Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 15 (2014) 238 – 246

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Emerging Market Queries in Finance and Business EMQFB2013

Predicting Consumer Behavior with Artificial Neural Networks

Laura Maria Badea (Stroie)[a,*]

[a]*The Bucharest University of Economic Studies, Calea Dorobantilor no. 15-17, Bucharest, 010552, Romania*

Abstract

Nowadays, facile access to information and advancements in processing power unfold opportunities for new decision support techniques used for financial and economic purposes. Artificial neural networks are machine learning techniques which integrate a series of features upholding their use in financial and economic applications. Backed up by flexibility in dealing with various types of data and high accuracy in making predictions, these techniques bring substantial benefits to business activities. This paper investigates how consumer behavior can be identified using artificial neural networks, based on information obtained from traditional surveys. Results highlight that neural networks have a good discriminatory power, generally providing better results compared with traditional discriminant analysis.

Selection and peer-review under responsibility of the Emerging Markets Queries in Finance and Business local organization

Keyworks: artificial neural networks; discriminant analysis; consumer behavior; supervised learning; market segmentation

1. Introduction

Market segmentation is a process that requires identifying homogeneous groups of consumers described by a set of similar characteristics, in order to improve marketing activities through a better allocation of resources and formulation of customizable strategies. When target groups are a-priori known, the problem becomes a classification task, under a process of supervised learning. Increased interest in identifying new liquidity sources forces financial institutions to investigate new ways of detecting individuals with high propensity towards saving money. Traditional statistical methods such as discriminant analysis have often been used in classification tasks, providing good results. However, the need to obtain even more accurate results has directed researchers' interest towards non-parametric classification techniques like artificial neural networks (henceforth also ANNs). Neural networks are Artificial Intelligence representations inspired by the way in which the

* Corresponding author. Tel.: +40-746-090-686
E-mail address: laura.maria.badea@gmail.com

Selection and peer-review under responsibility of the Emerging Markets Queries in Finance and Business local organization
doi:10.1016/S2212-5671(14)00492-4

human brain functions. They are statistical data modeling techniques in which interconnected elements (called nodes) process simultaneously the information, adapting and learning from past examples. Although introduced a long time ago (by McCulloch and Pitts, 1943), ANNs have gained popularity only recently, supported by the explosive expansion of computer usage and by those good results emphasized in a series of empirical studies (Zhang et al., 2008; Kim and Ahn, 2009; Hakimpoor et al., 2011; Štencl et al., 2012; Felea et al., 2012). The main purpose of this study is to analyze results obtained when building a model that identifies individuals with great chances of making bank deposits, using artificial neural networks. In order to make an assessment of ANNs' performance, classification results were compared with those obtained from applying linear discriminant analysis. The paper is structured as follows: section two provides the methodology used for building a classification model with discriminant analysis and ANNs. Section three details the aspects regarding the available data, clean-up processes, and the partitioning methodology employed. Section four describes the configuration steps performed for building the models. Section five provides the results obtained and the last section concludes in respect with the use of ANNs in consumer behavior applications, suggesting also further research directions.

2. Methodology

2.1. Discriminant analysis

Driven from linear probabilistic methods, discriminant analysis (DA), also known as Fisher linear discriminant analysis (Fisher, 1936), is commonly used in pattern recognition and dimensionality reduction. The basic principle stands in finding linear combinations of characteristics to classify objects within certain groups. The optimization of class separability within discriminant analysis technique is based on variance maximization between objects pertaining to different classes and variance minimization between objects within the same class. Thus, Fisher discriminant functions are linear combinations of predictors under the following form:

$$D(X) = \beta^T X \tag{1}$$

where, X is the vector of predictors, and β is the eigenvector of the matrix $\hat{\Sigma}_w^{-1} \cdot \hat{\Sigma}_b$, with $\hat{\Sigma}_w$ being the intra-class covariance matrix and $\hat{\Sigma}_b$ the inter-class covariance matrix.

