Croatian Operational Research Review
CRORR **7**(2016), 229–248

Predicting company growth using logistic regression and neural networks

Marijana Zekić-Sušac[1,†], **Nataša Šarlija**[1], **Adela Has**[1] **and Ana Bilandžić**[1]

[1]*Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000 Osijek, Croatia*
E-mail: ⟨{*marijuana, natasa, adela, anag*}*@efos.hr*⟩

Abstract. The paper aims to establish an efficient model for predicting company growth by leveraging the strengths of logistic regression and neural networks. A real dataset of Croatian companies was used which described the relevant industry sector, financial ratios, income, and assets in the input space, with a dependent binomial variable indicating whether a company had high-growth if it had annualized growth in assets by more than 20% a year over a three-year period. Due to a large number of input variables, factor analysis was performed in the pre-processing stage in order to extract the most important input components. Building an efficient model with a high classification rate and explanatory ability required application of two data mining methods: logistic regression as a parametric and neural networks as a non-parametric method. The methods were tested on the models with and without variable reduction. The classification accuracy of the models was compared using statistical tests and ROC curves. The results showed that neural networks produce a significantly higher classification accuracy in the model when incorporating all available variables. The paper further discusses the advantages and disadvantages of both approaches, i.e. logistic regression and neural networks in modelling company growth. The suggested model is potentially of benefit to investors and economic policy makers as it provides support for recognizing companies with growth potential, especially during times of economic downturn.

Keywords: company growth, factor analysis, logistic regression, neural networks, prediction model

Received: October 28, 2016; accepted: December 7, 2016; available online: December 30, 2016

DOI: 10.17535/crorr.2016.0016

1. Introduction

Besides innovation and venture creation, predicting company growth is one of the most important aspects of entrepreneurship research [7]. Most of the research in

† Corresponding author

high-growth enterprises is directed towards the potential for such growth in large companies [6]. This paper endeavors to fill this void by modelling the growth of Croatian SMEs based on data from financial statements. The first aim of the paper is to produce an efficient model that classifies companies into high-growth and non-high-growth categories using two data mining methods: logistic regression (LR) and artificial neural networks (ANNs). Going further, the second aim is to identify the determinants of SME growth in the most important aspects of the pre-modelling and post-modelling procedures. This particular methodology includes factor analysis in the pre-modelling stage, as well as LR and ANNs to establish the most accurate model using testing datasets incorporating the entire available and reduced input space. Sensitivity analysis was conducted in the post-modelling stage to determine the effects that each input variable has on company growth. The paper presents an overview of previous conducted research and subsequently a description of the methodology. The results of factor analysis, the LR and ANN models are described and discussed further on in the paper.

2. Review of previous research

In economic theory and practice, growth can be measured in many ways, such as in terms of revenue generation (sales), employment and asset growth, but also in terms of product quality and market position [30]. In this paper, the asset growth is a measure of company growth and is used to explore the various financial determinants important in predicting asset growth. Factors influencing the growth potential of an enterprise are usually viewed in through three main categories: the entrepreneur, the company itself and company strategy [29]. Education, risk acceptance, aspiration to grow, mid-management experience have been shown as important growth factors at the entrepreneur level [3]; [17]. At the company and strategy level, age and size of a company, strategic orientation, level of R&D, innovation are found to be relevant growth determinants [1], [10], [11], and [22]. Mateev and Anastasov [21] have shown that financial structure and productivity exert a positively influence on the potential for company growth. They investigated transition countries and found the following company-specific factors to be important: indebtedness, internal financing, future growth opportunities, process and product innovation, and organizational changes. Regarding financial growth determinants, findings from previous research vary across countries. Helmers and Rogers [13] have identified a company's capacity to invest, particularly in R&D as an important factor, while Becchetti and Trovato [2] have shown that the availability of external finance and internationalization are positively related to company growth. Sampagnaro [26] has shown that internal cash flows are relevant in predicting the company growth and success. Moreover, there is an certain tendency that external financing affects growth negatively.

The methodology used in previous research on modelling company growth relied on standard statistical methods such as multiple regression, logistic regression and discriminant analysis. Delmar et al. [8] use correlations and regression analysis to model company growth, whereas Geroski [12] uses static and dynamic optimizing models for company output choice, modelled production functions for corporate learning, modelled R&D competition and diversification, and examined their influence on corporate growth rates. Ma and Tang [19] use a stepwise logistic regression model to predict a default considering the misclassification loss. Logistic regression was also compared with and integrated into certain machine learning methods to predict company distress or bankruptcy. Hua et al. [14] combine LR and support vector machines to predict company financial distress by developing an integrated binary discriminant rule (IBDR) that decreases the empirical risk of support vector machine outputs and interprets and modifies the outputs according to the result of LR analysis. Other machine learning methods, such as ANNs, have not as yet been investigated adequately in modeling company growth.

