Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 32, 142-152, 2014

Research Article / Araştırma Makalesi

COMPARISON OF DATA MINING TECHNIQUES FOR DIRECT MARKETING CAMPAINGS

Esra AKDENİZ DURAN[*1], Ayça PAMUKCU[2], Hazal BOZKURT[2]

[1]*İstanbul Medeniyet University, Faculty of Sciences, Department of Statistics, Üsküdar-İSTANBUL*
[2]*Marmara University, Faculty of Arts and Sciences, Department of Statistics, Göztepe-İSTANBUL*

Received/Geliş: 27.07.2013 Revised/Düzeltme: 19.11.2013 Accepted/Kabul: 05.02.2014

ABSTRACT

The intensive increase in the competition of marketing campaigns over time reduced the impact of them on customer base. Economic pressures, intense competition in the industry, changing lifestyles of people and developing technology have caused marketing managers to adopt the concept of direct marketing by entering into new pursuits. The campaigns prepared in accordance with this understanding might be improved using a variety of data mining techniques. This study compares the performances of artifical neural networks, logistic regression and decision tree data mining techniques on a direct marketing campaign. The purpose of the study is to determine the best target group involved in the campaign by comparing estimation powers of the methods used for determining target groups. Based on the results of this study, it is revealed that artificial neural networks method is more reliable than decision tree and logistic regression analysis about estimating the likely responders in the campaign. This model can improve the efficiency of campaigns by determining of the main features that affect the success of the campaign, identifying the best target group and managing of resources.
Keywords: Direct marketing, data mining, decision trees, artificial neural networks.

VERİ MADENCİLİĞİ YÖNTEMLERİNİN DOĞRUDAN PAZARLAMA KAMPANYALARI İÇİN KARŞILAŞTIRILMASI

ÖZET

Pazarlama kampanyalarının sayısının artışı, düzenlenen kampanyaların müşteri kitlesi üzerindeki etkisini giderek azaltmıştır. Ayrıca ekonomik baskılar, sektördeki yoğun rekabet, insanların değişen yaşam tarzları ve gelişen teknoloji, pazarlama yöneticilerinin yeni arayışlar içerisine girerek, doğrudan pazarlama anlayışını benimsemelerine neden olmuştur. Doğrudan pazarlamada hedef kitlesinin ve kampanyanın olumlu sonuçlanmasını etkileyen faktörlerin belirlenmesi için veri madenciliği yöntemleri kullanılmaktadır. Bu çalışmada, bir bankanın doğrudan pazarlama kampanyası verileri yapay sinir ağları, lojistik regresyon ve karar ağaçları ile analiz edilmiştir. Çalışmanın amacı, kampanyaya katılabilecek hedef müşteri kitlesini belirleyebilmek için belirtilen yöntemlerin tahmin güçlerini karşılaştırılarak kampanyayı en iyi açıklayan modelin belirlenmesidir. Çalışma sonucunda, yapay sinir ağları yönteminin kampanyaya katılacak müşteri kitlesinin tahmininde, karar ağaçları ve lojistik regresyon modellerine göre daha iyi sonuç verdiği bulunmuştur. Bu model, hedef müşteri kitlesinin en iyi şekilde seçilmesinde, kampanyanın başarısını etkileyen temel özelliklerin belirlenmesinde ve kaynakların yönetiminde etkili olarak, kampanyaların etkinliğini arttırabilir.
Anahtar Sözcükler: Doğrudan pazarlama, veri madenciliği, karar ağaçları, yapay sinir ağları.

