

Models and Tools for Technology Start-Up Companies Success Analysis

Boyan Yankov^{*}, Petko Ruskov^{**},
Kaloyan Haralampiev^{***}

Summary:

The designs presented in the article are fastened in the authors' years-long research on entrepreneurship and business model innovations. A quantitative research was performed to derive a model for predicting the success of Bulgarian start-up companies. The authors started this research with in-depth inquiries of start-up companies in Bulgaria. Under our guidance, several research analysts investigated each start-up using approximately 100 questions. The preceding research stages included an overview and an analysis of existing success prediction models, a new abstract success prediction model, a venture creation process model and a qualitative research. The abstract success prediction model was extended with measurable variables with the help of a quantitative research of Bulgarian entrepreneurs. The current dataset of companies has been enriched with more cases and has been analyzed using data mining software: IBM SPSS Modeler, which automatically tests different models and suggests the best performing ones and also with the open source product Weka. The best derived model is a classification tree that correctly predicts the success of technology start-ups from the dataset

in 83,76% of the test cases. The analysis revealed the answers to challenges and questions that start-up companies face and implemented a model that was deployed into an information system for start-ups success prediction. The developed information system will help to predict the success of start-ups. The software will evolve iteratively, and by involving more companies to use it, will grow its database.

Key words: technology start-ups, success prediction, prediction tool, start-ups analysis

JEL Classification: M13, C38

1. Introduction

Start-up companies are founded in a fast changing and fuzzy environment and their owners make significant efforts to create innovative products and services, and to improve the business processes, which is often costly and time-consuming. The creation of a start-up requires a huge upfront investment in many aspects: from initial research and development to dedicated resources, new models, processes and equipment (Price 2004, Blank 2007, Amit 2012, Cosentino 2014). Future returns of these investments are always uncertain and unsafe. More start-up companies are now projecting towards innovative business models as an alternative or a complement to a product or services innovation (Hamel 2012, Majamäki 2013, Tsoleva 2014). The

^{*} PhD student, Sofia University, e-mail: boian_yankov@abv.bg

^{**} Professor, Sofia University, e-mail: petkor@fmi.uni-sofia.bg

^{***} Professor, Sofia University, e-mail: k_haralampiev@hotmail.com

creation of start-up companies has also the social impact of creating job opportunities and stimulating the world economic growth.

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 (Huang 2010). 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.

2. Research design

The research is based on the CRISP-DM data mining methodology that reduces the time required for large data mining projects, improves efficiency and helps identify new successful patterns (CRISP-DM 1.0, 2000).

The methodology consists of the following iterative steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

The authors' definition of a successful startup is identified by conducting interviews with small business owners, entrepreneurs and entrepreneurship educators. The selected start-up companies are small to medium ventures that were started 0 to 5 years ago, and a *successful start-up is defined as one that has survived during the last five years of operation and has also increased its size.*

The success prediction model was introduced by Sandberg (1986). The model can be illustrated with the formula (1) where: *NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy.*

