SYNTHESIS OF PREDICTIVE MODELS FOR START-UP COMPANIES

B. Yankov

Department of Software Engineering at Sofia University "St. Kliment Ohridski",
tel. +359887940844, e-mail: boian_iankov@abv.bg

Abstract: A quantitative research is performed to improve the accuracy of a previously derived model for predicting the success of Bulgarian start-up companies. The preceding research stages include an overview and analysis of older success prediction models, creation of a new abstract success prediction model, a venture creation process model, a qualitative and a quantitative research. The dataset is extended with 31 more cases and is analyzed in higher detail using the IBM SPSS Modeler software. The previously derived models are compared to the new ones in terms of accuracy of the success prediction and the predicting variables.

Key words: entrepreneurship, start-up companies, success prediction, business model, new ventures, SPSS Modeler

INTRODUCTION

New venture success prediction is used to assess companies for venture capital funding and on business plan competitions. The accuracy of the intuitive prediction can be improved by using a prediction model. The goal of this research is to improve a previously created model for predicting the success of Bulgarian start-up companies.

For the goals of the research, start-up companies are considered small to medium companies that were started 0 to 5 years ago. Company success is defined as the company survival and growth. Companies which have increased in size and survived during the last five year are considered successful, the ones that survived but did not grow in size are neither successful, nor unsuccessful and the ones that stopped their operation are unsuccessful.

After an analysis of existing success prediction models [1], a pattern has been identified. It was introduces by Sandberg [2] in his model from 1986 which can be illustrated with the formula:

$$NVP = f(E, IS, BS) \quad (1)$$

Where NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy.

By analyzing the requirements for a new venture prediction model and the venture creation process model [1, 2, 3], an extended new venture success prediction model [4] based on Sandberg [2] is proposed. The model is presented with the formula:

$$NVP = f(E, IS, BS, R) \quad (2)$$

R is a new variable representing the available resources and the other variables are similar to the ones from Sandberg's model. Each of the main categories in the company success prediction model is decomposed into subcategories [4] as shown in Fig. 1 - derived by the author.

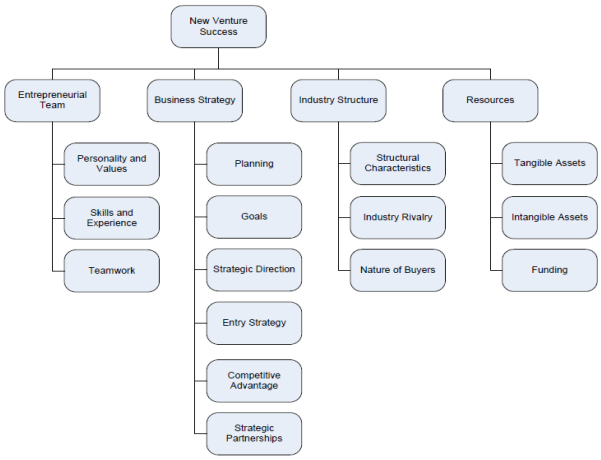

Figure 1. New venture success prediction model proposed by the author

The new venture success prediction model has been revised with the help of a qualitative research [4] by conducting in-depth interviews with durations of 0:30 to 2:30 hours with 5 non-representative cases – owners of young Bulgarian companies. Then an initial qualitative research on a dataset of 107 companies has identified the success factors and has derived a success prediction decision tree.

THE UPDATED QUANTITATIVE RESEARCH DATA

The current quantitative research of the new venture success prediction model uses a dataset of 137 companies which is 30 more than the data used for the previous research. The data collection has taken 5 months to gather responses from owners and managers of Bulgarian firms of various sizes.

The goal of the research is to analyze the success of SME (small and medium enterprises). For that reason, the companies that are just starting (and we have no information about their success) as well as the big companies have been eliminated from the analyzed sample. Some of the fields in the dataset are free text inputs and the information contained in them has been analyzed and categorized in higher detail compared to the previous analysis.

