

INTELLIGENT ALGORITHM FOR IMMEDIATE FINANCIAL STRATEGY FOR SMES

Carlo Lanzolla, Giuseppina Colasuonno, Katia Milillo and Gabriele Caputo

ORMA LAB srl, Corso Giuseppe Garibaldi n. 40, CAP:70027, Palo del Colle (Ba), Italy

ABSTRACT

In this work is discussed a scientific methodology concerning an intelligent algorithm oriented on financial strategy for SMEs. The paper follows the research guidelines of 'Frascati' manual about knowledge gain by innovative algorithms. Specifically has been applied a Support Vector Machine (SVM) algorithms predicting financial score of Small and Medium Enterprises –SMEs-. For the output results has been executed a Rapid Miner workflow. The used approach represents a methodology to follow in order to improve a research project about financial technologies.

KEYWORDS

Support Vector Machine, Data Mining, Frascati Guideline, Financial Score, SMEs financial features.

1. INTRODUCTION: BASIC STATE OF THE ART

Frascati Research and Development R&D guidelines [1] affirms that knowledge gain can be improved by innovative algorithms. Data mining –DM-, sometimes named as artificial intelligence, can be adopted for this purpose by predicting some features or attributes. Concerning financial applications data mining algorithms could be applied in order to improve financial information [2]-[3]. In particular DM are suitable for the estimation of loan risks [4]. Analysis of patterns of dataset to process [5], is an important step to follow before to apply a DM algorithm. Starting to a well structured dataset different DM can be applied in order to obtain hidden information or new knowledge. For example Decision tree algorithm has been executed in literature to predict prospective business sectors for lending in retail banking [6], besides artificial neural networks –ANNs- has been adopted in finance for cash forecasting [7] or for market stock prediction [8]. An important application of ANN is also for credit risk prediction [9]. Data mining can be applied for a variety of financial applications [10]-[13] thus confirming that it can represent an important tools for financial service innovations. Starting to model able to describe credit risk [14] or other risk typologies, can be utilized different open source tools for financial information gain [15]-[25]. One of these tool is Rapid Miner [25] having the characteristic to model a financial problem by means of a structured workflow. A workflow is a graph describing the data processing during each stages such as data pre-processing, data processing and output reporting. In this paper is described a Rapid Miner workflow implementing a Support Vector Machine [25] –SVM- data mining algorithm about SMEs risk prediction thus explaining how data can be processed in order to improve the knowledge about financial strategies. In particular the paper is structured as follows:

- is designed and discussed a Rapid Miner workflow implementing SVM algorithm about SMEs risk financial evaluation;

- is described the experimental dataset;
- is executed the SVM workflow by discussing results;
- is concluded the paper by showing the research topics.

The proposed algorithm has been tested within the framework of an industry project called "INTELLIGENT ALGORITHM FOR IMMEDIATE FINANCIAL STRATEGY FOR SMEs: I.A.F." concerning the study of a prototype system mainly based on predictive algorithms, Big Data and financial tech techniques, able to suggest some SME to the Business Management of SMEs business-financial strategies to be applied. The "I.A.F." intelligent system based on the company data received different data such as balance sheet data, sector data, reference market, etc., and uses of the business intelligence –B.I.- in order to develop an innovative corporate strategy suggesting also the tax and financial benefits applicable to SMEs.

The "I.A.F." system will allow the companies to

- reduce time-to-market;
- speed up corporate decision-making processes;
- reduce the risk of corporate insolvency;
- prevent the risk of corporate insolvency;
- maximize the expected profit of the entrepreneur;
- trace a corporate strategy in order to improve balanced data sheet, and consequently to improve the judgment of the banks;
- transforms data analysis into immediate added value.

