

Article

Risk Assessment of Polish Joint Stock Companies: Prediction of Penalties or Compensation Payments

Aleksandra Szymura 

Department of Econometrics and Operational Research, Wrocław University of Economics and Business,
Komandorska 118/120, 53-345 Wrocław, Poland; aleksandra.szymura@ue.wroc.pl

Abstract: Corporate misconduct is a huge and widespread problem in the economy. Many companies make mistakes that result in them having to pay penalties or compensation to other businesses. Some of these cases are so serious that they take a toll on a company's financial condition. The purpose of this paper was to create and evaluate an algorithm which can predict whether a company will have to pay a penalty and to discover what financial indicators may signal it. The author addresses these questions by applying several supervised machine learning methods. This algorithm may help financial institutions such as banks decide whether to lend money to companies which are not in good financial standing. The research is based on information contained in the financial statements of companies listed on the Warsaw Stock Exchange and NewConnect. Finally, different methods are compared, and methods which are based on gradient boosting are shown to have a higher accuracy than others. The conclusion is that the values of financial ratios can signal which companies are likely to pay a penalty next year.

Keywords: machine learning; supervised learning; classification; corporate finance; financial analysis; financial indicators



Citation: Szymura, Aleksandra. 2022. Risk Assessment of Polish Joint Stock Companies: Prediction of Penalties or Compensation Payments. *Risks* 10: 102. <https://doi.org/10.3390/risks10050102>

Academic Editors: Tianyang Wang, Jing Ai and Xiufang Li

Received: 22 March 2022

Accepted: 22 April 2022

Published: 12 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The assessment of activities in an enterprise is especially important, not only from the point of view of management but also counterparties or investors ready to invest their capital as well as other interdependent companies. Before deciding to grant a loan, many financial institutions, such as banks, are obliged to confirm the credibility of both individual and corporate customers. To support the analysis of company activities, advanced credit scoring algorithms are created. Their main objective is to streamline the evaluation process and minimize potential losses for business entities resulting from erroneous and costly decisions.

The evaluation of companies has been an important and widely studied topic for decades. A vast number of algorithms have been proposed (Beaver 1966; Altman 1968; Ohlson 1980; Zmijewski 1984; Betz et al. 2014; Mselmi et al. 2017; Pisula 2017; Shrivastav and Ramudu 2020). Nowadays, given the recent growth of big data, the most popular method is the implementation of machine learning techniques (Barboza et al. 2017; Le and Viviani 2018; Petropoulos et al. 2020; Jabeur et al. 2021; Pham and Ho 2021). In creating evaluating algorithms, it is also important to discover the reasons why one company is riskier than another and which variables have an impact on the final prediction results. Using advanced machine learning methods, it is difficult to obtain such information directly. Model explainability is a significant aspect of modelling. Many researchers focus only on the effectiveness of models, not their explainability. This approach makes the models more effective, but not easily interpretable.

Companies are evaluated in many ways, but most often this is through the analysis of the probability of events such as bankruptcy or losses (Jabeur et al. 2021; Pham and Ho 2021; Pisula 2017), which has been proven in the literature. The main purpose of this paper

was to present an original solution for the evaluation of companies in terms of predicting a negative event—the payment of penalties or compensation. This study is based on a different approach by focusing on a different, but also significant, dependent variable. An in-depth analysis of the methods used to assess companies led to the development of the following hypothesis: the values of financial indicators signal, one year in advance, the occurrence of a negative event in the form of penalties or compensation, which is reflected in the financial situation of the business entity. Financial problems may translate into delays in the delivery of products and services, which in turn leads to sanctions. For example, in 2018 and 2019, the Polish company Elektrobudowa SA had to pay large penalties and compensation to another Polish company because of delays in the fulfilment of a contract. Consequently, this was one of the factors that contributed to the drop in the company's financial indicators and its financial collapse, which eventually led to its bankruptcy¹. To verify this hypothesis, machine learning methods were used. The problem under investigation is related to classification and thus supervised learning methods were implemented. Another goal of this paper was to propose a solution to point out financial indicators which signal the occurrence of an analyzed negative event. The final results could support firms in decision-making processes.

Section 2 of the article reviews the literature and presents the latest scientific achievements related to the use of classic statistical or machine learning methods in the assessment of the activities of entities, companies or individual bank customers and the likelihood of negative events occurring in their activities. This section also presents evaluations of individual credit applicants in order to demonstrate that similar algorithms are used and thus the same methods can be implemented to assess both corporate and individual clients. Among other things, it shows which methods were used in the early years of machine learning and which have recently gained in importance. In the review, the focus was also put on the data used in the studies in question, which showed that this is a global issue rather than a regional one. This is followed by a comparison of the use of variables. Based on the literature review, it can be concluded that there is a trend involving the creation of models based on a limited set of popular dependent variables. The same publicly available datasets are used online on platforms such as Kaggle or the UCI Machine Learning Repository (Marqués et al. 2012; Tsai and Wu 2008). This work attempts to include a dependent variable that is not commonly used and to demonstrate that the information on penalties and compensations paid is a valid dependent variable in scoring modelling.

The third section presents the research methodology. It describes the individual stages of the research procedure, i.e., the selection of variables for the model and the correlation analysis between them, the sampling for modelling and its balancing as well as the presentation of chosen supervised learning methods and techniques for evaluating the developed models. Section 4 presents the results of the conducted research, with a description of the chosen dataset, which contains indicators calculated mostly based on financial information from financial statements of companies. These indicators are also used, among others, in fundamental analysis. Since it was readily available, we used information on business entities listed on the Warsaw Stock Exchange and NewConnect and carried out an analysis of their descriptive statistics and distributions. The effects of the study are presented in both tabular and graphic forms. To select a set of the most important variables that help to predict the studied phenomenon, the SHAP approach was used, which is based on Shapley values originating from game theory. All experimental analyses were implemented using Python.

The final sections include a summary of the study and proposals for further research. The topic covered in this article has not been exhausted. Each year, more and more data are collected that ought to encourage the use of new and more advanced analytical techniques. The assessment of entities also attracts the interest of other researchers.

