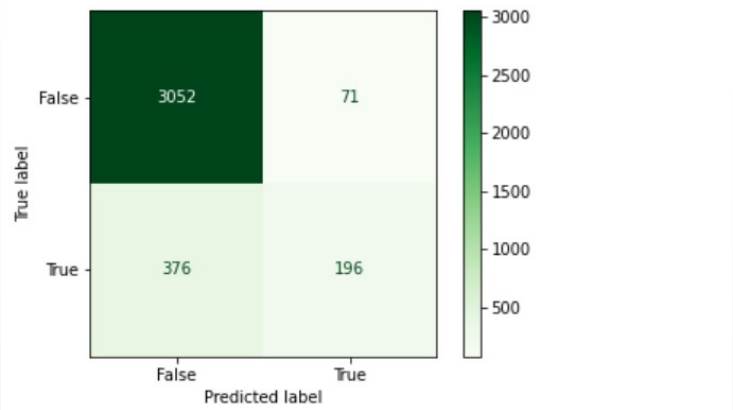
Logistička Regresija

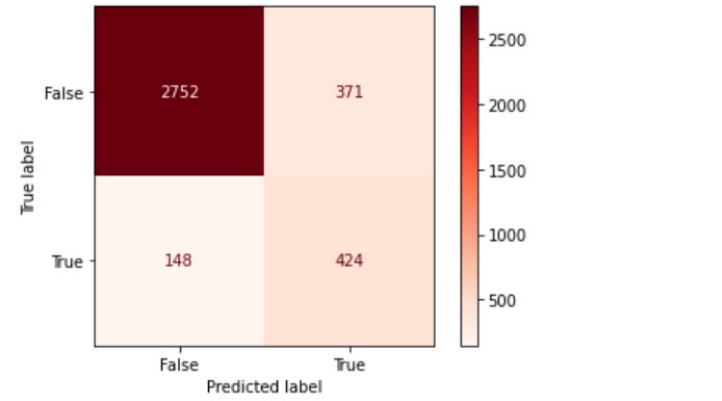
- **Upotrebljiva samo za binarnu klasifikaciju**
- **GridSearchCV**
- **C: 4.0**




```
Train result: 0.8872389791183295
Test result: 0.8790257104194857

              precision    recall  f1-score   support

           0       0.89      0.98      0.93      3123
           1       0.73      0.34      0.47       572

    accuracy                           0.88      3695
   macro avg       0.81      0.66      0.70      3695
weighted avg       0.87      0.88      0.86      3695
```

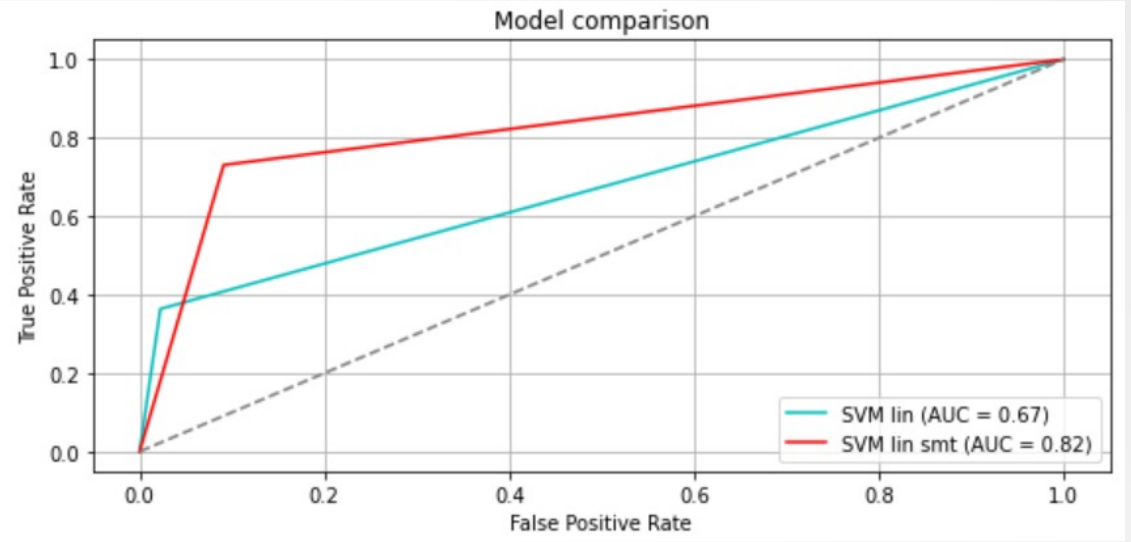
```
Train result: 0.8456629151797969
Test result: 0.8595399188092017

              precision    recall  f1-score   support

           0       0.95      0.88      0.91      3123
           1       0.53      0.74      0.62       572

    accuracy                           0.86      3695
   macro avg       0.74      0.81      0.77      3695
weighted avg       0.88      0.86      0.87      3695
```





