Journal of Life Economics

Cilt / *Volume* **7,** Sayı / *Issue* **4, 2020**, *pp.* **297-308**
E - ISSN: **2148-4139**
URL: https://www.journals.gen.tr/jlecon
DOİ: https://doi.org/10.15637/jlecon.7.022
Araştırma Makalesi/*Research Article*

ESTIMATION OF THE SECTORS OF THE INVESTMENTS MADE ON VENTURE CAPITAL COMPANIES WITH ARTIFICIAL NEURAL NETWORKS AND MULTIPLE LOGISTIC REGRESSION ANALYSIS

Kubilay ERİSLİK* & Özlem DENİZ BAŞAR**

** Arş. Gör., Istanbul Commerce University*
TURKEY, e-mail: kubilayerislik@ticaret.edu.tr
ORCID ID: https://orcid.org/0000-0002-0744-4435

*** Prof. Dr. Istanbul Commerce University,*
TÜRKİYE, e-mail: odeniz@ticaret.edu.tr
ORCID ID: https://orcid.org/0000-0002-9430-8975

Received: 15 October 2020; Accepted: 31 October 2020

ABSTRACT

Venture capital companies undergo three different phases as core, growth and maturity phases as of their establishment. There are different stages in these phases in terms of providing the finance. The stage of providing finance for the first introduction of the product to the market in the core phase is called Serial A, the stage of providing the increasing finance need during the continuation of the growth is called Serial B and the stage of providing the finance needed in the growth and maturity phases is called Serial C and it continues as Serial D. In this study, it has been aimed to estimate the sectors of the venture capital companies by benefiting from the phases and amounts of the investments made by the investors to the venture capital companies. In the study, 5 sectors with the highest investment from investors have been selected and the investment data of 709 venture capital companies taking place in this sector have been benefited. Artificial Neural Networks and Multiple Logistic Regression Analysis have been used in the estimation of the sectors covering the companies with the data attained from the investment series. When the attained results have been examined, it has been determined that the results attained with Artificial Neural Networks are more successful than the results attained with Multiple Logistic Regression analysis.

Keywords: *Artificial Neural Network, Multiple Logistic Regression, entrepreneur, classification*

Jel Codes: *C38, C39, C45, G24, L26*

1. INTRODUCTION

Venture in daily life expresses the status of acting, starting and attempting to do a work, entrepreneur is used in the meaning of the entrepreneur person taking place in such a situation. These concepts are actually handled in an economic frame. In this frame, entrepreneur is seen as the person directing the supply and demand and seeking for market and the entrepreneurship is rather seen as the activity of economically mobilizing and prompting the resources (Aytaç and İlhan, 2007).

Venture capital companies have three developmental phases as of their establishment. These are called as core, growth and maturity phases. According to the companies, the dimension and speed of the growth in these phases show differences. The number of companies that could successfully complete all of the three phases is too few (Harvard Business Review, 2019). When the venture capital companies are examined, the companies having investment support at these phases have been seen to have had a higher survival ratio than those not having any investment. The investment process contains a series of activities starting with the suggestion of the new venture and continuing until sufficient income is successfully attained.

The first one of the finance collection phases is called the core phase. The stage in which the finance necessary for the acceleration of the work of introducing the first product to the market is met is called Serial A. The need for finance increases more while the company continues to grow and this phase is called Serial B. Meeting the finance needed in the growth and maturity phases of the company continues as Serial C and Serial D. Generally, angel investors undertake the investments needed by the venture capital companies during the core and Serial A phases. The financial needs at the phases of Serial B, Serial C and Serial D are met by the risk capital companies (Harvard Business Review, 2019).

