Start-up Companies Predictive Models Analysis

Boyan Yankov, Kaloyan Haralampiev, Petko Ruskov

Abstract: *A quantitative research is performed to derive a model for predicting the success of Bulgarian start-up companies. The preceding research stages included an overview and analysis of older success prediction models, a new abstract success prediction model, a venture creation process model and a qualitative research. The abstract success prediction model is extended with measurable variables which are included in a survey. The survey is currently in progress with 105 responses by owners and managers of Bulgarian companies. The current dataset was analyzed using the IBM SPSS Modeler software which automatically tests different models and suggests the best performing ones. The best derived model is a decision tree model that predicts the success of the start-up companies from the dataset with 91,86% probability using 11 variables.*

Keywords: *entrepreneurship, start-up companies, success prediction, business model, new ventures, SPSS Modeler*

JEL: *M13, C38*

Introduction

Start-up companies create job opportunities and are important for the Bulgarian economy. The efficiency of the new venture creation process can be improved by increasing the returns and minimizing the risks with the help of a model for predicting the success of start-up companies. Success prediction models and software tools for Bulgarian start-ups would be useful to entrepreneurs, business owners, business incubators, university start-up centers, business consultants, venture capitalists and investors.

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, not unsuccessful and the ones that stopped their operation are unsuccessful.

After an analysis of 42 success prediction models [1] a pattern has been identified. The pattern was introduces by Sandberg [2] in his model from 1986 as shown of Fig. 1.

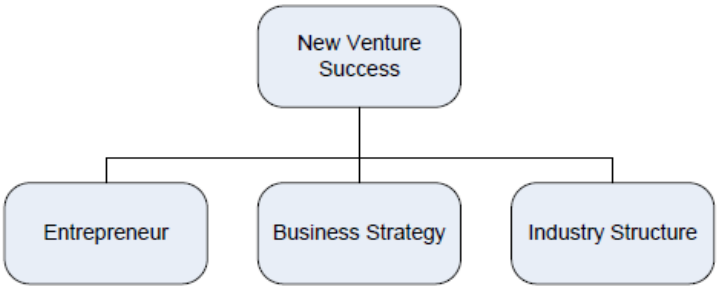

Figure 1. New venture success model by Sandberg

The model by Sandberg can be illustrated with the formula:

NVP = f(E, IS, BS) *(1)*

Where NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy. Later studies [3] based on Sandberg include other factors: the entrepreneurial team, the interaction of the company strategy, the industry structure and the available resources.

The Proposed NVP Prediction Model

By analyzing the requirements for a new venture prediction model and the venture creation process model [1, 3, 4], an extended new venture success prediction model [5] based on Sandberg [2] is proposed. The model is presented with the formula:

NVP = f(E, IS, BS, R) *(2)*

R is a new variable representing the available resources. The other variables are similar to the ones from Sandberg's model: NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy. Each of the main categories in the company success prediction model is decomposed into subcategories [5] as shown in Fig. 2 - derived by the author.