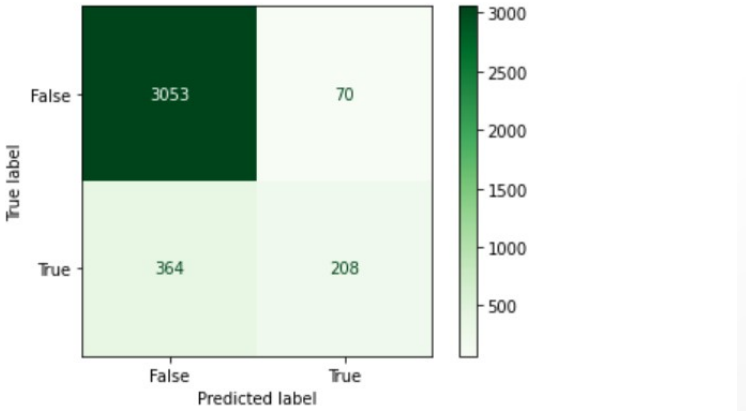
Linearni SVM

- **GridSearchCV**
- **C: 1**



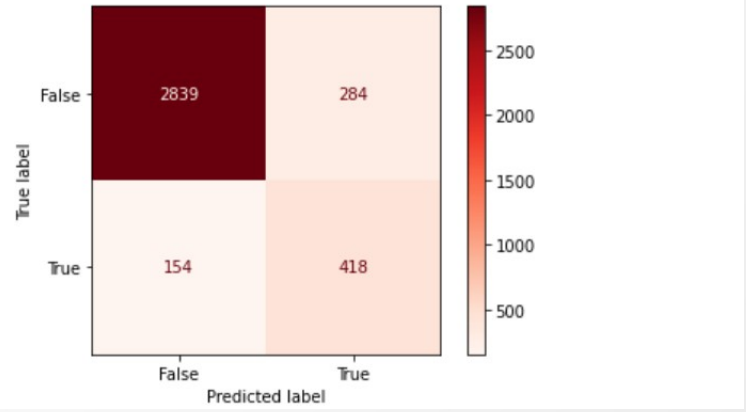
Train result: 0.8881670533642692
Test result: 0.8825439783491205

	precision	recall	f1-score	support
0	0.89	0.98	0.93	3123
1	0.75	0.36	0.49	572
accuracy			0.88	3695
macro avg	0.82	0.67	0.71	3695
weighted avg	0.87	0.88	0.86	3695



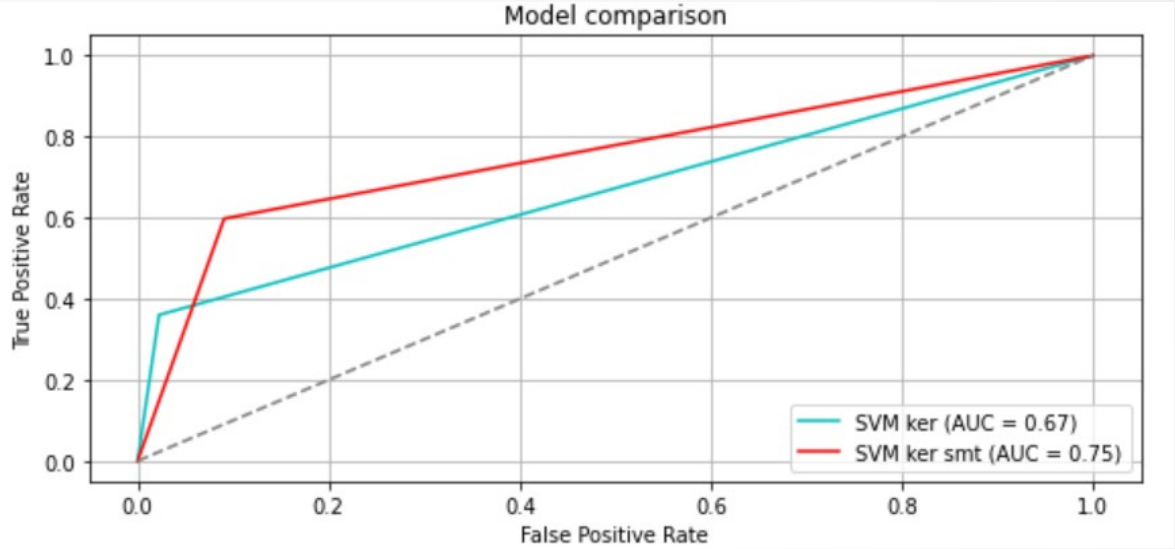
Train result: 0.8583584957452649
Test result: 0.8814614343707713

	precision	recall	f1-score	support
0	0.95	0.91	0.93	3123
1	0.60	0.73	0.66	572
accuracy			0.88	3695
macro avg	0.77	0.82	0.79	3695
weighted avg	0.89	0.88	0.89	3695



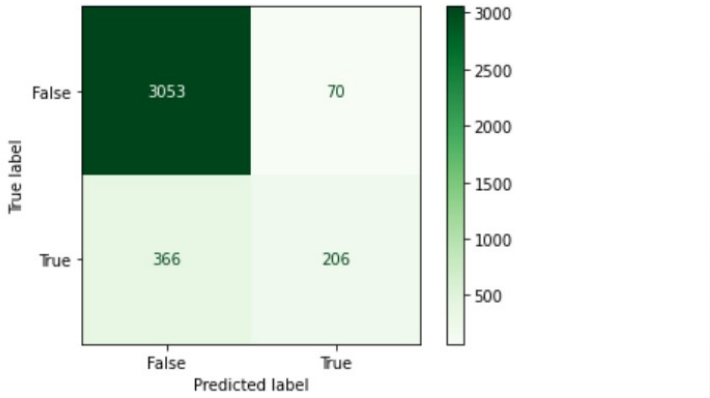
SVM sa kernelom

- **GridSearchCV**
- **C: 2.0,**
- **kernel: rbf**



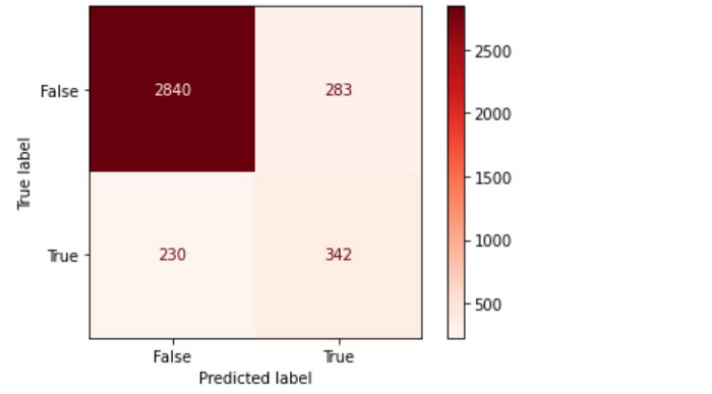

```
Train result: 0.9155452436194895
Test result: 0.8820027063599458
```

	precision	recall	f1-score	support
0	0.89	0.98	0.93	3123
1	0.75	0.36	0.49	572
accuracy			0.88	3695
macro avg	0.82	0.67	0.71	3695
weighted avg	0.87	0.88	0.86	3695