In this study, it has been aimed to estimate the sectors of the venture capital companies by benefiting from the phases and amounts of the investments made by investors to the venture capital companies. Investors are affected from the economic, social or personal reasons while reaching an investment decision to the venture capital companies. The economic and social reasons could be expressed as the vision, aim and sector of the company. In the study, 5 sectors having the highest investments from the investors have been selected and the investment data of 709 venture capital companies in this sector have been benefited. The sectors used in the analysis are Cloud Services, Big Data and Machine Learning, E-Commerce, Mobile Applications and Social Media sectors. The investment amounts of the ventures belonging to the sectors specified in the conducted analysis in 6 investment series have been used. These investment series are respectively Core, Serial A, Serial B, Serial C, Serial D and the investments independent of the series.

Artificial Neural Networks and Multiple Logistic Regression Analysis have been used in the estimation of the sectors containing the companies with the data attained from the investment series. As a result of the study, the estimations obtained from Artificial Neural Networks and Multiple Logistic Regression Analysis have been compared.

2. MATERIAL AND METHOD

The data regarding the data and the used methods are specified below in this study in which the comparison of the findings attained with artificial neural networks and multiple regression analysis has been aimed.

2.1 Data

The data used in the study have been taken from Crunchbase website in which the data belonging to the investments taken by the venture capital companies take place. 29538 companies taking place in the cloud services, big data and machine learning, e-commerce, mobile applications and social media sectors having investment between 2010 – 2015 form the universe of the study. 709 ea. venture capital companies have been included in the study using simple random sampling from the universe. 141 out of 709 randomly selected companies take place in cloud services sector, 84 of them take place in big data and machine learning sector, 154 of them take place in e-commerce sector, 245 of them take place in mobile application sector and 85 of them take place in social media sector. The amounts of the investments made to the companies during the investment process have been benefited in the sector estimation. The variables used in the analysis are given in Table 1.

Table 1. Variables Used in the Analysis

Code	Dependent Variable
Y	Sector of the Venture Capital Company
	Independent Variables
X_1	Investment Amounts Made at the Core Stage
X_2	Investment Amounts Made in Series A
X_3	Investment Amounts Made in Series B
X_4	Investment Amounts Made in Series C
X_5	Investment Amounts Made in Series D
X_6	Investment Amounts Independently of the Series

Sectors and their codes used in the analysis are given in Table 2 for the artificial neural networks and logistic regression analysis to be able to be applied to the data.

Table 2. Sector Codes

Codes	Sectors
1	Cloud Services
2	Big Data and Machine Learning
3	E-Commerce
4	Mobile Applications
5	Social Media

Artificial neural networks and logistic regression analysis are the methods statistically similar to each other. Both classification methods conduct classification using the pattern recognition model. The obtainment of correct results from artificial neural networks and logistic regression analysis depends on three factors: the quality of the data set, model parameters and the criteria of the modeling results.

2.2 Classification Methods

In this part, Artificial Neural Networks and Multiple Logistic Regression Analysis used in the study have been explained in detail.

2.2.1 Artificial Neural Networks

Machine learning is used in many fields such as the web searches in modern society, content filtering in social networks and the products suggested to you in e-commerce websites. Machine learning is the process of estimating new data by forming a model or algorithm from the current data. Artificial Neural Networks (ANN) is one of the methods of machine learning used in the education part. ANN are the computer systems developed for the purpose of automatically revealing the skills of human brain such as being able to produce new information via the way of learning being from the characteristics of human brain without any help (Öztemel, 2012).

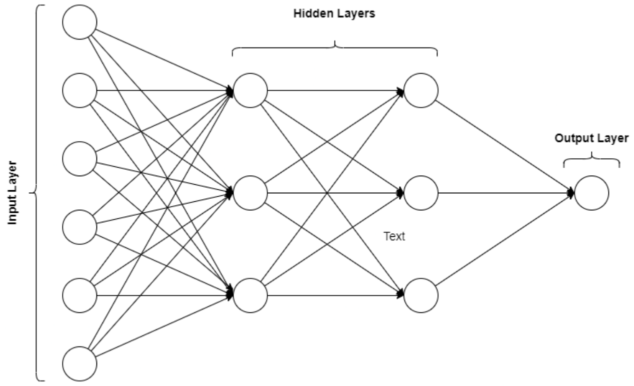
ANN is more valuable than other methods used in the estimation processes thanks to some of its distinguishing properties. The first one of these properties is that it is a self-excited method based on data, not based on assumptions in contrast to the traditional model based methods. ANN learn from the examples and catch the fine functional relations among the data although it is hard to define. In this meaning, ANN could be considered as one of the multivariate non-linear statistical methods (Hornik, Stinchcombe and White, 1989; Cheng and Titterington, 1994).

