User/Customer Engagement and Sales Conversion in a Social Media Context

Lynd Bacon[1], Danielle Murray[2], Peter Lenk[3]
[1]Loma Buena Associates, P.O. Box 620960, Woodside CA 94062 USA
[2]Shutterfly, Inc., 2800 Bridge Parkway, Redwood City, CA 94065 USA
[3]Stephen M. Ross School of Business, University of Michigan,
701 Tappan St., Ann Arbor MI USA

Abstract
An important challenge for brands and marketers is how to get, hold, and grow the attention of customers. This is particularly true when commerce depends on customers' social media platform activities. Switching costs for new users may be low, and brand messaging can be mitigated by the flows of information between customers and across platforms. We describe a hierarchical in parameters, latent trait structural equation model of engagement on a social media site. It scales a hierarchy of platform activities, and allows the prediction of behaviors in terms of escalating engagement. Parameter estimation is by way of MCMC procedures. We apply the model to a firm's customer activity and purchase data, and relate engagement and attitudinal measures collected from a sample of customers.

Key Words: customer engagement, CRM, latent trait, marketing, Hierarchical Bayes, social media.

1. Background

User engagement with web sites and social media platforms is an issue of interest to any company whose business model depends on the duration and nature of user interaction. "Engagement" is a fuzzy concept, and also usually a site or platform design goal. Measures of engagement are typically based on behavioral data. Those in use range from relying on a single, manifest measure like the number of unique visits to a site or the total number of minutes users spend on a site, to multiple indicator measures based combinations of summaries of behavior measures.

The objective of our application is to develop and deploy a multiple indicator, latent variable methodology that provides a measure or measures predictive of sales conversions that is as simple as possible, that scales well with respect to data on millions of users. Our measure need not be computable in real time, an advantage that allows using informative procedures at the cost of some computational intensity.

2. Modeling Approach

The variety of alternative methods we might use for this project is large. It ranges from psychometric procedures like confirmatory factor analysis to machine intelligence methods like Bayes nets.

We chose to apply a type of model developed in the area of standardized educational testing called an "Item Response Theory" (IRT) model(Sijtsma, & Junker. 2006; Wainer, Bradlow & Wang, 2008). IRT models have been used in marketing research(e.g., Bacon & Lenk, 2008; Bacon, Lenk & Durall, 2004; Balasubramanian & Kamakura, 1989; De Jong, Steenkamp, Fox & Baumgartner, 2008; De Jong, Steenkamp & Veldkamp, 2009;

Kamakura & Balasubramanian, 1989), and also in political science research(Clinton, Jackman & Rivers, 2004). For the present application, IRT models have several desirable features, including the following:

1. One or more latent variables, or "traits," can be measured with binary or ordinal observed measures, "indicators," or "items."
2. The items measuring a latent variable can be considered to represent a hierarchy of some sort, such as difficulty or involvement.
3. People (cases, or subjects), have locations on the latent variables.
4. The items have parameters describing their relationship to the latent variables.
5. If need be, their computation as Hierarchical Bayes models can be made relatively scalable using methods already developed in educational testing, or by applying approximations like particle filtering (Ridgeway & Madigan, 2002), or closed form approximations to distributions like the logistic (Miller, Bradlow & Dayaratna, 2006).

Simple IRT models can also be easily extended into what might be called a latent trait structural equation model (Bacon et al. 2004; De Jong et al., 2008; Fox & Glas 2001), a feature that we take advantage of for the application at hand.

A very common form of the basic IRT model is the two-parameter model:

$$p(\text{response}_{\text{ik}}=\text{'correct'}|\theta_i,\alpha_k,\beta_k)=\frac{1}{1+\exp\left[-\beta_k(\theta_i-\alpha_k)\right]},$$

where:

$$\begin{aligned}\theta_i &= \text{person i's score on the latent variable,}\\ \beta_k &= \text{item k's discrimination parameter,}\\ \alpha_k &= \text{item k's difficulty parameter}\end{aligned}$$

In this model, the probability of a "correct" response by person i on item k is a function of the person's location on the latent variable, θ_i, and the item parameters α_k and β_k. α_k is often called the "difficulty" parameter, and β_k the "discrimination" parameter. Another common form is the Rasch, or one-parameter, model in which β_k is fixed equal to 1.0. IRT models with ordinal items have α_k's for the cut-points between response categories. Model identification is accomplished by setting the scale of θ or by constraining item parameters. The normal CDF is often used instead of the logistic.

Model items are often summarized by plotting their "item characteristic curves," or ICCs. An ICC plots the probability of a positive response on an item as a function of the item parameters. Figure 1 gives example ICCs for three items in a two-parameter model. The latent variable is on the abscissa, and the probability of a positive response, P(α_k, β_k), is on the ordinate. Items with larger positive values of β_k more accurately discriminate between people with high and low scores on the the latent variable. Scores for two people, A and B, are shown in Figure 1 for illustrative purposes. You can see by comparing the locations of A and B that B is more likely to provide a positive response on all three items than A is.

