Binarna klasifikacija: Predviđanje zadovoljstva putnika

Strojno učenje

U Zagrebu, 15.lipnja 2022.

Nikola Sunara
Matematički odsjek
Prirodoslovno–matematički fakultet
Zagreb, Hrvatska
nikola.sunara@student.math.hr

Fran Vojković
Matematički odsjek
Prirodoslovno–matematički fakultet
Zagreb, Hrvatska
fran.vojkovic@student.math.hr

Alen Živković
Matematički odsjek
Prirodoslovno–matematički fakultet
Zagreb, Hrvatska
alen.zivkovic@student.math.hr

***Sažetak*—Rad proučava dataset o zadovoljstvu putnika neke avio–kompanije. Predočena je eksploratorna analiza podataka koje promatramo. Kasnije je proveden postupak odabira značajki relevantnih za opisani model te su opisani rezultati implementiranog modela.**

***Index Terms*—Airline Satisfaction, Passenger satisfaction**

SADRŽAJ

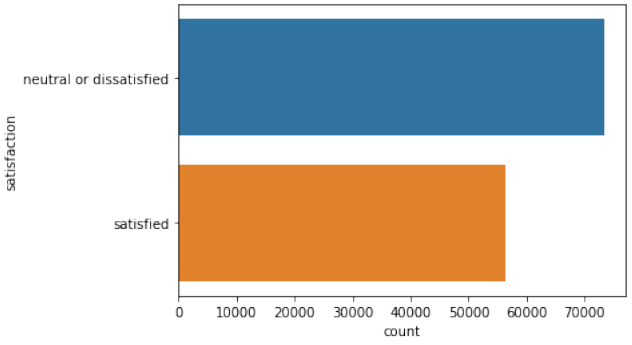
I. UVOD

U projektu se bavimo problemom predviđanja zadovoljstva putnika neke avionske kompanije. Proučavamo navedeni binarni klasifikacijski problem u kojem promatrano zadovoljstvo putnika može biti jedno od sljedećih: neutralno ili nezadovoljavajuće (*neutral or dissatisfied*) te zadovoljavajuće (*satisfied*). Koristimo set podataka o zadovoljstvu putnika jedne aviokompanije (dostupan na [1]) koji sadrži 130 000 redova podataka raspoređenih na 25 značajki. Promatrani skup podataka sadrži i kategoričke i kontinuirane značajke. Kategoričke značajke smo *encodali* na razne načine onako kako se činilo ispravno te je provedeno čišćenje podataka od nedostajućih vrijednosti i previše koreliranih značajki. Također provedeno je nekoliko metoda za odlučivanje o odbacivanju značajki. U poglavlju III prikazan je odabir deset najrelevantnijih značajki za naš model. Također, prikazali smo i rezultate najtočnijeg modela te rezultate modela slabije točnosti.

II. Analiza dataseta

A. Vizualizacija dataseta

Promatrani dataset sadrži sljedeće relevantne značajke: *id*, *Gender*, *Customer Type*, *Age*, *Type of Travel*, *Class*, *Flight Distance*, *Inflight wifi service*, *Departure/Arrival time convenient*, *Ease of Online booking*, *Gate location*, *Food and drink*, *Online boarding*, *Seat comfort*, *Inflight entertainment*, *On-board service*, *Leg room service*, *Baggage handling*, *Checkin service*, *Inflight service*, *Cleanliness*, *Departure Delay in Minutes*, *Arrival Delay in Minutes*, *satisfaction*. Putnici su nakon leta ocijenili svoje zadovoljstvo promatranom avio–kompanijom te su dobiveni rezultati prikazani na slici 1. Kategorički

