Binarna klasifikacija: Predviđanje zadovoljstva putnika

Prirodoslovno–matematički fakultet

Nikola Sunara, Fran Vojković, Alen Živković

Lipanj 2020

Uvod

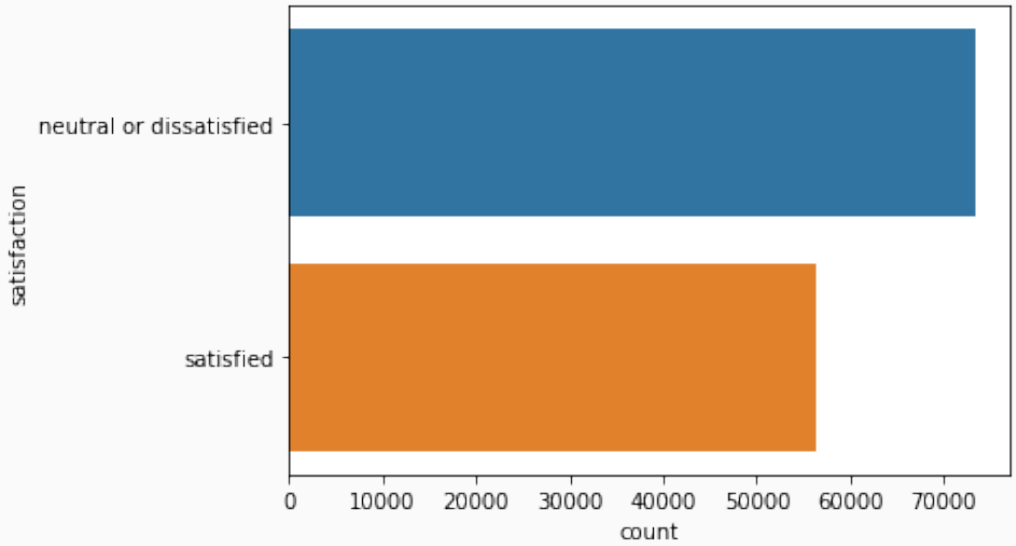
Uvod i

- Bavimo se problemom predviđanja zadovoljstva putnika avionske kompanije.
- Klasifikacijski problem gdje zadovoljstvo može biti: neutralno ili nezadovoljno te zadovoljno
- Proučavamo dataset sa Kaggle-a (Airline Passenger Satisfaction).
- Dataset sadrži 130 000 redova podataka raspoređenih na 25 značajki.
- Proveli smo čišćenje podataka od nedostajućih vrijednosti i previše kolinearnih značajki te odbacili smo manje relevantne značajke.

Analiza dataseta

Vizualizacija podataka i

- Putnici su nakon leta ocijenili svoje zadovoljstvo promatranom avio–kompanijom (slika 1).
- Kategorički stupci (*Type of travel*, *Class*, *Gender*, *Customer Type*) u datasetu imaju jedinstvene vrijednosti(*satisfied* ili *dissatisfied*) koje smo *encodali* u nule ili jedinice.

