Customer Churn Prediction Using a Meta-Classifier Approach; A Case Study of Iranian Banking Industry

Davoud Gholamiangonabadi
Department of Industrial Engineering and Management Systems
Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
dgholamian@gmail.com

Sanaz Nakhodchi
Department of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
S.nakhodchi@gmail.com

Ammar Jalalimanesh
Iranian Research Institute for Information Science and Technology (IRANDOC),
Tehran, Iran
jalalimanesh@irandoc.ac.ir

Adele Shahi
Department of Industrial Engineering and Management Systems
Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
Shahi.adele@gmail.com

Abstract

In this paper, a new approach is presented to identify customers churn in the banking industry. The purpose of this study is to increase the accuracy of customer churn identification. In order to predict it, the neural network methods of multilayer perceptron, radial basis function, support vector machine, and generalized regression are used. Then, the accuracy is increased by using a Naïve Bayes as a meta-classifier. The results show that this approach has led to a significant improvement in the prediction of customers churn. In addition, we figured out that if classifier techniques can achieve good results, the meta-classifier will boost the accuracy considerably.

Keywords

Customer churn , Data Mining, MultiLayer Perceptron Neural Network, Radial Basis Function Neural Network, Generalized Regression Neural Network, Support Vector Machine, Naïve Bayes

1. Introduction and Literature Review

A customer churn is someone who stop using a products or services of organization and use the products and services of other competitors.
Many researches have been done on the importance and necessity of predicting customer behavior, and their churn in particular. For example:

- The research (Chu et al., 2007) states that the cost of absorbing a new customer is 5 to 10 times more than keeping the older ones.

© IEOM Society International

- If customer retention rate increases by 5%, the organization's profit will increase from 25% to 85% (Feinberg and Trotter, 2001). According to the results, the cost of obtaining and attracting new customers is more than five times the cost of maintaining old customers (Slater and Narver, 2000; Colgate and Danaher, 2000).
- Older customers buy more, take less time of the organization, less sensitive to price fluctuation, and bring new customers to the organization (Ganesh et al., 2000).

In particular in the banks and financial institutions

- A bank is able to increase its profits up to 85% with a 5% improvement in its customers churn (Reichheld and Sasser, 1990).
- Financial effect of a 1% increase in the customer retention rate has been calculated and it is revealed that the company's benefit could be increased (Van den Poel and Lariviere, 2004).
- Generally, older customers spend more time with the bank and have a longer life cycle and more value for the bank (Benoit and Van den Poel, 2009).

Above aforementioned, attracting new customers for each organization is more costly than maintaining current customers (Reinartz and Kumar, 2003). If the causes of customer churn can be identified, strategies and plans for keeping such customers can be developed. The first step is to identify a customer churn. Hence, the model of customer churn prediction is important. A brief literature review has presented in the Table 1.

Table 1. A brief literature review on customer churn

Paper	Year	Scope	Description
(Wei and Chiu, 2002)	2002	Telecom Calls	Use the decision tree method and consider the customer details impact
(Meer, 2006)	2006	Financial institution	Use the Click flow analysis Online activities have an important role in predicting customer behavior
(Chu et al., 2007)	2007	Telecommunication	Using the C5 tree method and clustering with the GHSOM method Preparing policies for each cluster
(Burez and Van den Poel, 2007)	2007	TV	Using the regression method Developing incentive strategies Increase profits by using customized strategies
(Coussement and Van den Poel, 2008)	2008	Newspaper subscribers	Using support vector machines, logistic regression and random forest The support vector machine has a better accuracy The process of optimizing the input variables has a significant impact on increasing accuracy
(Xia and Jin, 2008)	2008	Telecommunication	Using support vector machines, C4.5, logistic regression, and Bayesian networks The support vector machine has a higher accuracy
(Tsai and Lu, 2009)	2009	Telecommunication	Using two models, better achievement in using two MLP neural networks simultaneously than the SOM and MLP neural networks
(Xie et al., 2009)	2009	Bank	Use of modified randomized forest algorithms and normalized data This method has improved accuracy compared to the multilayer perceptron neural network, decision tree, and support vector machine methods
(Karahoca et al., 2009)	2009	Mobile operator	Using decision tree, Bayesian network and ANFIS The ANFIS algorithm has improved the results and accuracy
(Hosseini et al., 2010)	2010	SAPCO company to supply vehicle parts	Clustering with k-means method

