The Korean Journal of Applied Statistics (2010)
23(4), 739–745

Computing the Repurchase Index Based on Statistical Modeling

Whasoo Bae[1] · Wooseok Jung[2] · Youngbae Lee[3]

[1]Department of Data Science, Inje University; [2]Institute of S/W technology, NURI solution, Inc.
[3]Department of Data Science, Inje University

(Received March 2010; accepted June 2010)

Abstract

This paper computes the repurchase index based on statistical modeling. Using the transaction record of a certain product, the repurchase index is obtained by fitting the Poisson regression model. The customers are classified into 5 groups based on the index giving the information about the propensity to repurchase.

Keywords: Poisson regression model, transaction record, repurchase index.

1. Introduction

It is important for a company to keep a beneficial relationship with the customers and trying to make present customers loyal in buying their products. If the possibility of the next transaction can be estimated from their past record, it will help to plan the proper services which suit the customer needs, keeping the mutual relationship close. Especially if the customers are classified into several groups based on the measured scale of purchase propensity, the target marketing for each customer group maintains the customer relationship management effectively (Lee *et al.*, 2005).

Many types of indexes that explain cultural or social situations can be found in various fields and most of them are generated by combining the various indicators which show the simple statistics of the occurrences, like the frequency, the percentage. However Land (1975) mentioned the importance of the modeling process accompanied with proper explanatory variables to relate the status to a certain policy. Kim and Choi (2008) used the weighted average of frequencies for the concepts obtained from the survey to get the digital culture index. Choi *et al.* (2008) developed the foodborne index via the statistical approach, which is better in explaining the foodborne disease occurrence than the original index given by the Korean Meterological Administration.

This work is to compute the repurchase index using statistical modeling. In Section 2, the proper modeling variables are obtained by analyzing the transaction data to fit the statistical model. The Poisson regression model is fitted to estimate the repurchase possibility in Section 3. Section 4 shows how to derive the repurchase index and the customer classification based on the index, interpreting the trait of each group near future purchases and Section 5 discusses the concluding remarks.

[1]Corresponding author: Professor, Department of Data Science/Institute of Statistical Information, Inje University, 607 Obangdong, Gimhae, Gyeongnam 621-749, Korea. E-mail: statwbae@inje.ac.kr

Table 2.1. Variable summary

variable name	definition	data period	variable type
distance	number of days between the last purchase and 01.31.2006	Period 1	explanatory variable
freq	the number of purchases		
mean_int	mean number of days between the purchases		
freq(rep)	the number of repurchase	Period 2	target variable

Table 2.2. Descriptive statistics of the explanatory variable

variable	repurchase	N	mean	sd	min	max
mean_int	no	207,969	352.06	229.84	0.00	550.00
	yes	47,151	190.18	204.61	0.00	550.00
freq	no	207,969	2.17	2.23	1.00	99.00
	yes	47,151	4.94	5.36	1.00	259.00
distance	no	207,969	257.83	154.30	0.00	548.00
	yes	47,151	117.11	129.68	129.68	548.00

2. Computing the Variables

The raw data used here is the transaction record of a certain product which contains the customer code, order code, purchase amount and discount amount of 266,897 customers of 21 months from 03.2004 to 04.2006.

Taking 01.31.2006 as base time point, the data before 01.31.2006 is defined as the **data of Period 1** which is used to compute three explanatory variables. The first explanatory variable, **distance**, is computed by the number of days between the last purchase date and 01.31.2006. The second one, **freq** is defined as the number of purchases occurred. The variable **mean_int** is defined to be the mean of the purchasing intervals between two purchases, computed by (last purchasing date − first purchasing date) / (**series freq** − 1) in the Period 1. When **freq** = 1, **mean_int** is not defined because the customers are new. In this case, **mean_int** is set to be 550 days, which is the number of days in the Period 1.

For the computed explanatory variables, the multicolinearity among the variables might occur. Since the maximum of VIF(variance inflation factor) of the new variables is 1.61 and the maximum of condition index is 6.28; the multicolinearity problem dose not occur in this case.

Using the data after 01.31.2006, called the **data of Period 2**, the target variable, **freq(rep)** is computed as the number of purchases, which is regarded as the number of repurchases in the near future. Table 2.1 summarizes the variables.

Table 2.2 shows the descriptive statistics of the explanatory variable partitioning customers into two groups(with and without repurchase). The mean of **mean_int** for the customers without repurchase is 352.06 days, which is longer than that of customers with repurchase, 190.18 days. The customers with repurchases has a larger mean mean of the variable **freq** than the customers without purchase by two times. The mean of the variable **distance** for without repurchase group is 257.83 days, which is longer than that of the repurchase group, 117.11 days.