The second property of the artificial neural networks is that they could correctly complete the deficient observations in the data. In this way, they also enable the conduction of future estimation by benefiting from the past observations. The third property of the artificial neural networks is that they have more general and flexible functional patterns than the traditional statistical methods. A basic relation is assumed between the inputs and outputs in the traditional statistical estimation models. These assumptions mostly cannot be provided in the complexity of the problems in real life. Therefore, artificial neural networks could be a very good alternative for the estimation processes.

Artificial neural networks are not linear. Generally linear statistical methods have been used in the estimation processes. It is easy to understand, analyze and interpret the linear methods. However; the problems in real life are mostly not linear. There are many non-linear estimation models. However; it is very hard to apply a non-linear model to a certain data set; because, there are many possible non-linear models and the non-linear model that has not been previously determined may not be successful enough to catch the important properties in the data. Artificial neural networks which are data-based could conduct non-linear modeling without having any background knowledge thanks to the relations between the input and output variables.

Artificial neural networks consist of the bonding of neural cells to one another in various ways. Many artificial neural network structures have been developed according to the bonding ways, learning rules and transfer functions of the cells (Arıkan Kargı, 2015). Generally multilayer artificial neural networks are used in the classification and estimation analyses. There are input layer, hidden layers and output layer in the artificial neural network model in multilayer artificial neural networks. The data prepare from the multilayer sensor network are served to the network from the input layer, pass from the hidden layers, reach the output layer and the response of the network is transmitted to the outer world in return for the inputs served to the network (Öztemel, 2012). Weights are constantly changed for the purpose of minimizing the error between the real value and the output value produced by the network at the learning stage in multilayer networks. Learning process is realized at the moment when the error is minimized (Velo, López and Maseda, 2014). The sensor network with one output layer, two hidden layers and one output layer is shown in Figure 1.

Figure 1: Two-Hidden Layer Sensor Network with Single Output Layer



The most important property of the artificial neural networks is that the layers are not designed by human beings. The data entering the system from the input layer are examined, the important aspects among the data are reinforced and the irrelevant aspects are suppressed. The data in strong relation to one another are assigned to the same class, the classification process is conducted and in this way, the process of learning is realized with the help of the data.

In the classification process in artificial neural networks, any input is demanded to determine which one of k separate categories it is included in. The learning algorithm should determine a $f : \mathbb{R}^n \to \{1, \cdots, k\}$ function to be able to solve this task. Such that the algorithm will assign an input shown with x to the category shown as y in a way that $y = f(x)$ (Goodfellow, Bengio and Courville, 2016).

The artificial neural networks applied by using computers will be a method that will be used very much in the future due to the easiness in the calculation process. In this way, the increase in the data amount does not harden the applicability of the analysis. For this reason, ANN are preferred instead of many classification methods.

2.2.2 Multiple Logistic Regression Analysis

Logistic regression analysis is a technique used for measuring the cause and effect relation between the categorical dependent variable and the constant or categorical independent variables and for classifying the decision units (Burns and Burns, 2008). As per its structure, logistic regression analysis is similar to the simple linear regression, but its difference from the simple linear regression is that the dependent variable is discontinuous. Logistic regression analysis is separated into three depending on the structure of the dependent variable. If the dependent variable consists of two categories, it is called as binary; if it consists of more than two categories, it is called as multiple logistic regression and in the event that there is a superiority or order among the categories, it is called as ordinal logistic regression.