THE ANALYSIS IN IBM SPSS MODELER

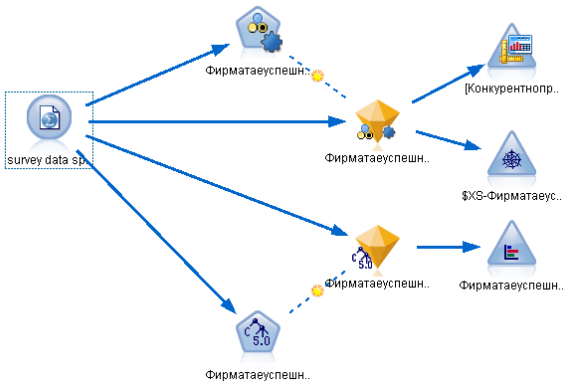

Figure 2. The updated model in IBM SPSS Modeler

The updated model in SPSS Modeler is shown in Fig. 2. The data is loaded using a Statistics File node (with the circle shape on the left) which reads data from a .sav file format used by SPSS Statistics. The Auto Classifier node (the pentagon shape on the top) then estimates and compares models for the selected target using a number of different methods. The pentagon shaped node on the bottom is a customized C5.0 modeler used to test various options and their results. The selected target is the company success and the other variables are used as inputs (Fig. 3).

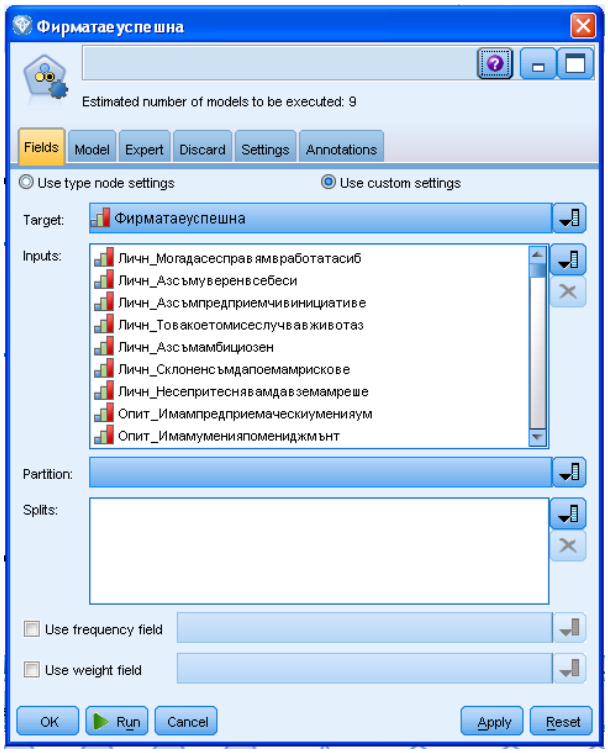

Figure 3. Selecting fields for the Auto Classifier node - inputs and targets

The Auto Classifier node ranks candidate models based on the target and saves the best 3 models for further analysis. The resulting models are visible in a container called a model nugget as shown in Fig. 4.

