

Conventional Machine Learning Methods for Detecting Tax Evasion among Inactive or Insolvent Companies

Alexandru SIMION*

Bucharest University of Economic Studies, Bucharest, Romania

*Corresponding author, simionalexandru25@stud.ase.ro

Adrian COSTEA

Bucharest University of Economic Studies, Bucharest, Romania

adrian.costea@csie.ase.ro

Abstract. Previous scientific literature suggests that further research into the tax behaviour of inactive or insolvent companies is required. This paper applied conventional machine learning models rather than advanced ones to maintain explainability and trust in machine learning. One of the challenges of using machine learning to detect tax evasion is the limited availability of private information that can be used as input data. This paper uses public data published by the Romanian Ministry of Finance, the Romanian Trade Register and the Romanian tax authorities. The research questions focus on the preference for supervised versus unsupervised models for tax evasion detection and the statistical significance of the available explanatory variables. Given the significant imbalance in the dataset, this study shows that the performance metrics of classical machine learning models, such as logistic regression and fuzzy c-means, are comparable to those of other models. This paper proposes a two-layer ensemble that is optimised based on an increased weight granted to the recall indicator, with the aim of improving the performance of tax evasion detection. The results show that the business activity code, solvency ratio and company life period are the main explanatory variables of tax evasion in the selected dataset, and that models selected in this research are many times more effective than a random choice. From an economic point of view, this highlights the prolonged survival of inefficient companies, confirms the susceptibility of certain industries to tax evasion, and indicates that financial distress is a driver of tax non-compliance. The contribution of this paper is identifying the main explanatory variables for an automated system for predicting tax evasion among inactive and insolvent companies based on public data. The future emphasis shall be on identifying potential explanatory variables from other sources to improve the performance of robust and explainable machine learning algorithms.

Keywords: machine learning, inactive companies, insolvent companies, tax evasion, DBSCAN, logistic regression, fuzzy c-means.

Introduction

Concerns about tax evasion remain pressing in Eastern European countries and several studies have gathered factors that contribute to this phenomenon and possible policies that could be implemented for an equitable tax system. The difficulty lies in defining what constitutes tax evasion, which is largely left to the provisions transposed or implemented in the local legislation of each jurisdiction. Data mining and machine learning (“ML”) methods are important because they provide a solution to efficiency issues by bringing more structure to the topic. The Romanian tax authorities (“RTA”) already have the infrastructure to manage large amounts of data related to taxpayers, and previous research (Barbu, 2024) shows, for example, how information mentioned in tax receipts from electronic cash registers can be used to detect potential fraudulent cases.

The first question in this paper assesses whether it is preferable to use unsupervised learning methods rather than supervised methods for a dataset with public information. The second question aims to identify the main features that can be used to detect potential cases of tax evasion among inactive or insolvent companies.

The first hypothesis is that unsupervised ML methods are the favoured choice in predicting tax evasion, because obtaining labels for tax evasion is very difficult (e.g., large-scale results regarding tax evasion cases are not publicly available and cannot be incorporated into the research). To test this hypothesis, we first used an unsupervised ML method to determine whether there is any non-random links between outliers and the target group associated to tax evasion. After that, we compared the performance metrics of a different predictive models created from both supervised and unsupervised ML methods. The second hypothesis assumes that explanatory variables extracted from public data are not statistically significant. There are many unknown behavioural factors contributing to tax evasion that are not reflected in companies' public data. Therefore, the initial assumption is that the variables in standard regression models may be subject to significant error and cannot be used to create ML prediction models.

Literature review

In terms of applications for supervised ML models, there are two main types: classification models and regression models. In a classification model, the target variable is binary (i.e. true or false, 1 or 0), whereas in a regression model, the target variable is continuous, although some regression models can be transformed into classification methods (Costea & Eklund, 2003).

One previous research (Charizanos, 2025) proposes a new type of classifier, starting from the Bayesian logistic regression and fuzzy numbers. The author argues that traditional ML methods are significantly impacted by imbalance and separation problems in the data, and that fuzzy Bayesian logistic regression is a better option due to strong classification performance. It should be noted that in this paper multiple datasets have been used and the maximum imbalance between the binary dependent variables was recorded in the dataset for Australian Forest Fires (i.e., $Y=0$ for 92% and $Y=1$ for 8%).

One previous research reviewed Finish limited liability companies that have defaulted on social security contributions or VAT (Höglund, 2017). The dataset in this case included the total number of defaulting companies which were matched with non-defaulting companies on year, industry, and pre-default year total assets. The objective of the study was to customise a genetic algorithm to predict companies that default with outstanding tax liabilities. For this purpose, a variation of linear discriminant analysis (LDA) was used. LDA-based models were considered not to be obsolete, as they can achieve over 90% predictive accuracy.

