Establishing the Potential Clients Using Artificial Neural Networks

Z. Pamuk, Y. Yurtay, and O.Yavuzyilmaz

***Abstract*—Today, technologies retrieving forward-looking information from the existing data are available. In this study, whether the clients would open a deposit account was estimated using the data in the marketing campaign of a bank in Portugal for its clients. The purpose of the study was to create a decision support system to determine the potential clients in future. The data set collected from 4.512 subjects consists of 16 input attributes (job, age, balance, etc.) and 1 output attribute (yes/no). In the study, the 6-fold cross validation method was used. The data obtained from 3.760 people were used for the training process and the data obtained from 752 people were used for the testing process. As classifiers; Feed Forward Neural Networks (FFNN), Probabilistic Neural Network (PNN) and k Nearest Neighbor (kNN) were used. At the end of the study, success ratios of different algorithms were compared by Receiver Operating Characteristics (ROC) analysis method. Feed forward neural network yielded the best result with an accuracy rate of 95.74%.**

***Index Terms*—Bank data, deposit, artificial neural network, k nearest neighbor algorithm.**

I. INTRODUCTION

TODAY, data can be digitally collected and stored due to the rapid development of computer systems. Together with the high increase of data and the need for obtaining significant extractions, the concept of data mining emerged. The main objective of data mining is to find out information such as relationships between data, patterns, changes, deviations and trends and certain structures with the combinations of mathematical theories and computer algorithms and obtain valuable information through the interpretation of this information.

Data mining is considered to consist of four main topics including classification, categorization, estimation and visualization. Marketing campaigns occupy an important place for banking institutions. In this study, the marketing campaign is based on telephone conversations.

Z.Pamuk, Electrical Electronic Engineering Department, University of Sakarya. 54100, Sakarya, Turkey, (e-mail: ziynet@sakarya.edu.tr).

Y.Yurtay, Computer Engineering Department, University of Sakarya, Sakarya, Turkey,(e-mail: yyurtay@sakarya.edu.tr)

O.Yavuzyilmaz, Kocaeli University, Gazanfer Bilge Vocational School, Department of Public Relations and Publicity, Kocaeli, Turkey, (e-mail:oguz.yavuzyilmaz@kocaeli.edu.tr)

Telephone conversations have a significant impact on the decision-making process of clients. A classification was made using the client information obtained in the telemarketing campaign the employees of the Portugal Bank carried out for its clients and whether they would be deposit account clients was tried to be determined.

In some studies carried out using bank data, whether the clients would be a deposit account subscriber or not was predicted using the information obtained from the clients and data mining techniques. For example; Moro et al. made analyses using data mining techniques in line with the information obtained as a result of telemarketing campaigns. They used support vector machine, decision tree and naive bayes classifier data mining models. In the study, the best results were obtained from the Support Vector Machine Model [1]. Moro et al. studied on the same topic in the study titled "A data mining approach for bank telemarketing using the rminer package and R tool" they carried out in 2013. They performed that study using the r tool extension of rapidMiner program [2]. In the study titled "Bank direct marketing based on neural network" carried out by Elsalamony Hany A. and Elsayad Alaa. M., the data obtained as a result of the marketing campaign were analyzed with MLPNN (Multi-Layer Perceptron Neural Network) and Decision Tree-DT models [3]. In the study titled "Evaluating marketing campaigns of banking using neural networks" carried out by Qeethara Kadhim Al-Shayea, the prediction about whether the clients would be a deposit account holder or not was made in line with the information obtained from them through ANN and DT techniques [4].

II. METHOD

In the first stage, the data were preprocessed. Thus, it was aimed to obtain accurate results from the study. At this stage, data cleaning and data integration steps were applied. The missing data were deleted and the data were tried to be made consistent. In order to find out the relationships between the data and obtain information through the interpretation of them, ANN and as a different classifier apart from ANN; k nearest neighbor algorithm, which is simple and frequently used in the literature, were used. *k* nearest neighbor algorithm is abbreviated as *kNN*. In the stage of evaluation of the results, Receiver Operating Characteristic –ROC Formulas were used.

A. *Pre-processing*

Pre-processing is quite important to obtain the most accurate result from the data. At this stage, first data cleaning, next data integration processes were carried out. In the final stage, data conversion process was performed.

B. *Data cleaning*

Approximately seventy thousand people were interviewed in the data recording process. However, not all these data were processable. In order to increase the usability of the data, the data including missing information and incorrect data were cleaned and 4.512 processable data were obtained.

