Analiza I formiranje modela za predviđanje nad podacima o iznajmljivanju bicikala

Aleksandra Borisavljević, IN17/2018, aleksandra.borisavljevic99@gmail.com
Vladimir Trpka, IN41/2018, vtrpka@gmail.com

I. Uvod

Tema izveštaja je analiza predviđanje broja iznajmljnenih bicikala u zavisnosti od vremenskih uslova koji su zabeleženi. Sistemi za iznajmljivanje su postali veoma popularni među populacijom koja živi u gradovima. Ovi servisi su takođe korisni zato što pružaju potpunu brigu što se tiče odlaganja i servisiranja. Meteorološki uslovi su u velikoj meri povezani sa brojem iznamljivanja i na osnovu toga se može napraviti koristan model koji bi mogao da vrši predviđanja. Njegova primena dovela bi do poboljšanja u svakodnevnom poslovanju.

II. Baza podataka

Uporedo se obrađuju dva skupa podataka, day.csv i hour.csv. U prvom skupu podataka jedan uzorak predstavlja izmerene vremenske prilike i broj iznajmljivanja u toku jednog dana, dok uzorak u drugom skupu podataka predstavlja izmerene iste vremenske uslove i broj iznajmljivanja, ali na nivou jednog sata. Day.csv sadrži 731 uzorak i 16 obeležja. Dok hour.csv sadrži 17 379 uzoraka i 17 obeležja. U oba skupa podataka nalaze se uzorci iz 2011. i 2012. godine. Međutim u hour.csv nedostaju uzorci koji su vezani za pojedine sate tokom dana.

Kategorička obeležja koja se javaljaju su: season (godišnje doba), yr (godina, 0 je 2011., 1 je 2012.), mnth (mesec), hooliday (da li je praznik), weekday (dan u nedelji), workingday (da li radni dan), weathersit (vremenski uslovi), hr (sati), dteday (datum). Dok su numerička obeležja: instant (redni broj uzorka), temp (temperatura u °C), atemp (lični osećaj temperature u °C), hum (vlažnost vazduha u %), windspeed (brzina vetra u mph), casual (broj neregistrovanih korisnika), registered (broj registrovanih korisnika) i cnt (broj korisnika).

III. Analiza podataka

Kao prvi korak analize skupa podataka, potrebno je odbaciti obeležja koja ne pružaju dodatnu informaciju o uzorku. Takvo obležje je *instatne*. Predstavlja redni broj uzorka, jedinstveno je za svaki uzorak i ne daje nam dodatnu informaciju, stoga je suvišno i potrebno ga je ukloniti iz skupa podataka.

A. Analiza null vrednosti

Nakon uklanjanja suvišnih kolona sledi analiza null vrednosti, kao izuzetno bitan segment svake analize podataka. Ispitivanjem se zaključilo da ne postoje null vrednosti koje su eksplicitno navedene. Međutim to ne znači da u skupu podataka ne postoje null vrednosti koje nisu implicitno navedene. Stoga je neophodno na druge načine proveriti njihovo postojanje. Jedan od načina je ispitivanjem nekih statističkih osobina, kao što su maksimalna i minimalna vrednost, medijana, srednja vrednost i mnoge druge.

Što se tiče obeležja za temperaturu, vlažnost vazduha i brzinu vetra, date su normalizovane vrednosti koje su dobijene tako što je svaka vrednost podeljena sa maksimalnom vrednošću za to obeležje.

Najveća vrednost za obeležje temperature iznosi 41 °C, dok je minimalna vrendost koja se pojavljuje 0.82 °C. Vrednositi za medijanu i srednju vrednost su veoma približne, što svakako ukazuje da su ove vrednosti validne, odnosno postojanje autlajera je odbačeno kao mogućnost. Dok nam vrednosti za minimum i maksimum govore da su temperature u ovom regionu veoma povoljne tokom cele godine. Takođe može se primetiti veoma slično ponašanje sa vrendostima za obeležje lični osećaj temperature.