In Belgium, corporate tax residence fraud has been targeted through a model based on support vector machines and Naive Bayes (Junqué et al., 2014). This model incorporated both structured data (e.g. sector, city, geographical location, industry type, start-up date) and transactional private data (e.g. incoming and outgoing invoicing, invoice logs between Belgian and foreign companies). The selected ML models were evaluated separately and then, in a final step, combined into a stacked model. The authors stated that, since most jurisdictions use audits to ensure compliance with tax laws, accurately selecting the most likely fraudulent cases is crucial to maintaining the efficiency of tax inspections. Their model produced results that were three times better than a random selection.

Another study in Finland (Abedin et al., 2021) argued that previous approaches to tax payment defaults had relied mainly on conventional statistical models, and that advanced data analytics were needed. The authors of that paper found that feature transformation (e.g., logarithmic and square-root transformation) plays a very important role improving the metrics of the ML model. They also found that extreme gradient boosting and multiple decision trees prove to be better in terms of accuracy of the classification model. In their case, the variables that were more important in making the prediction were the equity ratio, liquidity ratio, and debt-to-sales ratio.

Previous research (Pop & Coroiu, 2022) also covers Romanian defaulting companies. The paper included a comparison between multiple ML methods to detect and predict early signs of bankruptcy but mainly focused on artificial neural networks due to increased accuracy. Private data used in that research required an agreement with the Romanian authorities.

One survey-based study made in Romania (Surugiu et al, 2025) used labelled data and a logistic regression to prioritize between variables integrated in the supervised ML model. The emphasis in this paper was investigating tax compliance patterns among individuals in Romania, with the objective to understand factors influencing the behaviour of individuals related to tax conformity in Romania. The results of this research showed a surprising result, namely that higher education levels correlate with lower tax compliance. The authors concluded that the result in their article contradicts previous research stating that education has a positive impact on tax compliance and raised concerns related to targeted interventions and educational campaigns and the need for fiscal transparency in the effectiveness of the administrative system in Romania.

Tax policy concerns in Romania were also raised by another research paper (Coita and Mare, 2021) based on a survey of 520 respondents. The neural network, logistic regression, and other ML models were used to obtain the results. The purpose of that paper was to predict the propensity of any individual to show an intention to evade taxes based on their trust in the governmental administration. The results obtained in the paper by running a binary logistic regression show constraints in applying it to behavioural data. The author preferred the MLP neural network because it showed more accurate results. The conclusion raised concerns about transparency in public spending, clearer tax law followed by legal predictability, and socio-economic development sustained by public spending.

For the purposes of this research paper, the term 'tax evasion' should be understood as referring to instances where this phrase was identified in insolvency reports issued by the Romanian Trade Registry. In some cases, the phrase 'tax evasion' may have been used in relation to matters under investigation or at different stages of the process, as set out in the local procedures of authorities. Therefore, the term 'tax evasion', as used in this paper, should not be taken to mean that any of the companies or their shareholders have been definitively sentenced. The dataset included in this paper reflects the 2025 status of potential ongoing cases, which are updated regularly by the relevant authorities, resulting in subsequent changes.

Considering the gap in the literature regarding the tax behaviour of inactive companies, this paper takes a new approach by using a public dataset and compares multiple conventional ML methods for tax evasion detection purposes. Previous research in Romania has focused on surveys or local in-depth analysis of private data, which is difficult to obtain in the case of taxpayers. This research takes a new perspective by managing the limitations of using public data in Romania, the trade-offs embedded in supervised and unsupervised ML models, and the ethical concerns raised in the financial sector.

Methodology

This research paper applies multiple ML methods on a dataset obtained from three different public sources (i.e., financial statements from the Orbis database, insolvency reports from the Romanian trade registry and outstanding tax obligations published by the RTA as of 31.12.2024). The software used was Python (Scikit-Learn modules).

To test the first hypotheses, the unsupervised ML method density-based clustering algorithm (“DBSCAN”) has been used. This method focusses on clustering items in the dataset based on the distance between each point and the core points identified in that dataset. Conversely, the classifiers shown in Figure 1 have been used to test both hypotheses.

