Abstract. Issue of enterprise financial distress represents the actual and interdisciplinary topic for the economic community. The bankrupt is thus one of the major externalities of today's modern economies, which cannot be avoided even with every effort. Where there are investment opportunities, there are individuals and businesses that are willing to assume their financial obligations and the resulting risks to maintain and develop their standard of living or their economic activities. The decision tree algorithm is one of the most intuitive methods of data mining which can be used for financial distress prediction. Systematization literary sources and approaches prove that decision trees represent the part of the innovations in financial management. The main propose of the research is a possibility of application of a decision tree algorithm for the creation of the prediction model, which can be used in economy practice. Paper main aim is to create a comprehensive prediction model of enterprise financial distress based on decision trees, under the conditions of emerging markets. Paper methods are based on the decision tree, with emphasis on algorithm CART. Emerging markets included 17 countries: Slovak Republic, Czech Republic, Poland, Hungary, Romania, Bulgaria, Lithuania, Latvia, Estonia, Slovenia, Croatia, Serbia, Russia, Ukraine, Belarus, Montenegro and Macedonia. Paper research is focused on the possibilities of implementation of decision tree algorithm for creation of prediction model in the condition of emerging markets. Used data contained 2,359,731 enterprises from emerging markets (30% of total amount); divided into prosperous enterprises (1,802,027) and non-prosperous enterprises (557,704); obtained from Amadeus database. Input variables for model represented 24 financial indicators, 3 dummy variables and countries GDP data, in the years 2015 and 2016. The 80% of enterprises represented training sample and 20% test sample, for model creation. The model correctly classified 93.2% of enterprises from both the training and test sample. Correctly classification of non-prosperous enterprises was 83.5% in both samples. The result of the research brings the new model for identification of bankrupt of enterprises. The created prediction model can be considered sufficiently suitable for classifying enterprises in emerging markets.

Keywords: prediction model, decision tree, emerging markets.
**Introduction.** Nowadays, in a dynamic economic environment and increasing competitive struggle, it is more difficult to maintain its market position and achieve prosperity for enterprises (Konigova, et al., 2012; Rajnoha and Lorincova, 2015; Afonina, 2015; Zyka and Drahotsky, 2019). For this reason, it is desirable for an enterprise to continually investigate its financial health; to know its strengths and weaknesses; to know the direction to prevent the enterprise from failing. It is precisely the prediction of bankruptcy that has been the subject of many economists around the world for many years. The results of their works are predictive models, by means of which it is possible to predict the financial health of the company for several years in advance. The bankrupt is thus one of the major externalities of today’s modern economies, which cannot be avoided even with every effort. Where there are investment opportunities, there are individuals and businesses that are willing to assume their financial obligations and the resulting risks in order to maintain and develop their standard of living or their economic activities. (Fialova and Volvarcn, 2020)

Nowadays, the problem with the failure of business entities is a very actual topic in the economy. One of the biggest business risks is credit risk (Belas et al., 2012; Cipovova and Belas, 2012; Kljuncnikov et al., 2016), which relates with secondary insolvency of business entity. The failure of the business entity has a negative influence on all subject with a relationship with this business entity. Possibility of evaluating and predicting credit risk and financial situation of business entities is an advantage for creditors, investors and business owners. Based on the assessment and forecasts of credit risk and financial health of business entities, it was possible to take necessary corrective action, in time. Economists are able to evaluate and predict credit risk and financial health of business entities due to ex-ante financial analysis - through prediction models. Prediction models can early predict the probability of failure of the business entity. The problem which is necessary to solve, for application of prediction models, is the selection of prediction model which could provide the best evaluation of the probability of failure of analyzed business entity and of course, the model which could provide exact and relevant results. Over the years, five basic groups of prediction models have been developed: 1) Statistical models, e.g. one-dimensional analysis, multidimensional discrimination analysis, logit, probit; 2) Mathematical programming, e.g. linear programming, multi-criteria decision-making; 3) Artificial Intelligence, e.g. fuzzy models, decision trees, neural networks; 4) Credit risk models, e.g. Merton’s model, KMV model; and 5) Alternative theoretical models, e.g. multi-logit models, cash-flow models, chaos theory and the like. The paper assumption is a possibility of application of a decision tree algorithm for the creation of the prediction model, which can be used in economy practice. The decision tree algorithm is one of the most intuitive and most commonly used methods of data mining, especially since this approach provides explicit classification rules. These rules are very simple and easy to interpret, leading to a quick evaluation of the results. Decision trees deal fairly well with varied or missing data, as well as non-linear effects between variables. When creating predictive models, decision trees are (not only) currently the only major competitor of classical approaches, namely multidimensional discriminant analysis and logistic regression. It should be stressed that decision trees are at the interface between predictive and descriptive method because they create the classification structure of the data set to which they apply. Therefore, they belong to the class of hierarchical grouping methods. A decision tree is a graph structure in the form of a tree that is used to divide a relatively large heterogeneous data set into smaller, more homogeneous subsets by successively applying simple decision rules. This tree structure contains a root node, internal nodes, and terminal nodes. Paper main aim is to create a comprehensive prediction model of enterprise financial distress based on decision trees, under the conditions of emerging markets. Emerging markets included 17 countries: Slovak Republic, Czech Republic, Poland, Hungary, Romania, Bulgaria, Lithuania, Latvia, Estonia, Slovenia, Croatia, Serbia, Russia, Ukraine, Belarus, Montenegro and Macedonia. The countries from emerging markets were chosen for their common economic conditions and similar bankrupt and financial distress reasons, which are described in the literature review. Paper data were obtained from the Amadeus database for 2015 (for
all independent variables) and 2016 (for the dependent variable). Data overall contained 2,359,731 enterprises from emerging markets; divided into prosperous enterprises (1,802,027) and non-prosperous enterprises (557,704). They presented the dependent variable. Input variables for model represented 24 financial indicators, 3 dummy variables and countries GDP data. They represented independent variables. The decision tree was constructed in IBM SPSS Statistics 24 using algorithm CART with the top-down method. The tree has always been pruned to prevent overfitting. The 80% of enterprises represented training sample and 20% test sample, for model creation.