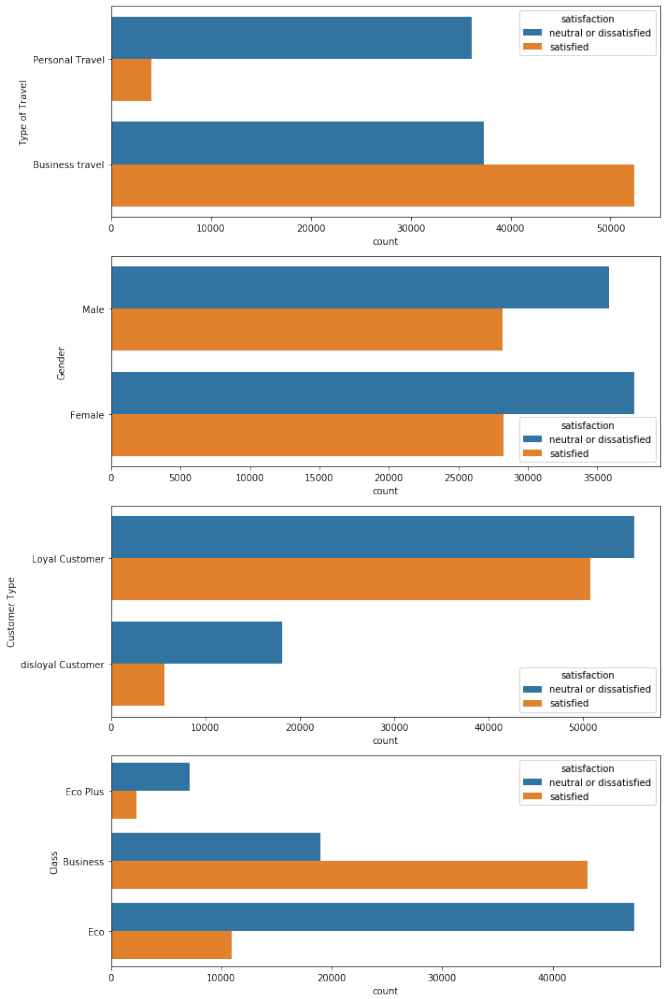


Slika 1. Zadovoljstvo putnika

stupci, *Type of travel*, *Class*, *Gender*, *Customer Type*, u datasetu imaju jedinstvene vrijednosti vidljive na slici 2 koje smo *encodali* u nule ili jedinice. Kategorički stupac *Class* smo *one hot encodali* da bi izbjegli ono što bi model mogao interpretirati kao težinu između tri kategorije. Nadalje, slika 3 prikazuje pojedinačne histograme za svaku od značajki i za svaki ishod ciljane varijable.

B. Obrada podataka

U promatranom datasetu postoje vrijednosti koje nedostaju, kako je njihov postotak približno 0.30% te se pojavljuju u samo jednom stupcu, navedene podatke možemo izostaviti. Primjenom *FAMD* (*Factor analysis of mixed data*) metode izračunali smo objašnjenu varijancu iz čega zaključujemo da prve četiri značajke imaju velik udio u objašnjenoj varijanci. Standardizirali smo sve značajke funkcijom *StandardScaler* koja oduzima srednju vrijednost i standardizira sve značajke tako da imaju jediničnu