2. DATA MINING SVM WORKFLOW MODELLING

The workflow model proposed in this paper can be applied to predict financial risk. Specifically it predict the risk of counterparty credit default by training a Support Vector Machine (SVM) model on credit default data, optimizing its core parameters C and gamma, and scoring risk on new data. The model is structured as a workflow working automatically at each of the following stages (named steps):

- Stage1 (loading input data to process): it loads the counterparty risk data containing attributes of firms and past default observations useful for the training dataset of the model. For the firms where a default observation is missing, the default risk will be predicted.
- Stage 2: Edit, transform & load (ETL) - it splits data into rows having a label value and where the label value is missing. The rows with labels are adopted for training a model which will predict the default risk for the rows without a labels representing the queries of the financial interrogation.
- Stage 3: Train and optimize –SVM- model to predict credit risk; this optimization operator varies the important SVM parameters C and gamma to return a model with maximum prediction accuracy (important aspect for the estimation of the research efficiency).
- Stage 4: using the optimized SVM model, it will predict a possible credit failure (credit risk).

- Stage 5: it converts the model optimization log to a data set.

In Fig. 1 is illustrated the experimental Rapid Miner workflow implementing each of stages above described.

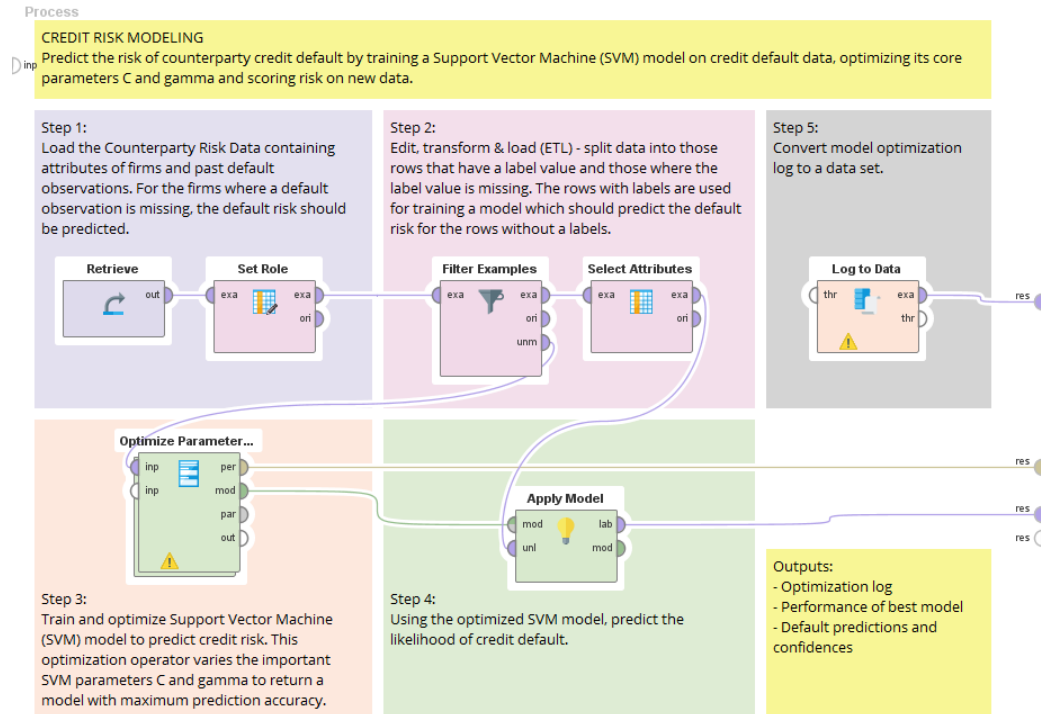


Figure 1. Rapid Miner workflow: SVM workflow modelling.

We provide more information about the operators represented in the workflow of Fig. 1. The ‘Retrieve Operator’ loads a Rapid Miner input dataset into the Process. This dataset can also be a data collection or a data model. The role of an attribute describes how other operators handle this attribute (operator ‘Set Role’). The default role is regular, other roles are classified as special. An input dataset can have many special attributes, but each special role can only appear once. If a special role is assigned to more than one attribute, all roles will be changed to regular except for the last attribute. In Fig. 2 is shown the panel regarding ‘Set Role’ parameter setting.



