Estimation of the customer mean survival time in subscription-based businesses

Z. Mohammed, S. Maritz & D. Kotze
Department of Statistics, University of the Western Cape, Cape Town, South Africa

Abstract

Two main components must be estimated in order to estimate customer lifetime value. The first component is customer survival time (time from subscription to cancellation of the service) and the second is the customer monthly margin. While the customer monthly margin could be estimated directly from the accounting model, the challenge will be the estimation of time to cancellation. In this study 30000 customers were selected from a well established subscription-based company. The customers were classified (segmented) according to their demographic and usage related characteristics. The stratified Cox model was used to identify the significant variables and to calculate hazard ratios. Both nonparametric and parametric survival analysis techniques were used to estimate the mean survival time. The results showed that gender, age and direct marketing city are significant in predicting the hazard of cancellation of the service. Young customers (age less than 26 years) have significantly shorter mean survival time than other age groups. A large difference between the restricted and unrestricted mean survival time was found; this may be due to the extreme right censoring of 85% that exists.
Key words: subscription-based businesses, customer lifetime value, Kaplan-Meier product limit, Cox proportional hazard model, parametric survival regression.

1 Introduction

Customer survival time is defined as time from subscription to cancellation of the service. The study of customer survival time has two important roles to play. The first in identifying important covariates that affect customer relationship


WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070281

with the service provider; this enables the service provider to design retention strategies. The second role is that it facilitates the estimation of customer lifetime value via the estimated survival probabilities (or estimated mean time to cancellation). In this study we investigate customer survival time in subscription-based businesses. From a number of demographic and usage related factors, the aim is to identify the significant ones and to estimate the mean time from subscription to cancellation of the service.

2 Literature review

The focus on the customer rather than the brand has resulted from the transformation of the world economy. The old economy was transaction in its nature and product-based thinking while the new is subscription in its nature and customer based-thinking. The evaluation of firm's performance is then determined on the basis of its customers' value rather than brand value (Rust et al. [1]); this has resulted in the introduction of the customer equity concept which was introduced by Blattberg and Deighton [2]. The customer equity is defined as the total future revenue that a firm expects from its relation to its customer today. In order to measure customer equity, customer lifetime value (the total expected revenue that a firm expect from its relation to its customer) is used. Two main components are needed in order to estimate customer lifetime value, namely, customer survival time (time from subscription to cancellation of the service) and the customer monthly margin. While the customer monthly margin could be estimated directly from accounting model, the challenge will be the estimation of time to cancellation.

Customer survival time was discussed by a number of researchers. Will Potts [3] outline the application of survival analysis to predictive modelling. He included a discussion on discrete-time logistic models and piece-wise hazard models. The discussion involved the applicability and parameter estimation in each of these models in addition to Weibull and piece-wise exponential models. Lu [4] discussed the estimation of customer lifetime value using survival analysis. He built a model for customer lifetime value in the telecommunication industry, emphasizing that the customer survival curve and the customer monthly margin are the most important components in modelling customer life time value. Aspects of applying the model such as sampling, variable reduction, model estimation, and model validation were considered. Linoff [5] pointed out that the nature of the survival curve in a business application might have special properties that come from the particular nature of and the practice of business. Termination of contracts, non-payment, and end promotion can lead to a sharp drop in the survival curve or a non-smooth spiky hazard function.

3 Methodology

From a database of an established internet service provider a sample of 30000 customers was selected. Time from subscription to cancellation of the service (or the date of data extraction) was recorded. Demographic variables such as,

gender, age, language, and marketing city, as well as usage related variables such as, IT background, WiFi usage, and customer segment were also recorded.

The Cox proportional hazard regression model (Cox, [6, 7]) was used to identify the significant factors that affect service cancellation. To test the proportionality assumption in Cox regression, Schoenfeld residuals [8] and scaled Schoenfeld residuals were used while the link test was used to test the correctness of Cox model specification. The generalized gamma model (Klein and Moeschberger, [9]) was used to plot the survival curve by customer age group within customer segment. The mean time from subscription to cancellation of the service was estimated using the restricted mean based on the Kaplan-Meier survival function [10]. The unrestricted (extended) mean was estimated by exponential extrapolation of the Kaplan-Meier estimate. The estimated mean time from subscription to cancellation of the service was obtained from the generalized gamma model so as to enable the comparison between the parametric and nonparametric methods. Both SPSS and STATA were used to analyze the data.

