Predviđanje korištenja bicikala u gradskom sustavu posuđivanja bicikala

Neven Grubelić[1], Zvonimir Iveković[1]

Sažetak
Za gradski sustav posuđivanja bicikala radimo model koji će za ulazne parametre predviđati broj posuđivanja po satu. Ovdje koristimo stablo uvjetnog zaključivanja(conditional inference tree) kao bolju opciju od klasičnog stabla odlučivanja.

Ključne riječi
regresija, predviđanje, stablo odlučivanja

[1] *Matematički odsjek, PMF, Sveučilište u Zagrebu*

Sadržaj

1. Uvod

Sustavi posuđivanja bicikala postali su jako popularni, pogotovo u velikim metropolama gdje gust promet može dovesti do velikih gužvi. U takvim uvjetima čini se bolje kretati se biciklom ukoliko je udaljenost koju trebamo proći mala i krećemo se samo središtima grada. S obzirom da je sustav poprilično korišten te sadrži veliku količinu podataka o korištenju koji su opisani egzaktnim vrijednostima, kao što su duljina trajanja vožnje, prijeđena udaljenost, vrijeme posuđivanja bicikla, vrlo je zanimljiv za istraživanje mobilnosti unutar grada.

U ovom radu željeli smo, za dani skup podataka o broju posuđenih bicikala u sustavu te vremenske prilike u trenutku posuđivanja, odrediti broj bicikala koji će biti posuđen u određenom trenutku s obzirom na vremenske prilike.

2. Opis problema

Skup podataka koji smo koristili je dan u obliku *.csv* datoteka već unaprijed podijeljenih na trening set i testni set. Datoteka trening seta sadrži podatke o broju posuđivanja bicikala u svakom pojedinom satu u danima između 1. i 19. u mjesecu posljednjih dvije godine. Testni set sadrži istu vrstu podataka samo za period od 20. dana u mjesecu pa sve do kraja mjeseca.

Svaki unos(redak) u bazi podataka je sadržavao slijedeći skup podataka:

- sat i datum sa vremenskim žigom
- godišnje doba
- oznaku je li danas praznik
- oznaku je li danas radni dan
- vremenske prilike označene na slijedeći način:
  - 1: vedro, malo oblaka, djelomično oblačno
  - 2: maglovito, maglovito i djelomično oblačno
  - 3: jako oblačno, lagana kiša, grmljavinsko nevrijeme, susnježica
  - 4: gusta magla, magla i snijeg, tuča, obilna kiša
- temperatura u Celzijusima
- ”osjetilna”(”feels like”) temperatura
- vlažnost zraka
- brzina vjetra
- broj neregistriranih korisnika koji su unajmili bicikl
- broj registriranih korisnika koji su unajmili bicikl
- ukupan broj unajmljivanja bicikala u tom satu

Bitno je naglasiti da smo kao mjeru uspješnosti procjene našeg modela uzeli ”*Root Mean Squared Logarithmic Error* (RMSLE)” koji se računa na slijedeći način:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1)-log(a_i+1))^2} \tag{1}$$

pri čemu je:

- n je broj sati u testnom skupu
- p_i je predviđen broj posuđivanja
- a_i je stvarni broj posuđivanja
- $log(x)$ je prirodni logaritam

Dakako, cilj je bio minimizirati gornju vrijednost za dani model.

3. Korištena metoda i dobiveni rezultati

Cijeli smo projekt radili u **R**-u. Trening skup nam se sastojao od 10886 unosa, dok je testni skup bio veličine 6493 unosa. Što se tiče obrade podataka, napravili smo analizu featurea, te tako dobili, smatramo, preciznije predviđanje od onog koje bismo dobili sirovim podacima. Korištena je faktorizacija atributa za obradu nenumeričkih vrijednosti. Tako smo, npr., atribut *datetime*, koji sadrži datum i sat, rastavili na dva atributa te oba faktorizirali. Nadalje, procijenili smo da nam sami datumi nisu toliko bitni, te da bismo korisniji predviditelj dobili grupirajući ih u skupine po danu u tjednu. Pri tome smo dobili da nedjelja iskače kao outlier(vidi sliku 1), pa smo smatrali korisnim dodati novu varijablu koja bi mogla biti koristan predviditelj.

```
> aggregate(train_factor[,"count"],list(train_factor$day),mean)
     Group.1        x
1    četvrtak 197.2962
2    nedjelja 180.8398
3       petak 197.8443
4 ponedjeljak 190.3907
5     srijeda 188.4113
6      subota 196.6654
7      utorak 189.7238
>
```

Slika 1 - Prosječan broj posuđivanja po danu u tjednu

Zbog prirode podataka (malo nezavisnih varijabli), smatrali smo da bi kao model moglo poslužiti stablo odlučivanja. U **R**-u postoji više načina na koji se može doći do takvog modela. Mi smo odabrali funkciju **ctree** iz paketa **party** zbog načina na koji bira varijable koje će koristiti za gradnju stabla odlučivanja. Naime, ta funkcija koristi framework za testiranje permutacija da bi sprovela statističku analizu na kovarijacijskim vrijednostima varijabli, te tako odabire varijable najpogodnije za izgradnju modela, za razliku od drugih metoda koje imaju inicijalnu pristranost u odabiru varijabli, kao npr. metoda **rpart**().

4. Zaključak

Naš je model imao testnu vrijednost **0.4952**. Valja napomenuti da smo analizom dobivenih podataka došli do zaključka da se varijabla *hour*, dobivena analizom iz vrijednosti varijable *datetime*, pokazala kao najbitniji predviditelj za model, što opravdava samu analizu.

Model smo usporedili s onim dobivenim pomoću funkcije rpart().Takav model ne koristi statističku analizu za dobivanje najpogodnijih varijabli, nego pokušava sagraditi stablo koje ima što više dijeljenja (”splits”). Takav je model imao testnu vrijednost oko 0.6. zaključili smo, stoga, da je inicijalna pristranost u našem slučaju utjecala na točnost ovog modela.

S obzirom na jednostavnost modela, smatramo da je točnost zadovoljavajuća.