Sara Borzić

GDi STEM nagrada za izvrsnost:
Primjena modela strojnog učenja za predikciju trajanja istovara vozila

Dokumentacija

Listopad 2020. godine, Zagreb

1. Opis rješenja

Glavne biblioteke korištene prilikom izrade programskog rješenja su *pandas* i *scikit learn*. Nakon obavljenog učitavanja datoteke *dataset_GDi_STEM_2021.rpt,* sadržaj je pohranjen u *DataFrame* objekt biblioteke *pandas* koja sadrži mnoge operacije i funkcionalnosti za rad sa skupovima podataka. Skup podataka se sastoji od narudžbi te svaki redak sadrži sljedeće značajke: *date, vehicleID locationID, orderID, scheduledDeliveryTWStart, scheduledDeliveryTWEnd, scheduledDeliveryTWDuration, realDeliveryServiceStart, realDeliveryServiceEnd, realDeliveryServiceDuration, totalOrderVolume, totalOrderWeight, totalOrderQuantity i totalOrderDistinctQuantity*.

Značajka *orderID,* tj. ID narudžbe nije koristan u predikciji te je izbrisan taj stupac. Različiti stupci, čije formate *pandas* nije automatski prepoznao, su na početku prebačene u ispravan format – decimalni brojevi su prebačeni u izraz s decimalnom točkom, a ne sa zarezom te su značajke koje prikazuju datum ili vrijeme prebačene u *datetime* oblik.

Zbog postojanja više narudžbi čiji je istovar započeo istovremeno, koje su dostavljene istim vozilom i na istu lokaciju te im je jednako trajanje, postojala je potreba za grupiranjem, tj. zbrajanjem količine, težine, volumena i različitosti artikala u takvim narudžbama radi točnijeg skupa podataka i bolje iskoristivosti istoga za zadatak predikcije trajanja istovara. Kada to ne bi bilo napravljeno, model bi krivo protumačio da, na istoj lokaciji, u isto vrijeme, narudžbi s 2 predmeta male mase i volumena, istovar traje jednako kao i narudžbi s 10 predmeta velike mase i volumena.

Nove značajke *realDeliveryStartHour* i *scheduledStartHour* su kreirane iz postojećih značajki *realDeliveryServiceStart* i *scheduledDeliveryServiceStart* radi lakšeg baratanja podatcima o vremenu početka zakazanog i stvarnog istovara. Korištena pretpostavka je da vrijeme početka istovara također uvjetuje dužinu trajanja - primjerice, u vremena kada je veći promet, dostavljaču bi se moglo žuriti što brže obaviti narudžbu; pri kraju radnog vremena bi mogao postati umoran te bi to također moglo utjecati na vrijeme istovara, itd. Također, i zakazani početak istovara bi mogao imati utjecaj na vrijeme istovara, kao i vremensko kašnjenje u odnosu na zakazani rok istovara (dostavljaču će se žuriti dostaviti što prije u slučaju većih kašnjenja). Pritom je kreirana značajka *lateness*, tj. kašnjenje koja predstavlja broj minuta koliko je istovar počeo kasnije od predviđenog kraja istovara, odnosno koliko je istovar narudžbe zakasnio.

Kontinuirane značajke je bilo važno skalirati, a zbog postojećih *outlier*-a, odlučeno je da će se odviti pomoću standardnog ili normalnog skaliranja.

Uslijedila je analiza zavisnosti kontinuiranih značajki, tj. ciljne značajke*, realDeliveryServiceDuration* i preostalih kontinuiranih pomoću matrica zavisnosti. Korištene su dvije metode: računanje Pearsonovog koeficijenta korelacije i Spearmanovog koeficijenta korelacije ranga. Pearsonovim koeficijentom se mjeri linearna zavisnost dviju značajki, a Spearmanovim koliko je moguće odnos dviju značajki prikazati monotonom funkcijom. Iz tablice izlistane u priloženom programskom kodu može se iščitati prvi redak matrice Pearsonovog koeficijenta korelacije (Tablica 1), odnosno Spearmanovog koeficijenta (Tablica 2).

	Trajanje istovara	Količina	Masa	Volumen	Količina različitih artikala	Kašnjenje narudžbe	Očekivano trajanje narudžbe
Trajanje istovara	1.000	0.031	0.043	0.052	0.043	-0.060	0.110

Tablica 1.

