Uticaj metoda uzorkovanja na imputaciju nedostajućih vrednosti

Stefan Kerkoč

Uvod u teoriju uzoraka

Uvod

Imputacija podataka, odnosno proces procene nedostajućih vrednosti u skupovima podataka, igra ključnu ulogu u poboljšanju pouzdanosti donošenja odluka u različitim modelima baziranim na metodu odabira uzorka. Dostupnost potpunih i tačnih podataka direktno utiče na efikasnost i stabilnost modela zaključivanja i analitičkih alata. Ovaj rad istražuje uticaj tehnika imputacije podataka na procese donošenja odluka unutar različitih modela baziranih na metodu odabira uzorka. Analizom kako imputirani podaci utiču na tačnost modela, njegovu stabilnost i opšti kvalitet donošenja odluka, ovaj rad ima za cilj da pruži uvid u optimizaciju strategija uzorkovanja pri ovom pristupu. Koristili smo metod višestruke imputacije, a testirali smo na algoritmu logističke regresije i stablu odlučivanja.

Teorijska osnova

U ovom poglavlju ćemo se baviti time šta je imputacija, metod imputacije i algoritme koje smo koristili u ovom radu.

Imputacija

Imputacija je statistička tehnika koja se koristi za procenu nedostajućih vrednosti u skupovima podataka. Nedostajući podaci su čest problem u istraživanjima i praktičnim primenama, nastali iz različitih razloga kao što su nedostatak odgovora u istraživanjima ili greške pri unosu podataka. Popunjavanjem nedostajućih vrednosti procenjenim vrednostima, imputacija ima za cilj da poboljša potpunost i pouzdanost skupova podataka, čime se unapređuje robustnost kasnijih analiza i procesa donošenja odluka [5].

Višestruka imputacija

U ovom radu ćemo koristiti višestruku imputaciju. Višestruka imputacija je sofisticirana tehnika imputacije koja uključuje stvaranje više verovatnih vrednosti za svaki nedostajući podatak, čime se uzima u obzir nesigurnost u procesu imputacije. Ovi višestruki skupovi podataka se analiziraju odvojeno, a rezultati se kombinuju koristeći specijalizovana pravila koja odgovaraju uvođenju varijabilnosti koju uvodi imputacija. Višestruka imputacija ne samo da pruža tačnije procene nedostajućih vrednosti, već takođe daje pouzdanije standardne greške i intervale poverenja u poređenju sa metodama jednostruke imputacije. [4].

Logistička Regresija

Logistička regresija je statistička metoda koja se koristi za modeliranje odnosa između binarne zavisne varijable i jedne ili više nezavisnih varijabli. Za razliku od linearne regresije, logistička regresija se koristi kada je zavisna varijabla kategorička, obično binarna (npr. uspeh/neuspeh, da/ne). Logistički regresioni model je predstavljen logističkom funkcijom, poznatom i kao sigmoidna funkcija, koja preslikava predviđene vrednosti u verovatnoće:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

U ovoj jednačini, $P(Y = 1)$ je verovatnoća da zavisna varijabla Y iznosi 1, β_0 je intercept, a $\beta_1, \beta_2, \ldots, \beta_n$ su koeficijenti nezavisnih varijabli $X_1, X_2, \ldots, X_n$. [2].

Pretpostavke Logističke Regresije

1. **Linearnost Logita**: Logiti zavisne varijable su linearna kombinacija nezavisnih varijabli.
2. **Nezavisnost**: Posmatranja su nezavisna jedna od drugih.
3. **Bez Multikolinearnosti**: Nezavisne varijable nisu previše visoko korelisane jedna s drugom.

Stabla odlučivanja

Stabla odlučivanja su supervizovane mašinske metode učenja koje se koriste za klasifikaciju i regresiju. Ove metode konstruišu model u formi stabla sa čvorovima i granama, gde svaki čvor predstavlja test na određenoj karakteristici, a grane predstavljaju moguće rezultate tog testa. Cilj je razdvajanje skupa podataka na što homogenije podskupove u pogledu ciljne promenljive.

Stabla odlučivanja rade na principu rekurzivnog deljenja skupa podataka na manje podskupove, sve dok se ne postigne određeni kriterijum zaustavljanja, kao što su maksimalna dubina stabla ili minimalni broj instanci u listovima. Ovi modeli su transparentni i intuitivni za interpretaciju, što ih čini popularnim alatom za istraživanje podataka i analizu.

Prednosti stabala odlučivanja uključuju sposobnost rada sa numeričkim i kategoričkim podacima, otpornost na preprilagođavanje (overfitting) pri pravilnom podešavanju parametara, kao i mogućnost da rade sa nedostajućim podacima koristeći odgovarajuće tehnike imputacije.

U praksi, stabla odlučivanja često služe kao osnova za složenije ansamble modela kao što su slučajne šume (random forests) ili gradijentno pojačavanje (gradient boosting), čime se poboljšava performansa i generalizacija modela.[1].

