Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 4, August 2025, pp. 3029~3036
ISSN: 2302-9285, DOI: 10.11591/eei.v14i4.9556

Feature selection for support vector machines in imbalanced data

Borislava Toleva[1], Ivan Ivanov[1], Vincent Hooper[2]
[1]Faculty of Economics and Business Administration, Sofia University St. Kl. Ohridski, Sofia, Bulgaria
[2]SP Jain Global School of Management, Academic City, Dubai, United Arab Emirates

Article Info

Article history:

Received Nov 14, 2024
Revised May 31, 2025
Accepted Jul 5, 2025

Keywords:

Analysis of variance
Bankruptcy prediction
Class imbalance
Feature selection
Support vector machines

ABSTRACT

Addressing the effects of class imbalance on feature selection models has become an increasingly important focus in academic research. This study introduces a novel support vector machine (SVM)-based algorithm specifically designed to handle class imbalance during the feature selection process. Using the Taiwan bankruptcy dataset as a case study, the algorithm incorporates the ExtraTreeClassifier() to manage class imbalance and identify a reduced set of relevant variables. To validate the selected features, SVM is applied within the imbalanced data context. Subsequently, analysis of variance (ANOVA) ranking is employed to further refine the variable set to three key features. An SVM model tailored for class imbalance is then constructed to assess the effectiveness of the final feature set. The proposed model significantly outperforms existing approaches in terms of classification performance. Specifically, it achieves a Type I error of 1.17% and a Type II error of 22.9%, compared to 4.4% and 39.4% reported in prior research. In terms of overall accuracy, our method reaches 83.1%, surpassing the 81.3% achieved by earlier studies. These results demonstrate that the proposed feature selection algorithm not only improves SVM accuracy but also outperforms other feature selection techniques when used in conjunction with SVMs, particularly under conditions of class imbalance.

Corresponding Author:

Borislava Toleva
Faculty of Economics and Business Administration, Sofia University St. Kl. Ohridski
Sofia, Bulgaria
Email: vrigazova@uni-sofia.bg

1. INTRODUCTION

Feature selection has become a key focus in machine learning, offering a means to enhance algorithm quality by removing redundant features. This process not only improves algorithm efficiency but also aids in revealing a small group of factors that significantly impact an event. However, the effectiveness of feature selection algorithms relies on the fulfillment of certain assumptions by the data. For example, class imbalance in the target variable can lead to overfitting of the model.

While much academic literature addresses handling class imbalance in the final classification stage, this research seeks to address it at an earlier stage: feature selection. We propose a new algorithm designed to handle class imbalance during feature selection, with a focus on improving the performance of support vector machines (SVM). Our algorithm aims to effectively identify a subset of features that not only improve SVM predictions but also help explore the connection among the independent variables and the target variable.

This study presents a novel approach to addressing class imbalance within the feature selection process, providing a valuable tool for enhancing prediction accuracy and interpretability in machine learning models. The following sections delve into current academic research on the topic, the methodology behind our proposed algorithm, and the results of our experiments.

Identifying the key factors that influence events is crucial across various scientific fields, including economics. A prominent area of interest within economics is understanding the factors contributing to a company's bankruptcy. To address this, credit scoring algorithms have been developed in [1]. However, creating a universal algorithm for identifying bankruptcy factors faces challenges such as class imbalance in datasets [2], insufficient data [3], and changes in the economic and regulatory conditions [4]. Consequently, researchers often develop bankruptcy prediction models tailored to specific countries or economic sectors, such as public companies [5], US companies [6], Chinese companies [7], and Taiwanese companies [8].

In the context of identifying factors influencing company bankruptcy, two primary categories of algorithms are commonly employed: machine learning and deep learning. These algorithms are frequently combined with feature selection techniques to identify the most influential factors in corporate bankruptcy. Zhao *et al.* [9] offer a comprehensive review of various algorithm groups for predicting corporate bankruptcy. They categorize these algorithms into several groups, including:

- Multivariate discriminant algorithms [10];
- Regression algorithms [4];
- Stochastic process-based algorithms [11];
- Decision trees [12];
- Neural networks [13], [14];
- Ensemble learning algorithms [15].