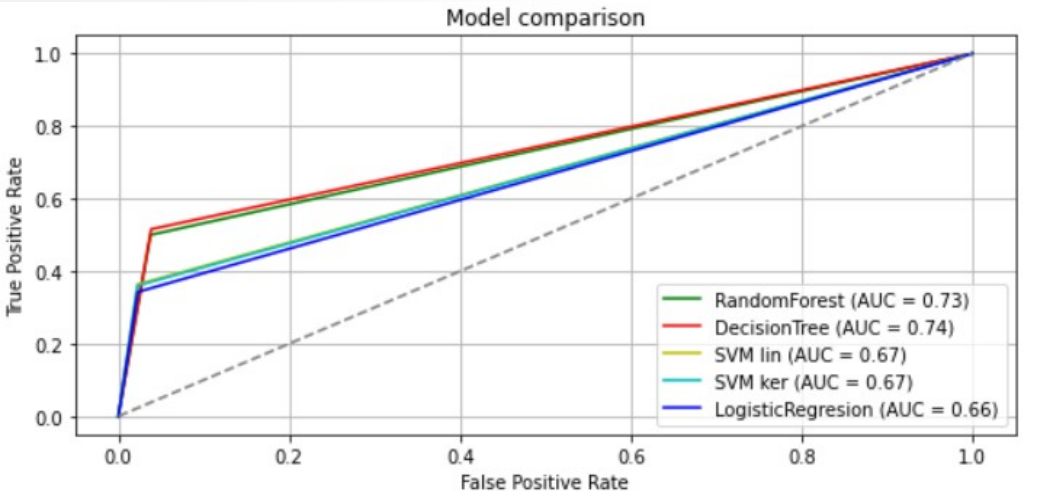

```
Train result: 0.9398847104035136
Test result: 0.8611637347767253
```

	precision	recall	f1-score	support
0	0.93	0.91	0.92	3123
1	0.55	0.60	0.57	572
accuracy			0.86	3695
macro avg	0.74	0.75	0.74	3695
weighted avg	0.87	0.86	0.86	3695

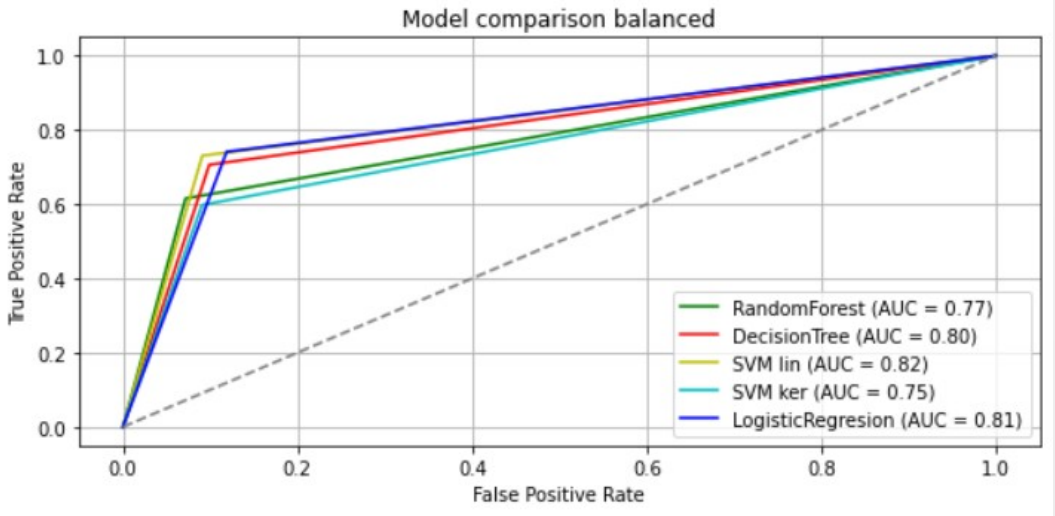


Poređenje modela kalsifikacije

Classifier	Balanced	Train Score	Test Score	Precision 0	Precision 1	Recall 0	Recall 1	F1-scr 0	F1-scr 1
Random Forest	✗	0.99	0.89	0.91	0.71	0.96	0.50	0.94	0.59
Random Forest	✔	0.99	0.88	0.93	0.61	0.93	0.62	0.93	0.61
Decision Trees	✗	0.90	0.89	0.92	0.71	0.96	0.52	0.94	0.60
Decision Trees	✔	0.94	0.87	0.94	0.57	0.90	0.71	0.92	0.63
Logistic Regression	✗	0.89	0.88	0.89	0.73	0.98	0.34	0.93	0.47
Logistic Regression	✔	0.85	0.86	0.95	0.53	0.88	0.74	0.91	0.62
SVM linear	✗	0.89	0.88	0.89	0.75	0.98	0.36	0.93	0.49
SVM linear	✔	0.86	0.88	0.95	0.60	0.91	0.73	0.93	0.66
SVM kernel	✗	0.91	0.88	0.89	0.75	0.98	0.36	0.93	0.49
SVM kernel	✔	0.94	0.86	0.93	0.55	0.91	0.60	0.92	0.57