2. Literature Review

Methods for evaluating the probability of entities experiencing a negative occurrence have already appeared in the literature in the twentieth century. One of the first models, which is still used, was created by Altman, and is known as the Z-score (Altman 1968). It was created using discriminant analysis, and nowadays it is used to predict corporate bankruptcy. Out of the initial twenty-two variables, five were selected for its construction (Altman 1968). The model is constantly in development and is used to predict the insolvency of companies. Almamy et al. (2016) used the Altman model to estimate the probability of such an event among British companies during the financial crisis. Despite the passage of years, it has been proven to still be precise. The Altman model encouraged other researchers to raise the issue of evaluating selected aspects of business activities. Over time, more and more advanced analytical methods were developed. This in turn, attracted more interest in the use of algorithms to assess entities.

In the twenty-first century, machine learning has played an increasingly important role in the construction of such algorithms. Colloquially, machine learning is defined as the ability of machines to learn without being programmed directly. This definition was coined by Arthur Samuel in 1959 (Awad and Khanna 2015). In the case of evaluation algorithms, they are based on a branch of machine learning known as supervised learning, which is similar to learning with the help of a teacher. This method consists of the computer model learning how to assign labels to input data that contain labels previously classified by humans (Chollet 2018). However, this does not exclude the use of other methods in scoring algorithms, such as unsupervised learning. The main purpose of unsupervised learning is to discover dependencies and patterns in data (Chollet 2018). Such methods are the basis for the segmentation and construction of recommendation systems. An example of the use of unsupervised learning methods in the assessment of taxpayers is a publication by Colombian researchers (de Roux et al. 2018). They analyzed the declarations of the Urban Delineation tax in Bogota to detect under-reporting taxpayers and based their calculations on a sample of 1367 tax declarations. They divided them into smaller groups using a spectral clustering technique. Then, they marked the declarations that stood out from other declarations in each of the created clusters. Finally, they submitted the selected declarations for expert verification (in-depth analysis).

Among supervised learning methods, a very popular trend in entity evaluation models is the use of the logistic regression (Barboza et al. 2017; Mselmi et al. 2017; Le and Viviani 2018; Zhou 2013; Zhao et al. 2009; Zizi et al. 2020) and support vector machine (Barboza et al. 2017; Geng et al. 2015; Harris 2015; Mselmi et al. 2017; Xia et al. 2018; Zhou 2013; Shrivastav and Ramudu 2020) models. Additionally, neural networks are also used. Tsai and Wu (Tsai and Wu 2008) followed this path in their research. They used neural networks to predict bankruptcy and evaluate creditworthiness using credit data from three countries: Germany, Japan and Australia. Neural networks also appeared in (Zhou 2013), except that the focus was on American (years 1981–2009) and Japanese (years 1989–2009) non-financial companies. In recent years, however, there has been an increase in the use of ensemble classifiers, such as random forests (Ala'raj and Abbod 2016; Barboza et al. 2017), as well as boosting-based methods, such as gradient boosting (Tian et al. 2020; Pham and Ho 2021), adaptive boosting—AdaBoost (Sun et al. 2020; Marqués et al. 2012; Pham and Ho 2021), extreme gradient boosting—XGBoost (Chang et al. 2018; Xia et al. 2018), or the increasingly popular categorical boosting—CatBoost (Jabeur et al. 2021). To construct the abovementioned methods, several weaker classifiers are combined. As a result, the most powerful classifier is created, which, by definition, increases accuracy (Bequé and Lessmann 2017). Similar methods have been used to assess both types of entities—companies and individuals. In both cases, the effectiveness of the models was satisfactory. Regarding the increasing attention on the assessment of companies, advanced machine learning methods have recently grown in importance.

In the literature, the use of datasets from different parts of the world can be observed, which shows that this topic is a global one. Pisula (2017) used and compared different

ensemble classifiers to assess the phenomenon of production companies going bankrupt in a Polish region, based on a sample of 144 records. In his work (Harris 2015), Harris decided to compare the results of machine learning methods using two historical credit scoring datasets. In both cases, the information concerned credit applicants with and without creditworthiness. The author used a sample of 1000 observations from Germany with 20 variables and a credit union dataset from Barbados with 21,620 observations and 20 variables. In their empirical studies, Spanish researchers (Marqués et al. 2012) used six datasets. In a similar way to the aforementioned Tsai and Wu, they used credit datasets from Germany, Japan and Australia and also supplemented their calculations with information from the United States, Iran and Poland.

The common denominator of the analyzed publications is the use of financial indicators to predict the occurrence of negative events in companies (Sahin et al. 2013; Pham and Ho 2021; Patel and Prajapati 2018; Park et al. 2021; Monedero et al. 2012; Harris 2015; Zizi et al. 2021). Individual customer scoring was instead built based on information about a person's life, such as their gender, marital status and location, etc.

The literature review has shown that individuals were mainly assessed with regard to their creditworthiness. For companies, the dependent variable was based on whether the business entity declared bankruptcy or made a profit or loss. The results presented in the literature indicate that a very important aspect has been neglected. The dependent variable could be powered by other information, for example, on whether the business has had to pay a penalty or compensation to another firm. The purpose of this article was to verify whether this aspect is significant in the assessment of a company's performance. Information regarding the payment of significant penalties or compensation may indicate the relevant intentional or accidental actions of the enterprise, affect its credibility and signal whether it can successfully navigate the current reality and regulations. This is particularly important for companies listed on the stock exchange, where such information can have an impact on investor relations and the transparency of businesses. Moreover, it may shake their financial liquidity, especially if penalties are imposed by supervisory institutions such as The Polish Financial Supervision Authority or The Office of Competition and Consumer Protection in Poland.

