Projektni zadatak: Iznajmljivanje bicikala

Lazar Mitrovic, IN20/2018, lazar.mitrovic@gmail.com
Aleksa Skrbic, IN29/2018, aleksa.shrbic@gmail.com

I Uvod

Izvestaj se bavi analizom meteorolskih podataka I njihovom uticaju na iznajmljivanje bicikala u Njujorku tokom 2011-te I 2012-te godine. Analizom nama dostupnih podataka moguce je kreirati model za predikciju sto moze doprineti optimizaciji celokupnog procesa iznajmljivanja bicikala I samim tim maksimizacijom profita.

II Baza podataka

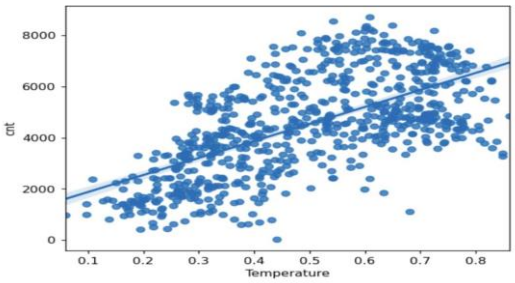
Na raspolaganju imamo imamo 2 dataseta s nazivom **day** I **hour**. **Hour** dataset sadrzi podatke o broju iznajmljenih biciklova za svaki sat zajedno sa meteoroloskim podacima za dati sat. Sadrzi **17378** uzoraka I **17** obelezja od kojih je **5** numericko I **12** kategorickih. Kategoricka obelezja su redni broj uzorka(**instant**), datum(**dteday**), godisnje doba(**season**), godina(**yr**), mesec(**mnth**), sat(**hr**), informacija o tome da li je bio praznik ili ne(**holiday**), dan u nedelji(**weekday**), informacija o tome da li je bio radni dan ili ne (**workingday**), ocena kvaliteta vremena (**weathersit**).Numericka obelezja su ukupan broj iznajmljenih bicikala za dati sat (**cnt**), broj iznajmljenih bicikala neregistrovanih korisnika(**casual**), broj iznajmljenih bicikala registrovanih korisnika(**registered**), temperatura(**temp**), vlaznost vazduha(**humi**), brzina vetra(**windspeed**) I subjektivni osecaj temperature(**atemp**). Dataset **day** sadrzi agregirane podatke dataseta **hour** odnosno sabrani su brojevi iznamljivanja bicikala za svaki sat I spojeni u jedan dan. Obelezja ova 2 dataseta su skoro identicna s tim da day ne sadrzi obelezje **hr**.

III Analiza podataka

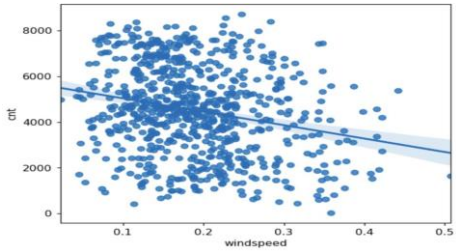
Nijedan od datasetova ne sadrzi nedostajuce vrednosti. Obelezja **casual** I **registered** su uklonjena jer cemo se fokusirati na predikciju I analizu samo ukupnog broja bicikala za sve korisnike. Takodje, obelezje **instant** je uklonjeno jer predstavlja samo redni broj uzorka I ne doprinosi analizi, obelezje datum je uklonjeno jer vec imamo podatke o mesecu I godini.Na ***slikama 1,2…6*** bice analizirana obelezja dataseta **day.**

A. Uticaj vremenskih prilika na broj iznajmljenih bicikala

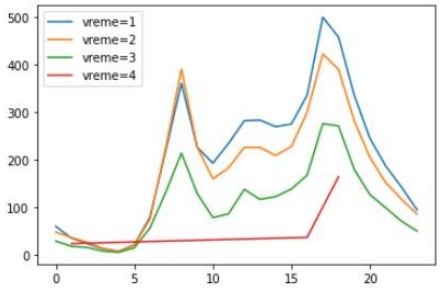
Posmatrajuci ***sliku 1*** mozemo uociti pozitivnu korelaciju izmedju temperature I broja iznajmljenih bicikala odnosno porastom temperature raste I broj iznajmljenih bicikala sto je I logicno jer ako je vreme lepse, vise ljudi ce zeleti da vozi. Na ***slici 2*** vidimo suprotan slucaj, porastom brzine vetra opada broj iznajmljenih bicikala sto takodje ima smisla jer jaci vetar ometa I smanjuje kvalitet voznje bicikla.



