Improving customer loyalty evaluation methods in the grocery retail industry: a data mining approach

Samira Khodabandehlou*

Master of Information Technology Engineering,
Department of Electronic and Computer Engineering,
Graduate University of Advanced Technology,
Kerman, Iran
Email: s.khodabandehlou@student.kgut.ac.ir
*Corresponding author

Ali Akbar NikNafs

Faculty of Computer Engineering Department,
Shahid Bahonar University of Kerman,
Kerman, Iran
Email: niknafs@uk.ac.ir

Abstract: Evaluating customer loyalty is an issue, which has gained a lot of attention in recent years due to modern facilities and tools for gathering and analysing data. These evaluations have had great and significant effects on improving business processes. Accordingly, data mining methods present significant capabilities. On the other hand, common methods for evaluating customer loyalty have been developed only based on three components, including recency (*R*), frequency (*F*) and monetary (*M*). In this study, it has been tried to add some other effective factors including number of bought products, number of returned products, amount of discount and delivery delay to the analysis in order to measure the impact of each one of them on the quality of the evaluation. The ideas and opinions of experts and the current available literature on the subject have been used as criteria for assessing quality. While implementing the methods, machine-learning tools such as artificial neural networks and support vector machine have been utilised. The results show that the method where the four factors are simultaneously fed into the RFM presents the highest possible accuracy in evaluating customer loyalty and among the learning models, the MLP-boosting method provides the highest accuracy.

Keywords: customer loyalty evaluation; data mining; neural networks; RFM; support vector machine.

Reference to this paper should be made as follows: Khodabandehlou, S. and NikNafs, A.A. (2016) 'Improving customer loyalty evaluation methods in the grocery retail industry: a data mining approach', *Int. J. Electronic Customer Relationship Management*, Vol. 10, Nos. 2/3/4, pp.158–178.

Biographical notes: Samira Khodabandehlou has received MS degree in Kerman Graduate University of Advanced Technology, Iran. Her research interests are information technology, data mining, electronic commerce and customer relationship management.

Ali Akbar NikNafs is an Associate Professor in the Department of Computer Engineering at Shahid Bahonar University of Kerman, Iran. He received his PhD degree in applied mathematics from the Shahid Bahonar University of Kerman, Iran. His research includes data mining, business intelligence and DNA computing.

1 Introduction

The majority of business owners investigate the factors affecting customer loyalty based on available information to attract more customers, satisfy their needs and make strategic decisions for achieving higher profits (Katsifou, Seifert and Tancrez, 2014). Considering the complexity and wideness of business practices, the organisation's information is very critical for achieving competitive advantage and the information obtained from the customers is considered as a source for dealing with business opportunities and challenges (Ravasan and Mansouri, 2015).

The previous studies have shown that the costs of attracting new customers are five times more than retaining current customers (Tsai and Chiu, 2004). Moreover, based on the literature, the probability of an organisation's success for repeat selling to an active customer is about 60–70% while the probability of success in selling to a new customer is approximately between 5 and 20% (Griffin and Lowenstein, 2002).

Identifying customers and gauging their levels of loyalty is a critical concern for creating competitive edge for an organisation (Tontini, 2003). By obtaining a general sense of the customers and then segmenting them into different classes, organisations can better optimise marketing plans, reach customer satisfaction and increase organisation's profit (Chen and Li, 2009; Soudagar, 2012). Customer segmentation enables organisations to improve the effectiveness as well as the efficiency of their marketing programs (Niraj et al., 2001).

One of the best methods for customer segmentation and extraction of their behavioural patterns is to use data mining algorithms. Data mining is one of the artificial intelligence techniques used for analysing a huge amount of data in order to discover augmented patterns and meaningful rules (Edelstein, 1997).

By determining the value each one of the customers has brought during their life cycle (the value of customer's life cycle), the companies can optimally allocate limited resources, use appropriate marketing strategies and ultimately reach a profitable management along with customer relation management. Using a suitable method for customer segmentation is critical for identifying the relationship between the customer value and his or her behaviour in CRM (Wang et al., 2014). The RFM model is one of the most common methods for determining the customer's life-cycle value which is based on the three measures of trade recentness, the number of trades and the monetary value of trades (Kafashpour and Alizadeh Zavarem, 2012). Because of the successful results obtained from the RFM method, many studies have dealt with customer segmentation using RFM parameters (Cheng and Chen, 2009; Chang and Tsai, 2011; Miglautsch, 2000; Yeh, Yang and Ting, 2008). RFM analysis is used for modelling customers' purchasing behaviour and measuring their loyalty, profitability and purchasing potential (Chang and Tsai, 2011; Wu and Lin, 2005). Therefore, in the current study, the augmented RFM model is used for evaluating customer loyalty.

2 Problem statement

The business owners are trying to retain their customers in the long-run using different methods and strengthen their relationships with them so that in a tough competitive environment they can increase the efficiency and productivity of the organisation on a daily basis. In doing so, they have to evaluate and determine the extent of customer loyalty and factors affecting it using different methods so that they can organise and direct these factors in a way that they can increase the customer loyalty.

On the other hand, for business organisations, the main issue is to create, develop and retain their relationship with the customers. To know how organisations can identify their customers, what is really important for the customers, what are the real needs and requirements of the customers, how the customers get motivated to purchase and what will lead to the customer satisfaction and loyalty, we have to analyse the customer's purchasing behaviour.

In business field, considering hardware tools, software facilities and intelligent systems, doing this is considered one of the objectives and important challenges of information technology (IT). In the current study, it is tried to determine the available methods for evaluating customer loyalty and then improve these methods based on data mining facilities, artificial intelligence and machine learning.

One of the methods for gauging customer loyalty is to utilise RFM model; however, based on different studies, it is clear that this model cannot perfectly and most accurately determine the extent of this loyalty. Accordingly, some of the studies have tried to improve the predicting capability of the RFM model and they use the addition of extra variables for predicting the behaviour of customers or developing new models to realise this objective. Nevertheless, nowadays there are many factors that can be effective in identifying and evaluating customer loyalty, but have not been considered.

Based on the available database for the case study, in this paper, we are focusing on developing the RFM model and in so doing, we have added four extra factors including the total number of purchased products, the total number of returned products, the total amount of discount and the average delivery delay to the RFM parameters. The results of the current study can be effectively used for customer segmentation and designing marketing strategies to create customer satisfaction and increase their loyalty.

Based on the main objective of the current study, which is to propose a new approach for evaluating the loyalty of customers using data mining methods, it tries to determine what the effects of the four factors, the total number of purchased products, the total number of returned products, the total amount of discount and the average delivery delay, are on classification and identification of customers' loyalty.

3 Basic concepts

3.1 Data mining

Data mining is the process of "selecting, assessing and modeling a huge amount of data in order to detect latent patterns in the data for achieving competitive edge for the business" (Lewis and Thornhill, 2000; Jahromi, 2009). Data mining searches the data available in information banks to extract useful information and knowledge from these data; hence, it is one of the tools used for customer relation management (Tarokh and Sharifian, 2010) and involves selecting, assessing and modelling a huge amount of data to detect latent patterns and ultimately intelligible information from huge databases (Shaw et al., 2001; Jahromi, 2009).