Najmanja linearna zavisnost je očito između izračunate značajke Kašnjenje narudžbe (l*ateness*) i Trajanja istovara (*realDeliveryServiceDuration*), dok je Očekivano trajanje narudžbe (*scheduledDeliveryTWDuration*) značajka najbliža linearnoj ovisnosti o istoj.

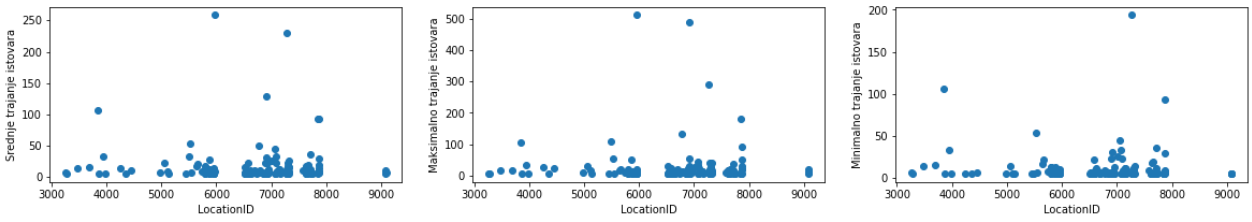
	Trajanje istovara	Količina	Masa	Volumen	Količina različitih artikala	Kašnjenje narudžbe	Očekivano trajanje narudžbe
Trajanje istovara	1.000	0.096	0.072	0.093	-0.001	-0.010	0.129

Tablica 2

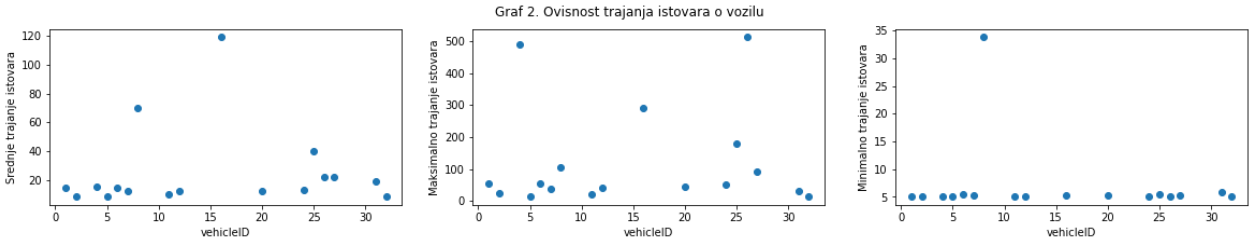
Spearmanov koeficijent govori kako se značajke Količina različitih artikala (*totalOrderDistinctQuantity*) i Kašnjenje narudžbe (*lateness*) ne mogu prikazati u monotonoj ovisnosti o značajki Trajanje istovara (*realDeliveryServiceDuration*), dok se ponovno u najvećoj mjeri može monotona ovisnost prikazati između navedene značajke i Očekivanog trajanja narudžbe (*scheduledDeliveryTWDuration*). To daje zaključak kako bi ta značajka mogla biti korisna u modelu. Spearmanovi koeficijenti između *realDeliveryServiceDuration*-a i preostalih značajki su pozitivni i međusobno vrlo slični te bi zbog toga bi mogla postojati slična zavisnost u svakom od tih parova.

Osim numeričke, pri odabiru modela, bila je korisna i grafička analiza zavisnosti komponenti. Graf 1 prikazuje ovisnost redom srednjeg, maksimalnog I minimalnog trajanja istovara za svaku lokaciju. Jasno je vidljivo da postoji ovisnost trajanja istovara o lokaciji.

