Comparison of Artificial Neural Network and SPSS Model in Predicting Customers Churn of Iran's Insurance Industry

Gholamreza Pirmohammadi
Payame Noor University
Department of Mathematics
Tehran, 19395-3697, Iran

Maryam Mast Zohouri
Payame Noor University
Department of Mathematics
Tehran, 19395-3697, Iran

ABSTRACT
This paper has three aims. The first aim is Prediction of Iran's insurance industry customers churn using data mining techniques and SPSS software. In this way, on the one place, in data mining part, multi-layer perceptron ANN[1] with 8 neurons in hidden layer has applied and the best performance of this network appears in epoch 10. Plus, the structural model of network is added. On the other place, Regression test has used in order to prediction customer churn by SPSS. As a result, the performance of predicting regression and neural network model are compared. Second, the difference between target and output values is presented based on the Root Mean Square Error (RMSE) and Mean Square Error (MSE) codes in Matlab. Indeed, MSE have less value rather than RMSE. Finally, in order to prevent the waste of financial and human resources, the K-Means method has used for clustering customers into two groups of churn and non-churn.

Keywords
Data Mining, Customer Churn, ANN, Perceptron, Prediction, Cluster.

1. INTRODUCTION
Nowadays, according to saturation of markets and the subject of CRM2, customers preservation, specially valuable customers is the center of manager's consideration. Insurance industry also because of low cost of movement, has encountered with customers with high risk of turning over and is not an exception. Data mining is one of the main branches of machine learning in data analysis and is one of ten developing knowledge in order to discovering hidden patterns, relationships and extract rules for predicting future behaviors in the database. Due to this, in this research, the problem of prediction of Iran's insurance industry customers churn is investigated using data mining. By view of [3], turned over customer is someone who cuts all his activities with the insurance company. The concept of CRM includes four dimensions: customer recognition, customer attraction, customer retention, and customer development (see [12]). Considering the importance of the issue, many researchers have carried out wide investigations in dealing with prediction of customers churn and have suggested a solution to control it. There have been some of these studies such as [2], [4], [5], [9], [10], [11], [14] using ANN which are composed of three layers, input, medium and output layer generally. These networks are applied as instruments to prediction, classification and clustering. Apart from this, in [2], a related prediction problem has done using neural network and logistic regression. Moreover, in [4], the author talk about Logistic Regression because of it's accuracy and flexibility. Where as, [6] uses Anfis networks in order to prediction customer churn. What is of paramount importance is that which model works well on your data set. In the present paper, multi-layer perceptron ANN model towards SPSS model applied for prediction and classification as well as clustering of customers. The data set used in the study consists 138 real customers of a branch of Iran's insurance company by neglecting 12 records of customers with our definition of customer churn. The studied case is one of private Iran's insurance company branch that have provided our initial data set. The data extraction interval was between the years 2010 to 2019 (nine years). Data preprocessing includes data quality, data cleaning, data integration, data reduction, and data transformation and data discretization has done (See [7], [8]). Customers who have entered in the last year of the indicated interval, have not the possibility for putting value. Hence, they have deleted of our data set. Some fields of data set also have neglected because of their little influence. A two-year purchase time is taken to define customers are churn or not. In exact words, if a customer had not bought non policy, he is regarded as a customer churn, but if he did not buy only in the last before, he had labeled as an customer non-churn. This study uses data mining and statical methods to predict customer churn rates and suggest timely recommendations to increase customer retention and thereby increase overall profitability. To achieve this goal, the Perceptron neural networks with 8 neurons in hidden layer has applied. Notice that sixteen attributions as inputs for the used network include not only quantitative variables but also qualitative variables. There should be mention there are eleven type of customers which are employers of metal, clothing, food, cement, hospitality, broadcasting, medical supplies, service, plastic manufacturing, municipality, and wood industries. The sixteen aforementioned variables have known as Number of insurance policies purchases, Fire insurance, Responsibility insurance, Complementary therapy insurance, Fire insurance's cost, Responsibility insurance's cost, Complementary therapy insurance's cost, Frequency of fire damages, Frequency of responsibility damages, Frequency of complementary therapy damages, Result of fire damages, Result of responsibility damages, Result of complementary damages, Average number of installments, Method to pay premium, The length of the contract extension, respectively. Of course, a variable as target for network, which is supposed as Customers status. Then we have normalized our data set in a no scaling method so that we transfer all components of each input by it's related scale into the interval [0,1]. After all data preprocessing, we