© IEOM Society International

(Huang et al., 2010)	2010	Telecommunication	Selection of variables with multi-objective optimization approach Using better input variables leads to better and accurate results
(Tsai and Chen, 2010)	2010	Multimedia company	Using association rules to get the most important variables Using decision tree and neural network to predict The use of association rules has made better results
(Abbasimehr et al., 2011)	2011	Telecommunication	Using ANFIS, C 4.5 and RIPER methods Clustering with using Fuzzy C-means Higher precision and less rules with the ANFIS method
(Kisioglu and Topcu, 2011)	2011	Telecommunication	Using the Bayesian network method and creating three scenarios Making discrete data with the CHAID method
(Keramati and Ardabili, 2011)	2011	Mobile services	Use of logistic regression method and new parameters in addition to literature
(Nie et al., 2011)	2011	Credit Cards	Using logistic regression and decision tree and considering 135 variables Logistic regression has better results than the decision tree
(Benoit and Van den Poel, 2012)	2012	Financial institution	Using social network variables, using kinship network approach and predicting with random forest algorithm
(Miguéis et al., 2013)	2013	retail	Using logistic regression and multivariate regression Multivariate regression has better results
(Abbasimehr et al., 2013)	2013	Telecommunication	Identifying high value customers, and clustering them with using k-means algorithm Predicting with using neural fuzzy inference systems, local linear fuzzy systems, multilayer perceptron neural network, and radial basis function neural network

2. Methodology

In this research, using neural networks, a new approach has been presented to study and predict the customer churn. As is views in the Figure 1, firstly pre-process is performed; next, clustering and finally, using neural networks and considering cost function, a solution is found the problem of customer churn.

2.1. Min-Max method

This method applies a linear transformation on a set of Continuous data. The goal of this, is to increase precision in the next phases. Assume that x_{min} and x_{max} are the minimum and the maximum of an attribute j respectively. Also, x'_{min} and x'_{max} are the new minimum and maximum for this attribute; then this transformation is conducted using the equation (1).

$$x'_{ij} = \frac{x_{ij}-x_{min,j}}{x_{max,j}-x_{min,j}}(x'_{max,j} - x'_{min,j}) + x'_{min,j} \quad (1)$$

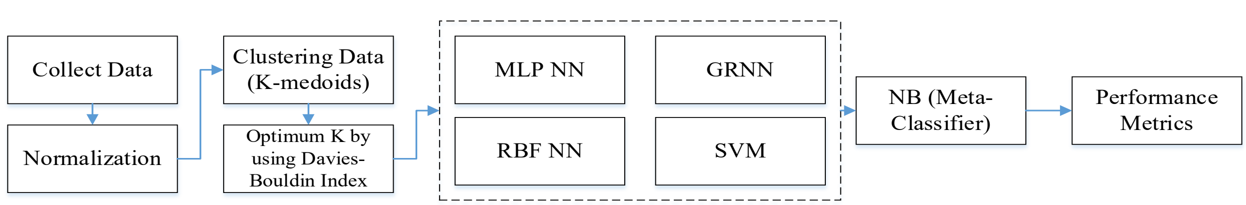


Fig 1. Proposed Methodology

2.2. K-medoids

© IEOM Society International

The K-medoids algorithm is well-known for its partitioning around medoids, is one of the expanded algorithms form K-mean. This method was proposed in 1987 (Kaufman and Rousseeuw, 1987). The purpose of it is reducing the sensitivity of the values generated using the mean of the cases, it uses the approximation of medoids as its suggested outcome.

• X_k is assigned to the cluster associated with the medoid u_h and we have equation (2).

$$dist(x_i, x_k) \leq \min_{u_e \in U, e \neq h} dist(u_e, x_k) \tag{2}$$

In such condition, the observation x_k- is chosen as the new medoid of x_i to represent the cluster C_h. The contribution of the substitution may be positive, negative or zero and is calculated using the equation (3).

$$R_{ihk} = dist(x_i, x_k) - dist(u_h, x_k) \tag{3}$$

• X_k is currently assigned to the cluster associated with the medoid u_h and we have equation (4).

$$dist(x_i, x_k) \leq \min_{u_e \in U, e \neq h} dist(u_e, x_k) \tag{4}$$

Now, the observation xi is assigned to another cluster and the contribution of this substitution is equation (5).

$$R_{ihk} = \min_{u_e \in U, e \neq h} dist(u_e, x_k) - dist(u_h, x_k) \tag{5}$$

• X_i is not yet assigned to the cluster associated with the medoid u_h and we have equation (6).

$$dist(x_i, x_k) \geq \min_{u_e \in U, e \neq h} dist(u_e, x_k) \tag{6}$$

• X_k is not yet assigned to the cluster associated with the medoid u_h and we have equation (7).

$$dist(x_i, x_k) \leq \min_{u_e \in U, e \neq h} dist(u_e, x_k) \tag{7}$$

In such condition, the observation x_k is to be assigned to the cluster C_h and the substitution contribution is equation (8).