2.2. Artificial neural networks

Artificial neural networks are non-parametric methods used for pattern recognition and optimization. They generate a signal or an outcome based on a weighted sum of inputs which is afterwards passed through an activation function as in equation (2):

$$Y = f(\sum WX) \tag{2}$$

where, X is the vector of inputs, W is the vector of weights, $f(.)$ is the activation function and Y is the output vector.

Of all ANN types used in classification matters, a viable option is the multilayer perceptron (MLP) which is organized in three types of layers: an input layer, hidden layers (usually not more than three) and an output layer. Within MLPs, the information flow is processed in a feed-forward manner, and all elements in a layer are fully connected to the nodes from the upcoming layer. The training process within MLPs is performed by back-propagating the errors and adjusting the network weights correspondingly, in order to decrease the deviations of the outputs from the target values. Activation functions are applied to the weighted sum of the inputs of a node to generate a certain outcome. As the non-linear character of ANNs is given by the form of the activation

functions, the most common types, especially within the hidden layers, are those taking a non-linear form. Among these, sigmoid ("*S*" shape) functions are often preferred for their continuous character which makes possible differentiation, an important feature when training with back-propagation, but also for their bounded range, which makes them easily interpreted. Logistic function ($f(x)=\frac{1}{1+e^{-x}}$) is a type of sigmoid function that has been previously employed in classification tasks (Felea et al., 2012), generating good results. Given a three layer MLP, with the back-propagation training algorithm, the learning is performed by employing several training cycles which consist of passing through the model a set of *m* input-output pairs $\left(x^d, t^d\right)$, $d=1,...,m$, which form the training sample, and adjusting the network weights to decrease the value of the deviations of the model outputs, y^d, from the target values, t^d. Initially, the network weights are set to small random values. Afterwards, the input pattern is applied and propagated through the network until a certain output is generated for the hidden layer:

$$h_j^d = f(net_j^d) = f(\sum_k w_{jk} x_k^d) \tag{3}$$

where, *d=1, …,m* is the number of input-output pairs available in the training set, h_j^d is the output of the hidden unit *j*, net_j^d is the input of the hidden node *j*, x_k^d is the input node *k*, w_{jk} is the weight given to input *k* for hidden node *j*, and $f(.)$ is the activation function used in the hidden layer.

These outputs of the hidden layer, h_j^d, are next used as entries for the output layer. Weighted and summed up, they are passed through an activation function to produce the final output:

$$y_i^d = g(net_i^d) = g(\sum_j w_{ij} h_j^d) = g(\sum_j w_{ij} \cdot f(\sum_k w_{jk} x_k^d)) \tag{4}$$

where, y_i^d is the exit value of the output unit *i*, net_i^d is the input of the output node *i*, w_{ij} is the weight given to the hidden node *j* for the output node *i*, and $g(.)$ is the activation function used in the output layer.

The way in which the network weights are modified to meet the desired results defines the training algorithm and is essentially an optimization problem. When the activation functions are differentiable, the error back-propagation algorithm becomes a good approach in progressing towards the minimum of the error function, $E(w)$. Consider the form of the error function bellow:

$$E(w) = \frac{1}{2}\sum_{d=1}^{m}\left(t^d - y^d\right)^2 \tag{5}$$

For two output nodes (*i ={1, 2}*), the error becomes:

$$E(w) = \frac{1}{2}\sum_{d=1}^{m}\sum_{i=1}^{2}\left(t_i^{\,d} - g(\sum_j w_{ij} \cdot f(\sum_k w_{jk} x_k^d))\right)^2 \tag{6}$$

Then, the errors are passed back through the network using the gradient, by calculating the contribution of each hidden node and deriving the adjustments needed to generate an output that is closer to the target value. For the hidden to output and for the input to hidden connections, the gradients are computed like in equations (7) and (8) respectively:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = \eta \sum_{d=1}^{m} (t_i^d - y_i^d) g'(net_i^d) \cdot (h_j^d) \quad (7)$$

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \sum_{d=1}^{m} \frac{\partial E}{\partial h_j^d} \cdot \frac{\partial h_j^d}{\partial w_{jk}} = \eta \sum_{d=1}^{m} \sum_{i=1}^{2} (t_i^d - y_i^d) g'(net_i^d) \cdot w_{ij} \cdot f'(net_j^d) \cdot x_k^d \quad (8)$$

where, η is the learning rate which controls the size of the step in each cycle, $g'(.)$ is the first order derivate of function $g(.)$, $f'(.)$ is the first order derivate of function $f(.)$.

