sciendo

A Quantitative Analysis of Default Risk Using Machine Learning and SHAP Value Interpretation

Coralia TANASUICA (ZOTIC)*
Academy of Economic Studies, Bucharest, Romania
Doctoral School of Economic Cybernetics and Statistics (SDCSE)
**Corresponding author, tanasuicacoralia22@stud.ase.ro*

Abstract. *In finance, creating a model that balances risk reduction with opportunity is essential. This investigation addresses the necessity for risk evaluation frameworks that combine efficiency with adaptability, thus preserving opportunities for transactions critical to some organizations. The present study identifies, within a factoring process involving two key players: the invoice seller and the debtor, the essential variables that determine the likelihood of the debtor defaulting on the invoice payment. The event of non-payment is most often associated with the debtor's inability to pay due to insolvency, making it crucial in this type of activity to emphasize an efficient credit scoring system capable of proactively highlighting a debtor company with a high risk of default. Nonetheless, some companies pass this filter and enter the factoring process, but end up being unable to pay. The study identifies them based on a set of real data and uses supervised machine learning techniques to select the optimal classification model, also highlighting the variables with a major impact on the target. The specialized literature is focused on identifying the models that perform best in the credit scoring activity or studies that identify the non-payment behavior of clients. What this work adds is the combination of these two dimensions, for example, it provides an additional filter to credit scoring, using parameters identified as essential in determining defaulters and using them as inputs for an unsupervised learning model, thus classifying the entire population of companies in Romania to identify clusters containing the highest proportion of non-payment companies.*

Keywords: Machine learning, Model interpretability, Payment Default Prediction, Clustering, Propensity to pay, Bad debt, Analytics

Introduction

How can be the credit risk model improved, so that it will be able to identify the specific subset of companies that, even though they were scored with low probability of defaulting, they finally did not pay their bills at the due date? The nonpayment behavior was determined by the fact that they became insolvent till the due date of the bills.

Factoring is an important financial service that is specifically designed to assist the company's lack of cash by providing immediate access to funds. The process of factoring offers the possibility to a company that needs cash, to sell its outstanding invoices to the factory company at a discounted rate in exchange for immediate funding. This practice gives the seller the needed cash and reduces the risk of non-payment, with this risk transferring to a third party.

The direction chosen was to construct an ensemble of models that, in addition to a credit scoring model based solely on the financial information of companies from their balance sheets and financial statements, could also incorporate insights obtained from analyzing customer non-payment behavior.

The case study was done on a sample of factoring transactions, of which 93% were successfully concluded with the invoices being paid, and 7% resulted in bad debt, with the invoices not paid and the transactions leading to insurance claims. Considering that the two classes, of good and bad payers, are strongly imbalanced, the need for using robust algorithms suited for imbalanced

DOI: 10.2478/picbe-2024-0020

data samples appeared. Boosting methods are known as being able to deal with imbalanced data as these tend to give more attention to minority classes. Lastly, in evaluating the models, the performance metrics used to test the accuracy of the models, were the once appropriate for imbalanced classes, such as the F1 score. The algorithms were tested on both raw and preprocessed data and using the Principal Component Analysis (PCA) technique, to explore all possible directions, having the final scope of maximizing the accuracy of the generated models. The supervised machine learning models were tested using the grid search option to identify the optimal model parameters for each model. The aim was to identify the model that best classifies the two population categories, from the perspective of the performance indicators analyzed.

The model selected as optimal was analyzed to identify which variables have a major impact on the debtor's final action within the factoring process, whether to pay the invoice or not. Once these variables were detected, they became the starting point for segmenting the entire active company population in Romania, aiming to identify those population segments statistically comparable to the two population samples analyzed with the help of the classification model. Subsequently, statistical methodologies were employed to ascertain whether there exists a similarity between the bad debt sample within the dataset and the cluster generated through the application of unsupervised machine learning models at the general level of all active companies. This approach essentially extends the analysis of local behavior to a macro level. Essentially, the segmentation model allows for additional filtering, successfully complementing the financial credit scoring model across the entire population of companies in Romania.

The research contains five chapters, each representing bricks for building an architecture of models. The final scope was to improve the overall performance of the credit scoring model and to better detect the defaulting companies. The second chapter of the article provides a summary of the relevant literature, focusing primarily on machine learning models used for estimating the probability of a company entering into insolvency, meaning that it is not capable anymore to pay its bills. The sections following contain a description of the methodologies and the models used in the case study and present the conclusions of this research, its limitations, and ideas for further development, followed by the final section containing the bibliographic resources used.