$$R_{ihk} = dist(x_i, x_k) - \min_{u_e \in U, e \neq h} dist(u_e, x_k) \tag{8}$$

$$T_{ih} = \sum_{x_k \notin U} R_{ihk} \tag{9}$$

2.3. Davies-Bouldin Index

This criterion was presented in 1979 to assess clustering algorithms (Davies and Bouldin, 1979). If X_j is assigned to the cluster C_j and the respective cluster center is shown by A_i then, S_i calculates the scatter in the cluster using the equation (10).

$$S_i = (\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q)^{\frac{1}{q}} \tag{10}$$

Now, if the separation between the clusters i and j (C_i and C_j) is called M, we have:

$$M_{i,j} = ||A_i - A_j||_p \tag{11}$$

If $M_{i,j}$ is defined as how good the clustering scheme, then we have:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{12}$$

$$D_i = max_{j;i \neq j} R_{i,j} \tag{13}$$

And if N is the number of the clusters, we have: The Davies-Bouldin criterion is the average value for all the clusters; in other words it equals:

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{14}$$

© IEOM Society International

2.4. Neural Networks

Artificial neural network is an adaption from the biological neural system, trying to mimic the process the data like human brain. The key element of this method is the new structure of data processing system. This system is consisted of large number of highly interconnected processing elements which act consistently for solving a problem. Alike humans brain, artificial neural networks, can be trained with examples. An artificial neural network is adjusted for doing a special task like detecting rules and categories in a period of learning process. In biological systems, learning is associated with adjusting the connections of synapsis which is located between nerves. Artificial neural networks exploit the same method, with their significant capacity to deduce meaning of complicated and vague data can extract the rules and methods which is difficult and elaborate for humans and other computer techniques to discover. It is used in different fields such as Face Recognition (Azami et al., 2013), Diabetes Diagnosis (Fiuzy et al., 2013), Bankruptcy Prediction (Bagheri et al., 2012), prediction changes in stock (Ghezelbash, 2012;Gholamiangonabadi et al., 2014), and prediction quality in devices (Gholamiangonabadi et al., 2015).

2.4.1. General Regression Neural Network

GRNN is one of the neural network type presented in 1991 (Specht, 1999). This is one of the radial basis neural networks. The advantage of this method is it can be used to train the network when we lack data. In addition, to train a network using this method, an iterative training procedure is followed instead of back propagation neural network. This network is able to approximate any given function between the input and the output.

A GRNN is consisted of four layers: the input layer, pattern layer, summation layer and the output layer. Assuming there are q neurons as the input layer- that equal the number of the input parameters, the output for this layer is considered as the input for the pattern layer where p neurons are designed. The output of the pattern layer is entered to the summation layer where two neurons named Denonminator and Numrator are considered. Clearly, each neuron in the pattern layer is connected to the two abovementioned neurons (S, D). The neuron S calculates the summation of weighted response associated with the pattern layer and the neuron D does the same for the un-weighted outputs. The output layer and the summation layer, together normalize the output set. To train such network, Radial Basis Function or Linear Basis Function can be used.

GRNN is widely used in detecting cancer, diabetes and heart disease, also in fraud detection. A short review of the calculations done in GRNN is presented below:

$$Y(x) = \frac{\sum_{i=1}^{n} y_i.exp(-D(x,x_i))}{\sum_{i=1}^{n} exp(-D(x,x_i))} \quad (15)$$

$$D(x,x_i) = \sum_{k=1}^{m} (\frac{x_i - x_{ik}}{\sigma})^2 \quad (16)$$

Where y_i is the weight connection between the ith neuron in the pattern layer and the S-summation neuron, n is the number of training patterns, D is the Gaussian function, m is the number of elements of the input vector, x_k and x_{ik} are the j −th element of x and x_i, respectively, r is the spread parameter, whose optimal value is determined experimentally.

2.4.2. Radial Basis Function Neural Network

Radial Basis Neural networks were presented by different researchers. The input neurons bear no weight, thus the first hidden layer receives the exact same values as the first layer. The function designed in the hidden layer are the Radial Basis type. The transfer function for the neurons of the hidden layer are non-monotonic. Then the output of these neurons are sent to the output layer by weights. The neurons of the output layers are actually, simple summations. Let us assume that there are H neurons in the hidden layer. The transfer function are mostly like Gaussian Density Functions. If this function is Gaussian:

$$a_{h,k} = exp(-\frac{||\hat{x}_h - x_k||^2}{\sigma_h^2}) \quad (17)$$

In which $a_{h,k}$ is the output of the hth neuron in the hidden layer. Also, $\hat{x}_h$ is the center of the radial function and is the distance scaling parameter which determines over what distance in the input space the unit will have a significant influence. Finally, the weighted average of the outputs associated with the hidden layer determines the output. In other words, the equation (18) shows this output value.