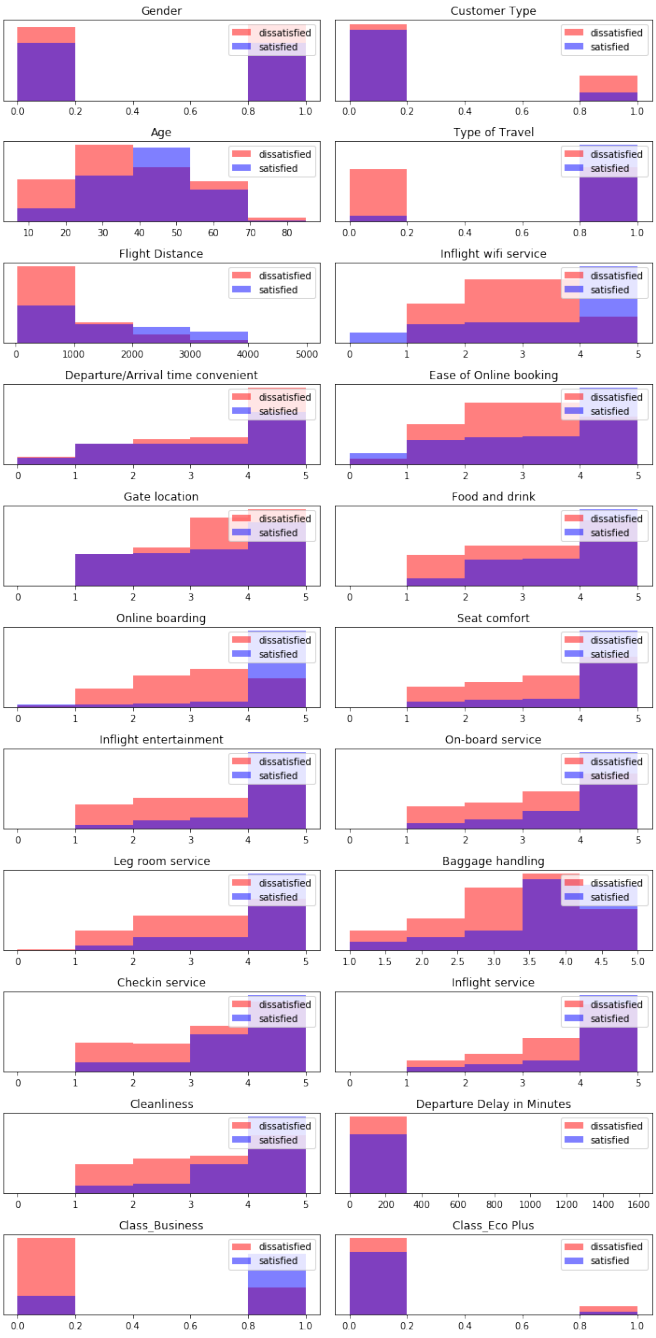


Slika 2. Jedinstvene vrijednosti kategorijalnih stupaca

varijancu. Također iskoristili smo analizu glavnih komponenti *PCA* (*Principal component analysis*) kako bi projicirali podatke u novi prostor značajki u kojem su one ortogonalne i sortirane po količini varijance koju objašnjavaju u ortogonalnim podacima. Na slici 4 prikazana je varijanca objašnjena sa svakom od glavnih komponenti zajedno sa varijancom objašnjenom uključivanjem svake naredne glavne komponente u projiciranom datasetu. Iz navedenog uočavamo da nećemo moći značajno reducirati broj značajki.

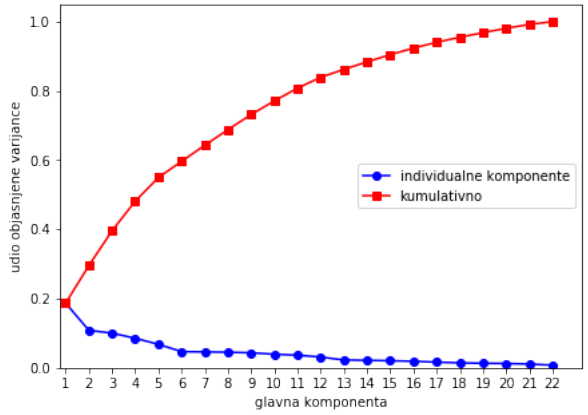
III. Odabir značajki

Kako bi odabrali podskup relevantnih značajki za naš model u svrhu snanjenja računalnih zahtjeva te poboljšanja prediktivnosti modela provodimo odabir značajki za model. Konstruirali smo *heatmap* svih



Slika 3. Histogrami značajki

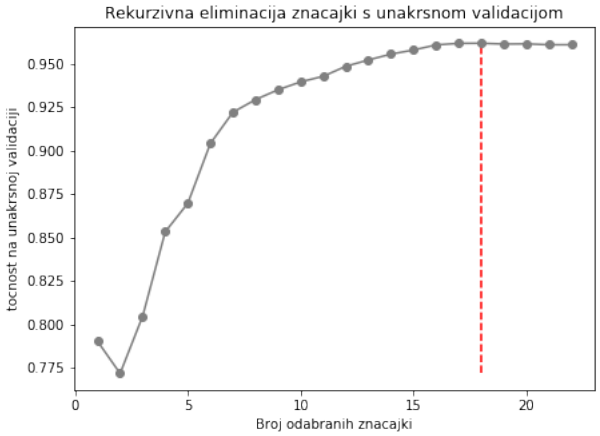
značajki, dostupan u [5], iz kojega vidimo korelaciju značajki promatranog dataseta. Iz priloženog *heat-mapa* uočavamo da je zadovoljstvo putnika visoko korelirano s vrstom putovanja, online boardingom te radi li se o business klasi, dok je najslabije korelirana s lokacijom vrata i spolom putnika.