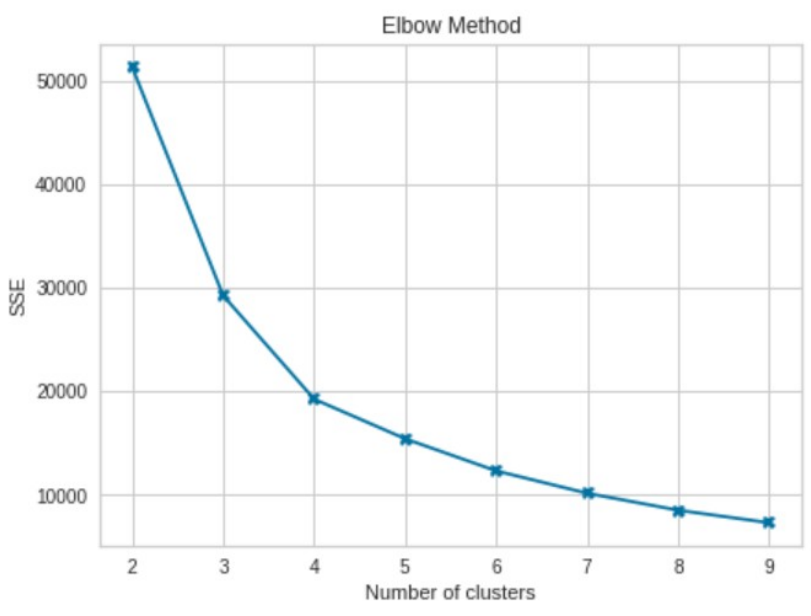


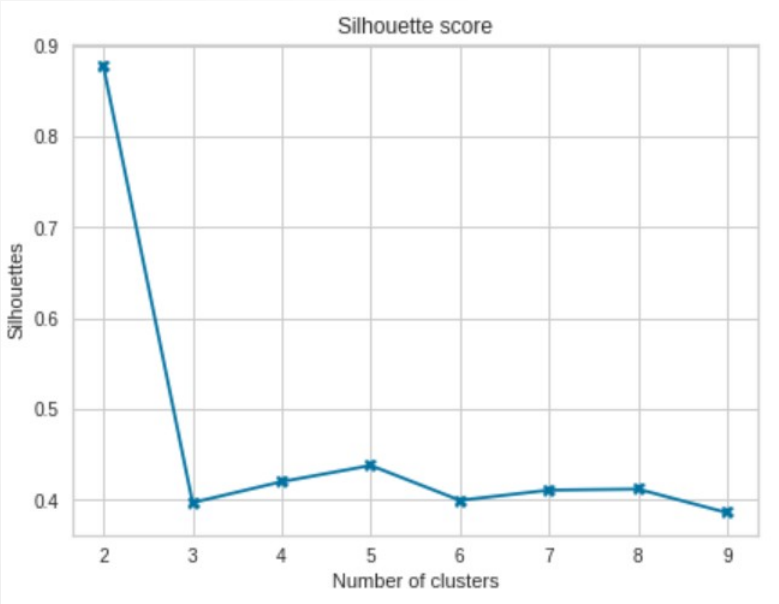
Pre SMOTE

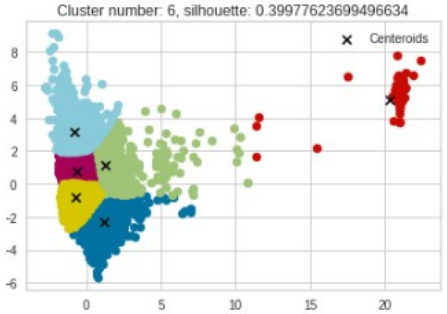


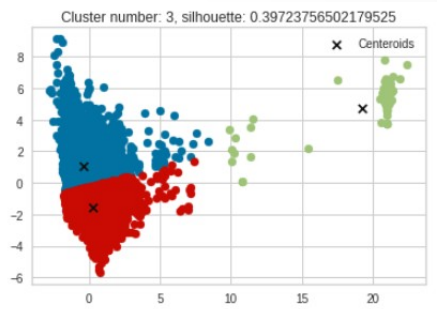
Posle SMOTE

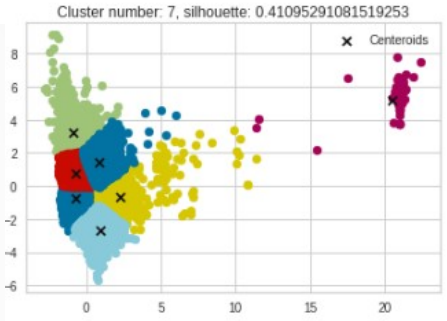
Klasterovanje – K sredina

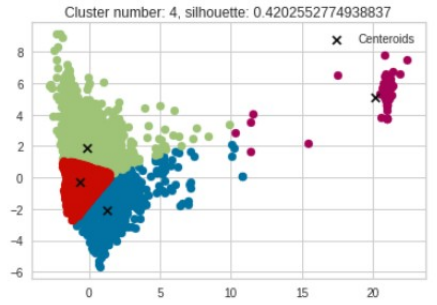


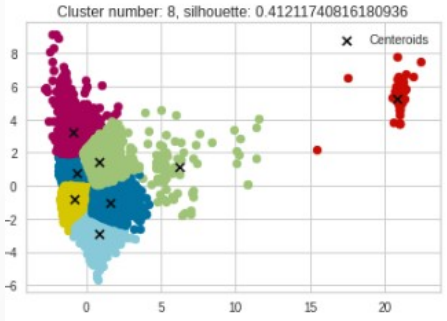


