Comparison of Multivariate Statistical Analysis and Machine Learning Methods in Retailing: Research Framework Proposition

Ivica Ćorić
Hera d.o.o, Bosnia and Herzegovina

Abstract

The aim of this paper is comparison of multivariate statistical analysis and machine learning methods based on the model used for the measurement of current and forecasting of the future customer profitability. Modern customer profitability analysis shows that customer-company relationship is burdened, beside costs of product, with many other different costs generated by business activities. Such costs generated by logistics, post-sale support, customer administration, sale, marketing etc. are allocated in customer's base in non-linear way. Allocation can vary significantly from customer to customer, making the reason why each different customer's monetary unit of revenue does not participate in profit in the same way. The research model uses RFM model to define forecasting variables and neural network, multivariate regression analysis and binary logistic regression as forecasting methods. This paper shows the ways how proposed methods can be used in process of forecasting customer profitability giving comparison of their application in that field.

Keywords: multivariate statistical analysis, RFM, machine learning, customer profitability, forecasting, knowledge
JEL classification: C45, C53

Introduction

The past years, from the beginning of the century, have been marked by the generation of huge amounts of data stored in the databases and data warehouses, intensive ICT business support and increased level of application of machine learning methods of these data (Ngai et al. 2009, Sonnenburg et al. 2010). Profitability analysis of market subjects shifts the focus from the level of products and services to individual customer (Ngai 2005). Management of companies finds a great value and basis for decision-making process in the profitability measures of individual customer. At the same time, reactive activity on the customer base does not show to be sufficient. Necessary condition for survival on the market, maintenance and growth of business results, imposes the obligation of predictive activity on the customer base. Possibility of such action is provided by large amounts of data in management information systems of companies, availability of data from the general business environment, and the development and application of artificial intelligence methods and multivariate statistical analysis.

This paper is on the trail of defined actual needs of the real sector. Research intention is to propose a model to measure current customer profitability in real time and to explore the predictive ability of such a model using machine learning methods and methods of multivariate statistical analysis. Main goal of paper is to define framework for comparing predicting possibilities of these methods.

This model predicts the use of neural network method as one of the methods of machine learning, and multiple linear regression and binary logistic regression as selected methods of multivariate statistical analysis. Elements of customer profitability were found in the RFM model of customer profitability, the dependent variables of the built model are defined on. One of the goals of the construction of the model is the verification of methods of logistic and multiple linear regressions in predicting RFM components of profitability, and comparison of predictive ability of neural network method in relation to the aforementioned methods of multivariate statistical analysis. *Methodology chapter* gives an overview of previous research in this field, defines variables of proposed model and declares machine learning methods which should be used in proposed framework. Results chapter presents data sets and performance indicators which should be used for building and testing compared methods. Discussion chapter gives an overview of scientific and pragmatic goals of research imposing some questions that require answers. At the end, conclusion chapter presents limitations of this research and gives some suggestions for further research.

Methodology

For the purpose of considering profitability, it is necessary to provide empirical indicators of the business relationship between observed company and every single customer or group of customers. Research in this area, in the end of the 1990s and the first decade of the 2000s, generated a series of models that were leaning on marketing or on the analysis of management accounting (Berger et al., 1998; Mulhern, 1999; Niraj et al.; 2001; vanRaaij et al., 2003; Krakhmal, 2006). The proposed model of prediction of customer profitability consists of independent and dependent variables and methods of machine learning and multivariate statistical analysis which are used to predict customer profitability indicators.

The independent variables of the model

The independent variables of the model are classified into the following groups according to similarity:

- Cost Policy (L1),
- Price Policy (L2),
- Features of the customer (L3).

The basis for calculating cost variables (T_n) (Table 1) was found in Niraj model (Niraj et al., 2001), which calculates the current profitability. Since this model was built over a company that deals exclusively with distribution and sale of goods, the model is expanded because of purposes of calculating the cost of production for companies engaged in the production and sale also and its products.

Product price is not fixed category and it differs from customer to customer. The independent variables (C_n) that define the deviation of the individual cost per customer than the standard price of the product makes this group of variables (Table 1).

