DOI:10.26524/jms.2016.2

Web mining for e-commerce websites using repeat-purchase models

Mihir Dash

Alliance University, Chikkahagade Cross, Anekal, Bangalore-562106, Karnataka, India

ABSRACT: e-retailing and database marketing are two emerging industries that require strong support from the CRM system. It is important for a website to keep its customers interested and come back frequently to visit. As web data and direct marketing data are available in huge volumes, data mining is an important and popular tool for both industries to develop good CRM systems to target loyal customers. Since most of this data is primary purchasing data, one could even go one step further to develop models to describe and predict behaviour of customers.

In this study two statistical models from the theory of repeat purchase behaviour were used to analyse customer loyalty. The models were able to predict the percentage of repeat-customers, and were able to identify marketing variables which affect the repeat-rate.

Keywords: Type your keywords here, seprated by semicolons;

INTRODUCTION

Loyal customers are those who make repeated purchases of a product/service. Studies have consistently shown that it is loyal customers that add the greatest value to a business and contribute the highest profits, particularly because it costs less to retain an existing customer than to recruit a new customer. This is captured in the concept of customer lifetime value (CLTV). Thus, targeting loyal customers based on customers' personal characteristics and purchase behaviour has become an important issue to be tackled by marketing managers.

Using data mining techniques to mine loyal customers is a logical conclusion to answer to this question, since there is a huge amount of data relating to the customers' opinions, perceptions, purchase behaviour, and day-to-day transactions available to managers. This information is extractable from the web-log file, which tracks every browse, visit, and transaction on the web, including the pages, topics, keywords, and sections the customer selects over time, together with some personal information obtained during the registration.

In particular, it is important for website marketers to keep customers interested and come back frequently to visit. As web data and direct marketing data are available in huge volumes, data mining is an important and popular tool for both industries to develop good CRM systems to target loyal customers. Since most of this data are genuine purchasing data, one could even go one step further to develop models to describe and predict behaviour of customers.

In this study, two statistical models from the theory of repeat buying are used to analyse the data from an e-commerce database. The models are used to identify differences in behaviour due to demographics, allowing the website marketer to segment customers according to demographic groups and formulate marketing strategies for different segments accordingly.

1. MODELS OF REPEAT PURCHASE BEHAVIOUR

The seminal work of Ehrenberg [1] has opened up the field of repeat purchase theory as a subfield of consumer behaviour. The theory applies particularly well to fast moving consumer goods (FMCG).

In particular, two models of repeat purchase behaviour are of relevance in the present context: the negative binomial distribution (NBD) model and the logarithmic series distribution (LSD) model. The negative binomial distribution is used to model purchase behaviour for a heterogeneous population, where the mean number of purchases varies across the population, while the logarithmic series distribution is used as a simplified version of the negative binomial distribution, when the customers who have made no purchase during the study period are removed [2].

According to the negative binomial distribution model, the proportion of customers who purchase *r* units (in one period) is given by the expression:

$$p(r;k,m) = \frac{\Gamma(k+r)}{\Gamma(k).r!}\left(\frac{k}{k+m}\right)^k\left(\frac{m}{k+m}\right)^r \text{, } r = 0, 1, 2, 3, \ldots$$

In particular, the parameters m and k determine the salient characteristics of the distribution: the proportion of non-buyers is given by $p(0;k,m) = \frac{1}{(1+m/k)^k}$, the mean number of units purchased is given by $E(r) = m$, and the variance of the number of units purchased is given by $V(r) = m.(1+m/k)$.

The negative binomial distribution implicitly assumes stationarity of customer behaviour over time, and can be used as a predictive tool in this case. On the other hand, if stationarity fails, the negative binomial distribution can be used to detect and compare changes that have taken place over time [2].

In empirical applications, the NBD is fit to the data using fitting by means and zeros: the parameter m is estimated by the sample mean of purchases (in one period), i.e. $\hat{m} = \bar{x}$, while the parameter k is estimated numerically using the sample proportion of non-buyers, i.e. numerically solving the equation: $1-\hat{p} = \frac{1}{(1+\bar{x}/k)^k}$ [3].