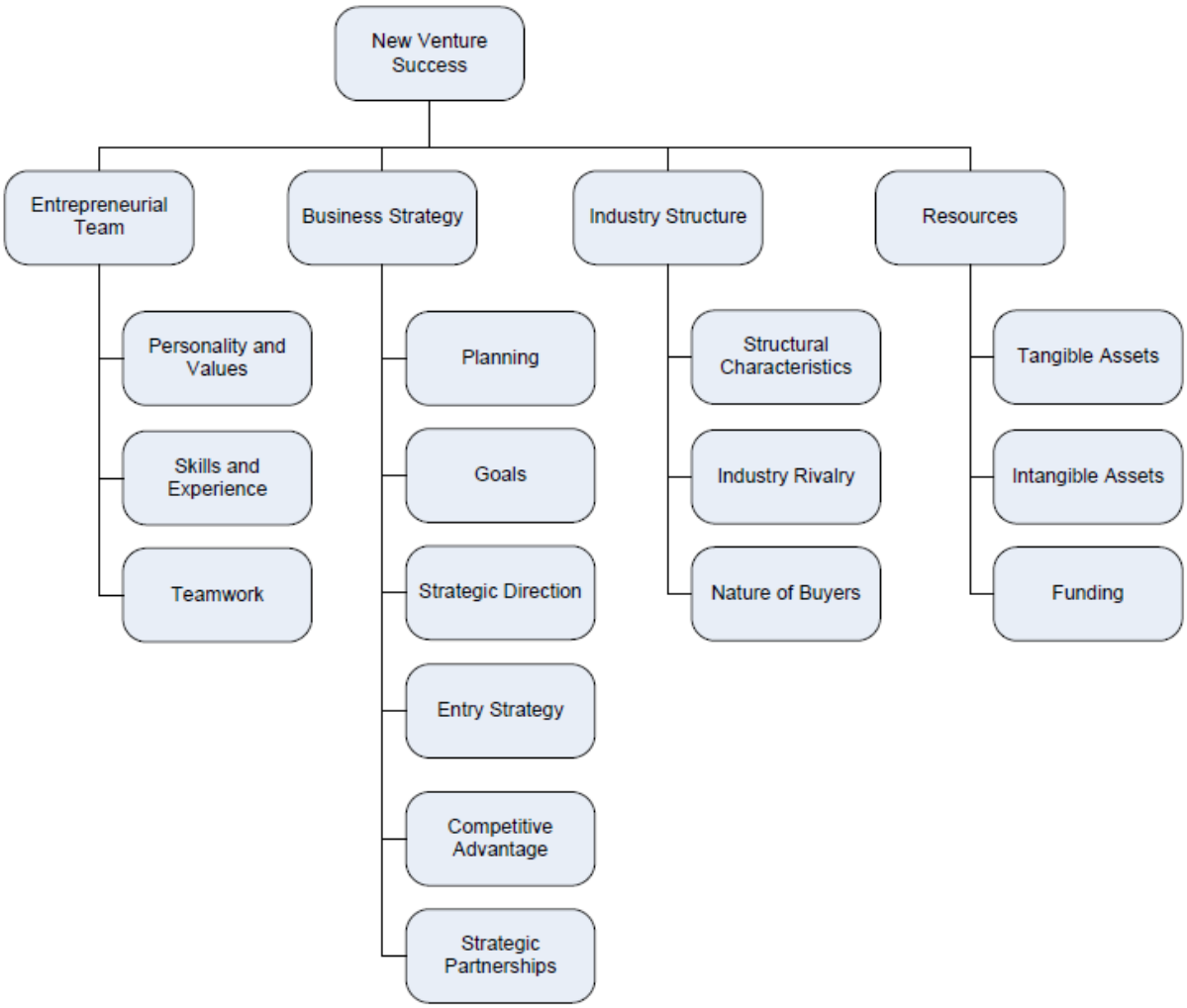

Figure 2. New venture success prediction model proposed by the author

The new venture success prediction model has been revised with the help of a qualitative research [5] by conducting in-depth interviews with durations of 0:30 to 2:30 hours with 5 non-representative cases – owners of young Bulgarian companies.

The Quantitative Research Data

A dataset has been collected for a quantitative research of the new venture success prediction model. The data has been collected with the help of a survey consisting of 107 variables based on the categories in the prediction model. The data collection has taken 3 months and is still in progress. The current sample contains data about 105 Bulgarian Companies of various ages and sizes (88% micro, 9% small, 2% medium, 1% big). The data has been collected using various sources: networks of business contacts, online social networks, business communities, start-up communities, targeted email marketing and targeted CPC advertising.

The CRISP-DM Methodology

The study is based on the CRISP-DM [6] methodology which is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance, as shown in Fig. 3.

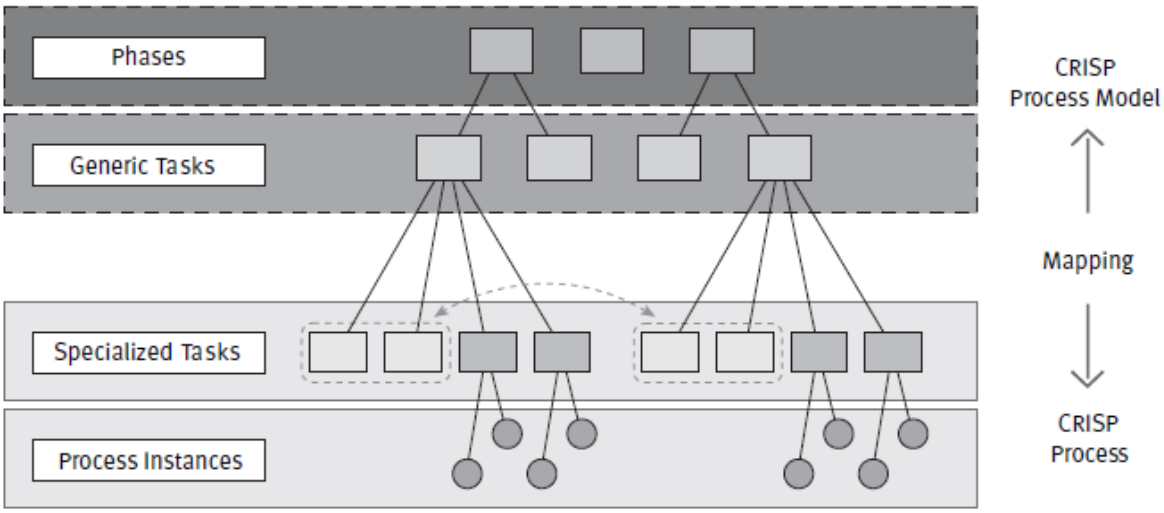

Figure 3. Four level breakdown of the CRISP-DM methodology

The CRISP-DM Reference Model

The life cycle of a data mining project contains the phases of a project, their tasks and relationships. The life cycle consists of six phases as shown in Fig. 4. The sequence of the phases is flexible allowing moving back and forth between different phases. The outcome of each phase determines which phase or task has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