C. *Data integration*

17 variables of the 4.512 data have different types such as symbolic, numeric and categorical. In the study, all variable types were digitized in order to facilitate the analyses on the data. The digitized values are indicated in Table 1 Input data and Table 2 Output data.

D. *Data Conversion*

The digitized data were normalized between -1 and +1 with the help of Matlab © code. Thus, the speed of processing increased while the data were trained.

E. *Data Set*

The data set was obtained from the telephone conversations between the employees of the Portuguese bank and the bank clients. The study was carried out using the client information obtained from the telemarketing campaign of the bank and whether the clients would be deposit account subscriber or not was tired to be determined with the help of the information given by the clients. The data belonging to a total of 4.512 clients are given in Table 1 and Table 2.

The data were divided into training and testing sets. The data were also divided into K=6 groups using the k fold cross validation technique. How the groups were determined is given in Table 4 in the Application section.

F. *Classification Methods*

ANN emerged as a result of the efforts to simulate the working system of human brain artificially. The first artificial neural network was realized by Warren McCulloch, a neurologist, and Walter Pitts, a mathematician, in 1943. McCulloch and Pitts modelled a simple neural network with electrical circuits being inspired by the computing capacity of human brain. They are information processing systems which generally simulate the working principles of human brain or central nervous system [5].

ANN has a simple structure and a directed graph format. Each node is an nth degree nonlinear circuit called “cell”. These nodes are defined as processing elements in ANN. There are links connecting the nodes to each other and each of them functions as a simplex communication path. A single output can feed several cells, in other words; each communication element can receive a desired number of input connections and a single output connection. The output of the processing element can be in a desired mathematical type. Continuously working input elements produce an output signal. The input signals carry information to ANN and the result can be obtained from the output signals. ANN is composed of three layers including Input Layer, Interlayers (Hidden Layers) and Output Layer [6].

Feed Forward Network (FFN) and Probabilistic Neural Network (PNN) were used in the application as available network types in ANN. The accuracy results of the data sets were analyzed giving different values to the variables of the number of neurons, the number of layers, goal and lr determined in FNN Network. Goal value is a value indicating how much of the error can be minimized. Generally a value near zero is given instead of giving zero. In this study, the value of 0.01 was given. It continues training till the network error value is 0.01 or the desired epoch, i.e. “cycle number” is reached. The LR value, i.e. learning rate should not be selected as a very high or very low value. If a very high value is selected, the network memorizes the data. On the contrary, if a very low value is selected, then the network either learns the data too slowly and causes loss of time or cannot learn it.

FFN includes a series of layers. The first layer has a connection from the network inputs. Each layer is connected with the previous layer. The final layer produces the network outputs. Feed Forward Networks can be used for output matching and any kind of input. A Feed Forward Network can be used in any finite input output matching problems with a hidden layer and sufficient neurons in the hidden layer [7].

PNN is a radial basis neural network. It is based on counselling learning. PNN is a neural network which uses Bayes Theorem in decision making. According to Bayes theorem, if an equality of vector x is accurate, it belongs to the 1^{st} class, if not, it belongs to the 2^{nd} class [8].

kNN method is one the learning methods which solves the classification problem. With this method, the similarities of the data to be classified to the data in the learning set is calculated, the average of the values of their k nearest neighbors is taken and they are assigned to a class according to their threshold value. In order to perform this assignment, the characteristics of each class should be clearly determined in advance.

TABLE 1.
INPUT DATA

id	Variables	Explanation	Type	Value (digitized value)
1	Age	Age when contacted	numeric	18 +
2	Occupation	Occupation of the contact person	categorical	unknown(1), administrator(2), unemployed(3), manager(4), servant(5), contractor(6) , student(7), worker(8), self-employed (9), retired(10), technician(11), service personnel(12)
3	Marital status	Marital status of the contact person	categorical	single(1) , divorced(0), married(-1)
4	Education	Educational status of the contact person	categorical	unknown(1),primary school(2), secondary school(3), undergraduate(4)
5	Non-performing loan	Does the client have a non-performing loan?	categorical	yes(1), no(-1)
6	Avg balance	Average annual balance of the client's current accounts in Euro currency	numeric	-3313euro < ; < 71188 euro
7	Mortgage loan	Does the client have a mortgage loan?	symbolic	yes(1), no(-1)
8	Personal loan	Does the client have a personal loan?	symbolic	yes(1), no(-1)
9	Communication type	What was used as a means of communication?	categorical	unknown(1), telephone(0), mobile phone(-1)
10	Last contact day	Last contact day when the clients were interviewed for the campaign	numeric	1<; <30
11	Last contact month	Last contact month of the year when the clients were interviewed for the campaign	categorical	January(1), February(2).....December(12)
12	Length of interview sec.	Length of the interview (in seconds)	numeric	4 sec < ; <3025 sec
13	Number of contacts established in the campaign	Number of contacts during the campaign	numeric	1 < ; <50
14	How many days after the last campaign the interview was made	How many days after the last campaign were the clients contacted?	numeric	1 < ; <871 ;
15	Number of pre-campaign contacts	Number of contacts before the campaign	numeric	0< ;<25
16	Previous campaign result	Result of the previous marketing campaign	categorical	unknown(1), imperfect(2), other(3), successful(4)