[1] Artificial Neural Network
[2] Customer Relationship Management

have applied Matlab 2018 for data mining part. The paper organized in seven section. In primary sections we state some auxillary informations. Section 6 contains the main results which is devided into two parts. The subsection 6.1 belongs to data mining model results. In details, in 6.1.1, the objective is to prediction of customers churn. In order to this, we have used multi-layer perceptron neural network. By choosing the right geometry for the network, that is, exactly 8 neurons in hidden layer and then training the network and finally testing the system, we have obtained the regression number as R=0.7021 (Of course, we tried the network by different number of neurons, but the best selection was eight). Added to this, we have shown that the best validation performance of network occurs at epoch 10. In addition, the structural model of neural network has also shown. The subsection 6.1.2 speaks about classification the result of network (output) and target based on RSME[3] and MSE[4]. In the sequel, in section 6.2, we have used SPSS models. Firstly, in 6.2.1, prediction of customers churn using Regression test have done. Next, in 6.2.2, we have applied K-Mean clustering method in order to cluster customers into two groups of churn or non-churn.

2. RESEARCH METHODOLOGY

2.1 Qualitative research method

In qualitative section, researcher identify the main indicators of customers churn via in-depth interviews with insurance industry executives, branch heads, deputies and insurers.

2.2 Quantitative research method

In this section, applying statical techniques, we have designed a model to predict customers churn while clustering insurers. In order to this, we have first applied algorithms of NNT methods and SPSS for prediction. To be more precise, we have used perceptron neural network in data mining part and regression exam in SPSS. Then comparing the relavant results, we have evaluated two methods. Second, we have shown the difference between target and output based on the Root Mean Square Error (RMSE) and Mean Square Error (MSE) via matlab's codes. Finally, using K-Means method in SPSS, we have clustered customers into two groups of churn and non-churn groups.

3. DESCRIPTIVE MODEL

One of strong methods for implementation of data mining projects is CRISP[5] method (see [1]). In the present paper, proposed model is based on CRISP which includes six steps. Each step includes itself some subsections so that each step's input has depended to the previous step's output. All of each steps have shown in Fig 1.

