Three Potential Models for Predicting Ride Rating

Will Jones

November 24, 2015

We have a set of data comprised of bicycles rides and binary ratings of those rides. We are seeking to find a model to predict ride rating using predictors that operate on different levels of observation. For example, weather conditions are ride specific, but the presence of a bike lane is specific to the different road segments travelled on during a ride.

Because we have data on different levels of observation—rides, riders, and segments—it will be appropriate to use a multilevel model.

For rides, the variables we have are:

- length (distance)
- time
- traffic
- weather

For segments, the variables we have are:

- bike lane / type of infrastructure
- grade
- road type (main street versus residential street)

For riders, we don't have specific variable right now, though it would be useful to consider their effects as random rather than fixed. Even with all these other variables being equal, we could hardly expect a rider to give the same rating for every ride on a route.

Furthermore, it will be worthwhile to explore differences in how riders rate rides: Do some simply tend to rate higher than others? Are some riders more sensitive to traffic conditions or weather conditions?

I will outline here a few potential models we could use for predicting ride ratings.

The simple model

For a very simple model, we can simplify or ignore information from the segments altogether. Here, we would consider the level of observation to be rides, with random effects from riders. Some segment data could still be used on the level of rides; for example, we could include a variable indicating whether the rider did an unprotected left-hand turn.

So, for a simple example, we would have a dataset of rides, where for the ith ride we have:

- rating $r_i \in \{0, 1\}$
- rider $j \in 1, \ldots, K$
- length $l_i \in \mathbb{R}_{\geq 0}$
- during rush hour indicator $t_i \in \{0, 1\}$
- raining indicator r_i
- temperature indicator T_i

The we can use a logistic regression model where

$$\mathbb{P}(r_i = 1) = \text{logit}^{-1}\left(\alpha_{j[i]} + \beta^{\text{length}} \cdot l_i + \beta^{\text{traffic}} \cdot t_i + \beta^{\text{raining}} \cdot r_i + \beta^{\text{temp}} \cdot T_i\right)$$

where the rider random effects are

$$\alpha_j \sim N(\mu_\alpha, \sigma_j^2).$$

The segments and riders model

We could improve the simple model by adding in information about the segments. Unlike rider observations, we actually have some data about segments we can use as predictors.

Here is one example of a model we could use here. Let $\mathcal{R}$ be the set of road segments in the road network, with $|\mathcal{R}| = S$, and for the kth road segment we have:

- indicator for bikelane, b_k
- indicator for no car road, c_k
- road grade, g_k

Then the contribution from road segment k to a riders rating can be written

$$\alpha_k^{\text{road}} \sim N(\beta_k^{\text{bike.lane}} \cdot b_k + \beta_k^{\text{no.car}} \cdot c_k + \beta_k^{\text{grade}} \cdot g_k, \sigma^2_{\text{segment}})$$

For the ith ride, there is an S-dimensional vector Ω_i where the kth component, ω_{ik}, is 1 if the k segment was a part of the ride, and 0 otherwise. Then the total contribution of the segments to the rating, α_i^{path}, is

$$\alpha_i^{\text{path}} = \sum_{k=0}^{S} \omega_{ik} \cdot \alpha_k^{\text{road}}.$$

Finally, we add this path term to the simple model:

The we can use a logistic regression model where

$$\mathbb{P}(r_i = 1) = \text{logit}^{-1}\left(\alpha_{j[i]}^{\text{rider}} + \alpha_i^{\text{path}} + \beta^{\text{length}} \cdot l_i + \beta^{\text{traffic}} \cdot t_i + \beta^{\text{raining}} \cdot r_i + \beta^{\text{temp}} \cdot T_i\right)$$

where the rider random effects are

$$\alpha_j^{\text{rider}} \sim N(\mu_\alpha, \sigma_j^2).$$

The rider experience model

The previous is model is potentially fruitful because is incoporates data about road segments and variation in and between riders. However, it assumes that the variation in rider rating is independent of both ride and segment level data.

This may not be the case: some riders may be more comfortable riding in traffic than others, and some may find cold temperatures less unpleasant than others.

Thus, we might consider a level in our model we might call 'rider experience' (i.e. the experience the rider had on the segments.) This might be done by having rider specific coefficients some some variables such as weather or traffic.

Here, we might consider there to be a rider vector of coefficients for certain variables. For example, say some exploratory analysis found that the effect of bike lanes and of temperature varied greatly between riders. Then we would want rider specific coefficients for those variables.