3. Methodology

3.1. Research design

Research aimed at finding the most successful model for predicting company growth on the Croatian dataset was conducted in three main stages. Given that company financial statements contain numerous financial ratios, the **first stage** of the research involved a variable reduction procedure to extract important financial features of companies from the dataset. Hence, factor analysis was performed for each particular year in the 2008-2010 period with a different factor of numbers and combinations. The factor scores were obtained using the package **psych** in R and the function **fa()**. While the factoring method exhibited a maximum likelihood, regression was used for obtaining factor scores. Not all variables showed correlation with at least one factor, hence they were excluded from factor analysis one by one until the desired result was obtained. Along with the factor loadings from function **fa ()** and some additional variables that covered profitability, **glm ()** was used to develop the final logistic regression model.
The **second stage** of research included modeling company growth using LR and ANNs. The LR was selected as one of the most exploited statistical methods in modeling company growth, financial distress, success, and similar outcomes. Due to the high popularity of machine learning methods in recent years, the challenging was to compare the performance of LR with ANNs as a machine learning method that has not (to our knowledge) been used previously in predicting company growth. Here, the two methods were used in a competitive way with the aim of deriving the most accurate model. Moreover, this paper also discusses and suggests integrating the two methods. A comparison of the accuracy

of LR and ANN accuracy was based on classification rates, ROC and Kolmogorov-Smirnov (KS) test obtained using both methods on the same validation dataset. Comparing the results of the LR and ANN models was based on several statistical tests: the t-test of differences in proportion, and McNemar's non-parametric test. The non-parametric test was used based on the suggestion of Luengo et al. [18] who note that a performance comparison of ANNs and other classification methods should be conducted using a non-parametric statistical test given that the validation of methods depends on the sampling procedure.
In the **third, post-modeling stage**, variable importance was investigated using sensitivity analysis with the aim of examining the effect of each company growth predictor, and its relevance for the obtained prediction model. This kind of analysis can guide the decision maker by highlighting the most influential predictors.

3.2. Data

The dataset was collected from the Croatian Financial Agency (FINA) and initially consisted of 53,434 SMEs that operated during the 2008-2013 period. A random sample of 1492 privately-owned SMEs was selected and used in further experiments. The measure of growth suggested by the OECD methodology (2010) defines a high-growth enterprise if experiencing average annualized asset growth exceeding 20% over a three-year period, specifically in the period from 2010-2013. When considering the total number of SMEs, 746 such enterprises fulfilled this criterion. The development sample included 650 high-growth SMEs, with the validation sample consisting of 96 high-growth SMEs. The other 746 non-high-growth SMEs were selected randomly from the entire data set, and categorized in the same way as high-growth enterprises.
Company financial ratios were used as independent variables which were computed for every enterprise in the sample for 2008, 2009 and 2010, and in addition, the percentage change of each ratio for the period 2008-2009 and 2009-2010 was calculated. In total, 111 variables were created. The best models were developed using financial ratios calculated for the year 2010. Descriptive statistics and a description of the variables for the year 2010 used in the research are given in Table 1.

Variable code	Description of variable	High-growth	Non-high-growth
		Median (interquartile)	Median (interquartile)
Liquidity ratios			
Lclnw*	current liabilities/equity	0.35 (3.06)	0.49 (1.91)

Lubrl	(current assets-inventory)/current liabilities		0.92 (1.82)	0.88 (1.85)
Lkiui*	current assets/total assets		0.83 (0.42)	0.76 (0.58)
Ctrenl*	cash/current liabilities		0.12 (0.57)	0.09 (0.47)
Casal*	current assets/sales		0.37 (0.51)	0.56 (0.78)
Kimovkobv*	current assets/current liabilities		1.201 (0)	1.121 (1.917)
Turnover ratios				
Aukupni*	total revenue/total assets		1.79 (2.58)	0.99 (1.61)
Adug*	total revenue /fixed assets		9.61 (26.26)	3.81 (14.31)
Akrat*	total revenue /current assets		2.39 (3.9)	1.71 (2.41)
Asalta*	sales/total assets		1.66 (2.52)	0.84 (1.56)
Asalwc	sales/net working capital		0 (6.07)	0.52 (4.52)
Aqasal*	(current assets-inventory)/sales		0.31 (0.42)	0.4 (0.57)
Anap1*	365/receivables turnover		36.55 (86.85)	54.48 (117.09)
Akrdob*	365/payables turnover		43.08 (120.8)	69.97 (147.53)
Azal1*	sales/inventory		9.66 (31.69)	5.39 (17.16)
Leverage ratios				
Zkz*	total debt/total assets		0.71 (0.59)	0.79 (0.87)
Zdk*	total debt/equity		0.44 (3.7)	0.77 (2.76)
Blta*	bank loan/total assets		0 (0)	0 (0.06)
Zlongdca*	long-term debt/current assets		0 (0.04)	0 (0.26)
Profitability ratios				
Pros	net income/sales		0.02 (0.08)	0.21 (0.07)
Pnpm	net income/total revenue (%)		1.17 (23.6)	1.16 (17.15)
Proa	net income/total assets (%)		1.9 (31.91)	0.84 (9.45)
Pnroe *	net income/equity (%)		23.64 (56.17)	8.38 (33.78)
Prearnta*	retained earnings/total assets		0 (0.78)	0.04 (0.35)
Other variables				
Nematimovina	intangible assets/total assets		0.02 (0)	0 (0)
Tech_djel[#]	industry sector	Agriculture	2.32	97.68
		Manufacturing	1.4	98.6
		Construction	1.33	98.67