Slika 1. Uticaj temperature na broj iznajmljenih bicikala



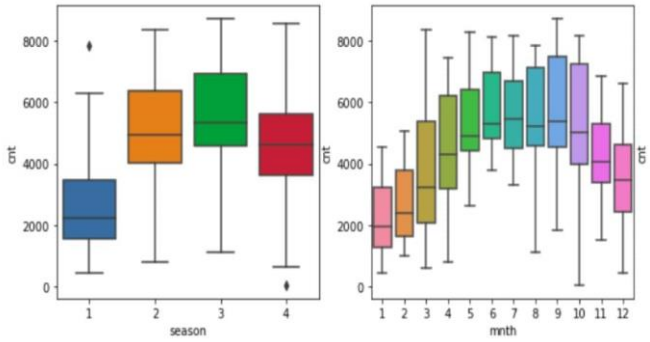
Slika 2. Uticaj brzine vetra na broj iznajmljenih bicikala



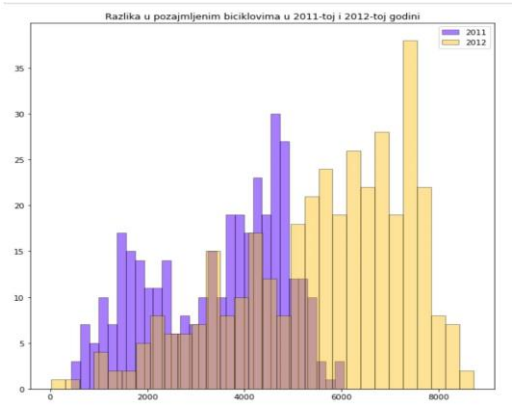
Slika 3. Prosecan broj iznajmljenih bicikala u periodu od 24 casa za razlicite vremenske uslove

Na ***slici 3*** vidimo kako se menja prosek iznajmljenih bicikala s ocenom kvaliteta vremena (1 je najbolja ocena, 4 najgora). Plavom bojom oznacena je kriva prosecnog broja iznajmljenih bicikala kada je vreme ocenjeno najbolje I tada je prosecno najvise biciklova iznajmljeno dok kada je vreme ocenjeno sa 4, prosek je ubedljivo najmanji.

Na ***slici 4*** vidimo da je potraznja bicikala najveca u jesen (mesec septembar), tada su I vremenski uslovi najbolji I imaju ocenu 1, takodje potraznja je visoka I na leto sto se pripisuje lepom I suncanom vremenu. Potrazivanje drasticno opada tokom zimskih meseci.

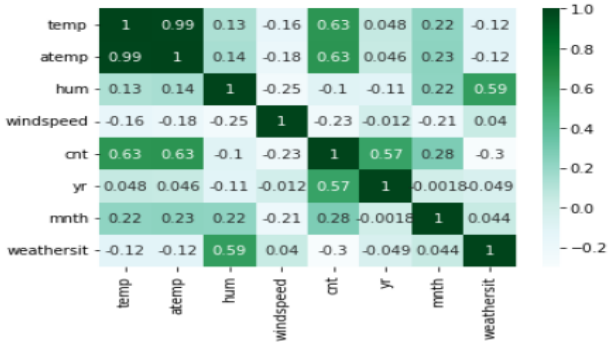


Slika 4 – Levo se nalazi broj iznajmljenih bicikala po godisnjim dobima dok je desno broj iznajmljenih bicikala po mesecima



Slika 5. Promena u broju iznajmljenih bicikala izmedju 2011-te I 2012-te godine

Sa ***slike 5*** uocljiva je promena u broju iznajmljenih bicikala u 2012-toj u odnosu na 2011-tu godinu. Jasno se moze videti da je dosta veci broj iznajmljenih bicikala u 2012-toj godini. Jedan od validnih zakljucaka moze biti da su gradjani svesniji kako upotreba fosilnih goriva unistava ekosistem I zagadjuje vazduh te su se opredelili za zdraviju I jeftiniju varijantu od voznje.