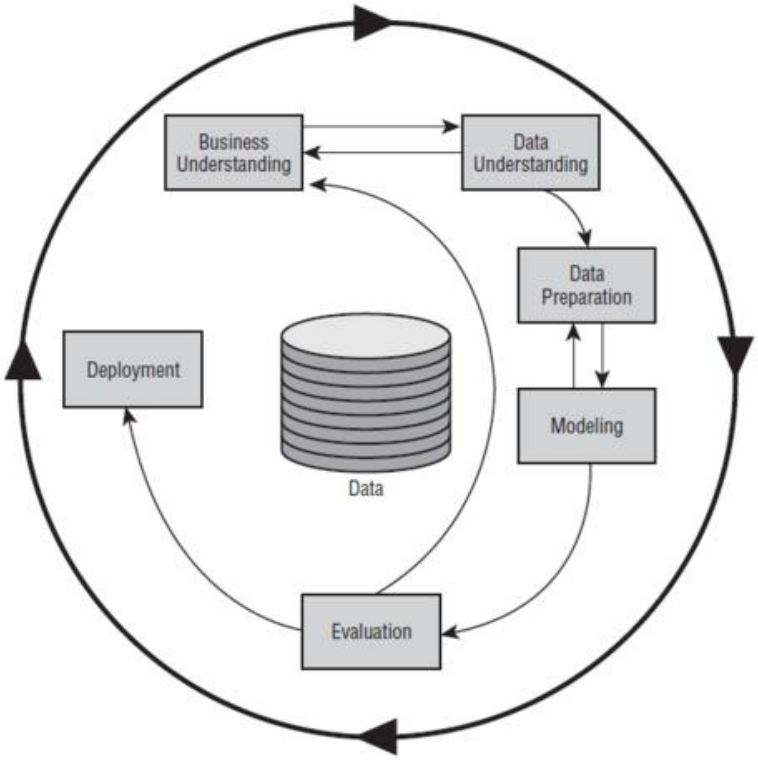

Figure 4. The life cycle of a data mining project

The outer circle in Fig. 4 symbolizes the cyclical nature of the data mining process which does not end when a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more-focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones.

The analysis in SPSS Modeler

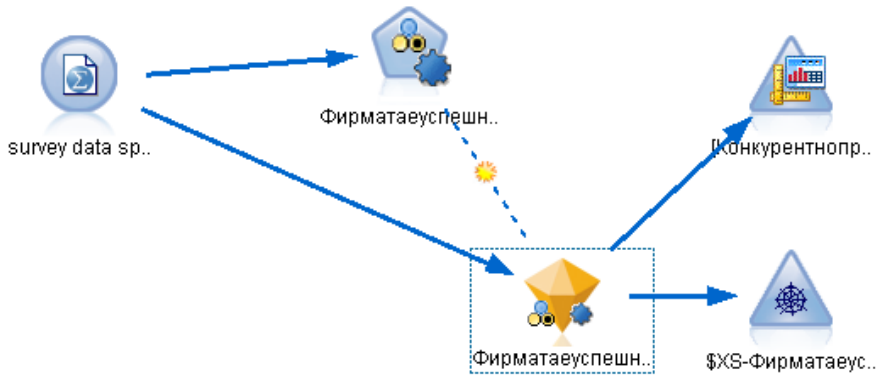

Figure 5. The model in SPSS Modeler

The model in SPSS Modeler is shown in Fig. 5. The data is loaded using a Statistics File node (with the circle shape) which reads data from a .sav file format

used by SPSS Statistics. The Auto Classifier node (the pentagon shape) then estimates and compares models for the selected target using a number of different methods. The selected target is the company success and the other variables are used as inputs as shown in Fig. 6.

