Journal of Engineering and Applied Sciences 11 (4): 828-836, 2016
ISSN: 1816-949X

Proposing a New Model for Determining the Customer Value Using RFM Model and its Developments (Case Study on the Alborz Insurance Company)

Mastooreh Moeini and Sasan H. Alizadeh
Department of Computer Engineering, Islamic Azad University, Qazvin Branch, Qazvin, Iran

Abstract: The purpose of this study is to provide a model for clustering insurance customers so that we can found any uncertainties and risks regarding customer classes, as well as influential variables in their behavior. In this regard, information related to the car insurance in Alborz Insurance Company has been collected from 2009-2013. At the first we have studied and preprocessed data and then used RFM technique and two ARFM and SRFMA approaches as well as studying the risk of customers that is due to the compensation, using k and fuzzy k-means clustering methods, customer loyalty has been reviewed. Next, we studied RFM techniques and proposed approaches as well as risk of customers in combination with each other. Numerical results affirm the superiority of fuzzy based approaches which outperforms traditional ones in term of the accuracy measures. Considering risk factor as an addition to proposed RFM scheme, the measure was enhanced and increased by 8.41%.

Key words: Client management, clustering, data mining, decision tree, k-means, RFM, risk

INTRODUCTION

Getting to know the different groups of clients and developing an effective relationship with them in such a way that the financial interests of the company would be guaranteed is one very important issue in modern business. Attracting new beneficial clients and maintaining the old ones are both principal problems which need the characteristics of the clients to be closely studied. This would be of even more importance when we find out that most companies lose 25% of their clients on average and also, the cost of attracting a new customer is five times the cost of keeping an old one. Clustering the clients and studying the different features of them is one way to gain knowledge about them. Establishing a long term relationship with the clients is the main purpose of client management. Client management is based on managing strategies and business technologies to understand the clients of a given company.

The necessity of this research: Now a days, clients are the most important business associates of a company. Due to the improvements in technology and the increase in competitive factors, companies have more need than ever to establish a meaningful relationship with their clients to attract them and to keep them. Understanding the clients and serving them the best way possible, on one hand and rating the clients and finding the most valuable among them, on the other hand are critical issues in succeeding in a business.

It is difficult to get to know the clients because of their different needs and tastes. Besides, companies are not able to serve all kinds of clients at once. Therefore, identifying the main features of the clients and clustering them would be a good strategy to find out the best way to treat each group of them.

Concepts and definitions in data mining: As the data stored in data banks increases, the need for fast, accurate methods of knowledge extraction is becoming more and more. Data mining is the collection of methods and techniques.

Gartner research group defines data mining as the process of extracting the meaningful algorithms and patterns using statistical and mathematical techniques applied on big data. There are other definitions for data mining too Table 1:

- Data mining is a search for new information among big data and it is a collaborative process between man and machine (Ngai *et al*., 2009)
- Data mining is the extraction of knowledge from data (Lefebure and Venturi, 2001)
- Data mining is the process of finding and processing databases for the purpose of obtaining knowledge (Khajvand and Tarokh, 2011)
- Data mining is the process of extracting valid, understandable and credible information from big databases and using them to make business decisions (Edelstein, 1999)

Corresponding Author: Mastooreh Moeini, Department of Computer Engineering, Islamic Azad University, Qazvin Branch, Qazvin, Iran

Table 1: Data mining applications in insurance industry

Applied model	Applied field	Main purpose	Researchers
K-means clustering	Car insurance	Finding common features between customers	Goe (2003)
CHIAD	Car insurance	Finding the importance of different risks	Goe (2003)
K-means clustering	Car insurance	Finding common features between customers	Yeo *et al.* (2001)
CHIAD	Car insurance	Predicting future customers, customers with damages and future damage amounts	Choobdar (2008)

- Data mining is the semi-automatic process of processing big databases for the purpose of finding useful patterns (Seifert, 2004)

Data mining in insurance industry: Data mining can help insurance companies gain commercial credit. For instance, by exploiting data mining techniques, companies are able to discover knowledge based on the clients' purchasing patterns. Data mining also helps decrease deception and increase risk management.

MATERIALS AND METHODS

The RFM model: It is a technique in market clustering to analysis client's behavior, e.g., purchase delay, purchase repetition, purchase cost (Khajvand and Tarokh, 2011). Tsai and Chiu (2004) did a research on developing a new methodology for client clustering based on variables like purchased items and their price. Loyal and profitable clients were identified on this basis. Hwange *et al.* (2004) proposed a model based on the profit share and the potential profit of each client to calculate the customer lifetime value and clustered the clients based upon these factors. Jonker *et al.* (2004) suggested another model in based on an application of the RFM model, to determine the optimal marketing policies and their results show that their model has been successful in clustering and managing the clients.