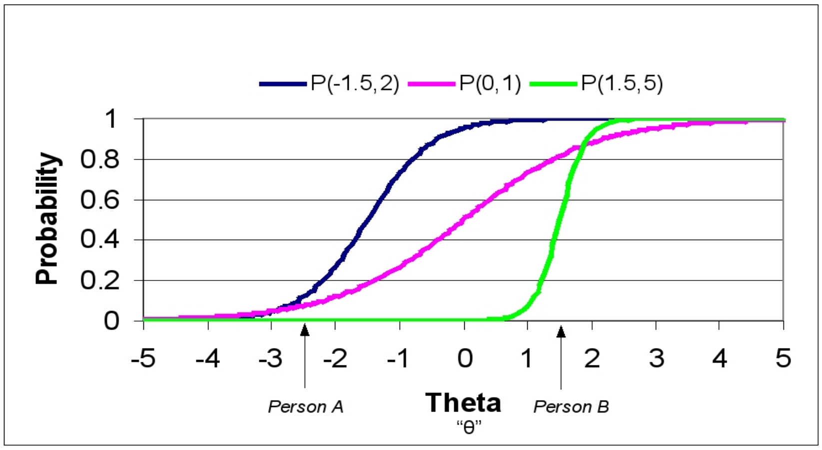

Figure 1. Item characteristic curves (ICC's)

1. The Data

Our application data are from an on line community site providing free content storage and sharing. The community owner sells various products that site users can create from their content. The data consist of 150 site activity count, or frequency, measures on 100,000 users selected randomly from the owner's data warehouse. The measures are proprietary, and so they and the other variables in use for this project cannot be explicitly defined here. But they capture user behaviors like the number of content uploads, the number of content shares, and the number of product purchases in various categories. For a subset of 1,893 of these 100,000 users, we also had a comprehensive set of attitude and usage measures collected in an on line survey. It was to this subsample that we ultimately applied the models of interest.

4. Creating Behavioral Items

From the 150 site activity count measures we selected a set of 30 that seem to span a range of difficulty or effort for users, and that seemed like they could be fallible indicators of user engagement. By "difficulty" we mean the extent to which effort or some other resource would be required in order to engage in a particular behavior. For example, we considered logging in to be a relatively easy activity, while sharing organized content with specific other site users would be harder.

The users in our data set had been registered to use the site for periods as short as a few weeks up to almost 10 years. To take varying tenures into account, we converted our count measures to rates per 30 days. We then created 30 binary variables from these rate measures by first calculating the overall median rate across the 30 rate variables. Then we scored each case's response on a binary item as positive when its rate on the corresponding rate variable exceeded the overall median rate, and as a negative, otherwise.

5. Item Selection, Model Specification and Estimation

We used a subsample of 2,000 cases exclusive of all 1,893 cases with survey data to select a subset of the 30 binary items that would be a good compromise between reliability and minimizing the number of binary items to be computed. To do this we fit unidimensional IRT models to item subsets using a Gibbs sampling, and by employing data augmentation(Albert & Chib, 1993). The parameters of all of these models were estimated by running 10,000 burn-in iterations, followed by 50,000 iterations and retaining every 50th sampled value.

The specification for these models was:

$$y_{i,j} = \beta_{1j}\theta_i - \beta_{0j} + \varepsilon_{i,j}$$

$$\varepsilon_{i,j} \sim N(0,1)$$

$$\theta_i \sim N(\mu, \lambda)$$

$$\beta_j \sim N(\beta_{0,} \Phi)$$

$y_{i,j}$, above, is a continuous latent variable that is greater than or equal to 0 when the observed response by person i on item j is positive, and is otherwise zero. β_{1j} is item j's discrimination parameter. β_{oj}/β_{1j} is item j's difficulty parameter as defined earlier. The error term $\varepsilon_{i,j}$ has a standard normal prior distribution. The person scores θ_i have a univariate normal prior with mean μ and precision λ, and the two parameters for each item, vector of length 2 $\beta_{j,}$ have a bivariate normal prior with mean β_0 and precision matrix Φ. Identification was by way of constraining the scale of the θ_i and by using proper priors. The $y_{i,j}$ were sample by drawing from a truncated normal distribution.

Based on the set of IRT models we fit we selected nine binary items that all reliably discriminated person scores on the latent variable in the sense that the expected value of their discrimination parameters was greater than zero. They also spanned a wide range of item difficulty, and were ordered in terms of difficulty in a logical manner given the user behaviors they reflected. We then used these nine items as indicators in IRT models with covariates. The specification for these models was the same as for the models described above, except that the person scores were expressed as a weighted linear combination of the covariates:

$$\theta_i \sim N(\Gamma Z, \lambda) \quad ,$$

where Γ is a matrix of regression coefficients for which we used a multivariate normal prior, Z is a covariates matrix, and λ is the prior precision of the θ_i.