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

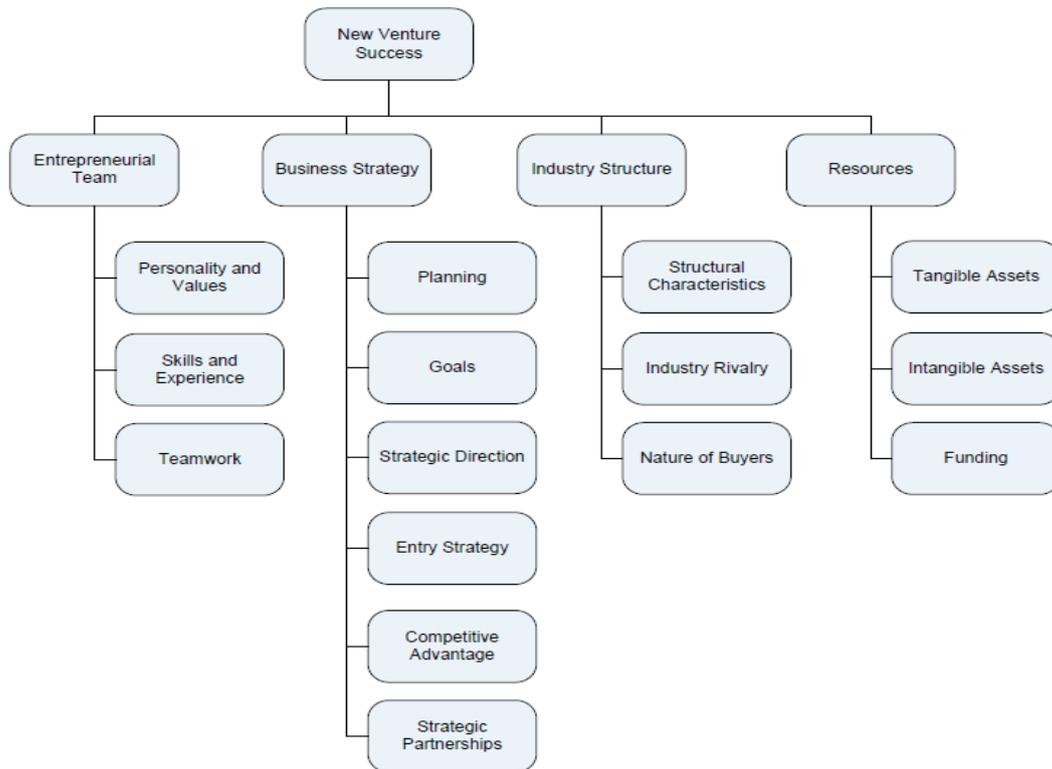


Fig. 1. New venture success prediction model proposed by Yankov

By analyzing the requirements for a new venture prediction model and the venture creation process model (Carland 2000, Yankov 2012), an extended new venture success prediction model (Yankov 2013) based on Sandberg is proposed. The model extends Sandberg's by also including the *available resources (R)* and is presented with the formula:

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

After an analysis of existing success prediction models, experiments and practical applications of them, a pattern of successful start-up has been established. Each of the main categories in the start-up company success prediction model is decomposed into subcategories as shown in Fig. 1.

The new venture success prediction model was revised and improved with the help of a qualitative research (Yankov, Haralampiev, Ruskov 2013) by conducting in-depth interviews with durations of 0:30 to 2:30 hours with 5 non-representative cases – owners of young Bulgarian companies.

3. Technology Start-ups Quantitative Research and Discussion

The current quantitative research of technology start-up companies uses an upgraded dataset of 142 companies which

is bigger than the data used for the previous research iterations. The questionnaire is based on the new venture success prediction model proposed by Yankov. The data collection has taken 12 months to gather responses from owners and managers of Bulgarian firms of various sizes and industries. The goal of the research is to analyze the success of SME (small and medium-sized 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 quantitative research was performed using the previously used data mining software - IBM SPSS Modeler and also with another open source data mining product - Weka. The modes derived using the two products are then compared in terms of their accuracy to select the best prediction model.

3.1. Creating prediction models with Weka

Weka is a set of Java algorithms for machine learning, applicable in data mining

Table 1. The resulting models when using Weka for classification

Algorithm used to create the model	Model Type	Accuracy (no cross validation)	Accuracy (cross validation)
J48	Tree	83.76%	66.67%
J48graft	Tree	83.76%	66.67%
DecisionTable	Rules	68.38%	64.10%
LMT	Tree	66.67%	64.10%
BayesNet	Bayes	65.81%	64.10%
BFTree	Tree	64.10%	64.10%
REPTree	Tree	69.23%	63.25%
SimpleCart	Tree	64.10%	63.25%
FT	Tree	100.00%	60.68%
DecisionStump	Tree	64.95%	60.68%
RandomForest	Tree	98.29%	59.83%
NBTree	Tree	91.45%	54.70%
NaiveBayes	Bayes	80.34%	52.14%
LADTree	Tree	83.76%	48.72%
RandomTree	Tree	99.15%	47.86%

projects (Hall, 2009), (Bouckaert, 2013). Weka is open source (distributed under the GNU General Public License) and contains tools for data pre-processing, classification, regression, clustering and visualization.

