Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 120 (2017) 23–30

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 24-25 August 2017, Budapest, Hungary

A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran

Mehdi Mohammadzadeh[a*], Zeinab Zare Hoseini[b], Hamid Derafshi[c]

[a]*Department of Pharmacoeconomy&Administrative Pharmacy, Shahid Beheshti University of Medical Sciences,POBOX 14155-6153,Tehran Iran*
[b]*Department of Engineering&Technology, Payame Noor University, , Tehran, PO BOX 19395-3697, Iran*
[c]*Department of Ophthalmology, School of medicine, Alborz University of Medical Sciences, Karaj, POBOX 314977-9453, Iran*

Abstract

Nowadays Health care industry has a significant growth in using data mining techniques to discover hidden information for effective decision making. Huge amount of healthcare data is suitable to mine hidden patterns and knowledge. In this paper we traced behavior of patients during the period of 3 years in three clinics of a big public sector hospital and tried to detect special groups and their tendencies by RFML model as a customer life time value (CLV). The main goal was to detect 'potential for loyal' customers for strengthen relationships and 'potential to churn' customers for recovery of the efficiency of customer retention campaigns and reduce the costs associated with churn. This strategy helps hospital administrators to increase profit and reduce costs of customers' loss. At first, K-means clustering algorithm was applied for identification of target customers and groups and then, decision tree classifier as churn prediction was used. We compared performance of three clinics based on the number of loyal and churn customers. Our results showed that Pediatric Hematology clinic had a better performance than that of other clinics, because of more number of loyal customers.

Peer-review under responsibility of the scientific committee of the 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception.

Keywords: Hospital information system (HIS); data mining; clustering; classification; RFM model; CLV.

* Corresponding author
E-mail address: m_mohammadzadeh@sbmu.ac.ir

1877-0509
Peer-review under responsibility of the scientific committee of the 9th International Conference on Theory and application of Soft Computing, Computing with Words and Perception.
10.1016/j.procs.2017.11.206

1. Main text

Business and marketing organizations may be ahead of healthcare in applying data mining to derive knowledge from data. This is quickly changing. Successful mining applications have been implemented in the healthcare arena. In recent years, data mining techniques are widely used in healthcare system. Three important mining applications include, Hospital Infection Control, Hospitals Ranking and High-Risk Patients' identification (Obenshain and MAT 2004). Most studies have investigated prediction of different types of disease and sickness and they were less focused on the relationship between patients as customer and hospitals as organization especially patients churn and loyalty. Customer churn are more attended in other sectors such as different companies, retails and shops (Khajvand et al. (2011) and telecommunication sectors Verbeke et al. (2012) and Amin et al. (2017). Ramanan et al. have determined theory of mind performance in Alzheimer disease using data mining study. This study was done on 48 Alzheimer's patients and 44 behavioral-variant Front temporal Dementia patients. They have highlighted the relevancy of data mining statistical approaches in clinical and cognitive neurosciences Ramanan et al. (2017). For neonatal jaundice in newborns, predictive models using Naive Bayes, multilayer perceptron, and simple logistic were applied Ferreira et al. (2012). The dataset consisted of 227 healthy newborns. Also in another study, Naive Bayes classifier and J48 decision tree algorithm were used for building predictive models for MERS-CoV infections Al-Turaiki et al. (2016). The dataset used consists of 1082 records. In this paper we have used other applications of data mining on patient's data records provided in Hospital Information System (HIS). We have supposed patients as customers and tried to detect special types of patients and their behavioral tendencies with transactional data. We have used RFM (Recency Frequency Monetary) model as customer lifetime value (CLV) Khajvand et al. (2011) and marketing analysis methods for segmentation and prediction of model. In 1995, RFM defined as Recency, Frequency and Monetary. Recency is the period since the last purchase. A lower value corresponds to a higher probability of the customers to making a repeat purchase. Frequency is a number of purchases made within a certain period and higher frequency indicates greater loyalty. Monetary defines as the money spent during a certain period; a higher value indicates that the company should focus more on that customer Bult and Wansbeek, (1995). In the healthcare system, higher frequency means referring of patients for all/more health needs to our hospital and this occurs when patients are satisfied with hospital services. Also the length of time that the patient has been in contact with us is a major factor that shows whether the patient has discontinued its relationship with us or not. We have added this factor to RFM model and used RFML model to evaluate patient's loyalty. The ultimate goal in the healthcare system is to have hospital and staffs who work well so that patients as customers refer again to this hospital if needed. In this way, patients will be satisfied as our loyal customers.