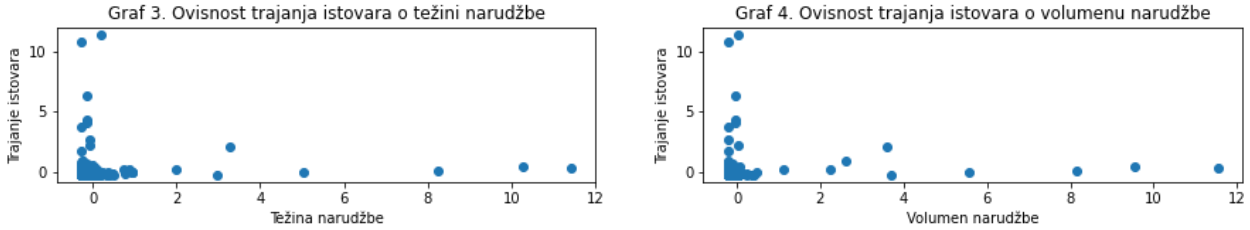
Graf 1. Ovisnost trajanja istovara o lokaciji



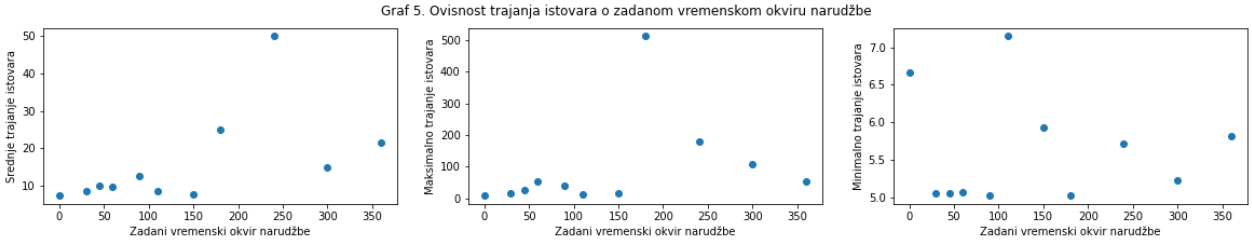
Graf 2 prikazuje ovisnost redom srednjeg, maksimalnog i minimalnog trajanja istovara za svako vozilo. Različita vozila ipak nemaju u tolikoj mjeri različite srednje, a ni minimalne iznose trajanja istovara te je odlučeno da podatak o vozilu neće biti uključen u konačni model predikcije.



Grafovi 3 i 4 prikazuju ovisnost trajanja istovara o težini, odnosno volumenu narudžbe te se može vidjeti da postoji velika razlika u trajanju istovara između različitih volumena i težina narudžbi. U skladu s prethodnim numeričkim analizama gdje je pokazano da postoji mala, ali pozitivna korelacija između ovih značajki i trajanja istovara, ove dvije značajke su također uvrštene u model.



Graf 5 prikazuje ovisnost redom srednjeg, maksimalnog i minimalnog trajanja istovara za svaki zadani vremenski okvir narudžbe. Kod prikaza ovisnosti srednjeg trajanja istovara o vremenskom okviru narudžbe, može se vidjeti naznaka ovisnosti koju smo mogli naslutiti kroz numeričke vrijednosti koeficijenata korelacije između ovih dviju značajki.



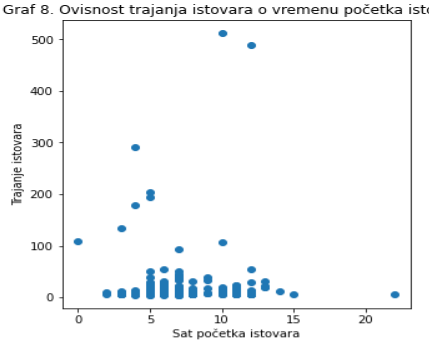
Graf 6 prikazuje ovisnost trajanja istovara o količini različitih artikala u narudžbi te je vidljivo kako je ovisnost uistinu vrlo mala te ta značajka nije uključena u model. Kod Grafa 7 može se vidjeti ovisnost sličnog izgleda kao i kod Grafova 3 i 4 te se količina artikala kao značajka uključuje u model.

Graf 6. Ovisnost trajanja istovara o količini različitih artikala u narudžbi

Graf 7. Ovisnost trajanja istovara o količini artikala u narudžbi

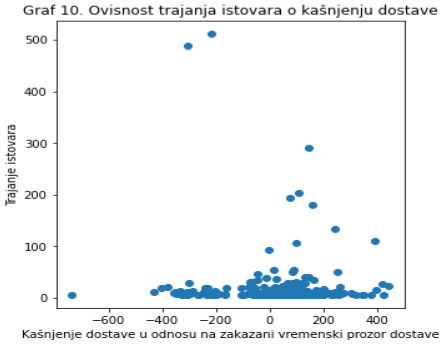
Graf 8 prikazuje ovisnost trajanja istovara o vremenu početka istovara te se vidi da postoji ovisnost, kao što se isto pokazuje i u Grafu 9 koji prikazuje trajanje istovara o zakazanom početku vremena istovara. Graf 10 prikazuje ovisnost trajanja istovara i kašnjenju i ta korelacija baš nije jasna, mnogo različitih vrijednosti kašnjenja ima isto trajanje istovara pa ta vrijednost nije uključena u model.