Paper consists of five chapters, namely: 1) literature review which includes the brief historical development of decision trees algorithm; 2) methodology which includes the methodology of decision trees creation and methods which were used in the paper; 3) empirical results which include results of empirical research and prediction model of enterprise financial distress based on decision trees; 4) discussion; 5) conclusions.

**Literature Review.** The first studies dealing with enterprise bankrupt has been arising since the 30s of the 20th century, during the Great Depression. According to Altman (1993), Smith and Winakor (1935), they were the first to deal with bankruptcy prediction in their studies seriously. Their findings were followed by Merwin (1942). Both studies pointed out that failing firms report significantly different values of selected financial ratios than successful firms. This fundamental principle was a huge breakthrough and offered considerable prospects for further research. Among the studies focused on the issue is also the work of Fitzpatricka (1932) dealing with significant differences between successful and failed businesses. This work has inspired many applied studies that began in the mid-1960s. Among the most important are the studies: Beaver (1966), Altman (1968), Edmister (1972), Tamariz (1976), Springat (1978), Ohlson (1980), Beerman (1982), Fulmer (1984), Zmijewski (1984), Bathory (1984), Argenti (1985), Zavgrenen (1985), Jones (1987), Salchenberger (1992), Shirat (2002), and others. The studies are described in detail in, for example, research articles. (Dimitras et al., 1996; Kumar and Ravi 2006; Bellovary et al., 2007; Balcaen and Ooghe 2004). The issue of enterprise default, financial distress, insolvency and bankrupt is still a managerial challenge to an interdisciplinary problem. Many authors have tried to find basic reasons for this negative economic situation. Enterprise is in bankrupt if it is insolvent. It means it is impossible to pay its two commitment at least to one creditor 30 days after their maturity. Slattery and Lovett in 1999 described several basic reasons for enterprise bankrupt in emerging markets. They created two basic groups, namely: endogenous and exogenous. Endogenous reasons include poor management and its failures, insufficient financial control, poor working capital management, high expenses, insufficient marketing, inappropriate financial policy of the company. Exogenous reasons include negative changes in market demand, competition, a change in prices of input commodities in an unfavourable direction. Mitroff (2011) defined eight basic reasons for worsening financial health of an enterprise in Emerging markets – economic reasons, information reasons, physical reasons, human resources, image, natural disaster, crime. Newton (2005) considers the causes of the enterprise financial distress in emerging markets to be: the inability to manage cash flow; the low value of equity; absence of a good business plan; determination of unrealistic goals; excessive optimism; the inexperience of management; organizational arteriosclerosis. The first studies devoted to bankruptcy prediction were based on one-dimensional analysis of the financial ratios. These studies simply analyzed financial ratios and compared the results of these indicators in creditworthy enterprises and enterprises in bankrupt. In 1930, the Bureau of Business Research presented a study that analyzed the development of 24 financial ratios from 29 industrial enterprises in bankrupt (Fitzpatrick 1932; Malin 2017).