In the standard CRISP-DM (cross industry standard process for data mining) model, which is a very well-known and popular method for data mining (Shearer, 2000). Data mining involves six main phases which include understanding the business, understanding the data, preparing data, modelling, evaluating and developing the model (Sen, Ucar and Delen, 2012).

3.2 Customer segmentation

Customer segmentation is an important notion in marketing which is easy to define and understand from a conceptual point of view; however, its implementation and execution are difficult (Casabayo, Agell and Sánchez-Hernández, 2015; Liu, Kiang and Brusco, 2012; Yankelovich and Meer, 2006). Customer segmentation is the process of dividing customers into distinct, meaningful and homogenous subgroups based on characteristics and features, which is used as a differentiating tool in marketing (Goyat, 2011; Khobzi and Teimourpour, 2015).

This method is a basis for marketing strategies since different groups of customers indicate the need for a combination of different marketing methods (Doyle, 1987; Kahreh et al., 2014). The objective of segmentation is to understand the customers, allocate resources and provide a combination of products and services as well as to develop new methods for the product or the market (Palmer and Millier, 2004; Teichert, Shehu and Wartburg, 2008). Generally, customer segmentation is carried out based on variables of character and life style, attitude, behaviour, utilisation of a product as well as the customer's purchasing pattern (Moriarty and Reibstein, 1986; Kahreh et al., 2014) and in this study, the segmentation has been done based on the customer's purchasing pattern. The previous studies have shown that conventional segmentation methods cannot effectively deal with differences parameters (Mason, 2002, 2003; Teichert, Shehu and Wartburg, 2008); hence, nowadays, segmentation based on data mining is carried out widely and the RFM model is one of its methods.

3.3 RFM model

Hughes was the first to introduce the idea and method of RFM where he used the historical data of the customers (Hughes, 1996).

RFM is a model that differentiates important customers based on huge transaction data. This method is very useful and effective for customer segmentation (Newell, 1997);

which is used for identifying and analysing customer behaviour based on the characteristics of their current behaviour (Madani, 2009; Sohrabi and Khanlari, 2007).

To identify significant customers, RFM uses the following three variables:

- recency (R): the time difference between the last purchase of the customer and the current time
- frequency (F): the total number of purchases during a certain time period
- monetary (M): the money spent during a certain time period.

The RFM score is calculated as follows:

$$RFM = Recency * W_R + Frequency * W_F + Monetary * W_M \tag{1}$$

where W_R W_R and W_M are, respectively, the weights of the variables of recency, frequency and monetary value.

3.4 Artificial neural networks

Artificial neural networks are strong constructs for control processes, data classification and information clustering. The weights of an artificial neural network vary phase by phase based on the information it receives and the greater the extent of this input information, the lower the network error will be (Menhaj, 2000).

In the current study, multi-layer perceptron (MLP) network and RBF network are utilised.

3.4.1 The boosting method

In the boosting method, the classes are taught using sets with completely different distributions. This method can be used for increasing the efficiency of any learning algorithm as well as the accuracy of the models (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

3.4.2 The bagging method

In the bagging method, the basic learners are trained using different learning data so that there is a little difference among them to improve the machine learning based on stability and classification accuracy (Freund and Schapire, 1996).

3.5 Support vector machine

Support vector machine is one of the supervised classification methods which works based on kernel functions to change the dimensions of problem space and it is utilised in data mining and machine learning (Kecman, 2005; Ryan, Kandanaarachch and Smith-Miles, 2015).

In fact, SVM is a kernel-based classifier which tries to create a super plane between the classes so that the distance between each plane and the super plane is maximised. Maximising the margin of the super plane leads to maximised separation between classes. The data points closer to the super plane will be used for measuring this distance. Hence, these data points are called support vectors (Vapnik, 1995).

After investigating different cores of the SVM, this study uses to core functions including RBF and polynomial function.

4 Background

Zalaghi and Varzi (2014) conducted a research entitled measuring customer loyalty using an extended RFM and clustering technique. In this research, a method is provided that obtains the customer behavioural characteristics using the extended RFM and the information related to the organisation's customers. Moreover, it classifies the customers through the *K*-means algorithm and ultimately calculates the score of customers in each cluster according to their loyalty. In the proposed method of this research, first, the customer records are clustered and then the RFM model parameters are defined via selecting the features affecting the customer loyalty level using a genetic algorithm. Finality, the customers in each cluster are rated based on their impact on the amount of loyalty.

Al-Shayea, Member and Al-Shayea (2014) conducted a research aimed to predict customer behaviour using neural networks. The research results showed that the use of neural networks to predict on the basis of recency, frequency, monetary value and time (RFMT) has been accurate and useful and also these networks are highly capable of learning the models relating to customers' RFMT method.

Noyan and Simsek (2014) conducted a research with the aim of creating a conceptual model to provide a clear understanding of customer loyalty. In this research, structural equation modeling is used as a tool for data analysis to analyse the data collected from 1350 customers of the big four supermarket chains. The research results showed that among other factors, the customer satisfaction is the most important factor affecting the customer loyalty. Customer loyalty depends on price, discount, product quality, service quality, satisfaction and value perception. Price, discount and customer satisfaction have a direct and positive impact on customer loyalty.

Zakaria et al. (2014) conducted a research entitled the relationship between loyalty program, customer satisfaction and customer loyalty in retail industry: a case study. This research was carried out based on quantitative correlation. The research results showed that there is a positive relationship between loyalty initiatives, customer satisfaction and customer loyalty. Also, the results showed that the store partnership program, awarding prizes, insurance coverage and price dramatically affect customer satisfaction and also store partnership program, members' daily program, discount and price significantly affect customer loyalty.

Nikumanesh and Albadvi (2014) conducted a research entitled customer's life-time value using the RFM model in the banking industry: a case study. In this research, they assigned weights to RFM variables based on expert opinion. In the proposed method of this research with the use of clustering methods, customer records were divided into three categories of VIP, main and ordinary customers. Moreover, on the basis of means of RFM parameters, customers were divided into eight groups. The results showed that the ordinary customers were 80% of the total customers with lowest profitability.

Moslehi, Kafashpour and Azimi (2014) conducted a study entitled the use of LRFM model for customer segmentation based on their life-cycle value. In the proposed process of this research, based on CRISP data mining methods, six general phases were designed and after determination of the values of the LRFM model indicators and assigning weight to them, customers were sorted into clusters. According to the research results, customers were divided into sixteen groups and five main categories of loyal, potential, new, lost and high consumption; and their lifetime value was determined.