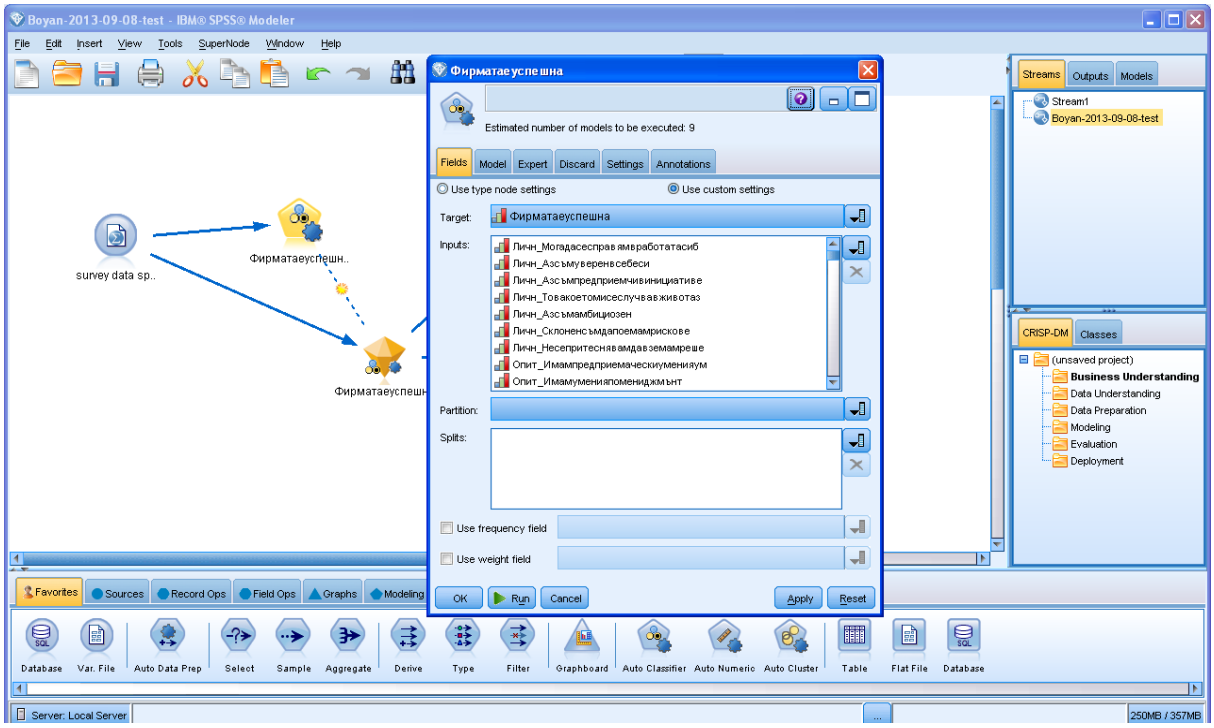

Figure 6. Selecting fields for the Auto Classifier node - inputs and targets

The Auto Classifier node explores every possible combination of options, ranks each candidate model 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, the main purpose of which is scoring data to generate predictions or to allow further analysis of the model properties. Opening a model nugget on the screen enables you to see various details about the model as shown in Fig. 7.

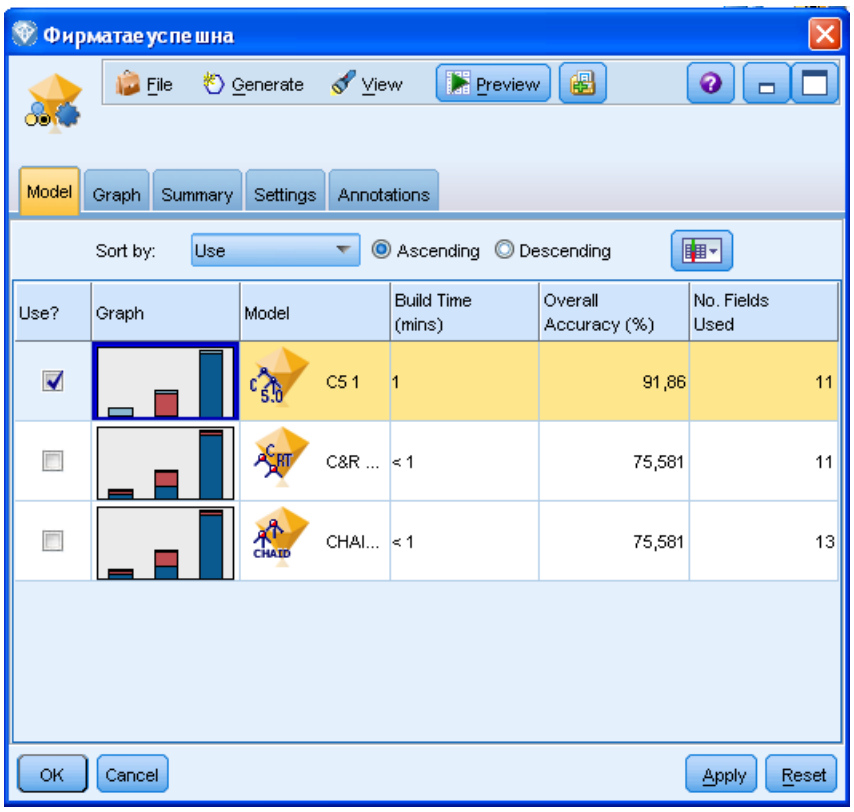

Figure 7. Resulting models

Comparison of Models

The C5.1 model is ranked first because it has the highest overall accuracy of 91,86% and uses 11 fields. The second suggested model is C&R Tree (classification and regression tree) with 75,58% accuracy and 11 fields. The third mode is CHAID which also has 75,58% accuracy, but uses more fields.

The models proposed by the tool are all based on the rule induction technique: C5.1, C&R Tree and CHAID. All of them derive a decision tree or 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. Another advantage of rule induction methods over other methods, such as neural networks, is that the process automatically eliminates any fields that are not important in making decisions.

The models C5.1, C&R Tree, CHAID, QUEST and Decision List use the rule induction algorithm. To explain how it works, let us think about making a decision to buy a house. The most important factor may be cost - ability to afford the property. The second may be what type of property you are looking for - a house or a condo.

The next consideration may be the location of the property, etc. Combining these questions produces a decision tree as shown in Fig. 8.