Table 1
Independent variables of cost policy (L1) and price policy (L2)

L1-Cost policy		L2-Price Policy	
Label	Description	Label	Description
T_1	The cost of sales and direct marketing	C_1	Individual discount
T_2	The cost of processing customer orders	C_2	Territorial discount
T_3	Shipping costs	C_3	Quantity discount
T_4	The cost of procurement of suppliers and warehousing		
T_5	Costs of order preparation from suppliers		
T_6	Manipulation costs and costs of transportation of goods from suppliers		
T_7	Storage and capital costs		
T_8	Cost of sold goods		

Source: Author's table

There are a number of variables that describe the customer, but for research as significant were found (Table 2) the characteristics of size (VK_n), complexity (KK_n) and behavior in the payment obligations (PP_n). Size of the customer is defined by the traffic and internal characteristics of the customer. Complexity refers to the structure and location of purchase dispersion. Behavior in the payment is conditioned by the way of meeting the financial obligations of the buyer.

Table 2
Independent variables of customer characteristics (L3)

L3.1-Customer size		L3.2-Customer complexity		L3.3-Behavior in the payment	
Label	Description	Label	Description	Label	Description
VK_1	Size of traffic with customer	KK_1	The number of different purchased products	PP_1	The average delay time in the payment of the customer
VK_2	The gross margin achieved in traffic with customers	KK_2	The number of orders that the customer made during a given time period	PP_2	The amount of receivables paid on time
VK_3	Customer size	KK_3	Number of delivery locations	PP_3	The amount of receivables charged with delay of 30 days
VK_4	Domicile / foreign company	KK_4	The value of the return of goods	PP_4	The amount of receivables charged with delay of 90 days
		KK_5	The number of transactions of the return of goods in the observed period	PP_5	The amount of receivables charged with delay up to 180 days
				PP_6	The amount of receivables charged with delay over 180 days
				PP_7	Due debt of the buyer at the end of the observed period

Source: Author's table

All defined variables of the model can be found in the database of the company's management information system. Normalization of the **categorical variables** is conducted by forming new binomial variables in a number equal to the number of categories of the observed variable. Normalization of **nominal variables** of the model is conducted by reducing the value range from -1 to 1, using the following equation:

$$Var_x = 2 * \left[\frac{(\text{var}_x - \min(\text{var}_x))}{(\max(\text{var}_x) - \min(\text{var}_x) + 1)} \right] - 1 \quad \text{(Equation 1)},$$

where:

***Var*x** -normalized nominal variable,
***var*x** - original value of nominal variables,
max (var*x*) **-** the maximum value of the variable varx ,
min (var*x*)**-** the minimum value of the variable varx .

The dependent variables of the model

RFM analytical model uses three equally important dimensions to segment customer base and, originally, it was introduced in 1994 by Hughes (Hughes, 1994). R (Recency) dimension is the length of time period elapsed since the last purchase. F (Frequency) refers to the number of purchases in a defined period of time of analysis of customer buying habits, and M (Monetary) refers to the monetary value of the purchase of observed time period. Gupta (Gupta et al., 2006) highlights the RFM model as the simplest and most powerful model for forecasting customer profitability, and in the literature is recognized as one of the most popular methods for measuring the value of individual customer on the basis of the historical records on the realized turnover (Reinartz and Kuman, 2003; Cheng and Chen, 2009).

Table 3
The dependent variables of the model (RFM)

Label	Description	Calculation	Possible values
R	Continued purchase of customer	1-Yes 0-No	Rϵ{0,1}
F	Number of purchases	0-≤median 1-≥median	Fϵ{0,1}
M	Monetaryvalue of customer transactions	Denormalized value(Equat.1)	0≤M≥1

Source: Author's table

R variable in the model represents a binary indicator of continuing customer purchases in the following time period. Number of purchases (F variable) is also a binary indicator of the number of repeat purchases compared to the median value of the frequency of purchases of individual customer. Cash statement of customer transaction is reduced to the range [-1,1] using the equation for denormalization of this data (Equation 1).