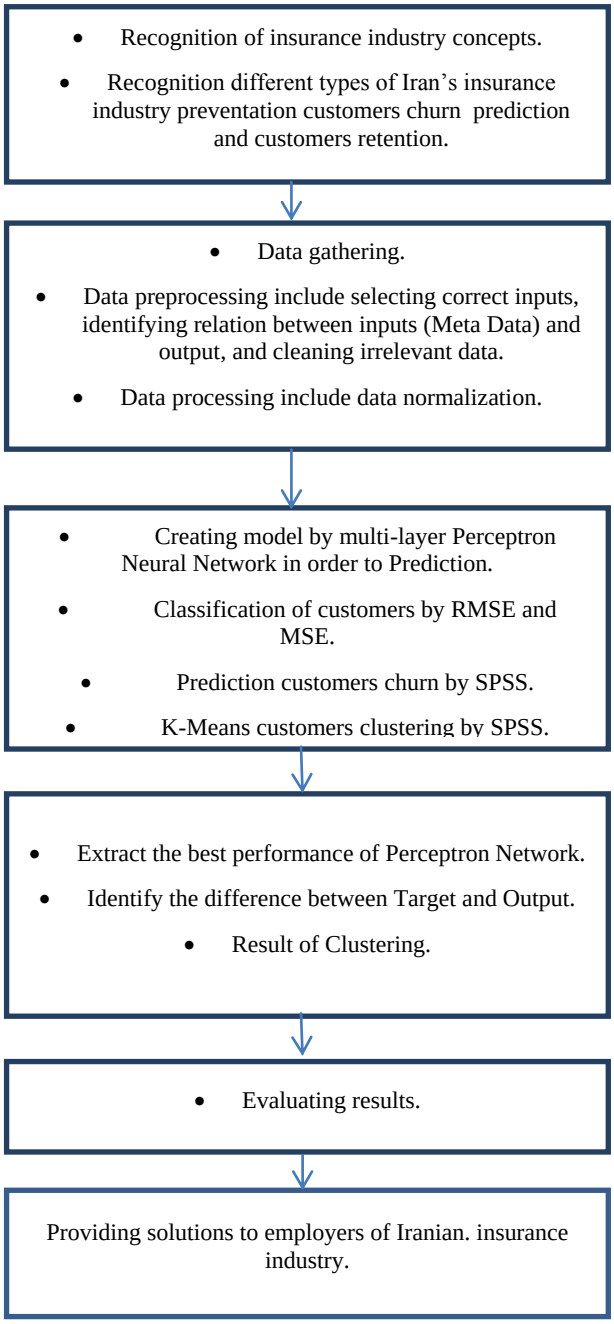

Fig 1. CRISP Model

4. RESEARCH VARIABLES

Here, the number of research's variables have listed.

3 Root Mean Square Error

4 Mean Square Error

5 Cross Industry Process for Data Mining

Table 1. Research Variables

Number	Variable's Notation	Variable's Name	Type
1	x1	Number of insurance policis purchases	Input
2	x2	Fire insurance	Input
3	x3	Responsibility insurance	Input
4	x4	Complementary theraphy insurance	Input
5	x5	Fire insurance's cost	Input
6	x6	Responsibility insurance's cost	Input
7	x7	Complementary insurance's cost	Input
8	x8	Frequency of fire damages	Input
9	x9	Frequency of responsibility damages	Input
10	x10	Frequency of complementary theraphy damages	Input
11	x11	Result of fire damages	Input
12	x12	Result of responsibility damages	Input
13	x13	Result of complementary damages	Input
14	x14	Avarage number of installment	Input
15	x15	Method of pay premium	Input
16	x16	The length of contract extension	Input
17	x17	Customers status	Target

5. RESEARCH QUESTIONS

- How we can predict insurers churn in Iran's insurance industry?
- How we can classify insurers into churn and non-churn groups?
- How we can cluster insurers into two clusters of churn and non-churn?
- How we can determine the effect of variables in customers churn?

6. RESULT AND DISCUSSION

6.1 Data mining model

Data mining is a young branch of science that find hidden patterns in big data set. Today, increasing data in the world enhanced the importance of data mining, so non-modern method of management is not a suitable method for analyzing big data. Therefore, data mining is a powerful knowledge in issuses such as prediction, classification, clustering, approximation and etc. Moreover, data mining is a way to solve complicated problems by giving an appropriate model. In insurance industry also we are concerned with similar problems in order to get out the waste of financial and human resources. In this way, we state the followings.

6.1.1 Prediction customers churn by perceptron neural network

Customer retention is very significant problem for a variety of companies and industries. Determining turnover and retention is one of applications of data mining in CRM6. The total process here is summarized in Fig 3. The research begins with the litreture study and data collection.