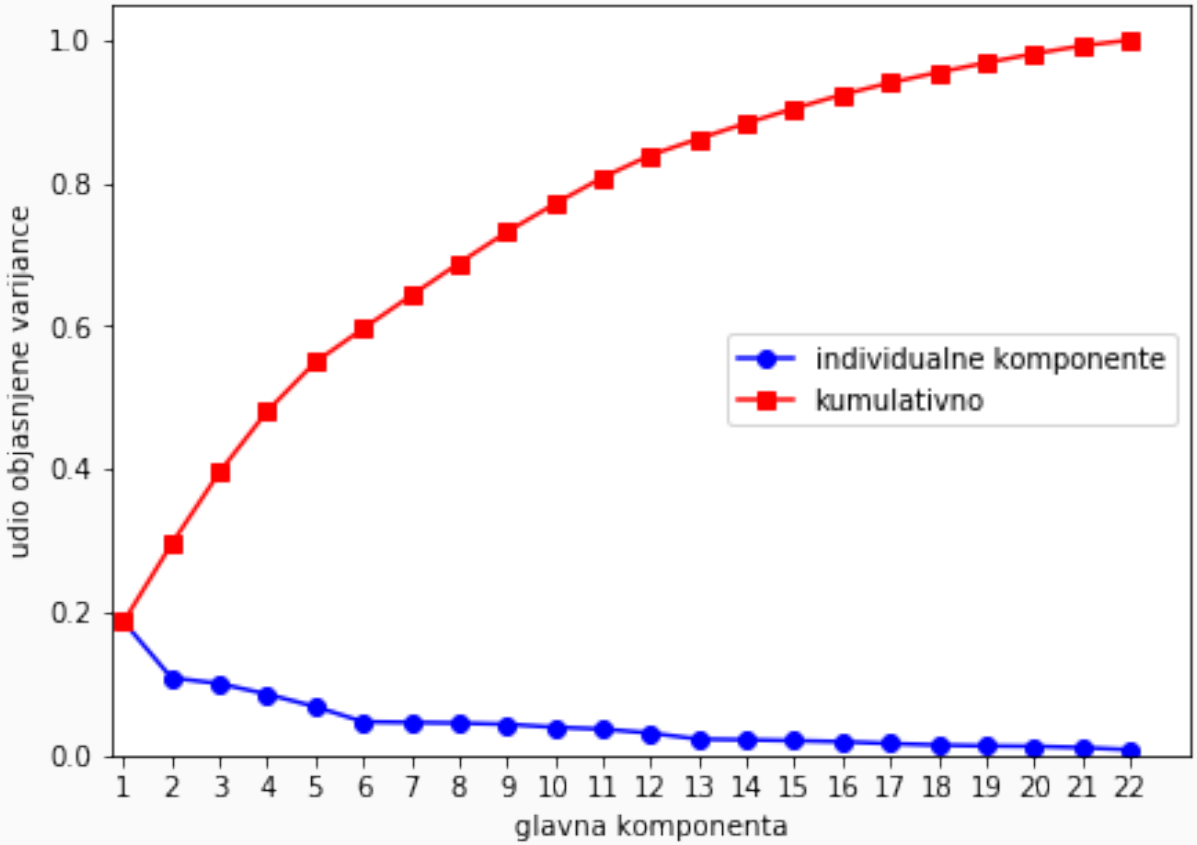


Slika 1:

Obrada podataka i

- Nedostajajuće vrijednosti u datasetu $\approx 0.3\%$, nalaze se u samo jednom stupcu.
- *FAMD* metodom izračunali smo objašnjenu varijancu te zaključujemo da prva 4 featurea imaju veliki značaj u varijanci.
- Standardizirane značajke tako da sve značajke imaju jediničnu varijancu (*StandardScaler*.
- Primjenom *PCA* uočili smo da nećemo moći značajno reducirati broj značajki.

Obrada podataka ii



Slika 2: *PCA*

Odabir značajki

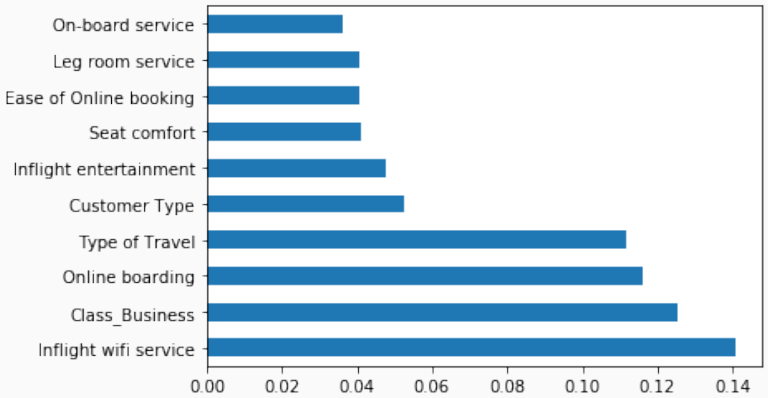
Odabir značajki i

- Iz *heatmap*-a svih značajki vidimo da je zadovoljstvo putnika visoko korelirano s vrstom putovanja, online-boardingom te radi li se o business klasi.
- Rekurzivnom eliminacijom značajki sa unakrsnom validacijom dobili smo da je optimalan broj značajki 18.