		Trade	1.21	98.79
		Transportation and storage	0.9	99.1
		Accommodation and food service	1.32	98.68
		Information and communication	2.12	97.88
		Financial activities	1.46	98.54
		Professional and scientific services	1.4	98.6
		Social, education and other services	1.71	98.29

* Statistically significant difference based on the Mann-Whitney test
# Percentage of high growth and non-high-growth enterprises in each industry sector

Table 1: *Descriptive statistics of the input variables*

A total sample of 1,492 companies was divided into two subsamples, i.e. the train and the validation sample along with the same distribution of companies according to their growth category. The train sample consisted of 1300 (87.13%) cases and the validation sample had 192 (12.87%) cases. Logistic regression was performed on both. For ANN modelling, the train sample was divided once more into a smaller train sample and test sample. The new train sample consisted of 1040 (70%) cases, and the test set had 260 (17.43%) cases which were used in a cross-validation procedure to optimize the ANN parameters such as training time and the number of hidden units, whereas the validation set was used for final testing and comparing the accuracy of the model. The sample structure is shown in Table 2.

Subsample for LR	Subsample for ANN	Output category		Total no. of cases	Total (%)
		0 Non-high-growth	1 High-growth		
Train	ANN Train	520	520	1040	69.70
	ANN Test	130	130	260	17.43
Validation	Validation	96	96	192	12.87
Total	Total	746	746	1492	100.00

Table 2: *Sampling procedure*

3.3. Factor analysis

A frequent problem is obtaining robust regression coefficients when building the model. This is mitigated by reducing a number of variables, which is achieved using factor analysis. Reducing the number of variables decreases the number of regression coefficients requiring estimation, as well as the group of correlated variables so as to address the problem of multicollinearity. The downside of this is the loss of some information and subsequently the model loses some of its predictability.

Factor analysis is a procedure for reducing the original p observed variables down to r unobserved variables, where every variable X_i could be described by the r new variables, or:

$$\begin{aligned} X_1 - \mu_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1r}F_r + U_1 \\ X_2 - \mu_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2r}F_r + U_2 \\ &\vdots \\ X_p - \mu_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pr}F_r + U_p \end{aligned} \quad (1)$$

F_j, $j = 1,2,\dots,r$ are new variables called common factors, and U_i , $i = 1,2,\dots,p$; are called unique factors, all of them are unobserved. Equivalently, the set of equations can be written as

$$(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{u} \quad (2)$$

where A is the factor pattern matrix consisting of its elements a_{ij} which are called factor loadings, $\boldsymbol{x}$ is the $p \times 1$ vector of elements X_i, $i = 1,2,\dots,p$ and $\boldsymbol{\mu}$ is vector of their means. While $\boldsymbol{f}$ is the $r \times 1$ vector of elements F_j, $j = 1,2,\dots,r$, they are assumed to have a mean of 0 and variance of 1, while U_i , $i = 1,2,\dots,p$ are ssumed to have mean of 0, but a variance of σ_i^2, $i = 1,2,\dots,p$, as they form the $p \times 1$ vector $\boldsymbol{u}$. Addionally, it is assumed that the unique and common factors are

uncorrelated. Hence, by marking the covariance matrix of $\boldsymbol{x}$ with $\boldsymbol{\Sigma}$, the previous equation becomes:

$$\Sigma = \mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'] = \mathbf{A}\mathbf{A}' + \boldsymbol{\Psi}, \tag{3}$$

where Ψ is the vector of variances of U_i. Since the right side of the equation consists only of unobserved data, this process is not unique, and different factors can be obtained [13].

3.4. Logistic regression

Logistic regression is used to develop a regression model when the dependent variable is categorical. It was developed in 1958 by David Cox [4]. There are three types of logistic regression: (1) binary, for a binary response variable, (2) multinomial - where the dependent variable has more than two non-ordered categories, and (3) ordinal - when the categories are ordered. Logistic regression is less demanding with respect to the relationship between the dependent and independent variables and need not be linear, the distribution of the variables need not be normal and variance assumptions need not be homoscedastic [15]. Although regression coefficients in logistic regression are not easy to interpret and understand as in linear regression, the advantage is being able to interpret whether the relationship is proportional or inversely proportional between each independent and dependent variable.