Keramati et al. (2014) conducted a study to evaluate data mining classification techniques including ANN, SVM, *K*-nearest neighbours (KNN) and decision tree (DT) and compared their performance. These researchers, with the analysis of the performance of these techniques, provided a combined method that highly improved the value of some of the evaluation criteria. The results showed that the combined method for Precision and Recall has a high accuracy over 0.95, and ANN significantly showed a better performance than the other three methods of DT, SVM and KNN.

Jahromi, Stakhovych and Ewing (2014) conducted a study based on data mining approach to make a model of customer churn in the field of B2B. In this study, several modeling techniques were compared in terms of their ability to model customer churn, and the findings of this research showed that the boosting method was the best way to predict customer churn.

Runge et al. (2014) conducted a study aimed at predicting the players' churn. The results found that neural networks show the best performance in predicting churn.

Murakani and Natori (2013) studied the customer relationship management through RFM+I analysis (I stands for influence). The researchers said that the traditional customer analysis was primarily based on customer's purchasing power that is pointed out by RFM. However, the importance of information distribution by customers through verbal communication leads to the creation of an important need to incorporate the ability of the customer to release information perspective in the customer management. The results showed that traditional customer management focuses on maximising profits obtainable from customers while through RFM+I analysis that takes into account the ability to disseminate information, in addition to maximising profit, new customers can be attracted on the basis of the relationship between customers.

Alvandi, Fazli and Abdoli (2012) conducted a research entitled *K*-mean clustering method for analysis customer lifetime value with LRFM relationship model in banking services. This research used the *K*-mean method for clustering, and also they applied the CRISP method on the actual data of a state-owned bank in Iran. The research results showed nine behavioural patterns in the customers' clusters, according to which some useful and practical strategies for customer relationship management have been provided.

Lai (2009) conducted a research with the title segmentation study on enterprise customers based on data mining technology. In this study, he initially defines the data mining and explains data mining methods used in CRM and analyses customer classification based on data mining and then sets up a model for customer classification. Ultimately, this research explores the application of clustering methods available in data mining for customer classification. The research results showed that the variables such as address, age, gender, income, occupation, education level, etc., play an important role in customer clustering based on RFM.

Coussement and Poel (2009) to predict customer churn, extended RFM by the two methods of eRFM (by adding social and demographic variables and other transaction variables to the RFM model) and eRFM-EMO (by adding emotional variables to the eRFM model). They used different methods of classification like support vector machine to compare the accuracy of these two models. The results showed that the eRFM-EMO model has a better performance than the eRFM model.

Yeh, Yang and Ting (2008) conducted a study with the aim of providing a comprehensive method for selecting targets in direct marketing and this method involved extending the RFM model to RFMTC through addition of two parameters, namely, the time period from the first purchase and probability of customer churn. This model can estimate the probability of customer's next purchase (in the near future) and the expected value of the total number of times that the customer will make a purchase in the future. The results showed that the offered method provided more accurate predictions than the RFM model.

Sohrabi and Khanlari (2007) conducted a study entitled customer lifetime value (CLV) measurement based on RFM model. The main objective of this research was to present a new CLV model and classify customers according to RFM model and also to provide customer retention strategies that were carried out through *K*-means, CLV and RFM methods. Research analysis steps included the following:

1 calculation of RFM variables

2 creating the CLV model

3 customer clustering using the *K*-means algorithm

4 suggested strategies for customer retention.

This research offered a CLV model according to the RFM that classifies customers into categories based on their lifetime value.

Aggelis and Christodoulakis (2005) conducted a research entitled customer clustering using RFM analysis. The main objective of the research was the calculation of RFM scoring for active electronic banking customers to evaluate the customer behaviour to make better decisions, predict future earnings and retain the most important customers. It was carried out using RFM method and *K*-means and two-step algorithms. The research results showed that through RFM scoring for active electronic banking customers, they can be rated according to the pyramid model. These results were emphasised using two clustering methods which shows the electronic banking department of a bank can easily use RFM to identify its most important customers.

Chang and Tsay (2004) in a study entitled integrating of SOM and *K*-mean in data mining clustering: an empirical study of CRM and profitability evaluation suggested the LRFM model by extending the length of relationship with the customer. After extraction of the model and clustering the data, they used a combination of two value matrices (combination of two indicators of F-M) and loyalty matrix (combination of two indicators of L-R) for their analysis and sorted the customers into five types and sixteen categories. The research results showed that adding this indicator improves the identification of loyal customers.

According to the research cited in the experimental research background, it can be said that a number of these researches have investigated the relationship between customer loyalty with other variables in a descriptive manner. These researches are mostly in the field of humanities, business administration and marketing and are separate from identifying loyal customers in the field of data mining and information technology;

and Zakaria et al. (2014) and Noyan and Simsek (2014) researches also belong to this scope. These researches measure the factors influencing the customer loyalty not the method of identifying and classifying various customers in terms of loyalty. A number of other studies only use the RFM model for customers classification and have not added any new variables to it; for example, Nikumanesh and Albadvi (2014), Tsai and Chiu (2004), Lai (2009) and Aggelis and Christodoulakis (2005). As we mentioned in section 'Background', these researchers use different sets of data to further demonstrate the credibility of RFM in customer identification and classification scheme and just use it without suggesting any new ideas and it could have been better if these researchers proposed new ideas for the development of this model.

Another group of cited researches have compared the accuracy of RFM method with other classification methods, for example, Chan (2008), and Sohrabi and Khanlari (2007). In these studies, the researchers have just compared the assessment accuracy in RFM model classification and none of them have developed the model with new variables; they have used the same three basic parameters for the evaluation and classification of customers. The fourth category of studies have added only one or two variables to the model, for example, Al-Shayea, Member and Al-Shayea (2014), Zalaghi and Varzi (2014), Murakani and Natori (2013), Alvandi, Fazli and Abdoli (2012), and Yeh, Yang and Ting (2008). In these studies, only one or two variables have been added to the RFM model which includes customer's ability to disseminate information, time and the customer lifetime, time passed from the first purchase and the possibility of the customer churn. The researchers believe that other variables should have been included.

Based on this, it can be said that in the past research in the field of data mining and customer loyalty appraisal and classifying them according to RFM model, this model has not been developed appreciably. There are other important but unaccounted factors such as the amount of discounts, the number of purchased goods, the number of sales prizes, the acceptance rate of customers' returned goods and the amount of delay in distribution of ordered goods that could help to identify loyal customers and to classify them better. Certainly these variables can play a major role in this regard so that by better identification of customers, one may manage them better and take effective steps towards their retention.

According to what was presented here, researchers have not yet provided a comprehensive model based on RFM to classify the customers and each researcher has added just one or a few variables to it or compared the RFM accuracy with other methods. So, there is a lot of research gaps in this area that would justify doing research to improve methods of evaluating customer loyalty in the field of business.

5 Proposed methods

5.1 Proposed methods for evaluating loyalty

Table 1 presents the variables added to the RFM model and establishes a new and more comprehensive method based on one or more new variables.