TABLE 2.
OUTPUT DATA

id	Variables	Explanation	Type	Value
1	Y	Has the client subscribed to a term deposit account?	symbolic	yes (1) , no (-1)

kNN is an instance-based learning algorithm and used for performing classifications over the available learning data when a new instance is encountered. The algorithm determines the class of the instance looking at its k nearest neighbor when a new instance is encountered [9]. Referring to the data whose classes are known, their distance to the data whose classes are unknown is calculated and it is based on the selection of k number of observations with the minimum distance.

In the example in Figure 1, the number of nearest neighbors was taken as 5 and in which class the unknown data would be included was determined. As the number of neighbors in Class A was higher than the number of neighbors in Class B, the unknown data was included in Class A. The k value in the figure is not the value in the application, which is 511.

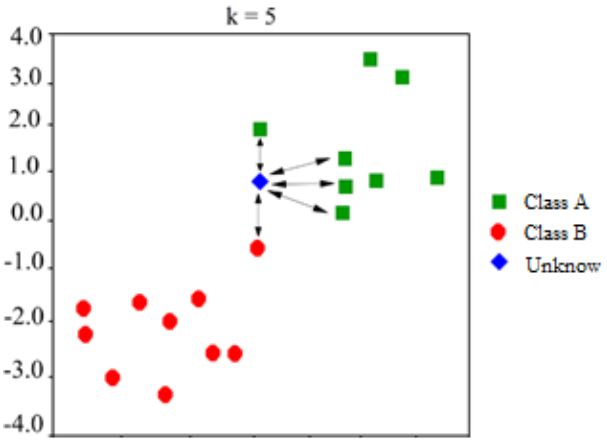

Fig 1. kNN representation

The steps of *kNN* algorithm are as follows.

1- The newcomer individual is added into the class.
2- *k* number of neighbors are looked at.
3- The distance is calculated using various distance function (Euclidean distance function).
4- The individual is assigned to the nearest place [10].

In equation 1, the Euclidean Distance Function formula, where d is the function of distance between two points, i and j are the indices of the points, x is the position of this point and the p value is 752, which is the test value, is given.

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2} \quad (1)$$

The distances of a point to all the other points are calculated separately, the rows are sorted and the smallest k number is selected. In which category the selected rows are is determined and the most repeated category is selected.

G. *Evaluation*

Receiver Operating Characteristic (ROC) was developed for the accurate identification of the signals detected on radar in Britain during the World War II and enabling the distinction between friend and foe. Lusted suggested the use of ROC analysis in decision making in medicine in 1967 and led to the use of it in medical imaging devices in 1969. In the following years, the use of ROC analysis in the evaluation of the performance of diagnostic tests in medicine gradually became widespread. The developments emerging with ROC analysis are a natural consequence of the need for the evaluation and comparison of statistical results [11].
ROC analysis is defined as "the receiver operating characteristic" or simply "ROC curve" in signal detection theory. The ROC curve is obtained with the ratio of sensitivity to specificity in cases where the distinction threshold value varies in binary classifier systems. In simpler terms, ROC can be expressed as the fraction of true positives to false positives [11].
While performing a ROC analysis, a classification should primarily be made. Meanwhile, the threshold values are determined and divided into two classes. These classes are positive (P) and negative (N) values according to the threshold value. When subjected to this classification, the estimated and actual values are divided into four different classes.