Other types of bankruptcy prediction algorithms also exist, including Altman models and credit scoring models [16]-[23]. Regardless of the algorithm used, a central topic is often identifying the factors behind corporate bankruptcy [24], [25].

SVMs is another type of machine learning algorithm that is widely applied for imbalanced classification due to its flexibility. For instance, Ye *et al.* [26] developed a novel SVM model aimed at handling class imbalance. The main assumption behind this novel approach is that the positive class conditional posterior probability density function is quadratic. However, this assumption is not fulfilled in all datasets. Research by Wei *et al.* [27] propose another SVM model aimed at handling class imbalance. This is the multilayer support vector machines (ML-SVM). However, this approach is targeted specifically at detecting fault signals from healthcare equipment in the case when the data for one of the class are dominating. This algorithm has not been tested on general datasets. Other versions of the SVMs for class imbalance also exist [28], [29] but they are also targeted at resolving a specific task in the healthcare. Maldonado and López [30] propose embedded feature selection for SVM in the case of class predominance in the target variable. However, the efficiency of [30] on corporate failure data has not been examined.

Given the ongoing challenges in identifying corporate bankruptcy factors [9], [24], [25], the goal of this paper is to introduce an effective and simple feature selection algorithm for SVMs to predict corporate bankruptcy in cases of highly imbalanced data. This algorithm was developed using the Taiwan bankruptcy dataset [25]. The proposed algorithm is effective as it can be used with large datasets, where feature selection is necessary. As the algorithm uses ExtraTreeClassifier and analysis of variance (ANOVA), the importance of the features is ranked, giving the researcher the opportunity to further explore the influence of various combinations of features on the corporate bankruptcy and the effects from the class imbalance. The application of the algorithm is simple as it involves using the built-in functions in the software package (e.g., Python). Also, the model can be easily adjusted to various datasets or newly added/removed features by parameters tuning. The following section details the proposed methodology, while sections 2 and 3 explain the results. Section 4 concludes.

2. METHOD

The Taiwan bankruptcy dataset consists of 6,819 rows and 95 independent (X) variables [25]. The dependent variable, 'bankrupt', exhibits class imbalance, with 0 being the dominant class. To address this imbalance and select variables for the classification algorithm, an appro-priate feature selection algorithm is necessary. This not only reduces computing time but also enhances algorithm accuracy, aiding in the analysis of key factors leading to bankruptcy.

The methodology involves identifying 29 variables that impact the bankruptcy of Taiwanese companies and fitting a classification algorithm that predicts bankruptcy cases accu-rately despite the class imbalance. The algorithm is implemented in Python 3.11 on Windows 11, utilizing an Intel Core i3 processor. We have performed many experiments until we find parameter values for the commands below that produce high classification metrics. However, the parameters shown in the methodology are the ones with which we obtained the best results. The steps are outlined as follows:

- Load the data and define the X and Y variables. An important note is that we do not standardize the data.
- Use feature selection to reduce the number of independent variables, considering class imbalance. Utilize ExtraTreesClassifier [31] with parameters (n_estimators=100, class_weight='balanced_subsample', bootstrap=True, random_state=250) to rank feature im-portance and SelectFromAlgorithm(clf, prefit=True) to select important features based on the rankings. Use ExtraTreesClassifier to rank features based on importance scores. Use SelectFromAlgorithm to select the most important features. ExtraTreesClassifier is a useful tool to perform feature selection in cases of class imbalance according to academic literature. Setting the parameters ‘class_weight’ to ‘balanced_subsample’ and ‘bootstrap’ to ‘True’ are effective way to handle class imbalance according to Python’s documentation.
- Set a classification algorithm (SVM) to be fitted on the selected features. SVM=SVC(C=250, class_weight='balanced'). We conducted experiments with other values for C but this is the one with which the classification metrics were the best.
- Set a cross-validator (kFold cross-validation). skf=KFold(n_splits=10, shuffle=True, random_state=seed).
- Run the classification algorithm and evaluate its performance. Calculate metrics such as accuracy, precision, recall, F1-score, and confusion matrix to evaluate its performance.
- Fit ANOVA feature selection on the 29 variables to select the three most important features. Unlike other feature selection methods, the ANOVA provides a ranking of the importance for each feature. Therefore, the researcher has a better understanding of the importance of a feature that has not been selected in the model. Also, knowing the importance of each feature would allow the research to conduct experiments with various combinations of features and easily check whether he/she should include or exclude the particular feature. Also, the ANOVA provides an easy way to grasp how feature importance changes when a new feature is added to the dataset or an existing one removed from the dataset. Therefore, changes in the classification metrics can be more easily tracked and analyzed whether their source is the change of features in the dataset or a change in the model setting. In view of all these advantages of the ANOVA model, its usage in the proposed setting aligns best with the purpose of this research. In Python we use SelectKBest(score_func=f_classif, k=3) for ANOVA feature selection.
- Tune the classification algorithm on the three selected features to confirm their predictive ability.
- Fit the tuned algorithm with cross-validation.
- Evaluate the algorithm's performance using the selected three features. The metrics used for models’ evaluations are confusion matrices, accuracy, precision, recall, and F1-score. Also, Types I and II errors have been calculated using the same methodology as in [25] to have comparable results to existing academic literature.

