The Impact of Marketing-Induced vs. Word-of-Mouth Customer Acquisition on Customer Equity Growth

Julian Villanueva, Shijin Yoo and Dominique M. Hanssens

Web Appendix

In this Web Appendix we show that, for our particular illustration, login behavior is highly correlated with purchase behavior. As explained earlier, our data come from a firm that initially provided the service for free. During the 70 weeks of our observation period, customers were not charged and did not know the firm was to change this policy later. Two weeks after the end of our observation period the firm announced by email to all of its customers that they had two months to switch to a basic service for a monthly fee, or they would have their service terminated. We obtained data on the customers who either terminated or remained after this trial period. Our objective is to study the relationship between the login activity of each customer when the service was free, and the probability of continuing after a change in the business model to a fee-based service. We have individual data on login behavior and some demographic information for the customers that registered during this 70 week period. We build a binary logit model, where the dependent variable is whether or not the customer switched. That is,

$$switch = \begin{cases} 1 \; if \; customer \; pays \\ 0 \; if \; customer \; abandons \end{cases}$$

We include in the model the following covariates: (1) *number of logins:* total binary logins during the first 20 weeks of a relationship. Since we observe customers joining the firm at different points in time, we decided to investigate login behavior during the first 20 weeks of the relationship with each particular customer[1]. This is a sufficient period of time to capture the level of use and interest in the service from a particular customer. Hence, we study the login behavior of customers that registered from week 1 to week 50 of the observation period. (2) *time trend:* week in which the customer registered. This is to test whether customers who joined earlier have a higher probability of switching to the fee-based service. (3) *retailer:* =1 if retailer, =0 otherwise. Most of the firm's customers are small companies trying to advertise or even sell through the Internet. Retailers are the most common business type and the main target of the firm, as it thought the service value would be higher for that group than for other businesses types. (4) *US based:* =1 if the customer is a US-based company, =0 otherwise. Although most of the firm's customers are based on the US, some were from abroad. By creating this dummy variable we test whether US-based firms have a higher probability of converting to the new business model. (5) *number of employees*. This is a potential predictor to identify those that will derive more value from this service and thus ultimately pay for it. The firm expected this service to be more suited to small firms, because of its simplicity. In addition, the service was difficult to customize for businesses requiring a sophisticated website. Therefore, the firm expected large firms to be less interested.

[1] If total logins for each customer during this 70-week period were used, we would certainly have more logins for those that registered earlier. Even using average weekly logins would be problematic since data exploration shows that a large percentage of customers only log in the first week and never return. In those cases, for instance, we could observe a customer registering at week 70 that would have an average weekly login of 1, even though she might never come back again.

We first estimate a binary logit model with the total population: 93,119 customers registering from week 1 to week 50 of the observation period. The conversion rates for this particular company were extremely low. Of these customers, only 1,030 actually stayed with the company paying for the service they used to receive for free. This makes the occurrence of ones in the *switch* variable an extremely rare event. This posits a problem when estimating a logit model in that the model would predict everyone to abandon the fee-based service, and obtain a 99% level of accuracy. This is what our results indicate. Nevertheless, we find all estimates to be statistically significant, and the *number of logins* has a positive impact on the probability of switching (see Table 2 in the manuscript). Therefore, we can conclude that individual logins are positively correlated with future probability of retention.

To overcome the problem of misclassification, we re?-estimate the model while deliberately under-sampling the defectors, so that we obtain a more balanced sample of ones and zeros in the *switch* variable. This sampling technique is called choice-based sampling (see Ben-Akiva & Lerman 1985 for a comprehensive discussion). The problem of using such model is that the intercept is not consistent if traditional maximum likelihood estimation is used. Two alternative solutions to overcome this problem have been suggested in the literature. Manski and Lerman (1977) develop a weighted endogenous sampling maximum likelihood (WESML) estimator, which accounts for the different weights in the zeros and ones from the population of interest. Nevertheless, this estimator has the undesirable property of increasing the standard errors of the estimates (Manski and Lerman 1977, Greene 2000). A second approach, which we follow, is to adjust the estimated intercepts for each alternative by subtracting from the exogenous

maximum likelihood estimates of the intercept the constant $\ln(S_g / P_g)$, where S_g is the percentage of observations for alternative g in the sample, and P_g is the percentage of observations for alternative g in the population[2] (Manski and Lerman 1977).

The estimation results using the choice-based sample are reported in Table 2 in the paper. The model correctly classifies 90.8% of defectors and 86.5% of paying customers. The average predicted probability of switching to the fee-based service for our choice-based sample is 0.475, which is very close to the observed 0.484 of the sample. Using the revised intercept, the predicted average probability of retaining customers for the population is 0.0107, which is also very close to the observed value of 0.0111 (1,030 out of 93,119). Therefore, the revised estimation results predict well for the population even if they were estimated using a choice-based sample in which the zeros were heavily under-sampled.

The results show a significant and positive effect of login activity of a customer on their subsequent purchase behavior. The more a customer logs in, the higher the probability of switching to the fee-based service. Therefore, for this particular company, we conclude that acquisition channels that have a higher increase in the login activity of the firm's customers will also increase the subsequent average conversion rates.

[2] Hence, for our particular estimation results, where we find an estimated intercept of –5.120, we have to revise this intercept through the following steps. We have $S_1 = 1,030 / 2,130 = 0.4836$, and $P_1 = 1,030 / 93,119 = 0.0111$. Thus we have to subtract from the estimated intercept $\ln(S_1 / P_1) = 3.77$. Similarly, we have $S_0 = 1,100 / 2,130 = 0.5164$ and $P_0 = 92,089 / 93,119 = 0.9889$. Therefore, we should subtract from the intercept of alternative 0 the constant $\ln(S_0 / P_0) = -0.6497$. The estimated new constants will be -5.12 - 3.77 = -8.90 for alternative 1, and 0 – (-0.65) = 0.65 for alternative 0. Finally, since we want to keep alternative 0 normalized to be 0, we should add the constant -0.65 to both alternatives. The resulting revised intercept will be -9.55. For an example implementing this approach for the Multinomial Logit Model see Ben-Akiva and Lerman (1985, p. 238).

There are other factors that influence the probability of switching to the fee-based service. Customers who registered earlier, retailers, US-based firms and firms with fewer employees are more likely to purchase the service. However, these demographic variables are found to add only 0.3 and 2.1 percentage points for correctly classified defectors and paying customers, respectively. In fact, a logit model in which number of logins and an intercept are the only covariates correctly classifies 90.5% of defectors and 84.5% of future buyers. Additionally, a model with only demographic variables as covariates correctly classifies 75.9% of the defectors and only 59.0% of the future buyers. Thus, we conclude that login activity provides researchers with sufficient information to be linked with subsequent probability of adopting the fee-based service.

REFERENCES

Ben-Akiva, Moshe, and Steven R. Lerman (1985), *Discrete Choice Models*. MIT Press.

Greene, William H. *Econometric Analysis*. 4th ed ed. Upper Saddle River, N.J: Prentice Hall, 2000.

Manski, Charles, and S. Lerman (1977), "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica,* 45, 1977-88.