Table 1 Proposed methods for evaluating customer loyalty

Variable	*Measure*
RFMG	RFM + total number of purchased goods (products)
RFMT	RFM + total number of returned goods
RFMS	RFM + total amount of discount
RFMD	RFM + average distribution delay
RFMGT	RFM + total number of purchased goods + total number of returned goods
RFMGS	RFM + total number of purchased goods + total amount of discount
RFMGD	RFM + total number of purchased goods + average distribution delay
RFMTS	RFM + total number of returned goods + total amount of discount
RFMTD	RFM + total number of returned goods + average distribution delay
RFMSD	RFM + total amount of discount + average distribution delay
RFMGTS	RFM + total number of purchased goods + total number of returned goods + total amount of discount
RFMGTS	RFM + total number of purchased goods + total number of returned goods + average distribution delay
RFMGSD	RFM + total number of purchased goods + total amount of discount + average distribution delay
RFMTSD	RFM + total number of returned goods + total amount of discount + average distribution delay
RFMGTSD	RFM + total number of purchased goods + total number of returned goods + total amount of discount + average distribution delay

5.2 Gathering and preparing the data

Each research has a set of initial data which will be turned into final data after some preparation stages. Accordingly, the data set chosen for this paper includes the data related to a wholesale food provider during a full year (2013–2014) in Alborz Province, Iran. This wholesale company has 607 regular customers. The main data set for the customers' transactions and purchases included 329900 records with the following fields: invoice number, invoice date, customer code, customer name, the name of the shop, salesperson's code, the salesperson's name, the inventory name, the product code, the product name, the volume/the number of boxes, the total number, unit price, total price, the discount, the net price, the number of sale prizes, the number of returned goods, sale type, payment method, identifying code, the account name and the distribution date.

During the data preparation phase, at first there was 607 customer profiles and after data cleaning, the incomplete profiles were detected and eliminated. Hence, the number of customers' transaction was reduced to 329878 records and the number of customers was decreased to 594. Later, only the fields relevant to the topic and objective of the study were selected and irrelevant fields (invoice number, customer name, the name of the shop, salesperson's name, inventory name, the name of the goods, the volume/the number of boxes and unit price) were eliminated. Table 2 presents the new variables and the method for calculating them.

Table 2 Calculation method for new variables obtained from initial data set

New variables	*Calculation method*
Total number of goods (*G*)	$G=\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} tn_i$
Total number of returned goods (*T*)	$T=\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} nr_i$
Total discount volume (*S*)	$S=\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} d_i$
Total number of sale prizes (*P*)	$P=\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} np_i$
Average distribution delay (*D*)	$D=\frac{\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} (id-dd)_1}{n*m}$
Recency (*R*)	R = (LDI) – (CD)
Frequency (*F*)	F = distinct count of (FI)
Monetary (*M*)	$M=\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} A_i$

In Table 2, the variable i indicates the i^{th} good (product); the variable j shows the number of times the i^{th} good has been bought by the customer; tn indicates the total number of goods; nr indicates the number of returned goods; d shows the discount amount; np indicates the number of sale prizes; dd indicates the distribution date; id indicates the invoice (order) date; LDI indicates the last invoice date for the customer; CD indicates the current date; FI indicates the invoice ID and finally A indicates the price amount.

During the data integration phase, the data set for customers transactions was integrated with the new fields the researcher had created and in the data conversion phase, the format of all the variables was converted to numeral format.

6 Implementation and testing the results

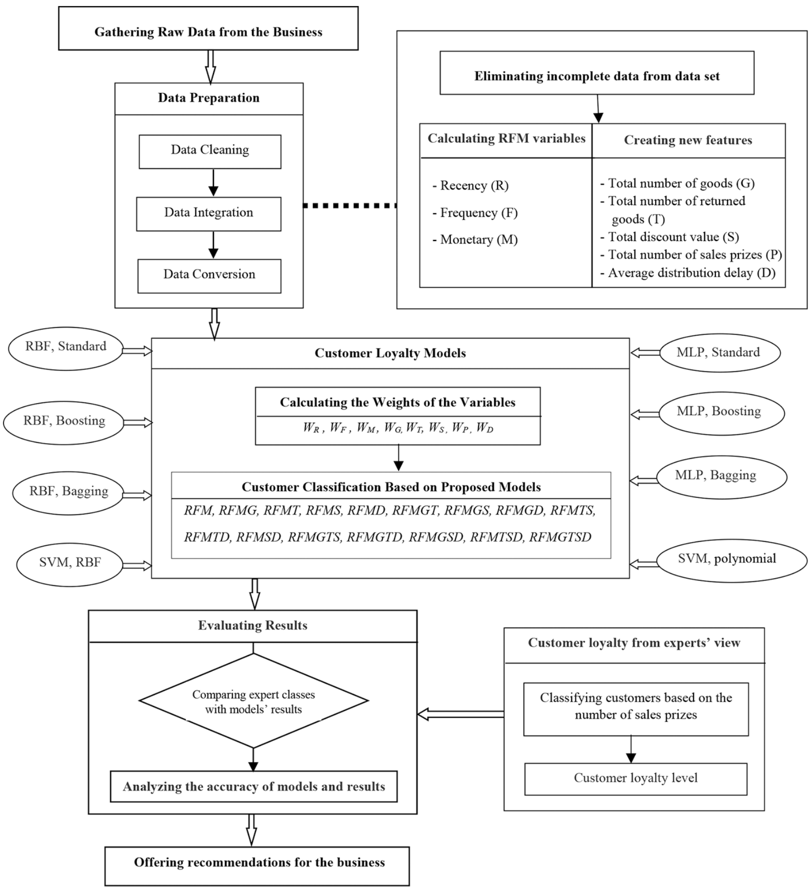
To analyse the data, IBM SPSS Statistics 21 and IBM SPSS Modeler 14.2 software applications were utilised.

To establish the RFM model and assign weights to its variables and also develop new models based on RFM, for the first time, the current paper uses support vector machine with RBF and polynomial core functions as well as MLP and RBF neural networks. In the current paper, to increase the accuracy and stability of the models, in these neural network models boosting and bagging methods are also used along with the standard model. These machine learning methods were trained with 84% (499 records) of the data set in a random manner and they were tested using the remaining 16%.

Based on the opinions of the experts, the most loyal customer of the shop is the one who has won the highest number of prizes. Based on their point of view, the total number of sales prizes can be an appropriate base for evaluating the extent of customer loyalty. Accordingly, the total number of sales prizes was selected as the reference for investigating and evaluating the performance of the proposed methods.

Since in the current study we used the RFM model as the base model to improve customer loyalty evaluation methods, in support vector machine and neural network methods, the variables of RFM were considered as the constant input variables and the classes of the total number of sales prizes were considered as the output. Besides the three variables of RFM, the variables of the total number of goods, the total amount of discount, the total number of returned goods and the average distribution delay were also considered as the variable inputs. Figure 1 presents the general framework of the proposed methods.