4 Results with discussion

To analyze and understand customer survival time, we take the problem through three levels of analysis. In the first level we identify the significant variables that affect the customer survival time; for this purpose we use the results obtained from a stratified Cox regression. In the second level hazard ratios and the results obtained from the generalized gamma regression are used to evaluate the effects of variables such as marketing city and age. In the third level of our analysis we look at the estimation of the mean time from subscription to cancellation of the service using both parametric and nonparametric (with and with out extrapolation) methods of estimation. As our data suffers from a high degree of censoring of 85%, the idea behind this level of analysis is to see how the estimates of the mean will differ in two scenarios; the first scenario is where we use the mean estimated from extrapolated Kaplan-Meier (unrestricted or extended mean) versus mean estimated from Kaplan-Meier with no extrapolation and the second scenario is the nonparametric versus parametric estimate of the mean.

4.1 Identifying the significant variables that affect customer survival time

Table 1 shows the hazard ratios, the confidence interval for the hazard ratio and the corresponding p-values for the stratified Cox regression model used.

The gender, age group, marketing city, IT background, and WiFi usage were found to be significant in predicting the hazard of cancelling the service. With respect to gender, females are less likely to be loyal compared to males (13% more risk of cancelling the service for females compared to males). Customers from the marketing city of Durban have the highest risk of cancelling the service when we the compare the listed cities to the unlisted cities (the category of other cities). The language factor was classified into three categories: English,

Afrikaans and other (other languages such as isiXhosa, isiZulu, etc). The hazard ratios were calculated for English and Afrikaans compared to the category of "other" (other languages). Regarding the factor of age group, the hazard ratios were calculated for two age groups: customers of age "less than 26 years" and customers of age "26 to 40 years" and were compared to the age group of "more than 40 years". Customers in the young age groups are more likely to be at risk of cancelling than the service of customers in old age groups The factor of marketing city was classified into 6 categories; these categories are Cape Town, Durban, Johannesburg, Pretoria, Wits and others (other cities in South Africa). The hazard ratio for each marketing city compared to the category of "other" (other cities in South Africa) was calculated.

Table 1: Hazard ratios, confidence interval for the hazard ratio and corresponding p-values.

The variable	Hazard ratio (95% CI)	P-value
Gender (Female)	1.13 (1.05 , 1.23)	0.002
Language (English)	0.69 (0.17 , 2.77)	0.603
Language (Afrikaans)	0.98 (0.24 , 3.95)	0.975
Age (Less than 26 years)	3.81 (3.43 , 4.23)	0.000
Age (From 26 to 40 years)	1.65 (1.52 , 1.80)	0.000
Marketing city (Cape Town)	1.01 (0.90 , 1.13)	0.860
Marketing city (Durban)	1.11 (0.96 , 0.29)	0.157
Marketing city (Johannesburg)	0.89 (0.79 , 1.00)	0.050
Marketing city (Pretoria)	1.07 (0.94 , 1.21)	0.316
Marketing city (Wits)	1.14 (1.02 , 1.28)	0.023
IT background	0.17 (0.02 , 1.21)	0.076
WiFi usage	0.42 (0.23 , 0.76)	0.004

4.2 Language, spatial and age effect on customer survival

In table 2a we present the hazard ratio of the language in the row compared to the language in the column. This enables the comparison of risk of canceling the service across languages. The table shows that customers who stated Afrikaans as their language has 1.42 time the risk of those who stated English as their language. Customers who stated Afrikaans were similar in their risk of service cancellation to those who stated other languages (such as Xhosa).