In 1935, R. F. Smith and A. Winakor verified the BBR study results. They analyzed financial indicators of 183 enterprises in bankrupt, and their results confirmed BBR study (Bellovary et al., 2007). W. H. Beaver was the first economist who used statistical methods for prediction of the financial health of enterprises. In the study «Financial Ratios of Prediction of Failure» (1966), under the certain number of financial
indicators the author divided enterprises into the two categories of creditworthy enterprises or bankrupt enterprises (Beaver, 1966; Jones, 1987). In 1968, E. I. Altman created one of the most famous and also the first bankruptcy prediction model, which is known as «Z-score». This model interconnected the explanatory power of several variables. This model is the basic stone of multiple discrimination analysis (Altman, 1968). Since this time, the number of bankruptcy models has risen. In the 70s of 20th century, were published 28 studies about prediction models, in 80s of 20th century were published 53 studies and in 90s of 20th century were published 70 studies about prediction models. Other researchers who have tried to improve Altman’s multiple discrimination analysis are Deakin (1972), Taffler (1974), Loris (1976), Springate (1983), Fernandez (1988), Neumaier and Neumaierova (1995, 1999, 2000, 2005), Gajdka and Stos (1996), Virag and Hajdu (1996), Chrastinova (1998), Binkert (2000), Gurcik (2002), Sharita (2003) and so on (Virag and Kristof, 2005; Mousavi et al., 2015; Gurcik, 2002; Agarwal and Taffler, 2007; Kubickova, 2015; Zavrgen, 1985). In the 70s of 20th century, also raised prediction models based on the logistic regression - logit and probit models. The first authors who used logistic regression for prediction the bankrupt of the enterprise were Santomero and Vinso (1977) and Martin (1977). They only analyzed bankrupt of American banks. In 1980, Ohlson analyzed bankrupt of enterprise by logistic regression in general. The result of his model is one value which directly determining the probability of bankrupt of the enterprise. Logit analysis was also analyzed in works of Casey and Bartczak (1985), Zavrgen (1985), Pantalone and Platt (1987), Jakubik and Teply (2006), Saijter (2008), Hurtosova (2009), Bredar (2014), Gurka (2016). (Zavrgen, 1985, Ohlson, 1990, Poddig, 1995, Siekelova, 2017, Hidlovsky and Kral (2014).


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination analysis</td>
<td>2</td>
<td>22</td>
<td>28</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td>68</td>
</tr>
<tr>
<td>Logit analysis</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>16</td>
<td>3</td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>Probit analysis</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Neural networks</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>4</td>
<td>15</td>
<td>55</td>
</tr>
<tr>
<td>Different</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>28</td>
</tr>
</tbody>
</table>

Source: developed by the authors.

According to data in previous Table 1, the most commonly used method for prediction model creation is discrimination analysis. At the second place are neural networks. However, between neural networks and logit analysis is very thin border – only nine studies. Logit analysis was used in 46 studies about prediction models, and neural networks were used in 55 studies about prediction models. Paper research
is focused on the possibilities of implementation of decision tree algorithm for creation of prediction model in the condition of emerging markets. The first computer-implemented algorithm of the decision tree was founded in 1963 by J. Sonquist and J. Morgan developed the Automatic Interaction Detection (AID) algorithm. It is the first computer-implemented algorithm for generating a decision tree, according to Morgan and Sonquist (1963). The extension of this algorithm is CHAID (Chi-squared Automatic Interaction Detection), which is still one of the most widespread algorithms. CHAID uses both nominal and ordinal variables to model the nominal output variable. Another is Exhaustive CHAID algorithm. The motivation for the creation of the Exhaustive CHAID algorithm was, in particular, shortcomings of the CHAID algorithms, following Biggs et al. (1991).

The ID3 (Interactive Dichotomizer) algorithm is the classic case of an algorithm that generates a tree from top to bottom (TDIDT). These trees are classification trees, i.e. the output and input variables are categorical, stated by Quinlan (1986). The ID4, ID5, and finally by ID5R algorithms are the incremental modification of the ID3 algorithm, stated by Utgoff (1989). The algorithm C4.5 is derived from the ID3 algorithm by the same author. Input variables can be discrete but also continuous, Quinlan(1993). In 1998 Quinlan created a newer version of this algorithm implemented as C5.0 (for Unix and Linus) or See5 (for Windows). Over time, other algorithms have been created, e.g. CART, RECPAM, FACT, GUIDE, BART etc. Finally, the total evolution of decision trees algorithms could be divided into four basic categories – generations. This paper works with algorithm CART, which will be explained detail in methodology. A comprehensive review of decision trees algorithms evolution captures Table 2.

### Table 2. Review of decision trees algorithms evolution

<table>
<thead>
<tr>
<th>1st generation</th>
<th>2nd generation</th>
<th>3rd generation</th>
<th>4th generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AID</td>
<td>Morgan and Sonquist (1963)</td>
<td>QUEST</td>
<td>Loh and Shih (1997)</td>
</tr>
<tr>
<td>MVPART</td>
<td>Zhang (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>De’ath (2002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Su et al. (2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID3</td>
<td>Quinlan (1986)</td>
<td>MOB</td>
<td>Zeileis et al. (2008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chipman et al. (2010)</td>
</tr>
</tbody>
</table>

**Source:** developed by the authors.

**Methodology and research methods.** This chapter provides brief characteristics of research methodology which were used in this paper. The chapter is divided four parts: 1) a general introduction to the methodology of decision tree creation, with emphasis on algorithm CART, which was used in this paper; 2) determination of splitting criterion, with emphasis on Gini index which was used in this paper; 3) methodology of pruning of decision tree; 4) data collection.
The methodology of decision tree creation. For creating a decision tree to classify cases into \( m \) classes, it is necessary to define the criterion for selecting the variable and condition that will best divide the cases of individual classes in the given node. Then one can apply this repeatedly in the individual «parent» nodes of the tree to create the «child» notes. These nodes are repeatedly divided until the stopping condition is met. The stopping condition of tree generation is usually a combination of multiple rules:

1. Nodes are clean, i.e. they only contain the cases with the same value of the output variable.
2. A predetermined depth limit of a tree (the number of levels of tree divisions) was reached.
3. A fixed maximum number of tree leaves was reached.
4. A minimum number of cases in each node (usually between 75 to 100) was reached, where it is assumed that the nodes cannot be divided anymore.
5. Further division of any node would lead to the creation of child nodes with a lower frequency than a predetermined value.

The decision tree model describes a tree graph consisting of nodes and directed edges. The tree generation begins with a root node, which is sequentially divided into (non-terminal) nodes and (terminal) nodes, the so-called leaves. Each node (apart from the leaves) is linked from top to bottom by directed edges that originate from it, with at least two child nodes. Leaf nodes do not have child nodes. The decision tree was created by CART algorithm. The growth algorithm CART (Classification and Regression Trees), according to Breiman (1984), is one of the most efficient and globally-used algorithms generating decision trees. The CART algorithm is a registered trademark of the company Salford Systems. However, similar names such as CRT and CART are often used for naming. The CART generates binary trees. To find the best splitting of nodes, it uses the Gini index. In addition to this option, CART algorithm makers provided several technical solutions that bring two main benefits: universality and performance. Universality is based primarily on the fact that the number of categories of a dependent variable can be either final or even infinite. While the CART can be used to create classification trees (categorical output variable) as well as regression trees (continuous output variable). For each type of output variable, there is an appropriate criterion for node splitting. Universality also lies in the possibility of processing the missing values by replacing each variable with the same splitting variable or the same pruning variable. The same splitting variable is such a variable that provides approximately the same node purity as the original variable. The same pruning variable is such a variable that distributes subjects in approximately the same way as the original variable. These variables could be used as «substitute» variables, but it is best to use the same pruning variables to ensure tree consistency. The performance of the CART algorithm is primarily secured by its pruning mechanism, which is more sophisticated than in the case of the CHAID algorithm. The maximum tree is designed to continue with the node splitting process until it is possible. The algorithm then derives several nested subtrees by sequential pruning operations, compares them, and then selects the one that has the lowest possible error rate measured by cross-validation. The performance is also related to the absence of arbitrary but fixed thresholds. The examples of that are the significance threshold or the \( X^2 \) significance test in case of the CHAID algorithm. Determining this threshold is always difficult because it is necessary to make the best choice between the threshold giving a too bushy tree that does not have enough robustness because it depends too much on the specimen and the threshold giving such a small tree with a lower prediction power. The performance of the CART algorithm also lies in its exhaustive search for all possible divisions of a given node, which guarantees optimal splitting. This search may take a long time, especially when it comes to qualitative variables with a large number of categories because it is necessary to test too many potential divisions (Svabova et al., 2018; Dvorsky et al., 2020).

Splitting criterion. The basic principle of tree growth is to increase the purity of the child nodes. When assessing this purity, the purity rates based on entropy are the most commonly used. In particular, the
information gain and relative information gain are used. The selection of the splitting variable in the case of a categorical output variable is carried out by various methods, e.g., entropy, Gini index, information gain or $x^2$ independence test. This research worked with a categorical output variable representing the prosperity or non-prosperity of an enterprise and paper worked with CART algorithm. For this reason, Gini index as a splitting variable was used. The use of the Gini index when choosing the optimal splitting variable is basically very similar to the use of entropy. The Gini index of the output variable $Y$ is defined by the relationship:

$$G(Y) = 1 - \sum_{j=1}^{m} p_f^j = 1 - \sum_{j=1}^{m} \left( \frac{n_j}{n} \right)^2$$

(1)

where $p_f^j$ is again the probability of the occurrence of the class $y_i$ - estimated as the relative frequency of the $j^{th}$ class $p_j = \frac{n_j}{n}$.

Furthermore, analogously to entropy, we determine the expected Gini indices for individual input variables $X_i$:

$$G(X_i) = \sum_{l=1}^{l} \frac{n(x_{il})}{n} G(x_{il})$$

(2)

where $G(x_{il})$ - is the Gini index on the set of the cases in the given node for which the variable $X_i$ takes the value $x_{il}$:

$$G(x_{il}) = 1 - \sum_{j=1}^{m} \left( \frac{n_j(x_{il})}{n(x_{il})} \right)^2$$

(3)

The variable $X_i$ with the smallest value of the Gini index $G(X_i)$ or with the highest reduction of the given index impurity will be used for branching of the given node:

$$Z_{Gini}(X_i) = G(Y) - G(X_i)$$

(4)

Pruning of decision trees. When creating a decision tree, it is necessary to achieve a detailed description of the relationships within the training set. It leads to the creation of a so-called maximal tree that faultlessly explains the relationships within the training set. However, these relationships are often not generally valid, and therefore, when applying the tree to other data, a high error rate can occur. We say that such a tree is overfitting and it is mostly graphically displayed with an excessive number of branches that are a thin and small number of leaf nodes.