By following these steps, the methodology aims to develop a robust feature selection algorithm that addresses class imbalance and accurately predicts bankruptcy using a subset of key variables.

3. RESULTS AND DISCUSSION

The results of the proposed methodology can be divided into two parts: the output from the feature selection for imbalanced data (steps 1-7) and the validation of the selected features' predictive ability (steps 8 and 9). The findings highlight the efficacy of the novel SVM-based feature selection method and its superior performance over alternative techniques.

In the first part of the analysis (steps 1-7), a two-tier feature selection algorithm is employed to address class imbalance. Using ExtraTreeClassifier with 'balanced_subsample' in steps 2 and 3 ensures that feature selection considers class imbalance, avoiding potential issues such as overfitting or underfitting. After applying step 2, 29 features are identified from the initial dataset of 95 independent variables.

Steps 3-5 utilizes a SVMs algorithm to validate the 29 selected features' ability to produce a robust model. Subsequently, ANOVA is applied to these 29 variables to select the three most important features based on their f-scores. The three selected variables are after-tax net profit growth rate (score-753), ROA(A) before interest and % after tax (score-593), and ROA(B) before interest and depreciation after tax (score–549).

In the second part of the algorithm (steps 6 and 7), the three selected variables are further validated for their predictive ability using the SVM model. The overall accuracy of steps 1-9 is 83.1%, indicating the effectiveness of the proposed methodology. Figure 1 shows the confusion matrix from steps 1-9, illustrating the model's ability to predict bankruptcy cases. It is demonstrated that the novel SVM-based feature selection algorithm is successful in handling class imbalance and outperforms existing methodologies in predicting bankruptcy.

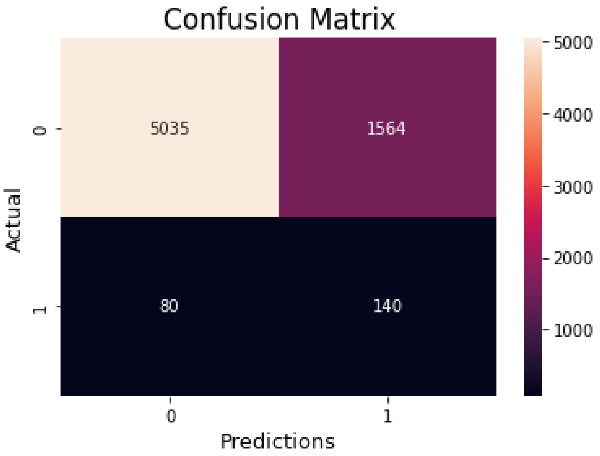

Figure 1. Confusion matrix from steps 3-9 authors' calculations

The high accuracy achieved by the algorithm, along with the detailed confusion matrix, confirms its effectiveness in predicting both majority (class 0) and minority (class 1) observations. Specifically, out of 6,699 observations for class 0, 5035 were correctly predicted, and out of 220 observations for class 1, 140 were correctly predicted. These results validate the importance of the 29 selected features for predicting both classes.