The selected target for the model is the company success and the other variables are used as inputs. The resulting classification models are sorted by their accuracy with cross validation and shown in Table 1.

The accuracy of the models without cross validation is calculated by testing the model on the training set. The accuracy with cross validation is more realistic and is calculated by using different data for training and testing the model. The most accurate model is a classification tree derived by applying the J48 algorithm but its accuracy of 66.67% is relatively low.

3.2. Creating prediction models with IBM SPSS Modeler

IBM SPSS Modeler has a visual environment and the sequence of actions

to derive prediction models is presented graphically (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 IBM SPSS Statistics. The selected target is the company success and the other variables are used as inputs. The above sequence on the figure shows the creation of classification models using the automated classifier. The below sequence on the figure shows the model creation using the C5.0 algorithm with manual settings.

The Auto Classifier node estimates and compares models for the selected target using a number of different methods and saves the best 3 models for further analysis. The resulting models (Fig. 3) are visible by clicking on a container called Model Nugget Node.

The C5.1 algorithm has created a classification tree model which ranks first for its highest overall accuracy of 88.03%. The C&R and CHAID algorithms have also created classification tree models but with a lower accuracy of 64.10%. The overall

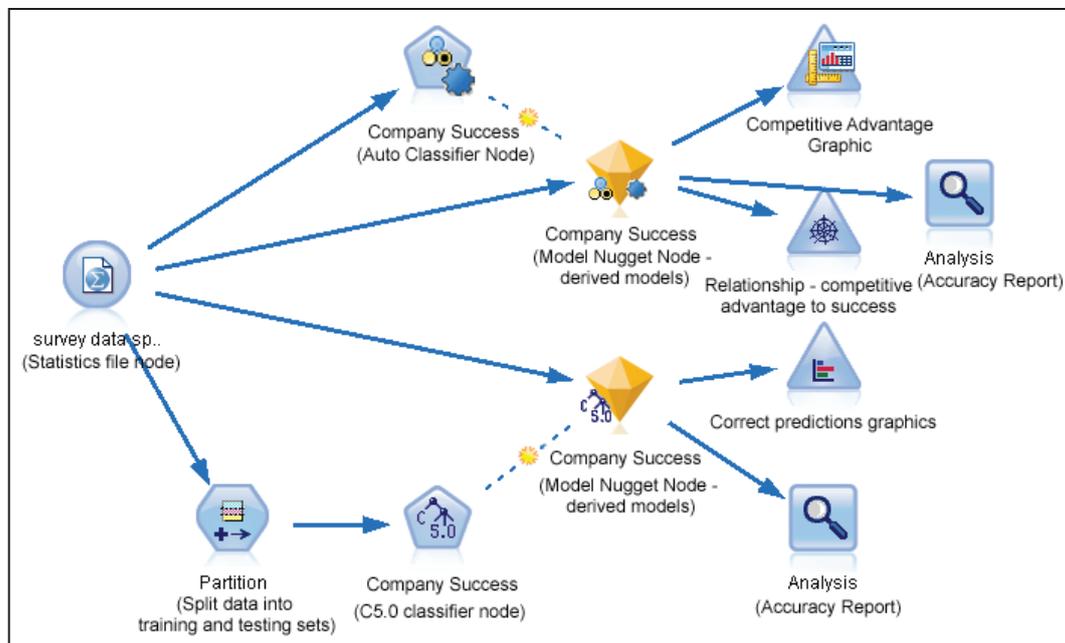


Fig. 2. The modelling sequence in SPSS Modeler

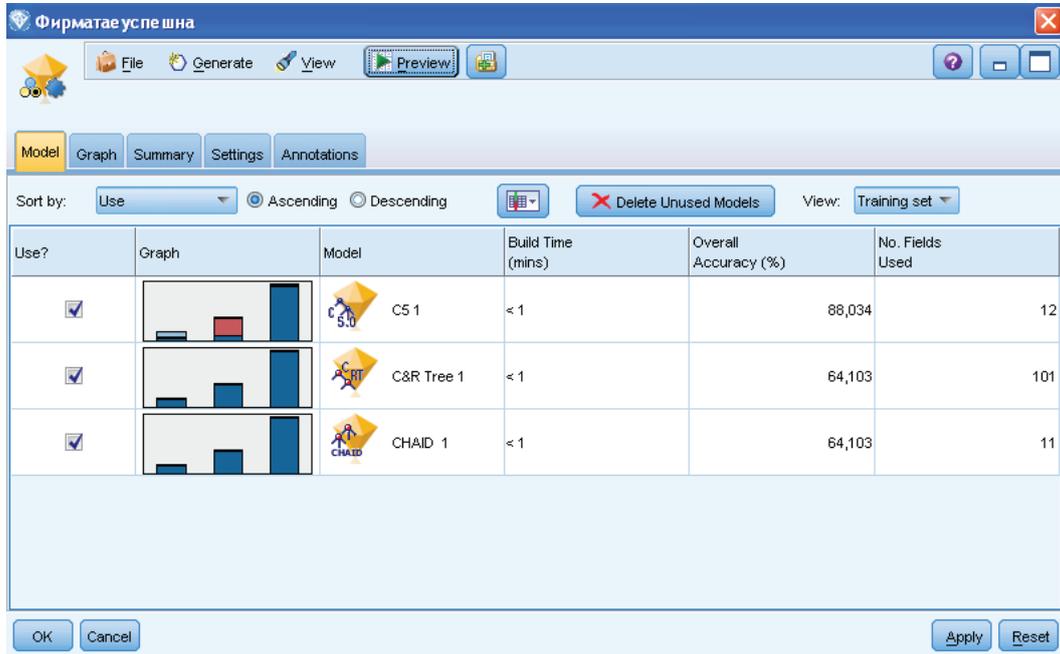


Fig. 3. The resulting models when using the Auto Classifier Node in IBM SPSS Modeler

accuracy is calculated without using cross validation – the same data set is used for training (model generation) and for testing, and therefore the calculated accuracy is an optimistic estimation.