Slika 4.

A. Rekurzivna eliminacija značajki s unakrsnom validacijom

Kako bi utvrdili optimalan broj značajki proveli smo rekurzivnu eliminaciju s unakrsnom validacijom koristeći funkciju `sklearn.feature_selection.RFECV` sa stratificiranom unakrsnom validacijom `sklearn.cross_validation.StratifiedKFold` te kao rezultat dobijemo 18 značajki kao što je prikazano na slici 5. Slika 6 ilustrira važnost

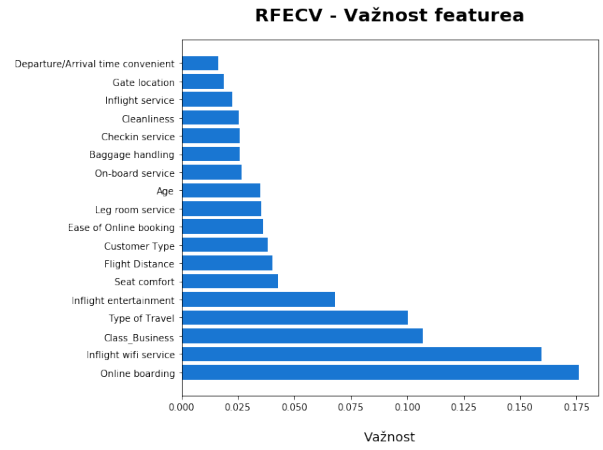


Slika 5.

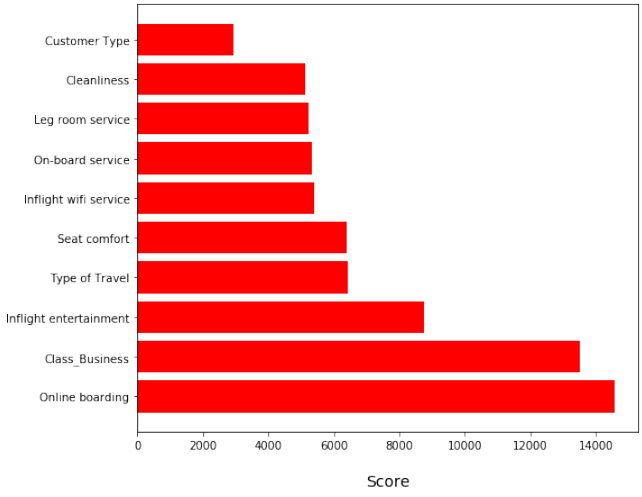
preostalih, odabranih značajki. Iz navedenog vidimo da posljednje četiri navedene značajke imaju znatno veću važnost od ostalih.

B. Chi²

Proveli smo *Pearsonov Chi-kvadrat test* nad kategorijalnim značajkama, prikazan na slici 7, kako