Literature review

Studies broadly focus on identifying models with the highest accuracy in scoring companies and individuals entering a financing process or identifying the non-payment behavior of borrowers. Some of these studies continue to identify factors with the most significant influence on the targeted outcome. One such study is focusing on predicting recovery rates for non-performing loans. It finds that AI rule-based algorithms, particularly Cubist, boosted trees, and random forests, significantly outperform other methods (Bellotti, 2021). The research brings a novelty in modeling the credit risk, by the usage of unconventional data such as socio-demographic data, loan information, and bank recovery history, in addition to traditional financial information and economic indicators. The methods tested in this article include linear and nonlinear models such as multivariate adaptive regression splines (MARS) and support vector machines, as well as rule-based models like regression trees and random forests. The data used by the authors come from a European debt collection agency. Their goals were to detect an optimal model for determining the recovery rate and to identify factors that significantly impact this rate.

In the same direction of highlighting the importance of model explainability in conjunction with its accuracy, some authors employ specific methods for visually identifying variables that impact the model's final output, such as the SHAP library (Ariza, 2019) from Python. The study

compares logistic regression with machine learning algorithms like decision trees, random forests, and XGBoost. The primary focus is on the performance and explainability of these models.

Other authors focus on a specific model and are trying to decode it, focusing on the model's interpretability (Carmona, 2022). The XGBoost model is used for predicting business failure, and is known as one of the high-performance boosting models. Utilizing data from 1760 French firms (1585 healthy and 175 failing) in 2018, the study identifies key indicators such as equity per employee, solvency ratio, current ratio, net profitability, and sustainable return on investment as critical for predicting business failure.

Beyond research focused on numeric data, there are also works attempting to identify the essential factors leading to non-payment behavior using exclusively qualitative data (Nst, 2023). The research detects the internal factors (weak credit administration, supervision, and information systems, as well as irregularities or fraud during the credit process) and also the external factors (economic downturns, business failures, and absconding partners) that contribute to the increase in bad debts.

Another segment of studies focuses on the use of new and innovative methods in attempting to identify the model that performs best in scoring defaulters. One such study presents a model combining the extreme learning machine (ELM), fuzzy c-means (FCM) algorithm, and confusion matrix for scoring credit quality (Pang, 2021), predicting default probabilities, and calculating default loss rates. It evaluates 7706 borrowers from Renren loans, achieving a high overall accuracy of 98.5%. This approach is novel and illustrates the potential of ML algorithms in improving credit quality assessment practices. The approach for detecting bad debt using fuzzy logic principles to handle uncertainty is also found in another work, where authors use an information-based fuzzy-partitioning method (Li, 2021), demonstrating the advantages of integrating fuzzy logic and domain expertise in financial risk assessment.

The same approach, of identifying the model that performs best in insolvency risk identification on exclusively financial data, is found in other authors' approaches who identify the Catboost model as performing best on these types of data (Jabeur, 2021). The study compares CatBoost with eight other machine learning models, emphasizing the importance of feature selection and the interpretability of machine learning predictions. The authors collected financial data from French firms, focusing on a variety of financial ratios for analysis. Another boosting algorithm with very good results is LightGBM (Li, 2021; Dong, 2022) which is used for its high performance when data are imbalanced.

Another innovative approach to identifying default is the use of graph theory to analyze trading interactions between companies (Yıldırım, 2021). This approach allows for expanding the model from using exclusively financial data to the possibility of including alternative indicators that can have a significant impact on a company's non-payment behavior. Using a comprehensive dataset from Turkish companies, the study demonstrates improved prediction accuracy through these new dimensions included.

The use of hybrid models, both supervised and unsupervised ML, is a less common approach in the literature, identifying that hybrid models, particularly those combining k-Means with decision trees or random forests, significantly outperform individual models in predictive accuracy, offering a novel approach for financial institutions to evaluate commercial credit risk (Machado, 2022). Another perspective on hybrid methods is represented by integrating expert knowledge and genetic algorithms for feature selection in credit risk assessment (Lappas, 2021). This framework aims not to improve machine learning performance per se but to incorporate expert

insight into the credit scoring problem, offering another perspective on feature selection and optimization in financial risk assessment.

Numerous studies analyze various algorithms to identify the one that performs best in predicting the risk of bankruptcy or bill non-payment (Saurabh, 2022; Megistu, 2023; Đukanović, 2023).

Another segment of studies focuses on the use of ML algorithms for profiling customers from the perspective of bill payment and identifying the probabilities that they will pay (Bashar, 2023). The research, leveraging a dataset from an Australian energy company, aims to predict customers' likelihood to pay energy bills on time. It integrates traditional and Bayesian neural networks (BNN) with decision trees and logistic regression to handle prediction under uncertainty.

While there are attempts to hybridize the algorithms used, or rather the customer scoring process, by combining methods and methodologies or by extending the input parameters from exclusively financial ones to alternative variables, there is still ample research space. This is especially true as each geographic region and field of activity has its specificity. This can mean that certain methods and methodology architectures yield very good results in a specific environment and perform less well under different input conditions.