The primary objective of this research is to develop an algorithm capable of selecting a significantly reduced set of features while maintaining effectiveness for company analysis. Table 1 displays the classification metrics, offering deeper insights into the algorithm's performance:

- Accuracy: an accuracy of 83.1% reflects a high overall prediction correctness.
- Precision: with a precision of 61.8%, the algorithm demonstrates a moderate level of false positive control.
- Recall (sensitivity): a recall of 63.6% suggests the algorithm has low levels of false negatives.
- F1-score: the F1-score for class 1 stands at 62.7%, indicating a well-rounded performance in detecting relevant cases.

Table 1. Classification metrics authors' calculations

Class	Precision	Recall	F1-score	Support
0	0.992	0.832	0.905	6599
1	0.138	0.805	0.235	220

These classification scores further validate the effectiveness of the proposed algorithm in selecting a relevant subset of features for company analysis, while maintaining a high level of predictive accuracy. Precision assesses the proportion of correctly predicted instances of a class among all predictions for that class. In this case, a precision of 0.992 for class 0 indicates that 99.2% of the instances predicted as class 0 were correct. Conversely, a precision of 0.138 for class 1 shows that only 13.8% of the predicted class 1 instances were accurate. This result is anticipated, given the class imbalance: class 1 represents a minority with only 220 instances, compared to 6,599 instances for class 0. Recall or sensitivity, measures the proportion of actual instances of a class that were correctly predicted. A recall of 0.832 for class 0 means the algorithm correctly identified 83.2% of class 0 cases. For class 1, a recall of 0.805 indicates the model successfully detected 80.5% of actual class 1 observations. These values demonstrate strong performance in detecting both majority and minority classes. When combined with an overall accuracy of 83.1% and the insights from the confusion matrix, these metrics confirm that the algorithm is effective in analyzing the three selected features: after-tax net profit growth rate, ROA(A) before interest and after tax, and ROA(B) before interest and depreciation after tax.

Table 2 presents a comparative analysis between the proposed method and the results from Liang *et al.* [25], which employed three SVM-based models. Liang's best-performing model (SVM (cost of financing (FC))), using the FC feature set, achieved an accuracy of 81.3%. In contrast, our algorithm—automatically selecting a diverse subset of only three features—achieved a higher accuracy of 83.1%, showcasing its superior performance. In addition, our SVM-based approach demonstrated a lower Type I error compared to Liang's best model, while maintaining a similar Type II error (as shown in Table 2). This indicates that our feature selection method, when integrated with SVM, outperforms Liang's approach in both accuracy and error reduction. Liang *et al.* [25] also explored various feature combinations using SVM-based selection, focusing on two main categories: financial ratios (FR) and FC. Table 3 provides a comparison of accuracy, Type I error and Type II error across these combinations, further highlighting the advantages of our proposed methodology.

Table 2. Comparison between our results and Liang's metrics [25]

Model	Accuracy (%)	Type I error (%)	Type II error (%)	Source
SVM (FRs)	79.1	20.2	21.6	Table 5 [25]
SVM (CGIs)	67.9	27.7	24.5	Table 5 [25]
SVM (FC)	81.3	17.8	19.7	Table 6 [25]
SVM (3 variables)	83.1	1.17	22.9	Authors' calculations

Table 3. Liang's results [16] on different cost ratios and feature selection+SVMs

Ratio	FR		FC	
	Type I error (%)	Type II error (%)	Type I error (%)	Type II error (%)
1	20	18.1	16	19.3
1.5	12.5	27	10	26.1
2	10.2	30.9	7.1	30.9
3	6.9	38.4	5.4	34.8
5	3.5	50.5	4.8	36.9
7.5	1.9	60.1	4.4	39.4
10	1.3	65.8	4.4	39.4
15	0.9	69.9	4.4	39.4
20	0.9	69.9	4.4	39.4
30	0.9	69.9	4.4	39.4

Tables 2 and 3 highlight the effectiveness of our algorithm compared to Liang's SVM [25] experiments using different ratios for FR and FC. Our algorithm achieves a comparable Type I error (1.17%) to Liang's SVM experiments with an FR ratio of 15/20 and 30, where Liang achieves the smallest Type I error using an FR ratio of 15/20/30. However, our model achieves this low Type I error using only 3 variables, whereas Liang's model requires more variables. Additionally, our model significantly outperforms Liang's in terms of Type II error, with our model achieving a much lower error rate of 22.9% compared to Liang's 69.9% with the FR ratio of 15/20/30.