Wu and Lin's researches in 2005 show that the more R and F, the more probable a new contract with the client. Kim (2006) calculated each customer's value in considering the importance of the knowledge about the customer in developing a long term relation with him, gaining his loyalty and his profits. They then clustered all the customers based on this factor.

Yeo *et al.* (2001) used the RFM variables as inputs to the clustering algorithm, to choose the target market in direct marketing and developed the RFM Model adding two more parameters: the date of the first purchase and the inset probability. The purpose of the study was to propose a general methodology to discover knowledge from the database which was done by using the Bernoulli distribution and the amount of the value of waiting for all the probable purchases. Their case study was the blood transition organization and their proposed model had a better prediction comparing to the RFM Model. Change also presented the RFMTC Model in 2009 by adding two factors of the date of the first purchase and the attraction probability (Table 2).

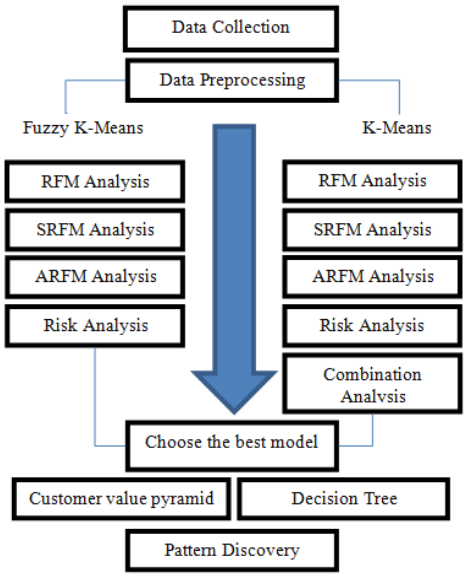

Fig. 1: General framework of the research

General framework of the research: The main purpose of this project is to present a new model for clustering the clients based on the developed RFM model by adding new variables of S (Sex) and A (Age) and using a combination of the model with risk models, using k-means and fuzzy k-means algorithm and processing the acquired data by decision trees and customer value pyramids (Fig. 1).

Preparation and the datasets: The data on the car insurance contracts at the Alborz insurance company within 2009-2013 are collected. There are 656583 records of data whose variables are summarized. A code is assigned to each customer so after summarizing, we found 205106 customers for which the number of purchases through the past year, the amount of purchase, the last purchase, age and sex are calculated.

Table 2: Some RFM developing methods

Model name	Developing method	Application
TRFM	Combination with seasonal information	Seasonal products developing
RFD	Combination with periods	Website visits analysis
RML	Combination with loyalty factor	Customer loyalty investigation
FRAT	Combination with amount and type of the purchase	Customer clustering improvement
RFR	Combination with amount of networks influence	Social networks analysis

Data sheet:

- Insurance contract's computer code
- Insurance contract's number
- Starting and ending date
- Insurer
- Duration
- Insurer's sex
- Year of birth
- Number of years without damage
- Car value
- City
- Car type code
- Car type
- Code in use
- Code in letters
- Vehicle type
- Number of cylinders
- Number of people on board
- Color
- Production date

RESULTS AND DISCUSSION

RFM analysis: The RFM technique is a common model in customer value analysis. It was suggested by Hughes in 1994 and it states the differences between customers using three variables of novelty, repetition and money value (Davis *et al*., 1992). This model analyses the behaviors of the customers and makes the prediction of their behavior using the data stored in databases (Azevedo, 2008).

To initiate the calculations, the optimum number of clusters is determined using the Davis Baldwin Method and then the clustering in applied on the RFM data by the k-means algorithm (Cheng and Chen, 2009). The Davis Baldwin method is an evaluation method in clustering which determines the dependency of each datum to each cluster by assigning a number to the datum. This number would specify a cluster to be more credible by assigning a smaller number to it. In some studies, researchers have turned this number into a "the more the better" factor by making some corrections and using similarity factors (Fig. 2).