1. Methodology

2.1 materials and preprocessing data

In this research a big public sector hospital (Shohadaye Tajrish Educational Hospital) has been studied. We selected three specialize clinics based on highest number of patients; the Cardiovascular, Neurology and Pediatric Hematology clinics. Three year Patients' data records were extracted from Hospital Information System (HIS). The numbers of transactions in these clinics were 19036, 41257, 3425 respectively that belonged to 8979, 22210, and 677 patients. Due to the nature of treatment industry and contrast between customer loyalty and patient loyalty, we selected outpatient data records. Outpatients select their preferred hospital with no referral by the doctor treating.
The raw dataset consisted of socio demographic characteristics of patients, services that used in visit date, admission type, insurance type and amount paid. As regards, we wanted to use RFML model for evaluation of patients' categories, so R (recency), F (frequency), M (monetary) and L (length) attributes must be calculated. Then based on clustering model, weighted Life Time Value (LTV) was calculated for each cluster. We believe that greater amount of LTV represents more faithful customers and lower amount shows churn customers. So with the help of LTV amount, four clusters have been assigned into four class label of patient (loyal, potential for loyal, potential for churn and churn) to compare patients of three clinics. Also we used R, F, M and L attributes for classification of the model to predict churn behaviors of new customers and behavioral analysis of special current patients.

2.2 *methods*

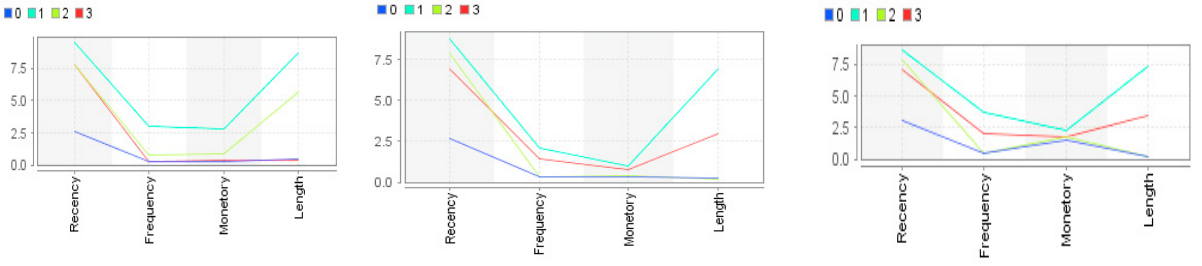
For identifying and comparison of different groups of patients in three clinics, k-means algorithm as a clustering technique was used. Also decision tree, naïve base and neural network as classification techniques were applied for the churn prediction and finding the behavioral tendency of current patients based on R, F, M, L attributes. RapidMiner data analysis software was used for preprocessing data records, attribute selection and finally applying proper algorithms, which is the world-leading open-source system for data mining. The result showed that all three classification algorithms had a similar performance on the data records.

2. Results

2.1 comparison of different groups of patients in three clinics: clustering model

At first, clustering techniques were used to identify and compare different groups of three clinics. Clustering is a good technique to create a general understanding, when we don't have much information about the study population. K-means algorithm (k=4 groups; loyal, potential for loyal, potential for churn and churn) was performed on normalized fields (R, F, M and L) of three clinics to group customers that are similar to each other. Normalization is a preprocessing technique used to rescale attributes values and fit them in a specific range. For a fair comparison, all attributes should have the same scale when the clustering algorithm uses numerical measure such Euclidean distance.

Plots and information of the groups in the Cardiovascular, Neurology and Pediatric Hematology clinics have been shown in Figure 1.



c. b. a.

Figure 1. Plots of clustering model on Cardiovascular clinic data (a), Neurology clinic data (b) and Pediatric Hematology data (c).

As represented in table 1, 2 ,3, 4, 5 and 6, average numbers of patients' visit, average of total cost paid, average of the length of time the patients has been in contact with us and average of patients' last visit in four categories (cluster 0, 1, 2 and 3) have been calculated. Then, customer life time values have been calculated for each cluster according to the following formula. Analytic hierarchy process (AHP) was used to determine the relative importance or weights of the RFML variables, WR, WF, WM and WL. We used the questioner and interviewed with administrative managers and medical directors of ST Hospital to determine efficient factors affecting on patient loyalty and weights of the extended model. According to the assessments, the relative weights of the RFML variables are 4.5, 8.3, 8.2 and 4.9 respectively (scale is 1-10).

$$CLV_{Ci} = NR_{ci} * W_R + NF_{ci} * W_F + NM_{ci} * W_M + NL_{ci} * W_L$$

Table 1. Average value of R, F, M and L attributes, CLV value and the numbers of members in each cluster in Cardiovascular clinic. Total members: 8979.