The derived models, when using the algorithms C5.1, C&R Tree and CHAID, are based on the rule induction technique (Chapman et al., 2000). These algorithms produce a classification 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 are actually trees that openly present the reasoning for each rule and can therefore be used to understand the decision-making process that produces a particular outcome. The classification trees start with the most important success predictors and split 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.

The automated classification serves as a base for shortlisting the best performing algorithms and for the creation of improved models with more realistic accuracy. The selected algorithm for the creation of an improved model by using manual settings and cross validation is C5.0 (a manual settings version of C5.1). By applying the algorithm on the training and testing datasets, it produces a model that has 83,76% accuracy and uses 9 variables (fig. 4). The accuracy is more realistically calculated compared to the initial automated

Results for output field CompanySuccess

Comparing \$C-CompanySuccess with CompanySuccess

Correct	98	83,76%
Wrong	19	16,24%
Total	117	

Fig. 4. Statistics for the model, derived using the C5.0 algorithm with cross validation

classification.

By examining the derived success prediction models from IBM SPSS Modeler

and Weka by their accuracy when using cross validation, the best one is a classification tree with 83,76% accuracy, created using the C5.0 classification algorithm in IBM SPSS Modeler. The model contains the following 9 variables that describe the successful company:

- Presence of a clear competitive advantage,

- Dependence on the environment as a key success factor,
- Founder's experience in a similar position,
- Goodwill (established business reputation),
- Type of market entry,
- Recognizable brand,
- Most companies in the industry of the start-up are profitable,