In Table 2.3, Welch's t-test to compare the means of two groups(without and with repurchase) shows significant differences in all the variables.

The whole data set is randomly partitioned into three parts, the training data set(30%), the testing data set(40%) and the evaluation data set(30%).

Table 2.3. Comparison of means in two groups(with and without repurchase)

variable	t value	pr $> \|t\|$
distance	205.02	$< .0001$
freq	-109.91	$< .0001$
mean_int	151.48	$< .0001$

Table 2.4. Repurchase status of three data sets

repurchase	training data	testing data	evaluating data	total
no	83,277 (18.61%)	62,364 (81.48%)	62,328 (81.44%)	207,969 (81.52%)
yes	**18,771 (18.39%)**	**14,172 (18.52%)**	**14,208 (18.56%)**	**47,151 (18.48%)**
total	102,048	76,536	76,536	255,120

Table 3.1. Estimates of the regression coefficients

variable	df	estimate	s.e	χ^2	Pr $> \chi^2$
Intercept	1	0.1062	0.0089	142.11	$< .0001$
distance	1	-0.0058	0.0001	11137.80	$< .0001$
freq	1	0.0194	0.0003	4649.66	$< .0001$
mean_int	1	-0.0019	0.0000	3320.12	$< .0001$

(See Kang *et al.*, 2003) The training data set is used in modeling and the other two data sets evaluate the fitted model. Table 2.4 compares repurchasing status of three data set and the whole data set to see whether the repurchase tendency is similar or not. It shows that there is little difference in repurchasing rate among the data sets, 18.48% for the whole data, 18.39% for the training data, 18.52% for the testing data and 18.56% for the evaluating data.

3. The Repurchase Model

3.1. Poisson regression model

Because the response variable, **freq(rep)** is the number of repurchases, which is the count variable for rare events, the Poisson regression model is considered to be proper in explaining the relationship between the target variable and the explanatory variables. Poisson regression is a method to model the frequency of event counts or the event rate. The counts are assumed to follow a Poisson distribution with other variables that are modeled as a function of the covariates (Agresti, 2007; Lindsey, 1995; Frome, 1983) The Poisson regression model is a special case of a generalized linear model(GLM) with a log link - this is why the Poisson regression may also be called Log-Linear Model (McCullagh and Nelder, 1989). Using the logarithm function as the link function, the fitted model is shown as in (3.1).

$$\begin{aligned}\hat{\mu} &= \hat{E}(\mathbf{freq(rep)}) \\ &= \exp(0.1062 - 0.0058 \times \text{distance} + 0.0194 \times \text{freq} - 0.0019 \times \text{mean\_int}),\end{aligned} \tag{3.1}$$

where $\hat{\mu}$ is the estimated mean of **freq(rep)**. All the regression coefficients are highly significant as shown in Table 3.1 so the model seems to explain the repurchase propensity quite well. Table 3.2 checks the model adequacy and the result seems satisfactory.

Using the Poisson probability function the probability that **freq(rep)** occurs x times is computed by

Table 3.2. Goodness of fit result

criterion	df	value	value/df
Deviance	102,044	88,666.9811	0.8689
Scaled Deviance	102,044	88,666.9811	0.8689
Pearson Chi-square	102,044	202,076.6449	1.9803
Scaled Pearson Chi-Square	102,044	202,076.6449	1.9803
Log Likelihood		−52,450.0914	

Table 3.3. The descriptive statistics of $\hat{p}$ (p:the possibility that repurchase occurs)

N	mean	sd	minimum	maximum
102048	0.2343	0.2064	0.0168	1.0000

Table 3.4. The distribution of $\hat{p}$

$\hat{p}$	percentage
$0.0 \sim 0.1$	38.90
$0.1 \sim 0.2$	16.77
$0.2 \sim 0.3$	12.75
$0.3 \sim 0.4$	8.20
$0.4 \sim 0.5$	6.96
$0.5 \sim 0.6$	8.48
$0.6 \sim 0.7$	6.53
$0.7 \sim 0.8$	1.22
$0.8 \sim 0.9$	0.16
$0.9 \sim 1.0$	0.03
Total	100.00

$P(\mathbf{freq(rep)} = x|\mu) = \mu^x e^{-\mu}/x!$, $x = 0, 1, 2, \ldots$. Hence the possibility that repurchase occurs can be obtained by the probability $p = P(\mathbf{freq(rep)} \geq 1|\mu)$. Using $\hat{\mu}$ obtained in (3.1), this probability, p is estimated as follows.

$$\hat{p} = \hat{P}(\text{freq(rep)} \geq 1|\hat{\mu}) = 1 - \hat{P}(\text{freq(rep)} = 0|\hat{\mu}) = 1 - e^{-\hat{\mu}}. \tag{3.2}$$

The descriptive statistics and the distribution of $\hat{p}$ is shown in Table 3.3 and 3.4, respectively. The mean of $\hat{p}$ is 0.234 and 40% of $\hat{p}$ are distributed in the range below 0.1, which shows that the repurchase possibility is generally quite low.