$$y_i = \sum_{i=1}^{n} w_i \times a_{h,i} \tag{18}$$

In which the w_i, k is the weight assigned to the neuron i −th in the hidden layer and the k −th neuron in the output layer. As this method is an observer learning method, the exact values for x_i and y_i are predetermined. Thus to have the weights in the second layer, in this research, the pseudo-inverse method is used, in which:

$$G = [\{g_{i,j}\}] \tag{19}$$

$$g_{i,j} = exp(\frac{-||x_i - v_j||}{2\sigma_j^2}) \quad i = 1,2,\dots,n; j = 1,2,\dots,p \tag{20}$$

And we have:

$$D = GW \tag{21}$$

Where D is the desired output for the trained data.

If G^{-1} exists, then we have:

$$W = G^{-1}D \tag{22}$$

If G is ill-conditioned (close to singularity) or is a non-square matrix, then:

$$W = G^{+}D \tag{23}$$

$$G^{+} = (G^T G)^{-1} \times G^T \tag{24}$$

2.4.3. Support Vector Machine

Support Vector Machine selects a number of observations as the representative of a certain class (Vercellis, 2011). These observations, determine the separation process in the classification of the feature space. If the space is linear, there are infinite numbers of lines and planes that can separate different classes. The optimized separation line, is a line that bear the best level of expansion, and the amount of error. The separation margin, here is twice the size of the distance between the trained data and separating hyper-plane. In addition, they are the support vectors that have the least distance from the separating hyper-plane. To determine these hyper-planes the pattern below is followed:

• If w is the coefficient vector associated with the hyper-plane and b is the bias, then the separating hyper-plane is:

$$w'x = b \tag{25}$$

• The two supporting focal hyper-planes are:

$$w'x - b - 1 = 0, \quad w'x - b + 1 = 0 \tag{26}$$

In which the separation margin is:

$$\delta = \frac{2}{||w||'}, \quad ||w|| = \sqrt{\sum_{j\in N} w_j^2} \tag{27}$$

• To determine w and b, an quadratic optimization problem with linear constraints is to be solved:

$$max_{w,b} \frac{1}{2}||w||^2 s.t. y_i(w'x_i - b) \geq 1, \quad i \in M \tag{28}$$

• The objective function seeks to maximize the separation margin by minimizing the inverse, and the constraints have each x_i stay at the associated class y_i.

• Thus the objective function and the constrains are:

$$min_{w,b,d} \frac{1}{2}||w||^2 + \lambda \sum_{i=1}^{m} d_i s.t. y_i(w'x_i - b) \geq 1 - d_i, \quad i \in M d_i \geq 0, i \in M \tag{29}$$

• The abovementioned optimization problem can be solved using Lagrangian duality:

$$L(w,b,d,\alpha,\mu)=\frac{1}{2}||w||^2+\lambda\sum_{i=1}^{m} d_i-\sum_{i=1}^{m} \alpha_i[y_i(w'x_i-b)-1+d_i]-\sum_{i=1}^{m} \mu_i d_i y_i(w'x_i-b)\geq 1-d_i,\quad i\in M d_i\geq 0, i\in M\alpha_i,\mu_i\geq 0 \tag{30}$$

To find the optimized solution, the partial deviations to b, d and w must equal zero. By the placement of the calculated values in the dual objective function and by applying the Kuhn-Tucker's conditions on the problem and the dual, we have:

$$\alpha_i[y_i(w'-b)-1+d_i]=0,\quad i\in M\mu_i(\alpha_i-\lambda)=0,\quad i\in M \tag{31}$$

Where (31) is to detect the support vectors. In this case, each new observation of x is classified as equation (32):

$$f(x)=sgn(\sum_{i=1}^{m} \alpha_i y_i x_i' x_i+b) \tag{32}$$

If the data are not linearly separable, the features can be transferred to a new space to make them separable by a line. In other words, if the function is the transfer function for the data from the non-linear space to a linear one, then in the Lagrangian duality would be:

$$L(w,b,d,\alpha,\mu)=\sum_{i=1}^{m} \alpha_i-\frac{1}{2}\sum_{i=1}^{m} sum_{h=1}^{m} y_i y_h \alpha_i \alpha_h K(x_i,x_h) s.t. \sum_i \alpha_i y_i=00\leq\alpha_i\leq C \tag{33}$$

Where $K(x_i,x_h)=\phi(x_i)^T\phi(x_h)$. Thus, to classify each new observation:

$$f(x)=sgn(\sum_{i=1}^{m} \alpha_i y_i K(x_i,x+b))b=\frac{1}{|S|}\sum_i [y_i-\sum_h \alpha_j y_h K(x_i,x_h)] \tag{34}$$