This paper aims to unify previously presented research directions: explainability and default propensity, and to extend the identified behavior through unsupervised machine learning to the entire population. Specifically, it begins with the default behavior of companies scored using ML algorithms and identified as having a low bankruptcy risk, i.e., companies with a very good credit score that still end up in default. This situation is prompted by the initiation of insolvency proceedings for these companies. The goal was to identify the most critical parameters in defaulters' identification. Once these parameters were identified, they were used as input variables for a segmentation model for the entire population of active companies in Romania. Upon generating the clusters, the initial credit scoring model could be adjusted, considering clusters that identify potential defaulters, even if the initial model did not identify them as such.

Methodology

This research is based on testing various Supervised Machine Learning (SML) algorithms to identify the optimal classification algorithm but also Unsupervised Machine Learning (USL) algorithms for segmenting the entire population of companies in Romania and identifying which is the comparable cluster to a sample of defaulting companies. An important technique used during the algorithms training phase was the so-called Grid Search technique. It simulates an entire range of possible model parameter combinations having the final scope of identifying the optimal combination of parameters for which the model produces the highest accuracy metrics. What does this technique consist of? It consists of building a matrix of parameters for the model, then training, and evaluating the performance of the model for each possible combination of parameters. After all the parameter combinations were tested, the Grid Search provides the set of parameters that produce the best model performance. At the end of the process, the model with the best-found parameter combination is trained using the entire dataset.

The SML algorithms trained and tested for the classification stage were: Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, AdaBoost, LightGBM, CatBoost, XGBoost, K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes. In the final stage of segmenting the active company base, the K-means algorithm was used. The programming language used for building and training the models was Python. In the stage of

identifying factors impacting the output of the classification model, the LIME and SHAP libraries were used.

Each algorithm used in the current research has its defining characteristics, which I will shortly present in the following rows.

The logistic regression was one of the algorithms chosen to be tested for the binary classification of the non-payer versus good payers. This algorithm only digests numerical data, but it can be trained on row data, not necessarily on normalized or standardized data. The quality of the input data directly affects the model. For data to be of good quality the noise in the data must be eliminated. The logistic regression uses as the core function the sigmoid function:

$$\sigma_{(z)} = \frac{1}{1+e^{-z}} \quad (1)$$

where z is the linear combination of the input variables and their respective coefficients:

$$z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \quad (2)$$

SVM is another powerful algorithm for both classification and regression. The scale of the variables directly affects the algorithm, transforming it into a sensitive method concerning the standardization of data. The algorithm's strategy is to identify a hyperplane that best separates the data into classes. The support vectors are essential in defining the margin, which is the distance between the hyperplane and the support vectors. they are the data points that are closest to the hyperplane. The "kernel trick" is an essential part of the algorithm. It is a kernel function, that permits the SVM algorithm to make classification in high-dimensional spaces. It helps in finding the separating hyperplane even for data that is not linearly separable in the original space.

Decision Trees are algorithms that are easy to interpret with good performances both on unstandardized input data and on standardized data. The core flow of this category of algorithms is that they use binary questions for dividing the feature space into distinct regions. The node in the tree represents a condition on a feature, while the branches are the answers to that specific condition. Following the branch, you get to a new node, until the leaf node is reached, which represents the model's prediction.

The decision tree algorithms can conduct to overfitting, meaning that they learn to suit the training data very well by also learning the noise in the data, and do not perform well on new unseen data. The Random Forest algorithm came as a solution, because it reduces the overfitting, by combining multiple decision trees. But it also reduces the interpretability of the model, compared to a decision tree algorithm. This ensemble learning algorithm creates a forest of trees, which also brings complexity and inefficiency if we have large amounts of data. The algorithm does not need standardized data. It also does not request the missing data to be managed. The algorithm generates the final prediction based on aggregated results of each tree in the forest.

For training the Naive Gaussian Bayes algorithm, supplementary data preparation steps are needed, since it does not perform well on correlated predictors. It is a simple algorithm, with good performance. the disadvantage of it is that it needs input data to be specifically prepared. It does not perform well on correlated data and needs the input data to be normally distributed.

K-nearest Neighbors (KNN) is one of the most intuitive SML algorithms. The scope of it is to label the new unseen datapoint, based on the labels of the K-nearest neighbors. The algorithm is inefficient if the input data are large datasets. The inefficiency comes from the calculus of distance between the new unseen point and all the other points in the training set.

If the input data consists of large and complex datasets, the classical tree algorithms will be too expensive from a computational perspective. So they will produce less accurate results or won't be efficient in training them from a computational cost point of view. There is a group of algorithms that outperforms in these conditions, the so-called “ensemble” algorithms. They combine several weak models into a stronger one, on a sequential basis, driving to an increased overall performance.

AdaBoost is one of the first boosting models, working based on adjusting the weights of the samples. Thus, samples misclassified by the previous model receive higher weights, so the next model focuses more on these. AdaBoost then combines the weak models to make the final predictions, where each model contributes based on its accuracy.