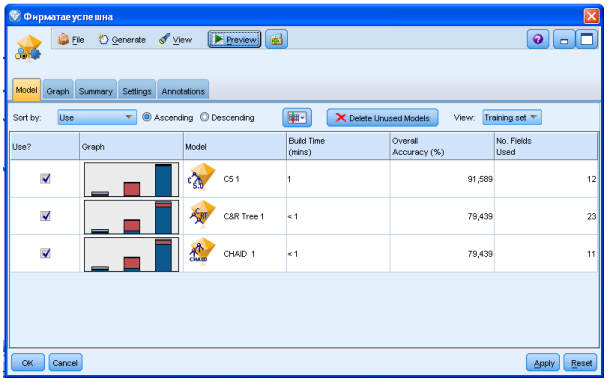

Figure 4. The resulting models

The C5.1 model is ranked first because it has the highest overall accuracy of 91,59% and uses 12 fields. The accuracy of the C5.1 model from the previous analysis was 91,86%. The C&R and CHAID models have an improved accuracy of 79,44% compared to the previously derived models, which had an accuracy of 75,58%.
By examining the suggested tree in the C5.1 model, a problem becomes evident: on level 2 of the tree (Fig. 5). The tree suggests that senior entrepreneurs (older than 60 years) have lower chance of success. The split is not balanced and this conclusion is not reliable because it is based on only 4 cases in Node 25. The advantage of younger entrepreneurs is possible but must be tested with mode data.

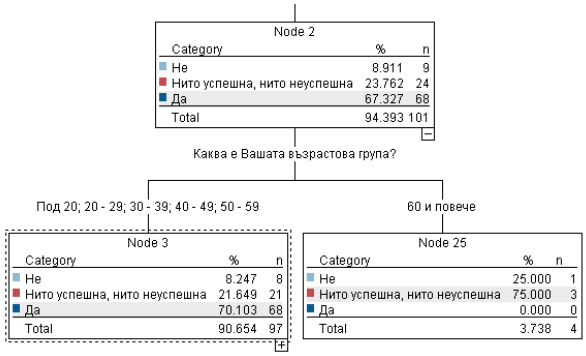

Figure 5: The Tree – Branch 2: Age Group of the Founder

The problem with the imbalanced split is probably related to the method settings. The algorithm by default tries to produce the most accurate tree possible [5]. In some instances, this can lead to overfitting and loss of generality, which can result in poor performance when the model is applied to new data. An attempt was made to solve this issue with C5.0 analysis with manual settings. When the minimum records per child branch were increased to 5, the model accuracy dropped. Better results have been achieved by excluding the variable “Age Group of the Founder” from the analysis which was selected as the best option. The resulting models are shown in Fig. 6. The C5.1 model has the same accuracy without the age variable.

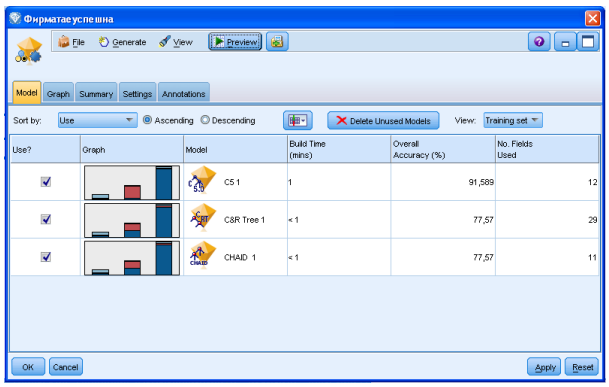

Figure 6. The resulting models – revised

The derived models: C5.1, C&R Tree and CHAID, are based on the rule induction technique [5]. All of them produce a decision tree based on a set of rules that describe distinct segments within the data in relation to the target field, which in our case is the company success. The models' outputs openly present the reasoning for each rule and can therefore be used to understand the decision-making process that drives a particular outcome. The tree (Fig. 7) starts with the most important success predictors and splits the cases into groups (represented by nodes) depending on the responses. The process continues until the case reaches an end (leaf) node which indicates the predicted value of the target – the company success.