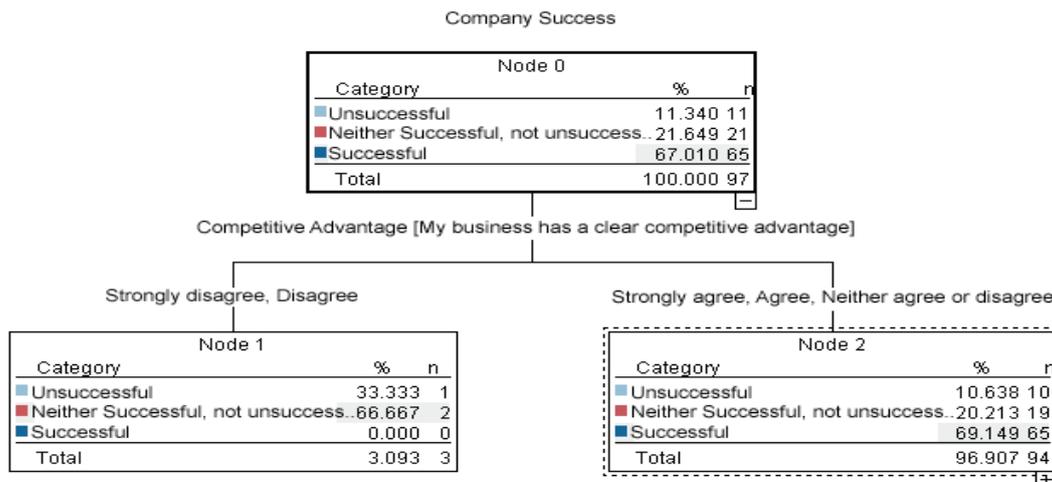


Fig. 5. The classification tree – 1st level: Presence of a clear competitive advantage