- If the estimated value is positive (P) and the actual value is also positive (P), it is called a true positive (TP).
- If the estimated value is positive (P) and the actual value is negative (N), it is called a false positive (FP).
- If the estimated value is negative (N) and the actual value is also negative (N), it is called a true negative (TN).
- If the estimated value is negative (N) and the actual value is positive (P), it is called a false negative (FN).

TABLE 3.
ERROR MATRIX FOR THE ANSWER "NO"

Error matrix		Classifier Outcomes	
		YES	NO
Actual Results	YES	TN	FP
	NO	FN	TP

Accuracy is the proximity of the measured value to the actual value. Error is the difference between the measured value and the actual value [12].

$$Accuracy=\frac{TP+TN}{(TP+TN+FP+FN)} \quad (2)$$

III. APPLICATION

All the data were divided into two groups as training and testing data to be given to the classifiers. While performing this division in the data set, the k-fold clustering algorithm was used. Firstly, a k value was selected to apply to the data set. For K=6, the data set including 4.512 data were divided into 6 equal parts (folds). The K-folds are illustrated in Table 4.

During fold partition, the numbers of "yes" and "no" belonging to the outcome variable were also taken into consideration. Thus, the state of excess of "yes" or "no" in any of the folds, which is possible in a random distribution, was eliminated. In case of such a situation, the outcomes of the model will also be affected negatively.

TABLE 4.
K-FOLDS

id	Yes	No	Total
Fold 1	86	666	752
Fold 2	86	666	752
Fold 3	86	666	752
Fold 4	86	666	752
Fold 5	86	666	752
Fold 6	86	666	752
Total	516	3996	4512

As illustrated in Table 5 randomly chosen 5 folds constitute the training data (3.760 pieces) and 1 fold constitutes the testing data (752 pieces). In Table 5, how the k-folds were formed is clearly expressed (k=6). In the k-fold technique, the data were named as A-B-C-D-E-F and 6 different data sets were created.

TABLE 3.
DATA SETS

Data set	Training Folds	Testing Folds
A	Fold1 - Fold 2 - Fold 3 - Fold 4 - Fold 5	Fold 6
B	Fold 1 - Fold 2 - Fold 3 – Fold 6 - Fold 5	Fold 4
C	Fold 2 – Fold 3 – Fold 4 – Fold 5 – Fold 6	Fold 1
D	Fold 1 – Fold 3 – Fold 4 – Fold 5 – Fold 6	Fold 2
E	Fold 1 – Fold 2 – Fold 4 – Fold 5 – Fold 6	Fold 3
F	Fold 1 - Fold 2 - Fold 3 - Fold 4 – Fold 6	Fold 5

The classification techniques applied in this study were realized using the data sets in Table 5. In the evaluation stage, the data set which gave the most accurate result to determine whether the client would be a deposit account subscriber in the future or not was found to be the data set B.

The numbers of the training and testing data sets used by ANN are shown in Table 6.

TABLE 6.
NUMBER OF DATA

	Yes	No	Total
Training	430	3330	3760
Testing	86	666	752
Total	516	3996	4512

The training set includes a total of 3.760 data records, whereas the testing set includes 752 data records. Each row in Table 1 corresponds to a record here.

The accuracy results of the data sets were examined by giving different values to the variables of the number of neurons, the number of layers, goal and lr determined in FFN network structure. The values which gave the best result were identified as follows; the number of neurons = 80, the number of layers = 1, goal = 0.01, lr = 0.005. With this identified network, the best accuracy result determined with ROC was found.

The vector state of the training output was found in PNN. The reason for the conversion of the training output into a vector matrix was that it would be used in creating a network. During the creation of the neural network, the training input, the training output vector and the distribution values were taken. Upon the creation of the network, the testing data were given to the network. As the training input values had been converted into the vector state initially, the results were produced in index format upon obtaining the testing output in vector state. The PNN model was experimented for the selected data set B, however the algorithm did not give a good result due to the excess number of data.

In the data set B, the data were classified using the kNN algorithm. The algorithm was implemented in the selected data set B. 3.760 data of the data set B were separated for training and 752 data for testing. The training input, training output and testing input data were given to the algorithm. The testing output was taken from the result of the algorithm. The classes of the data in the training input given to the algorithm were known, whereas the data whose classes were unknown belonged to the testing input data. The distance of each data in the set whose classes are known to the observation values whose classes are unknown is calculated according to the kNN algorithm. In the calculation of the distance, Manhattan distance function, Minkowski distance function and Euclidian distance function are used. K number of data with the minimum distance according to the calculation is selected. In this study, Euclidian distance function was used. K number was taken as 511. 511 data with the minimum distances calculated according to the algorithm were selected. To measure the reliability of the results, the obtained results were evaluated with ROC.