Figure 2. Rapid Miner parameter panel setting: set role parameter setting.

‘Filter Examples’ operator returns those dataset that match the given condition. ‘Select Attribute’ operator provides different filter types to make attribute selection easy. The invert selection

parameter reverses the selection. Special attributes (attributes with roles, like id, label, weight) are by default ignored in the selection.

Only the selected Attributes are delivered to the output port. The rest is removed from the input dataset. In Fig. 3 is illustrated the used parameters panel setting about attribute selection

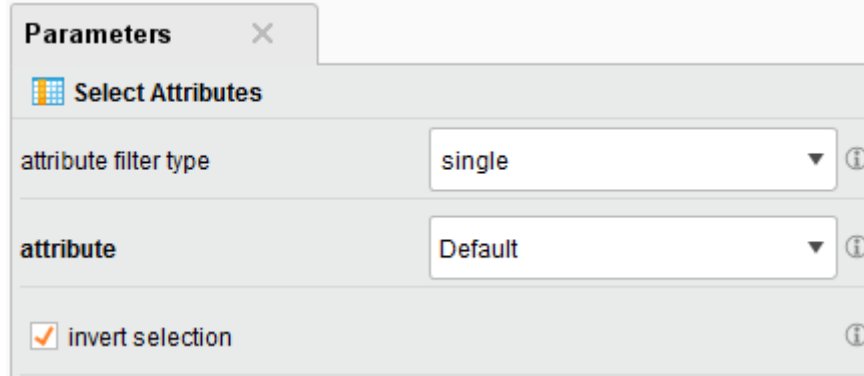


Figure 3. Rapid Miner parameter panel setting: attribute selection.

The 'Optimize Parameter' operator returns an optimal parameter set which can also be written to a file. This parameter set can be read in another process and then be applied using the 'Set Parameters' operator.

'Apply Model' is a model trained on a specific dataset (learning algorithm). The dataset upon which the model is applied, has to be compatible with the attributes of the model: this means, that the dataset has the same number, order, type and role of attributes as the dataset used to generate the model.

The 'Log to Data' operator stores information into the log table. This information can be almost anything including parameter values of operators, apply-count of operators, execution time etc. This operator is mostly used in order to see the values calculated during the execution of the process that are otherwise not visible. A large variety of information can be stored using this operator. The information stored in the log table can be viewed in the 'Results View' panel. The 'Log to Data' operator provides the information in the Log table in form of a dataset. This dataset can be used in the process like any other dataset. Rapid Miner automatically estimates the type of attributes of this dataset and all attributes have regular role. The type and role can be changed by using the corresponding operators.

Row No.	iteration	SVM.kernel_gamma	SVM.C	accuracy
41	41	0.775	1.000	0.751
42	42	1.443	1.000	0.751
43	43	2.686	1.000	0.751
44	44	5	1.000	0.751
45	45	0.010	10.000	0.951
46	46	0.019	10.000	0.956
47	47	0.035	10.000	0.954
48	48	0.065	10.000	0.946
49	49	0.120	10.000	0.921
50	50	0.224	10.000	0.823
51	51	0.416	10.000	0.754
52	52	0.775	10.000	0.751
53	53	1.443	10.000	0.751

ExampleSet (66 examples, 0 special attributes, 4 regular attributes)

Figure 4. Rapid Miner: results panel of the executed SVM workflow.

2.1.Experimental dataset

The experimental dataset is represented by the following attributes:

- Default: training dataset indicating if there is a credit risk or not; the training dataset is based on observed data;
- Long Term Financing on Working Capital;
- Working Capital Requirement;
- Debt Cash Flow Coverage Ratio;
- Liability to Equity;
- Net Debt to Equity Ratio;
- Debt to Capital Ratio;
- Long Term Debt to Asset;
- Long Term Debt To Tangible Asset;
- Interest Coverage Ratio;
- Net Profit Margin;
- Gross Profit Margin;
- Return on Invested Capital;
- Return on Equity;
- Fixed Asset to Debt Ratio;
- Short Term Debt to Sales Ratio;
- Expense to Revenue Ratio;
- Fixed Asset Turnover;
- Collateral;

- Firm Size.