Figure 1 General framework of proposed methods



It should be noted that apart from RFM, eight other variables were addressed for better predictions that include total number of goods, goods diversity, type of purchasing store (wholesaler, retailer), total amount of discount, total number of returned goods, buyer's spatial distance to store, season of the year (fall, winter, spring and summer) and the average distribution delay.

In this research, to select the variables important and influential on the output, we used two methods, that is, multivariate linear regression and feature selection algorithm (in SPSS modeller). Based on the regression analysis, the *P* value was lower than 0.05 (in addition to the RFM variables for the four variables of total number of goods, total amount of discount, total number of returned goods and average distribution delay) that shows the influence of these variables in prediction. Therefore, these variables were selected as influential variables. Also, in feature selection method, the four aforementioned variables, that is, total number of goods, total amount of discount, total number of returned goods and average distribution delay along with RFM were identified as important variables and were included in research analysis model.

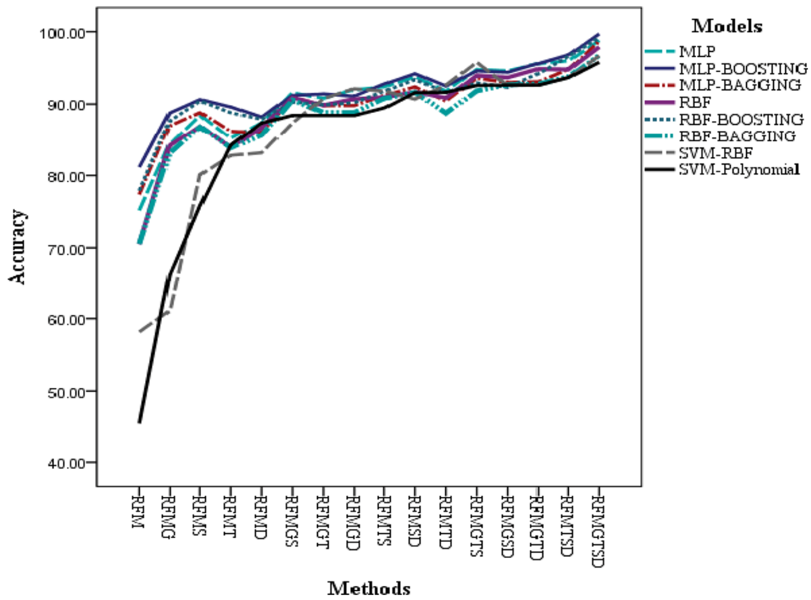
7 Findings

Table 3 and Figure 2 present the accuracy of the proposed methods compared to the RFM method for different machine learning models. Based on the obtained results among the studied algorithms, the MLP-boosting model provides the highest predicting accuracy among all the learning models and the SVM-polynomial has the lowest predicting accuracy. Moreover, the RFMGTSD method, where all the variables are simultaneously added to the RFM model, has the highest accuracy level among all the models for evaluating customer loyalty.

Table 3 Comparing the accuracy of the proposed methods using RFM method

Methods	*MLP-standard*	*MLP-boosting*	*MLP-bagging*	*RBF-standard*	*RBF-boosting*	*RBF-bagging*	*SVM-RBF*	*SVM-polynomial*	*Mean*
RFM	75.15	81.15	77.36	70.5	77.89	70.53	58.16	45.53	69.53
RFMG	84.37	88.79	86.99	84.21	87.68	83.16	61.05	66.32	80.32
RFMS	88.39	90.6	88.8	86.84	90.4	86.79	80.11	75.79	85.97
RFMT	85.19	89.6	86.2	83.89	88.8	83.84	82.79	84.21	85.57
RFMD	87.39	88.2	85.99	86.68	87.9	85.63	83.16	87.37	86.54
RFMGS	91.59	91.2	91	90.89	90.3	90.58	87.37	88.42	90.17
RFMGT	90.79	91.4	89.8	89.84	89.9	88.84	90.74	88.42	89.97
RFMGD	91.98	91.1	89.8	90.68	90.2	88.84	92.11	88.42	90.39
RFMTS	92.4	92.8	91.2	90.95	91.8	90.79	91.79	89.47	91.4
RFMSD	93.79	94.2	92.4	91.68	93.4	91.63	90.68	91.58	92.42
RFMTD	91.79	92.5	90.4	90.84	91.6	88.74	92.63	91.63	91.27
RFMGTS	94.79	94.6	93.6	93.95	92.95	91.84	95.79	92.58	93.76
RFMGSD	94.59	94.4	93	93.68	92.3	92.68	92.58	92.63	93.23
RFMGTD	95.59	95.6	93.1	94.84	94.2	92.79	92.63	92.63	93.92
RFMTSD	95.99	96.8	94.8	94.79	96.6	93.79	93.68	93.68	95.02
RFMGTSD	98.79	99.7	98.7	97.89	99.2	96.74	96.68	95.79	97.94
Mean	90.78	92.04	90.2	89.51	90.94	88.57	86.37	85.28	–

Figure 2 Comparing methods for evaluating customer loyalty (see online version for colours)



Moreover, based on the results shown in the table, the number of variables added to RFM has a direct relation with the evaluation accuracy of the proposed method so that by adding four variables, the accuracy reaches its highest level and as the number of the added variables decreases, the evaluation accuracy of the method decreases, too.

The results presented in Figure 3 show that generally the average accuracy of the MLP-boosting model is at the highest level and the SVM-polynomial provides the lowest level of accuracy. The average accuracies of the other learning models are presented in Figure 4.

Figure 3 General average accuracy based on different models of data analysis

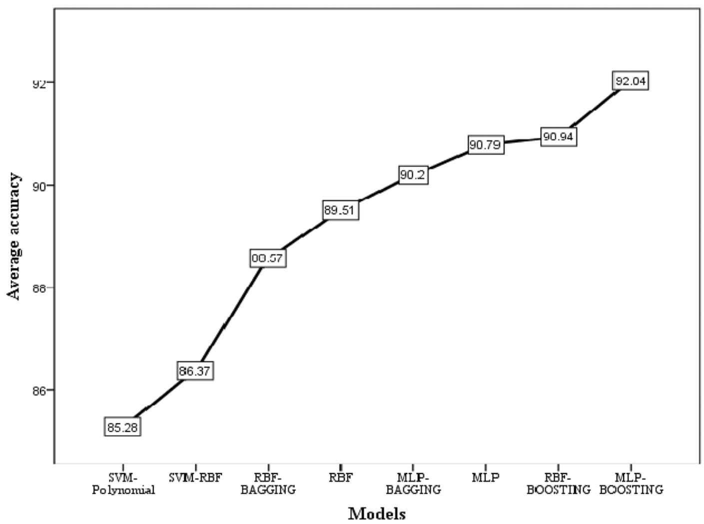
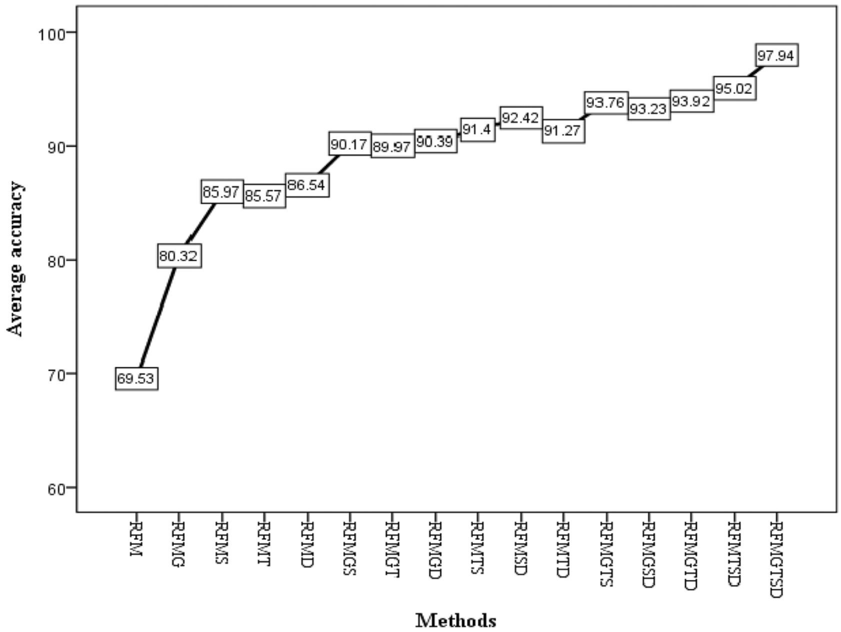