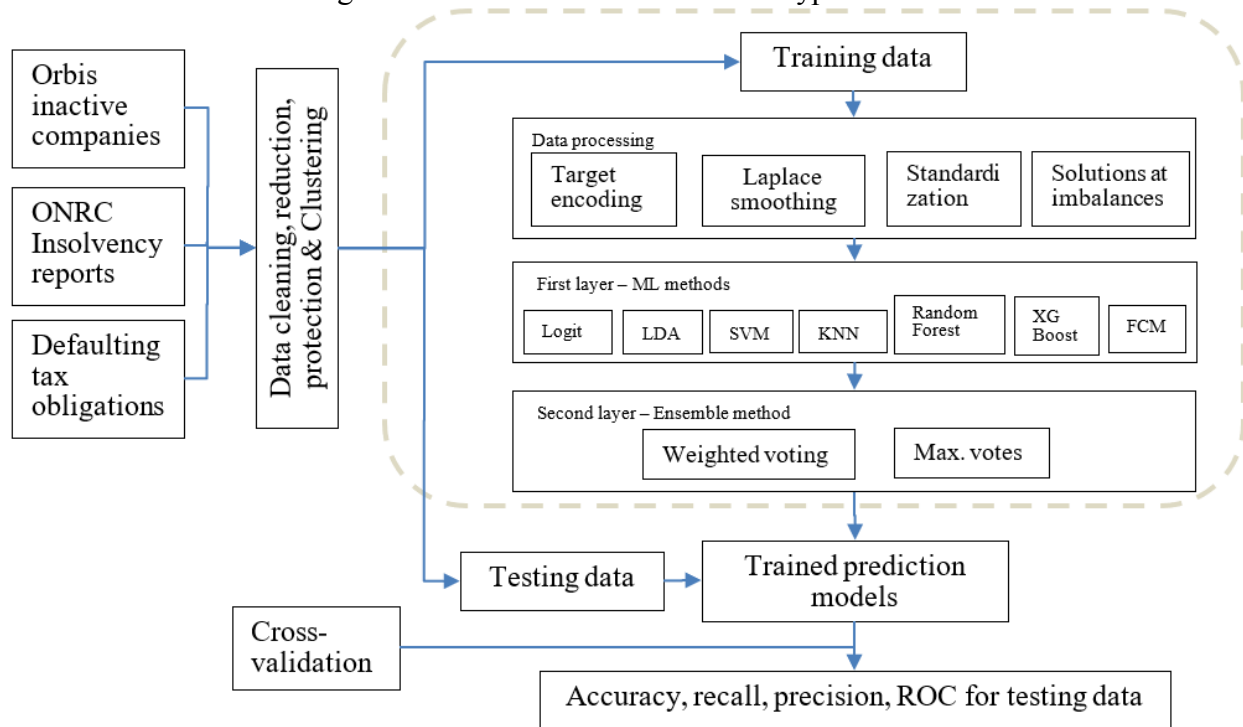


Figure 1. Visual representation of ML prediction method

Source: own research and structure drawn from Abedin et al, 2021.

The imbalanced dataset produced a Fisher Discriminant ratio of 0.4963, indicating that the minority class is cohesive. However, a nearest-neighbour non-linearity ratio of 0.0178 suggests that features are not distinct enough to distinguish between the classes. The appropriate resampling technique suggested in previous research (Carvalho et al, 2025) was the synthetic minority over-sampling and data cleaning with edited nearest neighbours (i.e., SMOTE-ENN).

The first layer includes the following ML methods: logistic regression (Logit), the linear discriminant analysis (LDA), support vector machines (SVM), K-nearest neighbour (KNN), random forest (RF) and extreme gradient boosting (XG Boost, in line with the previous research paper of Abedin et al, 2021).

Applying the Fuzzy C-Means (“FCM”) model

The FCM performs clustering based on a membership percentage allocated to each item. Each item has a membership degree to every cluster, and the sum of all membership percentages is 1. Previous research on FCM (Demirhan, 2025) shows that this method can be used also for categorical

variables (nominal or ordinal scale). In our specific case, we have a binary dependent variable, which denotes the existence or absence of a reference to “tax evasion” in the insolvency reports published by the Romanian trade registry.

Previous authors (Costea & Eklund, 2003) state that C-Means is a statistical clustering technique with the goal to minimize the sum of the variances within clusters. The objective function is defined as:

$$J = \sum_j^C \sum_{i \in S_j} \|x_i - c_j\|$$

The abbreviations used for the C-Means are the following:

- $\| \cdot \|$ is the measure of similarity, such as the Euclidian distance
- C represents the two clusters (1= tax evasion, and 0 = no tax evasion)
- x_i is the value of observation i
- c_j is the centroid for cluster j

The algorithm involved in the C-Means has the following steps:

- i. Define the number of clusters c_j
- ii. Select C observations as the centres for the C predefined clusters
- iii. Assign each observation from the dataset to the closest cluster based on the Euclidian distance
- iv. Recompute the centres of the clusters either after all assignments
- v. Repeat steps iii. and iv. until the number of maximum iterations has been reached (i.e., when the centroids that were already set do not change their position in a significant manner).