Based on the results, the optimum number of the clusters would be 2 because the maximum of the Davis index has been obtained at this number of clusters. Therefore, the data in divided into two clusters using the k-means method and the cluster graphs is presented in Fig. 3. After extracting comparison indices and investigating the gained scores, it is clear that cluster:

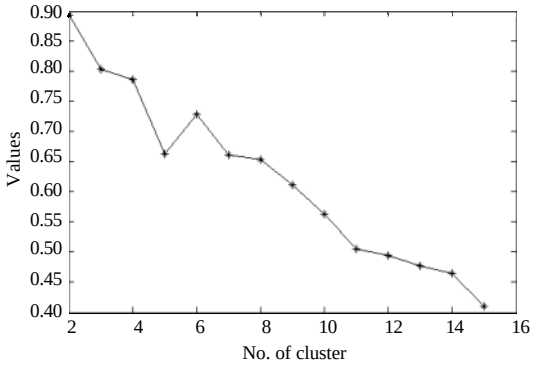

Fig. 2: Davies bouldin chart for different numbers of clusters

- Belongs to loyal customers and cluster
- Contains customers with less loyalty

By assuming label 1 for loyal customers and label 0 for less loyal customers, the decision tree of both clusters using the CART algorithm will be as presented in Fig. 4. Loyal customers are identified based on the same logic. Notice that X_1 shows variable R, X_2 shows variable F and X_3 shows variable M.

In fuzzy k-means, each datum belongs to each cluster at a specific dependency degree. Then by determining a splitting boundary, the members of each cluster will be identified. Fuzzy clustering will be done and the results of the data distribution based on the dependency degree will be as shown in Fig. 5. Extending the results for the fuzzy clustering makes it clear that the second cluster belongs to the loyal customers and the first cluster is for customers with less loyalty (Fig. 6). To build the decision tree and to analysis its performance, 70% of the data will be used as the training data and the remaining 30% will be our test data. The results are reported by the amount of sensitivity, clarity and the performance accuracy index (Fig. 7). The results are summarized in Table 3 in training and testing phases.

SRFM analysis: The SRFM analysis is discussed as an innovation aspect of the research. The main idea in this

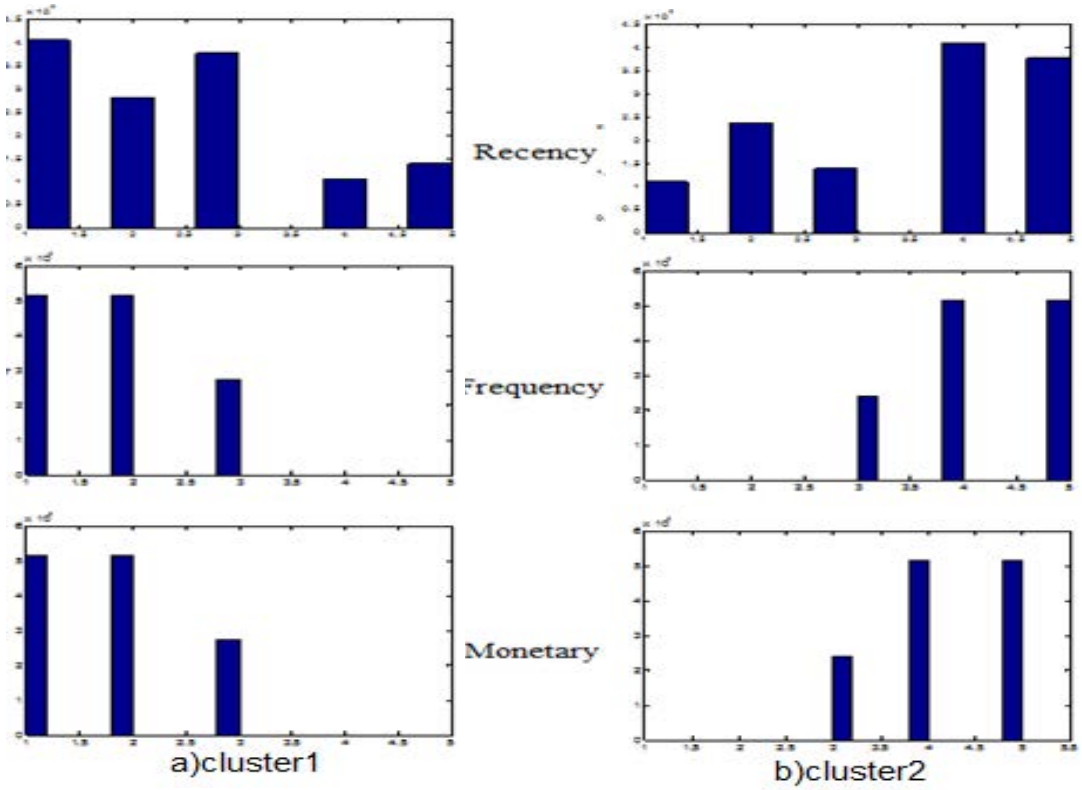

Fig. 3: Redundancy distribution histogram for R, F and M in a) cluster 1 and b) cluster 2