Praktični deo

Glavna ideja

Odabrali smo skup podataka takav da ima mali broj redova i kolona, kako bi uticaj imputacije bio upečatljiviji. U takav kompletan skup podataka smo uveli nedostajuće vrednosti kako bi imali kontrolu nad njihovim udelom. Razlog za ovo je što ćemo sami kreirati nedostajuće vrednosti u jednoj od nekoliko kolona tako što ćemo vrednosti na 1/5 nasumično odabranih redova staviti na NA. Odabrali smo baš jedan od pet jer se to već smatra velikim udeom nedostajućih vrednosti. [4].

Skup podataka koji smo odabrali je **nycflights13**, skup o letovima iz njujorka, s tim da smo izabrali samo 3 kolone: **depdelay**, **airtime**, **distance**, dok smo predviđali da li će avion kasniti (poleteti 15 nakon predviđenog vremena polaska).

Koristili smo prosto slučajno uzorkovanje i stratifikovano uzorkovanje. Testirali smo na skupu podataka sa imputovanim podacima i na skupu podataka gde su redovi sa nedostajućim vrednostima zanemareni.

Imputacija

Za imputaciju je korišćen paket **mice**, pravljenjem 5 višestrukih skupova podataka za svaki trening skup. Podaci su bili imputovani na trening skupu u zavisnosti metoda uzorkovanja, a model koji je istreniran tu je korišćen i za test skup.

Logistička Regresija

Logističku regresiju smo realizovali koristeći **glm** funkciju, podešavajući parametar **family** na **binomial**. Logistička regresija nam daje solidne rezultate po pitanju odlučivanja pri svakoj metodi, ali to ne znači da nam nije dala nikakve informacije po pitanju našeg problema. Naime, modeli pravljeni na imputovanim podacima su gotovo istog kvaliteta bez obzira na metod uzorkovanja, dok modeli pravljeni na podacima gde su izbačene nedostajuće vrednosti daju model sa najvećom i najmanjom tačnošću, sugerišući na to da se ovom metodom prave nestabilniji modeli. Sa manjim brojem podataka, stratifikacija daje, očekivano, znatno bolje rezultate. [3].

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	44	5
	Kasni	1	10

Tabela 1: Prosto slučajno uzorkovanje - imputovani podaci

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	43	5
	Kasni	2	10

Tabela 2: Stratifikovano uzorkovanje - imputovani podaci

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	35	6
	Kasni	0	7

Tabela 3: Prosto slučajno uzorkovanje - izbačeni redovi

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	35	3
	Kasni	0	10

Tabela 4: Stratifikovano uzorkovanje - izbačeni redovi

Stabla odlučivanja

Stablo odličivanja smo realiovali koristeći paket **rpart**. Ovo stablo je pomalo razočaravajuće u smislu kompleksnosti, jer uzima u obzir samo kolonu **depdelayed**, procenivši da ostale nisu bitne. Daje gore rezultate po pitanju tačnosti u odnosu na logističku regresiju. Što se tiče glavnog problema, daje nam isti zaključak kao i logistička regresija, a to je da nam imputacija daje stabilnost, dok izbacivanje može dati možda i najbolji model po pitanju tačnosti, ali moramo paziti na metod uzorkovanja i generalno zavisnost od podataka. [3].

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	43	8
	Kasni	2	7

Tabela 5: Prosto slučajno uzorkovanje - imputovani podaci

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	39	5
	Kasni	5	10

Tabela 6: Stratifikovano uzorkovanje - imputovani podaci

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	31	7
	Kasni	5	5

Tabela 7: Prosto slučajno uzorkovanje - izbačeni redovi

		Predviđeno	
		Na vreme	Kasni
Stvarno	Na vreme	34	3
	Kasni	2	8

Tabela 8: Stratifikovano uzorkovanje - izbačeni redovi

Potrebni R Paketi

Za pokretanje koda, potrebno je instalirati sledeće R pakete:

- **nycflights13**
- **mice**
- **caret**
- **rpart**
- **rpart.plot**

Zaključak

Korišćenjem robusne tehnike imputacije kao što je višestruka imputacija možemo jako dobro procentii originalni skup podataka, što nam može svakako pomoći u odlučivanju. Ovakva tehnika nam i omogućava da koristimo različite tehnike uzorkovanja koje nam u zavisnosti od problema mogu biti pogodne. Tehnika kao što je izbacivanje redova sa nedostajućim vrednostima nam možda u određenim situacijama može dati bolje rešenje, ali nam imputacija daje veću fleksibilnost.

Literatura

[1] Leo Breiman, Jerome Friedman, Charles J. Stone, and Robert A. Olshen. *Classification and Regression Trees*. CRC Press, 1984.

[2] Andy Field. *Discovering Statistics Using IBM SPSS Statistics*. Sage, 2013.

[3] Stefan Kerkoč. utu-projekat. `https://github.com/kerkoc01/utu-projekat/tree/main`, 2024.

[4] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2019.

[5] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.