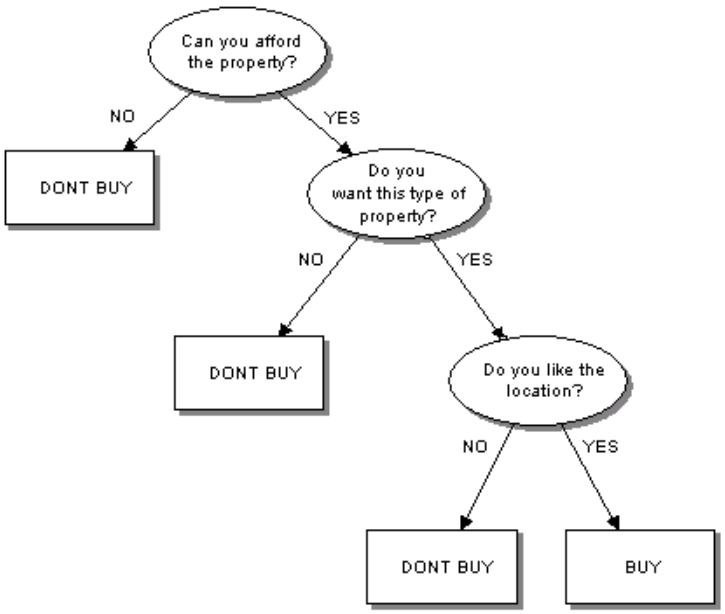

Figure 8: Graphical Representation of a Decision Tree

The rule induction or decision tree methods produce the decision tree by culling through a set of predictors by successively splitting a dataset into subgroups on the basis of the relationships between predictors and the target field.

Algorithms Differences

The C5.1, C&R Tree, CHAID models differ by the type of output they produce. C5.1 only uses categorical target fields while C&R Tree and CHAID support both categorical and continuous targets. In our case the target is categorical. When the models recursively loop through the data, they produce a different type of split into subgroups (on a predictor). C&R Tree support only binary (two group) splits, while CHAID and C5.1 support splits with more than two subgroups. The algorithms differ in the criterion used to drive the splitting. For C5.1 an information theory measure is used - the information gain ratio. When C&R Tree predicts a categorical field, a dispersion measure (the Gini coefficient by default) is used. CHAID uses a chi-square test.

All algorithms allow for missing values for the predictor fields, although they use different methods. C5.1 uses a fractioning method, which passes a fractional part of a record down each branch of the tree from a node that is split ~~is~~ based on a field for which the record is missing. C&R Tree uses substitute prediction fields, where

needed, to advance a record with missing values through the tree during training. CHAID makes the missing values a separate category and allows them to be used in tree building.

In our data there are a small number of fields with missing values. It is not advisable to completely remove these fields from the data mining process. Moreover, not answering a question about the business may have a meaning. The missing values (blanks) in the dataset were replaced with legal values.

Recognized Cases by the Selected Model

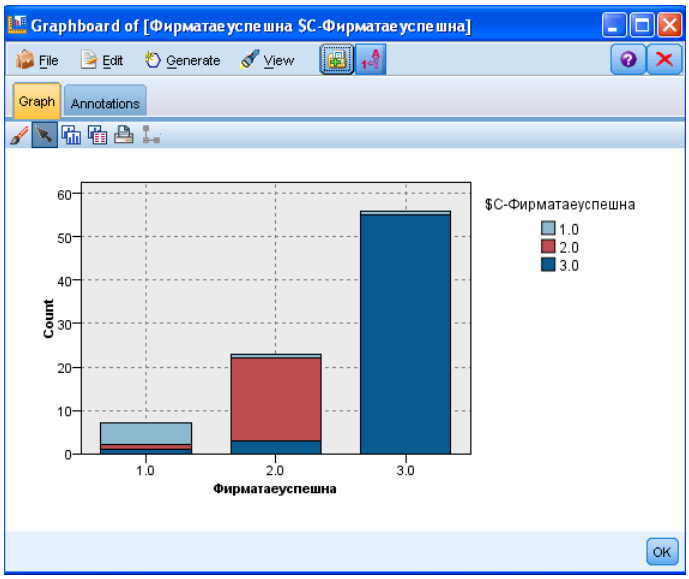

Figure 9: Recognized Cases by the Selected Model

The C5.1 model predictions are compared to the actual values of the company success variable and the results are shown in Fig. 9. Unsuccessful companies are presented with the number 1, neither successful nor unsuccessful are presented with 2 and successful are presented with 3. The largest segment of all bars presents the successfully recognized cases – 91,86% of all.

The Decision Tree

The C5.1 model generates a decision tree, as shown in Fig. 10. The tree 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.