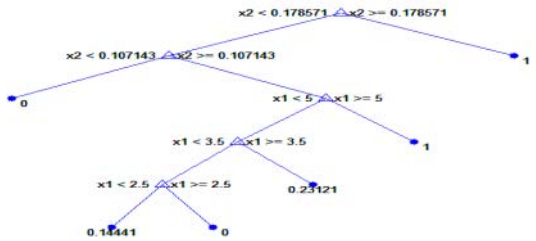

Fig. 4: The decision tree and its extracted rules for determining loyal customers

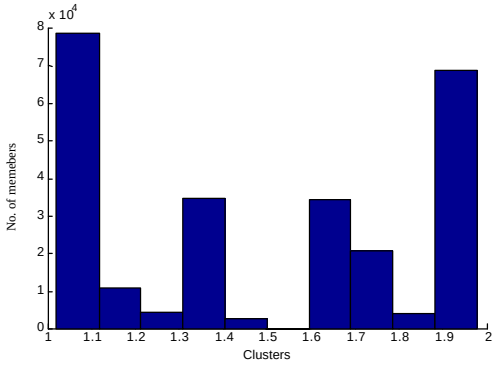

Fig. 5: Redundancy distribution of the data based on dependency degree

Table 3: Fuzzy clustering results for the decision tree

Parameter	Training phase	Testing phase
Accuracy	100	64.39
Sensitivity	100	100
Clarity	100	58.07
AUC	100	64.87

approach is to use the sex of the customer to study his/her loyalty. Since, the priority of the sexes is not defined, we are not able to reason about their value. Therefore, one of the states is assumed and the score 5 is assigned to male customers and score 1 to the female customers. Then the SRFM is applied onto them.

Cluster 2 belongs to loyal customers and cluster 3 to customers with medium loyalty and cluster 1 to customers with low loyalty (Fig. 8).

Cluster 2 belongs to loyal customers and cluster 3 to customers with medium loyalty and cluster 1 to customers with low loyalty (Table 4 and 5). By considering the sex of the customers as a new variable in the SRFM analysis, the accuracy of the decision tree is increased, therefore this analysis is preferred to classic one in terms of pattern recognition and cluster analysis (Table 6).

Arfm analysis: This approach is another innovation aspect of the current research in which age is also added as a variable to the RFM analysis. Logically, age might

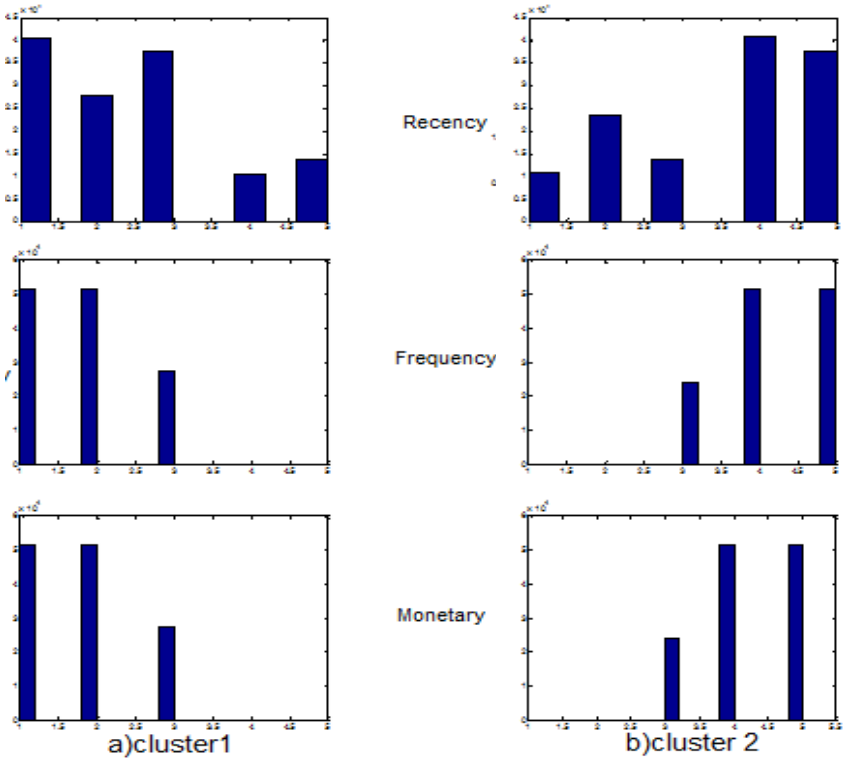

Fig. 6: Redundancy distribution histogram for R, F and M in fuzzy clustering for a) cluster 1 and b) cluster 2