In steady-state, the NBD yields some very useful predictions, including:

- the proportion of buyers in time period T: $b_T = 1 - \frac{1}{(1+Tm/k)^k}$
- the proportion of new buyers: $b_N = \frac{1}{(1+m/k)^k} - \frac{1}{(1+2m/k)^k}$
- the proportion of repeat buyers: $b_R = 1 - \frac{2}{(1+m/k)^k} + \frac{1}{(1+2m/k)^k}$
- the mean amount purchased by new buyers: $m_N = \frac{m}{(1+m/k)^{k+1}}$
- the mean amount purchased by repeat buyers: $m_R = m\left[1 - \frac{1}{(1+m/k)^{k+1}}\right]$.

According to the logarithmic series distribution model, the proportion of customers who visit *x* times (in one period) is given by the expression:

$$p(X = x) = \frac{-\theta^x}{x \log_e(1-\theta)}, x = 1, 2, 3, \ldots$$

where the parameter θ is a fixed positive real number, $0 < \theta < 1$, representing the proportion of visits from repeat buyers. In particular, the mean value is given by the expression $E(X) = \frac{-\theta}{(1-\theta).\log_e(1-\theta)}$.

In empirical applications, the LSD is fit to the data using fitting by moments: by equating the mean value above to the sample mean $\bar{x}$ [3].

As in the case of the NBD, the LSD yields some useful predictions, including:

- the proportion of visits by repeat visitors: $1 + \frac{\log_e(1+\theta)}{\log_e(1-\theta)}$
- the mean number of visits by new subscribers: $\frac{\theta}{\log_e(1+\theta)}$
- the proportion of sessions accounted for by repeat visitors visiting at least *r* times: θ^{r-1}.

2. DATA & METHODOLOGY

2.1 Normal or Body Text

The data used for the study was collected from the website of an online job search portal in 2012. The most time-consuming part of mining the website involved the capture, extraction, aggregation and preparation of server log data for analysis. This involved three types of data files: the web log file, subscriber profiles, and job profiles.

The web log file was the largest data file. This file contains records of every transaction between server and browsers, with the date and time, including the IP address of the server making the request for each page, the status of the request, the number of bytes transferred to the requester, and so on. Each day, on average, there could be hundreds of megabytes of data collected from the site. The second data file was the subscriber profile. This file contains personal information such as age, gender, type of present job, and so no for all the subscribers. Finally, the third data file was the job profile, containing information on job title, job type and job industry, qualification, and so on.

The subscribers selected for the analysis were the job seekers who were registered with the portal. The subscribers were grouped according to specialized fields. The sample consisted of three groups of job seekers: group A represented information technology (IT), information technology-enabled services (ITeS), and systems candidates, group B represented marketing and sales candidates, and group C represented operations-related candidates. A visit represents the event that a subscriber applied for a particular job profile on the portal. The focus of the analysis was on the distribution of the number of visits, both overall and in different groups. The sample distribution of the number of visits is summarized in the Table 1.

Table 1. Frequency Distribution of the Number of Visits in Different Groups

	number of subscribers			
number of visits	**group A**	**group B**	**group C**	**overall**
0	2137	2853	3557	8547
1	752	623	611	1986
2	348	247	228	823
3	195	107	55	357
4	107	68	17	192
5	61	37	19	117
6	31	27	9	67
7	19	12	1	32
8	9	11	1	21
9	6	4	2	12
10	2	1	0	3
Total	**3667**	**3990**	**4500**	**12157**

Source: primary data

3. ANALYSIS & INTERPRETATION

The results of fitting the NBD and LSD models for the overall sample and for each group are shown in Table 2 below.

Table 2. NBD and LSD Model Fitting for the Sample Distributions

NBD model				
	group A	**group B**	**group C**	**overall**
k	0.5929	0.3417	0.3676	0.3870
m	0.8811	0.5702	0.3293	0.5748
p_0	0.5828	0.7150	0.7904	0.7031
$\chi^2 cal$	2.3608	9.5206	20.1269	10.0526
p-value	0.8837	0.1463	0.0005	0.2614
LSD model				
	group A	**group B**	**group C**	**overall**
θ	0.7398	0.7155	0.5704	0.6994
$\chi^2 cal$	36.5040	5.8845	23.3826	24.6574
p-value	0.0000	0.4362	0.0007	0.0009

Source: author’s calculation

The NBD model was found to fit the overall sample distribution and the sample distributions for groups A and B, but not for group C. The lack of fit for group C could be attributed to the high rate of non-visit (79.04%) vis-à-vis the tail behaviour of the number of visits by those who did visit within this group: the tail frequencies were between 50% to 75% higher than would have been expected at the observed rate of non-visit (i.e. lower than expected kurtosis).