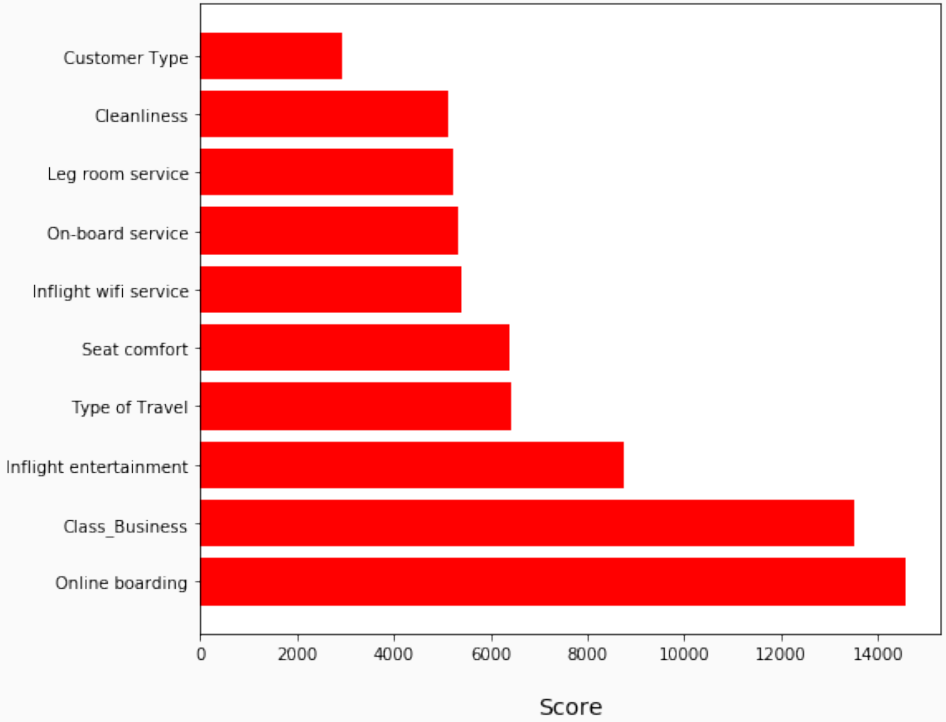
Slika 3:

Odabir značajki ii

- *Pearsonovim Chi-kvadrat* testom prikazanim na slici odredili smo 10 najznačajnijih značajki.
- Usporedili smo rezultate sa *Extra Trees Classifier* metodom.