LightGBM is an efficient boosting algorithm, that performs well on large datasets, which acts differently from most other boosting algorithms by extending the tree by leaves, not by level. The algorithm chooses for expansion, the leaf that minimizes the loss. The model can perform poorly on small datasets, due to a higher risk of overfitting. The growth at the leaf level can drive a complex model. A complex model means it has learned also from the noise in the data and is not able to generalize anymore on new unseen data.

CatBoost is a boosting model that works very well on categorical input data and does not require prior preprocessing through encoding methods. CatBoost builds the additive model sequentially, adding a new "decision tree" to the existing model at each step. The scope is to reduce the total prediction error.

XGBoost, like the other boosting methods, improves the model by sequentially adding decision trees that correct the errors made by the previous models, based on the gradient of the loss function.

The performance indicators used in validating and identifying the optimal model were accuracy on the testing and training sets, precision, recall, and the F1 score. In evaluating the models' accuracy, due to unbalanced data, the F1 score was considered the most important figure, which represents the harmonic mean between precision and recall.

The model used for segmenting the population was the K-means algorithm. For determining the central representation of a cluster, it uses the arithmetic mean of all the points in a cluster. This central representation is so called: the centroid. There are some useful methods for identifying the proper number of clusters to be used, like the Elbow method and the Silhouette score. While the Elbow method indicates the optimal number of clusters, the Silhouette score is more of an accuracy indicator, specifying if the clustering model is efficient in optimally managing the split of the population.

Results and discussions

The data sets used for the current research contain transactional data from a non-banking financial institution in Romania, consisting of a set of 113 factoring transactions. Just 106 transactions were completed, meaning that the debtors completed their payment obligations at invoice maturity. Only 7 transactions were categorized as 'bad debt' because the debtor did not pay the bill. The public financial data, published by the Ministry of Finance was the data set used for segmenting of the entire population of active companies in Romania but also for the comparative analysis. The dataset contains financial data from the balance sheets and income statements for all companies in Romania, covering the years 2018 to 2021.

Analyzing the factoring transactions within a non-banking financial institution (Invoicecash SA), it was identified that 7% of transactions were not paid due to the debtors entering insolvency. The institution uses a credit scoring model based on ML algorithms that scored a sample of

companies as having a low risk of entering insolvency, yet these companies were unable to pay the invoice at maturity precisely because they entered the insolvency process. The classification algorithms used aimed to profile and identify the essential characteristics that impact a company's membership in one of two categories: bad payers or good payers.

There were several simulations conducted, testing the ML algorithms on both unstandardized and standardized data. Variables with a correlation coefficient above 90% were eliminated. The Principal Component Analysis technique for dimensionality reduction was also tested, in an attempt to identify the optimal architecture (model and data) that would provide maximum accuracy.

The objective of the first phase - the binary classification - was to obtain the variables with the most significant influence on the target. The results that will be presented are the ones obtained for the simulation using unstandardized and uncorrelated data, but without applying the dimensionality reduction method. This method increases the overall accuracy but the model's explainability is much more decreased.

Table 1. Accuracy metrics: comparing the models

ID	Model	Accuracy Train	Accuracy Test	Precision	Recall	F1 Score	Fit Status
1	SVM	0.934066	0.913043	0.913043	1	0.954545	Balanced
2	Decision Tree	0.978022	0.913043	0.952381	0.952381	0.952381	Overfitting
3	Random Forest	0.978022	0.913043	0.952381	0.952381	0.952381	Overfitting
4	AdaBoost	0.978022	0.913043	0.952381	0.952381	0.952381	Overfitting
5	LightGBM	0.978022	0.913043	0.952381	0.952381	0.952381	Overfitting
6	CatBoost	0.978022	0.913043	0.952381	0.952381	0.952381	Overfitting
7	XGBoost	0.978022	0.956522	0.952381	1	0.976744	Balanced
8	KNeighbors	0.934066	0.913043	0.952381	1	0.954545	Balanced
9	Linear Discriminant Analysis	0.956044	0.913043	1	0.904762	0.95	Balanced
10	Gaussian Naive Bayes	0.912088	0.869565	0.95	0.904762	0.926829	Balanced

Source: Author calculation using Python.

Table 1 presents the accuracy results obtained through training and testing the models using the Grid Search method to identify the optimal combination of parameters within each.

A threshold of 5% was set for identifying the models that were potentially affected by overfitting. If there was a difference in accuracy score between training test sets higher than 5%, then the model was considered affected by overfitting and excluded from further simulations. This information is highlighted in the last column of the table (Table 1), named 'Fit status'. XGBoost model was the one recording the highest F1 score and it was chosen for the final modeling of the analyzed population. The variables used as input predictors for the model were: non-current assets, current assets, inventories, prepaid expenses, advanced payments, provisions, receivables, payables, gross profit or loss, age of the company in the market, employee count, equity, the ratio of the number of employees to the age of the company in the market, and the corresponding rates from the last year to the current year for all the variables.