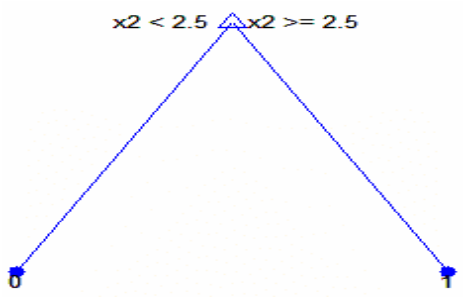

Fig. 7: The decision tree and its extracted rules for fuzzy clustering in RFM analysis

have some influences on the customer's loyalty. In this case, including age among the effective variables on the customer's loyalty, would increase the accuracy of the prediction. The results suggest an over fitting of the model such model would not be applicable in practice (Table 7).

Analysis of the risk caused by damage: This analysis assumes customers with less risk to be more loyal customers. We will first focus on the damage caused by the customer. Then, by adding RFM variables, customer's history, amount of the purchases and the number of the purchases would be included in the calculations and the age and the sex also will be combined using the ARFM and SRFM techniques. A loyal customer is considered to

Table 4: SRFM clustering comparison for the gained scores

	Cluster 1		Cluster 2		Cluster 3	
Parameter	Sum	Average	Sum	Average	Sum	Average
Gained score for all the variables	756500	3.31	1479760	4.21	1160774	2.57
Gained score for variable S	238576	4.17	359282	4.09	476220	4.21
Gained score for variable R	104404	1.82	379935	4.32	289977	2.56
Gained score for variable F	201161	3.52	374182	4.26	198977	1.76
Gained score for variable M	212359	3.71	366361	4.17	195600	1.73

Table 5: Results of the decision tree for clustering with SRFM

Parameter	Trainning phase	Testing phase
Accuracy	0.6175	0.9642

Table 6: Results of the decision tree for fuzzy clustering with SRFM

Parameter	Training phase	Testing phase
Accuracy	0.7958	0.1003

Table 7: ARFM analysis results

Parameter	Training phase	Testing phase
Accuracy (non-fuzzy clustering)	0.6469	0.0912
Accuracy (fuzzy clustering)	0.8471	0.0839

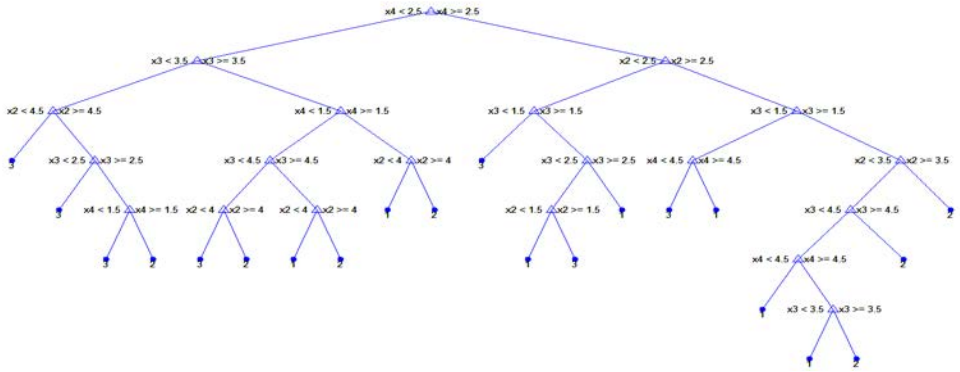

Fig. 8: The decision tree in clustering with variables R, F, M and S

be one who has used the insurance facilities less than others. Therefore, variables R, F and M would not be used and only the amount of the damage imposed by the customer while using the services (usually in 1 year) would be taken into account (Singh and Singh, 2016; Newstead and D'Elia, 2010).

To analysis the customer's loyalty using the risk factor, the optimum number of clusters in 2 and the calculated histogram shown in Fig. 9 suggests that the first cluster belongs to customers with less damage and the second cluster is for those with more damage. The decision tree based on this model and its accuracy, show this over fitting in Fig. 10.