3. Methodology

3.1. Analysis of Variables

The first step was the analysis of variables. Basic statistics were calculated, and variables were visualized using histograms and quantile–quantile plots. The normality of distributions was tested with the Shapiro–Wilk test, in which the null hypothesis refers to the normality of the studied distribution, while the alternative hypothesis refers to a distribution that lacks such normality. Regarding the analysis of the distribution of variables, it was also essential to compare the values of the calculated skewness coefficients. Variables with extremely high or low values have to be properly transformed. If values were negative, logarithmic transformation, which has been applied in some empirical studies (Feng et al. 2014), was rejected and another method had to be found. Finally, if a significant right-hand or left-hand skewness was detected, the transformation defined by Formula (1) was used, in which x denotes the variable undergoing transformation.

$$z = f(x) = \begin{cases} \operatorname{sgn}(x) \times \ln|x|, & x < > 0 \\ 0, & x = 0 \end{cases} \quad (1)$$

The next step was the analysis of the correlations between variables. The Spearman rank correlation, the Phi coefficient (ϕ) and the Fisher test were used. The Spearman rank correlation was chosen to explore the relationship between continuous independent variables, the Phi coefficient (ϕ) between continuous independent variables and the dependent variable and the Fisher test between the independent binary variable and any other independent (continuous) variables. Since the Phi coefficient (ϕ) and the Fisher test are used to study the correlation between two dichotomous variables, continuous variables

had to be transformed into dichotomous ones and the observations had to be classified into two new categories. After this transformation, the above methods were applied.

3.2. Sample Selection for the Modelling Process

Due to the fact that the data related to the financial activity of companies span three years and include businesses that have varying numbers of sales, it was decided that instead of simple random sampling, it would be more appropriate to employ the stratification of the sample in relation to other variables or additional information. Stratification allowed for the avoidance of a situation where only large companies were included in the training set and small companies in the test set. This could cause the model to learn sufficiently well on the training set but be overfitted or underfitted on the test set. The collected data was divided into several smaller subsamples, and an adequate percentage of “good” and “bad” examples was selected from each subsample. These percentages were assigned to a training set or a test set. The process was divided into three parts, which are described in detail in Section 4.4. This division resulted in an unbalanced sample with one class containing more records than the other. If used, such samples could lead to incorrect results and render the models useless. One of the most popular methods that is often cited in the literature (Wang et al. 2020; Sun et al. 2020; Ng et al. 2021; Maldonado et al. 2019; Zizi et al. 2021) is SMOTE (Synthetic Minority Over-sampling Technique). Zhou (2013) showed the superior effectiveness of models based on artificially balanced sets using SMOTE in comparison to other balancing set methods such as under-sampling. The purpose of this technique is to generate an appropriate, fixed number of “synthetic” minority class records. In the space of variables, any point from the class which is meant to be enlarged is selected. Then its k -nearest neighbors, also from the same class, are determined, and an additional point in a random location on the line between the chosen point and the one of the nearest neighbor is generated (Chawla et al. 2002).

3.3. Supervised Learning

In this paper, five supervised learning algorithms were used to make a prediction and assign labels to observations using a model trained on input data containing labels. The approach uses classic methods such as logistic regression and a decision tree as well as boosting-based methods such as gradient boosting, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and categorical boosting (CatBoost). These methods are briefly described below. All of them use the following libraries implemented in Python: Scikit-learn, xgboost, catboost and lightgbm². The choice of methods is not accidental. In many publications, boosting-based methods have been used to assess the probability of negative financial events in companies (Jabeur et al. 2021; Pham and Ho 2021). It was decided to verify their effectiveness in predicting the payments of penalties or compensation by Polish joint stock companies.

The first method that was used was logistic regression. It can be applied to predict both dichotomous and multiclass variables. It is not as rigorous in terms of principles as linear regression, but it is important to avoid using independent variables that are strongly correlated among themselves and collinear (Jabeur et al. 2021). In this method, coefficients are not interpreted directly. The most important aspect of the model are the odds ratios, which describe the ratio of the probability of success to the probability of failure. In addition to finance, logistic regression is also widely used in medical research, for example to assess the likelihood of re-infection of a specific disease or recovery from an illness (Fawcett and Provost 2014). The second method used in this paper was a decision tree, which is applied to solve both classification and regression problems. It is often used in the development of decision support tools (Al-Hashedi and Magalingam 2021). Decision trees are also applied in ensemble classifiers as a base classifier, on the basis of which more powerful classifiers are built. A decision tree is a construction consisting of a main node, branches and leaf nodes. Each node shows the feature, the branch, i.e., the decision, and the leaf, i.e., the result in a categorical or continuous form (Patel and Prajapati 2018). Along with logistic regression,

it is one of the most easily interpretable algorithms used in fraud detection (Monedero et al. 2012; Sahin et al. 2013). This paper uses the CART technique with the Gini impurity measure. In machine learning, the approach using ensemble classifiers plays a significant role. It relies on combining the results of single methods, the so-called base classifiers (Marqués et al. 2012). While choosing individual techniques is important, we should also know how to combine them (Sesmero et al. 2021). Multiple techniques are used to improve accuracy. It has been established that the combination of methods produces more accurate prediction results than single base classifiers (Dietterich 1997). In short, the process can be described as combining weak classifiers in order to obtain a more powerful one. It should be mentioned that this is associated with a decrease in their interpretability—ensemble classifiers are not as easy to interpret as the implementation of logistic regression or a single decision tree. There are three main methods for using ensemble classifiers: bagging, boosting and stacking. This research used algorithms based on boosting: gradient boosting, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and categorical boosting (CatBoost). All of them include a decision tree as the base classifier. The last three from the above list are recognized as very effective in predicting events in disciplines other than finance. This is confirmed by the fact that many winning models in competitions on the Kaggle website use these algorithms (Sagi and Rokach 2021). The first algorithm, based on the boosting method, is gradient boosting, also known as gradient boosting machine (GBM for short). This algorithm can be used to solve both classification and regression problems. It reduces prediction errors made by previous classifiers. It was first proposed by Friedman (Friedman 2002). Many algorithms, including XGBoost, LightGBM and CatBoost, are based on the gradient boosting method to improve scalability (Jabeur et al. 2021). XGBoost was developed by Chen and Guestrin (Chen and Guestrin 2016). This technique adds the regularization component to the loss function. As a result, the final complexity takes into account the predictability at each splitting. In addition, XGBoost can handle excessive overfitting by tuning multiple hyperparameters (Sagi and Rokach 2021). LightGBM, unlike many methods based on decision trees, relies on the leaf-wise tree growth algorithm, not a depth-wise one³. Not only are the resulting complex trees more accurate, but this method has proven to be faster than gradient boosting (Ke et al. 2017). One of the disadvantages of utilizing the leaf approach in the generation of a model is the possibility of overfitting in situations where the datasets are smaller. The CatBoost ensemble classifier is the youngest of all the algorithms mentioned in this paper. It is a modification of the standard gradient boosting algorithm. It was proposed by Yandex employees, who are continuously developing it (Prokhorenkova et al. 2018). It copes well with both numerical and categorical data and can also be used for small datasets (Jabeur et al. 2021).