In Fig. 5 are illustrated the dataset imported at the input of the workflow model

Row No. ↑	Default	Long Term Financing of Working Capital	Working Capital Requirement	Debt Cash Flow Coverage ...	Liability to Equity	Net Debt to
1	No	18.535	70.224	2.222	0.409	0.345
2	No	1.268	68.543	2.276	0.152	0.099
3	No	26.770	7.496	0.843	0.470	1.186
4	No	5.789	17.880	1.365	0.046	0.138
5	No	11.864	5.164	0.061	0.102	0.047
6	No	6.651	16.898	2.093	0.107	0.117

Row No. ↑	Default	Long Term Financing of Working Capital	Working Capital Requirement	Debt Cash Flow Coverage ...	Liability to Equity	Net Debt to
385	Yes	0.718	17.068	0.329	0.314	0.018
386	Yes	2.967	4.307	1.195	0.289	0.198
387	Yes	7.372	8.275	0.198	0.348	0.162
388	Yes	1.278	15.247	0.087	0.399	0.143
389	Yes	1.452	36.272	0.432	0.127	0.139
390	Yes	3.361	6.925	0.293	0.283	0.126
391	?	0.836	51.987	1.787	0.018	0.010
392	?	2.187	111.569	1.654	0.124	0.386
393	?	4.828	40.670	1.700	0.094	0.193
394	?	0.853	51.729	1.123	0.258	0.045
395	?	3.210	77.686	0.147	0.206	0.098
396	?	8.440	20.271	0.406	0.100	0.387
397	?	1.110	20.350	1.015	0.210	0.138
398	?	4.867	88.386	0.105	0.026	0.089

420	?	5.922	3.451	0.301	0.075	0.130
421	?	1.332	7.506	1.452	0.126	0.309
422	?	5.450	15.527	1.070	0.124	0.131
423	?	5.315	17.587	0.820	0.324	0.393
424	?	3.095	1.509	0.797	0.624	0.283

Figure 5. Screenshot of the dataset imported at the input of the workflow model.

3. SVM WORKFLOW RESULTS

The results are provided by selecting the result table of Fig. 6 indicating the row number (each row corresponds to a SME), the predicted risk (Yes or No risk), and the confidence estimation of each results. The calculus is performed by using a Kernel model with 390 as total number of Support Vectors, and a bias (offset) of -6.443.

Row No.	prediction(Default)	confidence(No)	confidence(Yes)
21	Yes	0.000	0.997
22	Yes	0.011	0.989
23	Yes	0.022	0.978
24	No	0.509	0.491
25	Yes	0.017	0.983
26	Yes	0.021	0.979
27	No	0.998	0.002
28	Yes	0.000	1.000
29	Yes	0.012	0.988
30	Yes	0.010	0.990
31	Yes	0.019	0.981
32	Yes	0.316	0.684
33	No	0.941	0.059
34	Yes	0.002	0.998

Figure 6. Table of results (first 34 records).

In table 1 are reported the estimated weight of each attribute used in the calculation.

Attribute	Weight
Long Term Financing of Working Capital	17.90645720899751
Working Capital Requirement	-55.98365096112798
Debt Cash Flow Coverage Ratio	-73.87482159307879
Liability to Equity	109.64944920207228
Net Debt to Equity Ratio	-71.99542822055058
Debt to Capital Ratio	-48.930525971855275
Long Term Debt to Asset	-36.90814717179048
Long Term Debt to Tangible Asset	-40.120912661218604
Interest Coverage Ratio	5.167521509849003
Net Profit Margin	88.97217780342255
Gross Profit Margin	-2.6473026640675448
Return on Invested Capital	17.2693594654948
Return on Equity	2.958762936381593
Fixed Asset to Debt Ratio	-7.902717912084212
Short Term Debt to Sales Ratio	40.11710834806478
Expense to Revenue Ratio	183.6714461471422
Fixed Asset Turnover	-38.78159844308285
Collateral	14.735624473146796

Table 1. Assigned weights during the SVM calculus.