The solution to the problem of overfitting of a decision tree is pruning. One option is to trim a tree already during the creation. In this case, the growth of some branches is prematurely terminated if there is a sufficiently high probability that the cases within this branch belong to the same classification class.
In practice, pruning of an existing maximum tree is more often used. In this case, when pruning a tree, it is assessed how its classification capability will deteriorate. When a tree grows, and when it is pruned; it is possible to use a training set, but more often a training set is used to generate a tree, and a testing set is used to trim it. From the set of all subtrees, the one that achieves the smallest degree of incorrect classification is selected (Alexander and Grimshaw, 1996; Reitano, 2015).

For creating prediction models, it was necessary to create the database for which were used the financial and statistical indicators from the Amadeus database for 2015 (for all independent variables) and 2016 (for the dependent variable). A comprehensive prediction model based on decision trees was created under the conditions of emerging markets, which include: Slovak Republic, Czech Republic, Poland, Hungary, Romania, Bulgaria, Lithuania, Latvia, Estonia, Slovenia, Croatia, Serbia, Russia, Ukraine, Belarus, Montenegro, and Macedonia. However, data from Belarus was removed from the database due to a large amount of missing data. Overall, the database contained 2,359,731 enterprises, which were divided into prosperous enterprises (1,802,027) and non-prosperous enterprises (557,704). For verification of the model, classification ability was enterprises in database divided into a training and test sample in the ratio of 80%/20%. The training sample was used to derive the discriminatory function, and the test sample was then used to verify the ability of the model to be classified.

**Results.** Paper main aim was to create a comprehensive prediction model of enterprise financial distress based on decision trees, under the conditions of emerging markets. This chapter provides characteristics of research results which were obtained in this paper. The chapter is divided into two parts: (I) model indicators; (II) data analysis.

Model indicators. Model input indicators (independent variables) represented quantitative and qualitative variables. Quantitative variables included 24 financial-economic indicators for every enterprises and country GDP. Qualitative variables (in model marked as dummy variables) included country data, the size of enterprises (small, medium and large according to Amadeus database criteria), and sector according to NACE classification. Model output indicator (independent variable) was defined by prosperity and non-prosperity of the enterprise.

For the determination of the independent variables to be used for the creation of the predictive bankruptcy model, was focused on indicators determined by the prominent author as key predictors of financial health. For this purpose, were used studies and research of Sharifabadi et al. (2017), Tian et al. (2015), Bellovary et al. (2007), Ravi Kumar and Ravi (2007), Dimitras et al. (1996), Agarwal et al. (2018) and Ciszewski and Nowakowski (2018). However, two studies were mainly considered, namely Bellovary et al. (2007) «A Review of Bankruptcy Prediction Studies» 1930 to present, and Ravi Kumar and Ravi (2007) «Bankruptcy Prediction in Banks and Firms through Statistical and Intelligent Techniques – A Review». The authors of the first study analyzed 165 prediction models until 2004. They state that 752 different variables were used in the models, with up to 674 of these variables being used in only one or two models. After the study, they present 42 variables which were used in more than five models. The authors of the second study followed the previously published studies by Calderon et al. (2002), Dimitras et al. (1996), O’Leary (1998) and Scott (1981), Afonina (2015). They completed these studies by their research. Together they analyzed 62 prediction models. Based on the completed analysis, Table 3 presents indicators selected for the model. Given the lack of required data in individual variables of individual countries, some variables from further investigation had to be discarded. These are financial indicators X03, X05, X06, X13, X14, X17, X19, X23, X28, X29, X31, X32, X33, X34.

Besides, for the construction of the model, data from Belarus were not used, since most of them were not available in the database, and, therefore, it was not possible to include this country in the creation of this model. Subsequently, enterprises for which the value of the dependent variable was not listed, also had to be excluded. Because it was not possible to determine whether an enterprise is or is not prospering.
Subsequently, enterprises for which the value of the dependent variable was not listed, also had to be excluded. The reason was the impossibility to determine whether an enterprise is or is not prospering.

The decision tree was constructed in IBM SPSS Statistics 24 using the CART algorithm. The created tree has been pruned to prevent it learned and to ensure clarity and easy to the interpretation of its results.