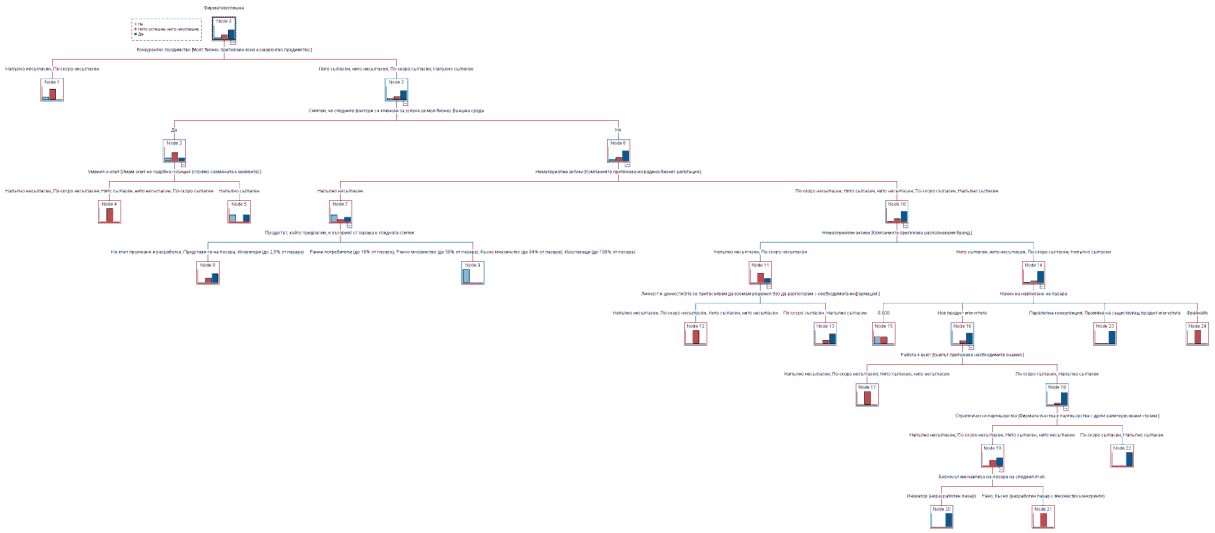

Figure 10. The decision tree generated by the model

The first level of the tree has two branches based on the “presence of competitive advantage” variable as shown in Fig. 11. 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 1) tend to be more successful that companies that do not (Node 2).

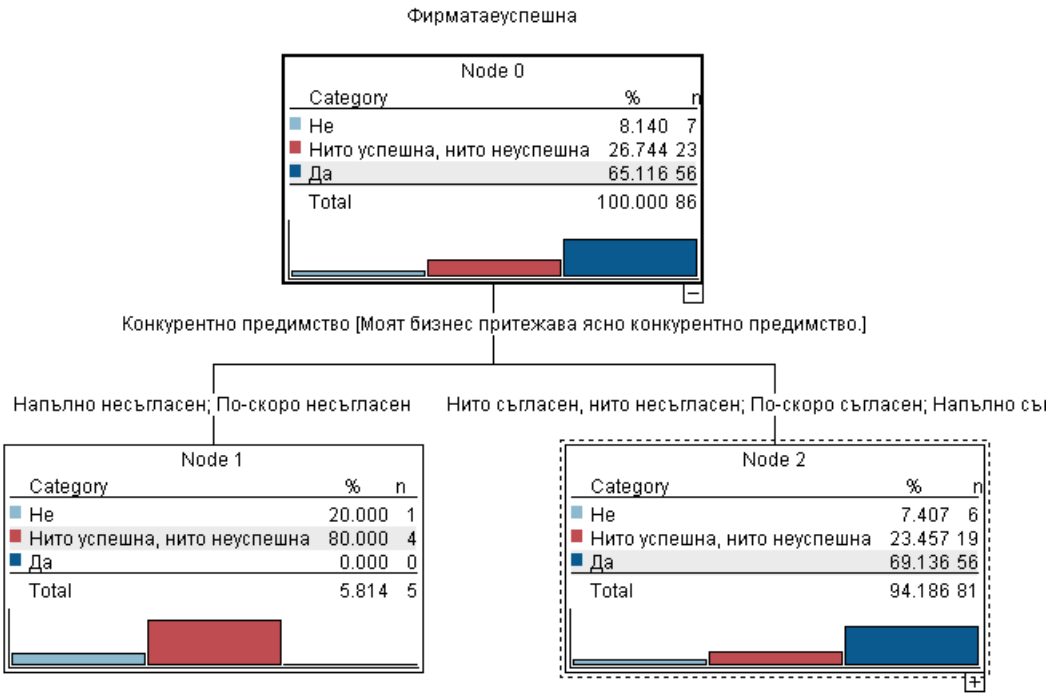

Figure 11. The Tree – 1st level: Competitive Advantage

Node 2 has two child nodes based on the next success predictor – the environment as a key success factor (Fig. 12). Those companies from the analyzed

sample that consider the environment as a key success factor are less successful (Node 3) than the others (Node 6). Node 3 has a few cases and we will not analyze its children.