Prediction Methods

The ability of neural networks to recognize the complex relations in the data set with a large number of variables makes them superior to conventional statistical methods. However, the ability of logistic regression are coming to the fore with a reduction in the complexity of the observed system, ie. after the reduction of the

number of input and output variables (Pourshahriar, 2012). Plenty of researches and literature in recent years clearly indicate neural networks as an excellent tool for the problem of classification and prediction (Pitts and Klepac, 2003; Campbell and Frei, 2004; Malthouse and Blattberg, 2005; Klepac and Mrsic, 2006; Donker et al., 2007; Pitts, 2007; Chiu and Tavella, 2008; Howson, 2008; Ngai et al, 2009; Vercellis, 2009; Rust et al., 2011). Almost all researches suggest the ability of approximations of any non-linear mathematical functions as an advantage over other methods of machine learning. They are treated as especially suitable for modeling of the systems that operate in the variables of which mutual relations are not known and are determined as a complex and unsuitable for processing to statistical methods. Some features which make their use difficult are inherent in them (the determination of the optimal number of layers, the number of neurons and the like) and make them time-consuming methods.

Results

On the trail of the presented variables and forecasting methods, this model provides testing and the use of neural network method for predicting the three dependent variables. Parallel to the use done will be a method of binary logistic regression for classification problem of predicting variables D and F, and method of multiple linear regressions to predict the variable M. The source of all the variables necessary for the application of empirical models is historical data from the companies' management information systems. The statistical and machine learning methods need to be set in the same conditions of operation on the data in order comparison of the methods to be credible. Thus, the model predicts the formation of two sets of data: for learning and testing ones.

Individual data set is made up of subsets relating to a calendar year. This subset consists of all values of the independent variables for each individual customer for a period of one year (t) and the value of dependent variables from the period of the next year (t + 1). The final architectures of particular methods are formed over the training data set and then applied to the testing data set. Results should then be compared by giving priority to the winning method. Evaluation is done by comparing the selected performance indicators generated by neural networks and logistic models:

- MSE (Mean Squared Error),
- NMSE (Normalized Mean Square Error),
- Correlation Coefficient r (Correlation Coefficient),
- The percentage of error - %Error,
- AIC (Akaike's information criterion).

Discussion

Scientific goal of presented research framework is defined through development of common theoretical model, based on existing scientific literature for measuring and predicting customer profitability using data from management informational systems. Research framework implies finding direction and intensity of relations between:

- Customer characteristics and customer profitability,
- Customer policy and customer profitability.

The purpose of research is defined by providing a contribution in finding the answers to the questions of how to measure and what methods to use to predict the

customer profitability. It also imposes a number of partial questions that require answers:

- What is the common framework to define customer profitability?
- How to compensate the lack of historical financial and transaction data needed to estimate the customer value?
- Are the methods of machine learning and multivariate statistical analysis usable in predicting customer profitability?
- What measures to use to express customer profitability?
- What methods to use in predicting customer profitability?
- What methods are better from the point of the predicting accuracy?

Conclusion

Existing studies of customer profitability prediction using machine learning methods and on the base of historical data that companies have in their databases, did not show a complete reliability and a great success. Some simpler methods of classical statistical analysis with using a reduced number of independent variables have proved to be just as good as and sometimes even better than those more complex ones. The designed model suggests several possible directions of research effort that would provide additional contributions to the field of study of machine learning methods in prediction of customer profitability and the improvement of the model. First of all, the machine learning method - method of neural networks is anticipated. There are many other methods that can probably just as well or even better model the classification and regression problems defined in this study.

The model has a relatively large number of independent variables. Great computing power of today's hardware resources available to the researcher and modern software packages for data processing enabled the modeling of problems with a bigger number of indicators that directly or indirectly affect the measures of customer profitability. There is space in this area to include some new variables in the model definition, and to exclude some which are considered insignificant in the model. The global business conditions, which have a recessionary tone in the case of this study, are the best example. They are a strong indicator that it would be desirable for further research to include indicators of the business environment in order to observe company and customers in an even wider and more global aspect. Of course, at this track, research in new industries will present some specific variables that characterize the business relationship with the customer.