3.4. Model Evaluation and Interpretation (SHAP Approach)

To evaluate the created models, measures based on the confusion matrix were used, such as AUC (Area Under the Curve) and Cohen's kappa, which is indirectly based on this matrix. These two popular evaluation techniques are used in classification models and to compare models. The higher their value, the more accurate the model is. AUC values range from zero (very bad model) to one (ideal model). A value of 0.5 indicates that the prediction of the model is random (Rachakonda and Bhatnagar 2021). The higher the value of this metric, the more accurately the created model assesses a random element of the positive class compared to a random element of the second class (negative). Cohen's kappa uses values between -1 and 1 . Values closer to zero determine the divergence of judgements, while those below zero indicate a worse classification than a random assignment to classes.

As the volume of data increases so does the usage of advanced and more complex predictive algorithms. This is linked to the difficulty of interpreting models due to their "black-box" character. The SHAP (SHapley Additive exPlanations) approach is becoming increasingly popular and has appeared in many publications, (Lundberg and Lee 2017; Mangalathu et al. 2020; Futagami et al. 2021; Dumitrescu et al. 2020; Bakouregui et al.

2021; Matthews and Hartman 2021), including papers related to fraud detection (Severino and Peng 2021). It is based on Shapley's value, which is derived from game theory. Its purpose is to distribute profit or pay-out to players depending on their contribution to the final result of a given game (Bakouregui et al. 2021). This approach was used in the interpretability of machine learning algorithms by Lundberg and Lee (Lundberg and Lee 2017). In this case, its purpose was to assess the contribution of the analyzed variable to the final result of the prediction (Futagami et al. 2021).

4. Results

4.1. Description of Data

All of the necessary information came from the financial statements of businesses. In addition, the dependent variable was constructed on the basis of lists of businesses that were given penalties, which are published on the official website of The Polish Financial Supervision Authority⁴. Based on the acquired data, dependent and independent variables were built. Since penalties or compensation are a consequence of actions, it was decided that the independent variables was to precede the dependent variable by one year. Independent variables were constructed based on data from the years 2016–2018 while the dependent variable was based on data from the years 2017–2019. The total number of collected observations was 928. Table 1 contains information about the total number of companies in each year relative to the dependent variable.

Table 1. Number of businesses included in the analysis by year.

Year	Number of Companies
2017	305
2018	311
2019	312

4.1.1. Dependent Variable

The dependent variable determines whether the company has paid a significant penalty or compensation to another firm ("bad") or not ("good"). It is based on the amount of penalty or compensation compared to the revenue from the same period. In some financial statements, the penalties are classified as "other operating expenses", which, apart from the amount paid, also include other information, such as the cost of court proceedings or administrative costs. However, since these amounts were insignificant, it was decided to build the dependent variable on the basis of "penalties/compensation paid" without considering "other operating costs".

The first stage involved calculating the ratio of penalties paid to revenues. If the sales revenues were zero, the ratio was also equal to zero. There were no cases where a business entity paid penalties and had sales revenues equal to zero in the same year. Due to the high value of the skewness coefficient (greater than 14), it was decided to transform this variable according to Formula (1), where x is the quotient of the amount of penalty/compensation divided by the sales revenues generated in the year when the penalty/compensation was paid.

Based on the transformation according to Formula (1), it was decided that the final shape of the dependent variable could be expressed using Formula (2). In addition, the amount of penalties was taken into consideration. All observations where the amount of penalty or compensation was lower or equal to 10,000 PLN were ignored and classified as "good". In other words, penalties below 10,000 PLN were arbitrarily classified as irrelevant. Formula (2) has the following form:

$$Y = \begin{cases} 1, & ((z \geq -8 \cap z < 0) \cup z > 0) \cap \text{Amount} > 10,000 \\ 0, & (z < -8 \cup z = 0) \cup \text{Amount} \leq 10,000 \end{cases}, \quad (2)$$

where z is the transformed quotient of the penalty paid divided by the sales revenues according to Formula (1), while Amount refers to the amount of the penalty or compensation paid to another company. A total of 284 records labelled as “bad” were obtained. Table 2 contains information about the number and percentage of records classified as “bad” by year.

Table 2. Number and percentage of companies included in the analysis and classified as “bad” by year.

Year	Number of “Bad”	Percentage of “Bad” [%]
2017	93	30.49
2018	95	30.55
2019	96	30.77

4.1.2. Independent Variables

On the basis of collected financial information, a total of nineteen independent variables were built (see Table 3). Many of them are financial indicators used, for example, in fundamental analysis and are mainly continuous variables. Only one of them, a characteristic indicating whether a company has made a profit or a loss, was classified as dichotomous. These variables were chosen because of their important role in company financial standing assessments. They describe significant aspects such as liquidity, profitability, investments, sales or debt. They have often been used to predict negative events such as bankruptcy or financial distress (Barboza et al. 2017; Mselmi et al. 2017), so it was decided to check their utility in predicting penalties or compensation payments.

Table 3. Description of independent variables used in the paper.