We applied this IRT with covariates model to the data for our subsample of 1,893 cases with survey responses, setting aside 20% of the cases for assessing out of sample prediction accuracy. We used the estimation procedure described previously to fit versions of the model that differed in terms of the number of covariates included. One model included all 78 covariates we had available, and one had no covariates at all, i.e., it was a simple two-parameter IRT model. A third had only the eight covariates for which the 95% highest probability density intervals of their coefficients did not include zero when they were in the model with all 78 covariates. Not too surprisingly, the latter model performed well compared to the other two. The log Bayes factor for preferring it over the simple IRT model was 39.3. The log Bayes factor for preferring it over the model with all 78 covariates was -1.05, a value that is generally considered to be small and not meaningful. This third model predicted the binary responses of the held out sample with 90% accuracy.

Figure 2 provides plots of the ICC's for the eight covariate model. These ICC's were computed using the mean values of the sampled and thinned difficulty and discrimination parameter estimate chains.

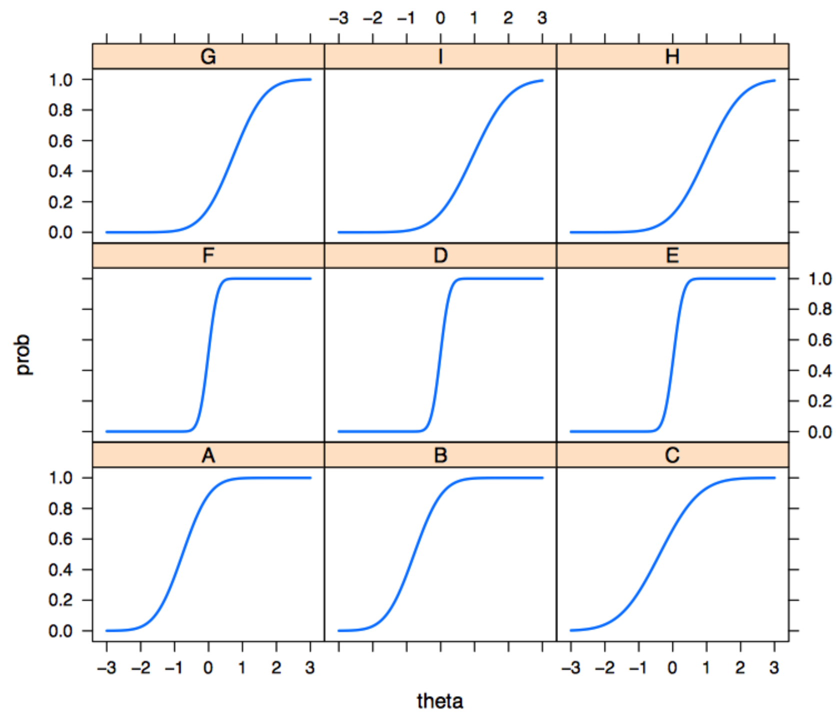

Figure 2: Item characteristic curves for the nine binary items A through H. The panels are main effects ordered from lowest difficulty at the lowest left to highest difficulty at the upper right. In each panel, values on the latent variable on on the abscissa, and the predicted probability of a positive response is on the ordinate.

The nine items in Figure 2 are labeled A through H. The individual ICC graphs in this figure are main effects ordered based on increasing value of their item difficulty parameters. So, for example, more positive responses would be expected for item A than for item C, and more would be expected for C than for H. It would appear from these ICC's that the number of items might be reduced as some have very similar difficulties. A few items were retained in this model despite their redundancy because of their substantive value and meaning.

Not surprisingly, the larger the number of positive responses for a person, the larger their person score. This is illustrated in Figure 3, where each person's θ_i is estimated by the mean of their thinned, sampled values.