Graf 8. Ovisnost trajanja istovara o vremenu početka istovara



Graf 9. Ovisnost trajanja istovara o zakazanom početku vremena istovara

Graf 10. Ovisnost trajanja istovara o kašnjenju dostave



Kategoričke značajke (‘*locationID’, ‘realDeliveryStartHour’, ’scheduledStartHour’*) su prebačene u *one-hot* vektorske reprezentacije. Značajke koje su nakon obrade i analize utjecaja stavljene u model su sljedeće: *realDeliveryStartHour, locationID, scheduledDeliveryStartHour, totalOrderQuantity, totalOrderWeight, totalOrderVolume, scheduledDeliveryTWDuration.*

Ukupni skup podataka se podijelio u skupove za treniranje i testiranje u omjeru 0.85 - 0.15 (kako bi imali što više primjera u skupu za treniranje zbog korištenja peterostruke *cross*-validacije), te se pretraživanjem po rešetci došlo do optimalnih hiperparametara *gamma* i *C* za model. Korišteni model je jezgreni stroj (*SVM, točnije SVR*) s *rbf* jezgrom. *Gamma* predstavlja preciznost modela, tj. što je veća *gamma*, to će model više razlikovati dva primjera, dok *C* predstavlja širinu pojasa *rbf* jezgre, tj. što je *C* veći, on će više kažnjavati pogreške te dovoditi do složenijeg modela. Pomoću pretraživanja po rešetci unakrsnom provjerom dolazi se do modela koji ne bi trebao biti ni prenaučen ni podnaučen, već optimalan. To upravo radi *GridSearchCV* metoda, kojom se došlo do parametara *C* i *gamma* koji su optimalni za model: *C*=20.0 i *gamma*=0.05. *GridSearchCV* metoda je korištena i na drugim modelima (*SVR* s polinomnom jezgrom, Algoritam *K* susjeda, *RandomForestRegressor*, *MLPRegressor*), no oni su davali svi lošije rezultate peterostruke *cross*-validacije od *SVR*-a u kombinaciji s *rbf* jezgrom. Svi su testirani s istim *random seed*-om kako bi rezultat bio vjerodostojniji.

Modeli su evaluirani pomoću RMSE metrike, a rezultati peterostruke *cross*-validacije, kao i iznos RMSE-a kada se odabrani model trenira nad cijelim skupom za treniranje te testira nad skupom za testiranje su prikazani u sljedećem poglavlju. Unakrsna provjera se radila na unaprijed odvojenom skupu za treniranje. On se na početku podijelio na pet jednakih dijelova te se kroz 5 iteracija mijenjao ukupni skup za treniranje (4/5 skupa za treniranje) i skup za validaciju (1/5 skupa za treniranje).

2. Rezultati peterostruke *cross*-validacije u metrici RMSE

Cross-validacija:

```
0.prolaz, RMSE = 0.400460
1.prolaz, RMSE = 0.242760
2.prolaz, RMSE = 1.630895
3.prolaz, RMSE = 0.349831
4.prolaz, RMSE = 1.533368

Srednji RMSE = 0.831463
```

Model treniran nad cijelim skupom podataka za treniranje:

```
Skup za testiranje: RMSE = 0.232549
```

3. Upute za pokretanje programskog koda

Programski kod je pohranjen u *jupyter* bilježnici, a sve se nalazi na poveznici: https://github.com/sara-borzic/gdi_task.git .Potrebno je bilježnicu otvoriti u Google Colabu ili na svom računalu te u istom repozitoriju gdje je i ona spremiti (odnosno na *session storage* učitati, ukoliko se koristi Google Colab za pokretanje) datoteku *dataset_GDi_STEM_2021.rpt* iz *git* repozitorija s poveznice. Ukoliko nedostaje python3 ili neka datoteka korištena u programskom kodu, potrebno ju je instalirati prije korištenja koda.