FCM, on the other hand, is an algorithm that can handle a certain degree of uncertainty in grouping items in the dataset. This allows for degrees of truth, implying that there are values between 'true' and 'false'. Therefore, FCM algorithms assign different membership degrees to the elements in the dataset. The objective of the FCM algorithm is represented by:

$$J(U, v) = \sum_{k=1}^n \sum_{i=1}^C (u_{ik})^m (d_{ik})^2$$

The abbreviations used for the FCM are the following:

- n is the number of data points (i.e., in our case 10,717 lines in the full dataset)
- C represents the two clusters
- U represents the fuzzy C-partition of the dataset X, where $U_{ik} \in [0,1]$ is the membership degree of observation x_k in cluster i, with $\sum_{i=1}^C u_{ik} = 1$, or $u_{i1} + u_{i2} = 1$ in the case of two clusters
- d_{ik} is the Euclidian distance represented as $d_{ik} = \|x_k - v_i\| = \sqrt{\sum_{j=1}^p (x_{kj} - v_{ij})^2}$
- v_{ij} is the centroid for cluster i and variable j, with p-number of input variables
- m represents the weighting component $m \in [1, \infty)$ that controls the extent of membership-sharing between the clusters. When $m \rightarrow 1$, then FCM converges to the C-Means, while when $m \rightarrow \infty$ the centres tend towards the centroid of the dataset (i.e., the centres tend to be identical).

- The FCM algorithm has the following steps:
- Initialize by providing the number of clusters (i.e., value of C), and weight (i.e. $m=2$ in our specific case). $U^{(0)}$ is also provided at initialisation phase.
 - Compute the C cluster centres $v_i^{(S)}$ at each iteration S , based on $U^{(S)}$ representing each iteration.
 - Calculate $U^{(S+1)}$ based on $v_i^{(S)}$
 - Compare $U^{(S+1)}$ to $U^{(S)}$: if $\|U^{(S+1)} - U^{(S)}\| < \varepsilon$ stop; otherwise return to step ii.

Once the centroids have been determined, along with the distance from each data point to its respective centroid, the membership degree can be calculated for each data point and cluster. The membership degrees are updated at each iteration until the distances are below the agreed threshold (i.e., ε).

Results and discussions

DBSCAN results for outlier detection

The outlier count is presented in Table 1. The Chi-square test of the dataset shows that real observations marked with tax evasion (i.e., $\text{Target}=1$) are more frequent in the outlier group obtained by the DBSCAN method than in any of the other clusters. Considering that DBSCAN is an unsupervised ML method, the results have a descriptive nature.

Table 1. Observation count obtained with DBSCAN method

Metrics	<i>Outliers</i>	<i>Inliers</i>
Total samples:	375 (3.5%)	10,342 (96.5%)
Target=1 among outliers	44 (11.7%)	71 (0.7%)
Chi-square p-value	$3.22 * 10^{-90}$	

Source: Authors' own research, DBSCAN using entire dataset.

The visual representation of the dataset in Figure 2 was made using the principal component analysis method, to reduce the number of dimensions.

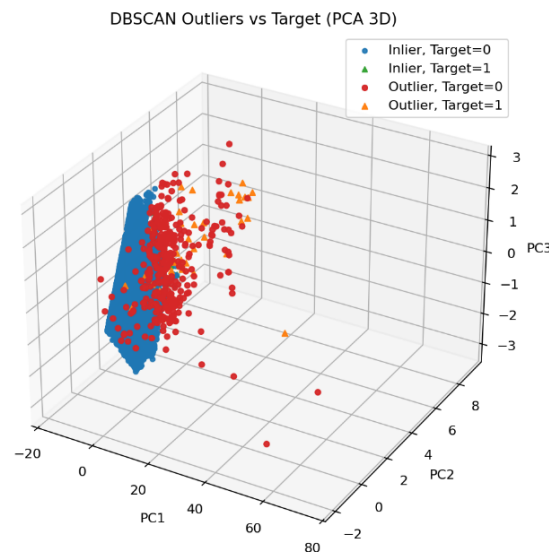


Figure 2. Visual representation of DBSCAN results for the entire dataset of 10,717 samples

Source: Authors' own research, DBSCAN using entire dataset.