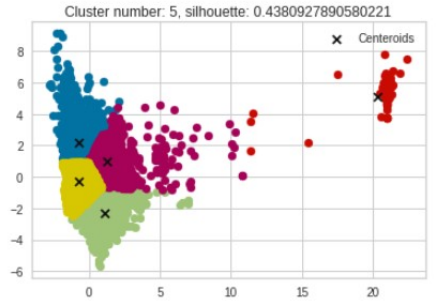


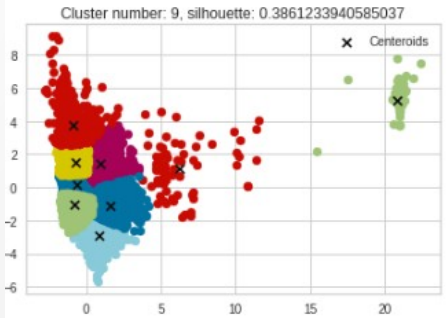






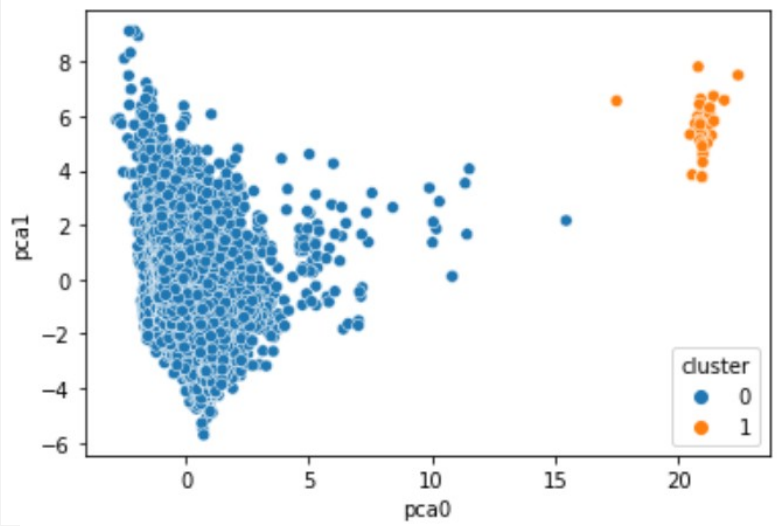


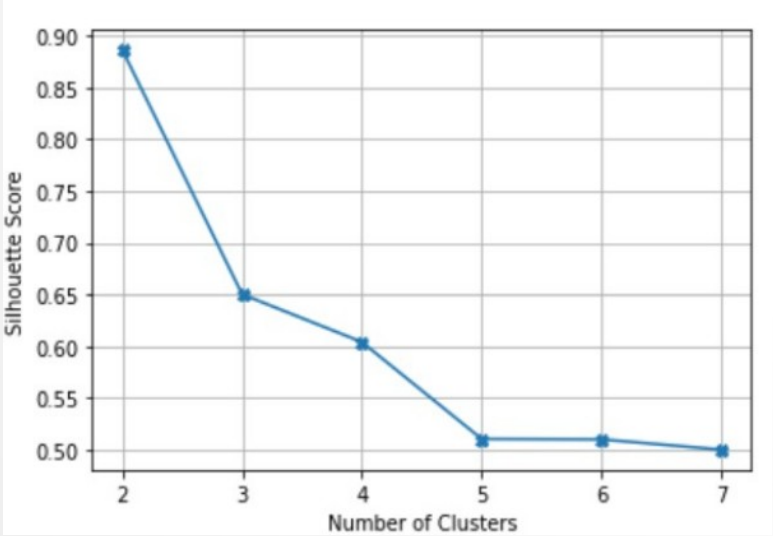


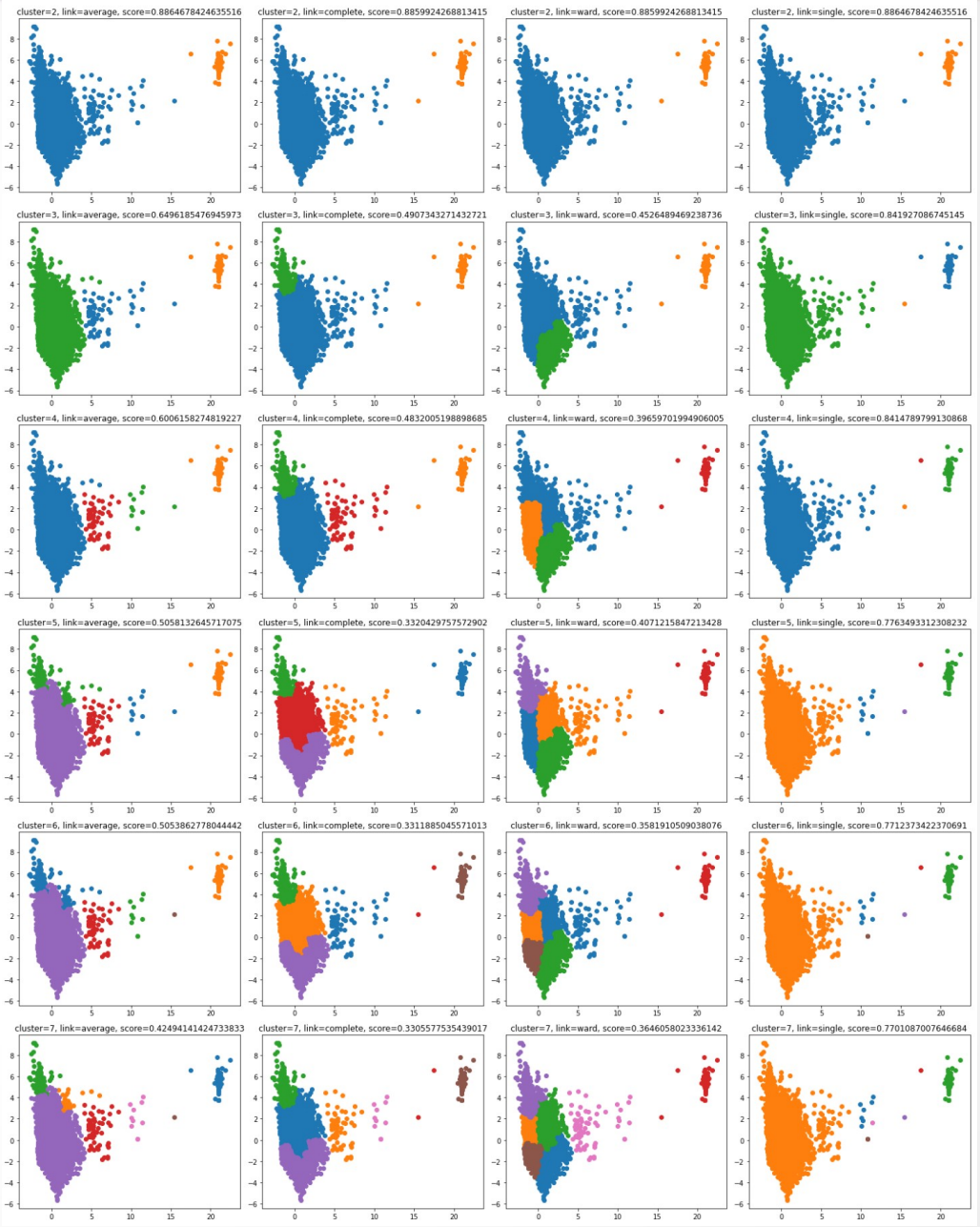


Sakupljajuće klasterovanje

Najbolji model: 2 klastera, average vezivanje