On the other hand, the LSD model was found to fit neither the overall sample distribution, nor the sample distributions for groups A and C, but did fit the sample distribution for group B. As before, the lack of fit for the sample distribution of group C could be attributed to lower than expected kurtosis, whereas the lack of fit for the sample distribution of group A and the overall sample distribution could be attributed to higher than expected kurtosis (i.e. peaking behaviour near the mean, and thinner than expected tails).

In view of the lack of fit of the LSD model, the analysis was subsequently focused on the NBD model. The predictions from the NBD model for the overall sample and for each group are shown in Table 3.

Table 3. Predictions from the NBD Model for the Sample Distributions

NBD model				
	group A	**group B**	**group C**	**overall**
proportion of non-visitors	58.28%	71.50%	79.04%	70.31%
proportion of new visitors	14.14%	10.94%	10.48%	11.66%
proportion of repeat visitors	27.59%	17.56%	10.47%	18.04%
average number of visits by new visitors	0.2065	0.1528	0.1373	0.1626
average number of visits by repeat visitors	0.6746	0.4174	0.1920	0.4122

Source: author's calculations

It was found that group A had a lower proportion of non-visitors (i.e. a higher proportion of visitors), a higher expected proportion of new visitors, a higher expected proportion of repeat visitors, and a higher expected number of visits by both new visitors and by repeat visitors than groups B and C had. This means that subscribers in group A have a higher tendency to apply for jobs on the portal than groups B and C.

A question that naturally arises from the preceding analysis is as to why this should be the case. The high attrition rates in IT and ITeS profiles as against that of other profiles in the period is an obvious factor, but the more pertinent question that arises is that of what attracted the group A candidates to revisit the portal, and why groups B and C were not sufficiently attracted to the portal.

In order to formulate appropriate strategies to address the differences observed between the groups of subscribers, web marketers must get a deeper understanding of the different subscribers' needs and expectations from the portal. Strategies to address specific segments of subscribers can then be developed appropriately [4,5].

4. CONCLUSIONS

Data mining has become very popular tool due to huge amount of data produced, collected and stored every day in all walks of life. The huge volumes of data involved in web mining and direct marketing certainly warrant the use of data mining. But one should not miss a more important issue: data collected from the web and direct marketing provide a much more accurate and detailed factual information about their customers. In marketing, there is already a full spectrum of literature on consumer behaviour but many of them rely on interview or panel data, which could have data quality problem as compared to web data.

This kind of study can also be applied to other areas, such as selecting appropriate mode of transport facility between any two given places in a given interval of time. It can also be used to suggest a better model to optmise the resource utilization and benefits for the organization, with greater customer satisfaction, and thereby increasing customer loyalty. This approach can be extended to identify the quality of service of any service-oriented industry.

An extension of the study is to investigate how marketing variables affect the parameters *m*, *k* and *p* of the NBD and the θ of the LSD. The findings could have significant financial implications to the companies involved in e-commerce and direct marketing.

5. REFERENCES

[1] Ehrenberg, A.S.C. 1988. Repeat Buying 2nd Ed. New York: Oxford University Press

[2] Lilien, G.L., Kotler, P. and Moorthy, K.S. 2003. Marketing Models. New Delhi: Prentice-Hall India

[3] Wani, J.K. and Lo, H.P. 1983. Selecting a power series distribution for goodness of fit. The Canadian Journal of Statistics 14(4), 347-353.

[4] Bult, J.R. and Wansbeek, T. 1995. Optimal selection for direct mail. Marketing Science 14(4).

[5] Jackson, R. and Wang, P. (1994), Strategic Database Marketing, Lincoln Wood, IL, NTC

*****