Figure 4 The general average for predicting accuracy based on the methods for evaluating customer loyalty



The results in Figure 4 show that generally (based on the average accuracies of all the proposed methods), the RFM method has the lowest accuracy and the RFMGTSD method provides the highest accuracy for evaluating customer loyalty. Moreover, there is a direct relation between the number of variables added to the RFM method and their accuracy in evaluating customer loyalty.

8 Conclusion

The results of the study show that among the studied variables, MLP-boosting has the highest accuracy, so it can be said that this model is the best method for data analysis as well as increasing the predicting power and it is recommended that in future research this model be used for data mining in order to decrease the calculation error.

The other results of the study show that among the methods and parameters added to the RFM model, the best method with the highest accuracy level is the RFMGTSD model in which four variables of the total number of purchased products, the total number of returned products, the total discount amount and the average distribution delay are simultaneously added to the RFM model. Hence, it can be said that when using the proposed RFMGTSD method for evaluating the customer loyalty and predicting their behaviour, the predicting accuracy will be added up to 97%.

The other results of the study show that by increasing the number of the new variables added to the RFM model, the accuracy level for evaluating loyalty will also increase.

When only one variable has been added to the RFM model, the highest accuracy is provided by the method where the variable of average distribution delay is added to the

RFM model and the RFMD method has been established. Accordingly, the average distribution delay can be more useful than other variables for the business owners to identify loyal customers.

Among the proposed methods with two variables added to the RFM model, the method where the amount of discount and the average distribution delay were added to RFM and created the RFMSD method had the highest accuracy. Therefore, when considering two variables for evaluating customer loyalty, the discount amount and the average distribution delay should be considered so that the accuracy of evaluating customer loyalty can be increased.

When adding three variables simultaneously, the highest prediction accuracy was provided when the total number of returned products, the discount amount and the average distribution delay were considered together for evaluating customer loyalty and the RFMTSD was established.

9 Discussion and review

Based on the results obtained from the research, the MLP-boosting method's prediction power is better than other methods and this result is in agreement with Al-Shayea, Member and Al-Shayea (2014), Runge et al. (2014), Jahromi et al. (2014), and Keramati et al. (2014). Since in boosting method, the forecast model is run several times with different training data and this type of analysis enhances the confidence level and accuracy of prediction and reduces the error, so MLP-boosting methods is one of the best and most powerful ways to predict customer behaviour on the basis of data.

According to the results, when the new and important variables such as the total number of purchased goods, the total number of returned goods, total amount of discount and average distribution delay are concurrently added to the RFM model, the prediction power goes up to about 28% and these results agree with the findings of Spring, Leeflang and Wansbeek (1999) who showed that the combined strategy is better than a simple RFM strategy for identifying customers; the findings of Zalaghi and Varzi (2014) who showed that the extended RFM obtained from the customers' information available in the organisation is more applicable in identifying customer loyalty; the findings of Hu, Huang and Kao (2013) who showed that the extended RFM methods are more useful for customer classification; and also the findings of Coussement and Poel (2009). This represents the inadequate nature of RFM model variables for prediction, and the important and significant role of including other variables used in increasing the prediction power, that is, when the total number of purchased goods is used as a variable next to the RFM, the amount of loyalty is predicted more accurately because usually the more loyal customers most likely purchase all required goods from one store so minding this parameter leads to more correct judgement about customer loyalty level. On the other hand, when the business owners accept the returned goods and goods in excess of the customers' needs, their level of loyalty to the store increases because they feel assured that if they purchase goods from this store and return it for reasons such as lack of demand, being defective, etc., the store accepts the goods and this would attract more buyers to the store and on the other hand, there is a high possibility that the people who buy more goods return them in higher amounts and this also shows these type of customers are permanent and loyal to the store.

Another factor that improves the accuracy of the model offered for predicting customer loyalty is the amount of discount the customers have been awarded by the store, that is, when the store grants discounts to customers against a certain amount of purchase, the customers are encouraged to buy more from the store and buy their required goods from only one store and with more purchases, they would get more discounts. Therefore, the higher discount rate secured by the customer is a sign of importance and loyalty of the customer and discounting strategy against a certain amount of purchase may create customer loyalty and this finding is confirmed by Zakaria et al.'s (2014) research that showed discount and price substantially influence the customer loyalty, and Noyan and Simsek's (2014) research that showed discount and price and customer satisfaction have a direct and positive impact on customer loyalty.

Another factor that has been added to the RFM is the product distribution time after the buyer's order, that is, if the store authorities take action immediately after goods order and deliver them to buyers, this would cause more loyalty but if the distribution of the ordered goods is delayed, customer loyalty decreases. Therefore, the goods distribution time may be an effective factor in accurate assessment of customer loyalty level. This variable depends on the retailers' performance and through its management they may increase their customer loyalty.

On the other hand, when the four listed variables are used at the same time besides RFM model to evaluate and predict the amount of customer loyalty, the reliability and accuracy in evaluating the level of loyalty significantly increase and model efficiency is improved since in the RFM model alone, the accuracy of customer loyalty prediction is almost 70%, so based on this, the accuracy of prediction and assessment without including other factors would be almost 30% with error but the other four variables extracted in this research increase the accuracy up to 98% and the error rate would be the minimum possible value, that is, 2%, which can be highly beneficial and valuable to evaluate the level of customer loyalty, and the organisations by accounting for these four variables along with RFM may accurately evaluate and predict their customers' loyalty. These results comply with the findings of Al-Shayea, Member and Al-Shayea (2014), Zalaghi and Varzi (2014), Spring, Leeflang and Wansbeek (1999), Hu, Huang and Kao (2013), Zakaria et al. (2014), and Coussement and Poel (2009).