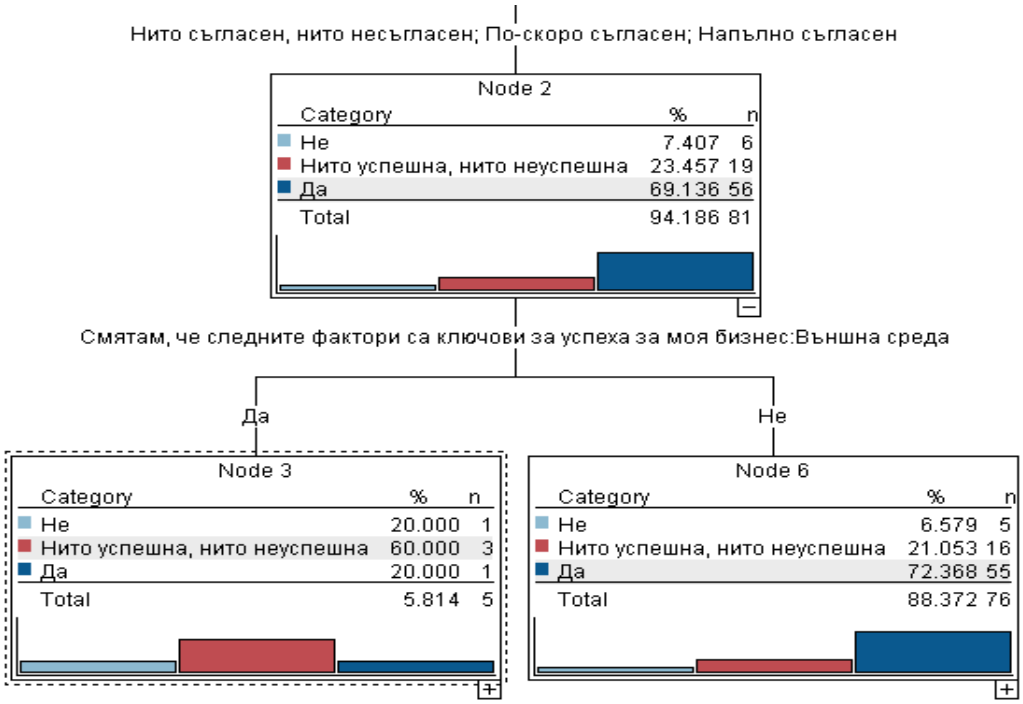

Figure 12: The Tree – Branch 2: Key Success Factors – Environment

Node 6 has two child nodes based on the next success predictor – the intangible asset – goodwill (Fig. 13). The companies from the analyzed sample that do not have established business reputation are less successful (Node 7) than the others (Node 10). Node 7 has a few cases and we will not analyze its children.

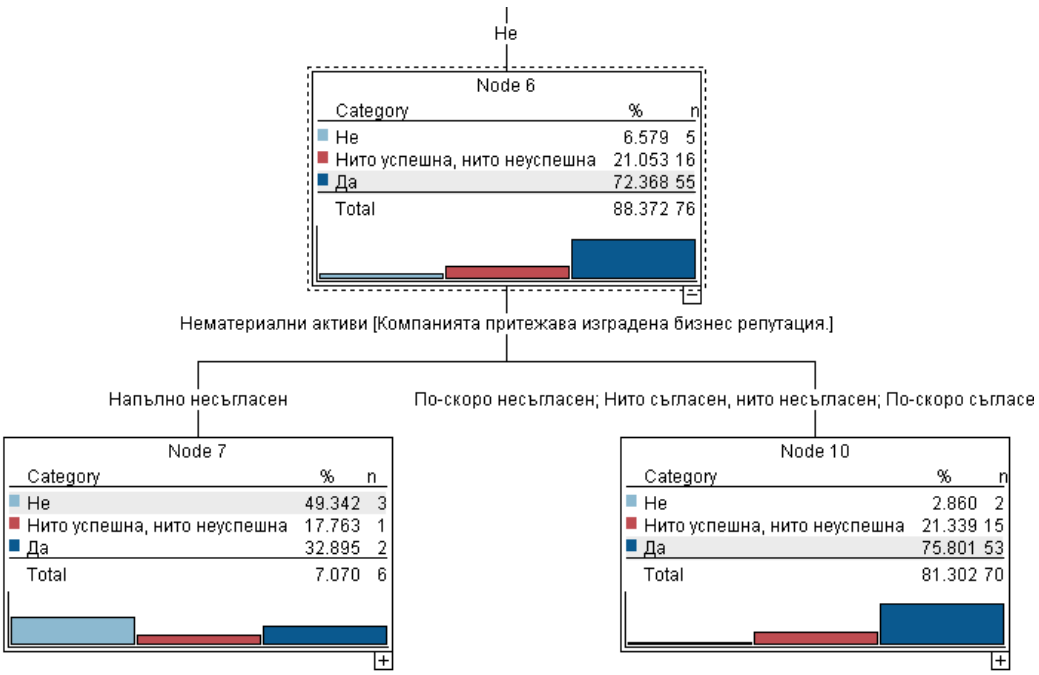

Figure 13: The Tree – Branch 2.2: Intangible Asset – Goodwill

Node 10 has two child nodes based on the next success predictor – the intangible asset – recognizable brand (Fig. 14). The companies from the analyzed sample that have a recognizable brand are more successful (Node 14) than the ones that do not (Node 11).