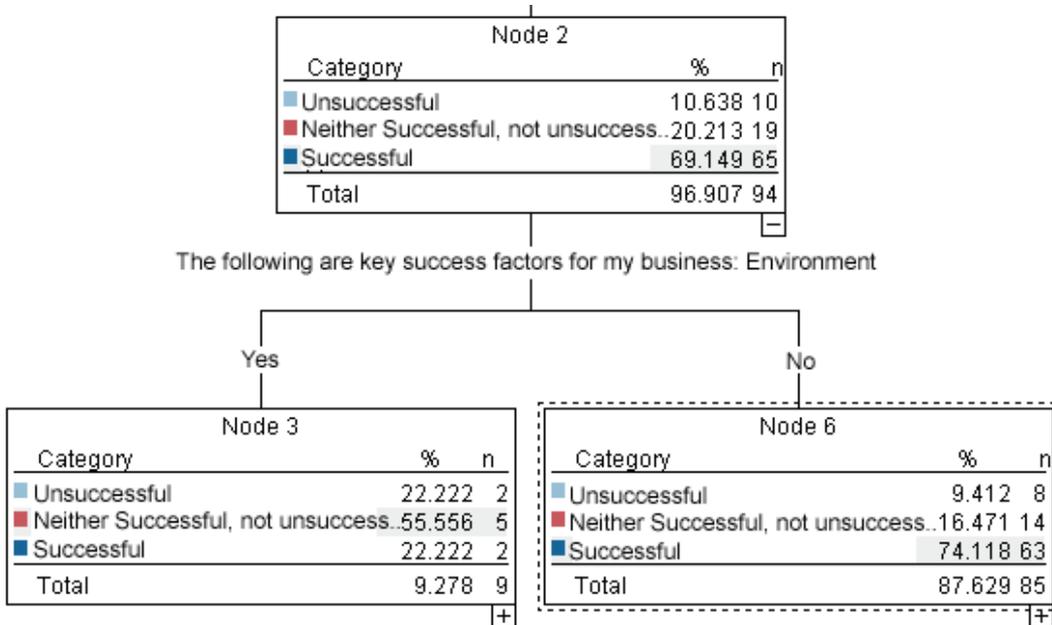


Fig. 6. The classification tree – Branch 2: key success factors – environment

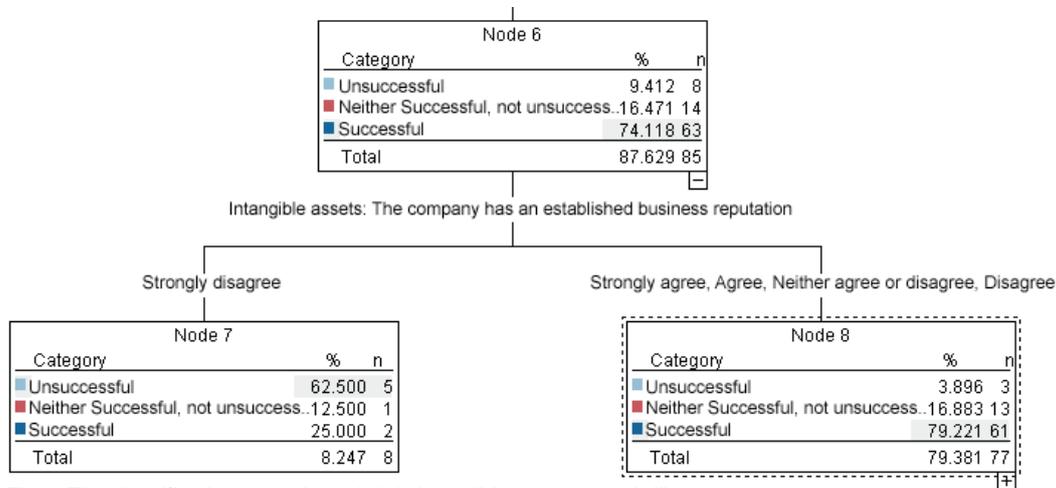


Fig. 7. The classification tree – branch 2.2: intangible asset – goodwill

- The company partners with third parties,
- There is a concentration of customers in the industry of the start-up.

By examining the classification tree, we can determine the importance of the factors. The most important ones are on the top and define the initial splits of the dataset. The first level of the tree has two branches based on the variable presence of a clear competitive advantage as shown in

next success predictor – the environment as a key success factor (Fig. 6). 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 few cases and we will not analyze its child nodes in detail.

Node 6 has two child nodes based on the next success predictor – the intangible asset goodwill (Fig. 7). The companies from the

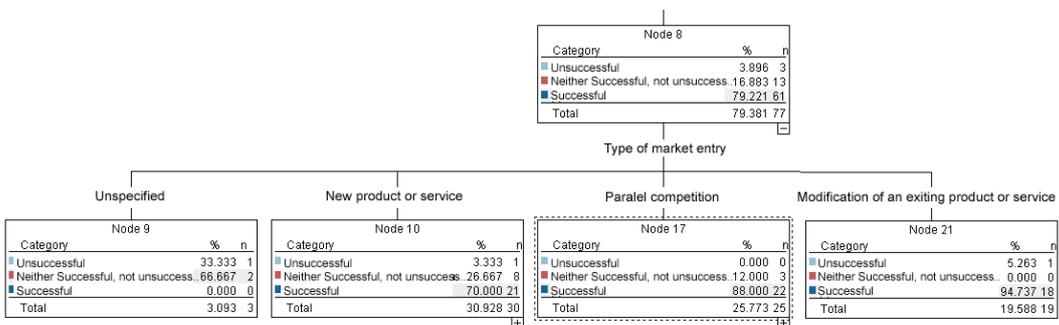


Fig. 8. The classification tree – branch 2.2.2: Type of market entry

Figure 5. This variable is the most important predictor of the success of the analyzed dataset of companies. Start-ups that have a clear competitive advantage (Node 2) tend to be more successful than companies that do not (Node 1).

Node 2 has two child nodes based on the

analyzed sample that do not have established business reputation are less successful (Node 7) than the others (Node 8).

Node 8 has multiple child nodes based on the next success predictor – the type of market entry of the start-up company. The modification of an existing product or

service is the most successful type of entry for the sample of companies (Fig. 8), but the companies that develop a new product or service or use an existing product in parallel competition also tend to be successful.

Further analysis of the tree reveals the other success predictors and their relative importance.

4. Information System for Start-ups Success Prediction

The described practices and processes of modeling the success prediction of start-ups were automatized and the Information

System for Start-ups Success Prediction (I3SP) was designed and developed by analyzing the business model. The I3SP will facilitate start-up companies to exercise the prediction model. Also, by growing the I3SP's database, it will evolve iteratively. It will make modeling easier and will automate the process of creating successful start-ups. The data flow of the engineered I3SP is illustrated in Figure 9. At the beginning, the user of the I3SP starts by filling in the Google survey. The responses data is stored in Google Drive. Then the user requests an analysis and prediction. The