Therefore, it is necessary that the business owners, especially foodstuff wholesalers, take care of the listed four variables concurrently and try to grant systematic discounts based on the purchase of a specific number of goods and by distribution of ordered goods in the least possible time (maximum 24 hours after the order), accepting customers' returned goods, and considering that all of these factors increase the number of purchased goods, significantly boost the level of customer loyalty and profitability of the store.

10 Offering recommendations and interpretation to the business

Based on the results of the current study as well as some previous works, it can be said that the augmented RFM model can be used for a more scientific and accurate identification and classification of the customers regarding their loyalty. Hence, it is recommended that the four factors, including the total number of purchased goods, the total number of returned goods, the total discount amount and the average distribution delay, along with the three factors of RFM be considered by the business owners as well as sellers to classify, identify and evaluate the customers.

It is recommended that in the future research, the MLP-boosting algorithm should be used to assess the extent of customer loyalty and their classification on the basis of RFM model among various neural network algorithms as it has the lowest error and highest level of accuracy.

To maintain the customer's loyalty to the organisation, when they return a certain product, the company should accept that and the company should consider special discounts for customers who buy more goods as well as regular customers.

Moreover, the following recommendations should be considered for attracting more customers and increasing the loyalty of the customers to the shops and business entities:

- Business owners should try to provide special discounts for their regular customers so as to encourage them to buy more goods.
- In case of buying a certain number of each product, one or more units of that product should be given to the customer as a prize.
- Business owners have to consider the number of visits as well as the number of purchased products in each visit to evaluate and identify the more profitable and loyal customers.
- The managers at shops who distribute the products too should try to use new technologies and experienced and skilful human resources to deliver the products to the customer as soon as possible.

11 Recommendations for future researchers

- Application of the developed model of this research based on the data from other commercial centres and comparing their findings with the present research to achieve a higher level of reliability appertaining to extendibility of the results.
- Study of individual variables affecting the evaluation of customer loyalty, such as age, gender, education, etc.
- using other algorithmic methods to estimate the accuracy of proposed methods.

12 Research limitations

- limitation of data to a single year
- limitation of data to a single foodstuff wholesaler, so there is a caveat in generalisation of the data
- Lack of demographic variables such as customer's age, gender, education, etc., in the store's database and as a result, exclusion of their role in prediction of customer loyalty.

References

Aggelis, V. and Christodoulakis, D. (2005) 'Customer clustering using RFM analysis', *9th WSEAS International Conference on Computers, Special Session Data Mining, Techniques and Application.*

Al-Shayea, Q.K., Member, I. and Al-Shayea, T.K. (2014) 'Customer behavior on RFMT model using neural networks', *WCE*, Vol. 1, pp.49–52.

Alvandi, M., Fazli, S. and Abdoli, F.S. (2012) '*K*-mean clustering method for analysis customer lifetime value with LRFM relationship model in banking services', *International Research Journal of Applied and Basic Sciences*, Vol. 3, No. 11, pp.2294–2302.

Casabayo, M., Agell, N. and Sánchez-Hernández, G. (2015) 'Improved market segmentation by fuzzifying crisp clusters: a case study of the energy market in Spain', *Expert Systems with Applications*, Vol. 42, pp.1637–1643.

Chang, H.C. and Tsai, H.P. (2011) 'Group RFM analysis as a novel framework to discover better customer consumption behavior', *Expert Systems with Applications*, Vol. 38, pp.14499–14513.

Chang, H.H. and Tsay, S.F. (2004) 'Integrating of SOM and *K*-means data mining clustering: an empirical study of CRM and profitability evaluation', *Journal of Information Management*, Vol. 11, No. 4, pp.161–203.

Chen, Y. and Li, X. (2009) 'The effect of customer segmentation on an inventory system in the presence of supply distributions', *Winter Simulation Conference*, pp.2343–2352.

Cheng, C.H. and Chen, Y.S. (2009) 'Classifying the segmentation of customer value via RFM model and RS theory', *Expert System with Applications*, pp.4176–4184.

Coussement, K. and Poel, D.V. (2009) 'Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers', *Expert Systemwith Applications*, Vol. 36, pp.6127–6134.

Doyle, P. (1987) 'Managing the marketing mix', in Baker, M. (Eds.):*The Marketing Book,* London, Heinemann, pp.287–313.

Edelstein, H. (1997) 'Data mining: exploiting the hidden trends in your data', *DB2 Online Magazine*, verano. Available online at: http://www.db2mag.com/9701edel.htm

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases', *Advances in Knowledge Discovery and Data Mining*, Vol. 17, No. 3, pp.1–34.

Freund, Y. and Schapire, R.E. (1996) 'Experiments with anew boosting algorithm', *International Conference on Machine Learning*, Vol. 96, pp.148–156.

Goyat, S. (2011) 'The basis of market segmentation: a critical review of literature', *European Journal of Business and Management,* Vol. 3, No. 9, pp.45–54.

Griffin, J. and Lowenstein, M.W. (2002) *Customer Winback: How to Recapture Lost Customers and Keep Them Loyal*, Jossey-Bass, San Francisco.

Hu, Y.H., Huang, T.C.K. and Kao, Y.H. (2013) 'Knowledge discovery of weighted RFM sequential patterns from customer sequence databases', *The Journal of Systems and Software,* Vol. 86, pp.779–778.

Hughes, A.M. (1996) 'Boosting response with RFM', *Marketing Tools*, Vol. 3, No. 3, pp.4–10.

Jahromi, A.T. (2009) *Predicting Customer Churn in Telecommunication Service Providers,* Master Thesis, Luleå University of Technology, Sweden.

Jahromi, A.T., Stakhovych, S. and Ewing, M. (2014) 'Managing B2B customer churn, retention and profitability', *Industrial Marketing Management*, Vol. 43, pp.1258–1268.

Kafashpour, A. and Alizadeh Zavarem, A. (2012) 'Implementing fuzzy Delphi analytical hierarchy process (FDAHP) and Hierarchical clustering analysis (HCA) in RFM model for determining the value of customer's life cycle', *Scientific and Research Periodical of Modern Marketing Research*, Vol. 2, No. 3, pp.51–68.

Kahreh, M.S., Tive, M. Babania, A. and Hesan, M. (2014) 'Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector', *World Conference on Business, Economics and Management*, Vol. 10, pp.590–594.

Katsifou, A., Seifert, R.W. and Tancrez, J.S. (2014) 'Joint product assortment, inventory and price optimization to attract loyal and non-loyal customers', *Omega*, Vol. 46, pp.36–50.

Kecman, V. (2005) 'Support vector machines - an introduction', *Studies in Fuzziness and Soft Computing*, Vol. 177, pp.1–47.