For the intensity of the correlation between the variables to be easier to perceive, the correlation matrix is presented in the form of a heat map which can be seen in the picture below. The predictors with a correlation index higher than 90% were eliminated, so the representation contains only the remaining predictors.

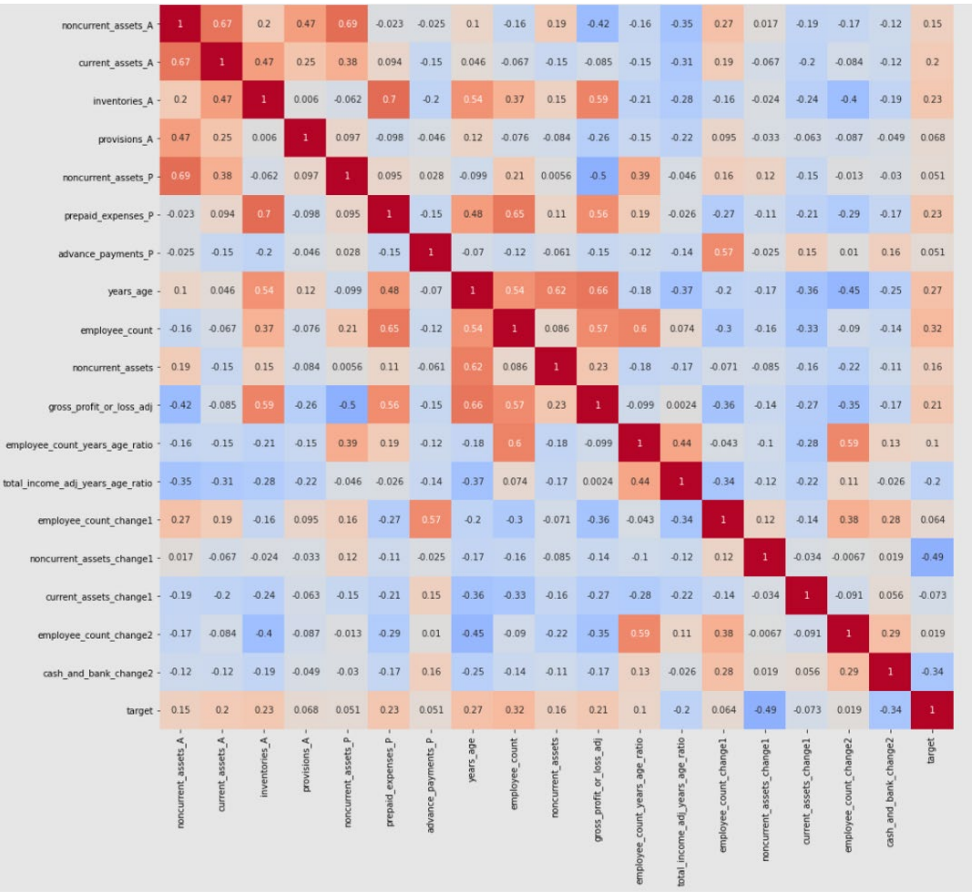

Figure 1. Correlation Matrix as heat map

Source: Author, generated using Python.

Subsequently, the SHAP ((SHapley Additive exPlanations)) library from Python was used to interpret the contributions of individual features to the model's predictions, providing valuable insights into how the model makes decisions based on the input data. The chosen graph for visualizing the impact of the features was the violin plot. The colors indicate the direction of a feature's influence on the prediction. Red means the highest values of the variable, while blue the low values of the input variable. The shape represents the density distribution of values for a specific feature. The wider part of the "violin" indicates a higher concentration of values, suggesting that many instances have similar values for that feature. The narrower part indicates a lower density, suggesting that fewer instances have values in that area. The distribution of values in the "violin" graph shows how different values of a feature influence the model's prediction. A feature with a symmetrical "violin" means its effect on the prediction is similar no matter if its value is high or low. An asymmetry indicates that the effect of the feature on the prediction varies depending on its value.