Variable	Variable Name	Character of Variable
X1	Net profit	Dichotomous
X2	Return on sales	Continuous
X3	Return on equity	Continuous
X4	Return on assets	Continuous
X5	Operating cash flow margin	Continuous
X6	Current Ratio	Continuous
X7	Quick Ratio	Continuous
X8	Absolute liquidity ratio	Continuous
X9	Debt ratio	Continuous
X10	Debt to equity ratio	Continuous
X11	Long-term debt to equity ratio	Continuous
X12	Operating profit margin	Continuous
X13	Sales profit margin	Continuous
X14	Basic earning power ratio	Continuous
X15	Net income to operating cash flow	Continuous
X16	Indicator of overall financial standing	Continuous
X17	Receivables to payables coverage ratio	Continuous
X18	Return on investment	Continuous
X19	Investment turnover ratio	Continuous

4.2. Analysis of Independent Variables

In the dataset, one of the variables was presented on a dichotomous scale. This is the variable that determines whether a firm has made a profit (indicated as “1”) or a loss or a result equal to zero (marked as “0”) from operations in a given year. Table 4 shows the number of occurrences of each category. Less than 29% of observations resulted in a loss in a specific year, while over 71% resulted in a profit. Most of the companies were analyzed three times because their data was collected for three subsequent years. Some companies recorded both a profit and a loss depending on the year of occurrence of this event.

Table 4. Number and percentage of companies that made a profit or a loss.

Category	Number of Occurrences	Percentage of Occurrences [%]
0—loss	263	28.34
1—profit	665	71.66

For continuous variables, descriptive statistics were calculated (see Table 5). The first stage of the analysis focused mainly on two metrics: the coefficient of variation and the skewness coefficient. The former provided information on variable diversification and the latter on how a variable’s distribution was formed.

Table 5. Values of basic statistics for continuous independent variables.

Variable	Mean	Minimum Value	Maximum Value	Median	Coefficient of Variation [%]	Skewness Coefficient
X2	−1387.96	−1,216,550.00	445,611.11	2.94	−3123.11	−22.71
X3	13.76	−5658.89	7278.58	7.70	3464.02	6.96
X4	−6.78	−7758.00	6325.71	3.05	−5122.08	−5.74
X5	−480.07	−246,550.00	5228.21	5.79	−1783.38	−26.17
X6	3.35	0.04	358.67	1.41	438.60	17.99
X7	2.86	0.03	358.67	1.01	512.24	18.20
X8	1.95	0.00	358.67	0.32	730.08	19.56
X9	55.73	0.28	2262.17	47.92	188.37	16.80
X10	129.55	−15,158.62	16,453.24	89.90	692.44	2.04
X11	46.34	−686.19	3314.26	18.74	388.58	12.59
X12	−986.23	−699,000.00	162,711.11	4.74	−2496.28	−24.90
X13	−581.19	−233,300.00	361.05	4.23	−1761.10	−21.28
X14	−1.86	−3998.00	2309.78	4.54	−8464.50	−14.40
X15	20,047.12	−35,700.00	18,374,400.00	46.37	3008.83	30.36
X16	−22.47	−13,123.06	3202.08	0.99	−2307.00	−20.48
X17	0.91	0.00	56.19	0.59	245.68	17.60
X18	−1.00	−292.41	146.86	0.20	−1859.29	−8.05
X19	36.83	−2.99	3866.35	7.99	474.30	14.70

Table 5 shows that each of the variables was characterized by a high variability exceeding 100% in absolute value. This is a desirable phenomenon—all variables were taken into account in the next stage.

However, what is not desirable is the significant right- and left-sided skewness. This is quite a common situation for economic (financial) data. Each of the variables had an absolute value of the skewness coefficient greater than 2. Most of the negative values that appeared had to be transformed using a different method than the popular logarithmic conversion. It was decided that the same transformation as in the case of the dependent variable, according to Formula (1), was to be used. Table 6 contains the skewness coefficient values for continuous variables after their transformation, while Figure 1 visualizes their distributions (after change).

Table 6. Values of skewness coefficient for continuous independent variables after transformation according to Formula (1).

X2	X3	X4	X5	X6	X7	X8	X9	X10
−1.20	−0.86	−0.84	−1.37	1.05	1.05	0.22	−1.19	−3.21
X11	X12	X13	X14	X15	X16	X17	X18	X19
−0.94	−1.43	−1.58	−1.09	−0.79	−1.51	−0.85	−0.15	−1.04