Cardiovascular clinic	R Avg value	F Avg value	M Avg value	L Avg value	CLV Avg value	Count	%
Cluster 0	3.007	0.426	1.449	0.148	29.682	4226	47%
Cluster 1	8.663	3.676	2.279	7.297	123.933	115	1.3%
Cluster 2	7.892	0.416	1.796	0.142	54.387	4397	49%
Cluster 3	7.088	2.059	1.709	3.455	79.922	241	2.7%

Table 2. Centroids of the clusters for each attribute in Cardiovascular clinic.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Recency	2.560	9.499	7.774	7.735
Frequency	0.206	3.011	0.789	0.228
Monetary	0.295	2.833	0.835	0.375
Length	0.436	8.711	5.649	0.350

Table 3. Average value of R, F, M and L attributes CLV value and the number of members in each cluster in Neurology clinic. Total members: 22210.

Neurology clinic	R Avg value	F Avg value	M Avg value	L Avg value	CLV Avg value	Count	%
Cluster 0	2.673	0.319	0.346	0.235	18.668	9995	45%
Cluster 1	8.803	2.117	0.993	6.928	99.277	632	2.9%
Cluster 2	7.895	0.319	0.375	0.208	42.265	10434	47%
Cluster 3	6.903	1.413	0.738	2.996	63.519	1149	5.1%

Table 4. Centroid of the clusters for each attribute in Neurology clinic.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Recency	2.672	8.802	7.894	6.903
Frequency	0.318	2.117	0.318	1.412
Monetary	0.346	0.993	0.375	0.737
Length	0.235	6.927	0.207	2.995

Table 5. Average value of R, F, M and L attributes CLV value and the number of members in each cluster in Pediatric Hematology clinic. Total members: 677.

Pediatric Hematology	R Avg value	F Avg value	M Avg value	L Avg value	CLV Avg value	Count	%
Cluster 0	2.560	0.206	0.296	0.436	17.796	183	27%
Cluster 1	9.500	3.011	2.834	8.712	133.663	106	15.6%
Cluster 2	7.775	0.789	0.835	5.649	76.065	148	21.9%
Cluster 3	7.736	0.229	0.376	0.351	41.510	240	35.5%

Table 6. Centroid of the clusters for each attribute in Pediatric Hematology clinic.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Recency	2.560	9.499	7.774	7.735
Frequency	0.206	3.011	0.789	0.228
Monetary	0.295	2.833	0.835	0.375
Length	0.436	8.711	5.649	0.350

The cluster with highest and lowest CLV considered as loyal group and churn customers respectively. In the Cardiovascular clinic, nearly half of patients (47%) were churned and only 1.3% was faithful customers who maintained their relationship with the hospital. These values in the Neurology clinic were 45%, 2.9% and in the Pediatric Hematology were 27% and 15.6% respectively. The Pediatric Hematology clinic performed better than both the Cardiovascular and Neurology clinics. This is because of the higher numbers of loyal customers and lower numbers of churn customers. In Table 7, percent of each cluster in three clinics have been compared.

Table7. Different categories of patients in three clinics.

	Loyal % - Avg CLV	Potential for loyal % - Avg CLV	Potential for churn % - Avg CLV	Churn % - Avg CLV
Cardiovascular clinic	1.3%	2.7%	49%	47%
Neurology clinic	2.9%	5.1%	47%	45%
Pediatric Hematology	15.6%	21.9%	35.5%	27%

Performance evaluation of clustering model was done by Avg. within centroid distance and Davies Bouldin criteria. In Table 8, the performance of the algorithm has been shown. The average within cluster distance was calculated by

averaging the distance between the centroid and all examples of a cluster. The less value is better. Also the algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index. The clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion (rapidminer.com).

Table8. Performance evaluation of clustering models with Avg. within centroid distance and Davies Bouldin criteria.

Performance Vector	Avg. within centroid distance	Davies Bouldin
Clustering model 1	-4.227	-1.096
Clustering model 2	-2.691	-0.962
Clustering model 3	-4.083	-0.811

2.2 Churn modeling/ churn customer prediction

There are different types of information about patients that can be used for churn prediction model such as socio-demographics and transactional information that are recorded in hospital HIS system. In this paper, we have used transactional data include Recency, Frequency, Monetary and Length information about patients. Here, customers that had not referred in the last eighteen months of period (the 3-years period), were considered as the lost customer or churn. In this way, we calculated churn field as class attributes for all patients and run decision tree, naïve Bayes and Neural network as classifiers. For more accurate modeling, all fields were discretizing to meaning groups (Table 9).