[*]Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: esra.duran@medeniyet.edu.tr, tel: (216) 280 34 75

1. INTRODUCTION

Data mining is becoming a strategically important tool for many business organizations including the banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the banks to look for hidden patterns in a group and discover unknown relationships in the data. Today, customers have so many options with regard to where to do their business. Data mining techniques facilitate useful data interpretations for the banking sector to avoid customer attrition. Customer retention is the most important factor to be analyzed in today's competitive business environment. The customers are exposed to different campaigns all the time, thus it is difficult to attract their attention by organizing mass campaigns [15]. Although direct marketing is a way to communicate straight to the customers, it can be very costly to organize customer specific campaigns considering the huge customer database. It is therefore more convenient to access to those who are likely responders. Data mining can provide an effective tool to reveal potential responders and thus reduce direct marketing costs [7].

Comparison of data mining methods have been widely applied in other fields successively. Das (2010) used data mining techniques for diagnosis of Parkinson disease, Khemphila and Boonjing (2010) compared the neural networks, the decision trees and the logistic regression model in classsifiying heart disease patients, Budak and Erpolat (2012) employed neural networks and logistic regression in the credit risk prediction. In our study, we used data from Portuguese marketing campaign related with bank deposit subscription which was first described and analyzed in Moro et al.(2011). Moro et al. (2011) compares naive bayes, decision tree and support vector machines methods in their paper. They reduce the size of the data set as some of the methods are computationally intensive. In our study, we use the full data set and compare performances of decision tree, logistic regression and neural network models on this data set. The classification goal is to predict whether the client will subscribe a term deposit which is a deposit held at a financial institution for a fixed term. We employed SAS Enterprise Miner 7.1 software for computational purposes.

The rest of the paper is organized as follows. Section 2 introduces data mining techniques and SAS nodes used in the study. Section 3 describes the case study of direct marketing in banks and presents the experimental results. Sections 4 summarizes the findings. Sections 5 presents the conclusion.

2. DATA MINING TECHNIQUES

The brief description of each node is given in the following.

Data Partition

The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. Data partition node in SAS was used to randomly partition the input data into train (60%) and validation data (40%) sets.

Transform Variables

Transform variables node was used to transform variables in data set which have highly skewed distributions. With this node logaritmic transformation etc. can be applied on variables with skewed distributions[5].

The data mining techniques that will be emploted are logistic regression, decision tree and neural network models.

Logistic Regression

Logistic regression analysis examines the influence of various factors on a dichotomous outcome by estimating the probability of the event's occurrence. In our data set the dependent variable

(DV) is binary therefore we used binomial/binary logistic regression method [11]. The following equation describes linear regression model with *k* independent variables (IV):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (1)$$

Odds ratio in the logistic regression is known as a ratio of the number of people incurring an event to the number of people who have non-events. The natural logarithm of odds ratio is called ‘‘Logit function’’. Relation between logit function and the linear regression model is given in the following equation:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon = \text{logit}(p) \quad (2)$$

This model assumes a linear relationship between the logit of the IVs and DVs. The logistic regression model may be written in terms of *p*, the risk of event as follows: y

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon)}} = \frac{1}{1 + e^{-\text{logit}(p)}} \quad (3)$$

Another popular data mining technique is decision trees.

Decision Tree

Decision trees are powerful and popular data mining techniques for both classification and prediction. Tree-based methods are simple and useful for interpretation. Each variable is represented by a node in the decision tree. Decision trees have various splitting criteria. Lift splitting criteria which we used in this study classifies the variables and allows us to obtain the optimum tree [5].

Which model is better depends on the problem at hand. If the relationship between the IVs and the DV is well approximated by a linear model then logistic regression will likely work well, and will outperform a method such as a decision tree that does not exploit this linear structure. If instead there is a highly non-linear and complex relationship between the IVs and the DV, then decision trees may outperform classical approaches. Another popular classification method is Neural Networks.

Neural Network

Neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. A neural network typically has an input layer, an output layer, and one or more hidden layers. Every neuron produces an output signal. All the input signals reach the neuron simultaneously, so the neuron receives more than one input signal, but it produces only one output signal. Every input signal is associated with a connection weight. The weight determines the relative importance the input signal can have in producing the final impulse transmitted by the neuron [14]. In this study, we used feed-forward neural network model. Neural networks are attractive as automatic model-building tools because they are able to represent any well-behaved function with arbitrary non, relatively robust to outliers and poor data. One disadvantage of neural network models is, it is almost impossible to interpret them however resistance to using these “black boxes” is gradually diminishing as more researchers use them.