2.4.4. Multilayer Perception

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs (Vercellis, 2011). A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. To train the network, Levenberg-Marquardt method is used that will be explained in the following: The training algorithm includes five steps.

step 1: Initialize weights and thresholds to small random values.
step 2: Choose an input-output pattern $(x^{(k)},t^{(k)})$ from the training data.
step 3: Compute the network's actual output $o^{(k)}=f(\sum_{i=1}^{l} w_i x_i^{(k)}-\theta)$. ($l$ is size of input vector or the size on input neurons). Adjust the weights and bias according to the Levenberg-Marquardt algorithm.
step 4: If whole epoch is complete, then pass to the following step; otherwise go to step 2.
step 5: If the weights (and bias) reached steady state ($\Delta w_i\approx 0$) through the whole epoch, then stop the learning; otherwise go to through one more epoch starting from.

2.4.5. Naïve Bayes

One of the probabilistic models is Bayesian method (Vercellis, 2011). Based on Bayesian theory, these methods calculate the posterior probability P (y | x) and determine the target class of x. It is assumed that the prior probability P (y) and the conditional probabilities of the class P (x | y) are known. Goal of Bayes categories are calculating probability of P (y | x). Therefore, the learning phase of a Bayesian classification is the analysis of the learned data and based on the probability values needed to perform the classification are estimated. Assume x is a record of the learning data and its target variable, y, can get H distinct values as $H=\{v_1,v_2,...,v_H\}$. Bayes theory is used to compute the posterior probability P (y | x) or the probability of observing the target class y if x is observed.

$$P(y|x)=\frac{P(x|y)P(y)}{\sum_{l=1}^{H} P(x|y)P(y)}=\frac{P(x|y)P(y)}{P(x)} \tag{35}$$

In order to categorize the new observation of x, the Bayes classification use the principle of posterior maximum, which calculates the posterior probability P (y | x) with relation (35) and Allocates observed x to the class with the highest value of P (y | x).

© IEOM Society International

$$y_{MAP} = arg \max_{y \in H} P(y|x) = arg \max_{y \in H} \frac{P(x|y)P(y)}{P(x)} \quad (36)$$

Since the denominator P (x) is independent of y, the posterior probability would be maximized of the numerator of the relation (36). So, the x record is assigned to the v_h class if and only if:

$$P(x|y = v_h)P(y = v_h) \geq P(x|y = v_l)P(y = v_l),$$
$$l = 1,2,\dots,H \quad (37)$$

The prior probability P (y) can be obtained by calculating the frequency of the class v_h, m_h, in the set D as the equation (37):

$$P(y = v_h) = \frac{m_h}{m}, \qquad h = 1,2,\dots,H \quad (38)$$

If the size of the sample is large, the estimation of relation (38) will be sufficiently precise for the prior probabilities.

3. Empirical Analysis

3.1. Data

Data selection has been made base on the literature review and the accessibility. To measure the customer churn, the data associated with 860 customers is gathered from Iranian banks. The input variables in this model are age, gender, residence, income, marital status, number of children, ownership of a car, ownership of a saving account, ownership of a current account and the mortgage status. The output variable is defined as retention of credit card. The data used in this research is gathered from an anonymous Iranian bank between February 2013 and June 2013. A short review of the data is shown in the Table 2.

Table 2. Desciriptive Data

Variables	Min	Max	Mean	standard deviation
Input				
Age	18	67	42.395	14.424
Gender	0	1	0.5	0.500
Region	0	3	1.37	1.008
Income*	830	63130	3215	1372
Married	0	1	0.66	0.474
Children	0	3	1.01	1.05
Car	0	1	0.49	0.5
Save-account	0	1	0.69	0.462
Current-account	0	1	0.758	0.482
Mortgage	0	1	0.348	0.476
Output				
usage	0	1	0.46	0.49

*: 1000 Rials

As can be seen in the Table 2, most of the observation is consisted of the young people. In addition, the income is not considered high for most of the cases. Most of these customers own current and saving account and nearly one third were in debt to the bank in form of mortgages.