Table 9. The numerical attributes discretizes into user-specified classes for the fair classification. The selected numerical attributes will be changed to nominal attributes.

	range1	Count Clinic1	Count Clinic2	Count Clinic3	range2	Count Clinic1	Count Clinic2	Count Clinic3	range3	Count Clinic1	Count Clinic2	Count Clinic3	range4	Count Clinic1	Count Clinic2	Count Clinic3
Frequency	once	7731	14527	312	up to 5	842	6961	195	up to 15	378	703	108	more	28	19	62
Monetary	Range1	2269	7219	210	Range2	2231	5053	128	Range3	2394	4391	169	Range4	2085	5547	170
Recency	Range1	1872	5671	101	Range2	2855	6264	128	Range3	4252	10275	448				
Length	Week	8077	16215	320	6M	573	4479	83	1y	178	797	80	More	151	719	194
Age	Teenager	163	1728	293	Young	1860	9055	307	middle age	3903	7442	52	old	3053	3985	25
Churn	0	5474	13081	511	1	3505	9129	166								

Classification modeling helps us to identify behavioral tendencies of different customers. In this model, we can recognize trends of loyal and churn patients. More importantly, potential patients who can cut ties with the organization and also new customers can be identified. As regards, customer attraction several times cost than customer retention, specification of these customers will be very valuable. To detect behavioral tendencies of special

customers of the clinics, we supposed 4 categories of patients; churn (C), potential for churn (PC), loyal (L) and new customer (N). In the following, the results of classification modeling by decision tree ID3 (the precursor to the C4.5 algorithm) for three clinics have been presented.

Table 10. Behavior of main categories of patients based on result of decision tree ID3

	churn (C)	potential for churn (PC)	loyal (L)	new customer (N)
Cardiovascular clinic	Length= week, Frequency=once/up to 5, Monetary=range1/2/3/4, Recency=range1. --- Length= 6m, Frequency=up to 5, Monetary=range1/2/3/4, Recency=range1. --- Length= week, Frequency= up to 15, Monetary=range1/2/3, Recency=range2. --- Length= week, Frequency= up to 15, Recency=range1	Length= week, Frequency=once/up to 5, Monetary=range1/2/3/4, Recency=range2. --- Length= 6m, Frequency=up to 5, Monetary=range1/2/3/4, Recency=range2. --- Length= 6m, Frequency=up to 15, Monetary=range1/4 --- Length= 6m, Frequency=up to 15, Monetary=range2/3 Recency=range1/2. --- Length= 1y, Frequency=up to 5, Monetary=range1/4 Recency=range2 --- Length= week, Frequency= up to 15, Monetary=range4, Recency=range2. --- Length= week, Frequency= up to 15, Recency=range3	Length=more --- Length= 1y, Frequency=up to 15, Monetary=range1/4	Length= week, Frequency=once/up to 5, Monetary=range1/2/3/4, Recency=range3. --- Length= 6m, Frequency=up to 5, Monetary=range1/2/3/4, Recency=range3.
Neurology clinic	Length= week, Frequency=once/up to 5, Monetary=range1/2/3/4, Recency=range1. --- Length=6m, Frequency=up to 15, Monetary=range4, Recency=range1.	Length= week, Frequency=once, Monetary=range1/2/3/4, Recency=range2. --- Length=1y, Frequency=up to 15, Recency=range2. ---------- Length=6m, Frequency=up to 15, Monetary=range4, Recency=range2. --- Length=6m, Frequency=up to 5, Monetary=range3/4, Recency=range2.	Length=more	Length= week, Frequency=once/up to 5, Monetary=range1/2/3/4, Recency=range3. --- Length=1y, Frequency=up to 15, Recency=range3. --- Length=1y, Frequency=up to 5, Monetary=range4, Recency=range3. --- Length=6m, Frequency=up to 15, Monetary=range4, Recency=range3.

Pediatric Hematology	Frequency=once, Length= week, Recency=range1, Monetary=range1. --- Frequency=up to 5, Length=6m, Recency=range1, Monetary=range3.	Frequency=one, Length= week, Recency=range2, Monetary=range1/2. --- Frequency=up to 5, Length=6m, Recency=range2, Monetary=range3. Length=1y, Recency=range2, Monetary=range4.	Length=more, Monetary=range4/2. Frequency=up to 5, Length=more, Monetary=range3/2	Frequency=once, Length= week, Recency=range3, Monetary=range1/2. Length= 1y, Recency=range3, Monetary=range4. --- Frequency=up to 5, Length= 6m/1y, Recency=range3, Monetary=range3. Frequency=up to 5, Length=6m, Recency=range3, Monetary=range3.