Vlažnost vazduha je parametar koji značajno utiče na vremenske prilike, pa samim tim i na odluku ljudi da kao prevozno sredstvo iskoriste bicikle. Maksimalna zabležena vrednost za ovo obeležje je 100%, dok je minimalna vrednost 0%. Vlaznost vazduha i temperatura su međusobno zavisne. Povećanje temperature i jačine vetra izaziva mnogo veće isparenja iz zemljišta i drugih površina, a samim tim se povećava količina vodene pare u vazduhu. Pri ovakvim uslovima ljudima je naporno da obavljaju i najmanje zahtevne fizičke aktivnost, stoga se većina odlučuje da ih odloži za vreme kada će biti povoljniji vremensi uslovi.

Najveća izmerena vrednost za brzinu vetra je 57 mph (milje na sat), dok su zabeleženi i dani kada vetra nije bilo. Brzina vetra od 55 mph pa naviše predstavlja izuzetno snažne vetrove. Oni mogu dovesti do značajnih strukturalnih oštećenja, raznošenje krovova kuća, stabala itd. Svakao ovakvi uslovi značajno onemogućavaju vožnju bicikala. Stoga je očigledno da ovakvi uslovi izuzetno utiču na broj iznajmljivanja.

Posmatrajući vrednosti za medijunu i srednju vrednost za obeležje casual, primećena je značajna razlika koja je ukazala na postojanje autlajera, koji uzimaju veće vrednosti. Analizom se došlo do zaključka da su u pitanju

dani kada je vreme izrazito povoljno. Vreme koje karakteriše minimalna oblačnost, bez jakih vetrova i padavina. Uglavnom je u pitanju nedelja u poslepodnevnim časovima. Sve navedene činjenice slažu se sa tvrdnjom da ljudi tokom vikenda i lepog vremena žele da provedu svoje vreme u prirodi uz rekreaciju. Interesantno je primetit da postoje uzorci kada nije bilo iznajmljenih bicikala od strane neregistrovanih korisnika unutar jedanog časa. Analizome se utvrdilo da su u pitanju rani jutarnji časovi ili kasni večernji, što se poklapa sa svakodnevnim ljudskim funkcionisanjem.

Kod broja registrovanih korisnika može se uočiti takođe postojanje određene količine izrazito većih vrednosti. Međutim može se uočiti da se razlozi za te vrednosti razlikuju od prethodno navođenih. Može se primetiti da su to radni dani u nedelji, uglavno u ranim jutarnjim časovima, ili u vreme kraja radnog vremena. Ovo ukazuje da su to stalni korisnici koji većinom koriste usluge iznajmljivanja radi svojih svakodnevnih obaveza. Kada se uporede dani kada su najviše iznajmljivani bicikli od strane registrovanih i neregistovanih korisnika, vreme je u oba slučaja povoljno za vožnju, odnosno bez padavina kao što su kiša, sneg, magla itd. Što se tiče veoma niskih vrednosti koje se pojavljuju, isti je slučaj kao kod neregistrovanih korisnika.

B. Kategorička obeležja

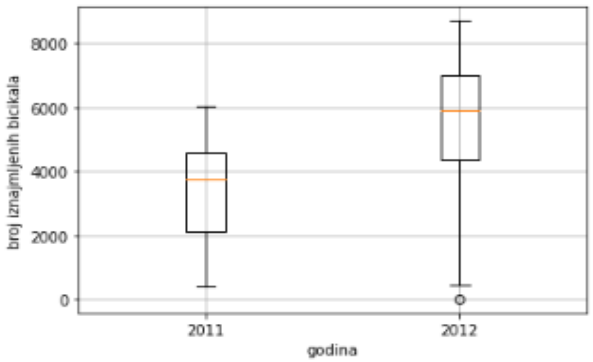
Sva kategorička obeležja predstavljena su numeričkim vrednostima. Stoga nije potrebna njihova konverzija i date vrednosti se mogu upotrebni.

IV. OPIS ZAVISNOSTI OBELEŽJA

Dobro uočavanje zavisnosti između obeležja kao i zavisnost broja iznajmljivanja od vremenskih uslova je nezaobilazan korak pravljenja uspešnog modela za predviđanje.