In our research, Y is an dependent variable, where 1 indicates a high-growth company and 0 indicates non-high-growth company. For r independent variables $x_1, x_2, \dots x_r$, the logistic function is:

$$p = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_rx_r}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_rx_r}} \tag{4}$$

where p is the probability that a company will experience high growth.

The goal is to obtain β_i, $i = 1,2,\dots.,r$. By denoting $g(x) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_rx_r$ and applying logistic transformation, we obtained a linear relationship between the log odds and independent variables:

$$\begin{aligned} logit(y) = \ln\frac{p}{1-p} &= \ln\frac{\dfrac{e^{g(x)}}{1+e^{g(x)}}}{1-\dfrac{e^{g(x)}}{1+e^{g(x)}}} \\ &= g(x) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_rx_r. \end{aligned} \tag{5}$$

For a sample of size n, for $i = 1, \dots n$, we denote y_i to be the observed variables if an enterprise is high growth or not, and $x_i{}' = (1, x_{i,1}, \dots, x_{i,r})$ to be the

corresponding r explanatory variables. The probability density function of Y is [15]:

$$f(y_i|\beta) = p_i^{y_i}(1-p_i)^{1-y_i} \tag{6}$$

where $p_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$.

For the entire sample, the likelihood function conditional on x_i is:

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{7}$$

To simplify the maximizing equation (7), its logarithm is used as such:

$$\begin{aligned} \ln L(\beta|\boldsymbol{y}) &= \ln \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \sum_{i=1}^{n} \ln p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \sum_{i=1}^{n} \ln p_i^{y_i} + \ln(1-p_i)^{1-y_i} \\ &= \sum_{i=1}^{n} (y_i \ln p_i + (1-y_i)\ln(1-p_i)) \end{aligned} \tag{8}$$

Further steps in maximizing the equation (8) includes partial differentiation using iterative processes [5].

Evaluating the quality of the model, KS (Kolmogorov-Smirnov) and the ROC (Receiver Operating Characteristic) curve required using [9], [25].

3.5. Neural networks

Artificial neural networks are machine learning methods that can be used for regression and classification type of problems [23]. The importance of this method is emphasized by Prieto et al. [24] showing that ANNs have been used in simulators, implementations, and real-world applications for a number of years, and has proven to be competitive in solving real-world problems. ANNs have also significantly contributed to the development of machine learning and other related areas. For classification purposes, they have often been used in comparison with logistic regression and discriminant analysis. Their advantages lie in robustness, the ability to work with missing data and the ability to approximate any nonlinear mathematical function [20]. In this paper, the multilayer perceptron (MLP) feed forward network was used in testing various optimization algorithms, such as backpropagation, conjugate gradient, and Broyden-Fletcher-Gordfarb-Shano, as well as two different activation functions: logistic and tangent hyperbolic.

The basic computation within an MLP neural network starts with the input layer where n input units, $i=1,2,\ldots, n,$ are multiplied by randomly determined initial weights w_i usually from the interval [-1,1]. Each unit in the middle (hidden) layer

receives the weighted sum of all x_i values as the input. The output of the hidden layer denoted as y_c is computed by:

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{9}$$

where f is the activation function. In this experiment, the logistic (i.e. sigmoid) activation function was used according to:

$$f(x_i) = \frac{1}{1+e^{g \cdot x_i}} \tag{10}$$

where g is the parameter defining the gradient of the function. To produce probabilities that could be modified to a binary output, a softmax activation function is added for classification purposes. The output is then compared to the actual output y_a, and the local error ε is computed. In the subsequent iterations, the process is repeated by adjusting the weights of the input vector according to a learning rule, usually the Delta rule [20], and an optimization algorithm is used to minimize error. Also, the number of hidden units was dynamically changed from 1 to 30 in a cross-validation procedure where the final number of hidden units was such that it produced the minimum error on the test set. The same procedure was used to optimize training time, while the maximum number of training epochs was set to 1000. In the postmodelling phase, sensitivity analysis was performed in order to analyze the relevance of input variables. To obtain the sensitivity coefficients, the value of the input variable was changed using a randomly selected percentage value (in the range of +-5%), while allowing all other input variables to retain their same values, leading to the observation of the model error changing. The sensitivity coefficient of each input was computed as the ratio of the average model error subject to changes of the examined input variable in relation to the model error without changes to the examined input variable. Having acquired the results of the sensitivity analysis, the user is then able to extract the important predictors, and also to perform a posteriori analysis of the predictor values that lead to the accurate prediction of output.