3.2. Results

According to the methodology in the second part, as it did not consist any missing data, and as the data were balanced- 45 percent were loyal, and 54 percent churning- first the data is normalized. After that part, the clustering has been done by k-medoids method. As can be seen in the Figure 2, one of the best clustering conducted was with the number of cluster 9. Thus this is chosen as the best number of clusters for the given set of data. Moreover, it's used to define the center and the width of the RBF neural network. Once the optimized clustering is done, using GRNN, SVM, MLP

NN and RBF prediction models of the customer churn has been built. To monitor the performance of the built models, sensitivity and specificity criteria were calculated using the equations (39) and (40), respectively. To analyze the precision of the classification the equation (41) was used:

$$Sensitivity = \frac{TP}{TP+FN}(\%) \tag{39}$$

$$Specificity = \frac{TN}{TN+FP}(\%) \tag{40}$$

$$Accuracy = \frac{\sum_{k=1}^{|C|} assess(c_k)}{|C|} \qquad c_k \in C \tag{41}$$

In which:

True positive (TP): Correctly identified
True Negative (TN): Correctly rejected
False Positive (FP): Incorrectly identified
False Negative (FN): Incorrectly rejected

3.3. Validation

To validate the model, 10-fold cross validation method is applied. In this method, firstly, the data is divided into ten equal parts, then the network is trained and tested for ten times. For example, for the first training of the network, the first nine part of the data is considered as the training and the last part as the test. For the second run, the first eight part and the 10th part of the data is considered as the training set and the ninth part for the test and so on.

As can be seen in the Table 3. associated with the GRNN, the network provide 74.2 percent precision. Using this method, the prediction of customer churn, compared to the prediction of the loyal customers show insignificant higher level of precision. Considering the Table 3 associated with the RBF Neural Network, show 83 percent precision. Again, the customer churn prediction is more precise compared to loyal customer prediction. The SVM results is reflected in Table 3 showing 85 percent precision in the predictions. Again the value associated with the prediction of customer churn is higher than the loyal customers. As can be observed in the Table 3 associated with the MLP Network, the average of the prediction related to customer churn and loyal customers. As can be observed in the Table 3 associated with the Naïve Bayes, the average of the prediction related to customer churn and loyal customers. Based on the Table 4 and Figure 2 it can be observed that the MLP and SVM Networks have shown better performance respectively than RBF and GRNN. As can be seen, Naive Bayes as a meta-classifier has the highest values among other methods in all three criteria namely, Specificity, sensitivity and accuracy.

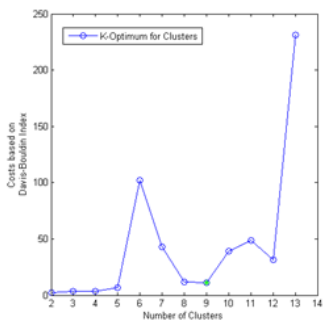
Table 3. Confusion Matrix From All Methods

Method		Predicted				
		Good/Bad		Sum	Percentage Correct (%)	Error (%)
	Observed	Good	Bad			
GRNN	Good	293	103	396	73.98	26.02
	Bad	118	346	464	74.57	25.43
	Sum	411	449	860	74.3	25.7
RBF NN	Good	317	79	396	80	20
	Bad	65	399	464	85.99	14.01
	Sum	382	478	860	83.25	16.75
SVM	Good	325	71	396	82.07	17.93
	Bad	53	411	464	88.57	11.43
	Sum	378	482	860	85.58	14.42
MLP NN	Good	337	59	396	85.1	14.9
	Bad	40	424	464	91.38	8.62
	Sum	377	483	860	88.48	11.52
Naïve Bayes (Meta-Classifier)	Good	368	28	396	92.9	7.1
	Bad	20	444	464	95.7	4.3
	Sum	377	483	860	94.4	5.6

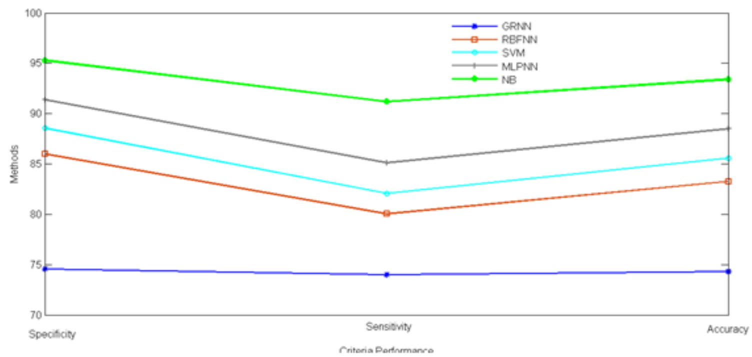
Table 4. Investigating Different Criteria for Different Methods

© IEOM Society International

Methods	Specificity	Sensitivity	Accuracy
GRNN	74.56	73.98	74.3
RBF NN	85.99	80.05	83.25
SVM	88.57	82.07	85.58
MLP NN	91.38	85.1	88.48
NB	95.7	92.9	94.4



Determine best K by Consider Davis-Bouldin Index



Investigating Different Criteria for Different Methods

Fig 2. Determine best K by Consider Davies-Bouldin Index and Investigating Different Criteria for Different Methods

4. CONCLUSION

Nowadays, Customer Relationship Management is one of the most important managerial concepts. The core of CRM is customer churn, and customer retention. Using data mining it is possible to recognize the hidden patterns of the data. The customer churn, also prediction the leaving customers pose huge amount of cost on the organization. In this research a new approach is presented to analyze the customer churn in Iranian bank. To form a more precise model, a meta-classifier method was suggested. And the results indicate the NB as a meta-classifier has well performance compared to MLP, SVM, RBF and GRNN in making the customer churn prediction.