Slika 6. Matrica korelacije za pojedina obelezja

Na ***slici 6*** vidimo matricu korelacije za pojedina obelezja. Radi preglednosti, neka od obelezja cija je korelacija s obelezjem **cnt** a I ostalim obelezjima jako mala nisu prikazana. Kao sto smo I mogli zakljuciti na osnovu slike 1 obelezja **temp** I **atemp** imaju vecu pozitivnu korelaciju s obelezjem **cnt** od svih ostalih obelezja. Takodje, obelezje **yr** je pozitivno korelisano s **cnt**. Obelezja **temp** I **atemp** su medjusobno jako visoko korelisana sto je I logicno jer predstavljaju skoro jednu te istu stvar I retko mnogo odstupaju jedna od druge.

U narednom delu izvestaja bavicemo se predikcijom obelezja **cnt** I to koristeci tehnike linearne regresije, stabla odluke I neuralnih mreza.Navedene tehnike bice primenjene na oba dataseta.

IV Linearna Regresija

A. Linearna regresija na day datasetu

Podaci su podeljeni na skup za trening koji sadrzi 85% od ukupnog broja uzoraka I test skup koji sadrzi 15% od ukupnog broja uzoraka. Za svaki od algoritama za obucavanje modela uradjena je I kros validacija na trening skupu s parametrom **n_splits**=5 te izracunata srednja vrednost za svaku od ocena performansi I prikazana na narednim slikama. Test skup od 15% je ostavljen za krajnje testiranje modela. Za obucavanje prvog modela koriscena je klasa **LinearRegression**, obucen je na trening skupu I testiran na test skupu. Takodje, izvrsena je I standardizacija obelezja koja je doprinela malim poboljsanjima performansi.Njegove performanse mozemo videti na ***slici 7.***

#	result
mse	669960.963
mae	618.832
rmse	817.515
r2 score	0.803
r2 adj	0.798

Slika 7. LinearRegression

#	result
mse	753980.332
mae	546.490
rmse	841.447
r2 score	0.786
r2 adj	0.781

Slika 8. Hipoteza samo interakcija

#	result
mse	618821.932
mae	517.097
rmse	769.324
r2 score	0.824
r2 adj	0.820

Slika 9. Hipoteza interakcije I kvadrata

#	result
mse	621365.605
mae	505.393
rmse	768.253
r2 score	0.824
r2 adj	0.820

Slika 10. Lasso regresija

Na ***slikama 7,8,9 I 10*** vidimo performanse 4 razlicita modela linearne regresije. Za sada najbolje performanse ima Lasso regresija za koju je koriscen parameter alfa=0.01.Ima najmanje odstupanje predvidjenih od stvarnih vrednosti I najveci r2 score odnosno pokriva najvise udela ukupne varijanse. Primecuje se da svi modeli osim prvog (***slika 7***) daju slicne performanse, s malim razlikama. Medjutim, najbolje performanse je dala Ridge regresija cije su performanse predstavljene na ***slici 11***.

#	result
mse	470363.497
mae	496.728
rmse	684.757
r2 score	0.862
r2 adj	0.859

Slika 11. Ridge regresija

#	result
mse	503400.507
mae	520.300
rmse	709.507
r2 score	0.882
r2 adj	0.865

Slika 12. Ridge regresija na krajnjem test skupu

Ridge regresija po svim parametrima pobedjuje ostale modele pa je ona testirana I na krajnjem test skupu. Treba uzeti u obzir da ceo dataset **day** ima samo 730 uzoraka te sam test skup nije veliki i performanse modela dobijene nad njime treba uzeti s rezervom. Performanse finalnog modela mozemo videti na ***slici 12***. Za Ridge regresiju koriscen je parametar **alfa=13** koji je dobijen metodom **RidgeCV**. Takodje, pokusana je redukcija dimenzionalnosti metodom **PCA**, medjutim doprinela je samo smanjenju performansi pa o njoj nije diskutovano.