4. Results

4.1. Results of factor analysis

Factor analysis was applied on the set of independent variables separately for every year. The best result was achieved for the financial ratios in the year 2010. An analysis of the variance of eigenvalues showed that three factors could be generated, representing 99% of the total variance. The results are presented in Table 3.

Variable	Factor 1 (Turnover Factor)	Factor 2 (Liquidity Factor)	Factor 3 (Leverage Factor)
current assets/current liabilities	-0.00012	**0.998667**	-0.0001
total debt/total assets	3.04E-06	-0.000000172	**0.998749**
equity/total asset	-3.00E-06	0.000000261	**-0.99875**
(current assets-inventory)/ current liabilities	-6.30E-05	**0.998669**	-0.0000527
total revenue/total assets	**0.99873**	0.000197	0.000205
total revenue/current assets	**0.998119**	-0.00016	-0.00016
sales/total asset	**0.998516**	-0.0000416	-0.0000468
cash/current liabilities	0.000301	**0.847054**	0.000251

Table 3: *Factors extracted by factor analysis*

In observing the factor loadings, it becomes obvious that variables are grouped in a theoretically sound way for assessing the three groups of business performance indicators: business activity (turnover ratios), liquidity and leverage. The three extracted factors were used in a reduced set of inputs (designated as ML1, ML2, and ML3) with the addition of three original variables not included in any of the three factors: ROS_10 (net income/sales in 2010), Pnroe_10 (net income/equity (%)), zapos_p10 (number of employees in 2010).

4.2. Logistic regression and neural network results

To create a growth-classification model for companies, logistic regression was tested separately with all available variables in the input space, and with the factors extracted using factor analysis. Table 4 shows the results from the validation sample using both LR models.

Model	Area under Curve (AUC)	Kolmogorov-Smirnov statistics	Classification rate for non-high growth (%)	Classification rate for high growth (%)	Total classification rate
1 – Financial ratios as inputs	0.735	0.445	76.54	60.94	69.66
2 – Factors as inputs	0.573	0.193	59.68	58.00	58.93

Table 4: *Logistic regression results*

As can be seen from Table 4, in cases where financial ratios were used as input variables the LR method performed better than in the model with factors as inputs. The more accurate model produced a total classification rate of 69.66%, whereas the classification rate for non-high growth was 76.54%, which is significantly higher than the classification rate for the category of high-growth companies (60.94%). When applying logistic regression modelling, several different models were developed, either solely based on financial ratios as input variables or as a combination of factors and financial ratios. The model with the best hit rates, the Kolmogorov Smirnov statistics and the area under curve (AUC) was selected. The results of the model are shown in Table 5.

Variable code	Description of variable	Regression coefficient
Liquidity ratios		
Lclnw	current liabilities / equity	0.0023
Lkiui	current assets / total assets	0.512**
Turnover ratios		
Asalta	sales / total assets	0.2919***
Adug	total revenue / fixed assets	$7.9x10^{-7}$
Leverage ratios		
Zkz	total debt / total assets	0.0513
Profitability ratios		
Pnroe	net income / equity (%)	0.0004*
Prearnta	retained earnings / total assets	0.0932*

Other variables		
NTA	intangible assets / total assets	1.130*
Industry sector	high-tech	0.2431*

* Statistically significant at 10%
** Statistically significant at 5%
*** Statistically significant at 1%

Table 5: *Results of the logistic regression for a potential high-growth prediction model*

As can be seen in Table 5, the potential for growth increases in line with an increase in liquidity, turnover and profitability. Moreover, companies that belong to the high-tech industry sector and have higher share of intangible assets in total assets exhibit a higher probability of becoming high-growth companies.
Neural networks were also tested separately using all available variables in the input space, and factor analysis applied for extracting the the factors. Two different activation functions, logistic and tangent hyperbolic, were tested alternatively in order to find the approximation with the highest accuracy. The results on the out-of-sample validation data are shown in Table 6.

Model	NN method and structure	Activation function	Area under Curve (AUC)	Kolmogorov-Smirnov statistics	Classification rate for growth category 0 (%)	Classification rate for growth category 1 (%)	Total classif-ication rate
1 – All available variables	MLP NN	Logistic	0.6836	0.46118	77.78	62.50	**71.03***
2 – Factors as inputs	MLP NN	Logistic	0.6748	0.34967	70.97	64	67.86

*The highest total classification rate

Table 6: *Neural network results*

It becomes evident that the most accurate model was obtained using NN with the logistic function. The network had 13 hidden units. The total classification rate of the best NN model was 71.03% for the validation set, while the accuracy was higher for growth category 0 (77.78%) than for category 1 (62.50%). Sensitivity analysis of the most accurate NN model is presented in Figure 1.