Model Comparison

Model comparison node allows comparing models more directly. We use the Model Comparison node to benchmark model performance and find a champion model among the Neural Network, Regression and Decision Tree nodes in our process flow diagram. Visual representations to compare models are provided with the receiver operating ROC curve/index and the

lift/cumulative lift functions. In addition, fit statistics allow making a more direct comparison between models [3].

Lift/Cumulative Lift Chart

The Lift Chart measures the effectiveness of models by calculating the ratio between the result obtained with a model and the result obtained without a model. It is the most commonly used metric to measure the performance of targeting models in marketing applications. The Cumulative Lift Chart represents the lift factor which shows how many times it is better to use a model in contrast to not using a model [14].

ROC Curve/Index

Receiver operating characteristic (ROC) curves are useful for assessing the accuracy of predictions. The area under the ROC curve indicates the discrimination power of the model. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space. The (0,1) point is also called a perfect classification.

ROC Index is measured by the area under the ROC curve. It takes avalue between "0" and "1". An area close to 1 represents a good model; an area close to 0.5 represents a worthless model [14].

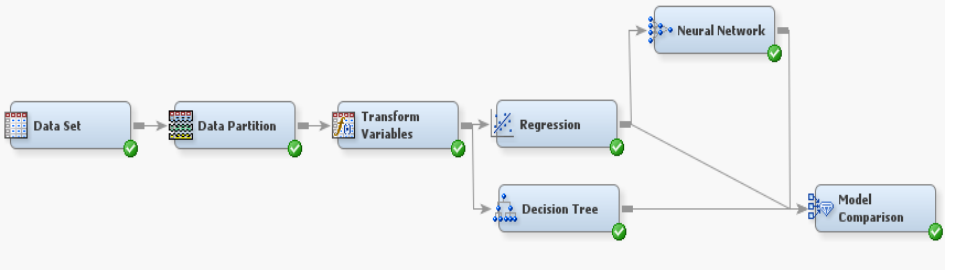

Figure 1. SAS Enterprise Miner diagram

SAS Enterprise Miner diagram in Figure 1 shows the nodes employed in the analyses. Next, the case study will be explained and the methods will be compared based on the metrics explained in this section.

3. APPLICATION

In this study, we used the data set collected from a Portuguese bank for a direct marketing campaign. It was first described and analyzed in Moro et al. (2011). The data set was extracted from missing values and ineffective variables so that there exists 45211 instances with 16 independent variables and 1 target variable in the model. The information about variables are given in Table 1.

Target variable is specified as whether the client will subscribe a term deposit. It is a binary variable where "Yes" represents the value for the clients who subscribed in the campaign and "No" represents the value for clients who did not subscribe in the campaign. 39922 person did not subscribe a term deposit while 5289 did subscribe. It is shown in Figure 2 below.

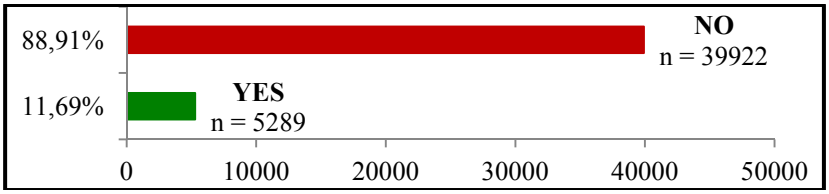

Figure 2. The number and the percentage of customers who participating in the campaign