IV. CONCLUSIONS

In this study, the probability of people to open a deposit account in the bank was estimated for marketing campaigns. 3.760 of the 4.512 data were used as the training data and 752 as the testing data. The data were studied using ANN classification models; FFN and PNN networks and also kNN. As shown in Table 7. ROC Accuracy Results, the method which gave the best result was FNN. For bank marketing campaigns, the probability of people to open a deposit account was estimated with the accuracy rate of 94.22% and the probability not to open a deposit account with the accuracy rate of 94.10%.

TABLE 7.
ROC ACCURACY RESULTS

Model Name	YES (%)	NO (%)
FFN	94.22535	94.10071
PNN	69.9468	69.9468
KNN	88.56382	88.56382

As a result of the study, the training process was completed with the training set and processes were performed with the testing set where whether the client became a deposit account subscriber or not was not known. As a result of the testing set, the FFN classifier identified the outcome of "Yes" for "Has the client become a deposit account subscriber?" correctly with a rate of 94.22% and identified the outcome of "No" for "Has the client become a deposit account subscriber?" correctly with a rate of 94.10.

FNN achieved a better result compared to the other methods. This method is thought to be utilized by banks and all the institutions which conduct marketing activities.

ACKNOWLEDGMENT

The study is selected from National Engineering Research Symposium 2015 (Ulusal Mühendislik Araştırmaları Sempozyumu) UMAS 2015 (Duzce University).

REFERENCES

[1] Moro S. , Laureano R., Cortez P. (2011). Using data mining for bank direct marketing: an application of the crisp-dm methodology. European Simulation and Modelling Conference.
[2] Moro S.(2013). A data mining approach for bank telemarketing using the rminer package and R tool. Working Paper. ISCTE - Instituto Universitário de Lisboa.
[3] Elsalamony Hany A., Elsayad Alaa. (2013). M. Bank direct marketing based on neural network. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-6, August 2013.
[4] Qeethara Kadhim Al-Shayea. (2013). Evaluating Marketing Campaigns of Banking Using Neural Networks, Proceedings of the World Congress on Engineering 2013, Vol II.
[5] Oztemel E., Artificial Neural Networks, Papatya Publishing, İstanbul, 2003.
[6] Kayhan Gulez. www.yildiz.edu.tr/~gulez/3k1n.pdf. 2015.
[7] Beale, M.H., Hagan, M.T., Demuth, H.B. Neural Network Toolbox. 383, 2012.
[8] Specht, D. F.(1990). Probabilistic neural networks, Neural Networks, Vol. 3 (1), pp 109-118.
[9] Nearest Neighbors Tutorial people.revoledu.com/kardi/tutorial/KNN/HowTo_KNN.html. 2015.
[10] Özkan, Y., Data Mining Methods. Papatya Publishing, İstanbul, 2013.
[11] http://tr.wikipedia.org/wiki/ROC. 2015.
[12] Fawcett, T.(2006). An introduction to ROC analysis, Pattern Recognition Letters, 27, 861-874.

BIOGRAPHIES

Ziynet PAMUK was born in Zonguldak, Turkey, in 1980. She has received the bachelor's degree and master's degree from the Sakarya University, Sakarya, Turkey, in 2002 and 2008, respectively. She has graduated Ph.D. at Department of Electrical and Electronics Engineering in 2014. She has been working at Sakarya University since 2004. Her interests are Biomedical Engineering, Classification Algorithms, and Artificial Immune System and Artificial Neural Network.

Yüksel YURTAY was born in Eskişehir, Turkey, in 1968. He has received the bachelor's degree from Anadolu University, Turkey, in 1992. He has received the master's degree from the Sakarya University, Turkey, in 1997. He has been working at Sakarya University since 1993. His interests are Software Engineering, Data Mining, and Numerical Analysis.

Oğuz YAVUZYILMAZ was born in Kocaeli, Turkey, in 1980. He has received the bachelor's degree from Marmara University, in 2003 İstanbul and master's degree Kocaeli University in 2008, Kocaeli, Turkey. Since 2011; he is continuing his PhD studies at Production Management and Marketing of Science program in Sakarya University, Sakarya, Turkey. He has been studying; Marketing, Postmodern Marketing, Womm, Data Mining, Production and Marketing analysis techniques.