B. Linearna regresija na hour datasetu

Hour dataset sadrzi puno vise uzoraka nego **day** pa u ovom slucaju test skup sadrzi 10% od ukupnog broja uzoraka dok ostalih 90% pripada trening skupu.Sve sto je radjeno na **day** datasetu ponovo je radjeno na **hour** datasetu , medjutim modeli dobijeni na hour datasetu daju slabije performanse nego na **day** pa se o njima u ovom izvestaju nece posebno diskutovati.

V Stabla odluke

A. Stabla odluke na day datasetu

U ovom poglavlju ce biti diskutovano o modelima obucenim na bazi stabala odluke, konkretno **DecisionTreeRegressor** I **RandomForestRegressor.** Zbog potreba ovih algoritama kreirane su dummy varijable za obelezja **season**, **mnth**, **weathersit** I **weekday**. Podela je kao I do sada izvrsena u odnosu 15% test skup I 85% trening skup, takodje radjena je kros validacija s 5 particija.

#	result
mse	885320.779
mae	732.156
rmse	938.049
r2 score	0.741
r2 adj	0.725

Slika 13. DecisionTreeRegressor

#	result
mse	771151.957
mae	687.400
rmse	876.557
r2 score	0.775
r2 adj	0.761

Slika 14.RandomForestRegressor

Na ***slikama 13. I 14.*** vidimo performanse modela dobijenih metodama **DecisionTreeRegressor** I **RandomForestRegressor**. Nesto bolje performanse daje RandomForestRegressor sto je u neku ruku logicno jer slucajna suma koristi vise stabala odluke kako bi dosla do boljih rezultata. Optimalni parametri za ove 2 metode dobijeni su koriscenjem metode **GridSearchCV** kojoj su prosledjene razlicite kombinacije parametara i uz pomoc malo rucnih korekcija. Optimalni parametri za **DecisionTree** su:
max_depth=3,**criterion**='friedman_mse',**max_features**='auto',**max_leaf_nodes**=None,**min_samples_leaf**=1,**min_weight_fraction_leaf**=0.1,**splitter**='best', dok je za **RandomForest** dobijeno:
n_estimators=100,**max_depth**=5,**max_features**=None,**max_leaf_nodes**=50,**min_samples_leaf**=1,**min_weight_fraction_leaf**=0.1,**criterion**='friedman_mse'.

Takodje, pokusana je redukcija dimenzionalnosti **PCA** metodom ali nije dovela do poboljsanja performansi.

B. Stabla odluke na hour datasetu

Podela je izvrsena tako da test skup sadrzi 10% a trening 90% uzoraka. Kros validacija je radjena na 3 particije zbog velicine dataseta. Ostala predobrada podataka I korisceni algoritmi su isti kao I na stablima odluke nad **day** datasetom. Rezultati dobijeni stablima odluke nad **hour** datasetom vidno su bolji od onih nad **day** datasetom jer po svojoj prirodi stable odluke I slucajne sume bolje rade na vise uzoraka pa se to I u ovom slucaju ispostavilo kao tacno. Performanse modela mozemo videti na ***slikama 15. I 16***.