Similar advantages are observed when comparing our results with Liang's experiments using different FC ratios. Liang's best results with FC ratios of 7.5/10/15/20/30 yield higher Types I and II errors (4.4% and 39.4%, respectively) compared to our results of 1.17% and 22.9%. These findings indicate that our feature selection algorithm for SVMs consistently outperforms Liang's results, even with varying FR and FC ratios.

Overall, the proposed methodology, which includes feature selection tailored for class imbalance, demonstrates superior performance in terms of accuracy, confusion matrix, and Type I and II errors. This suggests that our approach can effectively improve the performance of SVMs in analyzing complex datasets related to corporate bankruptcy.

4. CONCLUSION

This study introduces a novel feature selection algorithm tailored for SVMs to effectively address class imbalance. The algorithm is applied to the Taiwan company bankruptcy dataset (1999–2009) to identify the three most influential factors contributing to corporate bankruptcy in Taiwan. Unlike most academic studies that primarily aim to enhance prediction accuracy, this research places emphasis on uncovering the key drivers of bankruptcy. Despite the significant class imbalance present in the dataset, the proposed algorithm outperforms existing models, demonstrating strong capability in selecting meaningful features under imbalanced conditions. For instance, the proposed model achieves notable improvements in Type I error (1.17%) and Type II error (22.9%) compared to other studies, which report error rates of 4.4% and 39.4%, respectively. Additionally, while previous research reports an accuracy of 81.3%, the best result achieved by the proposed approach is 83.1%.

Future research could explore whether the influence of the three identified factors is systemic—affecting the broader corporate environment—or situational, depending on specific company or market conditions. Moreover, the proposed feature selection methodology could be applied to other classification algorithms and datasets to assess its generalizability and robustness. Overall, this study contributes to the field by presenting a feature selection approach that not only enhances model performance but also prioritizes interpretability and insight into the underlying causes of corporate bankruptcy.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Borislava Toleva	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Ivan Ivanov	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		
Vincent Hooper	✓			✓	✓		✓	✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available in D. Liang, C-C. Lu, C-F Tsai, and G-A. Shih, “Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,” European Journal of Operational Research, vol. 252, no. 2, pp. 561–572, 2016, doi: 10.1016/j.ejor.2016.01.012.

REFERENCES

[1] S. Guzmán-Castillo *et al.*, “Credit Risk Scoring Algorithm Based on the Discriminant Analysis Technique,” *Procedia Computer Science*, vol. 220, pp. 928–933, 2023, doi: 10.1016/j.procs.2023.03.127.

[2] F. Shen, Y. Liu, R. Wang, and W. Zhou, “A dynamic financial distress forecast algorithm with multiple forecast results under unbalanced data environment,” *Knowledge-Based Systems*, vol. 192, 2020, doi: 10.1016/j.knosys.2019.105365.

[3] E. Mattos and S. Dennis, “Bankruptcy prediction with low-quality financial information,” *Expert Systems with Applications*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121418.

[4] M. Á. Fernández-Gámez, J. Soria, J. Santos, and D. Alaminos, “European country heterogeneity in financial distress prediction: An empirical analysis with macroeconomic and regulatory factors,” *Economic Modeling*, vol. 88, pp. 398–407, 2020, doi: 10.1016/j.econmod.2019.09.050.

[5] C. Lohmann and S. Möllenhoff, “How do bankruptcy risk estimations change in time? Empirical evidence from listed US companies,” *Finance Research Letters*, vol. 58, 2023, doi: 10.1016/j.frl.2023.104389.

[6] Kalak, A. Azevedo, R. Hudson, and M. Karim, “Stock liquidity and SMEs’ likelihood of bankruptcy: Evidence from the US market,” *Research in International Business and Finance*, vol. 42, pp. 1383–1393, 2017, doi: 10.1016/j.ribaf.2017.07.077.