The external validation method with matching sets has been used (Palacio-Niño, 2019, P:7) to apply metrics valuations that are usually specific to supervised ML models in the appraisal of an unsupervised ML method. As we can see in Table 2, although the overall weighted average metrics of the model are very high, the prediction does not include any items targeted with tax evasion. This means that the model mainly predicted inliers, with no possibility of detecting real tax evasion cases among inactive or insolvent companies.

Table 2. Performance metrics of predictive model based on DBSCAN

Metrics @ DBSCAN		Precision	Recall	f1-score	accuracy	Count
Outlier detection	Binary 0	98.93%	100.00%	99.46%	98.93%	2,121
	Binary 1	0.00%	0.00%	0.00%		23
	Macro average	49.46%	50.00%	49.73%		2,144
	Weighted average	97.87%	98.93%	98.39%		2,144

Source: Authors' own research, DBSCAN using only the test set.

Although statistical significance was identified in the DBSCAN model, the relation between the outlier group and the group labelled with tax evasion was not strong enough to generate valid predicted results for this specific dataset. The results of this paper are in line with previous research papers (Kleinberg, 2019) that suggest difficulties in applying performance metrics for clustering methods.

Logistic regression results

The statistical results of the logistic regression analysis for all five explanatory variables and the dependent variable relating to tax evasion are presented in Table 3. Logistic regression can be considered a representative of supervised ML models, which are used to determine effectiveness compared to unsupervised models such as DBSCAN and FCM.

Table 3. Statistical metrics of the logistic regression

Feature	Coefficients	Standard error	p-value	Ci 95 low	Ci 95 high
Constant	-5.590	0.171	0.000	-5.926	-5.254
Nominal variable (NACE code)	31.957	2.427	0.000	27.201	36.714
Company life period duration	0.275	0.102	0.007	0.074	0.475
Solvency ratio (equity /assets)	-0.744	0.106	0.000	-0.953	-0.536
Outstanding tax obligations / assets	0.080	0.052	0.126	-0.022	0.182
Outstanding tax obligations / equity	0.009	0.090	0.919	-0.168	0.186

Source: Authors' own research: coefficients, standard error, p-value and confidence intervals.

The p-value probability of random chance for three out of five coefficients is below the 5% threshold, and the null hypothesis is partially rejected. The coefficients related to the amount of tax obligations are not statistically significant.

From an economic point of view, positive coefficients for NACE show that certain industries are susceptible to tax evasion in Romania (e.g., construction, accommodation, farming, wholesale trade, wood manufacturing, etc.). Due to their inherent nature, these business activities provide the tax authorities and policymakers with multiple insights that could be used to improve internal procedures and legal provisions.

The positive life period coefficient denotes that companies with tax evasion have been allowed to remain inactive or insolvent for a longer period of time. Policy implications might include targeting the prolonged survival of inefficient companies, imposing corrective fines, holding company representatives directly accountable, and increasing the focus on tax morale.

The negative coefficient for the solvency ratio indicates that as a company's solvency ratio decreases, the probability of it engaging in tax evasion increases. One policy implication might be to use solvency metrics to filter tax audits and differentiate between companies that evade taxes to prevent bankruptcy and those that engage in aggressive tax evasion.

The performance of the standard model of the logistic regression may be sub-optimal in the case of imbalanced datasets. The standard model of the logistic regression has a decision threshold of 0.5, that can be interpreted as the limit considered by the algorithm to distinguish between two groups based on the probability of each observation to belong to each of the two groups.

Table 4. Performance metrics of logistic regression with five explanatory variables

Threshold @ logistic regression		Precision	recall	f1-score	Accuracy	Count
Optimized F1-score: 0.9261	binary 0	98.97%	99.81%	99.39%	98.79%	2,121
	binary 1	2.00%	4.35%	7.14%		23
	macro average	59.49%	52.08%	53.27%		2,144
	weighted average	98.12%	98.79%	98.40%		2,144

Source: Authors' own research.

One option to optimize the logistic regression is to substitute the 0.5 threshold with the one used for obtaining improved performance metrics (e.g., the F1 score as presented in Table 4). The F1 score is typically used in imbalanced data scenarios.

Unsupervised FCM results

To gain a better understanding of the training data, it was necessary to visualise the centroids (Figure 3). The variables used in the model are the explanatory variables identified as statistically significant in the logistic regression. If the explanatory variables are not scaled, the largest-scale feature would dominate the distance calculation. Therefore, each explanatory variable was transformed to have 0 mean and a standard deviation of 1 (i.e., z-score normalization or standardization). As it can be seen in the graph below, the centroid of the tax evasion group (i.e. the group marked with red points) can be distinguished from the main group (marked with blue).