Specific behaviors of four considered categories (churn, potential for churn, loyal and new customer) were extracted from decision tree rules (Table 10). It should be reminded that, all transactions of three years were considered as the period time of this research. Churn customers with low frequency, short period length with different amount of monetary were admitted in first year. Loyal customers were admitted in whole period for different services (different amount of monetary) and frequency more than one. The main character of loyal patients was that in the whole period, they were not disconnected to the hospital. New customer (N) and potential to churn (PC) groups were sensitive groups for organization, because we're not going to lose the attracted customers. PC customers with different frequency number, different amount of monetary and period length one year and less than one year were admitted in second year of whole period. As regards, the last visit was one year ago and these customers had not visited during the last year, maybe organization loses this group or not. Therefore, it is better to adopt marketing strategies to return these customers. N group with different frequency number, different amount of monetary and period length up to one year were admitted in third year of the whole period. This group was new patients who were new asset for the hospital. Hospital could attract this group with better services and appropriate behavior of hospital staffs. Table 11 shows accuracy of ID3 tree algorithm in three clinics. It worth noting, that accuracy of Naïve bays and Neural Network algorithms are near to this one. Because of meaningful rules produces by Decision tree algorithm, this algorithm was selected.

Table 11. Accuracy of execution of decision tree (ID3) classifications algorithm in three clinics.

	Precision	Recall	Accuracy	Classification error	AUC	Correlation
Cardiovascular clinic	88.51%	90.11%	89.25%	10.75%	0.970	0.786
Neurology clinic	89.565%	88.49%	88.83%	11.17%	0.963	0.80
Pediatric Hematology	89.57%	87.95%	94.53%	5.47%	0.988	0.873

3. Conclusion

Nowadays, hospitals in healthcare industry could implement CRM (customer relationship management) to improve relationship with their patients and take marketing strategies consequently to improve their profit. In this study we evaluated three clinics in ST public hospital in Tehran. At first, all patients were categorized to four groups. Then based on average of weighted CLV value of each category, four groups include loyal, potential for loyal, potential for churn and churn was identified. As depicted in table 7, Pediatric Hematology clinic despite of less customer numbers, had a better performance than that of other clinics. Because of more numbers of loyal customers (15.6 in comparison with 2.9% and 1.3%) and less number of churn customers (27% in comparison with 45% and 47%). This is also true that Neurology had a better performance than that of Cardiovascular clinic. There are several reasons for this. However, managers of Neurology and Cardiovascular clinics with high rate of churn and potential *for churn* customers must take policies to improve the efficiency of customer retention campaigns and to reduce the costs associated with churn. In the second part of the paper, behavioral tendencies of special group of patients were evaluated by churn modeling. ID3 decision tree classifier with high accuracy predicted churn behavior of the customers. In this way characteristic of PC, N and C customers were considerable. Hospital managers can predict current status of patients and prevent churn of them.

4. References

Al-Turaiki, I., Alshahrani, M., Almutairi, T., 2016, Building predictive models for MERS-CoVinfections using data mining techniques, Journal of Infection and Public Health 9, 744-748.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., Huang, K., 2017, Customer churn prediction in the telecommunication sector using a rough set approach, Neuro computing, 237 242–254

Bult, J.R., Wansbeek , T., 1995, Optimal selection for direct mail, Marketing Science 14, 378-395.

Ferreira D, Oliveira A, Freitas A., 2012, Applying data mining techniques to improve diagnosis in neonatal jaundice. BMC Medical Informatics and Decision Making , 12, 143.

Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S., 2011, Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study, Procedia Computer Science 3, 57–63

Obenshain, M. K., MAT, 2004, Application of Data Mining Techniques to Healthcare Data, Infection Control and Hospital Epidemiology, 25, 8, 690-695

Ramanan, S., Cruz de Souza, L., Moreau, N., Sarazin, M., Teixeira, A. L., Allen, Z., Guimaraes, H.C., 2017, Determinants of theory of mind performance in Alzheimer's disease: A data-mining study, cortex 88 8-18

rapidminer.com, http://docs.rapidminer.com/studio/operators/validation/performance/segmentation/cluster_distance_performance.html

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B., 2012, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, European Journal of Operational Research ,218, 211–229.