#	result
mse	3211.717
mae	33.951
rmse	56.667
r2 score	0.901
r2 adj	0.901

Slika 15. DecisionTreeRegressor(**hour**)

#	result
mse	1883.863
mae	26.447
rmse	43.382
r2 score	0.942
r2 adj	0.942

Slika 16. RandomForestRegressor(**hour**)

Vidimo da je r2 score kod obe metode dosta veci nego na **day** datasetu odnosno da pokrivaju veci udeo od ukupne varijanse u odnosu na prethodne modele. Takodje se vidi da je odstupanje od stvarnih vrednosti dosta malo.Za odredjivanje optimalnih parametara takodje je koriscena metoda **GridSearchCV** uz malo rucnog podesavanja. Za **DecisionTreeRegressor** dobijeni su parametri:
max_depth=12,**max_features**='auto',**max_leaf_nodes**=None, **min_samples_leaf**=1,**splitter**='best',**criterion**='friedman_mse' a za **RandomForestRegressor** n_estimators=350, **max_depth**=None,**criterion**='friedman_mse',**max_features**='auto',**max_leaf_nodes**=None,**min_samples_leaf**=1.

PCA redukcija dimenzionalnosti ni ovoga puta nije dovela do poboljsanja performansi, stavise znatno ih je smanjila dajuci **r2 score** 0.38. Posto je **RandomForestRegressor** na **hour** datasetu dao najbolje performanse on jer testiran I na test skupu koji smo ostavili za krajnje testiranje. Performanse ovog modela mozemo videti na ***slici 17.***

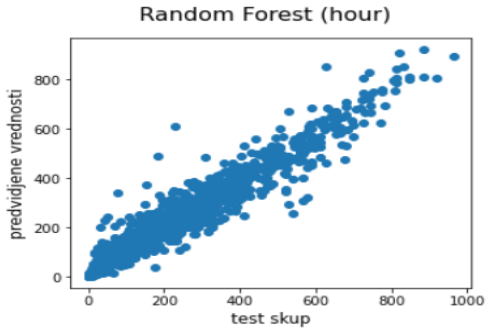
#	result
mse	1916.810
mae	26.339
rmse	43.781
r2 score	0.943
r2 adj	0.943

Slika 17. RandomForestRegressor

Primecujemo da su za **RandomForestRegressor** performanse slicne na kros validaciji I na krajnjem test skupu sto znaci da nas model dobro generalizuje I da je dobro obucen.

Na ***slici 18.*** mozemo videti odnos pravih I previdjenih vrednosti na krajnjem test skupu.Vidimo da su vrednosti jako slicne sto se I moze zakljuciti iz performansi finalnog modela

sa ***slike 17.*** I vidimo da bi se oni mogli lepo aproksimirati pravom.



Slika 18. Odnos pravih I previdjenih vrednosti u modelu dobijenom metodom slucajne sume

VI Neuralna mreza

U ovom poglavlju vice diskutovano u neuralnoj mrezi implementiranoj pomocu metode **MLPRegressor.** Test skup sadrzi 10% a trening 90% od ukupnog broja uzoraka. Takodje, I u ovom slucaju je radjena kros validacija sa 3 particije. Performanse neuralne mreze na datasetu **day** su se pokazale jako lose, sto je posledica manjka uzorka tog dataseta da bi se neuralna mreza adekvatno obucila.Stoga, u ovom poglavlju ce biti predstavljeni rezultati modela samo na datasetu **hour.**

Optimalni parametra za neuralnu mrezu dobijeni su opet metodom **GridSeachCV** I uz odredjenu dozu rucnih korekcija. Primeceno je da porastom broja skrivenih slojeva performanse modela se poboljsavaju dok ne dodje do 4 skrivena sloja gde performanse dostizu svoj peak.Takodje, porastom broja neurona u slojevima performanse se poboljsavaju.
Parametri **MLPRegressor-a** koji su korisceni su: **hidden_layer_sizes**=(300,300,300,300), **activation**='relu', **solver**='adam', **batch_size**=50, **learning_rate**='adaptive', **learning_rate_init**=0.001, **max_iter**=100, **shuffle**=True, **random_state**=42, **early_stopping**=True, **n_iter_no_change**=10,**validation_fraction**=0.1, **verbose**=False.