Table 1. Information about variables

Variables	Values	
Age	Numeric	Numeric
Job	Categorical	Categorical
Marital status	Married, Divorced/Widowed/Separated, Single	1- 3
Education	Unknown, Elementary, Middle/High, Higher	1-4
Payment status	Yes, No	1, 0
Annual balance	Numeric	Numeric
Loan	Yes, No	1, 0
Consumer credit	Yes, No	1, 0
Communication channel	Unknown, Landline, Cellular	1-3
Days since last contact	Numeric	Numeric
Months since last contact	Categorical	Categorical
Call duration (seconds)	Numeric	Numeric
Number of contacts	Numeric	Numeric
Days between a last campaign	Numeric	Numeric
The number of repeated contacts	Numeric	Numeric
Result of the last campaign	Unknown, Other, Unsuccessful, Successful	1-4

Table 2. Descriptive statistics of the numerical variables

Variables	Mean	Standard Deviation	Minimum	Maximum	Median
Age	41	10.619	18	95	39
Annual balance	1362.272	3044.766	-8019	102127	448
Number of contacts	2.764	3.098	1	63	2
Days since last contact	15.8	8.322	1	31	16
Call duration (seconds)	258.163	257.528	0	4918	180
Days between first-last campaign	40.198	100.129	-1	871	-1
The number of repeated contacts	0.580	2.303	0	275	0

The data is divided into two groups with data partition node where 60% is for training and 40% is for validation. Data segmentation and analysis of models are conducted in SAS Enterprise Miner 7.1 software package.

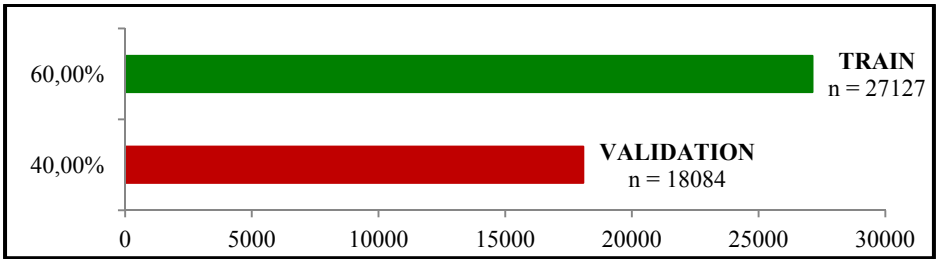

Figure 3. The number of customers allocated to train and validation

The variables "Call duration", "Annual balance" and "Number of repeated contacts" have skewed distributions therefore logarithmic transformation was applied to these three variables with Transform Variables node.

Lift criteria was selected as a splitting criteria for the decision tree so that unnecessary nodes were pruned and optimum tree was obtained. The results of the decision tree model are given in Table 3.

Table 3. Classification results of the decision tree analysis

		Train			Validation		
		Number of people		Classification Rate (%)	Number of people		Classification Rate (%)
Number of people		No	Yes		No	Yes	
	No	23112	754	96.8	15523	533	96.7
	Yes	1817	1444	44.3	1190	838	41.3
Total		27127		90.5	18084		90.5

Classification performance table of the decision tree model presented in Table 3 shows that both the train and the validation data set have classification accuracy of 90.5%. Classification accuracy for clients who answered "no" to the campaign is 96.5 % in train data set while it is 96.7% in validation data set. Customers who answered "yes" to the campaign are classified with 44.3 % accuracy in train data set and with 41.3% accuracy in validation data set.

Logistic regression analysis is another method we applied on our data set as the dependent variable is a binary variable. Stepwise method was used in logistic regression analysis in order to avoid overfitting . The results of the logistic regression model are given in Table 4.

Table 4. Classification results of the logistic regression analysis

		Train			Validation		
		Number of people		Classification Rate (%)	Number of people		Classification Rate (%)
Number of people		No	Yes		No	Yes	
	No	23249	617	97.4	15651	405	97.5
	Yes	2029	1232	37.8	1314	714	35.2
Total		27127		90.2	18084		90.5

Classification performance of the Logistic regression model presented in Table 4 shows that train data set has 90.2% classification accuracy and validation data set has 90.5% classification accuracy. Customers who answered "no" to the campaign are classified with 97.4 % accuracy in train data set and with 97.5% accuracy in validation data set. Customers who answered "yes" to the campaign are classified with 37.8 % accuracy in train data set and with 35.2% accuracy in validation data set.