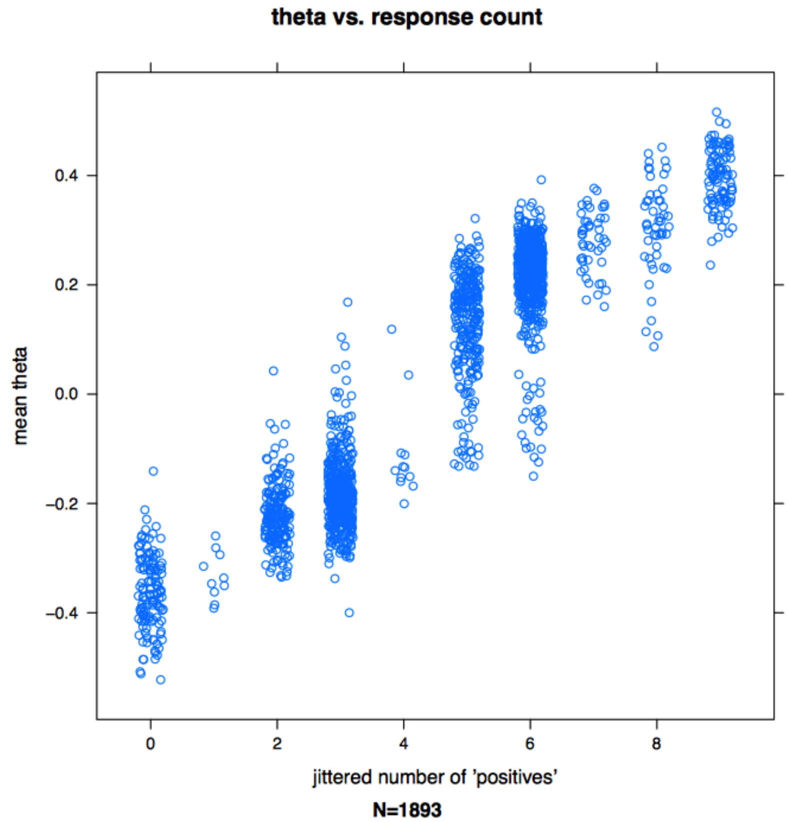

Figure 3: Person scores for the eight covariate model as a function of number of observed positive responses on the binary items. Each score is estimated as the mean of the person's thinned, post-burnin chain of sample values. Scores are on the ordinate, and the total number of positive responses are on the abscissa. The data points have been jittered to make the data density easier to appreciate. The cases from both the estimate sample and the held-out sample are included here.

The eight covariates we observed to predict the person scores reflected attitudes about the creation and use of user-generated content on the site. How each of the covariates related to the latent variable was sensible given the characteristics of the users and the purpose and features of the site. Descriptive analyses indicated that the latent variable was positively related to the rate of purchasing products in some categories and not in others, and also to tenure as a registered site user. The latter relationship would be expected to occur if facility with the site increases with experience, if time is required to amass enough user-generated content on the site to make use of it in a variety of ways, or if it takes a while for a user to accrue a large enough network of other users she knows and would like to share content with. Additional data is required to distinguish between these alternative hypotheses, of course.

4. Summary and Next Steps

Our decidedly simple application of a common univariate IRT model with covariates demonstrates that IRT methodology is a promising alternative for summarizing and understanding user behavior on web sites on which user generated content is shared. We're planning on pursing several model extensions and refinements as this project moves forward. These include incorporating ordinal indicators, developing and estimating a multi-period, time series form of IRT, and finally, scoring the site owner's customer data base of millions of users.

References

Albert, J. & Chib, S. (1993) "Bayesian analysis of binary and polychotomous response data." ***Journal of the American Statistical Association***, 88(422), 669-679.

Bacon, L. & Lenk, P. (2008) “Breaking the binary code: IRT models offer flexible marketing measurement.” ***Marketing Research Magazine,*** Winter, 7-10.

Bacon, L., Lenk, P. & Durall, J. (2004) "Item response theory (IRT) models: Basics, and marketing applications." ***Sawtooth Software Conference Proceedings***, 187-206, Sequim WA: Sawtooth Software, Inc.

Balasubramanian, S. & Kamakura, W. (1989) "Measuring consumer attitudes towards the market place with tailored interviews." ***Journal of Marketing Research***, XXVI, August, 311-326.

Clinton, J., Jackman, S. & Rivers, D. (2004) "The statistical analysis of roll call data." ***American Political Science Review***, 98(2), 1-16.

De Jong, M., Steenkamp, J., Fox, J. & Baumgartner, H. (2008) "Using item response theory to measure extreme response style in marketing research: A global investigation." ***Journal of Marketing Research***, XLV, February, 104-115.

De Jong, M. Steenkamp, J. & Veldkamp, B. (2009) "A model for the construction of country-specific yet internationally comparable short-form marketing scales." ***Marketing Science***, 28(4), 674-689.

Miller, S. Bradlow, E. & Dayaratna, K. (2006) "Closed-form Bayesian inferences for the logit model via polynomial expansions." ***Quantitative marketing & Economics***, 4, 173-206.

Ridgeway, G. & Madigan, D. (2002) "Bayesian analysis of massive datasets via particle filters." ***SIGKDD '02 Proceedings***, Association for Computing Machinery, 1-9.

Sijtsma, K. & Junker, B.W.(2006). "Item response theory: past performance, present developments, and future expectations." ***Behaviormetrika***, 33, 75-102.

Wainer, H., Bradlow, E. & Wang, X. ***Testlet Response Theory and Its Applications.*** New York: Cambridge University Press, 2007.