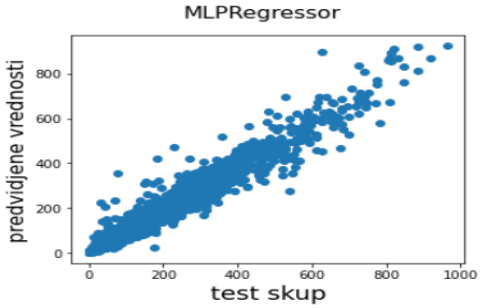
#	result
mse	1607.403
mae	24.888
rmse	40.061
r2 score	0.951
r2 adj	0.951

Slika 19. MLPRegressor

#	result
mse	1747.919
mae	26.413
rmse	41.808
r2 score	0.948
r2 adj	0.948

Slika 20. MLPRegressor(krajnji test skup)

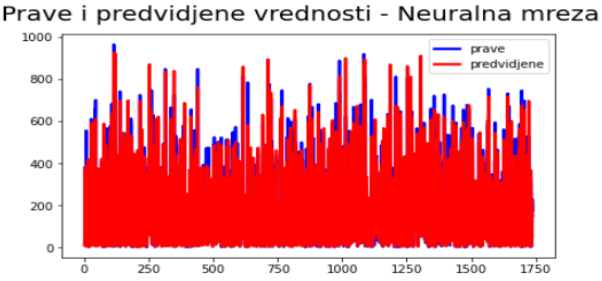
Pokusana je I redukcija dimenzionalnosti metodom **PCA** ali ni ovoga puta nije dala poboljsanja performansi, iako je ovaj put smanjenje performansi neznatno, dajuci **r2 score** 0.913. Rezultati dobijeni na kros validaciji I na konacnom test skupu se ne razlikuju puno sto znaci da model dobro generalizuje I da je dobro obucen.
Primecujemo da neuralna mreza daje najbolje performanse, bolje od svih ostalih algoritama. To moze biti posledica velikog broja adekvatnih parametara koriscenih za obucavanje iste uz dovoljan broj uzoraka da bi obuka bila uspesna. Prosledivsi joj dovoljan broj skrivenih slojeva I neurona mreza je sama uspela da nauci kompleksne veze izmedju obelezja cega su produkt najbolje perofmanse od svih modela.



Slika 21. Odnos pravih I previdjenih vrednosti u modelu dobijenom neuralnom mrezom

Na ***slici 21.*** Primecujemo istu stvar kao na ***slici 18.*** a to je da se prave I predvidjene vrednosti ne razlikuju puno I da se mogu aproksimirati pravom s tim da je rasipanje u ovom slucaju malo manje jer su performanse modela ipak malo bolje.

Na ***slici 22***. uocavamo veliko preklapanje izmedju predvidjenih I pravih vrednosti sto smo mogli I da zakljucimo u dosadasnjoj analizi.



Slika 22. Preklapanje pravih I predvidjenih vrednosti

VII Zakljucak

Na osnovu analize podataka iz treceg poglavlja mozemo zakljuciti da bi kompanija koja se bavi iznajmljivanjem bicikala trebalo da se fokusira na prosirivanje biznisa tokom letnjih I jesenjih meseci jer je potraznja tada najveca. Tokom zimskih meseci I loseg vremena bi trebalo vise da se fokusiraju na servisiranje I odrzavanje bicikala kako bi bili spremni onda kada ljudima najvise trebaju.Takodje, prognoziran je rast potraznje za iznajmljivanjem bicikala tokom godina sto moze biti u korelaciji s potencijalnim porastom ekoloske svesti kod ljudi.

Vidimo da metode **slucajne sume** I **neuralne mreze** daju vidno bolje performanse od linearne regresije ali isto tako zahtevaju veliki broj parametara cega je posledica puno vremena za njihovu obuku. Takodje, za neuralnu mrezu, konkretno **MLPRegressor** je potreban veci broj uzoraka da bi se obucila pa se nije mogao adekvatno iskoristiti na **day** datasetu. Stoga, mozemo zakljuciti da linearna regresija ce u ovom slucaju linearna regresija dati relativno solidan rezultat na datasetu manjeg broja uzoraka dok je za neuralnu mrezu I slucajne sume potreban veci broj uzoraka I vise vremena za obuku al ce dati znatno bolje performanse.