Multi-layer, feed- forward neural network model is also applied on the data set. A neural network typically has an input layer, an output layer, and one or more hidden layers. One disadvantage of artificial neural networks is that it does not have a variable selection procedure. We connected regression node before running artificial neural network node to do variable selection and thus avoid overfitting. As a result of this prosess, 12 input variables extracted by the stepwise regression analysis entered into neural network model. The variables used in both logistic regression and ANN models are presented in Figure 4. According to Figure 4; "call duration", "result of the last campaign", "months since last contact", "number of contacts", "consumer credit", "job", "the number of repeated contacts" , "marital status", "loan", "days between a last campaign ", "education" and "annual balance" are important variables that might affect the campaign.

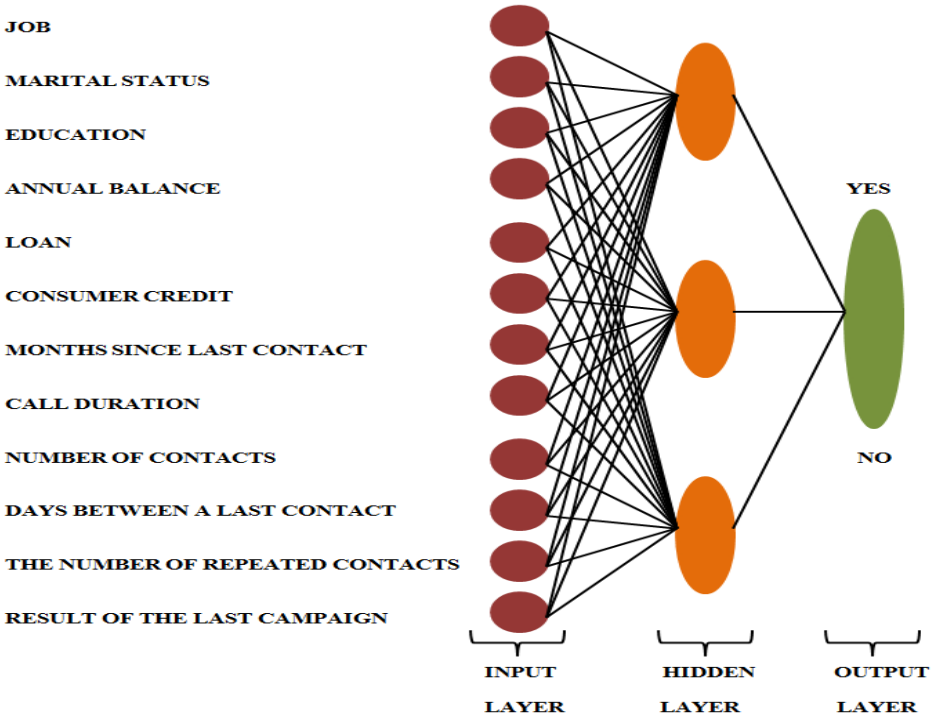

Figure 4. Artificial Neural Network model with the reduced variables

The results of the ANN model are given in Table 5.

Table 5. Classification results of the neural network analysis

		Train			Validation		
		Number of people		Classification Rate (%)	Number of people		Classification Rate (%)
Number of people		No	Yes		No	Yes	
	No	23102	764	96.8	15500	556	96.5
	Yes	1728	1533	47	1113	915	45.1
Total		27127		90.8	18084		90.8

The results of the artificial neural network model presented in Table 5 show that both the train and validation data set have 90.8% classification accuracy. Customers who answered "no" to the campaign are classified with 96.8 % accuracy in train data set and with 96.5% accuracy in validation data set. Customers who answered "yes" to the campaign are classified with 47 % accuracy rate in train data set and with 45.1% accuracy rate in validation data set.