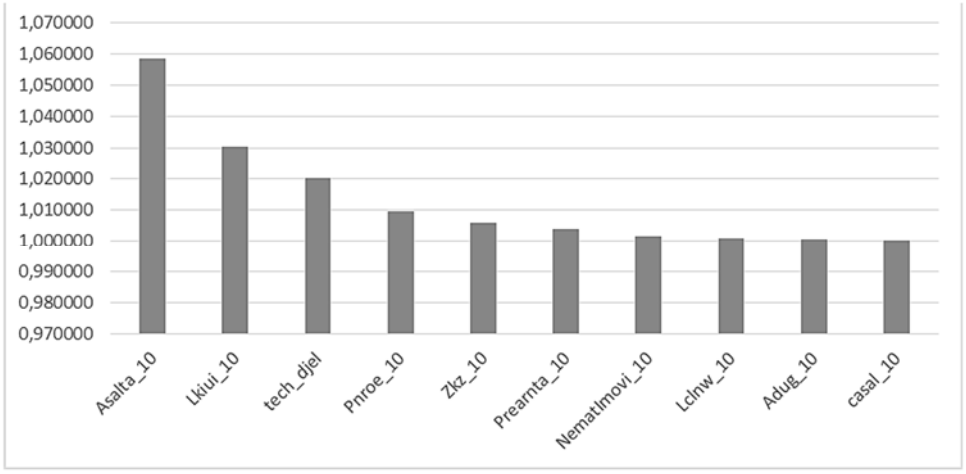


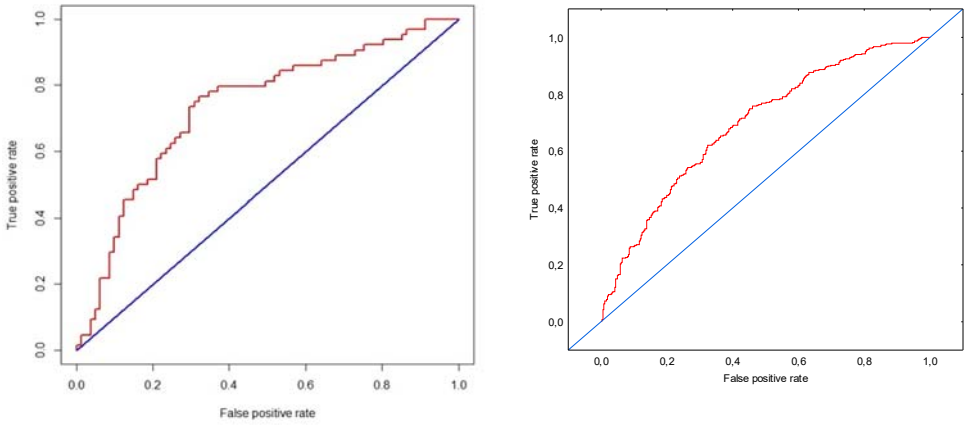
Figure 1: *Sensitivity analysis of the best NN model*

The graph depicting sensitivity ratios reveals that the variable with the highest impact on company growth is *Asalta_10* (sales/total assets), followed by *Lkiui_10* (current assets/total assets), *tech_djel* (industry sector), *Pnroe_10* (net income/equity), *Zkz_10* (total debt/total assets), *Preanrnta_10* (retained earnings/total assets) and *NematImovin_10* (intangible assets/total assets). Their sensitivity coefficient exceeds 1, which indicates that they contribute positively to accurady of the model. Other input variables have a sensitivity coefficient of 1 or lower, meaning that their absence from the model does not reduce the total classification rate of the ANN model.

4.3. Comparison of results

At first, it was interesting to see whether the LR model with all the original variables was statistically different in terms of accuracy than the LR model which uses factors as predictors. The t-test on difference in the proportions was used and the results showed (p=0.0283) that the LR model obtained from the original variables is significantly more accurate than the LR model obtained from factors. The same test was conducted for the NN models, and the obtained p-value (p=0.2790) shows that the NM model based on the original variables does not significantly differ in performance than the NM model that uses factors.
To compare the accuracy of the best LR and the best ANN model (the models with original variables), a statistical test of difference in proportion was conducted. The test produced a p-value of 0.3992 showing that there is no significant difference in the total classification rate between the best LR and the best NM model.