3.2. Evaluation of the model

In Section 3.1, the model from the training data explains the relationship between the target variables and the explanatory variables quite well. In this section, the model is evaluated whether the fitted model predicts the status of the repurchase using three partitioned data sets, the training data, the testing data and the evaluation data. To set the demarcation point of $\hat{p}$ to predict whether each customer will repurchase or not, the Heidke Skill Score (Heidke, 1926) is used. The Heidke Skill Score(HSS) is designed to increase the pure prediction power by subtracting the rate of random prediction from the rate of correct prediction. The HSS is computed by $\text{HSS} = (p1 - p2)/(1 - p2)$, where $p1 = (A + D)/N$, $p2 = (O_1F_1 + O_2F_2)/N^2$ in Table 3.5 and the pure prediction gets higher as the HSS tends to increase.

By maximizing the HSS, the demarcation point of $\hat{p}$ is determined as 0.48, and applying 0.48 to the training data, about 81.96%(= 73.11% + 8.85%) of right prediction rate was obtained in Table

Table 3.5. Classification table of repurchase status

frequency		observed repurchase status		Total
		yes	no	
forecasts	yes	A	B	$F_1 = A + B$
	no	C	D	$F_2 = C + D$
Total		$O_1 = A + C$	$O_2 = B + D$	$N = A + B + C + D$

Table 3.6. Classification table of repurchase status in the training data

frequency (percentage)		observed repurchase status		Total
		yes	no	
forecasts	yes	74,604(**73.11**)	9,744(9.55)	84,348(82.66)
	no	8,673(8.50)	9,027(**8.85**)	21,127(27.00)
Total		83,277(81.61)	18,771(18.39)	102,048(100.00)

Table 3.7. Classification table of repurchase status in the testing data

frequency (percentage)		observed repurchase status		Total
		yes	no	
forecasts	yes	55,525(**72.55**)	7,395(9.66)	62,920(82.21)
	no	6,839(8.94)	6,777(**8.85**)	13,616(17.79)
Total		62,364(81.48)	14,172(18.52)	76,536(100.00)

Table 3.8. Classification table of repurchase status in the evaluation data

frequency (percentage)		observed repurchase status		Total
		yes	no	
forecasts	yes	55,713(**72.79**)	7,310(9.55)	63,023(82.34)
	no	6,615(8.64)	6,898(**9.01**)	13,513(17.66)
Total		62,328(81.44)	14,208(18.56)	76,536(100.00)

3.6. Table 3.7 and 3.8 show that the prediction rates of testing data and the evaluation data are 81.40% and 81.80%, respectively. Three data sets have similar prediction rate, almost 82%, which is not bad.

This result of right prediction rate seems to show that the model may be used to compute the repurchase index.

4. The Repurchase Index

4.1. Computing the repurchase index

Since the fitted model was checked and found to predict the repurchase status reasonably in Section 3, $\hat{p}$ from the fitted model representing the possibility that the customer purchases the product in the near future may be used in obtaining the repurchase index. We change the scale of $\hat{p}$ to make the range between 0 and 100, multiplying by 100 and will use it as the repurchase index. The derivation of the repurchase index may be summarized in the following steps.

(1) Compute the proper explanatory variables and target variable for fitting the Poisson regression model.

(2) Obtain the estimated the mean number of the repurchase, $\hat{\mu}$.

(3) Using $\hat{\mu}$ in (2) and the Poisson probability function, estimate the probability that the repurchase

Table 4.1. Classification of the customer group based on the Rep_index

group	Rep_index	mean of repurchase frequency	repurchase status
1	86 ∼ 100	1.9661 ∼ 9.2103	Very High
2	63 ∼ 86	0.9943 ∼ 1.9661	Fairly High
3	40 ∼ 63	0.5108 ∼ 0.9943	Common
4	22 ∼ 40	0.2485 ∼ 0.5108	Low
5	0 ∼ 22	0.0000 ∼ 0.2485	Very Low

Table 4.2. Classified customers based on the Rep_index

group	Rep_index	number of customer	percentage(%)
1	86 ∼ 100	140	0.05
2	63 ∼ 86	13,683	5.36
3	40 ∼ 63	45,820	17.96
4	22 ∼ 40	47,448	18.61
5	0 ∼ 22	148,029	58.02

occurs by $\hat{p}$=$\hat{P}$(Y$\geq$ 1|$\hat{\mu}$)=1 − $e^{-\hat{\mu}}$, as in (3.2).