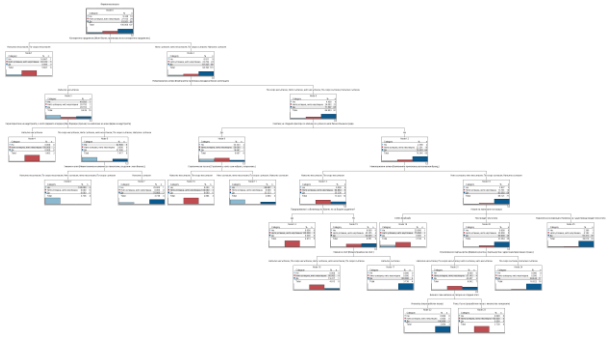
Figure 7: The decision tree

The first level of the tree has two branches based on the presence of a competitive advantage variable as shown in Fig. 8. This variable is the most important predictor of the company success of the analyzed dataset of companies. Companies that have a clear competitive advantage (Node 2) tend to be more successful than companies that do not (Node 1). This success factor remains the most important one, compared to results from the previous analysis.

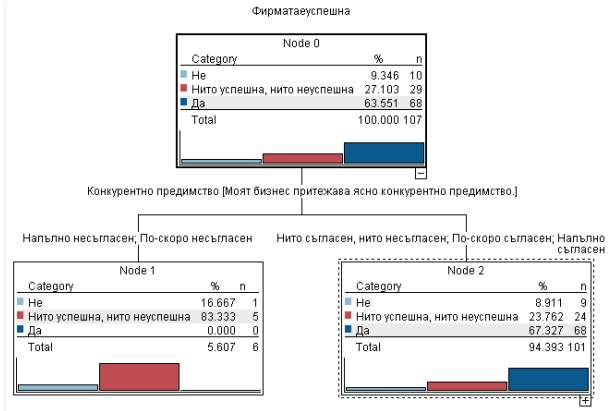

Figure 8: The tree – 1st level: Presence of a Competitive Advantage

Node 2 has two child nodes based on the next success predictor – the intangible asset – goodwill (Fig. 9). The companies from the analyzed sample that do not have established business reputation are less successful (Node 3) than the others (Node 8). The current analysis suggests that the established business reputation is more important than the key success factor – environment, which was more important in the previous tree [4]. Node 3 has a few cases and we will not analyze its children.

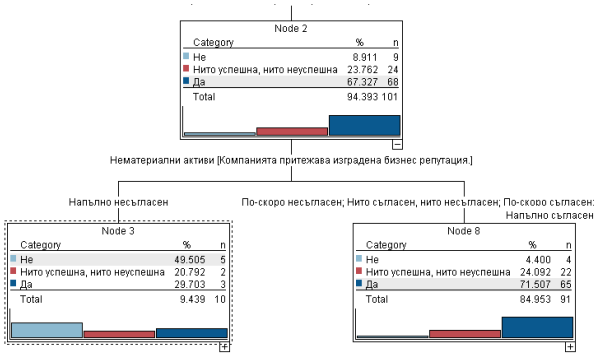

Figure 9: The Tree – Branch 2: Intangible Asset – Goodwill

Node 8 has two child nodes based on the next success predictor – the environment as a key success factor (Fig. 10). Those companies from the analyzed sample that consider the environment as a key success factor are less successful (Node 9) than the others (Node 12). Node 9 has a few cases and we will not analyze its children.

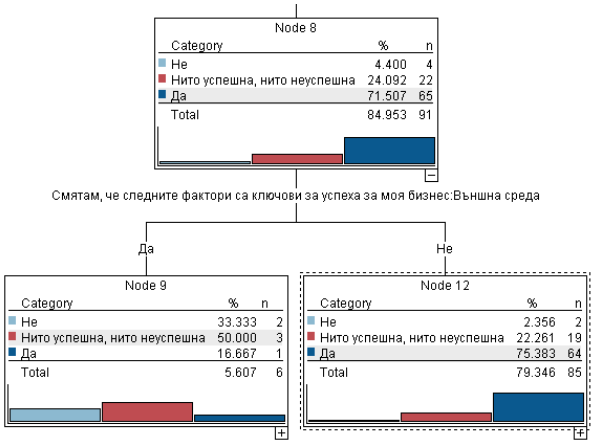

Figure 10: The tree – branch 2.2: key success factors – environment

Node 12 has two child nodes based on the next success predictor – the intangible asset – recognizable brand (Fig. 11). The companies from the analyzed sample that have a recognizable brand (Node 18) are more successful than the ones that do not (Node 13). The recognizable brand was also an important factor in the previous tree [4]. Node 13 has a few cases and we will not analyze its children.