In a multiple logistic regression analysis with dependent variable in M categories; it is necessary to calculate M-1 ea. equations for the purpose of defining the relation between the dependent and independent variables according to the reference category for each dependent variable category. If the dependent variable is considered to have consisted of 3 categories (0, 1, 2), it is necessary to calculate 2 equations. One of the three categories is selected as the reference category and logit function is calculated for the other categories. Afterwards, the calculated logit functions are compared. Logit functions are shown as follows.

$$g_1(x) = \ln\left[\frac{\Pr(Y=1|x)}{\Pr(Y=0|x)}\right]$$

$$= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p$$

$$= x^{'}\beta_1 \qquad \textbf{(3.1)}$$

and

$$g_2(x) = \ln\left[\frac{\Pr(Y=2|x)}{\Pr(Y=0|x)}\right]$$

$$= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p$$

$$= x^{'}\beta_2 \qquad \textbf{(3.2)}$$

The conditional probabilities of the common variable vector in each category are expressed as follows (Hosmer, Lemeshow and Sturdivant, 2013).

$$\Pr(Y=0|x) = \frac{1}{1+e^{g_1^{(x)}}+e^{g_2^{(x)}}} \qquad \textbf{(3.3)}$$

$$\Pr(Y=1|x) = \frac{e^{g_1^{(x)}}}{1+e^{g_1^{(x)}}+e^{g_2^{(x)}}} \qquad \textbf{(3.4)}$$

$$\Pr(Y=2|x) = \frac{e^{g_2^{(x)}}}{1+e^{g_1^{(x)}}+e^{g_2^{(x)}}} \qquad \textbf{(3.5)}$$

3. APPLICATION

The results attained with the artificial neural networks and multiple logistic regression analyses in the application are shown in detail in this part.

a) Findings Attained with Artificial Neural Networks Model

Data set is separated into three groups while establishing artificial neural networks models. These groups are called as the training group, test group and approval group. Because the number of companies used in the study is not so high, data set has been separated into 2 groups as training set and test set and the approval set has not been used. Two different network models have been used in the study. They have been arranged in a way that 70% of the data set will be training and 30% will be test data in the first one of them; and in a way that 80% will be training and 20% will be test data in the second one of them. The formed 70%-30% data set and 80%-20% data set have been compared to each other and the results have been interpreted.

RStudio 1.2.5001 software and R 3.6.1 version have been used while conducting artificial neural networks and multiple logistic regression analysis. The determination of hidden layer and neuron number in accordance with the model does not have any certain method while forming artificial neural networks model. Various models have been formed to determine the convenient hidden layer number and neuron number and the mean square errors has been calculated. The formed models and the mean square errors are given in Table 3. In this respect;

the best result has been determined as the model with 2 hidden layer and 3 neurons each in 70%-30% data set.

Table 3: Mean Square Errors of the Test Set According to the Models

	Mean Square Error	
Models	**%70-%30**	**%80-%20**
5-5-5	0.5604822	0.5502131
3-3-3	0.8741106	0.5480075
5-5	0.2305093	0.8699651
3-3	**0.2007436**	0.3677766
5-3	0.2157579	0.8251188
3-5	0.5612207	0.6603289

The properties belonging to the artificial neural network with the best formed result are given in Table 4.

Table 4: ANN Structure with the Best Result

Network Model	Multilayer
Learning Algorithm	Backprop
Neuron Number in Input Layer	6
Hidden Layer Number	2
Neuron Number in 1st Hidden Layer	3
Neuron Number in 2nd Hidden Layer	3
Neuron Number in Output Layer	1
Activation Function	Sigmoid Function

The activation function of the artificial neural network with the best result has been determined as sigmoid function. The results of the training set performances belonging to the model are given in Table 5.