The area under curve (AUC), which explains the probability that a model will accurately classify a randomly chosen high-growth company into the correct category, and the shape of ROC curves, implies that the LR model performs slightly better (LR AUC: 0.735, NN AUC: 0.6836), meaning that it is more stable in recognizing true positives, whereas Kolmorogov-Smirnov (KS) statistics is slightly higher for the NN model. The ROC curves for both the LR and NN model are presented in Figure 2.



a) b)

Figure 2: ROC *curves for* a) LR *model*, b) NN *model*

Examining the way the two estimators classify companies as potentially high-growth or non-high growth involved applying the McNemar test and comparing estimations of the two models. The McNemar test is generally used to assess an experiment in which a sample of n subjects is evaluated based on a dichotomous dependent variable and assumes that each of the n subjects contributes to two scores for the dependent variable [28]. For the purpose of this test, estimations obtained by LR and NN were grouped into two basic categories designated as 1 if a company was assigned to a high-growth category , and 0 otherwise. The hypotheses used in our research: H0: $pb = pc$, and H1: $pb - pc$ were tested for each pair of two estimators using McNemar's test:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{11}$$

where b is the number of companies assigned to category 1 using LR and those to category 0 using the NN model, where c is the number of companies assigned into category 0 using the LR model and those to category 1 using the NM model. If

the two estimators tend to assign different companies to category 1, there should be a significant difference in the probabilities of the distribution table for the positions relating to scores b and c. If the difference is not shown, the conclusion is that both estimators categorize companies with the same probability. In fact, the value of McNemar's test was 0.00, meaning that there is no significant difference in the way the ANN model and the LR model classify company growth. Although the ANN model provides greater accuracy than the LR model, some similarities, advantages and disadvantages of both approaches are noticeable. Both methods are more accurate in recognizing low-growth companies than high-growth companies, indicating that the input variables do not contain enough information as an indicator of high growth. The advantage of the ANN model lies in its accuracy, but the LR model is more informative in explaining the relationship among the input variables.

Based on the variables retained in the models, it is evident that both of our models extracted liquidity, turnover of assets, return on equity, retained earnings, intangible assets and high-tech industry as important determinants for growth prediction. There are some similarities and differences in comparison to previous research. Sampagnaro [26] points out that liquidity has a positive correlation to growth. He also emphasizes the importance of total assets turnover for high-growth companies that rely on their capacity to generate sales from assets. Khan, Theunissen [16] tested the relationship between ROE and growth on Belgium companies, and realized that ROW has no significant effect on growth. Mateev, Anastasov [21], Sampagnaro [26] and Segarra-Blasco, Teruel [27] show that external financing resources have a negatively impact growth. This is consistent with our findings where we have shown that financing through retained earnings increases potential for growth. Furthermore, we have also shown that intangible assets as a proxy for R&D have a positive impact on high growth which in turn has also been proven by Helmers and Rogers [13].

5. Discussion and conclusion

This paper outlines the creation of a model for predicting SME growth using logistic regression and neural networks. The input space consisted of real financial data from Croatian companies, where growth was a categorical dependent variable, indicating whether an enterprise was high-growth if its annualized growth in assets exceeded 20% over a three-year period. In the first stage of modelling, factor analysis was conducted to extract the most important input components. Logistic regression and artificial neural networks were used to establish the most accurate model for providing the highest classification rate on the validation set. The results showed that factor analysis did not improve the accuracy of both LR and ANN models, and that the accuracy of the best neural network and the best logistic regression is not statistically different, meaning that

either method can be used for predicting growth. When testing the significance of the difference in their results, the consistency of the parametric t-test and the McNemar's nonparametric test, as well as that of the ROC curves became clear. When analyzing the variable importance in the third stage of the research, the observation is that the most important predictor of company growth on the observed dataset is turnover ratio computed based on sales/total assets, followed by the liquidity ratio computed using current assets/total assets, industry sector, two profitability ratios: net income/equity and retained earnings/total assets, a leverage ratio computed by total debt/total assets and intangible assets/total assets ratio. However, there are advantages to the LR approach in its explanatory power, while ANNs are more sensitive to the sampling procedure. The results also bring a question the possible integration of both methods so as to improve accuracy. Suggestions for further research are the use ANNs in the first phase as a nonlinear variable selector for important predictors, while the LR can be used to produce a relationship among such predictors. Furthermore, other machine learning methods can be tested, such as support vector machines and decision trees. In selecting variables, what could also be interesting is the inclusion of other variables in order to increase accuracy of the models, such as the capacity to invest, import, export, productivity. In addition, measuring a company growth in terms of sales, number of employees or productivity is certainly a noteworthy test. The research may assist investors and economic policy makers in providing them with tools for classifying companies as having growth potential, especially those small and medium enterprises.

Acknowledgement

This study is funded by the Croatian Science Foundation under Grant No. 3933 Development and application of growth potential prediction models for SMEs in Croatia.

References

[1] Barringer, B. R., Jones, F. F., and Neubaum, D. O. (2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders, Journal of Business Venturing, 20(5), 663-687.

[2] Becchetti, L., Trovato, G. (2002). The determinants of growth for small and medium sized firms. The role of the availability of external finance, Small Business Economics, 19(4), 291-306.

[3] Cassia, L., Cogliati, G. M., Paleari, S. (2009). Hyper-Growth Among European SMEs: An Explorative Study, 2009, Available at SSRN 1389521.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1389521 [Accessed 01/09/2016]

[4] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), 20(2), 215-242.