The degree of importance of the variables are also found with decision tree analysis and shown in Table 6.The first six variables that has the highest impact on the decision of the customer within the campaign are given in the following table.

Table 6. Degree of importance fort he first 6 variables

Variables	Degree of importance
Call duration (seconds)	1
Result of the last campaign	0.702534
Months since last contact	0.483706
Age	0.238228
Communication channel	0.209023
Loan	0.200484

In addition to the individual model performance indicators, we wanted to compare performances of models with each other. Model Comparison node in SAS Enterprise Miner 7.1 is used to compare the predictive power of our methods. As a result, Lift and Cumulative lift charts, ROC Curve and ROC Index are obtained as follows.

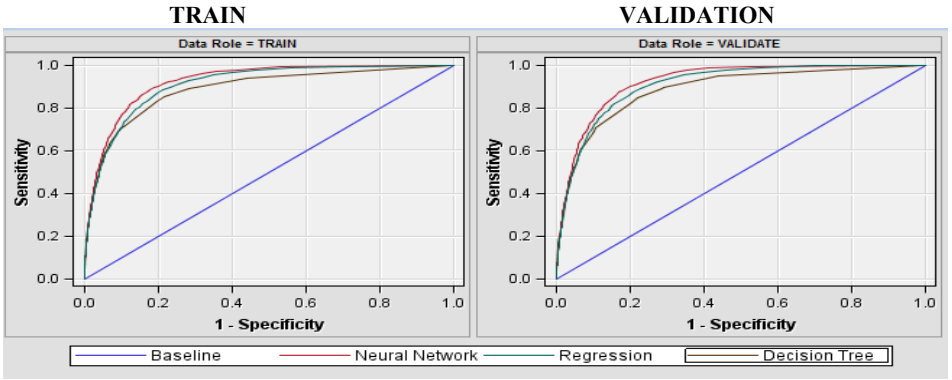

Figure 5. ROC Curves for the compared models

Receiver Operating Characteristic(ROC) Curves for all models are presented in Figure 5. The area under the this curve indicates the discrimination of accuracy rate from the customers who paticipated and who did not participate to the campaign. The area under the ROC curve is represented the ROC index values. ROC Indices for the compared models are given in the Table 7.

Table 7. ROC Index for comparable models

MODEL	TRAIN	VALIDATION
	ROC Index	ROC Index
Neural Network	0.92	0.92
Logistic Regression	0.91	0.91
Decision Tree	0.89	0.88

In table 7 ROC Index of ANN model is 0.92 for both train and validation data set. ROC Index for logistic regression model is 0.91 for both training and validation data set. And finally ROC Index of decision trees is 0.89 for train and 0.88 for validation data set.

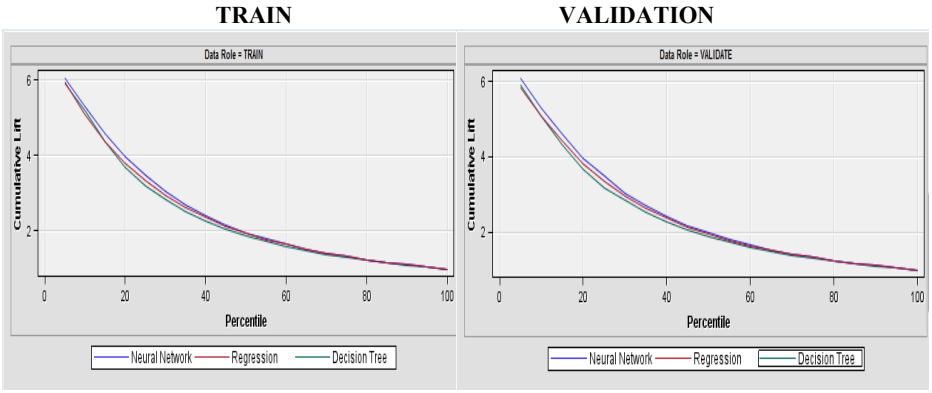

Figure 6. Cumulative Lift charts for the compared models

Cumulative lift chart in Figure 6 has a lift value of 3.80 for the first 20% slice of the logistic regression model, 3.69 for the first 20% slice of the decision trees model and 3.98 for the first 20% slice of the artificial neural network model.

