SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO - MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Strojno učenje

BINARNA KLASIFIKACIJA: PREDVIĐANJE ZADOVOLJSTVA PUTNIKA

PROJEKTNI PRIJEDLOG

Martina Gaćina
Nikola Sunara
Fran Vojković
Alen Živković

travanj, 2020.

1 Uvodni opis problema

U ovom projektu bavit ćemo se problemom predviđanja zadovoljstva putnika neke avionske kompanije. Radi se o binarnom klasifikacijskom problemu u kojem traženo zadovoljstvo može biti jedno od: neutral or dissatisfied (neutralni ili nezadovoljni) ili satisfied (zadovoljni). Koristit ćemo skup podataka pronađen na kaggleu (Airline Passenger Satisfaction) koji sadrži oko 130 000 redova podataka raspoređenih na 25 značajki. Dataset sadrži i kategorične i kontinuirane značajke pa smo kategorične encodali na različite načine onako kako se činilo ispravno te proveli čišćenje podataka od nedostajućih vrijednosti i previše koreliranih značajki kao i nekoliko metoda za odlučivanje o odbacvianju značajki.

2 Cilj i hipoteze istraživanja problema

Cilj je ispitati nekoliko standardnih metoda strojnog učenja na ovom skupu podataka te konstruirati najbolji mogući model za klasifikaciju zadovoljstva danog putnika.

3 Pregled dosadašnjih istraživanja

Budući da se radi o relativno novom skupu podataka, imamo tek nekoliko dosadašnjih istraživanja koja možemo analizirati. Koristili su se razni klasifikatori kako bi se stvorio željeni model poput: Random Foresta, LightGBM, Catboost, XGBoost, linearne regresija i tako dalje. Držimo da sva ova istraživanja problema dosad nisu provela adekvatnu analizu značajki, kao ni njihovo odbacivanje te se na tom području nadamo znatno poboljšati naše modele. Općenito se binarno klasificiranje zna rješavati i kompliciranijim metodama poput neuronskih mreža ili strojem s potpornim vektorima, ali ovakve metode znaju zahtjevati jake računalne resurse.

4 Materijali, metodologija i plan istraživanja

Pokušat ćemo konstruirati najbolji mogući model za dan skup podataka koristeći metode nadziranog učenja. Planiramo koristiti barem tri metode za binarnu klasifikaciju, to jest: Random Forest, linearnu regresiju i naivnog Bayesa, ali razmišljamo i o drugim mogućim metodama za postizanje rezultata. Naravno, ovo ćemo provesti u Pythonu. Svjesni smo da bi neuralna mreža mogla dobro funkcionirati na ovom skupu podataka, ali postoji mogućnost da će njena upotreba zahtjevati preveliko odbacivanje značajki i tako izgubiti na vjerodostojnosti. Nastojat ćemo raznim metodama za odabir značajki odabrati one najutjecajnije tako da model ostane precizan, ali i postane efikasniji. U priloženoj analizi podataka već smo proveli nekoliko takvih metoda i primijetili zanimljive uzorke koje valja dalje istražiti. Skup podataka kojim se bavimo već je podjeljen na testni i trening skup, no odlučili smo ih spojiti, provesti analizu i već prije spomenute transformacije (brisanje nedostajućih vrijednosti, odbacivanje jednog od para koreliranih značajki, encodanje kategoričnih varijabli) na čitavom skupu te ih onda pomoću train test splita podijeliti na trening i test skupove. Planiramo koristiti nekoliko načina za evaluaciju modela, a među njima i roc_auc metriku koja bi trebala dobro funckionirati na relativno balansiranom skupu podataka po pitanju mete.

5 Očekivani rezultati predloženog projekta

Teško je reći koja su realna očekivanja projekta, budući da se radi, kako smo već rekli, o relativno novom skupu podataka. Na temelju nekoliko već provedenih istraživanja, nadamo se da ćemo uspjeti konstruirati model točnosti barem 90 posto na testnom skupu podataka.

Literatura

[1] Towards Data Science, Sarang Narkhede. *Understanding AUC - ROC Curve*
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[2] Towards Data Science, Ren Jie Tan. *A starter pack for exploratory data analysis with python, pandas, seaborn and scikit learn*
https://towardsdatascience.com/a-starter-pack-to-exploratory-data-analysis-wi
249d

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

[4] Data camp, Manish Patak. *Handling categorical data in Python*
https://www.datacamp.com/community/tutorials/
categorical-data

[5] Machine Learning Matery, Jason Brownlee. *How to perform feature selection with categorical data*
https://machinelearningmastery.com/feature-selection-with-categorical-data/

[6] Towards Data Science, Dario Radečić. *Feature selection in Python - Recursive Feature Elimination*
https://towardsdatascience.com/feature-selection-in-python-recursive-feature-