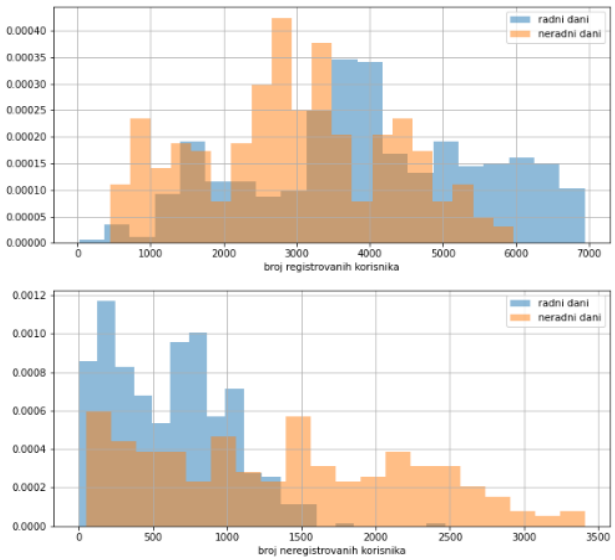
U zavisnosti od obeležja koje analiziramo, biće korišten jedan od dva skupa podataka, radi boljeg razumevanja tumačenja i dovedenih zaključaka.

Na samom početku detaljnije analize, radi boljeg razumevanja, predstavićemo kako se kretao broj iznajmljivanja toko 2011. i 2012 (Slika 1). godine. Može se uočiti značajniji porast u broju sveukupnog iznajmljivanja u 2012. godinu u odnosu na prethodnu. Može se uočiti da je 2011. godine zabeleženo 75% uzoraka koji su imali manje od 4700 rezervacija tokom jednog dana, dok je 2012. godine u 75% uzoraka zabeleženo 6900 i manje, broja iznajmljivanja u toku jednog dana. Pretpostavka je da je došlo do razvoja ovog biznisa kao i veće zainteresovanosti i potrebe ljudi koji žive u gradovima, za ovakvom vrstom prevoznog sredstva. Takođe može se uočiti da je svaki dan bilo zabeleženo iznajmljivanje bicikala. Ovo nam govori da su ovakvi servisi veoma koriste i da je njihova uloga veoma bitna. Međutim samo jedan dan zabeležen je broj inajmljivanja koji je iznosio 22. Obradom podataka došlo se do zaključka da je ovaj dan bio izuzetno nepovoljan za vožnju. Ovaj uzorak se odnosi na oktobarski dan sa mogućom pojavom kiše i grmljavine praćenu vetrom. Zabeležena je izuzetno velika relativna vlažnost vazduha, čak 88%.



Slika1: Broj iznajmljivanja u odnosu na godinu

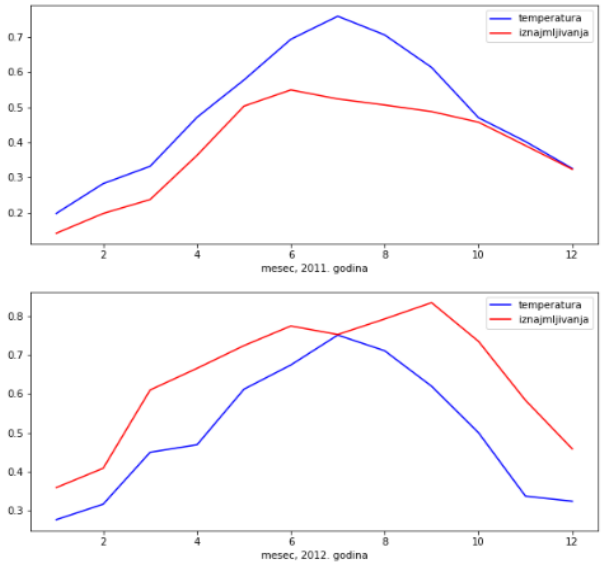
U zavisnosti da li je radni ili neradni dan u nedelji, razlikuje se broj iznajmljivanja od strane registrovanih i neregistrovanih korisnika (Slika 2). Kod registrovanih korisnika je broj iznajmljivanja veći tokom radnih dana, dok je kod kod neregistrovanih korisnika broj iznajmljvianja mnogo veći tokom neradnih dana. Registrovani korisnici svoje svakodnevne obaveze obavljaju korišćenjem iznajmljenih bicikala i upravo zato je nihova brojnost veća tokom radnih dana. Može se primetiti da se kod registrovanih korisnika ne razlikuju u velikoj meri broj iznajmljivanja tokom ranih i neradnih dana, koliko je to slučaj kod neregistrovanih korisnika. Kod njih je mnogo izraženija razlika. To svakako govori da i registrovani korisnici svoje slobodno vreme rado provode uz rekreaciju.