[7] Y. Yuxia, Y. Congyuan, L. Zhiya, and T. Yanting, “Initiative for China to establish a dual algorithm of mixed corporate governance on bankruptcy reorganization: An empirical analysis based on 93 listed companies,” *Heliyon*, vol. 8, no. 12, 2022, doi: 10.1016/j.heliyon.2022.e12007.

[8] H. Mateika, J. Jia, L. Lillard, N. Cronbaugh, and W. Shin, “Fallen angel bonds investment and bankruptcy predictions using manual algorithms and automated machine learning,” *arXiv Preprint*, 2022, doi: 10.48550/arXiv.2212.03454.

[9] J. Zhao, J. Ouenniche, and J. Smedt, “Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction,” *Machine Learning with Applications*, 2024, doi: 10.1016/j.mlwa.2024.100527.
[10] F. Habermann and F. Fischer, “Corporate social performance, and the likelihood of bankruptcy: Evidence from a period of economic upswing,” *Journal of Business Ethics*, vol. 182, no. 1, pp. 243–259, 2023, doi: 10.1007/s10551-021-04956-4.
[11] Z. Li, J. Crook, G. Andreeva, and Y. Tang, “Predicting the risk of financial distress using corporate governance measures,” *Pacific-Basin Finance Journal*, vol. 68, 2021, doi: 10.1016/j.pacfin.2020.101334.
[12] T. M. Alam *et al.*, “Corporate bankruptcy prediction: An approach towards better corporate world,” *The Computer Journal*, vol. 64, no. 11, pp. 1731–1746, 2021, doi: 10.1093/comjnl/bxaa056.
[13] P. Du Jardin, “Forecasting bankruptcy using clustering and neural network-based ensembles,” *Annals of Operations Research*, vol. 299, no. 1, pp. 531–566, 2021, doi: 10.1007/s10479-019-03283-2.
[14] J. Uthayakumar, N. Metawa, K. Shankar, and S. K. Lakshmanaprabu, “Financial crisis prediction algorithm using ant colony optimization,” *International Journal of Information Management*, vol. 50, pp. 538–556, 2020, doi: 10.1016/j.ijinfomgt.2018.12.001.
[15] X. Du, W. Li, S. Ruan, and L. Li, “CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection,” *Applied Soft Computing*, vol. 97, 2020, 10.1016/j.asoc.2020.106758.
[16] A. Dasilas and A. Rigani, “Machine learning techniques in bankruptcy prediction: A systematic literature review,” *Expert Systems with Applications*, vol. 255, 2024, doi: 10.1016/j.eswa.2024.124761.
[17] D. Veganzones and E. Severin, “Corporate failure prediction models in the twenty-first century: A review,” *European Business Review*, vol. 33, no. 2, pp. 204–226, 2021, doi: 10.1108/EBR-12-2018-0209.
[18] A. D. Voda, G. Dobrotă, D. M. Țîrcă, D. D. Dumitrașcu, and D. Dobrotă, “Corporate bankruptcy and insolvency prediction model,” *Technological and Economic Development of Economy*, vol. 27, no. 5, pp. 1039–1056, 2021, doi: 10.3846/tede.2021.15106.
[19] T. Noga and K. Adamowicz, “Forecasting bankruptcy in the wood industry,” *European Journal of Wood and Wood Products*, vol. 79, pp. 735–743, 2021, doi: 10.1007/s00107-020-01620-y.
[20] M. Salehi and M. D. Pour, “Bankruptcy prediction of listed companies on the Tehran Stock Exchange,” *International Journal of Law and Management*, vol. 58, no. 5, pp. 545–561, 2016, doi: 10.1108/IJLMA-05-2015-0023.
[21] J. Almamy, J. Aston, and L. Ngwa, “An evaluation of Altman’s Z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: Evidence from the UK,” *Journal of Corporate Finance*, vol. 36, pp. 278–285, 2016, doi: 10.1016/j.jcorpfin.2015.12.009.
[22] K. S. Ékes and L. Koloszár, “The efficiency of bankruptcy forecast models in the Hungarian SME sector,” *Journal of Competitiveness*, vol. 6, no. 2, pp. 56–73, 2014, doi: 10.7441/joc.2014.02.05.
[23] M. Karas, M. Reznakova, V. Bartos, and M. Zinecker, “Possibilities for the application of the Altman model within the Czech Republic,” in *Recent Researches in Law Science and Finances*, pp. 203–207, 2013.
[24] C-F. Tsai, “Feature selection in bankruptcy prediction,” *Knowledge-Based Systems*, vol. 22, no. 2, pp. 120–127, 2009, doi: 10.1016/j.knosys.2008.08.002.
[25] D. Liang, C-C. Lu, C-F Tsai, and G-A. Shih, “Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,” *European Journal of Operational Research*, vol. 252, no. 2, pp. 561–572, 2016, doi: 10.1016/j.ejor.2016.01.012.
[26] J. Ye, Z. Yang, Y. Zhu, Z. Zhang, and Q. Wen, “Kernel-free quadratic surface SVM for conditional probability estimation in imbalanced multi-class classification,” *Neural Networks*, vol. 188, 2025, doi: 10.1016/j.neunet.2025.107480.
[27] J. Wei, H. Chen, Y. Yuan, H. Huang, L. Wen, and W. Jiao, “Novel imbalanced multi-class fault diagnosis method using transfer learning and oversampling strategies-based multi-layer support vector machines (ML-SVMs),” *Applied Soft Computing*, vol. 167, 2024, doi: 10.1016/j.asoc.2024.112324.
[28] M. Ganaie and M. Tanveer, “K-nearest neighbor weighted reduced universum twin support vector machine for class imbalance learning,” *Knowledge-Based Systems*, vol. 245, 2022, doi: 10.1016/j.knosys.2022.108578.
[29] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, “New imbalanced bearing fault diagnosis method based on sample-characteristic oversampling technique (SCOTE) and multi-class least squares support vector machine,” *Applied Soft Computing*, vol. 101, 2021, doi: 10.1016/j.asoc.2020.107043.
[30] S. Maldonado and J. López, “Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification,” *Applied Soft Computing*, vol. 67, pp. 94–105, 2018, doi: 10.1016/j.asoc.2018.02.051.
[31] scikit-learn developers, “ExtraTreesClassifier,” *scikit-learn documentation*, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html. (Date accessed May 2025).