In table 2b we present the hazard ratio of the city in the row compared to the city in the column, this enables the comparison of risk of cancelling the service across cities. A large differences were seen between city of Wits and Johannesburg (the chance that a customer who subscribed in Wits will cancel the service is 1.28 time the one who subscribed in Johannesburg) followed by the difference between the Durban and the city Johannesburg (customers that

subscribed from city of Durban have 1.25 risk of cancelling the service than those who subscribed in Johannesburg). The spatial effect may be due to the differences in the income levels and usage patterns across different cities.

In table 2c we present the hazard ratio of the age group in the row compared to the age group in the. This enables the comparison of risk of canceling the service across age groups. Table 2c shows that customers of age less than 26 years have a high risk of canceling the service compared to other age groups. They have 3.81 times the hazard of more than 40 years age group customers and 2.31 times the hazard of age group 26-40 years. The reason that the young age group customers have this high risk of canceling the service could be due to lack of having a sustainable source of income or could be due to their enthusiasm to look for the new possibilities. Either way of explanation, a careful plan has to be set in place in order to retain the young age group customers.

Figure 1 above shows less survival chance for young age group customer than old age group customers. Also, it show that customer uses the service for private purposes has better chance of continuing than those who use it for both business and private purposes.

5 Conclusion and recommendations

In this study we investigated customer survival time in subscription-based businesses. From a number of demographic and usage related factors, the aim was to identify the significant factors and to estimate the mean time from subscription to cancellation of the service. The gender, age group, marketing city, IT background, and WiFi usage were found to be significant in predicting the hazard of cancelling the service. With respect to gender, females are less likely to be loyal compared to males (13% more risk of cancelling the service for females compared to males); it could be that females need a product that is personalized to meet their own preferences. Concerning the marketing city, huge

Table 2: (a) The hazard ratios of cancellation of the service by language (The calculated hazard ratios are for language in the row compared to the language in the column); (b) The hazard ratios of cancellation of the service by marketing city (The calculated hazard ratios are for city in the row compared to the city in the column); (c) The hazard ratios of cancellation of the service by age group.

(a)

		Language		
		English	Afrikaans	Other
Language	English	1	0.70	0.69
	Afrikaans	1.42	1	0.98
	Other	1.45	1.02	1

(b)

Marketing city	Marketing city						
	Cape Town	Durban	Johannesburg	Pretoria	Wits	Others
Cape Town	1	0.91	1.13	0.94	0.89	1.01
Durban	1.10	1	1.25	1.04	0.97	1.11
Johannesburg	0.88	0.80	1	0.83	0.78	0.89
Pretoria	1.06	0.96	1.20	1	0.94	1.07
Wits	1.13	1.03	1.28	1.07	1	1.14
Others	0.99	0.90	1.12	0.88	0.88	1

(c)

Age group	Age group		
	Less than 26 years	26 to 40 years	More than 40 years
Less than 26 years	1	2.31	3.81
26 to 40 years	0.43	1	1.65
More than 40 years	0.26	0.61	1