Keramati, A., Jafari-Marandi, R. Aliannejadi, M. Ahmadian, I. Mozzafari, M. and Abbasi, U. (2014) 'Improved churn prediction in telecommunication industry using data mining techniques', *Applied Soft Computing, 24, 994–1012*.

Khobzi, H. and Teimourpour, B. (2015) 'LCP segmentation: a framework for evaluation of user engagement in online social networks', *Computers in Human Behavior*, Vol. 50, pp.101–107.

Lai, X. (2009) 'Segmentation study on enterprise customers based on data mining technology', *Database Technology and Applications, IEEE First International Workshop*, pp.247–250, doi:10.1109/DBTA.2009.96.

Lewis, P. and Thornhill, A. (2000) *Research Methods for Business Students*, Prentice Hall, SAS Institute, Best practice in churn prediction, A SAS Institute White Paper.

Liu, Y., Kiang, M. and Brusco, M. (2012) 'A unified framework for market segmentation and its applications', *Expert System with Applications,* Vol. 39, No. 11, pp.10292–10302.

Madani, S. (2009) *Mining Changes in Customer Purchasing Behavior, a Data Mining Approach,* Master Thesis, Luleå University of Technology, Sweden.

Mason, K. (2002) 'Future trends in business travel decision making', *Journal of Air Transportation*, Vol. 7, pp.47–69.

Mason, K. (2003) 'Cleared for take-off: structure and strategy in the low fare airline business', *Journal of Air Transport Management*, Vol. 9, pp.69–70.

Menhaj, M.B. (2000) *Fundamentals of Neural Networks*, Amir Kabir Industrial University, Tehran.

Miglautsch, J. (2000) 'Thoughts on RFM scoring', *Journal of Database Marketing*, Vol. 8, pp.67–72.

Moriarty, R.T. and Reibstein D.J. (1986) 'Benefit segmentation in industrial markets', *Journal of Business Research,* Vol. 14, pp.463–486.

Moslehi, S.N., Kafashpour, A. and Azimi, Z.N. (2014) 'The use of LRFM model for customer segmentation based on their life cycle value', *Researches of Public Management,* Vol. 7, No. 25, pp.119–140.

Murakani, K. and Natori, S.H. (2013) 'New customer management technique: CRM by "RFM + I" analysis', *NRI Papers*, No. 186.

Newell, F. (1997) *The New Rules of Marketing: How to Use One-to-One Relationship Marketing to be the Leader in Your Industry*, McGraw-Hill, New York.

Nikumanesh, E. and Albadvi, A. (2014) 'Customer's life-time value using the RFM model in the banking industry: a case study', *Internal Journal of Electronic Customer Relationship Management*, Vol. 8, No. 1-2-3, pp.15–30.

Niraj, R., Gupta, M. and Narasimhan, C. (2001) 'Customer profitability in supply chain', *Journal of marketing*, Vol. 63, No. 3, pp.1–16.

Noyan, F. and Simsek, G.G. (2014) 'The antecedents of customer loyalty', *WC-BEM*, Vol. 109, pp.1220–1224.

Palmer, R.A. and Millier, P. (2004) 'Segmentation: identification, intuition, and implementation', *Industrial Marketing Management*, Vol. 33, pp.779–785.

Ravasan, A.Z. and Mansouri, T. (2015) 'A fuzzy ANP based weighted RFM model for customer segmentation auto insurance sector', *International Journal of Information Systems in the Service Sector*, Vol. 7, No. 2, pp.71–86.

Runge, J., Gao, P. Garcin, F. and Faltings, B. (2014) 'Chum prediction for high-value players in casual social games', *In: Proc. 2014 IEEE Conference on Computational Intelligence and Games. IEEE Comp. Soc. Press*, Washington, DC.

Ryan, S., Kandanaarachch, S. and Smith-Miles, K. (2015) 'Support vector machines for characterising whipple shield performance', *Procedia Engineering*, Vol. 103, pp.522–529.

Sen, B., Ucar, E. and Delen, D. (2012) 'Predicting and analyzing secondary education placement-test scores: a data mining approach', *Expert Systemwith Applications*, Vol. 39, pp.9468–9476.

Shaw, M., Subramaniam, C. Tan, G. and Welge, M. (2001) 'Knowledge management and data mining for marketing', *Decision Support Systems*, Vol. 31, No. 1, pp.127–137.

Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing*, Vol. 5, pp.13–22.

Sohrabi, B. and Khanlari, A. (2007) 'Customer lifetime value (CLV) measurement based on RFM model', *Iranian Accounting and Auditing Review,* Vol. 14, No. 47, pp.7–20.

Soudagar, R. (2012) *Customer Segmentation and Strategy Definition in Segments*, Master Thesis, Luleå University of Technology, Sweden.

Spring, P., Leeflang, P.S.H. and Wansbeek, T. (1999) 'The combination strategy to optimal target selection and offer segmentation in direct mail', *Journal of Marketing Focused Management*, Vol. 4, pp.187–203.

Tarokh, M.J. and Sharifian, K. (2010) 'Application of data mining to improve customer relationship management', *Scientific and Research Periodical of Industrial Management Studies*, Vol. 6, No. 17, pp.153–181.

Teichert, T., Shehu, E. and Wartburg, I. (2008) 'Customer segmentation revisited: the case of the airline industry', *Transportation Research Part*, Vol. 42, pp.227–242.

Tontini, G. (2003) 'Deployment of customer needs in the QFD using a modified Kano model', *Journal of the Academy of Business and Economics*, Vol. 2, No. 1, pp.559–612.

Tsai, C.Y. and Chiu, C.C. (2004) 'A purchase-based market segmentation methodology', *Expert Systemwith Applications,* Vol. 27, pp.265–276.

Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*, Springer Verlag, New York.

Wang, Z., Tu, L. Guo, Z. Yang, L.T. and Huang, B. (2014) 'Analysis of user behaviors by mining large network datasets' *Future Generation Computer Systems*, Vol. 37, pp.429–437.

Wu, J. and Lin, Z. (2005) 'Research on customer segmentation model by clustering', *7th international conference on Electronic commerce (ICEC), ACM*, pp.316–318.

Yankelovich, D. and Meer, D. (2006) 'Rediscovering market segmentation', *Harvard Business Review*, Vol. 84, No. 2, pp.1–10.

Yeh, I.C., Yang, K.J. and Ting, T.M. (2008) 'Knowledge discovery on RFM model using Bernoulli sequence', *Expert Systemwith Applications*, Vol. 36, pp.5866–5871.

Zakaria, B., Rahman, A.B. Othman, A. Yunus, N. Szulkipli, M.R. and Osman, M.A. (2014) 'The relationship between loyalty program, customer satisfaction and customer loyalty in retail industry: a case study', *International Conference on Innovation, Management and Technology Research*, Vol. 129, pp.22–30.

Zalaghi, Z. and Varzi, Y.A. (2014) 'Measuring customer loyalty using an extended RFM and clustering technique', *Management Science Letters*, Vol. 4, pp.905–912.