Table 3. Model input variables (independent variables)

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Calculation Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>X01</td>
<td>Sales/Total Assets</td>
</tr>
<tr>
<td>X02</td>
<td>Short-term Assets/Short-term Liabilities</td>
</tr>
<tr>
<td>X03</td>
<td>Gross profit/Total Assets</td>
</tr>
<tr>
<td>X04</td>
<td>Net income/equity</td>
</tr>
<tr>
<td>X05</td>
<td>EBITDA/Sales</td>
</tr>
<tr>
<td>X06</td>
<td>(Long-term + Short-term Liabilities)/EBITDA</td>
</tr>
<tr>
<td>X07</td>
<td>Net Income/Total Assets</td>
</tr>
<tr>
<td>X08</td>
<td>Working Capital/Total Assets</td>
</tr>
<tr>
<td>X09</td>
<td>Operating profit/Total assets</td>
</tr>
<tr>
<td>X10</td>
<td>(Long-term + Short-term Liabilities)/Total Assets</td>
</tr>
<tr>
<td>X11</td>
<td>Short-term assets/Total assets</td>
</tr>
<tr>
<td>X12</td>
<td>Cash and Cash Equivalents/Total Assets</td>
</tr>
<tr>
<td>X13</td>
<td>Cash flow/Total Assets</td>
</tr>
<tr>
<td>X14</td>
<td>Cash flow/(Long-term + Short-term Liabilities)</td>
</tr>
<tr>
<td>X15</td>
<td>Short-term Liabilities/Total Assets</td>
</tr>
<tr>
<td>X16</td>
<td>Short-term Assets/Sales</td>
</tr>
<tr>
<td>X17</td>
<td>Operating Profit/Interest Paid</td>
</tr>
<tr>
<td>X18</td>
<td>Inventories/Sales</td>
</tr>
<tr>
<td>X19</td>
<td>Cash flow/Sales</td>
</tr>
<tr>
<td>X20</td>
<td>Net income/Sales</td>
</tr>
<tr>
<td>X21</td>
<td>Long-term Liabilities/Total Assets</td>
</tr>
<tr>
<td>X22</td>
<td>Cash and cash Equivalents/Short-term Liabilities</td>
</tr>
<tr>
<td>X23</td>
<td>Cash flow/Short-term Liabilities</td>
</tr>
<tr>
<td>X24</td>
<td>Working Capital/Sales</td>
</tr>
<tr>
<td>X25</td>
<td>Current ratio</td>
</tr>
<tr>
<td>X26</td>
<td>(Short-term Assets - Inventory)/Short-term Liabilities</td>
</tr>
<tr>
<td>X27</td>
<td>Return on Assets (profit before tax/total assets)</td>
</tr>
<tr>
<td>X28</td>
<td>Return on equity (profit before tax/equity)</td>
</tr>
<tr>
<td>X29</td>
<td>Equity/Long-term Liabilities</td>
</tr>
<tr>
<td>X30</td>
<td>Equity/(Long-term + Short-term Liabilities) * 100</td>
</tr>
<tr>
<td>X31</td>
<td>Cash flow/Operating Revenues</td>
</tr>
<tr>
<td>X32</td>
<td>Turnover of Net Assets</td>
</tr>
<tr>
<td>X33</td>
<td>Interest Paid</td>
</tr>
<tr>
<td>X34</td>
<td>Gross Profit/Operating Income</td>
</tr>
<tr>
<td>X35</td>
<td>Profit before Tax/Operating Income</td>
</tr>
<tr>
<td>X36</td>
<td>Short-term Assets - Short-term Liabilities</td>
</tr>
<tr>
<td>X37</td>
<td>Working Capital</td>
</tr>
</tbody>
</table>

Source: developed by the authors.

The maximum number of tree levels was 7, which proved to be the optimal choice in this case. The maximum size of the parent node was 100 cases. It means, if the tree node contained more than 100 enterprises, it was further divided into child nodes if the maximum number of tree levels has not been reached yet. If the tree node contained lower than 100 enterprises, was no longer divided into child nodes. The minimum size of the child node was 50 cases. It means, the division of the parent node into two child
nodes would indicate that one of them would contain lower than 50 enterprises, this division would not take place. For selection and detection of splitting variable was used Gini index.

A minimum node reduction was 0.0001, which was used to the branching of each node. The classification ability of the created decision tree was quantified using a classification table. For verification of the model, classification ability was enterprises in database divided into a training and test sample in the ratio of 80%/20%. The training sample was used to derive the discriminatory function, and the test sample was then used to verify the ability of the model to be classified.

Variable X10 ([(long-term+short-term liabilities)/total assets]) was used like the most significant classification (splitting) variable. The breakpoint was 1. In this case, enterprises with the value of X10 lower or equal than 1, were classified as non-prosperous. The classification error at this tree point was 0.5%. This branch of the tree had another splitting variable X28 (return on equity), with breakpoint -0.47830. Enterprises with the value of X28 lower or equal than -0.47830 were definitely classified as non-prosperous. In turn, enterprises with the value of X28 higher than 1 were classified as prosperous with further splitting according to variable X28 again. In this tree branch, are two another splitting variables X30 (equity/(long-term + short-term liabilities)*100) and X36 (short-term assets-short-term liabilities).