REFERENCES

Abbasimehr, Hossein, Mostafa Setak, and M. J. Tarokh. "A neuro-fuzzy classifier for customer churn prediction." *Int J Comput Appl* 19, no. 8 (2011): 35-41.

Abbasimehr, Hossein, Mostafa Setak, and Javad Soroor. "A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques." *International Journal of Production Research* 51, no. 4 (2013): 1279-1294.

Azami, Hamed, Milad Malekzadeh, and Saeid Sanei. "A new neural network approach for face recognition based on conjugate gradient algorithms and principal component analysis." *Journal of mathematics and computer Science* 6, no. 3 (2013): 166-175.

Bagheri, Mahnaz, M. Valipour, and V. Amin. "The Bankruptcy Prediction in Tehran share holding using Neural Network and it's Comparison with Logistic Regression." *The Journal of mathematics and computer science* 5, no. 3 (2012): 219-228.

Benoit, Dries F., and Dirk Van den Poel. "Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services." *Expert Systems with Applications* 36, no. 7 (2009): 10475-10484.

Benoit, Dries F., and Dirk Van den Poel. "Improving customer retention in financial services using kinship network information." *Expert Systems with Applications* 39, no. 13 (2012): 11435-11442.

© IEOM Society International

Burez, Jonathan, and Dirk Van den Poel. "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services." *Expert Systems with Applications* 32, no. 2 (2007): 277-288.

Chu, Bong-Horng, Ming-Shian Tsai, and Cheng-Seen Ho. "Toward a hybrid data mining model for customer retention." *Knowledge-Based Systems* 20, no. 8 (2007): 703-718.

Colgate, Mark R., and Peter J. Danaher. "Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution." *Journal of the academy of marketing science* 28, no. 3 (2000): 375-387.

Coussement, Kristof, and Dirk Van den Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." *Expert systems with applications* 34, no. 1 (2008): 313-327.

Davies, David L., and Donald W. Bouldin. "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence* 2 (1979): 224-227.

Feinberg, Richard, and Mike Trotter. "Immaculate deception: the unintended negative effects of the CRM revolution: maybe we would be better off without customer relations management." *Defying the limits* 2 (2001).

Fiuzy, Javad Haddadnia Hadi Varharam Mohammad, Azam Qarehkhani, J. Haddadnia, J. Vahidi, and H. Varharam. "Introduction of a method to diabetes diagnosis according to optimum rules in fuzzy systems based on combination of data mining algorithm (dt), evolutionary algorithms (aco) and artificial neural networks (nn)." *The Journal of Mathematics and Computer Science (JMCS)* 6, no. 4 (2013): 272-285.

Ganesh, Jaishankar, Mark J. Arnold, and Kristy E. Reynolds. "Understanding the customer base of service providers: an examination of the differences between switchers and stayers." *Journal of marketing* 64, no. 3 (2000): 65-87.

Ghezelbash, Ali. "Predicting changes in stock index and gold prices to neural network approach." *The Journal of mathematics and computer science* 4, no. 2 (2012): 227-236.

Gholamiangonabadi, Davoud, Seyed Danial Mohseni Taheri, Afshin Mohammadi, and Mohammad Bagher Menhaj. "Investigating the performance of technical indicators in electrical industry in Tehran's Stock Exchange using hybrid methods of SRA, PCA and Neural Networks." In *Thermal Power Plants (CTPP), 2014 5th Conference on*, pp. 75-82. IEEE, 2014.

GholamianGonabadi, Davoud, Seyed Mohamad Hosseinioun, Jamal Shahrabi, and Mohammad AliMoradi. "Investigating Performance and Quality in Electronic Industry via Data Mining Techniques." *International Journal of Data Mining Techniques and Applications*, Volume: 04 Issue: 02 December 2015, Page No. 20-24.

Hosseini, Seyed Mohammad Seyed, Anahita Maleki, and Mohammad Reza Gholamian. "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty." *Expert Systems with Applications* 37, no. 7 (2010): 5259-5264.

Huang, Bingquan, Brian Buckley, and T-M. Kechadi. "Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications." *Expert Systems with Applications* 37, no. 5 (2010): 3638-3646.