Slika 2: Zavisnost broja iznajmljivanja u odnosu na radne i neradne dane

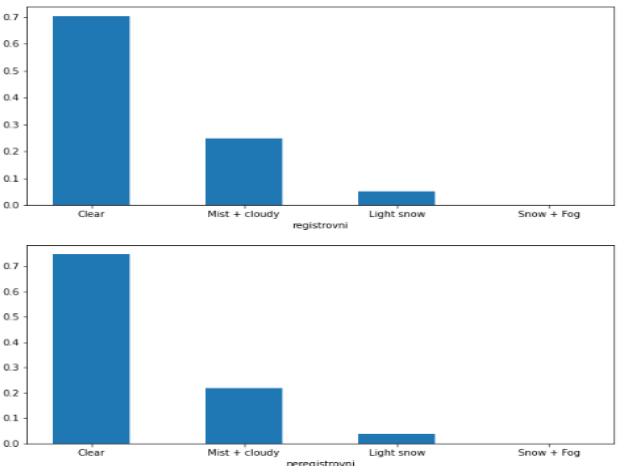
Kao što je prethodno komentarisano, postoji korelacija između tempereture i broja iznajmljivanja. Kako temperatura raste tokom godine broj iznajmljivanja takođe raste (Slika 3). Vidi se da je najveća temperatura zabeležena tokom letnjih meseci. Zimske mesece odlikuju manje temperature, a samim tim i izraženije manji broj iznajmljivanja. Može se primetiti da je najveći broj iznajmljivanja u 2011. godini bio u prolece, kada nisu

izmerene maksimalne temeprature. U 2012. godini najveći broj iznajljivanaj bio je krajem maja i u septembru, odnosno u proleće i jesen kada su temperature umerenije.



Slika 3: Odnos broja iznajmljivanja i temperature

Kada se posmatraju sveukupni vrmeneski uslovi (weathersit) mogu se uočiti razlike između registrovanih i neregistrovanih korisnika (Slika 4). Broj iznajmljivanja je srazmeran u odnosu na registrovane i neregistrovane, međutim uočavaju se minimalna odstupanja. Kada se posmatraju najpovoljniji vremenski uslovi može se uočiti da je tada procentualni broj iznajmljivanja bio veći od strane neregistrovanih korisnika, dok kada se posmatraju lošiji uslovi procentualni broj iznajmljivanja je veći kod registrovanih korisnika



Slika 4: Odnos između broja iznajmljivanja i sveukupnih vremenskih uslova

V. FORMIRANJE REGRESIONIH MODELA

Formiranje modela će se raditi upotrebom day.csv skupa podataka. Na samom početku neophodno je iz skupa uzoraka izbaciti obeležje koje treba da se predviđa, odnosno obeležje *cnt*. Zatim je neophodno podeliti skup podataka na skup za obuku i skup za testiranje modela. Skup za obuku dobio je 90% uzoraka.

A. *Linearna regresija*

Prvi model za regresiju koji je korišten je linearna regresija. U cilju pronalaženja modela koji daje najbolja predviđanja korištene su različite hipoteze. U nastavku (Tabela 1) je dat je pregled mera uspešnosti primenom različitih hipoteza. Može se primetiti poboljšanje kada se u hipotezu uključe zavisnosti između obeležja kao i povećanje stepena. Međutim osim promene hipoteze postoje i drugi načini kako mogu da se poboljšaju vrednosti, a da se očuva i ispravnost modela, odnosno da ne dođe do preobučavanja.