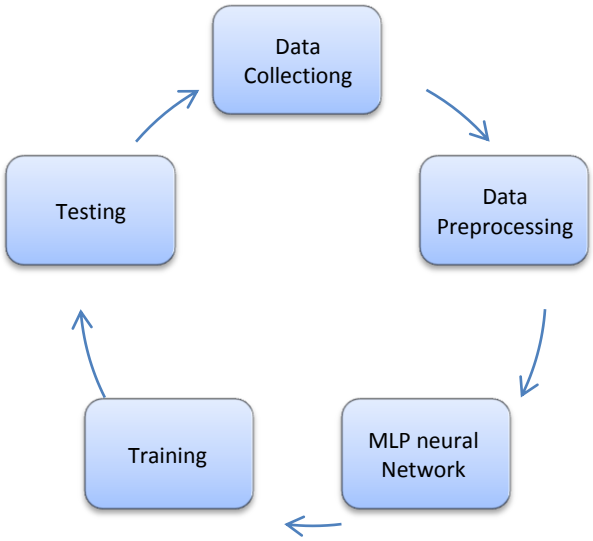

Fig 3. ANN Model

Then we collected 138 records of relevant customers after preprocessing include data cleaning and remove outliers, aggregating, data normalizing, and data discretization. In more details, we have used metadata in data cleaning step too. Also, in order to handling data reduction and integration, dimensionality reduction was very efficient. Moreover, in data normalization, we have applied Min-Max method. Then we have recalled data set in Matlab2018 and use multi-layer perceptron neural network with 8 neurons in hidden layer. As a result of implementation, the structural model of neural network figured out in Fig 4.

6 Customer Relation Management

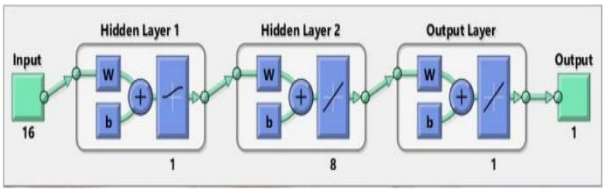

Fig 4. Structure of Network

The measured regression number is equal to R=0.7021 based on accurated geometry of the network, as we can see in Fig 5.

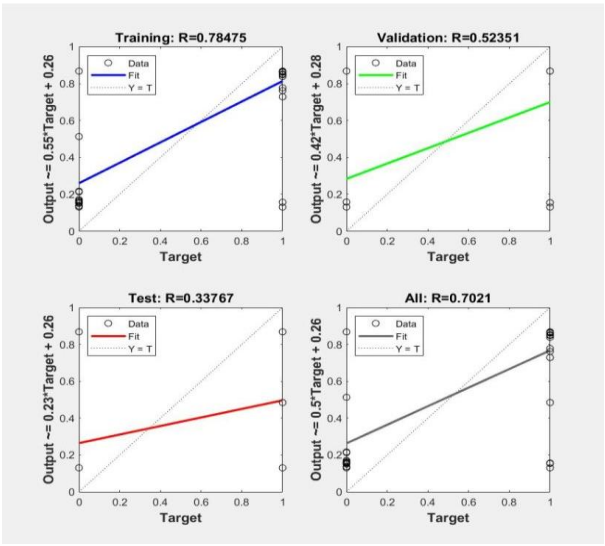

Fig 5. Neural Network Training Regression (PlotRegression)

Consequently, Fig 6 shows the best performance of network which occurs in epoch 10.

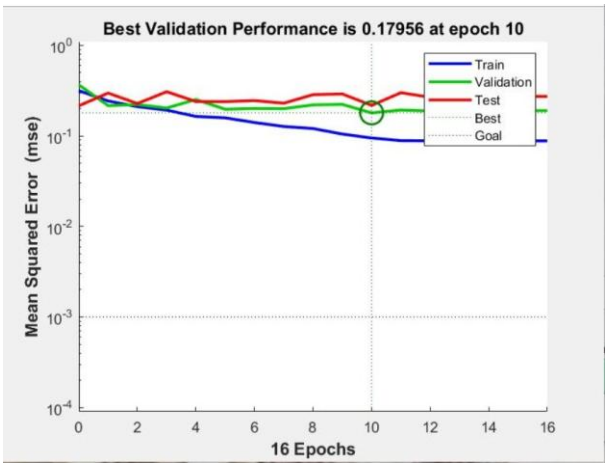

Fig 6. Neural Network Training Performance

6.1.2 Comparing differences between output and target

In order to classify and present the differences between output and target of the network, we writted some codes based on the RSME and SME methods. The selected criteria is the best model which is based on the lowest predict error. Indeed, the MSE or RMSE criterion is the most commonly studied criterion as follows:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{T}(y_i - \hat{y}_i)^2}{T}$$

and RSME is the positive root of SME, in which T is the number of input variables and y_t and $\hat{y}_t$ are real and predicted values of the intended variables respectively. Indeed, $y_t - \hat{y}_t$ shows the differences between these two amounts. In the following, we figured out the result.