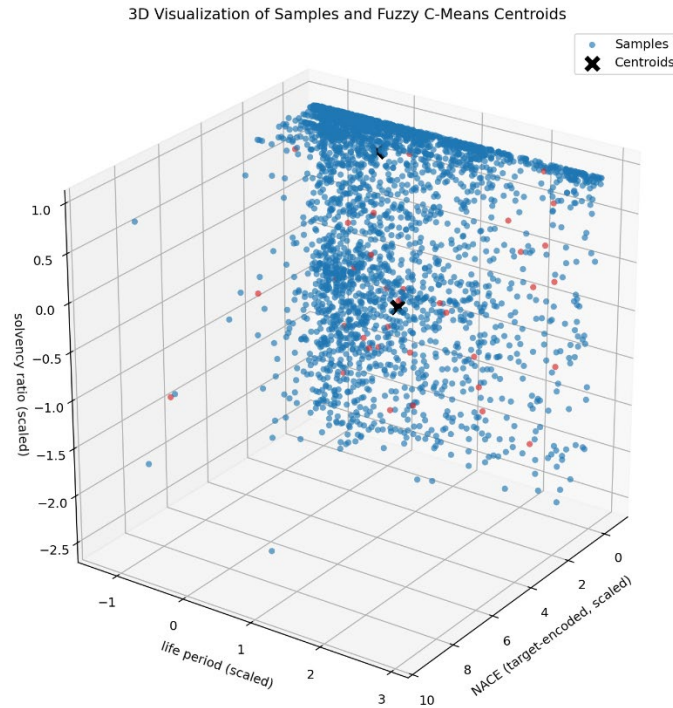


Figure 3. Training data with FCM centroids showing red dots associated with tax evasion

Source: Authors' own research.

The FCM is mainly an unsupervised ML method that can perform the clustering of the dataset provided as input. Initially, the FCM grouped the data points into two clusters according to the natural geometry of the observations. A binning process was performed with five linguistic terms: very high (VH), high (H), medium (M), low (L), and very low (VL). Similar to previous research (Costea & Eklund, 2003) the linguistic terms were obtained by applying FCM to each individual vector represented by explanatory variables, containing only certain observations. If the frequency of one or two linguistic terms exceeded the 33% threshold, this indicated that they characterised the cluster.

Table 5. Explanatory variables discriminating between FCM clusters

Binning @ FCM	Life period	Solvency ratio	Tax obligations to assets	Tax obligations to equity	NACE	Cluster type
Cluster 1	<i>M-L</i>	<i>L</i>	<i>VL</i>	<i>L</i>	<i>L</i>	<i>Tax evasion</i>
Cluster 0	<i>L</i>	<i>VH</i>	<i>VL</i>	<i>L</i>	<i>VL</i>	<i>Standard</i>

Source: Methodology drawn from Costea & Eklund, 2003

The cluster type is determined by the number of positives in the explained variable. The unsupervised FCM clustering method matches the allocation based on the number of positives in the explained variable (tax evasion is represented by a lower solvency ratio, longer life period and higher frequency of business activity codes associated with cases of tax evasion). Results on the binning process are similar to previous research (Costea & Eklund, 2003) and show that FCM cluster types can be determined without using the explained variable.

From an economic point of view, the unexpected result was that the outstanding tax obligations did not clearly discriminate between the clusters. This suggests that the existence or

value of outstanding tax obligations is not a key factor in determining whether a company is in the tax evasion group. Therefore, having unpaid fines or tax obligations following a tax audit may indicate that a company is inactive or insolvent, regardless of whether it is part of the tax evasion group.

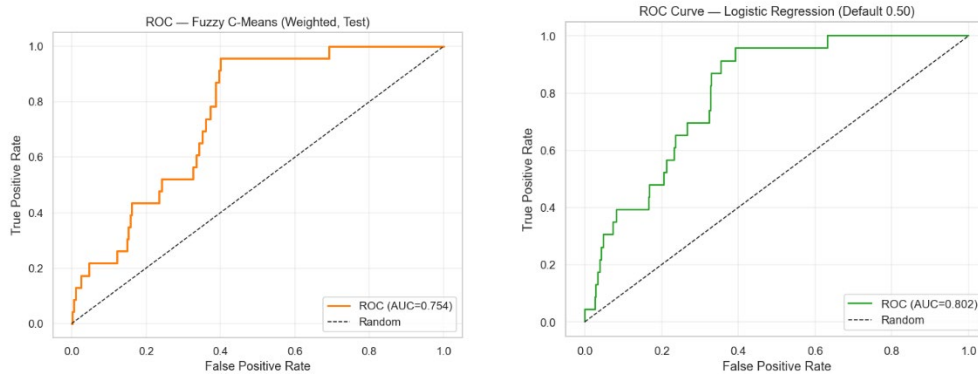


Figure 4. Receiver operating curve from binary classification results of FCM and Logit

Source: Authors' own research.