Considering the risk in the RFM analysis: Generally, increasing the number of variables in the RFM analysis (such as age and sex) or using a combination of variables with other approaches would not decrease the accuracy of the clustering but it may increase the complexity of the model. Therefore, if the amount of increase in the accuracy is not considerable, simpler methods are preferred. On the other hand, if the accuracy is increased considerably, the combined model would have better results. Due to the over fitting of the fuzzy model and also the existence of some extent of ambiguity in its results, all our calculations are done in certain situations. Notice that this study includes one of the main purposes of the research as the combination of the risk and the RFM Model.

Finally, all the variables such as age, sex, RFM variables and also the risk are combined together and the method is applied to them and the best model is chosen for the decision tree (Fig. 11).

It is obvious that that accuracy of this model is higher than all the other ones and therefore the decision tree based on this model would be the best model to analysis the customer's behaviour. Based on this fact, this decision tree will be more precisely studied Table 8-10.

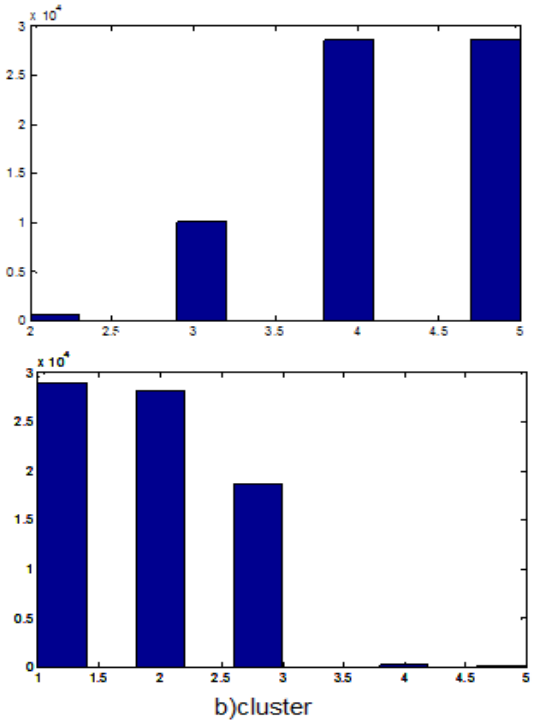

Fig. 9: Risk analysis results for a) cluster 1 and b) cluster 2

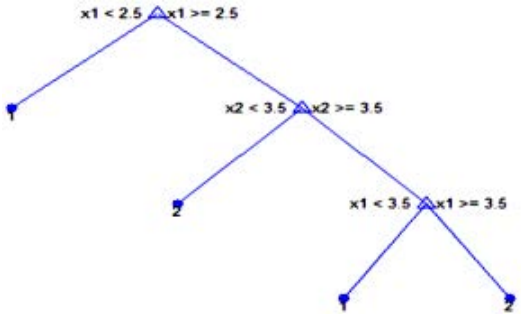

Fig. 10: The decision tree and the accuracy in risk analysis

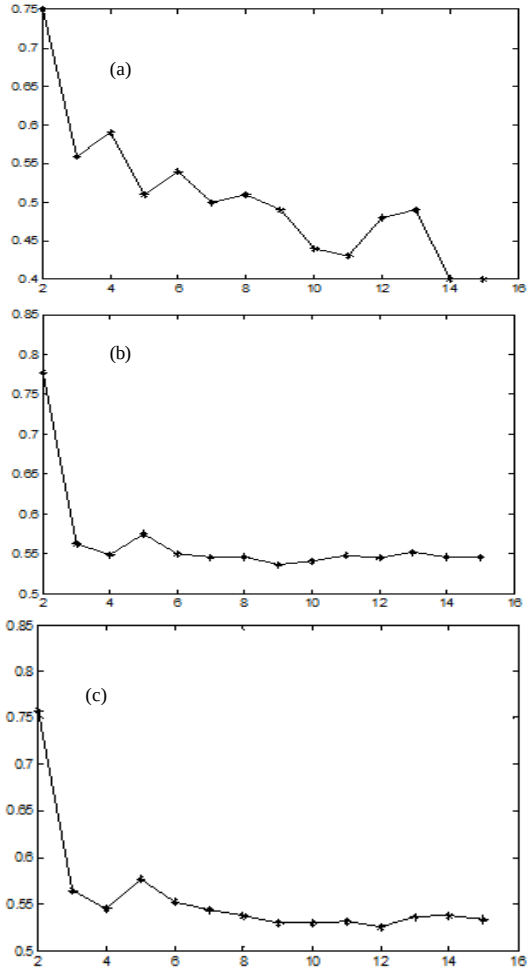

Fig. 11: Analyzing risk combined with: a) RFM accuracy level at training phase 0.7103 at testing phase 0.5427; b) ARFM accuracy level at training phase 0.7534 at testing phase 0.6127 and c) SRFM accuracy level at training phase 0.7018 at testing phase 0.5948

