Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Customer Targeting Framework: Scalable Repeat Purchase Scoring Algorithm for Large Databases

Biswajit Pal [1], Ritwik Sinha [1], Abhishek Saha [1], Peter Jaumann [2] and Subhasish Misra[1] [+]

[1] Global Analytics

[2] Corporate Marketing-Customer Intelligence

Abstract. A major question in database marketing is that of identifying the customers who are most likely to make a repeat purchase in the near future. This information helps us to preferentially target these customers in any marketing campaign. The likelihood of a customer to make a repeat purchase depends on two questions: "Has the customer churned?" and "What is the frequency of transactions of the customer?" In this paper we develop a methodology using a Bayesian analysis to answer these questions; we further design a scalable regression method that approximates the Bayesian model. Using the answer to these questions as inputs we predicted the likelihood of a customer making a transaction within a time span into the future (e.g. in the next six months). Our algorithm can score massive databases for repeat purchase in real time. The proposed algorithm was tested on a recent Back to School campaign run by consumer exchange and has yielded promising results.

Keywords: Repeat purchase, Bayesian analysis, Regression, Real time scoring, Targeted marketing

1. Introduction

In the context of targeted marketing to consumers, the ability to tell which customers are more likely than others to make a purchase with HP in the near future greatly enhances effectiveness of any marketing campaign. It helps to rank customers on their propensity to re-purchase, and leads to preferential treatment of the right customers. It also reduces the likelihood of bombarding customers, who are less likely to purchase, with marketing material (over email or postal mail), possibly alienating them from future interest in HP. The propensity to make a repeat purchase depends on two parameters unique to each customer, the probability of churn and the frequency of transactions. The availability of rich customer databases, storing a wealth of information about the customer (transactional, demographic, psychographic), means that predictive models can be built to predict future purchase. However, most of those techniques are not easily extensible or readily applicable. Moreover, almost all these methods are not amenable on big transaction data.

The customer repeat purchase modeling framework we propose, based on a regression based approximation to a Bayesian hierarchical model, answers an important question for marketers in a scalable fashion. Further, since it uses only transaction data, it is readily applicable to a wide array of customer segments across different business units. Our repeat purchase framework is a unified version of two data mining techniques, a Bayesian Hierarchical Model (BHM) and Regression based approximation of BHM applied on the transaction data available for customers. The algorithm provides the probability of a customer repeating in next k periods, where k may be any unit of time (e.g. 30 days, 6 months or 1 year).

2. Our Solution

2.1. Bayesian Hierarchical Model(B.H.M)

[+] Corresponding author. Tel.: + 91-9986745023.
E-mail address: subhasish.misra@hp.com.

In this step a Bayesian Hierarchical model was built with the parameters below. Prior distributions were chosen using historical data.

Number of transaction made by the j^{th} customer follows a Poisson process with rate parameter λ_j [1]

The probability of dropping out after each purchase as Binary with probability p_j

The underlying idea behind the analysis is shown in figure 1

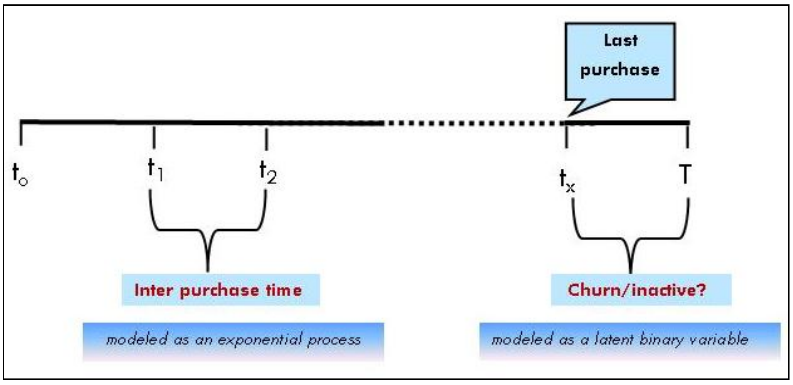
Last purchase
t₀
t₁
t₂
tₓ
T
Inter purchase time
modeled as an exponential process
Churn/inactive?
modeled as a latent binary variable

Figure 1: Modeling the frequency of transactions

The sufficient data used for sampling the conditional distributions in the MCMC chain for a customer are first transaction date (t_1), last transaction date (t_x) and the number of transactions (x). All variables are readily obtained from the transaction history of customers. No demographic/psychographic variables were used in this model- and hence this model is quite efficient from the data sufficiency perspective. We used the medians of the posterior distributions, sampled by the MCMC algorithm, as estimates of λ_j and p_j [3]. Then, the probability of customer j making a purchase in the next k periods is given by

$$(1- p_j)(1-exp\{-k\lambda_j\}) \qquad [i]$$

While the BHM was effective in correctly identifying customers who are likely to repeat, it is not efficient when scoring millions of customers (each customer requiring hundreds of iterations of the MCMC). It is common for an organization like HP to deal with millions of customers; for example, individual consumers, micro-businesses or small and medium businesses, customers on Snapfish.

2.2. Regression based Approximation.

The computational complexity of the BHM means that even though the results are promising, they are not ready to be operationalized. We next propose a fast scalable regression based approximation to this model. Note that the Bayesian model depends only on three variables for each customer, t_1, t_x and x. Further, any good scoring algorithm will have the property that two customers with the same values of t_1, t_x and x will have very similar scores. Because of these two properties, any sufficiently parameterized regression model will capture all pattern including non-linear effects and interactions.