Neki od načina su: da se razmisli koja obeležja utiču na obeležje koje se predviđa, normalizacija obeležja, upotreba regularizatora. Regularizatori kažnjavaju pojave dominantnih obeležja.

TABELA 1: Mere uspešnosti linearne regresije

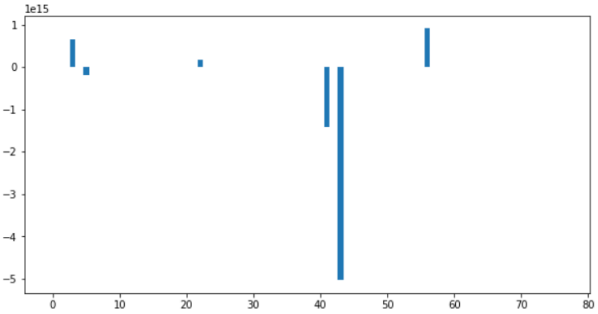
Hipoteza : y=b0+b1x1+b2x2+…	
Koren srednje kv. greške	807.98
R2 greška	0.85
y=b0+b1x1+b2x2+..c1x1x2+c2x1x3...	
Koren srednje kv. greške	660.92
R2 greška	0.90
y=b0+b1x1+...+bnxn+c1x1x2+...+d1x1^2+...+dnxn^2	
Koren srednje kv. greške	632.5
R2 greška	0.91
Ridge	
Koren srednje kv. greške	604.35
R2 greška	0.92
Lasso	
Koren srednje kv. greške	634.86
R2 greška	0.91

Kako bi mogli da primenimo neke od prethodno navedenih načina za poboljšanje modela, neophodno je da primenimo normalizaciju nad obeležjima. Primenom normalizacije može se doći brže do rešenja, ali ne obavezno i do boljeg.

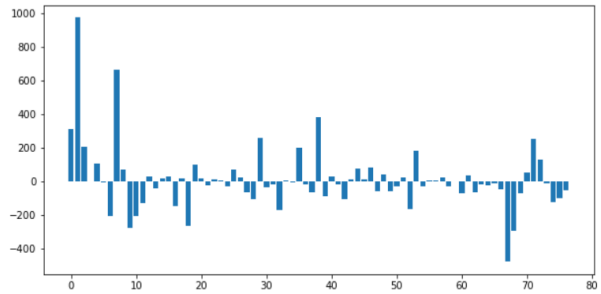
Primenom kompleksnijih hipoteza kod kojih su uključene zavisnosti i stepeni, vrednosti za mere uspešnosti su primetno poboljšane, međutim bitno je primetiti da su težinski koeficijenti neravnomerno raspoređeni (Slika 5). Postoje neka obeležja kojima se pridaje izrazito veća važnost, odnosno postoje dominantna obeležja. Ovakav problem rešava se primenom regularizatora.

Primenjene su dve vrste regulaziratora, Ridge i Lasso. Lasso regularizacija je specifičnija, odnosno pored uloge same regularizacije radi i selekciju. Ako se desi da je težina za neko obeležja jako blizu nuli, uvešće se pretpostavka da to obeležje ne utiče na predviđanje i izbaciće se iz dalje analize.

Analizirajuće sve dobijene regresione modele, zaključeno je da je Ridge regresor model koji daje najbolje rezulate, u vidu vrednosti mera uspešnosti, ali i ravnomerno raspoređenih težinskih koeficijenata (Slika 6).



Slika 5: Težinski koeficijenti pre primene Ridge regresije



Slika 6: Težinski koeficijenti nakon primene Ridge regresije

Zatim je upotrebljena unakrsna validacija nad Ridge regresorm. Kroz unakrsnu validaciju, ispitivan je parametar alfa, koji predstavlja stepen kažnjavanja dominantnih obeležja. Utvrđeno je da je najbolja vrednost za ovaj parametar 20. Prosečna vrednost za RMSE (koren srednje kvadratne greške) iznosi 776.16, dok je prosečna vrednost za R2 grešku 0.836.