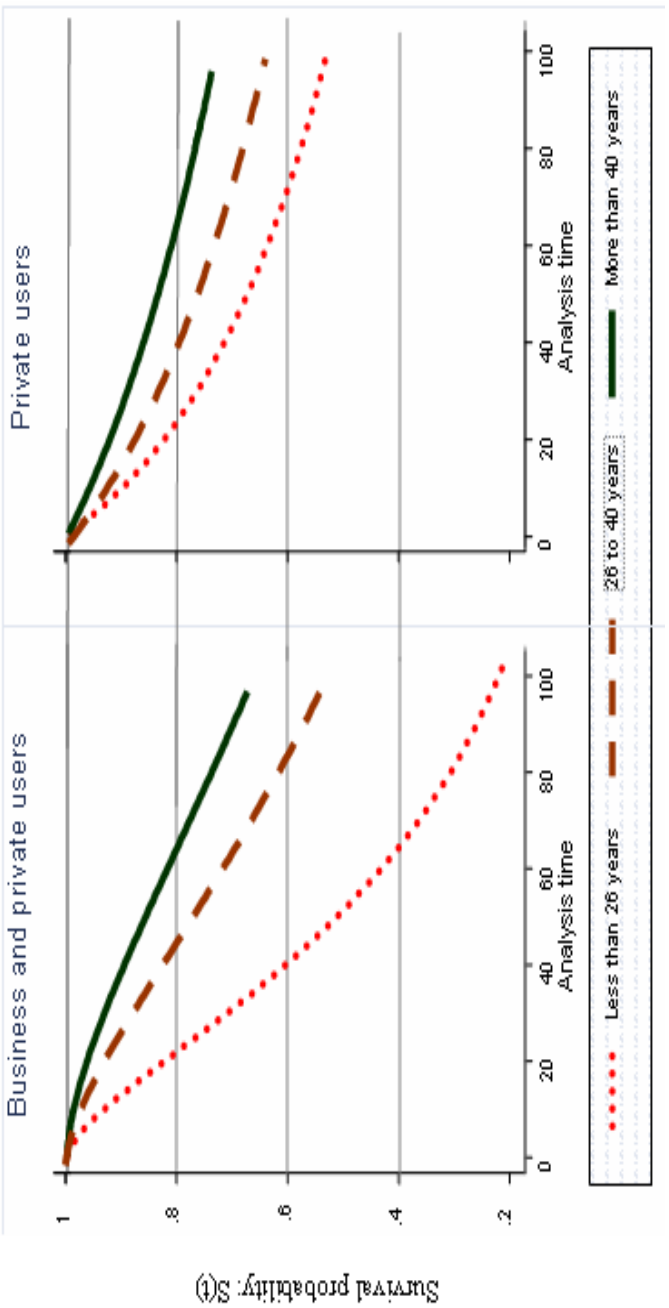

Figure 1: Customer survival curve for different age groups and by customer segment.

differences in the risk of customers cancelling the service were seen between customers who subscribed from the city of Wits and those who subscribed from the city of Johannesburg (the chance that a customer who subscribed in Wits will cancel the service is 1.28 time the one who subscribed in Johannesburg) followed by the difference between the Durban and the city Johannesburg (customers that subscribed from city of Durban have 1.25 risk of cancelling the service than those who subscribed in Johannesburg). The spatial effect may be due to the differences in the income levels and usage patterns across different cities. Customers in the young age groups were found to have lower survival chance and smaller mean time from subscription to cancellation of service than customers in old age groups.

In the process of estimating the mean time from subscription to cancellation two results were observed. The first one is the large difference between the restricted and the unrestricted mean survival time when we used Kaplan-Meier method. The second one is the large difference in the mean estimated using Kaplan-Meier and the mean estimated from the generalized gamma model. It seems that the high degree of censoring and method of extrapolating the survival curve beyond the empirical distribution have a major role in this difference. We are still busy studying the problem of extrapolating the Kaplan-Meier survival curve and we recommend further study of this issue.

References

[1] Rust, R. et al, Driving customer equity, The Free Press, 2000.
[2] Blattberg, R. & Deighton, J., Manage marketing by the customer equity test, Harvard Business Review, 74(4), pp. 136-144, 1996.
[3] Potts, W., Survival data mining (white paper), http://www.data-miners.com/resources/Will%20Survival.pdf
[4] Lu, J., Modelling customer lifetime value using survival analysis, SUGI (28), pp 120-128, 2003.
[5] Linoff, G., Survival data mining for customer insight, Intelligent Enterprise (7), 2004.
[6] Cox, D. R., Regression Models and Life-Tables, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. pp. 187-220, 1972.
[7] Cox, D. R., Partial likelihood, Biometrika, Vol. 62, No. 2. , pp. 269-276, 1975.
[8] Schoenfeld, D., Partial residuals for the proportional hazards regression model, Biometrika, Vol. 69, No. 1. , pp. 239-241, 1982.
[9] Klein, J. P and Moeschberger, M. L., Survival Analysis: Techniques for Censored and Truncated Data (Statistics for Biology and Health), Springer, 1997.
[10] Kaplan, E. L., and Meier, P., Nonparametric estimation from incomplete observation, Journal of the American Statistical Association, vol. 80, 1958.