The new weights can be adjusted taking also into account the modification from the previous cycle, this method being called back-propagation with momentum rate. This parameter is used to speed up the convergence process in flat regions, or to diminish the jumps in regions of high fluctuations, by adding a fraction of the previous weight change.

$$\Delta w(n+1) = -\eta \frac{\partial E}{\partial w} + \alpha \Delta w(n) \quad (9)$$

where, α is the momentum rate, $\Delta w(n+1)$ is the weight modification from cycle *n+1*, and $\Delta w(n)$ is the modification from the previous cycle, *n*.

3. Dataset and variables

The present study is based on the data collected by the Romanian Academy's Institute of World Economy, under a program sponsored by the World Bank. The survey, referenced as "ROU_2010_FLS_v01_M" has its data available with public access under the name "Romania - Financial Literacy and Financial Services Survey 2010" on the World Bank Microdata Repository, being downloaded on 14th of March 2013 from the following URL: "http://microdata.worldbank.org/index.php/catalog/1027/datafiles". The survey contains two datasets, one collected on individuals and the second collected on households, but the latter is more generous in terms of available variables. Therefore, “Household_RO” questionnaire was further used for analysis, as it also provides information at individual level like age, occupation, income, education and so on. The original database consisted of 2,389 observations collected at household level. After performing filtering operations such as removing observations with missing information, 1,671 records were kept for the analysis, generating a utility rate of about 70% of the initial dataset. These records were afterwards divided into two distinct classes of individuals: “with bank deposit” and “without bank deposit”. The class was assigned using the fields SF1, SF2, SF3, SF4, SF5, and SF6 available in the survey, which give information whether a household has, or does not have a bank deposit. As the survey initially contained 310 variables, for further evaluation, a primary space reduction was performed based on the relevance towards the analyzed problem. After this, 14 categorical factors and two continuous variables were retained (see Fig. 1). Numerical variable “respondent monthly income” was further split into five categories, the first two having as borders the minimum and respectively the average incomes in local currency (RON) at national level provided by the National Institute of Statistics from Romania for April 2010, the point when the survey was done. Variable “age” was also split into eight categories, using an expert judgment approach. Further modifications consisting of regrouping the available categories, were performed on the following fields: “occupation”, “education” and “nationality”.

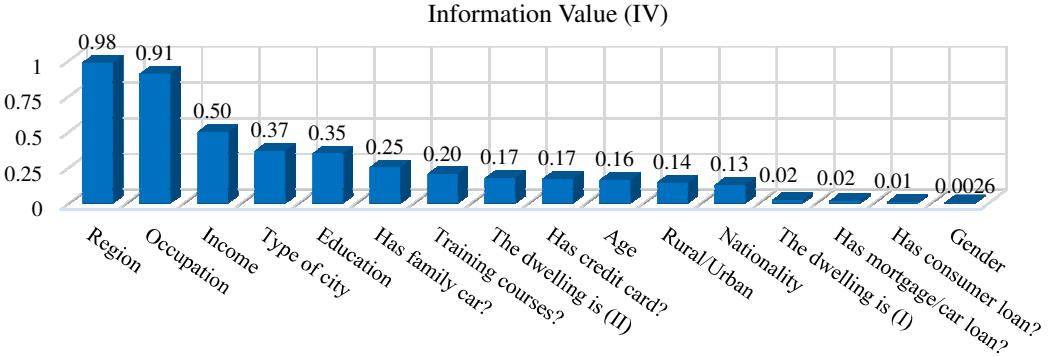
The relevance of each independent variable towards the end class was tested using Information Value (*IV*) as a metric for uni-variate analysis, helping to better understand the available data:

$$IV_{x_k} = \sum_{c=1}^{p} \left(\left(\%event_{x_k^c} - \%non\_event_{x_k^c} \right) \cdot WOE_{x_k^c} \right) \quad (10)$$

where, IV_{x_k} is the Information Value calculated for variable x_k, $k = 1,2, ..., q$, x_k^c denotes the category c of variable x_k, $c = 1,2, ..., p$, p shows the number of categories for a given variable x_k, $\%event_{x_k^c}$ shows the ratio of records from class "with bank deposit" in category c of variable x_k within the total number of records from class "with bank deposit", $\%non\_event_{x_k^c}$ shows the proportion of records from class "without bank deposit" in category c of variable x_k within the total number of records from class "without bank deposit", and $WOE_{x_k^c}$ is the Weights of Evidence ratio calculated for a certain category c of variable x_k which is determined like in equation (11):

$$WOE_{x_k^c} = \ln\left(\frac{\%event_{x_k^c}}{\%non\_event_{x_k^c}} \right) \quad (11)$$

Usually, an Information Value above 0.1 reveals a good discriminatory power (Siddiqi, 2006). Fig. 1 provides the results of *IV* metric for all 16 independent variables, sorted in descending order, from the most relevant to the least one.



Source: own calculation
Fig. 1. Information Value

For development purposes, the initial database was randomly split into three data samples, each having a specific role in the training process. As there is no general accepted rule for selecting the partitioning ratios, the choice is usually made by the expert considerations. However, it should be noted that the final model results are strongly related to the partitioning process, especially when the overall database is rather limited. Thus, for the current analysis, the database consisting of 1,671 observations was randomly split using the following ratios:

- Training set – 50% of the initial dataset, which was used for model development;
- Validation set – 25% of the initial dataset, used for model assessment when training ANN model. The error on this set will be compared with the training error, controlling the generalization capacity of the model.

- Test set – the remaining 25% of the data, for out-of-sample evaluation of the model built.

4. Model building

4.1. Discriminant analysis

Using discriminant analysis for data modeling requires performing certain data transformations and hypotheses testing. To concord with the requirement of having continuous predictors, all available variables were transformed from nominal into interval using the metric *WOE*, calculated as in equation (11). This was considered to be a good approach of ordering all categories in a given variable by their observed probability of pertaining to class "with bank deposit". Further these interval ratios were treated as continuous variables. To test the level of significance of available variables, forward stepwise method was used, considering a 0.05 significance level for both, p-value entry and p-value exit. Results of forward stepwise regression indicate the following factors as relevant for further model development with discriminant analysis: "region", "type of city", "age", "occupation", "income", "has credit card?" and "has family car?". A-priori class probabilities were those estimated from the available data. The model was built using training and validation datasets, leaving the test set for out-of-sample evaluation. Table 1 shows Fisher discriminant functions coefficients, which also indicate the importance of each variable included.

Table 1. Discriminant function coefficients

Variable	Discriminant function coefficients
Region	0.513265
Income	0.421798
Has credit card?	0.325027
Type of city	0.242234
Has family car?	0.229067
Age	0.216652
Occupation	0.197305