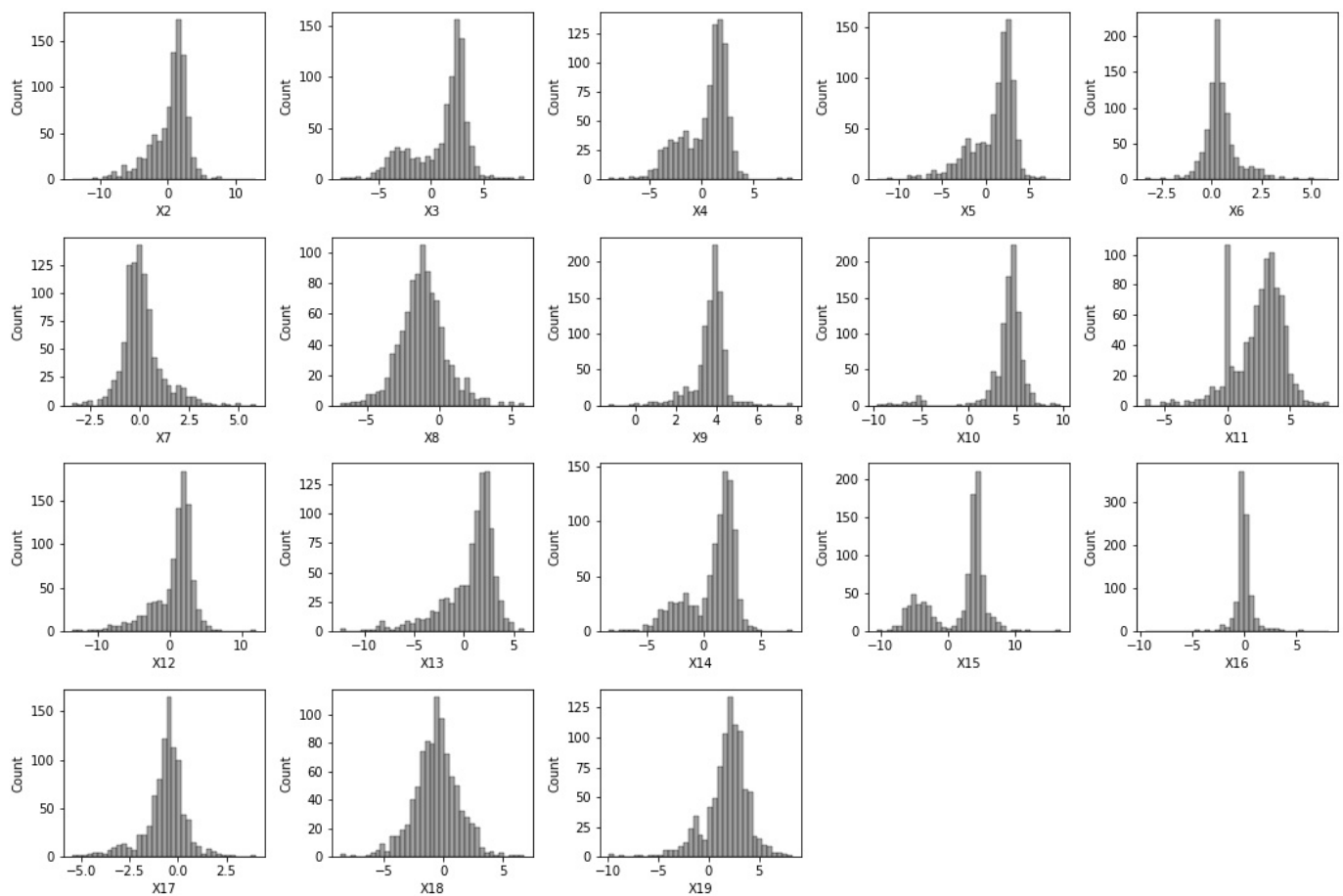


Figure 1. Histograms of the distributions of continuous variables after their transformation according to Formula (1).

After transformation, it was noticed that the absolute value of the skewness coefficient increased for one variable instead of decreasing. This variable, labelled X10, represents the debt to equity ratio. It was concluded that the transformed version of this variable should be included in subsequent stages of the analysis despite its increased skewness. The distributions of this variable before and after conversion were visualized (Figure 2). The increase in the skewness coefficient, despite the transformation, can be explained by the fact that the distribution of the transformed version of the variable labelled as X10 is a mixture of two distributions. The subsequent stages of this research took into account the modifications of variables according to Formula (1).

To investigate the normality of distributions of each variable after transformation, quantile–quantile graphs were used (Figure 3). Based on these graphs, it was concluded that the distributions of all variables were not normal. This fact was also confirmed by the Shapiro–Wilk test results, where the null hypothesis had to be rejected in favor of the alternative hypothesis for each variable due to the very low p -values ($p < 0.01$). It could be observed that despite transformations, the variable distributions were not close to normal. The lack of normality is not a huge problem for all supervised learning methods mentioned in this paper because they can cope with it. It was decided that variables after transformation were to be used in the modelling process, because using variables with very high skewness coefficient values could worsen the model's performance.

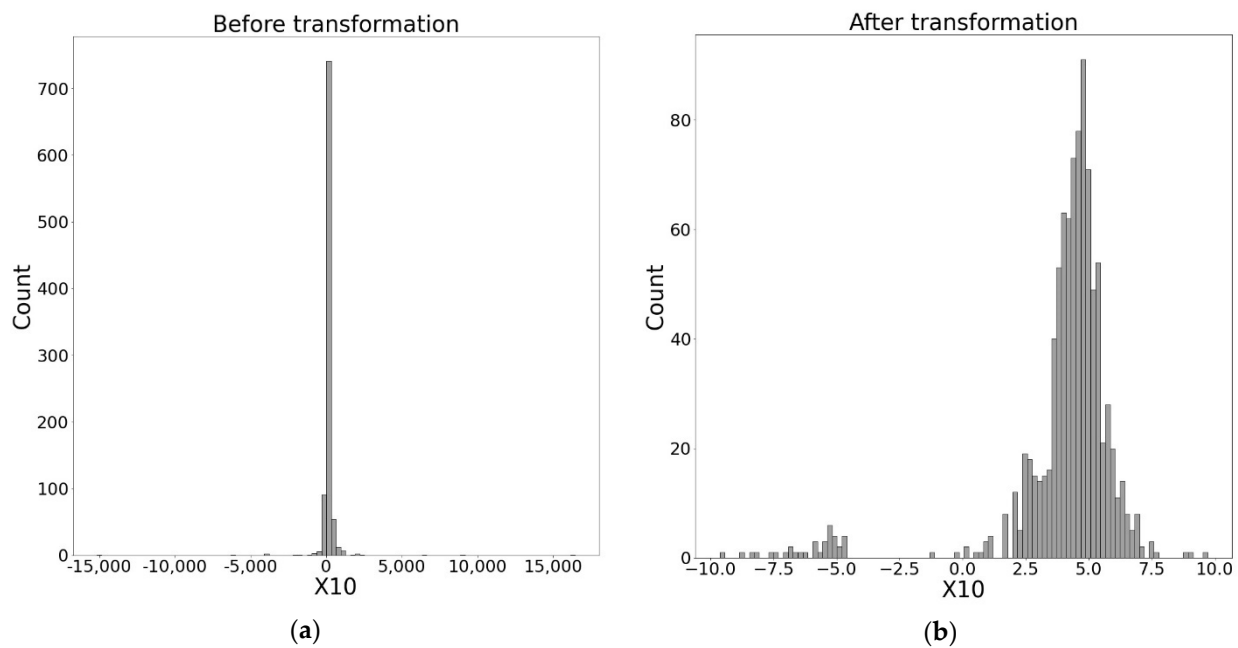


Figure 2. Histogram of the X10 variable before (a) and after (b) transformation according to Formula (1).

