NYU STERN

Samuel Forman <srf366@stern.nyu.edu>

Fwd: Invitation: Review Citi Bike logistic regression @ Wed Apr 4, 2018 4:30pm - 5pm (EDT) (jlk635@nyu.edu)

Jon Kastelan <jlk635@nyu.edu> Wed, Apr 4, 2018 at 6:16 PM
To: Samuel Forman <srf366@stern.nyu.edu>

Sam,

Good to sync up and see the missing bike logistic classifier you're building.

I'm attaching a couple of notebooks which could provide some ideas on how you may want to visualize the classifier, along with accuracy and precision. As discussed, the odds calculated as exp(coeff) are a good way to describe:

- Bikes rented in ZIP ZZZZZ are XX times more likely to go missing than ZIP 10001
- Bikes rented as part of an annual membership are AA less likely to go missing, compared with bikes rented on a 1-day pass

The notebook on called Session_07_practice is the Logistic classifier details. Let me know if you need any of the references datasets, I can send them through also.

The confusion matrix is another output I recommend putting together to provide a summary of model fits. It compares Actual values (missing bikes, and not missing) against model predicted values (predicted missing, predicted not missing). It's the True positive, False positive, True negative, False negative analysis.

I mentioned transforming the variables, and if you're interested in building a clear boundary between bikes which go missing vs. regular bikes, you may consider a technique like Principle Component Analysis. It is used to reduce the number of dimensions considered for analysis (but the flipside is, it's not as interpretable as analysis on the untransformed variables). I'm attaching an example of the PCA, in the Homework 4 notebook (questions 4 and 5).

I'll send a note to Alex also, and copy you in.

Have a good weekend, and we'll see you next week.

Cheers,
Jon

[Quoted text hidden]

2 attachments

Session_07_practice.ipynb
681K

Homework_4_jlk635.ipynb
859K