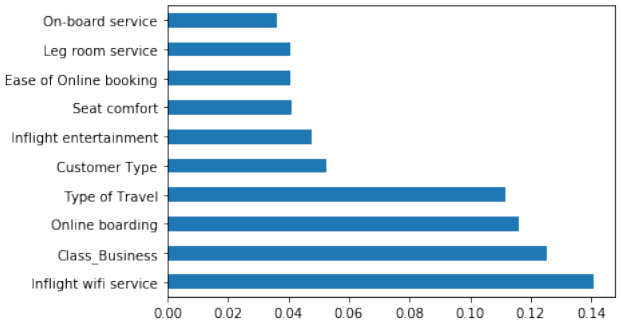


Slika 6. Važnost značajki



Slika 7. Chi-kvadrat test

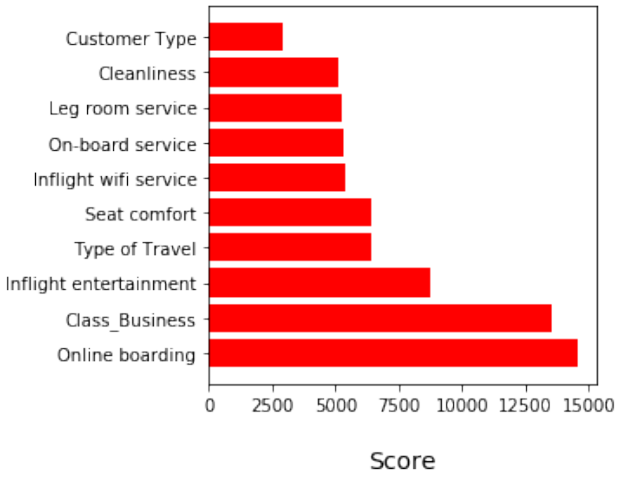
bi odredili 10 najznačajnijih značajki. Iz navedenog uočavamo da se kao najvažnije značajke pojavljuju opet *Online boarding* i *Class_business* kao i u prethodnom poglavlju III-A. Također proveli smo i metodu *Extra Trees Classifier* za deset najrelevatnijih značajki kako bi mogli usporediti rezultate. Rezultat metode prikazan je na slici 8. Iz navedenog vidimo da se rezultati algoritama razlikuju no postoje i neke jednako rangirane značajke. Značajke za business klasu i vrsta putovanja su jednako rangirane pomoću obje metode, što nam ukazuje da su najrelevantnije za naš model. Značajka *wifi service* malo je manje relevantna jer je slabije rangirana u Chi^2 metodi dok je značajka za online boarding relevantna u obje metode budući da se nalazi na prva tri mjesta u oba slučaja.



Slika 8. Extra Trees Classifier

C. *Mutual information feature selection*

Primjenili smo i metodu *Mutual information selection* koja računa redukciju neodređenosti (neizvjesnosti) jedne varijable za danu vrijednost druge varijable. Navedena metoda je implementirna u `mutal_info_classif()`. Kao i u prethodom poglavlju III-B, navedenu metodu možemo primjeniti kako bi dobili n najrelevantnijih značajki za naš model. Kao i u prethodnoj metodi, dobivamo slične



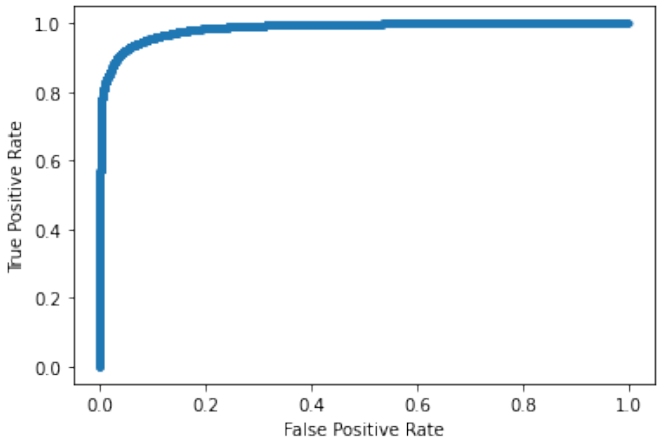
Slika 9. Mutual information feature selection

rezultate, koji su prikazani na slici 9.

IV. Rezultati istraživanja

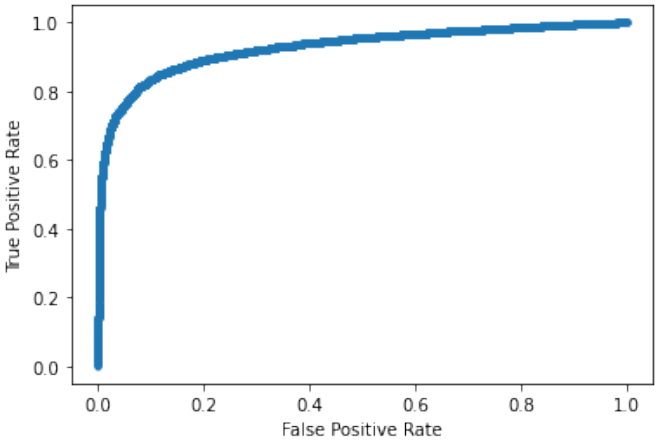
Konstruirali smo model za promatrani skup podataka u kojem smo koristili tri metode za binarnu klasifikaciju: *Random Forest*, *linearnu regresiju* te *naivnog Bayesa*. Navedene rezultate te samu implementaciju moguće je naći u [5]. Za evaluaciju modela koristili smo *Reciever Operating Characteristic* ili *ROC* krivulju kao vizualizaciju performansi