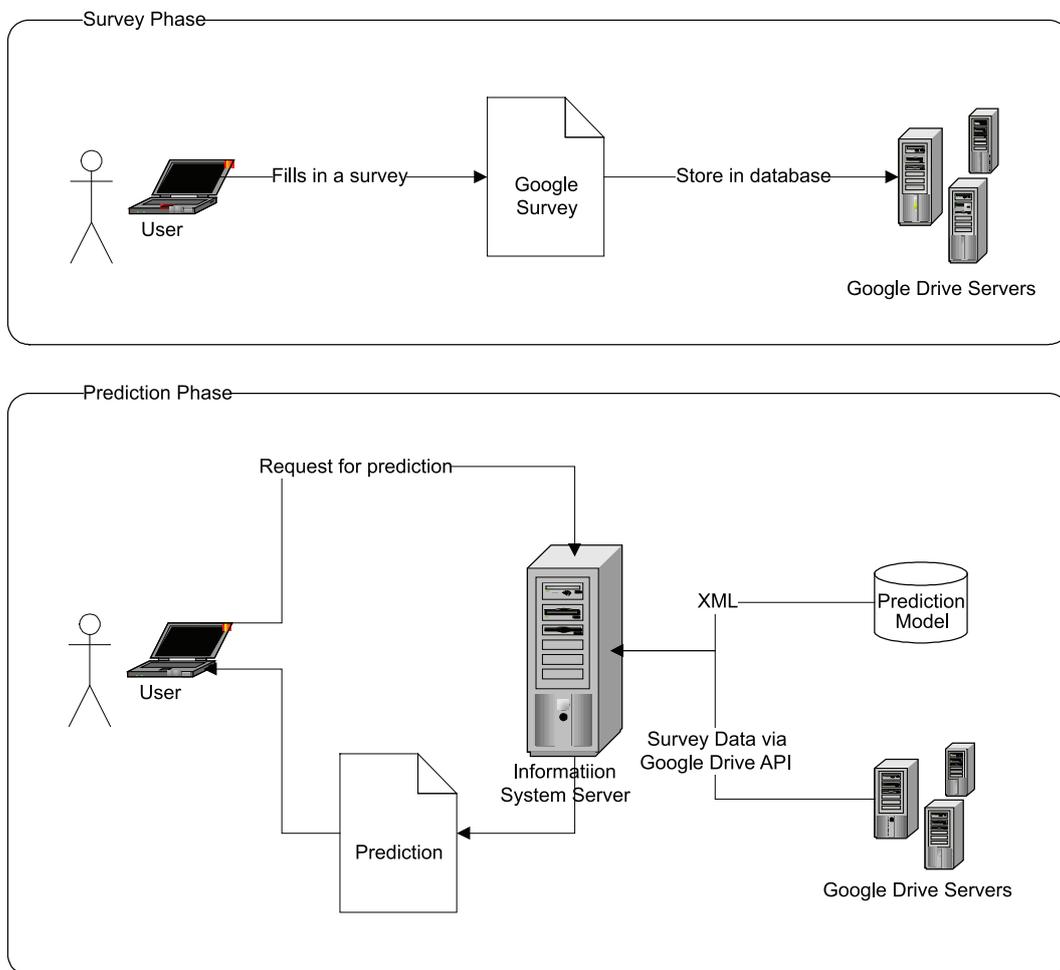


Fig. 8. Data Flow Diagram of the I3SP

Articles

I3SP uses the responses data to generate and return a prediction result – the company success probability and the populated classification tree. The I3SP calculates the prediction by matching the user's data with the classification tree from the data mining software.

The I3SP is currently operating in a closed beta version. Future plans for improvement of the investigations include further increasing the accuracy of the model by growing the data and involving more companies to use the system.

Conclusion and future plans

The paper has presented a quantitative investigation and creation of success prediction models and of a software tool based on the answers of the challenges and questions that start-up companies face. The analysis of the expanded data set, by using different data mining solutions and applying cross validation, produces various success prediction models with a good overall quality. The most accurate model is a classification tree which reveals the start-up success factors and their importance. The generated models are compared in terms of their accuracy and algorithms. As a result, the best models are used for developing the Information System for Start-ups Success Prediction.