BIOGRAPHIES OF AUTHORS

Borislava Toleva is a Ph.D. in data science at Sofia University, Bulgaria. She obtained a master’s degree in Statistics, Financial Econometrics, and Actuarial studies in 2015 after a bachelor’s degree in Economics at the same university. Her research areas include practical applications of machine learning algorithms for prediction and how their performance can be boosted. Also, applications of big data techniques to small datasets in the field of economics as alternative to traditional econometrics theory. She challenges traditional econometric modelling techniques used to find connections among variables from institutional economics by combining feature selection methods and big data prediction models. As a result, new applications of machine learning techniques to economic data appear. She can be contacted at email: vrigazova@uni-sofia.bg.

Ivan Ivanov is a Dr.Sc. of Mathematical Studies at Sofia University. Currently, he is head of the Data Science Laboratory at the Faculty of Economics and Business Administration, Sofia University. He is a co-founder of the master's degree in Business Analytics as well as the Ph.D. program in Data Science, both at Sofia University, Bulgaria. He has rich experience in the field of applied mathematics, having a numerous publication recognized by academics. His research interests in the field of machine learning are related to the optimization of algorithms' performance so that their practical advantages might be enhanced. He can be contacted at email: i_ivanov@feb.uni-sofia.bg.

Vincent Hooper 35 years' experience in Tertiary Education at top business schools in UK, Australia, China, Albania, Greece, Malta, Malaysia, Saudi Arabia including Exeter, Nottingham, Richmond, Surrey, Xiamen, ANU, and UNSW, University Campus of Football Business (Wembley Stadium, London). External Examiner for Master's in Oil and Gas. Ph.D. examiner in Australia for top universities. He is fellow of the Higher Education Academy UK. FHEA (2024) and holds a Ph.D. in Multinational Finance from Plymouth University, UK (1994). He can be contacted at email: vincent.hooper@spjain.org.