modela. *ROC score* grafički reprezentira *True Posotive Rate* naprema *False Positive Rate*. Proveli smo logističku regresiju nad polaznim podacima te smo dobili *ROC Score* od 0.9132 sa točnošću 0.8711. Kako bi poboljšali naš model očistili smo podatke od manje relevantnih značajki, kao što je opisano u prethodnom poglavlju III te smo za 10 najrelevantnijih značajki dobili sljedeće rezultate. *ROC Score* od 0.9854 sa točnošću 0.9379. Iz navedenog

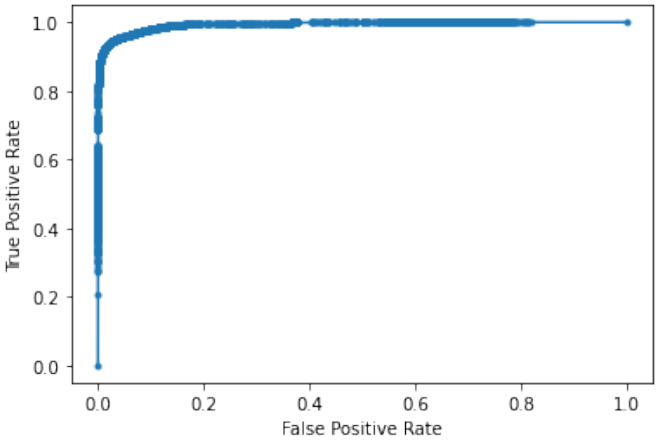


Slika 10. *ROC Curve* za logističku regresiju koristeći 10 najrelevantnijih značajki

vidimo očito značajno poboljšanje u vidu točnosti predikcije te *ROC Scorea* nad očiščenim podacima. Nadalje, proveli smo i naivnog Bayesa za polazne podatke te smo dobili *ROC Score* 0.0917527618 sa točnošću 0.8587. Kada reduciramo značajke na naših 10 najrelevantnijih, dobijemo *ROC Score* 0.9273 sa točnošću 0.8708 te pripadnom *ROC* krivuljom prikazanom na slici 11. Zasada zaključujemo da nam naivni Bayes ne predstavlja pouzdaniji model naspram prethodno opisane regresije. Kako bi poboljšali točnost našeg modela implementirali smo i *Random Forest* model. Za navedni model *ROC Score* iznosi 0.993401750 sa točnošću 0.960344 te pripadnom *ROC* krivuljom prikazanom na slici 12. Iz navedenog zaključujemo da nam *Random Forest* daje najbolji model za promatrane podatke.



Slika 11. *ROC Curve* za naivnog Bayesa koristeći 10 najrelevantnijih značajki



Slika 12. *ROC Curve* za *Random Forest* koristeći 10 najrelevantnijih značajki

LITERATURA

[1] *Airline Passenger Satisfaction*. URL: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction.

[2] Jason Brownlee. *How to perform feature selection with categorical data*. Travanj 2020. URL: https://machinelearningmastery.com/feature-selection-with-categorical-data/.

[3] *Extra Tree Classifier for feature selection*. Travanj 2020. URL: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/.

[4] Sarang Narkhede. *Understanding AUC - ROC Curve*. Travanj 2020. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[5] NS, FV i AŽ. „Projekt Binarna klasifikacija: Predvidanje zadovoljstvaputnika u Pythonu".

[6] Manish Patak. *Handling categorical data in Python*. Travanj 2020. URL: https://www.datacamp.com/community/tutorials/categorical-data.

[7] Dario Radečić. *Feature selection in Python - Recursive Feature Elimination*. Travanj 2020. URL: https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d1.

[8] Ren Jie Tan. *A starter pack for exploratory data analysis with python, pandas, seaborn and scikit learn*. Travanj 2020. URL: https://towardsdatascience.com/a-starter-pack-to-exploratory-data-analysis-with-python-pandas-seaborn-and-scikit-learn-a77889485baf.