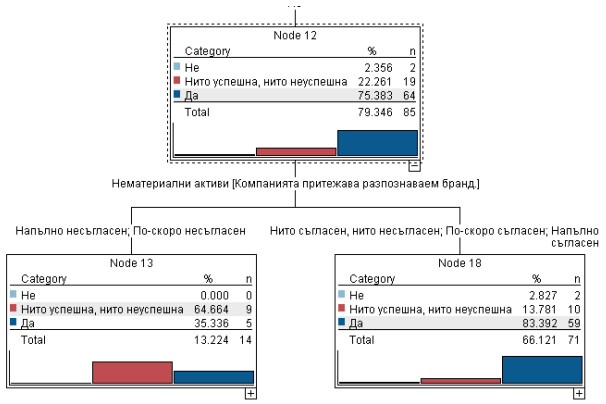

Figure 11: The tree – branch 2.2.2: intangible asset – recognizable brand

Node 18 has three child nodes based on the next success predictor – the type of entry of the company (Fig. 12). The companies whose type of entry is a parallel competition or a modification of an existing product or service (Node 25) are the most successful. Companies that develop a new product or service (Node 20) are less successful. The franchise type of entry (Node 19) is the least successful but this is based on a few cases. The type of entry was a success predictor in the previous tree [4] but the current data features more cases which allows for analysis of different types of entry.

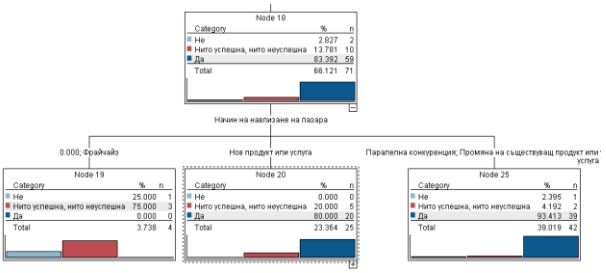

Figure 12: The tree – branch 2.2.2.2: type of entry

Further analysis of the tree reveals more success predictors but they are based on fewer cases which would decrease the reliability of the conclusions that we could make.

CONCLUSION

The analysis of the new data shows little differences compared to the previous success prediction model. The updated models have a good overall accuracy. The decision tree contains the success predictors: presence of a competitive advantage, intangible asset – goodwill, environment as a key success factor, intangible asset – recognizable brand, type of entry of the company, partnerships with 3rd parties, barriers for entry in the industry, order of entry on the market, management experience, technical skills related to the business, entrepreneurial parents, aggressiveness of strategy. The age of the founder could possibly be a success predictor but this must be tested with more data.

Future plans for improvement include increasing the accuracy of the model, deriving models with other methods, making comparisons. The model will be used for developing a software for predicting the success of start-up companies.

ACKNOWLEDGEMENT

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014). This work was supported by IBS Bulgaria, IBM Premier Business Partner.

REFERENCES

1. Yankov, B. (2012), Overview of Success Prediction Models for New Ventures, International Conference Automatics and Informatics'12, ISSN 1313-1850, pp 13-16.
2. Sandberg, W. R. (1986). New venture performance: The role of strategy and industry structure. Lexington, MA: Lexington Books
3. Carland, J.W. and Carland, J.A. (2000), A New Venture Creation Model, Western Carolina University.
4. Yankov, B., Haralampiev, K., Ruskov P.: Start-up Companies Predictive Models Analysis, Vanguard Scientific Instruments in Management '2013 (VSIM:13), ISSN 1314-0582, (2013)
5. CRISP-DM 1.0, Step-by-step data mining guide, Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000