Table 5: Training Set Performance Results

		Real Group				
		Cloud Services	**Big Data and Machine Learning**	**E-Commerce**	**Mobile Applications**	**Social Media**
Estimated Group	**Cloud Services**	90	7	0	0	0
	Big Data and Machine Learning	0	61	0	0	0
	E-Commerce	0	9	99	0	0
	Mobile Applications	0	10	0	166	0
	Social Media	0	1	0	0	53
Correct Classification Ratio		$= (90 + 61 + 99 + 166 + 53)/496 = 0{,}9455$				

When Table 5 is examined, it has been determined that the total correct classification ratio has been determined as 94,55% for the training set. The artificial neural network formed with the training set has been compared to the test set and the results are given in Table 6.

Table 6: Test Set Performance Results

		Real Group				
		Cloud Services	**Big Data and Machine Learning**	**E-Commerce**	**Mobile Applications**	**Social Media**
Estimated Group	**Cloud Services**	41	3	0	0	0
	Big Data and Machine Learning	0	23	0	0	0
	E-Commerce	0	5	40	0	1
	Mobile Applications	0	6	0	63	0
	Social Media	0	0	0	1	25
Correct Classification Ratio		$= (41 + 23 + 40 + 63 + 25)/213 = 0{,}9014$				

The formed ANN test has been controlled with the data set. 41 out of 44 companies in cloud services sector have been correctly classified by ANN and 3 of them have been classified wrongly. All of 23 companies in the sector of big data and machine learning have been classified correctly. 40 out of 46 companies in the sector of e-commerce, 63 out of 69 companies in the sector of mobile applications and 25 out of 26 companies in the sector of social media have been classified correctly. According to the test data set; ANN correct classification ratio has been found as 90%.

b) Findings Attained with Multiple Logistic Regression Analysis

Multiple logistic regression analysis has been applied to the data and regression model has been formed. One of the existent categories should be selected as reference category in the multiple logistic regression analysis. In the conducted study, category no.1 has been selected as the reference category and the model has been formed according to this category. The data belonging to the formed regression model are given in Table 7 and Table 8.

Table 7: Coefficients of Multiple Logistic Regression Analysis

Category	Constant	Independent from Series	Core	Series A	Series B	Series C	Series D
2	-0,172	1.446e-8	-7.245e-7	-1.777e-8	-3.841e-9	-2.565e-8	1.815e-7
3	0,026	1.752e-8	-3.123e-7	7.141e-9	-5.679e-8	4.413e-9	1.639e-7
4	-0,162	1.752e-8	-1.434e-8	9.454e-9	-9.644e-9	-1.454e-8	1.593e-7
5	0,180	-6.908e-9	-4.137e-7	1.971e-7	5.742e-8	-5.665e-8	1.859e-7

When the coefficients in Table 7 are examined, the investments made independent of the series mostly affect the sectors of E-Commerce and Mobile Applications. The investments made at the core phase affect the sector of Big Data and Machine Learning.

Table 8: Standard Errors of Multiple Logistic Regression Analysis

Category	Constant	Independent from Series	Core	Series A	Series B	Series C	Series D
2	6.412e-14	8.091e-9	1.716e-7	3.176e-8	1.526e-8	1.068e-8	3.821e-8
3	2.641e-14	8.042e-9	1.113e-7	1.415e-8	1.452e-8	5.141e-9	3.785e-8
4	3.282e-14	8.042e-9	8.247e-8	1.508e-8	9.848e-9	6.019e-9	3.779e-8
5	4.198e-14	1.240e-8	1.734e-7	4.721e-8	1.369e-8	1.591e-8	3.872e-8

The investments made at the phases of Serial A, Serial B, Serial C and Serial D mostly affect the sector of Social Media. The estimations made by benefiting from the formed multiple logistic regression model are given in Table 9.