[5] Czepiel, S.A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation, http://czep.net/stat/mlelr.pdf [Accessed 13/07/16].

[6] Davidsson, P., Delmar, F., Wiklund, J. (2006). Entrepreneurship as growth; growth as entrepreneurship. in Davidsson, P, Delmar, F, Wiklund, J (Eds.) Entrepreneurship and the Growth of Firms. Cheltenham, United Kingdom: Edward Elgar Publishing, 21-38.

[7] Delmar, F. (2006). Measuring growth: methodological considerations and empirical results. In Davidson, P., Delmar, F., Wiklund, J., Entrepreneurship and the Growth of Firms. Cheltenham, United Kingdom: Edward Elgar Publishing, 62-84.

[8] Delmar, F., Davidsson, P., Gartner, W. B. (2003). Arriving at the high-growth firm, Journal of Business Venturing, 18(2), pp. 189-216.

[9] Fawcett, T. (2006). An introduction to ROC analysis, Pattern Recognition Letters, 27(8), 861-874.

[10] Fischer, E., Reuber, A. R., Hababou, M., Johnson, W., Lee, S. (1997). The role of socially constructed temporal perspectives in the emergence of rapid growth firms. Entrepreneurship Theory and Practice, 22(2), 13–30.

[11] Freel, M. S., Robson, P. J. (2004). Small firm innovation, growth and performance evidence from Scotland and Northern England. International Small Business Journal, 22(6), 561-575.

[12] Geroski, P.A. (2005). Understanding the implications of empirical work on corporate growth rates, Managerial and Decision Economics, 26(2), 129–138.

[13] Helmers, C., Rogers, M. (2011). Does patenting help high-tech start-ups? Research Policy, 40(7), 1016-1027.

[14] Hua, Z., Wang, Y., Xu, X., Zhang, B., Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression, Expert Systems with Applications, 33(2), 434-440.

[15] Jobson, J. (2012). Applied multivariate data analysis: volume II: Categorical and Multivariate Methods, New York: Springer Science & Business Media.

[16] Khan, A.A., & Theunissen, L. (2012). Determinants of firm growth: evidence from Belgian companies, Gent University, http://lib.ugent.be/catalog/rug01:001893465 [Accessed 01/09/2016]

[17] Kolvereid, L., Bullvag, E. (1996). Growth intentions and actual growth: The impact of entrepreneurial choice, Journal of Enterprising Culture, 4(1), 1-17.

[18] Luengo, J., García, S., Herrera, F. (2009). A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests, Expert Systems with Applications, 36, 7798–7808.

[19] Ma, R., Tang, C. (2007). Building up Default Predicting Model based on Logistic Model and Misclassification Loss, Systems Engineering - Theory & Practice, 27(8). 33-38.

[20] Masters, T. (1995). Advanced Algorithms for Neural Networks, A C++ Sourcebook, New Jersey: John Wiley & Sons.

[21] Mateev, M., Anastasov, Y. (2010). Determinants of small and medium sized fast growing enterprises in central and eastern Europe: a panel data analysis, Financial Theory and Practice, 34(3), 269-295.

[22] Morone, P., Testa, G. (2008). Firms growth, size and innovation - An investigation into the Italian manufacturing sector, Economics of Innovation and New Technology, 17(4), 311-329.

[23] Paliwal, M., Kumar, U.A. (2009). Neural networks and statistical techniques: A review of applications, Expert Systems with Applications, 36(1), 2–17.

[24] Prieto, A., Prieto, B., Martinez Ortigosa, E., Ros, E., Pelayo, F., Ortega, J., Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges, Neurocomputing, in press, Available online 8 June 2016, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2016.06.014.

[25] Řezáč, M., Řezáč, F. (2011). How to measure the quality of credit scoring models, Finance a úvěr: Czech Journal of Economics and Finance, 61(5), 486-507.

[26] Sampagnaro, G. (2013). Predicting rapid-growth SMEs through a reversal of credit-scoring principles, International Journal of Entrepreneurship and Small Business, 18(3), 313-331.

[27] A. Segarra-Blasco, and M. Teruel, M. (2009). Small firms, growth and financial constraints, XREAP, November 2009. SSRN: http://ssrn.com/abstract=1825064 [accessed 01.09.2016]

[28] Sheskin, D.J. (2004). Handbook of Parametric and Nonparametric Statistical Procedures, Washington D.C.: Chapman & Hall/CRC.

[29] Storey, D. (1994). Understanding the small firm sector. Routiedge, London.

[30] Weinzimmer, L. G., Nystrom, P. C., Freeman, S. J. (1998), Measuring organizational growth: Issues, consequences and guidelines, Journal of Management, 24(2), 235-262.