The authors' experience and results show that the main challenges Bulgarian high-tech start-ups face, following open and disruptive innovations, include identifying innovations that have a market potential, obtaining know-how and expertise to develop new products and services, getting adequately funded at the initial stages of the business, marketing the product and gradually building a reputation, partnerships and a social network. The research also reveals that successfully overcoming the challenges identified requires a comprehensive understanding

of the industry, of the technologies involved in developing new products and the incalculable nature of disruptive innovations, the ability to evaluate critically, tolerance of risk and ambiguity, and carefully planned budget and growth predictions.

The study also offered several suggestions for further research, such as the problems related to the early-stage funding from investors' perspective, the existence of a possible correlation between disruptive innovation and growth, etc. Insights from the research could be used to help Bulgarian and international 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). The work on this paper has been also sponsored by IBS Bulgaria, IBM Premier Business partner.

References

- Amit R., Zott C., 2012, Creating Value through Business Model Innovation, *MIT SLOAN MANAGEMENT REVIEW*, Spring 2012, Reprint 53310, Vol. 53, No 3, pages 41-49.
- Blank, S. 2007, *The Four Steps to the Epiphany: Successful Strategies for Products that Win*. 3rd edition. Foster City, California: Cafepress.com.
- Bouckaert, R. F. (2013). *WEKA Manual for Version 3-6-10*. Hamilton, New Zealand: University of Waikato.
- Carland, J.W. and Carland, J.A., 2000, *A New Venture Creation Model*, Western Carolina University.
- Chrisman, J., Bauerschmidt, A. and Hofer, C., 1998, *The Determinants of New Venture Performance: An Extended Model*.

Articles

Cosentino T., 2014, Business Analytics in 2014,: Trends and Possibilities, Ventana Research, JAN 23, 2014, Accessed on 24 January 2014, http://www.information-management.com/blogs/business-analytics-in-2014-trends-and-possibilities-10025261-1.html?utm_campaign=daily-jan%2024%202014&utm_medium=email&utm_source=newsletter.

CRISP-DM 1.0, 2000, Step-by-step data mining guide, Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R..

Hall, M. F. 2009, The WEKA Data Mining Software: An Update. SIGKDD Explorations Volume 11, Issue 1

Hamel, G. 2012, What Matters Now: How to Win in a World of Relentless Change, Ferocious Competition, and Unstoppable Innovation. San Francisco, California: Jossey-Bass.

Huang, G. T. 2010, How to Predict Whether a Startup Will Succeed or Fail: Testing the "Disruptive Innovation" Model. Xconomy. Accessed on 24 January 2014, http://www.xconomy.com/seattle/2010/04/28/how-to-predict-whether-a-startup-will-succeed-or-fail-testing-the-disruptive-innovation-model/?single_page=true.

Models and Tools for Technology Start-Up Companies Success Analysis

Majamäki L., 2013, MASTERING DISRUPTIVE INNOVATION, Troubleshooting for Finnish High-Tech Start-Ups, Master's Thesis, International Business Management School of Business and Services Management, JAMK Centre for Competitiveness, December 2013.

Price, R. W. 2004, Roadmap to Entrepreneurial Success: Powerful Strategies for Building a High-Profit Business. New York: AMACOM.

Sandberg, W. R., 1986, New venture performance: The role of strategy and industry structure. Lexington, MA: Lexington Books.

Tsolova, S., 2014, Algorithm of Innovative Strategic Management E-System for Technology New Ventures, *The Journal of MacroTrends in Technology and Innovation*, Vol 2, Issue 1

Yankov, B., 2012, Overview of Success Prediction Models for New Ventures, International Conference Automatics and Informatics'12, ISSN 1313-1850, pp. 13-16.

Yankov, B., 2013, A Model for Predicting the Success of New Ventures, Vth International Scientific Conference e-Governance, ISSN 1313-8774, pp. 128-135.

Yankov, B., Haralampiev, K., Ruskov P., 2013, Start-up Companies Predictive Models Analysis, *Vanguard Scientific Instruments in Management*, 2013 (VSIM:13), ISSN 1314-0582, pp. 275-285.