Table 9: Performance Results of Multiple Logistic Regression Analysis Model

		Real Group				
		Cloud Services	Big Data and Machine Learning	E-Commerce	Mobile Applications	Social Media
Estimated Group	Cloud Services	44	0	1	8	2
	Big Data and Machine Learning	1	6	5	1	1
	E-Commerce	5	7	25	9	5
	Mobile Applications	87	67	119	223	51
	Social Media	4	4	4	4	26
Correct Classification Ratio		$= (44 + 6 + 25 + 223 + 26)/709 = 0{,}4570$				

According to the formed regression model, 44 out of 141 companies in cloud services sector have been correctly classified by the regression analysis and 97 of them have been classified wrongly. Only 6 out of 84 companies in the sector of big data and machine learning have been classified correctly. 25 out of 154 companies in the sector of e-commerce, 223 out of 245 companies in the sector of mobile applications and 26 out of 85 companies in the sector of social media have been classified correctly. The correct classification of the companies handled as a result of the established multiple logistic regression model is 46%. This result is too low when compared to the result of the artificial neural networks.

4. RESULT

Rapidly developing business life has caused to the occurrence of different concepts and rapid application of these concepts. The concept of venture whose appearance is not so back in the past is one of the best examples that could be shown for this situation. Venture in daily life expresses the status of acting, starting and attempting to do a work, entrepreneur is used in the meaning of the entrepreneur person taking place in such a situation. Although the number of those completing all of them successfully is very few, the venture capital companies conduct their lives within three developmental phases: core phase, growth phase and maturity phase. Companies have finance collection stages covering different periods in the process of these phases. The first stage in which finance is met is called as Serial A, the second stage in which the finance need is met while continuing to develop is called as Serial B and the processes following in this way are called as Serial C and Serial D.

In this study, it has been aimed to estimate the sectors of the venture capital companies by benefiting from the phases and amounts of the investments made by the investors to the venture capital companies. For this purpose; the sectors of Cloud Services, Big Data and Machine Learning, E-Commerce, Mobile Applications and Social Media having the highest amounts of investments from investors have been selected and the investment data of 709 venture capital companies in these sectors covering 2010-2019 have been benefited. These investment data are respectively Core, Serial A, Serial B, Serial C, Serial D and the investments independent of the Series.

In this study conducted upon the assumption that the investment amounts made by the investors change on sectoral basis, ANN and Multiple Logistic Regression analysis methods have been used in the estimation of the sectors. While the correct assignment ratio of 90% has been provided with ANN, a success by the ratio of approximately 47% has been provided with the help of Multiple Logistic Regression Analysis. ANN method has been detected to have given more successful results than Multiple Logistic Regression analysis in this study in which sectors have been classified in terms of the financial investment amounts.

REFERENCES

ARIKAN KARGI, V. S. (2015) *Yapay Sinir Ağ Modelleri ve Bir Tekstil Firmasında Uygulama*. Bursa: Ekin Yayınevi.

AYTAÇ, Ö. and İLHAN, S. (2007) 'Girişimcilik Ve Girişimci Kültür: Sosyolojik Bir Perspektif', *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (18).

BURNS, R. B. and BURNS, R. A. (2008) *Business research methods and statistics using SPSS*. Los Angeles ; London: SAGE.

CHENG, B. and TITTERINGTON, D. M. (1994) 'Neural Networks: A Review from a Statistical Perspective', *Statistical Science*, 9(1), pp. 2–30. doi: 10.1214/ss/1177010638.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016) *Deep learning*. Cambridge, Massachusetts London, England: The MIT Press (Adaptive computation and machine learning).

Harvard Business Review (2019) *Girişimcinin Elkitabı*. 1st edn. Translated by L. Göktem.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989) 'Multilayer feedforward networks are universal approximators', *Neural Networks*, 2(5), pp. 359–366. doi: 10.1016/0893-6080(89)90020-8.

HOSMER, D. W., LEMESHOW, S. and STURDIVANT, R. X. (2013) *Applied Logistic Regression*. Third edition. Hoboken, New Jersey: Wiley (Wiley series in probability and statistics, 398).

ÖZTEMEL, E. (2012) *Yapay Sinir Ağları*. İstanbul: Papatya Yayıncılık Eğitim.

VELO, R., LÓPEZ, P. and MASEDA, F. (2014) 'Wind speed estimation using multilayer perceptron', *Energy Conversion and Management*, 81, pp. 1–9. doi: 10.1016/j.enconman.2014.02.017.