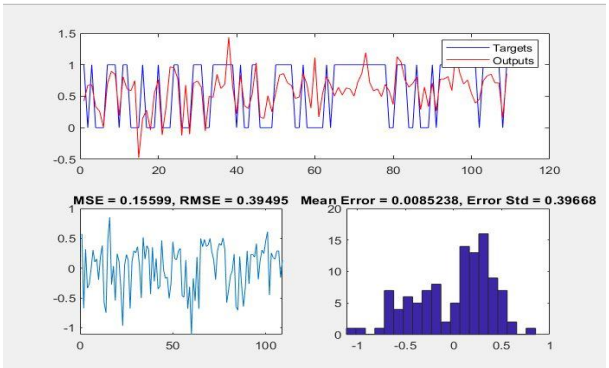

Fig 7. Differences between Output and Target

6.2 SPSS Model

In this section, after normalizing the data set, all data transferred to scale type same as the data mining section. This means that nominal and ordinal type of variables transferred to scale type. However, we could work with a combination of quantitative and qualitative variables without any processing for data set and apply Logistic Regression for prediction customers churn by SPSS, but we prefer to use a unique form of data set for two models, that is, NNT and SPSS models. Here we have applied Multiple Linear Regression method instead.

6.2.1 Customers churn prediction by regression Test

The main objective of multiple regression is to make linear combination of independent variables so that they cause the maximum correlation with dependent variable. Thus this linear combination can be used for predicting of dependent variable and the importance of each independent variable can be evaluated. Note that each independent variable has the same role in prediction. In addition, in multiple linear regression, the dependent's variable values estimated from independent variables. The normality of dependent variable is particularely necessary in order to use regression test, which is examined by K-S[7] test. Here, we have sixteen independent variables as we have indicated before. Hence, we enterprate the output of multiple linear regression test which is done by SPSS as follows. Fig 8 states that 28.6 percent of the whole changes of the dependent variable, Y, is depend to sixteen independent variables and the rested percent, that is, (71.4) is affected by variables which are out of the model.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.535[a]	.286	.192	.440

a. Predictors: (Constant), x16, x7, x12, x15, x5, x2, x14, x6, x11, x13, x3, x10, x8, x9, x4, x1

Fig 8. Model Summary

[7] Kolmogrov Smirnov Test

In the sequel, amount of F=3.035 in Fig 9, shows that the proposed regression model combined of sixteen independent variables and one dependent variable can not be an appropriate model to prediction.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.413	16	.588	3.035	.000[b]
	Residual	23.456	121	.194		
	Total	32.870	137			

b. Predictors: (Constant), x16, x7, x12, x15, x5, x2, x14, x6, x11, x13, x3, x10, x8, x9, x4, x1

Fig 9. ANOVA

As a data visualization, Fig 10 emphsises particularely the inefficiency of the model. This is a strong reason to apply ANN model in order to prediction.