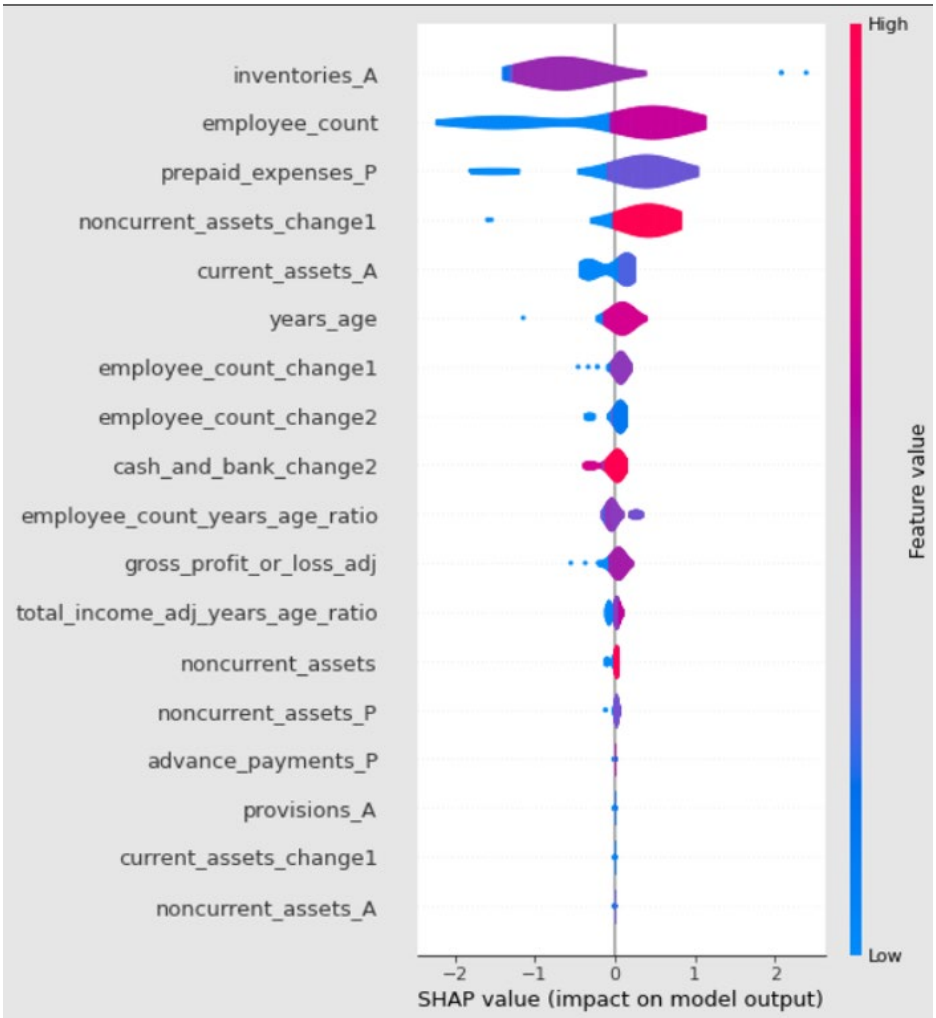

Figure 2. Contribution of individual features to models prediction

Source: Author, generated using Python.

The SHAP output (Figure 2) offers the opportunity to better visualize the predictors' influence on the target. Some predictors are pointed to have a higher impact on the target of the model. As can be seen in the figure, these highly correlated predictors are the number of employees, the year-on-year change in non-current assets, the company's lifetime on the market, and the values of current and non-current assets. The larger the number of employees a company has, the higher the probability that the invoice will be paid, or a longer presence in the market suggests a greater chance of the invoice being paid.

Some predictors with a high impact on the model target are the year-on-year change in non-current and current assets. The number of employees and the receivables have also a high impact on the model output. They were used as segmentation directions for the entire population of active companies in Romania. The input data set contains public data, published by the Ministry of Finance of Romania. Z-scores were calculated for each of the four variables. It was an important step, followed to identify for each element the deviation from the mean and the dispersion of the population. A threshold of 3 was set, so it was assumed that values exceeding 3 standard deviations from the population mean are extremely unusual and could be considered outliers. The correspondent records were not included in the analyzed population. The data were scaled, and

then the K-means algorithm was used for segmenting the population. The elbow method identifies an optimal number of clusters, which in this case is 5 (Figure 3).

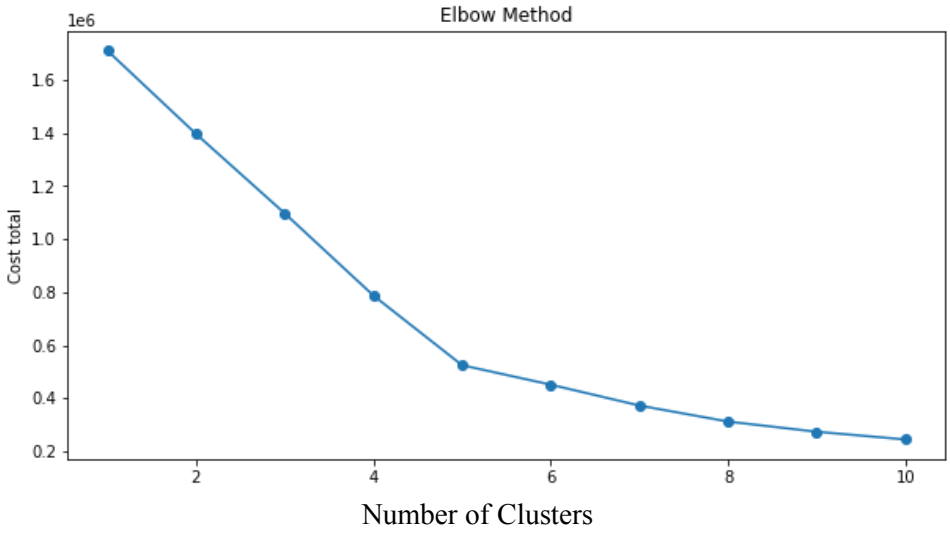

Figure 3. Elbow Method for detecting the optimal number of clusters

Source: Author, generated using Python.