Table 8: Phase analysis

Parameter	Trainning phase	Testing phase
Accuracy	1	0.5045

Table 9: Results of analyzing the customer value pyramid

Category	Number of customers	Average score
Platinum	1442	14.095
Gold	5776	13.172
Silver	21619	12.507
Lead	115308	9.507

Table 10: Phases of model

Parameter	Trainning phase	Testing phase
Accuracy	0.8332	0.7604

To fulfil the purposes of our research, the customer value pyramids are also analyzed. In this pyramids,

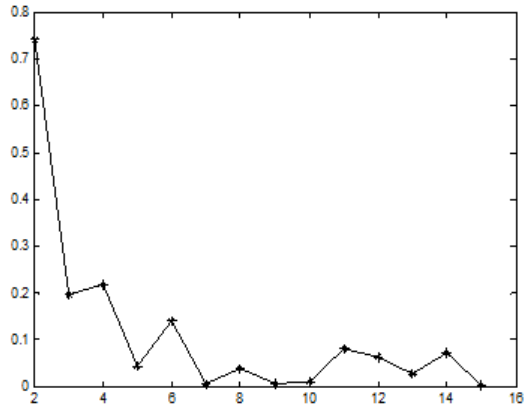

Fig. 12: Clustering with the combination of risk and ASRFM the accuracy level at training phase is 0.7018 at testing phase 0.5948

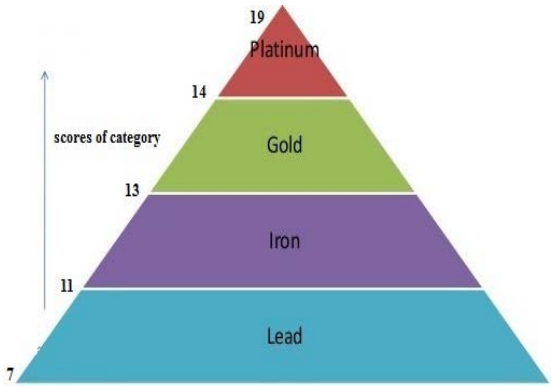

Fig. 13: Customer value pyramid based on our data

customers are divided into four categories of platinum, gold, silver and lead. Customers with the most scores of the R, F, M and risk would be the platinum customers and would have the most values. Respectively, the rest of the customers are also assigned to other categories. Based on the literature review of this research, 1% of the best customers are platinum, 4% of them are gold, 15% of them are silver and the remaining 80% are lead customers with the least value. In this approach, each customer belongs to a category based on his/her score (the RFM and the risk scores in this research) (Van Raaij *et al*., 2003). Using this same approach, results in building a pyramid which is shown in Fig. 12 and its analysis is presented in Table 8. Figure 13 shows the shape of the pyramid based on the minimum and the maximum scores in each category.

After investigating all the approaches above, the best model is chosen and the decision tree based on this model is analyzed and the patterns for loyal customers are