By observing data of Table I, the attributes of “Liability to Equity” and “Expense to Revenue Ratio” have the major weight for the financial risk calculus. By using the graphical libraries it is possible to plot graphically the output results. Focusing the attention on the attributes with the major weights, we plot in Fig. 7 the predicted risk and no-risk conditions. A full view of the predicted results is observed in Fig. 8 where is plotted the correlation between prediction and the four attributes having the major calculated weights

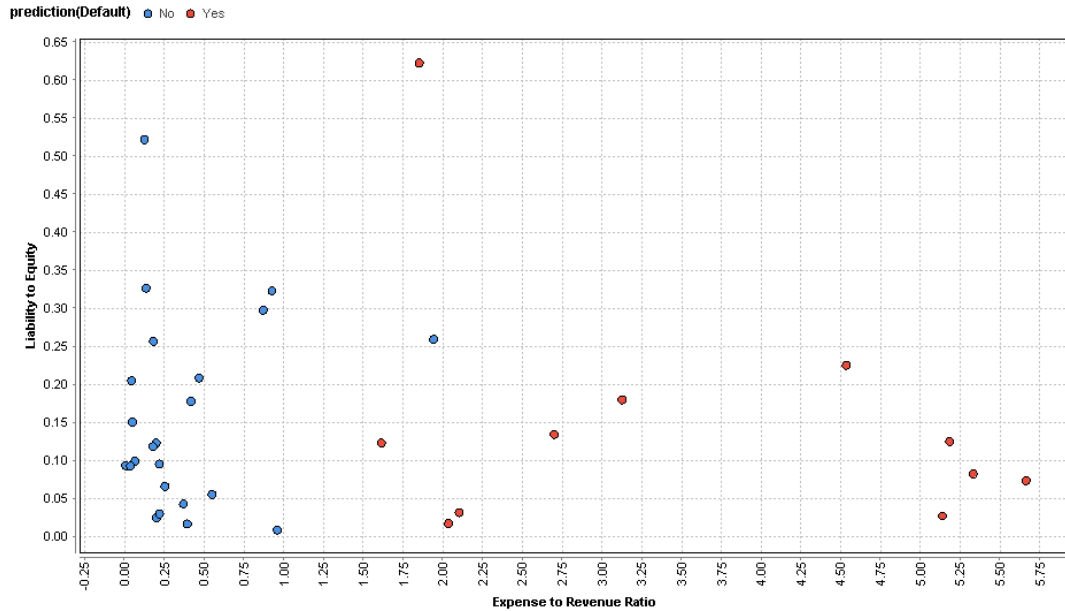


Figure 7. “Liability to Equity” attribute versus “Expense of Revenue Ratio” attribute.