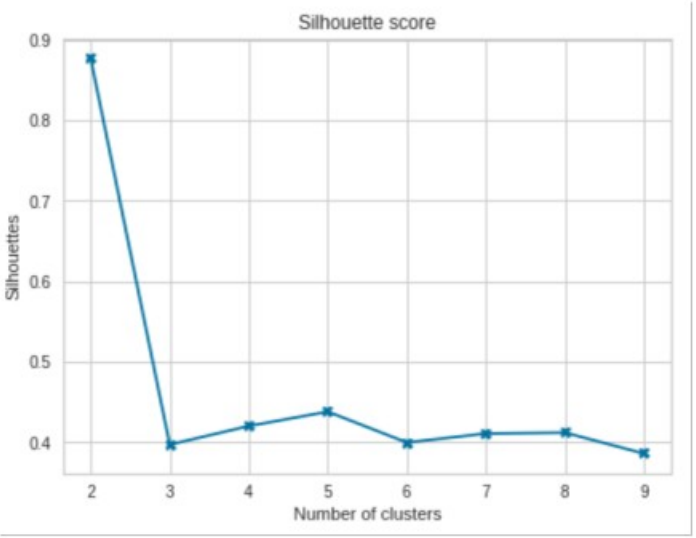




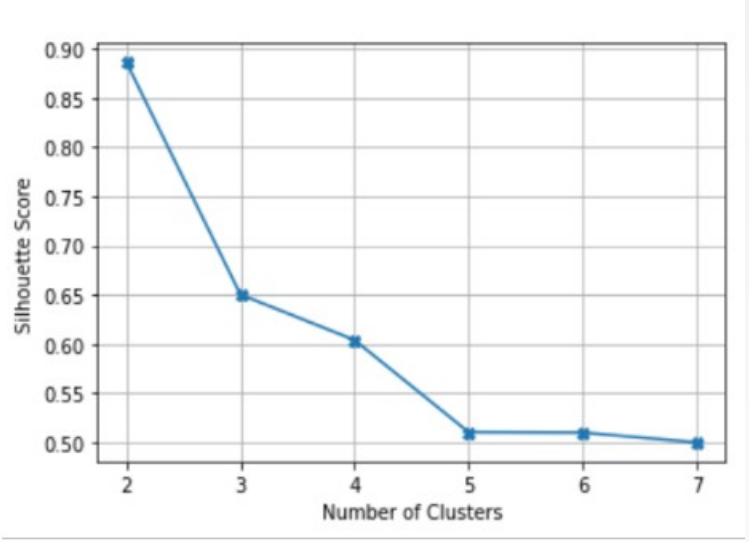


average complete ward single

Poređenje modela klasterovanja

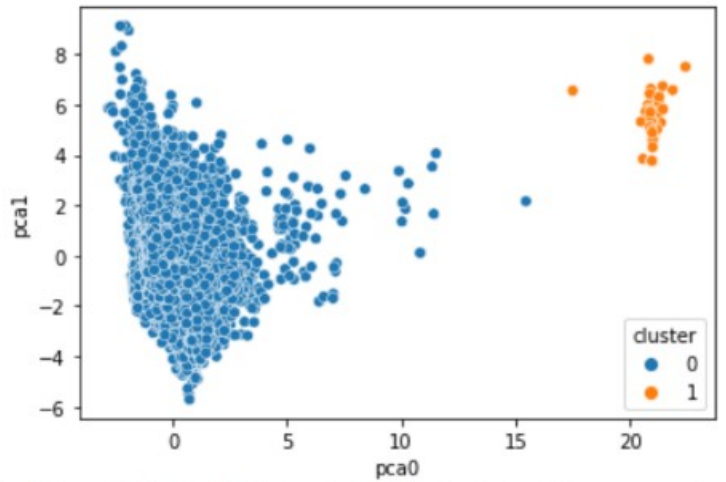


Slika 40.1: Rezultati siluete za K sredina



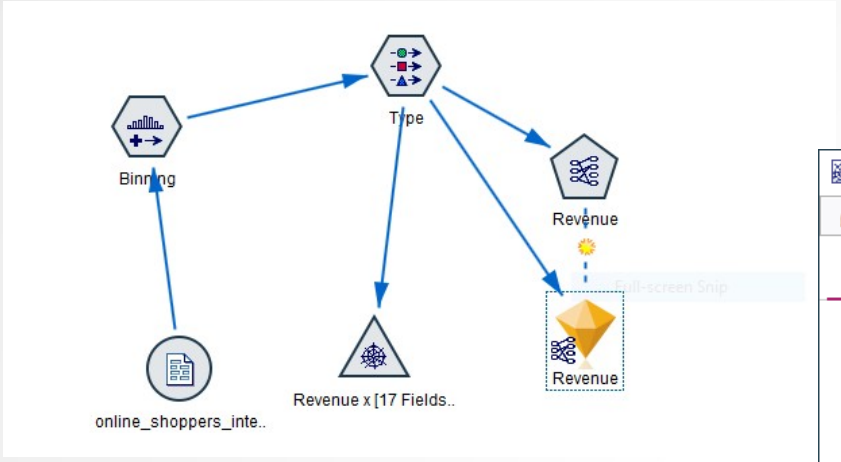
Slika 40.2: Rezultati siluete Sakupljajućeg klasterovanja