By analyzing the customer segments—those who paid their bills and those who did not—it can be discerned critical characteristics that distinguish the two groups with the aid of boxplots. The median for the "Bad DEBT" segment lies near the lower limit (Figure 4), suggesting a significant number of companies with a minimal number of employees, often below 5. The median for the segment containing paying customers is near the upper limit (Figure 5), which indicates that most of these companies have a larger number of employees, typically over 20.

Further investigation reveals consistent patterns among good-paying customers in segments 1 and 3, while segments 4 and 0 (Figure 6) appear to represent the characteristics of “Bad DEBT” customers more accurately, who are less likely to pay their bills. This analysis is critical for identifying the segment of the population that includes good clients, particularly when considering the employee count indicator. By integrating these insights, financial institutions can better segment their customer base and predict payment behavior, which is invaluable for risk assessment and decision-making processes.

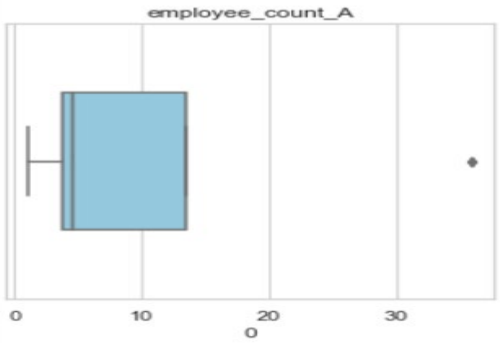

Figure 4. Boxplot for Bad Debt for the variable Number of Employees

Source: Author, generated using Python.

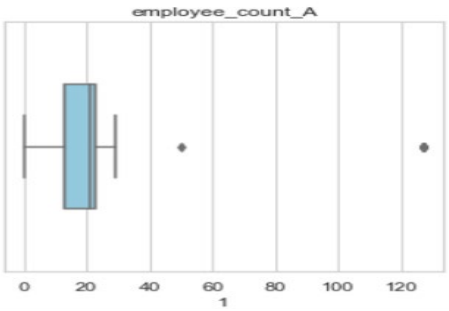

Figure 5. Boxplot for good payers for the variable Number of Employees

Source: Author, generated using Python.

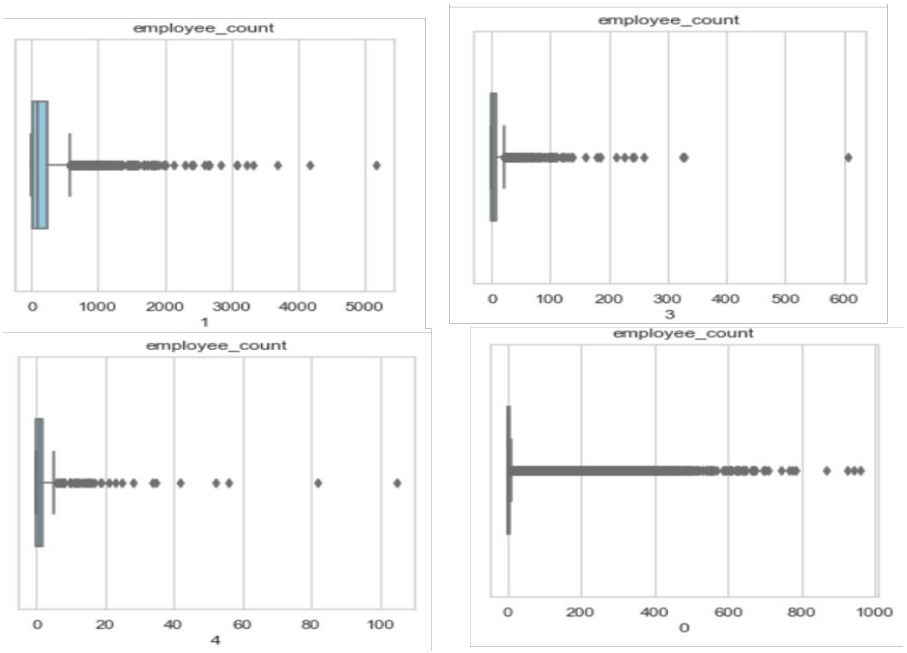

Figure 6. Boxplot for cluster 1 for the variable Number of Employees

Source: Author, generated using Python.