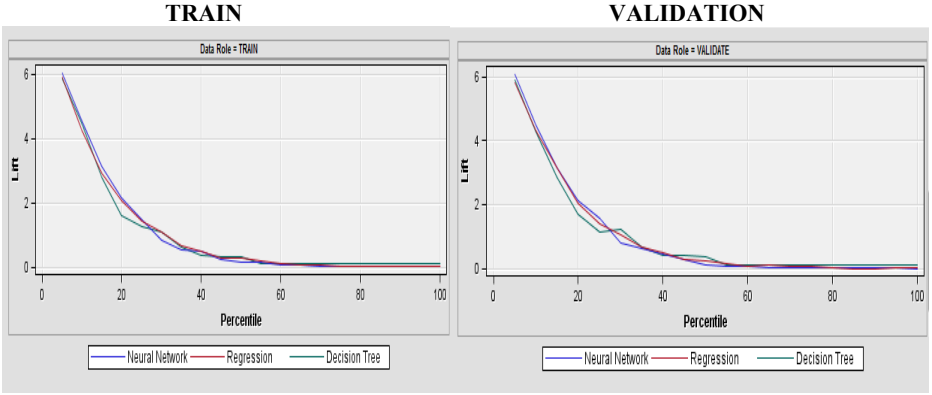

Figure 7. Lift charts for the compared models

Thus the model is predicted 3.80 times better and more accurate in target values of this range according to a random group in the data set.

4. RESULTS

The success of banks depends on the effective use of direct marketing strategies. In this study, we applied artificial neural networks, decision trees and logistic regression data mining methods in the banking sector and compared the predictive power of these methods. This section describes the efficiency of models that we built.

Analysis of the classification performances of the methods is examined; the logistic regression model has 90.2% and 90.5% classification accuracy for train data set and validation data set respectively. The decision tree model has the classification accuracy of 90.5% both for train and validation data sets. And finally, artificial neural network model has 90.8% classification accuracy both for train and validation data sets. The percentage of true classification of artificial neural networks is slightly higher than logistic regression and decision trees methods. In addition, SAS Enterprise Miner 7.1 software package chose artificial neural network model as the best model automatically.

ROC curves and indices of the models are also examined. The area under the ROC curve (ROC Index) is widely recognized as the measure of a diagnotic test's discriminatory power. The ROC curve of artificial neural network model is slightly higher than other models both for train and validation data sets. Thus, artificial neural network has a slightly higher discriminatory power than other models.

Cumulative lift and lift charts are other visual aids for measuring model performance. The cumulative lift and lift chart (Figure 6,7) show how much more likely we are to receive responses from customers that were selected from predictive model than if we contact a random sample of customers. For example; looking at the Cumulative Lift chart for train data, we see that the lift zone of the first 20% is 3.80 for logistic regression model, 3.69 for decision tree model and 3.98 for artificial network model. These results illustrate for example using artificial meural networks model the customers at top 20% decile are 3.98 times more likely to respond than would be expected. As can be seen from model performance criterias, artificial neural networks model is slightly better than decision trees and logistic regression models.

As a result, customers' participation in the campaign is affected by some of the variables in the model. In Table 6, we see that "call time" seems to be most effective variable on the success of the campaign. This result suggests that Bank of Portugal may increase the rate of participation in the campaign by doing interviews with customers 258 seconds on average in future campaigns. The other important variables are “result of the last campaign”, “months since last contact”, “age”, “communication channel” and “loan”.