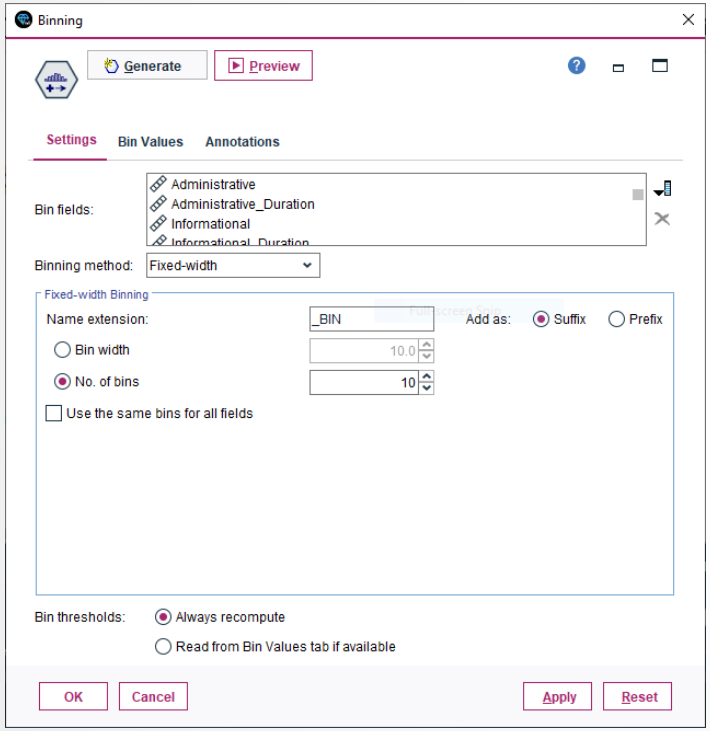
Slika 40.3: Najbolji model K sredina klasterovanja