(4) Compute the repurchase index by multiplying $\hat{p}$ by 100.
Rep_index = $\hat{p} \times 100$.

4.2. Interpretation of the repurchase index

It would be better for understanding if we can relate the Rep_index to the specific number showing the repurchase propensity. Since the Poisson regression model can relate $\hat{p}$ with $\hat{\mu}$, we will interpret the Rep_index in terms of the mean number of repurchases.

Based on the Rep_index and $\hat{\mu}$, customer groups are classified into 5 groups and interpreted their propensity for repurchases relating the index with the estimated mean number of repurchases in Table 4.1. Table 4.2 shows the distribution of grouped customers shown in the data. The 76.63% of customers belonged to the group 4 and 5 which shows little possibility in repurchasing. But it is encouraging that the portion of group 3 which is considered to have the potential possibility shows 17.96%. If some target marketing plan helps this group repurchase and get into group 2, then more than 20% of customers will repurchase the product so only 5.4% loyal customer group increase its portion by 4 times.

5. Concluding Remarks

The repurchase index implying the possibility of customer's repurchase is derived by the statistical modeling using the transaction record of a certain product for 266,897 customers from 03.2004 to 04.2006.

The whole data set is randomly partitioned into three parts, training data for modeling, testing data and evaluation data for checking the model. Computing three explanatory variables and the target variable from the raw data, the Poisson regression model is fitted to estimate the possibility of repurchases, $\hat{p}$ and the repurchase index, Rep_index is obtained from $\hat{p}$ by multiplying 100 to make the range between 0 and 100. Based on Rep_index, the customer group is classified into 5 groups and interpreted its propensity strength in terms of the estimated mean number of repurchases.

The data shows that about 50% of customers purchased only one time and the variable information

for **mean_int** from them might weaken the model prediction, which gives the distribution of $\hat{p}$ skewed, showing 40% in the interval that $\hat{p}$ is less than 0.1. Adding more data for the next time period and changing the base time point in various ways may give the proper estimated value for **mean_int** to new customers.

Computing the proper variables helps build a more adequate model and is another important task that we have to work on. Also, the possible measurement error of explanatory variables in making new variables, should be relegated in future research.

With these works done, the distribution of the Rep_index would be less skewed and the index predicts the repurchase status more accurately so the classified group has more reliable information to be used. Also various types of modeling might be used for the count data and the model may be compared the right prediction rate of repurchase status. As well in dealing with huge data set, the choice of k in k-fold cross-validation is another part to work on in building and checking the model.

In this work, the repurchase index was computed by the statistical methods for certain products, but this method may be used in other fields because the index usually relates the occurrence status of the target variable, which is the count variable. Society is now full of information everywhere and it becomes an important issue to quantify and interpret the information of a given status. As done in this work, the statistical modeling with proper variable information could enrich the information to predict and to make decisions in the near future. Computing the sensible index of various fields in the informational society will be a job which statisticians should participate in constantly to help society.

References

Agresti, A.(2007). *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc Joboken, New Jersey.

Choi, K., Kim, B., Bae, W., Jung, W. and Cho, Y. (2008). Developing the index of Foodborne Disease Occurrence, *The Korean Journal of Applied Statistics*, **21**, 649–658.

Frome, E. L. (1983). The analysis of rates using Poisson regression models, *Biometrics*, **39**, 665–674.

Kang, H., Han, S. and Shin, H. (2003). On the development of customer lifetime value evaluation model by using data mining techniques, *Journal of the Korean Data Analysis Society*, **5**, 927–936.

Kim, M. and Choi, D. (2008). Digital culture index development research, *Research Report of Korea Agency dor Digital Opportunity and Promortion*, 07–08.

Heidke, P. (1926). Berechnung des Erfolges und der Gute der Windstar kevorhersagen im Sturmwarnungs-dienst, *Geografiska Annaler*, **8**, 301–349.

Land, K. C. (1975). Theories, models and indicators of social change, *International Social Science Journal*, **27**, 7–37.

Lee, S., Choi, S., Kim, G., Kim, O. and Kang, C. (2005). A study on development of the profit model using customer segmentation in medical field, *Journal of the Korean Data Analysis Society*, **7**, 523–532.

Lindsey, J. K. (1995). *Modeling Frequency and Count Data*, Clarenden Press, Oxford.

McCullagh, P. and Nelder, J. A. (1989). *Generalised Linear Models (2nd edition)*, Chapman & Hall.

Welch, B. C. (1947). The generation of "Student's" problem when several different variances are involved, *Biometrika*, **34**, 28–35.