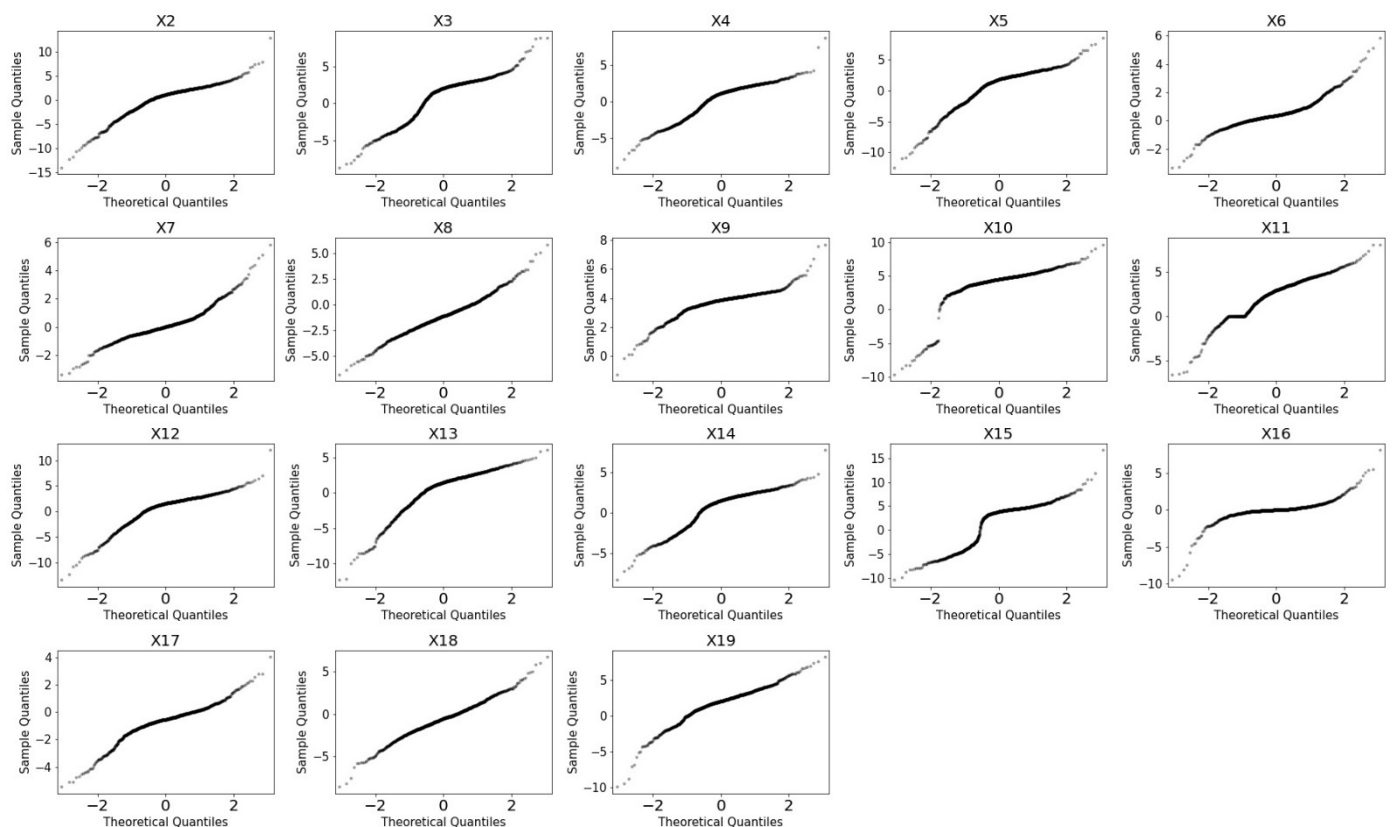


Figure 3. Quantile–quantile plots of continuous independent variables after transformation according to Formula (1).

4.3. Correlation Analysis

The first stage of correlation analysis between variables consisted of calculating the values of Spearman rank correlation coefficients for all pairs of continuous independent variables. The results are presented in graphical form (Figure 4). To study the dependencies

between independent variables and the dependent variable, the Phi coefficient (ϕ) was used after the continuous variable had been divided into two ranges.