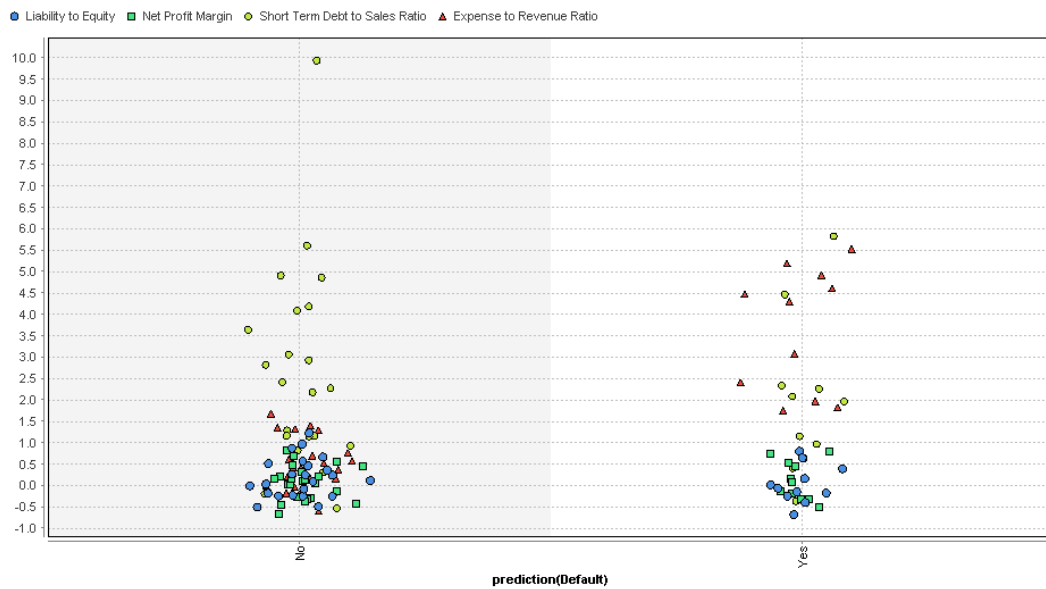


Figure 8. Other main attributes versus risk prediction estimation.

Below are reported the data concerning the estimated Performance Vector:

Confusion Matrix (see Fig. 9):

True: No Yes

No: 286 10

Yes: 7 87

accuracy: 95.64% +/- 3.25% (micro average: 95.64%)

precision: 93.55% +/- 7.14% (micro average: 92.55%) (positive class: Yes)

recall: 89.56% +/- 13.94% (micro average: 89.69%) (positive class: Yes)

AUC (optimistic): 0.990 +/- 0.012 (micro average: 0.990) (positive class: Yes)

AUC: 0.990 +/- 0.012 (micro average: 0.990) (positive class: Yes)

AUC (pessimistic): 0.990 +/- 0.012 (micro average: 0.990) (positive class: Yes)

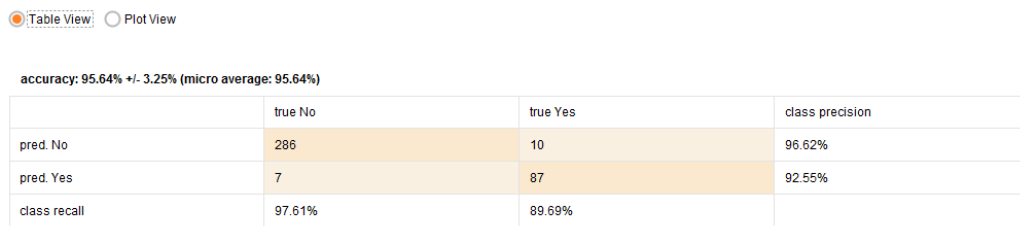


Figure 9. Estimated Confusion matrix.

The Performance Vector data and the ROC curve of Fig. 10 show that the training dataset is enough in order to execute a SVM model performing reliable results.