References

1. Berger, P.D., Nasr, N.I. (1998), "*Customer Lifetime Value: Marketing Models and Applications*", Journal of Inter- active Marketing, Vol. 12 No.1, pp. 17-30.
2. Campbell, D.,Frei, F. (2004),"*The persistence of customer profitability: Empirical evidence and implications from a financial service firm*", *Journal of Service Research, Vol. 7 No.2, pp. 107–123.*
3. Cheng C.-H., Chen Y.-S. (2009), "*Classifying the segmentation of customer value via RFM model and RS theory*", Expert Systems with Applications No.36 (2009), pp. 4176–4184.
4. Chiu, S., Tavella, D. (2008), "*Data Mining and Market Intelligence for Optimal Marketing Returns*", First Edition, Elsevier.
5. Donkers, B., Verhoef, P.C., de Jong, M. (2007), "*Modeling CLV: A test of competing models in the insurance industry", Quantitative Marketing and Economics, Vol. 5 No.2, pp. 163–190.*
6. Gupta, M., Foster, G.,Sjoblom, L. (1996),"*Customer Profitability Analysis: Challenges and New Directions*", Cost Management Spring, pp. 5–17.

7. Howson, C. (2008), "*Successful Business Intelligence: Secrets to Making BI a Killer App*", McGraw-Hill.
8. Hughes, A. M. (1994), "*Strategic database marketing*", Chicago: Probus Publishing Company.
9. Krakhmal, V. (2006), "*Customer profitability analysis in service industries*", BAA Annual Conference, Apr 2006, pp.11-13, Portsmouth, UK.
10. Mulhern, F. J. (1999), "*Customer Profitability Analysis: Measurement, Concentration, and Research Directions*", Journal of Interactive Marketing, Vol.13 No.1, pp. 25-40.
11. Ngai, E.W.T (2005), "*Customer relationship management research (1992-2002): An academic literature review and classification*", Marketing Intelligence, Planning 23, pp. 582-605.
12. Ngai, E.W.T, Xiu, L., Chau, D.C.K (2009), "*Application of data mining techniques in customer relationship management: A literature review and classification*", Expert Systems with Application, Vol. 36 (2009), pp. 2592-2602.
13. Niraj, R., Gupta, M., Narasimhan, C. (2001),"*Customer profitability in a supply chain*". Journal of Marketing, Vol. 65(July), pp.1–16.
14. Panian, Ž., Klepac, G. (2003), "*Business Intelligence" (Poslovnainteligencija)*, Zagreb: Masmedia.
15. Panian, Ž. (2007),"*Business Intelligence, case studies from Croatian practice" (Poslovna inteligencija, studije slučajeva iz hrvatske prakse)*, Zagreb: Narodne Novine.
16. Pourshahriar, H. (2012), "*Correct vs. accurate prediction: A comparison between prediction power of artificial neural networks and logistic regression in psychological researches*", Procedia - Social and Behavioral Sciences, Vol. 32 (2012), pp. 97 – 103.
17. Reinartz,W. J., Kumar, V. (2003),"*The impact of customer relationship characteristics on profitable lifetime duration*", Journal of Marketing, Vol.67, pp.77–99.
18. Rust, R.T., Kumar, V., Venkatesan, R. (2011), "*Will the frog change into a prince? Predicting future customer profitability*", International Journal of Research in Marketing, Vol. 28 (2011), pp.281–294.
19. Sonnenburg, S., Ratsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., de Bona, F., Gehl, C., Binder, A., Franc, V. (2010), "*The SHOGUN Machine Learning Toolbox*", Journal of Machine Learning Research, Vol.11, pp. 1799-1802.
20. vanRaaij, E., Vernooij, M., van Triest, S. (2003),"*The implementation of customer profitability analysis: A case study*", Industrial Marketing Management, Vol. 32 No.7, pp.573-583.
21. Vercellis, K. (2009), "*Business Intelligence: Data Mining and Optimization for Decision Making*", John Wiley and Sons.

About the author

Ivica Ćorić is a PhD student at the Faculty of Economics, University of Mostar. His research interests include databases, machine learning, business information systems and software application in business and educational sector. He is co-founder of a "Hera" software company in Mostar and has more than two decades of experience in developing and implementing business information systems. The author can be contacted at **ivica.coric@hera.ba**.