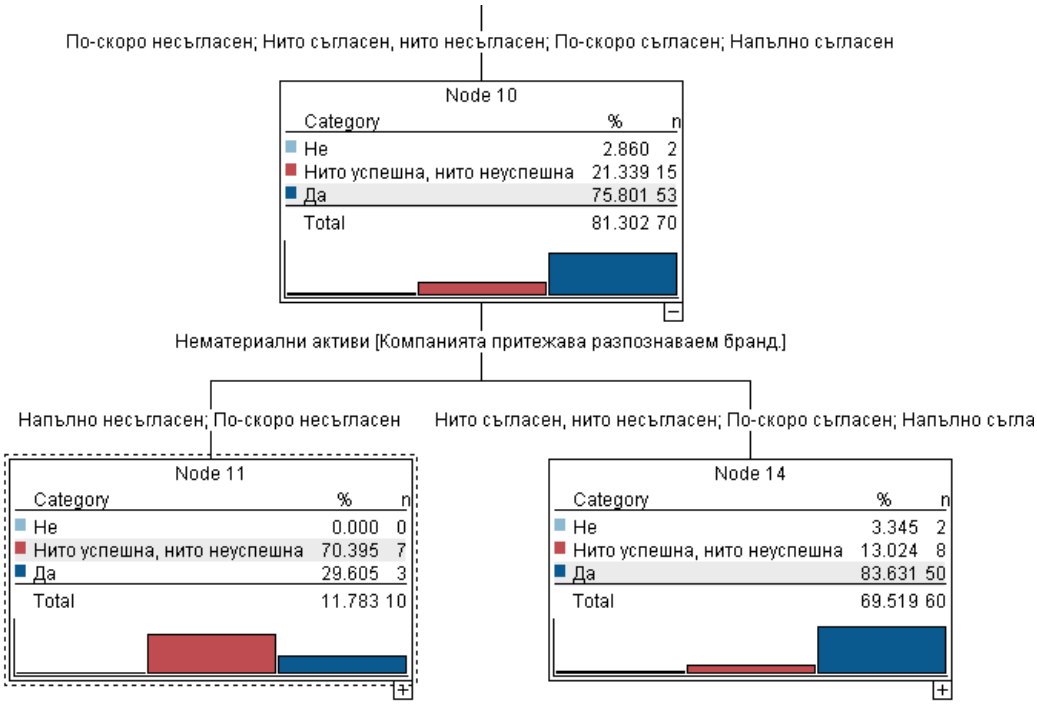

Figure 14: The Tree – Branch 2.2.2: Intangible Asset – Recognizable Brand

If we continue the analysis of the decision tree, more success factors become evident: the type of entry of the company (new product, improved product, parallel competition or franchise), the ability to take management decisions without the necessary information, the team knowledge and the strategic partnerships with 3rd parties.

Relation of Competitive Advantage to Company Success

Fig. 15 shows the relation of the most important variable – presence of a competitive advantage (the dots on the lower left corner) to the predicted value of the target – the company success (the dots on the upper right corner). Relations are indicated with lines where the thickness indicates the strength of the relationship. The presence of a competitive advantage (4 and 5) connects to company success (3).

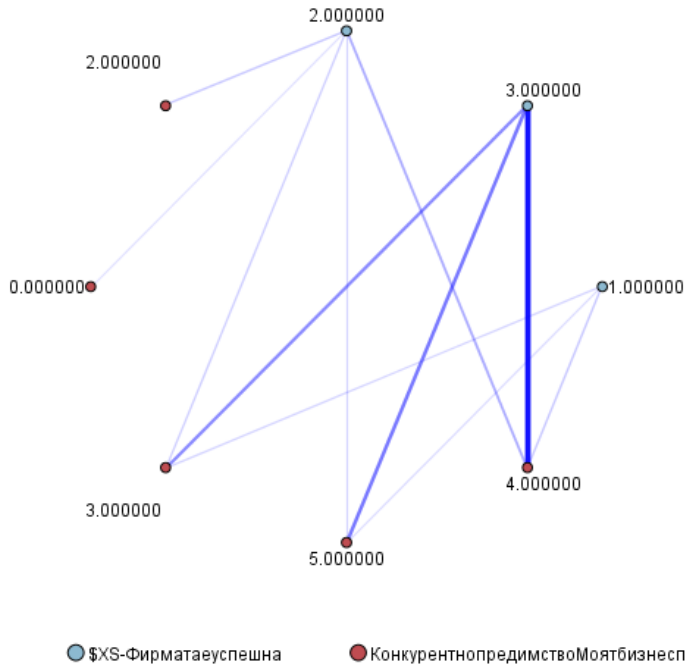

Figure 15: Relation of Competitive Advantage to Company Success

Conclusion

The models derived by the analysis with the SPSS Modeler have a good overall accuracy. The best model is C5.1 with 91,86% accuracy and 11 variables. It produces a logical decision tree containing the variables: presence of a competitive advantage, environment as a key success factor, intangible asset – goodwill, intangible asset –recognizable brand, type of entry of the company, ability to take management decisions without the necessary information, team knowledge, strategic partnerships with 3rd parties, etc. However the accuracy of the analysis still needs improvement as it is based on responses from only 105 companies.

Future plans for improvement include increasing the accuracy of the model, deriving models with other methods and making comparisons. Insights from the research could be used to help Bulgarian start-up companies succeed.

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. Chrisman, J., Bauerschmidt, A. and Hofer, C. (1998), The Determinants of New Venture Performance: An Extended Model
4. Carland, J.W. and Carland, J.A. (2000), A New Venture Creation Model, Western Carolina University.
5. Yankov, B. (2013), A Model for Predicting the Success of New Ventures, V International Scientific Conference "e-Governance", 2013, ISSN 1313-8774, pp.128-135
6. 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