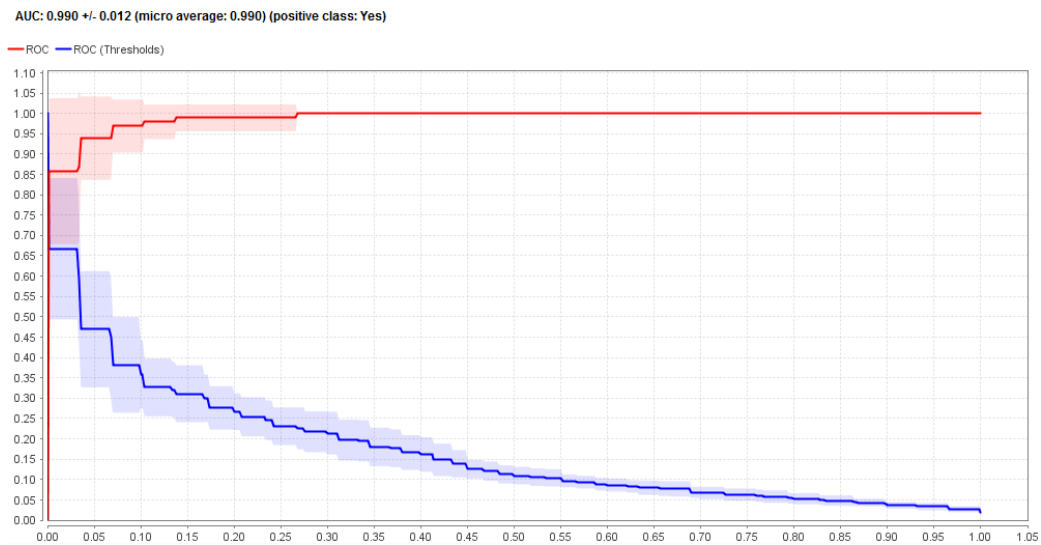


Figure 10. Calculated ROC Curve.

Finally in Fig. 11 is represented the global reporting of the SMEs dangerous for financial credit. About the experimentation 97 SMEs are characterized by an high risk and 293 SMEs are in a no dangerous zone.

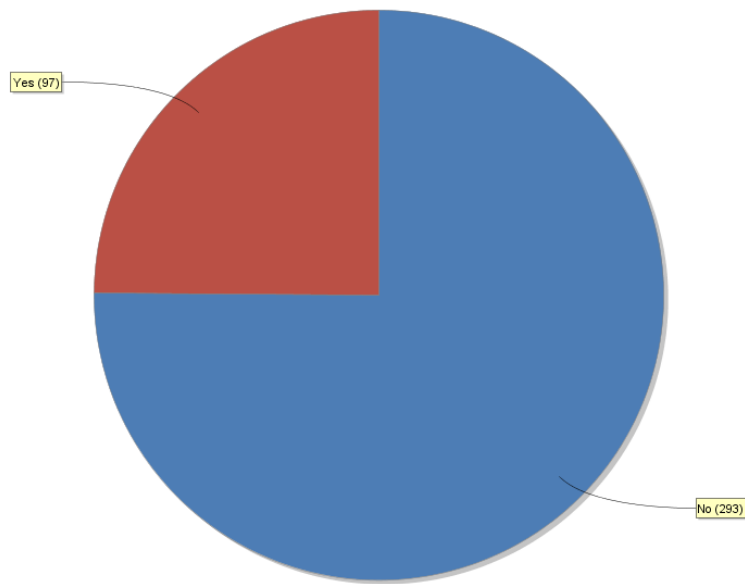


Figure 11. Pie plot of the global predicted results.

6. CONCLUSION

The results are developed within the framework of a research project. Other algorithms can be applied for the calculus of financial scores of SMEs using other technologies such as big data systems able to collect massive data useful to create a more reliable model. The goal of the paper is to discuss a methodological approach for the gaining of the knowledge base about financial score by following R&D 'Frascati' guideline. In the example shown in this paper a good ROC curve proves the accuracy of the applied Support Vector Machine –SVM- algorithm predicting financial behaviour of SMEs. Other innovative algorithms can be applied to other dataset by improving new financial services oriented on SMEs.

ACKNOWLEDGEMENTS

The work has been developed in the frameworks of the Italian projects: "INTELLIGENT ALGORITHM FOR IMMEDIATE FINANCIAL STRATEGY FOR SMEs: I.A.F.". Authors gratefully thanks Dr. Eng. Alessandro Massaro for his support about the writing of the paper.