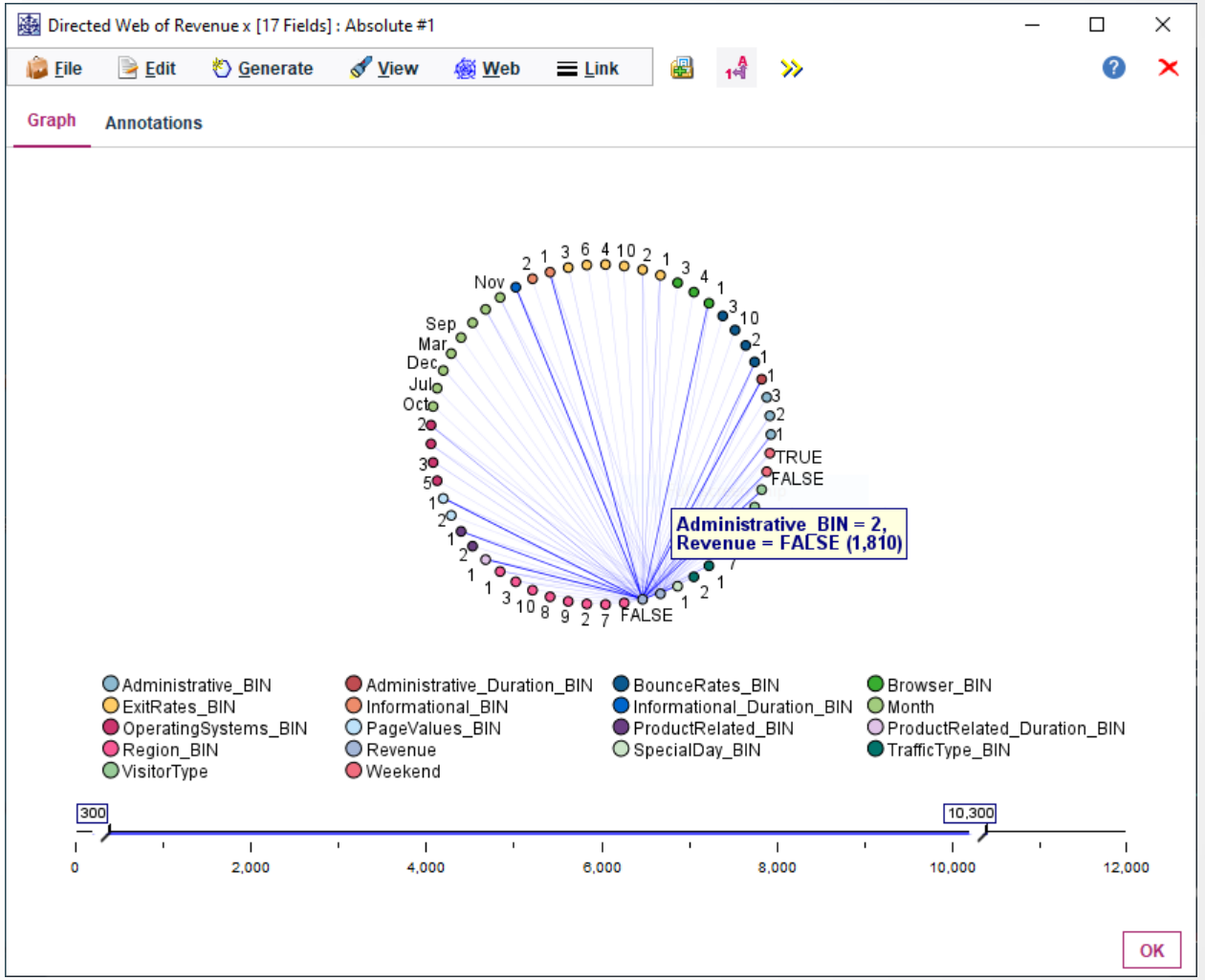


Slika 40.3: Najbolji model Sakupljajućeg klasterovanja

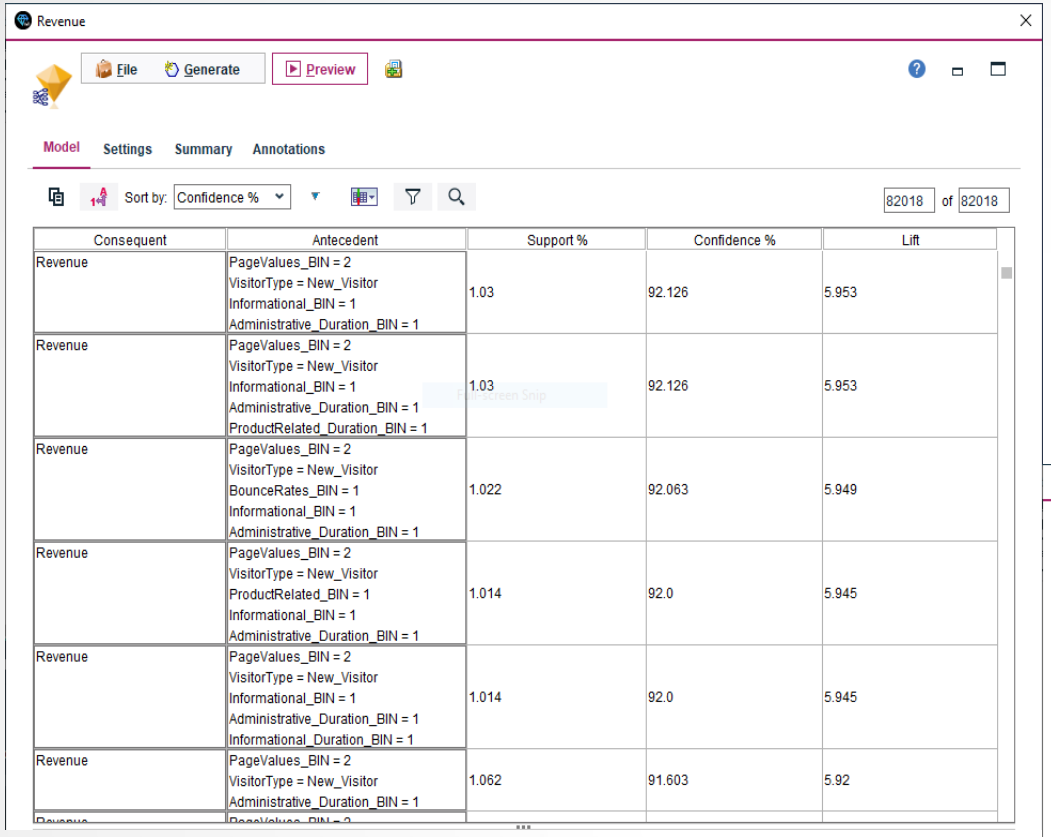
Pravila pridruživanja - Apriori







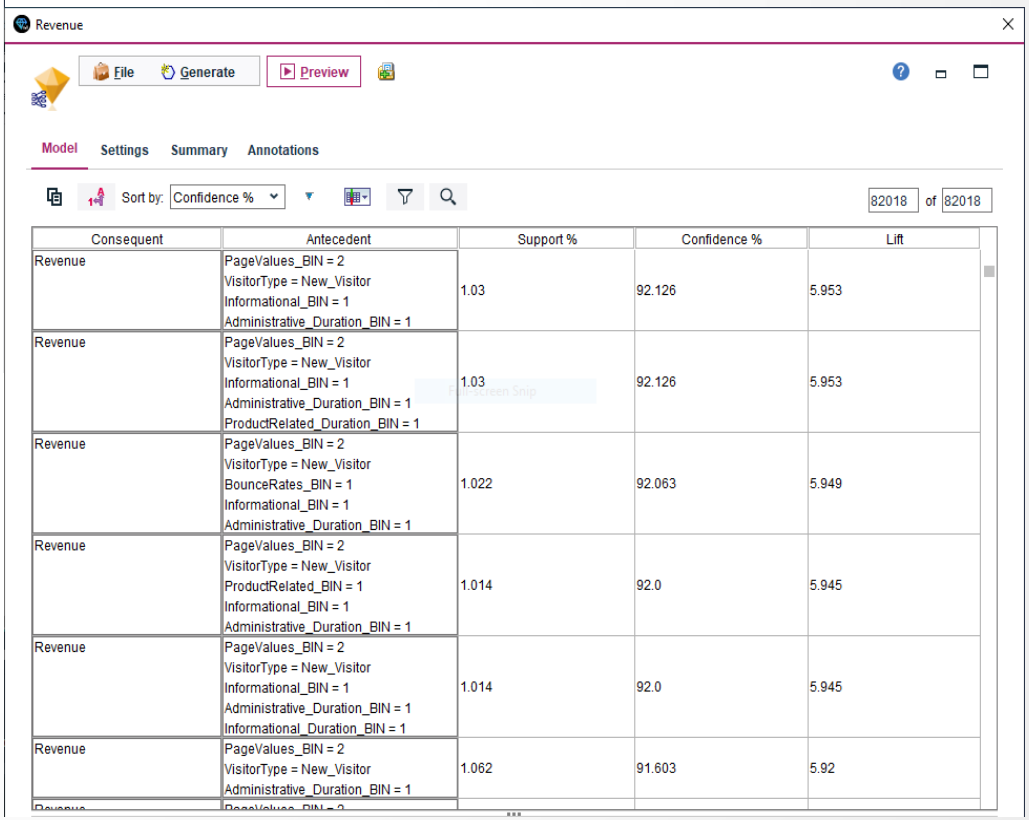
Pravila pridruživanja - Apriori



Consequent	Antecedent	Support %	Confidence %	Lift
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1	1.03	92.126	5.953
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1 ProductRelated_Duration_BIN = 1	1.03	92.126	5.953
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor BounceRates_BIN = 1 Informational_BIN = 1 Administrative_Duration_BIN = 1	1.022	92.063	5.949
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor ProductRelated_BIN = 1 Informational_BIN = 1 Administrative_Duration_BIN = 1	1.014	92.0	5.945
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1 Informational_Duration_BIN = 1	1.014	92.0	5.945
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Administrative_Duration_BIN = 1	1.062	91.603	5.92

PageValues interesantan, kao i činjenica da je VisitorType jedank novom korisniku

Lift vrednosti su jako visoke što ukazuje da su pravila zastupljenija nego očekivano.



Consequent	Antecedent	Support %	Confidence %	Lift
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1	1.03	92.126	5.953
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1 ProductRelated_Duration_BIN = 1	1.03	92.126	5.953
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor BounceRates_BIN = 1 Informational_BIN = 1 Administrative_Duration_BIN = 1	1.022	92.063	5.949
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor ProductRelated_BIN = 1 Informational_BIN = 1 Administrative_Duration_BIN = 1	1.014	92.0	5.945
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Informational_BIN = 1 Administrative_Duration_BIN = 1 Informational_Duration_BIN = 1	1.014	92.0	5.945
Revenue	PageValues_BIN = 2 VisitorType = New_Visitor Administrative_Duration_BIN = 1	1.062	91.603	5.92

Hvala na pažnji!