B. *Knn regresija*

Treći model za regresiju koji je korišten je kkn regresor.. Primenom unakrsne validacije izvršena je evaluacija parametara. Parametri koje je moguće podešavati kako bi se dobio što bolji model su broj suseda koji se uzima u obzir i metrika koja se koristi za određivanje rastojanja.

Analiziran je broja suseda od 1 do 20 i euklidska i menhetn metrika (Tabela 2). Došlo se do zaključka da su najbolje vrednosti za mere evaluacije dobijene za menhetn metriku i broj suseda 5. Prosečna vrednost za RMSE meru je 980, dok za R2 grešku iznosi 0.738.

TABELA 2:Prosečne mere uspešnosti KNN regresije

Metrika i susedi	RMSE	R2
Manhttan, 3	1009.56	0.72
Manhttan, 5	980.6	0.74
Manhttan, 15	1088.21	0.67
Euclidean, 3	1028.39	0.71
Euclidean, 5	985.88	0.73
Euclidean, 15	1148.84	0.64

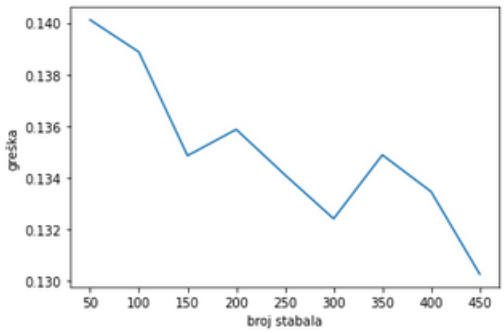
C. *Stablo odluke i slučajna šuma*

Treći model za regresiju koji je obrađen je stablo odluke i slučajna šuma. Kao prvi korak ovog postupka, unakrsnom validacijom se ispitivala optimalna dubina stabla odluke. Došlo se do zaključka da je najpovoljnija dubina među ispitivanim bila 7. Kao mera evaluacije korištene su srednja kvadratna greška i r2 score.

Prednost stabla odluke je jasniji prikaz kako je regresor pravio odluke i na osnovu čega je došao do zaključka. Međutim zbog rada na samo jednom stablo može doći do nadprilagođavanja. Kako bi se izbegao ovaj problem analiziran je metod slučajne šume.

Dalje se ispitivalo ponašanje vrednosti mera evaluacija za slučajnu šumu. Unakrsnom validacijom se pronalazio optimalan broj stabala unutar šume, dok je za dubinu uzeta prethodno utvrđana optimalna vrednost. Takođe ispitivanje optimalne brojnosti drveća u šumi vršeno je analizom greške na *out of bag* uzorcima (Slika 7). To su uzorci koji nisu uzeti u obzir prilikom obuke stabla u bootstrap postupku. U oba postupka utvrđeno je da je optimalan broj stabala u šumi 450.

Kao najbolji regresioni model u ovom slučaju odabrana je slučajna šuma sa parametrom dubine stabla 7 i sa brojem stabala u šumi 450, uz upotrebu bootstrap postupka. Vrednost za srednju kvadratnu grešku iznosi 683, dok je vrednost za meru r2 score 0.868.



Slika 7: Odnos broja stabala i out of bag greške

D. *Poređenje dobijenih regresionih modela*

Formirajući tri regresiona modela, mogu se uočiti razlike između njih. U nastavku (Tabela 3) mogu se videti razlike u merama uspešnosti. Najbolje vrednosti za obe mere daje model slučajne šume. Jedan od razloga zašto se ovaj metod pokazao kao najbolji je sama njegova kompleksnost. Slučajna šuma se sastoji od više stabala odluke. Svako stablo odluke predstavlja model za sebe. Kombinovajući veliki broj stabala odluke, odnosno modela, svakako može da se proizvede složenija i dublja analiza samih podataka.

TABELA 3: Mere uspešnosti 3 regresiona modela

	RMSE	R2
Linear regression	776	0.836
KNN	980	0.738
Desision tree	683	0.868