Previous research papers (Belkin, 2019) underlined the issue with bias within the data, including overfitting problems, an issue managed with cross-validation in this research. After applying cross-validation, the logit and FCM models have similar receiver operating curves.

Results of the two-layer ensemble

The ML methods specified in the second layer of Figure 1 have been optimised using the F2-score, which balances precision and recall. The F2-score gives twice as much weight to the recall indicator, enabling the model to identify as many positive cases as possible.

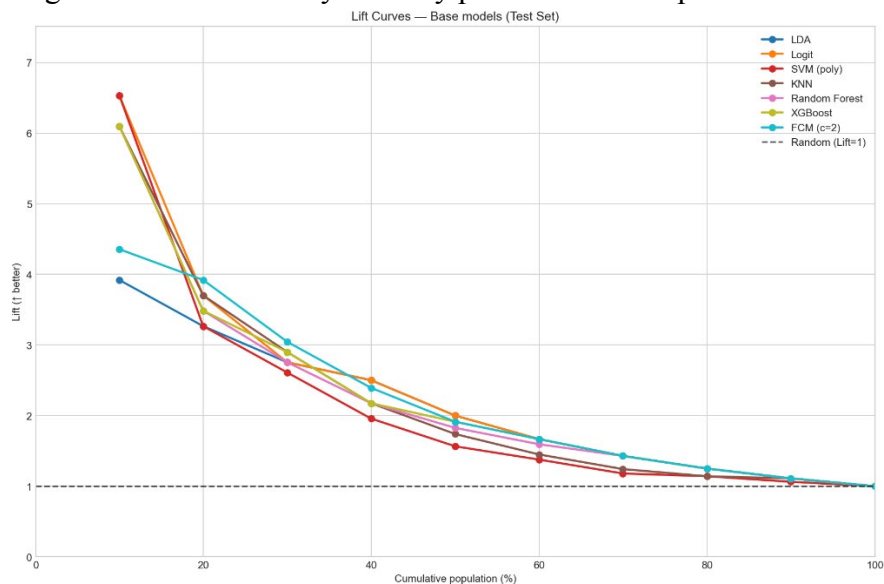


Figure 5. Lift curves for classification ML methods optimized with F2-score

Source: Authors' own research.

Due to the increased costs for RTA associated with false negative predictions, the recall indicator has become more significant. As shown in Figure 5, the focus is on the lift of each model, i.e. how much better it is than random (represented by 1 on the lift axis), when targeting the highest-ranking members. This is because tax evasion analysis investigates companies deemed to be the most suspicious. The results show that the supervised ML models selected in this research are many times more effective than a random choice. Since FCM performs unsupervised clustering based on natural allocation of observation, its predictions may not capture positives detected by supervised ML models.

The binary predictor variables obtained from each ML model in the first layer were used as input variables for the ensemble. The rules presented in Table 6 outline the two options considered in the ensemble: cumulating positive results from the first-layer ML models and applying a voting mechanism based on the F2 scores obtained by the first-layer ML models.

Table 6. Performance metrics of two-layer ensemble on F2-score optimization

Two-layer ensemble	Accuracy	Precision	Recall	F1 score	F2 score
Max. votes (OR strict rule)	69.82%	3.01%	86.96%	5.82%	13.23%
Weighted voting (F2)	90.44%	7.08%	65.22%	12.77%	24.67%

Source: Authors' own research.

The results of the two-layer ensemble show that the ensemble model has comparable performance metrics as the initial logit model, with improvements in terms of the recall indicator. Although there is the disadvantage of multiple false positives, those instances can be filtered by a team of tax specialists. These results are consistent with previous research (Junqué et al., 2014), which demonstrated that industry type and start-up date could be used to detect tax fraud. Given the potential for subsequent filtering of items associated with a risk factor, as well as the ethical concerns related to complex ML systems, the two-layer ensemble could be considered a good starting point.

Conclusion

Regarding the first hypothesis, supervised ML methods proved to be more effective in predicting tax evasion among inactive or insolvent companies than unsupervised methods. However, the unsupervised FCM model showed similar performance metrics when the imbalanced dataset was adjusted with SMOTE-ENN. A two-layer ensemble offers slight improvements compared to individual classification methods. The second hypothesis was rejected because three of the coefficients used in the logistic regression analysis (i.e. the NACE code, the solvency ratio and the company life period) were found to be statistically significant.