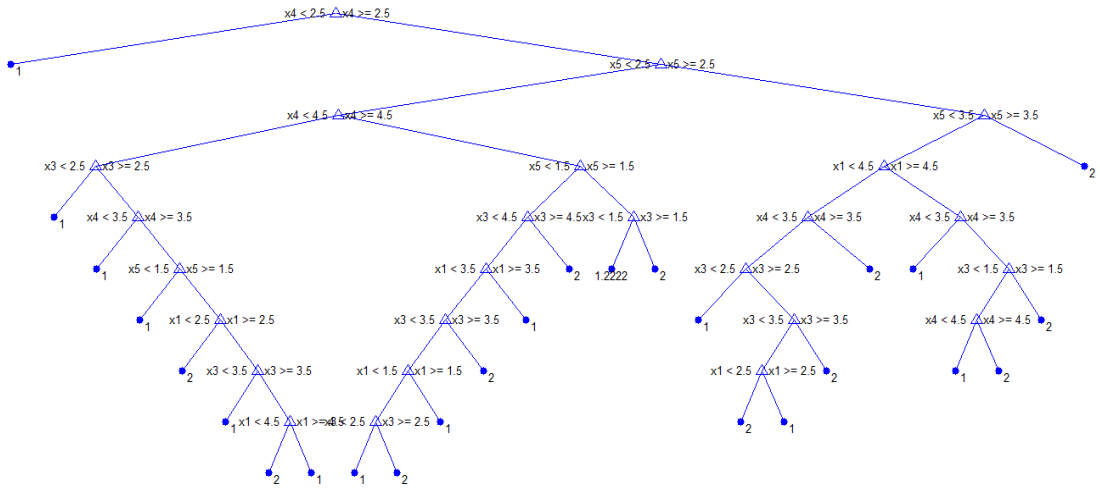

Fig. 14: The decision tree for the combination of risk and ASRFM

determined. Figure 14 shows, the number of the optimum clusters is 2 which are the loyal and the less loyal customers.

Some of the patterns are as follows:

- If the average number of the customer's visits is <2.5, the customer has low value
- The value of the customer is independent from its sex (since there is no variable related to sex in the decision tree)
- On average, younger customers are more valuable than the old ones
- The variable M has the most influence in determining the customer value. R, F and age are the next ones
- Customers who have not visited in the past 2.5 years, are considered less valuable if they are old. In case they are young, they are considered valuable if their last purchase has been >2500000

CONCLUSION

The purpose of this research is combining risk models and RFM for clustering customers. The approach used in this research was the RFM method combined with age and sex variables. The ARFM approach (using age) and the SRFM approach (using sex) have then been applied to the data in the fuzzy and certain Methods. Generally, the results present an over fitting caused by the fuzzy approach and show that the certain approach is more suitable in decision trees. For all these approaches, the optimum number of clusters is defined by the Davis Baldwin index. After analyzing the clusters and categorizing them, the decision tree is built for the model and the patterns are extracted. The risk factor is then added to the calculations, based on the fact that the more the variables, the better the performance of the model. Finally, a combination of the risk factor and the RFM analysis are applied to the data which resulted in more accuracy of the decision tree.

REFERENCES

Azevedo, A.I.R.L., 2008. KDD, SEMMA and CRISP-DM: A parallel overview. http://recipp.ipp.pt/bitstream/10400.22/136/1/KDD-CRISP-SEMMA.pdf.

Cheng, C.H. and Y.S. Chen, 2009. Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst. Appl., 36: 4176-4184.

Davis, R.H., D.B. Edelman and A.J. Gammerman, 1992. Machine-learning algorithms for credit-card applications. IMA J. Manage. Math., 4: 43-51.

Edelstein, H.A., 1999. Introduction to Data Mining and Knowledge Discovery. 3rd Edn., Two Crows Corporation, USA., ISBN-10: 1892095025, Pages: 36.

Jonker, J.J., N. Piersma and D. van den Poel, 2004. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Syst. Applic., 27: 159-168.

Khajvand, M. and M.J. Tarokh, 2011. Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Comput. Sci., 3: 1327-1332.

Lefebure, R. and G. Venturi, 2001. Data Mining: Gestion de la Relation Client, Personnalisation de Sites Web. 2nd Edn., Eyrolles, Paris, France, ISBN-13: 9782212091762, Pages: 408.

Newstead, S. and A. D'Elia, 2010. Does vehicle colour influence crash risk? Saf. Sci., 48: 1327-1338.

Ngai, E.W.T., L. Xiu and D.C.K. Chau, 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Applic., 36: 2592-2602.

Seifert, J.W., 2004. Data mining: An overview. https://fas.org/irp/crs/RL31798.pdf.

Singh, S. and S. Singh, 2016. Accounting for risk in the traditional RFM approach. Manage. Res. Rev., 39: 215-234.

Tsai, C.Y. and C.C. Chiu, 2004. A purchase-based market segmentation methodology. Expert Syst. Applic., 27: 265-276.

Van Raaij, E.M., M.J.A. Vernooij and S. van Triest, 2003. The implementation of customer profitability analysis: A case study. Ind. Market. Manage., 32: 573-583.

Yeo, A.C., K.A. Smith, R.J. Willis and M. Brooks, 2001. Modeling the Effect of Premium Changes on Motor Insurance Customer Retention Rates Using Neural Networks. Proceedings of the International Conference on Computational Science, May 28-30, 2001, San Francisco, CA., USA., pp: 390-399.