4.2. Artificial neural networks

Modeling data with artificial neural networks allows a flexible approach towards independent variables. Thus, unlike in the case of discriminant analysis, ANNs did not require a very thorough data preprocessing. Consequently, all 16 variables available were kept in the model in a categorical form, to shape the input layer of the neural networks through the available categories (70 categories). Nevertheless, building models with ANNs requires several configuration steps which relate to setting elements like: the network type, the number of hidden layers and hidden nodes, the activation functions in the hidden and output layers, the error function, the optimization algorithm used to adjust the synaptic weights, and the stopping criterion. For this analysis a multilayer perceptron was used to model consumer data for the Romanian market. Although the number of hidden layers generally ranges between one and three, previous studies (Knerr et al., 1992) have shown that ANNs with a single hidden layer can estimate any differentiable function, provided that they have enough hidden units. Moreover, a high number of layers would significantly increase the processing time and the adjustments required during network training. Therefore, only one hidden layer was included. The number of nodes in the hidden layer was varied between a minimum of two units and a maximum of 12 units. Featuring a total number of 70 input nodes corresponding to all categories available for those 16 variables included in the analysis, the maximum number of hidden neurons tested would be: $N_h = \sqrt{N_i \cdot N_o} = \sqrt{70 \cdot 2} = \sqrt{140} \approx 12$

(Antkowiak, 2006), where, N_h is the number of hidden nodes, N_i is the number of input nodes, and N_o is the number of output nodes. In the hidden layer, logistic sigmoid activation function was employed. This is a continuous differentiable function, generating results within range [0,1]. For the output layer, the activation function considered was softmax: $f(x) = \frac{e^x}{\sum e^x}$, generally used when the output must take the form of probabilities. Regarding the error function, cross entropy was applied: $E = -\sum_{d=1}^{m} t^d \ln\left(\frac{y^d}{t^d}\right)$. This is a type of error function suitable for binary classification tasks, when the activation function in the output layer is softmax. The weights were initialized using a uniform distribution within [-0.5, 0.5] range. Afterwards, these were adjusted through the learning process until they reached one of the stopping criteria: a maximum number of 300 adjusting cycles, or a variation in the average error below 0.000001 for 20 consecutive cycles. For each number of hidden neurons tested ({2, 3… 12}), 20 different networks were generated in which only initial weights were changed, all other elements being kept constant. Thus, the total number of ANNs trained was eventually 220. The training method by which weights were adjusted in this classification task is the gradient descent with momentum. The values for the learning rate and the momentum rate were set to 0.1 each, as changes in their values did not result into significant modifications of final results. The training was performed in a "batch" mode, meaning that weights were adjusted only after presenting all training records to the network, and to reach a final classification model, several cycles were performed until the meeting a certain stopping criterion. Out of the 220 neural networks trained, the best in terms of detection rates for class "with bank deposit" on the test set was retained. The selected neural network model is *MLP 70-7-2*, containing 70 nodes in the input layer, seven nodes in the hidden layer, and two output nodes. The evolution of the training set error was compared with the evolution of the validation error in order to make sure that the final model is not over-fitted, a characteristic that reduces the generalization capacity of the model when tested on new data.

5. Results

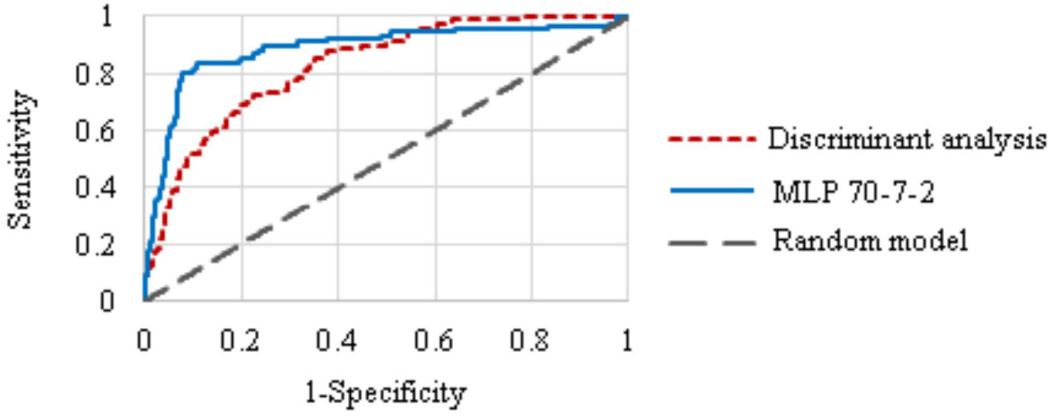
Considering that the main purpose of this application is identifying individuals that would save money through bank deposits, there is clearly an advantage of ANNs over discriminant analysis when analyzing results available in Table 2. Neural networks generate a higher detection rate for class "with bank deposits" on both data samples, development (training + validation) and test, giving proof that these techniques are very flexible even when dealing with traditional survey information.