The paper's main contribution is identifying features that can be used to predict tax evasion among inactive or insolvent companies, as well as the comparison between multiple conventional ML models to detect tax evasion among inactive or insolvent companies using information from public sources. The results suggest the existence of a prolonged survival of inefficient companies, confirm that certain industries are susceptible to tax evasion and show that financial distress acts as a driver for tax non-compliance. The findings have policy implications emphasizing the need to integrate solvency metrics into risk-scoring algorithms. By using early intervention mechanisms for distressed companies and streamline the liquidation of inefficient companies, regulators can mitigate market distortions and reduce the gap of lost public revenues.

Research papers on tax behaviour are important because they can be used by both the tax authorities and taxpayers. Possible applications of this research for tax authorities are to identify potential targets for tax audits before a company is liquidated and the probability of recovering tax liabilities is reduced. Taxpayers can use the research to review future collaborators internally, such as client onboarding procedures and supplier selection. There are at least two possible extensions to the current research paper. First, early signs of outstanding tax liabilities could be identified through benchmarking studies based on industry sectors, similar to previous research papers (Höglund, 2017 and Pop & Coroiu, 2022). Second, explanatory variables of VAT code annulment could be identified by extending the methods used in the current research.

One important limitation of this paper is the fact that explanatory variables from both private and public sources could be added to the dataset in order to improve performance metrics. Another limitation of this research is related to incomplete company information extracted from public sources. As future directions for this research, a better integration with the primary source of information would be necessary to identifying other explanatory variables for robust and explainable ML algorithms.

This paper was co-financed by the Bucharest University of Economic Studies during the PhD program.

References

- Abedin, M. Z., Chi, G., Uddin, M. M., Satu, Md. S., Khan, Md. I., & Hajek, P. (2021). Tax Default Prediction Using Feature Transformation-Based Machine Learning. *IEEE Access*, 9, 19864–19881. <https://doi.org/10.1109/ACCESS.2020.3048018>
- Barbu, D.-C., Bâra, A., & Oprea, S.-V. (2024). Impact of Electronic Cash Registers on Tax Collection. In *Proceedings of 22nd International Conference on Informatics in Economy (IE 2023)* (pp. 25–37). Springer. <https://doi.org/10.1007/978-981-99-6529-8>
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>
- Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, 12(1). <https://doi.org/10.1186/s40537-025-01119-4>
- Charizanos, G., Demirhan, H. and İcen, D. (2025) “Binary classification with Fuzzy-Bayesian logistic regression using Gaussian fuzzy numbers,” *Intelligent Systems with Applications*, 26. Available at: <https://doi.org/10.1016/j.iswa.2025.200494>
- Coita, I. F., & Mare, C. (2021). The Utility of Neural Model in Predicting Tax Avoidance Behavior. In *Smart Innovation, Systems and Technologies* (Vol. 238, pp. 71–81). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-16-2765-1_6
- Costea, A. and Eklund, T. (2003) “A two-level approach to making class predictions,” in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. IEEE*, p. 9 pp. Available at: <https://doi.org/10.1109/HICSS.2003.1174207>
- Demirhan, H. (2025) “Mixed fuzzy C-means clustering,” *Information Sciences*, 690, p. 121528. Available at: <https://doi.org/10.1016/j.ins.2024.121528>

- Höglund, H. (2017). Tax payment default prediction using genetic algorithm-based variable selection. *Expert Systems with Applications*, 88, 368–375. <https://doi.org/10.1016/j.eswa.2017.07.027>
- Junqué De Fortuny, E., Stankova, M., Moeyersoms, J., Minnaert, B., Provost, F., & Martens, D. (2014). Corporate residence fraud detection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1650–1659. <https://doi.org/10.1145/2623330.2623333>
- Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. *arXiv*. <https://doi.org/10.48550/arXiv.1905.05667>
- Pop, I. D., & Coroiu, A. M. (2022). Predicting bankruptcy in Romania using artificial neural network. *International Journal of Modern Manufacturing Technologies*, 14(3), 211–218. <https://doi.org/10.54684/ijmmt.2022.14.3.211>
- Kleinberg, J. (2019). An Impossibility Theorem for Clustering. *Neural Information Processing Systems: Proceedings of the 16th International Conference on Neural Information Processing Systems*. <https://dl.acm.org/doi/10.5555/2968618.2968676>
- Surugiu, M.-R., Vasile, V., Surugiu, C., Mazilescu, C. R., Panait, M.-C., & Bunduchi, E. (2025). Tax Compliance Pattern Analysis: A Survey-Based Approach. *International Journal of Financial Studies*, 13(1), 14. <https://doi.org/10.3390/ijfs13010014>