Slika 4: Extra Trees Classifier



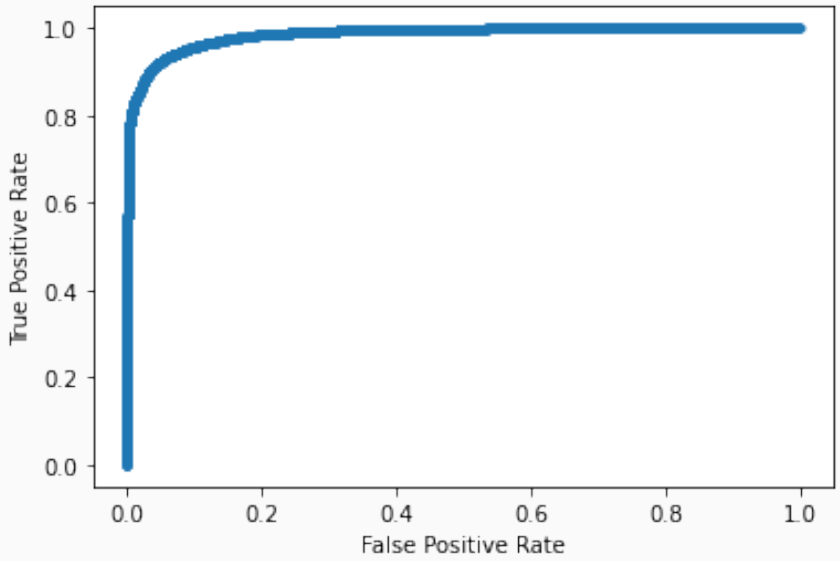
Slika 5: Chi-kvadrat test

Rezultati istraživanja

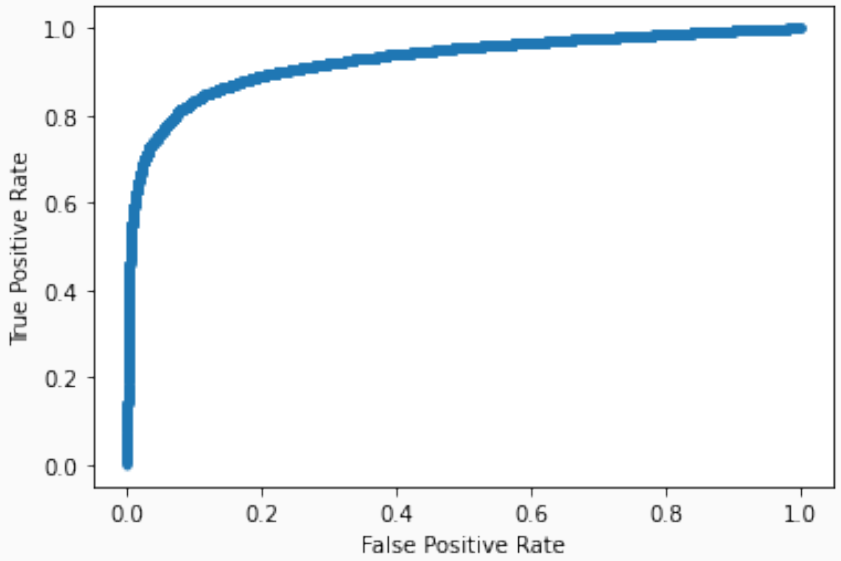
Rezultati istraživanja i

- Koristili smo tri metode za binarnu klasifikaciju: *Random Forest*, *linearnu regresiju* te *naivnog Bayesa*.
- Za evaluaciju modela koristili smo *Reciever Operating Characteristic* ili *ROC* krivulju kao vizualizaciju performansi modela.
- Za regresiju nad originalnim podacima dobili smo *ROC Score* od 0.9132 sa točnošću 0.8711.
- Za regresiju nad očišćenim podacima dobili smo *ROC Score* od 0.9854 sa točnošću 0.9379.
- Za *naivnog Bayesa* dobili smo *ROC Score* 0.9273 sa točnošću 0.8708

Rezultati istraživanja ii



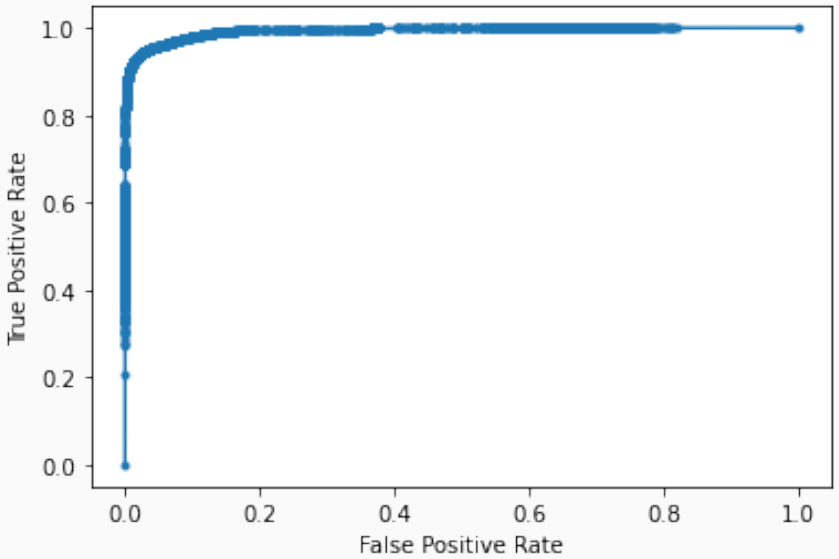
Slika 6: *ROC Curve* za logističku regresiju koristeći 10 najrelevantnijih značajki



Slika 7: *ROC Curve* za naivnog Bayesa koristeći 10 najrelevantnijih značajki

Rezultati istraživanja iii

- Implementirali smo i *Random Forest* model kako bi postigli bolje rezultate.
- Za navedni model *ROC Score* iznosi 0.993401750 sa točnošću 0.960344.
- Zaključno, navedeni model nam očito daje najbolje rezultate.



Slika 8: *ROC Curve* za *Random Forest* koristeći 10 najrelevantnijih značajki

Hvala na pažnji ☺