We build two polynomial regression models for the parameter estimates of λ_j, p_j. To accomplish this, we first estimated λ_j, p_j for 3 million customers using the Bayesian Model and then used the logarithms of estimates as response in two polynomial regression models with predictors' t_1, t_x and x (and some interaction variables). To guide our choice of the degree of polynomial fit we explored the non-parametric fit from a Generalized Additive Model or GAM [2] (where the response is modeled as a sum of spline functions of individual predictors). A polynomial fit is also faster than a GAM, and lends easily to scoring large databases.

Once the estimated values of λ_j and p_j for every individual are obtained from the regression model they are fed into equation [i] to obtain the probability of each customer making at least a purchase in next k time period.

3. Evidence that the Solution Works.

To validate the utility of our algorithm, we formulated different hypothesis revolving around the power of predicting repeat customers, speed with which it can be executed in big databases and the ability to provide solution for multiple marketing problems across various domains. The two main hypotheses we considered were: (1) is our model doing better than random targeting (2) is our algorithm fast so that it can be applied on large data bases that different business units of HP possesses.

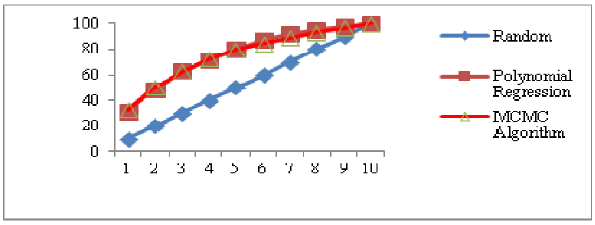

Fig.2:Effectivness of the models

Figure2 highlights that our models are giving considerably better results than random targeting and also the fact that regression approximation doesn't lead to any reduction in performance. The diagonal (baseline) denotes the case of random targeting (i.e. if 10% or 1 decile of customers are targeted one can expect to reach about 10% of the repeat purchasers). The curve above the diagonal shows the percentage of purchasers the model will successfully capture if we target the top (in terms of purchase probability) X deciles of customers. The top 4 groups here capture 75% of the buyers. This could be of strategic importance, i.e. when implementing a campaign, targeting the top 4 deciles captures a vast majority of repeat purchasers.

Models	Time taken for obtaining repeat purchase score for 1 MM customer
Regression based Approximation model	10 minutes
Bayesian Hierarchical Model	1800 minutes

Table. 1: Efficiency metric

Table1 indicates that how efficient the Regression based approximation model is over BHM. Regression method which was derived using the output of Bayesian method is 180 times faster over BHM.

The repeat algorithm provided a very satisfactory result when back tested on the 2011 Back to School (BTS) campaign. The top 10% of customers (based on the ranks from our model), captured customers who contribute 37% ($2.5 MM) of the total revenue in the BTS campaign. This says that, the model when used for customer selection identifies the most valuable customers. It further helps to reduce the cost of a campaign and the percentage of incorrect targeting which may lead to a bad customer experience.

4. Competitive Approaches.

Logistic regression can model events such as purchase/no purchase in a given period (e.g. 1 year) and has often been used to answer some of the questions discussed above. However, for each value of *k* a different regression needs to be built and validated. Sometimes, recency-frequency-monetary (RFM) analysis also is used to segment customers according to their transaction history, by rating their loyalty or value. This is a rather crude method, without any modeling assumption and or much predictive value

5. Current Status and Next Steps.

Churn is an important part of any customer intelligence knowledgebase and because of the non-contractual relationship between HP and its customers, is very difficult to predict. Presently we are in a process of validating the churn coming out as an output of the model. If comes out significant, will be a huge value adds in understanding customer lifecycle.

Also, we are working on building models that utilize the Bayesian modeling approach to predict the value and volume of future transactions. This will help us to infer product affinity for individual customers as well as cross-sale patterns for different products. This, when done and integrated with the approach discussed in this paper should prove useful for a marketer in inferring the customer lifetime value of any individual in his database.

6. Legal Disclaimer

“This document is published only for information purpose. The document is provided without any representation and in no event shall HP will be liable with the use of the information made available in the document. The information has been provided on an "as is" basis without warranty of any kind. The document is made available for information purpose only, further copying, reproduction or redistribution of the document is expressly prohibited. In no event shall HP be liable for lost profits or any direct, indirect, special, punitive, incidental or consequential damages arising out of or in usage of the document. HP is registered trademark of Hewlett Packard Company and its subsidiary”.

7. Acknowledgements

We would like to acknowledge the help from our managers – Ahmar Abdullah, Manav Shroff , Gitali Halder and Prasanna Dhore in helping us with the project and the paper.

8. References

[1] Peter.S.Fader , Bruce Hardie , Ka Lok Lee. “Counting your customers” the Easy Way: An alternative to the Pareto/NBD model. *Marketing Science*. Vol. 24:No. 2, Spring 2007, pp. 275-284.

[2] Dong Xiang. Fitting Generalized Additive Models with GAM procedure. *SAS Institute Inc*.

[3] Jayanta Kumar Pal, Abhisek Saha, Subhasish Misra. Customer repeat purchase modeling- A Bayesian Hierarchical Framework. *HP Labs technical report*. No 85, July 2010.