Splitting criterion X10>1 also created the second branch of a tree which leads to the classification of an enterprise into a group of non-prosperous enterprises. This classification branch is easier. The classification error in this tree node is also 0.5%. Subsequently, the splitting variable was X4 (net income/equity). The breakpoint is -0.66663. Enterprises with the value of X4 higher than -0.66663 were definitively classified as non-prosperous. On the other side, enterprises with the value of X4 lower or equal than -0.66663 were classified as prosperous with further splitting according to same variable X4 in breakpoint -1.74686.

The classification ability of the created decision tree was expressed using a classification table. The classification table provides the number of correctly and incorrectly ranked enterprises into a group of prosperous or non-prosperous enterprises.

The model correctly classifies 93.2% of enterprises in training sample as well as in the test sample. The correct classification of non-prosperous enterprises is 83.5% in the training sample as well as in the test sample. Thus, type I error is 16.5% in this model, which can be considered a sufficient classification model ability.

The ROC curve also expresses the classification ability of the created model. The area of AUC under the curve is 0.933. From this point of view, the created prediction model can be considered sufficiently suitable for classification of enterprises to prosperous or non-prosperous in emerging markets.

The decision tree technique has several advantages as well as disadvantages. The first advantage is that the results are expressed as simple and, therefore, very easy to understand explicit conditions for input variables. The resulting model could be easily programmed. Then, the classification of new cases is very fast. A great advantage of this method is also that the technique of the decision tree creation is non-parametric, which means that independent variables may not have a certain probability distribution. These variables may even be collinear. If one of the variables is not suitable for the classification, it will simply not be used in a tree. Besides, the dependence of the output variable on the input variables may be non-linear or even nonmonotonic. The preparation and data selection phases are considerably easier compared to other methods. It is because decision trees have no problems with extreme cases that are isolated in small nodes without a relevant effect on the overall classification.

Figure 1. A comprehensive prediction model based on a decision tree model for a training sample of emerging markets

Source: developed by the authors.
Figure 2. A comprehensive prediction model based on a decision tree model for a test sample of emerging markets

Source: developed by the authors.

Most tree classifiers, such as CART and C5.0, can handle any type of variable (continuous, discrete, or qualitative). They also can reasonably cope with missing values, either by creating a new category of a variable with missing values or by using a substitute variable. This research was also addressed by Siekelova et al. (2017), Valaskova et al. (2019) and Kliestikova et al. (2017). On the other hand, the disadvantage of decision trees is the «divide and conquer» rule used to create the tree itself. Variables that appear in the first division conditions have much greater weight (they separate...

a lot more cases) and affect the impact of other variables in the tree. Even a small change within these variables may, but may not, lead to a large change of the tree itself, and, therefore, its prediction capability.

Table 4. Classification table of a prediction model based on a decision tree model for emerging markets

<table>
<thead>
<tr>
<th>Classification</th>
<th>Sample</th>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1,38,7044</td>
<td>54,188</td>
</tr>
<tr>
<td>Training</td>
<td>1</td>
<td>73,726</td>
<td>372,655</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>77.4%</td>
<td>22.6%</td>
<td>93.2%</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>347,258</td>
<td>13,537</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>77.5%</td>
<td>22.5%</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

Source: developed by the authors.

![Figure 3. ROC curve of a prediction model based on a decision tree model for emerging markets](image)

Source: developed by the authors.

As a further disadvantage, we can mention the fact that trees detect the actual impact of the variable gradually and not simultaneously. The selection of the splitting criterion is not checked as feedback. It may be a problem, as some trees prefer multi-category variables. This lack of robustness of the decision tree can sometimes be unacceptable and can be overcome by resampling, but that means losing its simplicity. The resulting decision tree models form rectangular areas in the space of variables, which do not have to correspond to the division into cases. The cases that do not have a rectangular distribution are especially difficult to classify. Another disadvantage is the fact that a relatively wide data sample is needed to create a tree because otherwise it is threatened by relatively quick overfitting. This research was also addressed by Balcerzak et al. (2018), Salaga et al. (2015) and Huxley and Mouwafac (2018).

Conclusion. Issue of enterprise financial distress represents the actual and interdisciplinary topic for the economic community. Nowadays, the problem with the failure of business entities is a very actual topic in the market economy. One of the biggest business risks is credit risk, which relates to secondary insolvency of the business entity. The bankrupt is thus one of the major externalities of today's modern economies, which cannot be avoided even with every effort. Economists are able to evaluate and predict credit risk and financial health of business entities due to ex-ante financial analysis - through prediction models. Prediction models can early predict the probability of failure of the business entity. The paper
assumption is a possibility of application of a decision tree algorithm for the creation of the prediction model, which can be used in economy practice. Paper main aim is to create a comprehensive prediction model of enterprise financial distress based on decision trees, under the conditions of emerging markets.