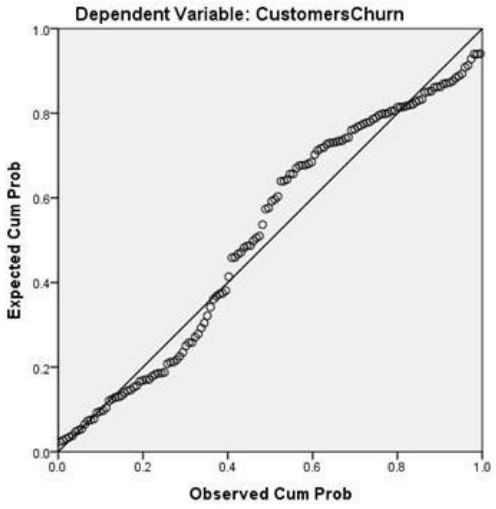

Fig 10. Normal P-P Plot of Regression Standardize Residual

6.2.2 Customers Clustering by K-Means Method

The process of grouping a collection of data and placing it in a class of similar samples is called clustering. A cluster is a set of data that is similar to other data in the same cluster but are different of instances from other clusters. In the business world, clustering helps marketers find different groups among their customers or identify customers based on purchase patterns. One of the most important methods of clustering is the K-Means method which is based on data patition. In this method, the assumption is that the number of clusters is known. Here, it is supposed that the number of clusters is two. After we have done the the K-Means method in SPSS, two variables are caused in Variable View which are labeled as "Cluster Number of Case" and "Distance of Case from its Classification Cluster Center" respectively. The variable QCL_1 determines that each customer belongs to which group and the variable QCL_2 determines that how much the distance of each customer of the center's cluster is. Now, we are going to analyze the output. Fig 11 Shows the centers of the initial clusters.

	Cluster	
	1	2
x1	.67	.33
x2	1.00	.00
x3	1.00	.00
x4	.00	1.00
x5	.0730	.0000
x6	.0000	.0000
x7	.0000	.0115
x8	1.0000	.0000
x9	.4	.0
x10	.0000	.4782
x11	1.00	.00
x12	.830	.000
x13	.000	.500
x14	.583	1.000
x15	1	0
x16	.888	.222

Fig 11. Initial Cluster Centers

In addition, Fig 12 explains each customer belongs to which group. For example, the 13nd customer belongs to cluster 1, that is churned and the customer of number 127 belongs to cluster 2, that is non-churned. The forth column in Fig 12, shows the distance of each customer of the center of the related cluster and whatever this number is greater, the customer is more far of the cluster's center.

Cluster Membership

Case Number	GroupName	Cluster	Distance
1	Metal industry	1	.891
2	Metal industry	1	1.399
3	Cloth industry	2	.812
4	Cloth industry	1	1.267
5	Metal industry	2	.817
6	Food industry	2	.956
7	Cement industry	2	.968
8	Hospitality	1	1.080
9	Broadcasting industry	1	.952
10	Broadcasting industry	1	1.246
11	Medical supplies industry	1	1.447

12	Service industry	2	.979
13	Plastic industry	1	1.356
14	Plastic industry	1	.822
15	municipality	1	1.583
16	municipality	2	1.282
17	municipality	2	1.100
18	Service industry	2	1.005
19	Service industry	2	1.193
20	Wood industry	1	1.153
21	Cloth industry	2	1.134
22	Metal industry	1	.839
23	Wood industry	1	1.007
24	Food industry	1	.971
25	Plastic industry	1	1.032
26	Cloth industry	2	1.180
27	Broadcasting industry	2	1.073
28	Service industry	2	.965
29	Cloth industry	2	.837
30	Food industry	1	1.475
31	Food industry	1	.863
32	Metal industry	2	.776
33	Cloth industry	2	1.255
34	Metal industry	2	1.271
35	Wood industry	2	.998
36	Service industry	1	.862
37	municipality	1	1.332
38	Food industry	1	1.093
39	Food industry	1	1.150
40	Metal industry	2	.758
41	Plastic industry	2	1.265
42	Plastic industry	1	.846
43	Broadcasting industry	2	.993
44	Wood industry	2	.891
45	Food industry	2	1.453
46	Service industry	2	1.291
47	Hospitality	1	1.261
48	Cloth industry	1	.838
49	Cloth industry	2	.748
50	Wood industry	2	.975
51	Hospitality	1	.979
52	Food industry	1	.944
53	Cement industry	2	1.036
54	Hospitality	2	.972
55	Broadcasting industry	1	.992
56	Broadcasting industry	1	.869
57	Medical supplies industry	2	1.114
58	Service industry	2	.843
59	Food industry	2	.769
60	Plastic industry	2	1.034
61	municipality	2	1.119
62	Service industry	2	.758
63	municipality	1	.923
64	Service industry	1	.799
65	Wood industry	1	1.389
66	Cloth industry	1	.957
67	Plastic industry	2	.997
68	Wood industry	2	.798
69	Cloth industry	2	.803
70	Plastic industry	2	1.298
71	Cloth industry	1	1.591
72	Broadcasting industry	2	1.141
73	Medical supplies industry	1	1.084
74	Service industry	1	1.113
75	Service industry	2	.990
76	Food industry	2	.803
77	Metal industry	1	1.397
78	Cloth industry	2	.896
79	Metal industry	1	1.507
80	Service industry	2	1.232
81	Service industry	2	1.222
82	Cloth industry	1	.982
83	Hospitality	1	1.018
84	Food industry	1	1.213
85	Cloth industry	1	.826
86	Metal industry	1	1.112
87	Plastic industry	1	.965
88	Cloth industry	2	.839