Table 2. Detection rates

Subset	Technique	Detection rates		
		"without bank deposit"	"with bank deposit"	Total
training + validation	Discriminant analysis	75.69%	72.22%	75.44%
	MLP 70-7-2	78.26%	85.56%	78.79%
test	Discriminant analysis	75.45%	73.33%	75.30%
	MLP 70-7-2	73.64%	76.67%	73.86%

Fig. 2, providing the Receiver Operating Characteristic (ROC curve) for the training and validation data sets taken together, shows the performance of the models as the class cut-off varies. Sensitivity ratio gives the detection rate for class "with bank deposit", while specificity denotes the detection rate for class "without bank deposit". In the *low sensitivity - high specificity* area, ANNs generate significantly better results compared with discriminant analysis, while in the area of *high sensitivity - low specificity*, ANNs perform less good. Nevertheless, given the primary objective of this analysis, a high specificity is preferred, because this way the

model identifies as many records as possible within class "with bank deposit" before having type I errors, which represent those records wrongly classified within “with bank deposit” category. Hence, marketing costs and efforts towards those individuals who do not intend to save money through bank deposits are largely reduced.



Source: own calculation
Fig.2. ROC chart for training and validation datasets taken together

6. Conclusions and further research

This study has shown that when using survey data, classification results with ANNs are superior to those reached by classical discriminant analysis. Although the available dataset was rather limited, neural networks generated high detection rates on the target category and provided good results when tested on out-of-sample data, being, thus, a good choice for improving marketing strategies and decision making processes. The fact that ANNs are more time consuming in respect with the model configuration steps is counterbalanced by less prior data transformations and hypotheses testing required compared with discriminant analysis. However, care must be taken when training ANNs, as they can be exposed to over-fitting phenomenon. Future directions in analyzing the performance of ANNs in classification matters may consider using second derivative optimization algorithms when adjusting the network weights as these may provide superior results compared with the classical gradient descent back-propagation method.

References

Antkowiak, M., 2006. Artificial Neural Networks vs. Support Vector Machines for Skin Diseases Recognition, Master thesis in Computer Science, Umea University, Sweden.

Felea, I., Dan, F., Dzitac, S., 2012. Consumers Load Profile Classification Corelated to the Electric Energy Forecast. Proceedings of the Romanian Academy, Series A, 13(1), 80-88.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188.

Hakimpoor, H., Arshad, K.A.B., Huam, H.T., Khani, N., Rahmandoust, M. 2011. Artificial Neural Networks Applications in Management. World Applied Sciences Journal 14(7), IDOSI Publications.

Kim, J., Ahn, H., 2009. A New Perspective for Neural Networks: Application to a Marketing Management Problem. Journal of Information Science and Engineering 25, p. 1605-1616.

Knerr, S., Personnaz, L.., Dreyfus, G., 1992. Handwritten Digit Recognition by Neural Networks with Single-Layer Training. IEEE Transactions on Neural Networks 3(6), p. 962-968.

McCulloch, W.S., Pitts, W.H., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, p. 115-133.

Siddiqi, N., 2006. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. SAS Institute, p. 79-83.

Štencl, M., Popelka, O., Šťastný, J., 2012. Forecast of Consumer Behaviour Based on Neural Networks Models Comparison. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 60, p. 437-442.

Zhang, T., Yuan, B. Liu, W.H., 2008. Predicting Credit Card Customer Loyalty Using Artificial Neural Networks. In Proceedings of the 11th Joint Conference on Information Sciences, AISR 7, Atlantis Press, Shenzhen, P.R. China.