REFERENCES

- [1] Frascati Manual 2015: The Measurement of Scientific, Technological and Innovation Activities- Guidelines for Collecting and Reporting Data on Research and Experimental Development. OECD (2015), ISBN 978-926423901-2 (PDF)..
- [2] Boris Kovalerchuk, Evgenii Vityaev, "DATA MINING FOR FINANCIAL APPLICATIONS," Springer, Data Mining and Knowledge Discovery Handbook, pp 1203-1224, 2005, ch. 57.
- [3] G. Cenk Akkaya , Ceren Uzar, "Data Mining in Financial Application," Journal of Modern Accounting and Auditing, ISSN 1548-6583, December 2011, Vol. 7, No. 12, 1362-1367.

- [4] Gerritsen, R. (1999). Assessing loan risks: A data mining case study. *IT Professional*, 1, 16-21.
- [5] Amritpal Singh, Amrita Kaur, Jasmeet Kaur, Ramandeep Singh, Shipra Raheja, "Pattern Analysis On Banking Dataset," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 06, JUNE 2015*.
- [6] Md. Rafiqul Islam, and Md. Ahsan Habib, "A DATA MINING APPROACH TO PREDICT PROSPECTIVE BUSINESS SECTORS FOR LENDING IN RETAIL BANKING USING DECISION TREE," *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, March 2015*.
- [7] PremChand Kumar, Ekta Walia, "Cash Forecasting: An Application of Artificial Neural Networks in Finance," *International Journal of Computer Science & Applications © 2006 Technomathematics Research Foundation Vol. III, No. I, pp. 61 – 77*.
- [8] Selvan Simon and Arun Raoot, "ACCURACY DRIVEN ARTIFICIAL NEURAL NETWORKS IN STOCK MARKET PREDICTION," *International Journal on Soft Computing (IJSC) Vol.3, No.2, May 2012*.
- [9] Sihem Khemakhem, and Younés Boujelbène, "Credit risk prediction: A comparative study between discriminant analysis and the neural network approach," *Accounting and Management Information Systems, Vol. 14, No. 1, pp. 60-78, 2015*.
- [10] Adam Fadlalla, Chien-Hua Lin "An Analysis of the Applications of Neural Networks in Finance," *INTERFACES 31: 4 July–August 2001 (pp. 112–122)*.
- [11] S.B. Soumya, N. Deepikam, "Data Mining With Predictive Analytics for Financial Applications," *International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-1, January 2016*.
- [12] Cüneyt Dirican, "The Impacts of Robotics, Artificial Intelligence On Business and Economics," *Procedia - Social and Behavioral Sciences 195 (2015) 564 – 573*.
- [13] Filippo Fuggitti, "Tecniche e Metodologie di Data Mining: Analisi applicata alle Charities inglesi," *tesi LUISS, 2016*.
- [14] Greta Di Fabio, "Applicazioni dei modelli di gestione del rischio di credito," *tesi LUISS, 2011*.
- [15] Hayden Wimmer, Loreen M. Powell, "A Comparison of Open Source Tools for Data Science," *2015 Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA*.
- [16] Paško Konjevoda and Nikola Štambuk, "Open-Source Tools for Data Mining in Social Science," *Intech, ch. 8, Business, Management and Economics » "Theoretical and Methodological Approaches to Social Sciences and Knowledge Management", 2012*.
- [17] Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3), 299-314.
- [18] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., . Starič, A. (2013). Orange: data mining toolbox in python. *the Journal of machine Learning research*, 14(1), 2349-2353.
- [19] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Wiswedel, B. (2008). *KNIME: The Konstanz information miner*: Springer.

- [20] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.
- [21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- [22] Holmes, G., Donkin, A., & Witten, I. H. (1994, 29 Nov-2 Dec 1994). WEKA: A machine learning workbench. Paper presented at the Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on.
- [23] Rakotomalala, R. (2008). Tangara.
- [24] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.
- [25] Vijay Kotu, Bala Deshpande, "Predictive Analytics and Data Mining," Elsevier 2015.