Karahoca, Adem, Dilek Karahoca, and Nizamettin Aydın. "Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference System in GSM Churn Management." In *Data Mining and Knowledge Discovery in Real Life Applications*. InTech, 2009.

Kaufman, Leonard, and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

Keramati, Abbas, and Seyed MS Ardabili. "Churn analysis for an Iranian mobile operator." *Telecommunications Policy* 35, no. 4 (2011): 344-356.

Kisioglu, Pınar, and Y. Ilker Topcu. "Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey." *Expert Systems with Applications* 38, no. 6 (2011): 7151-7157.

Meer, Geoffrey Van. "Customer development and retention on a Web-banking site." *Journal of Interactive Marketing* 20, no. 1 (2006): 58-64.

Miguéis, Vera L., Ana Camanho, and João Falcão e Cunha. "Customer attrition in retailing: an application of multivariate adaptive regression splines." *Expert Systems with Applications* 40, no. 16 (2013): 6225-6232.

Nie, Guangli, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. "Credit card churn forecasting by logistic regression and decision tree." *Expert Systems with Applications* 38, no. 12 (2011): 15273-15285.

Reichheld, Frederick F., and Jr WE Sasser. "Zero defections: Quality comes to services." *Harvard business review* 68, no. 5 (1990): 105-111.

© IEOM Society International

Reinartz, Werner J., and Vita Kumar. "The impact of customer relationship characteristics on profitable lifetime duration." *Journal of marketing* 67, no. 1 (2003): 77-99.

Slater, Stanley F., and John C. Narver. "Intelligence generation and superior customer value." *Journal of the academy of marketing science* 28, no. 1 (2000): 120.

Specht, Donald F. "A general regression neural network." *IEEE transactions on neural networks* 2, no. 6 (1991): 568-576.

Tsai, Chih-Fong, and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." *Expert Systems with Applications*36, no. 10 (2009): 12547-12553.

Tsai, Chih-Fong, and Mao-Yuan Chen. "Variable selection by association rules for customer churn prediction of multimedia on demand." *Expert Systems with Applications* 37, no. 3 (2010): 2006-2015.

Van den Poel, Dirk, and Bart Lariviere. "Customer attrition analysis for financial services using proportional hazard models." *European journal of operational research* 157, no. 1 (2004): 196-217.

Vercellis, Carlo. *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons, 2011.

Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." *Expert systems with applications* 23, no. 2 (2002): 103-112.

Xia, Guo-en, and Wei-dong Jin. "Model of customer churn prediction on support vector machine." *Systems Engineering-Theory & Practice* 28, no. 1 (2008): 71-77.

Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying. "Customer churn prediction using improved balanced random forests." *Expert Systems with Applications* 36, no. 3 (2009): 5445-5449.

Davoud Gholamiangonabadi received his Master degree in Industrial Engineering from Amirkabir University of Technology (aka Tehran Polytechnique). He has broad interdisciplinary research interests covering topics in Big Data, Data Mining, Machine Learning, Meta-Heuristic Algorithms, Image Processing, and Artificial Neural Networks. He is considered an active member of the data mining research community and has published more than ten articles and conference posters.

Sanaz Nakhodchi received her B.Sc degree in Information technology from Islamic Azad University, Mashhad, Iran, 2010. She also has 4 years experience (2014-2018) at CERT (computer emergency response team) Lab in Ferdowsi University of Mashhad as a research assistant. She has joined the University of Guelph as a master student in computer science and a member of CyberScience Lab in 2018. Her research interests are Cloud Computing, Artificial Intelligence, IOT, Security and threat hunting. Her e-mail address is: nakhodcs@uoguelph.ca

Ammar Jalalimanesh received his BSc in industrial engineering from Buali-Sina University Hamedan, in 2004. He got his MSc in industrial engineering from KNT University of Technology, Tehran, in 2007. He has a PhD in system engineering from Amirkabir University of Technology (Tehran polytechnic). In 2015, during his PhD, he won a scholarship from DAAD, Germany and worked as a research scholar at Technical University of Munich (TUM) for one year. Currently, he is an assistant professor of information systems research group at IranDoc. His research interests include complex system modeling and agent-based simulation, computational biology, evolutionary optimization and soft computing. He has published more than 20 papers in scientific Journals and conferences.

Adele Shahi received her BSc in Statistical from Isfahan University of Technology (IUT), in 2008. She got her MSc in industrial engineering from Amirkabir University of Technology (Tehran Polytechnic), in 2015. Currently, she is a Corporate Performance Management (CPM) researcher at FANAP Company. Her research interests include Functional modeling, Dimensional modeling and Cube Design of Measures in SQL for Entrepreneur Resource Planning Systems Modules, Public Relationship Management, and Customer Relationship Management.

© IEOM Society International