Conclusions

The research has made important contributions to the field of financial risk assessment by examining a vast dataset of 430K entities of active companies from Romania to achieve effective segmentation. An important part of the research was the utilization of SHAP values for the model to be easier to interpret. The method allowed a better understanding of the black box of ML algorithms and the impact of each feature on the model's output. By using SHAP values, it was possible to quantify the contribution of each parameter but also to gain insights into the factors that drive the model's decisions.

SHAP values were important in identifying the key alerting signals that serve as predictors for financial risk. Some examples could be entities with fewer than 5 employees, entities with current assets below 1,260,000 RON, with receivables under 375,000 RON, and so on. Additionally, a constant or reduced employee count year-over-year has been noted as a potential alert signal on default risk.

Furthermore, this work attempted to establish a connection between customer payment behavior and credit risk scoring systems by identifying the most contributive variables for classifying the population of clients. Further, these identified variables serve as segmentation directions for the entire population of active companies.

The segmentation model can be used to develop specialized scoring systems for different market segments. For instance, a credit scoring model for small enterprises might weigh the number of employees more heavily than one for larger corporations. This allows for a more precise approach to credit risk modeling.

This study concludes that a credit scoring system composed of multiple models built at the level of population clusters has the potential to outperform a credit scoring model constructed for the entire population of active companies in Romania. This approach can lead to a better understanding of the probability of default and of non-paying the bill due to insolvency risk. The increase in accuracy and understanding is determined by taking into consideration the specific characteristics and behavior patterns within each subgroup. A step forward in the process of finding the best strategy for reducing the risk of a company default is to tailor the models to specific segments of the population. This approach together with the segmentation algorithms drives an architecture of models working together to increase the overall accuracy. This study advocates for a transition towards segmented credit scoring models to enhance the precision and reliability of financial assessments in the dynamic economic landscape.

The study has some limitations. The small amount of data used for the initial classification model is one of them. Another one is that the test of the similarity between groups is only based on boxplots, but while they provide a valuable visual representation of data distribution and can highlight median trends and outliers, they may not capture the complex, multi-dimensional relationships between variables across different segments. Hence, the visual determination of similarities between diverse populations using boxplots alone may not be sufficient. For future research, the aim is to expand the analysis beyond boxplot comparisons to include descriptive statistical features of all segments and samples, transforming them into vectors to serve as inputs in a classification model. The goal will be to differentiate the two customer segments—bad debt and good payers—into distinct clusters, ultimately identifying which segments of the general population will fall into these new customer clusters.

Acknowledgments

This work was supported by the project "Societal and Economic Resilience within multi-hazards environment in Romania" funded by European Union – Nextgeneration EU and Romanian Government, under National Recovery and Resilience Plan for Romania, contract no.760050/ 23.05.2023

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

References

Bellotti A., Brigo D., Gambetti P., Vrins F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, 37(1), 428-444.

Carmona P., Dwekat A., Mardawi Z. (2022). No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. *Research in International Business and Finance,* 61.

Pang S., Hou X., Xia L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change*, 165.

Jabeur S.B., Gharib C., Mefteh-Wali S., Arfi W.B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166.

Li A., He J., Liu Z. (2022). An Information Based Fuzzy Partitioning Approach (IBFP) for "Bad" Credit Detection. *Procedia Computer Science*, 199, 1160-1167.

Yildirim M., Okay F.Y., Özdemir S. (2021). Big data analytics for default prediction using graph theory. *Expert Systems with Applications*, 176.

Machado M.R., Karray S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, 200.

Lappas P.Z., Yannacopoulos A.N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, 107, 428-444.

Zanin L. (2020). Combining multiple probability predictions in the presence of class imbalance to discriminate between potential bad and good borrowers in the peer-to-peer lending market. *Journal of Behavioral and Experimental Finance*, 25.

Saurabh A., Sushant B., Survesh S., Nassa V.K. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. *Materials Today: Proceedings*, 51(1), 110-117.

Bashar A., Nayak R., Astin-Walmsley K., Heath K. (2021). Machine learning for predicting propensity-to-pay energy bills. *Intelligent Systems with Applications*, 17.

Schoonbe, L., Moore, W.R., Van Vuuren, J.H. (2022). A machine-learning approach towards solving the invoice payment prediction problem. *South African Journal of Industrial Engineering*, 33(4), 126-146.

Nst, A.R., Sari, M.M., Ramadhani, A., Maramis, B.C., Nst, H.K., & Ghazali, M.R. (2023). Analysis of factors affecting bad debts. *World Journal of Advanced Research and Reviews*.

Ariza-Garzón M.J., Arroyo J., A. Caparrini and M. -J. Segovia-Vargas. (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*, 8, 2169-3536 (Online).