89	Broadcasting industry	2	1.132
90	Service industry	2	.937
91	Broadcasting industry	1	1.040
92	Plastic industry	1	1.132
93	Metal industry	2	.858
94	Food industry	1	1.113
95	Hospitality	1	1.027
96	Service industry	2	1.186
97	Service industry	1	1.244
98	Wood industry	1	.993
99	Cloth industry	1	1.133
100	Metal industry	1	.968
101	Wood industry	2	1.249
102	Food industry	2	1.250
103	Food industry	1	1.063
104	Cloth industry	2	1.046
105	Wood industry	1	.974
106	Medical supplies industry	1	.978
107	Service industry	2	.938
108	Hospitality	1	1.198
109	Broadcasting industry	1	.955
110	Wood industry	1	1.166
111	Metal industry	1	.991
112	Cloth industry	1	.858
113	Service industry	2	1.028
114	Medical supplies industry	1	1.118
115	Hospitality	1	1.031
116	Food industry	2	.941
117	Plastic industry	2	.915
118	Metal industry	2	.963
119	Cloth industry	1	1.417
120	Food industry	1	1.232
121	Medical supplies industry	1	.770
122	Plastic industry	1	1.127
123	Broadcasting industry	2	1.286
124	Metal industry	2	1.085
125	Hospitality	1	1.148
126	Wood industry	2	1.024
127	Broadcasting industry	1	1.055
128	Service industry	2	1.008
129	Service industry	2	1.101
130	Food industry	1	.921
131	municipality	1	1.098
132	Cement industry	2	.986
133	Food industry	2	.849
134	Wood industry	1	.963
135	Service industry	1	.937
136	Plastic industry	1	1.288
137	Medical supplies industry	1	1.114
138	Cloth industry	1	1.384

Fig 12. Cluster Membership

Moreover, Fig 13 determines that which variable has the most significant and the lowest role in clustering. Notice to amount of F. Indeed, x1 has the most crucial role and x13 has the lowest role in clustering, as we have seen.

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
x1	4.225	1	.031	136	137.851	.000
x2	.102	1	.145	136	.702	.404
x3	34.319	1	.000	136	.	.
x4	.025	1	.253	136	.099	.754
x5	.000	1	.000	136	.368	.545
x6	.000	1	.000	136	.101	.751
x7	.000	1	.000	136	.236	.628
x8	.042	1	.013	136	3.311	.071
x9	2.100	1	.029	136	71.480	.000
x10	.064	1	.094	136	.688	.408
x11	.842	1	.150	136	5.624	.019
x12	8.524	1	.073	136	116.705	.000
x13	.000	1	.106	136	.000	.983
x14	.060	1	.045	136	1.315	.254
x15	.446	1	.202	136	2.204	.140
x16	.058	1	.034	136	1.690	.196

Fig 13. Cluster Membership

Finally, Fig 14 illustrates that the number of customers in cluster 1 is equal to 74 and the number of customers in cluster 2 is equal to 64.