The decision tree algorithm is one of the most intuitive methods and the most commonly used data mining method. Moreover, that approach provides explicit classification rules. These rules are very simple and easy to interpret, leading to a quick evaluation of the results. The decision tree algorithm could be used as a prediction model for enterprise bankrupt. Paper main aim was to create a comprehensive prediction model of enterprise financial distress based on decision trees, under the conditions of emerging markets. Emerging markets included 17 countries: Slovak Republic, Czech Republic, Poland, Hungary, Romania, Bulgaria, Lithuania, Latvia, Estonia, Slovenia, Croatia, Serbia, Russia, Ukraine, Belarus, Montenegro and Macedonia. However, data from Belarus was removed from the database due to a large amount of missing data. For creating prediction models, it was necessary to create the database for which were used the financial and statistical indicators from the Amadeus database for 2015 (for all independent variables) and 2016 (for the dependent variable). Used data overall contained 2,359,731 enterprises from emerging markets; divided into prosperous enterprises (1,802,027) and non-prosperous enterprises (557,704). They presented the dependent variable. Input variables for model represented 24 financial indicators, 3 dummy variables (country, size of enterprise and NACE classification) and countries GDP data. They represented independent variables. The decision tree was constructed in IBM SPSS Statistics 24 using algorithm CART with the top-down method. The tree has always been pruned to prevent it overfitting. The maximum tree levels were 7. The minimum size of «parent» nodes was 100 cases, and «child» nodes 50 cases. The Gini index was used to selection and calculation of the optimal splitting variable. Model classification ability was quantified using a classification table. The 80% of enterprises represented training sample and 20% test sample, for model creation. The model correctly classified 93.2% of enterprises from both the training and test sample. Correctly classification of non-prosperous enterprises was 83.5% in both samples. The created prediction model can be considered sufficiently suitable for classifying enterprises in emerging markets. Paper findings bring new possibilities for bankrupt enterprise prediction with emphasis on the conditions in Emerging markets. Research aim and purpose have been fulfilled


Funding. The research leading to these results has received funding from the project titled «Creation of new paradigms of financial management at the threshold of the 21st century in conditions of the Slovak Republic» in the frame of the program of Slovak Scientific Grant Agency VEGA under the grant agreement number VEGA 1/0428/17.

References


Agarwal, V., & Taffler, R. J. (2007). Twenty-five years of the Taffler z-score model: Does it really have predictive ability?. Accounting and Business Research, 37(4), 285-300. [Google Scholar] [CrossRef]


Marketing and Management of Innovations, 2020, Issue 3

http://mmi.fem.sumdu.edu.ua/en

289


http://mmi.fem.sumdu.edu.ua/en


Іванна Подгорська, Ph.D., Харківський університет, Словачка Республіка
Яромир Врбка, Ph.D., Інститут технологій і бізнесу в Чеських Будейовиця, Чеська Республіка
Георгій Лазароу, Ph.D., доцент, Інститут комунікативної науки, США. Університет Спіру Харет, Румунія
Марія Ковацова, Ph.D., Жилінський університет, Словачка Республіка

Інновації у фінансовому менеджменті: модель рекурсивного прогнозу на основі алгоритму дерева рішень

У статті працювали ефективність інновацій у фінансовому менеджменті підприємств. Авторами зазначено, що банкрутство є одним із наслідків не ефективного фінансового менеджменту компанії. При цьому менеджмент компанії їх час розроблення систем фінансових інструментів повинен враховувати навички зовнішні та внутрішні ризики діяльності команії. У статті розглянуто алгоритм дерева рішень, як один з найбільш інтуїтивних методів збору даних, який можна використовувати для прогнозування ймовірності настання фінансових ризиків. Результати систематизації наукового дослідження з даним напрямом засвідчили, що експерти використовують дерево рішень у єкспертній інноваційному інструменту фінансового менеджменту. Статтю присвячено аналізу можливостей зосередження алгоритму дерева рішень для створення моделі прогнозування банкрутства компанії. Авторами розроблено комплексну модель прогнозування фінансової байдики підприємств на основі дерева рішень з використанням алгоритму CART. Для формування моделі прогнозування використано дані 2 359 731 підприємств (30% від загальної суми) із 17 країн, в тому числі Словаччини, Чехії, Польщі, Угорщини, Румунії, Румунії, Болгарії, Сербії, Росії, України, Венеції, Чорногорії та Македонії. У цьому зазначено, що із 802 027 компаній, що працювали та 557 704 – не працювали. Статистичні дані згенеровані з бази даних Амстердама. Вихідні змінні моделі обрано 24 фінансові показники, з допоміжні змінні та дані про ВВП країн у 2015 та 2016 роках. Вибірку для розроблення моделі сформовано на основі данів 80% підприємств, тобто як дані 20% – для її тестування. Отримана модель видалала класифікування 93,2% підприємств як процвітаючі та 83,5% – не процвітаючі. Авторами наголошено, що запропонована модель прогнозування є пріоритетною для класифікації підприємств за ринком ефективності їх діяльності на зростаючих ринках.

Ключові слова: модель прогнозу, дерево рішень, зростаючі ринки.

Manuscript received: 10.06.2020
© The author(s) 2020. This article is published with open access at Sumy State University