In this study we determined the model which is more effective on direct marketing campaigns and also determined the variables which affect the rate of participation in the campaign. Artificial neural network model is found to have slightly better classification accuracy than logistic regression and decision tree analysis. As a result, the bank can eliminate potential participating customers by using the artifical neural networks model and can take position with respect to the important variables to increase efficiency of the campaign.

5. CONCLUSION

We demonstrated that data mining is an effective tool for direct marketing which can improve bank marketing campaigns. Most research papers focus on computational and theoretical aspects of direct marketing though little efforts have been put on technological aspects of applying data mining in the direct market process. The complexity of the data mining models makes it difficult for marketers to use it, hence; we outlined a simplified framework to guide marketers and managers in making use of data mining methods and focus their advertising and promotion on those categories of people in order to reduce time and costs. We explained all the steps and tasks that are carried out at each stage of the data mining framework. The case study we evaluated shows the practical use and usefulness of the model.

REFERENCES / KAYNAKLAR

[1] Moro, S., Laureano, R.M.S., Cortez, P., “Using Data Mining for Bank Direct Marketing: An Application of the CRISP – DM Methodology”, Proceedings of the European

Simulation and Modeling Conference, ESM'2011, 117-121, Guimarães, Portugal, October, 2011.

[2] Budak, H.,Erpolat, S., (2012), Kredi risk tahmininde yapay sinir ağları ve lojistik regresyon analizi karşılaştırılması, AJIT-e: Online Academic Journal of Information Technology, 3 (9), 23-30.

[3] Das, R., (2010), A comparison of multiple classification methods for diagnosis of Parkinson disease, Expert Systems with Applications, 37 (2010), 1568-1572.

[4] Khemphila, A., Boonjing, V., "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients", International Conference on Computer Information Systems and Industrial Management Applications (CISIM), AGH University of Science and Technology, 193-198, Cracow, Poland, October, 2010.

[5] SAS, (2007), "Applied Analytics Using SAS® Enterprise Miner™ 5 Course Notes, NC, USA, SAS Institute Inc..

[6] Berry, M.J.A., Linoff, G.S., (1999), "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", John Wiley&Sons, Inc., 3rd Edition, p. 888.

[7] Nash, E., (2000), "Direct Marketing: Strategy, Planning, Execution", McGraw Hill Professional, 4th Edition, p. 600.

[8] Emel, G.G., Taşkın, Ç., (2005), Veri madenciliğinde karar ağaçları ve bir satış analizi uygulaması, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6 (2), 221-239.

[9] Tan, P.N., Steinbach, M., Kumar, V., (2013), "Introduction to Data Mining: Pearson New International Edition", Pearson Higher& Professional EMA, First Edition, p. 736.

[10] Yalçın, Ö., (2008), "Veri Madenciliği Yöntemleri", Papatya Yayıncılık Eğitim A.Ş., İstanbul, 2. Edition, p. 216.

[11] Gujarati, D.N., Porter, D.C., (2012), "Temel Ekonometri" (G.G. Şenesen, Ü. Şenesen, Çev.), Literatür Yayıncılık, İstanbul, 5th Edition, p. 972.

[12] Aktaş, C., (2009), Lojistik Regresyon Analizi: Öğrencilerin sigara içme alışkanlıkları üzerine bir uygulama, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 26, 107-121.

[13] Ling, C.X., Li, C., (1998), Data Mining for Direct Marketing: Problems and Solutions, In The Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, USA, 73-79.

[14] Giudici, P., (2003), "Applied Data Mining: Statistical Methods for Business and Industry", John Wiley&Sons, Inc., First Edition, p. 376.

[15] Chitra, K., Subashini, B., (2013) , Data Mining Techniques and its Applications in Banking Sector, International Journal of Emerging Technology and Advanced Engineering, 3 (8), 219-226.

Electrical-Electronics Engineering Article
/
Elektrik-Elektronik Mühendisliği Makalesi