Cluster	1	74.000
	2	64.000
Valid		138.000
Missing		.000

Fig 14. Number of Cases in each Cluster

7. CONCLUSION AND FUTURE RESEARCH

Findings on customers churn prediction through ANN by Matlab software show that the regression number using multi-layer perceptron neural network with 8 neurons in hidden layer is equal to 0.7021. Added to this, the best performance of ANN is occurred in epoch 10. In comparison, R2 as a criteria to prediction in SPSS method is equal to 0.286 which is not as acceptebale as prediction criteria in ANN model. As an statical performance of ANN model, in order to classify output and target value, SME and RSME are computed which are equal to 0.15599 and 0.39495 respectively. One of our future works is to cluster customers by ANN with K-Means un-supervised algorithm. Furthermore, however, we have tried to a suitable prediction model using a special ANN, Anfis, but as the number of input variables were almost too large, it makes the network complicated and it causes problem in training of network. Although, if we optimize the model, the problem may be eradicated. Thus the other future intention is hand in an optimized Anfis model in order to prediction customers churn.

8. ACKNOWLEDGEMENT

The authors would like to thank Dr. S.A. Mirroshandel of Guilan University for his useful comment.

9. REFERENCES

[1] Alizadeh, S., Ghazanfarian, M., and Teimourpour, B. 2011 Data Mining and knowledge discovery. Tehran: Publication of Iran University of Science and Technology, 2nd ed, p.p. 70-250.

[2] Bagheri, M. and Valipour, M., and Amin, V. 2012 The bankruptcy prediction in Tehran share holding using neural network and it's comparison with logistic regression. TJMCS, Vol. 5, No. 3, p.p. 219-228.

[3] Baicker, K. and Chandra, A. 2006 The labour market effects of rising health insurance premiums. Journal of Labour Economics, Vol. 24, No. 3.

[4] Berry, M.J.A. and Linoff. G.S. Data Mining Techniques for Marketing, Sales and Customers Relationship Management, 2nd edition, Wiley Publishing Inc, Indinana, USA.

[5] Bigus, J.P. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support Paperback, McGraw-Hill, Inc. Hightstown, NJ, USA.

[6] Chih, R. and Yu.Hsin, L. 2009 Customer churn prediction by hybrid neural networks. Expert Systems with Applications, Vol. 36, p.p. 12547-12553.

[7] Dasu, T. and Johnson, T. 2003 Exploratory Data Mining and Data Cleaning, John Wiley.

[8] Han, J., Kamber, M., and Pei, J. 2011 Data mining: Concepts and techniques, 3rd ed. United States, Morgan Kaufman.

[9] Ismail, M.R., Awang, M.Kh., Rahman, M., and Makhtar. M. 2015 A multi-layer perceptron approach for customer churn prediction. International Journal of Multimedia and ubiquitous engineering, Vol. 10, No. 7, p.p. 213-222.

[10] Izadparast, S. M., Farahi, A., Fath Nejad, F., and Teimourpour, B. 2012 Using data mining thechniques to predict the detriment level of car insurance customers. Iranian Journal of Information Processing and Management, Vol. 27, Issue 3.

[11] Mehregan, S., and Samizadeh, R. 2012 Customers retention based on the number of purchase: A data mining approach. Int. J. Manag. Bus. Res., 2 (1), p.p. 41-50.

[12] Ngai, E.W.T., Xiu, L., and Chau, D.C.K. 2009 Application of data mining techniques in customer relationship management: A Literature review and classification. Expert Systems with Applications, Vol. 36, p.p. 2592-2602.

[13] Pendar, M., and Haji, M. 2017 Comparison of neural network models, vector auto regression (VAR), Bayesian vector-autoregressive (BVAR), generalized auto regressive conditional heteroskedasticity (GARCH) process and time series in forecasting inflation in Iran. IJIM, Vol. 9, No. 2.

[14] Sabbeh, S.F. 2018 Machine-learning techniques for customer retention: A comparative study. IJACSA, Vol. 9, No. 2.

[15] Sharma, A., and Panigrahi, P.K. 2011 A neural network based approach for predicting customer churn in cellular network services. International Journal of Computer Applications, Vol. 27, No. 11.

[16] Tarokh, M.J., and Sharifiyan, 2007 K. Usage of data mining in improvement CRM. Industrial management studies, Vol. 6, p.p. 153-181.