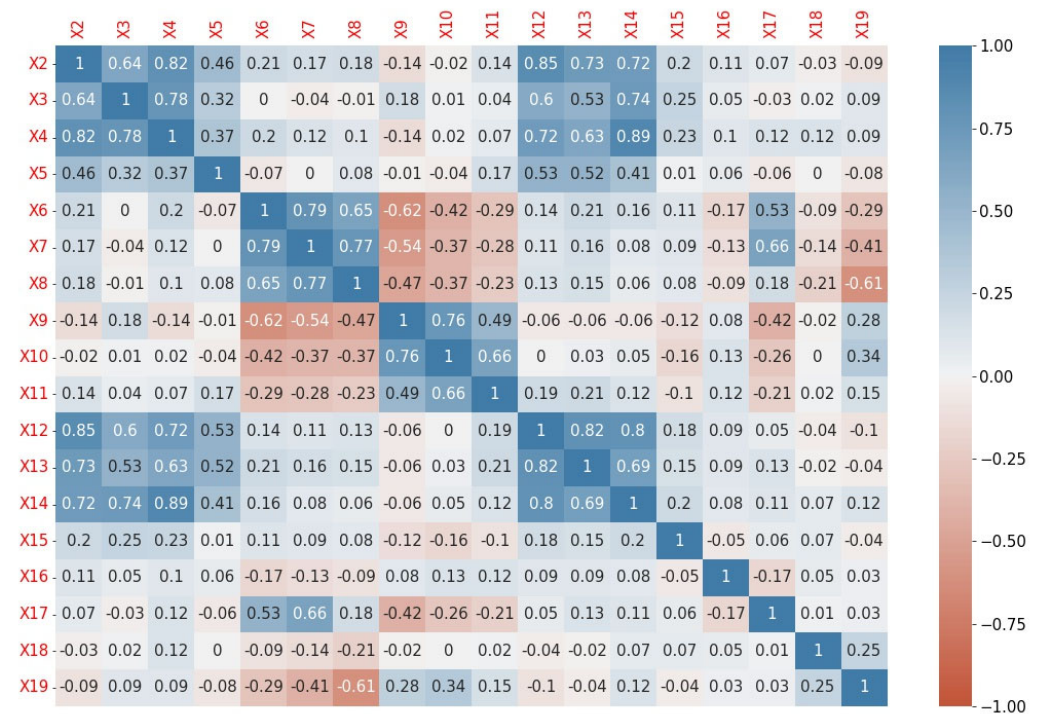


Figure 4. A plot of the Spearman rank correlation coefficient values between continuous independent variables.

Because of their low correlation with the dependent variable, the X3 (return on equity) and X4 (return on assets) variables were removed. Not only were their relationships exceptionally weak compared to other variables but these variables were found to have an insignificant impact on the dependent variable. In subsequent iterations, variables were firstly eliminated based on the adopted threshold values of the Spearman rank correlation coefficient. Strongly correlated variables were those with Spearman rank values lower than -0.7 and greater than 0.7 . The following independent variables were eliminated in subsequent steps:

- Step 1: Similar to the case of X12, the X2 variable correlated in accordance with the approved threshold with three other variables. There was also a strong correlation between X12 and X2, but X2 had a lower impact on the dependent variable than X12 based on the value of the Phi coefficient;
- Step 2: the X7 and X12 variables correlated in accordance with the approved threshold with two other variables;
- Step 3: X9 correlated with one variable (labelled X10), but its correlation with the dependent variable based on the value of the Phi coefficient was weaker than for X10.

To investigate the correlation of all independent variables with one binary independent feature (X1), the Fisher test was conducted. Continuous variables had already been transformed into the dichotomous type. The results are presented in Table 7. The X1 variable was strongly correlated with most of the variables, as evidenced by very low p -values. At a significance level of 0.05 , the null hypothesis of the independence of the examined variables was rejected. The X1 variable was eliminated. Finally, a set of variables was selected for further processing. These were: X5, X6, X8, X10, X11, X13, X14, X15, X16, X17, X18 and X19.

Table 7. The *p*-values obtained from the Fisher test.

X5	X6	X8	X10	X11	X13	X14	X15	X16	X17	X18	X19
0.0000	0.0000	0.0197	0.0000	0.0000	0.0000	0.0000	0.0000	0.5602	0.2839	0.0000	0.0000

4.4. Division of Data into a Training Set and a Test Set

The sample was stratified by entering additional information, such as the year of payment of the penalty or compensation, the revenues in that year and the dependent variable. The whole process was divided into three stages and the final objective was to obtain a training set and a test set.

Stage 1: Division of the sample according to the year of payment of the penalty/compensation

The sample was divided into three subsamples according to the year of payment of the penalty/compensation. Table 8 shows number of records of each new subsample.

Table 8. Number of observations and companies classified as “bad” by year.

Year	Companies	“Bad”
2017	305	93
2018	311	95
2019	312	96

Stage 2: Division of data according to the sales revenues in the year when the penalty was paid

Each of the subsamples created in step 1 were divided according to the amount of revenue obtained in the year when the penalty/compensation was paid and further divided into the following quartile-based groups:

- I group: <minimum value; I quartile>
- II group: (I quartile; median>
- III group: (median; III quartile>
- IV group: (III quartile; maximum value>

Tables 9–11 show the number of observations grouped by category after two stages in the sample division process

Table 9. Number of observations by category of the dependent variable and group based on quartiles of revenue volumes in 2017.

Category	I Group	II Group	III Group	IV Group
0	67	55	43	47
1	10	21	33	29

Table 10. Number of observations by category of the dependent variable and group based on quartiles of revenue volumes in 2018.

Category	I Group	II Group	III Group	IV Group
0	69	54	44	49
1	9	24	33	29

Table 11. Number of observations by category of the dependent variable and group based on quartiles of revenue volumes in 2019.

Category	I Group	II Group	III Group	IV Group
0	73	51	43	49
1	5	27	35	29

Stage 3: Division of data based on the dependent variable

The final stage of the division process was based on the dependent variable. A total of 75% of records from each category (“good”, “bad”) within a specific subgroup created in stage 2 were assigned to the training set and the other 25% to the test set. Table 12 contains information about the number of records in each category in a specified set.

Table 12. Number of records by set type and category of dependent variable after stratification.

Character of Set	Category	Number of Records
Training set	1	206
	0	478
Test set	1	78
	0	166

Based on the information contained in Table 12, it is clear that the created sets were unbalanced, as the number of observations in one category exceeds the number of records in the other category. It was decided one of the methods for set balancing described in Section 3.2, in particular the SMOTE method, was to be used. In this analysis, as in other publications (Maldonado et al. 2019), the value of parameter k was 5. According to the adopted modelling principles, this process was carried out only on the training set, with the test set remaining unchanged. Table 13 shows the number of observations from each set by category after applying the SMOTE method.

Table 13. Number of records by set type and category of dependent variable after stratification and SMOTE method.

Character of Set	Category	Number of Records
Training set	1	478
	0	478
Test set	1	78
	0	166

4.5. Supervised Learning

The whole procedure, consisting of drawing observations for the training and test sets as well as launching a particular method and calculating the evaluation indicators, was performed ten times. This process is a modification of cross-validation, in which the entire dataset is divided into a training set and a test set and then the training set is divided into n subsamples. An evaluation set consisting of one subsample and a training set containing n-1 subsamples were obtained. Each of the n groups in the whole process became, in a sense, a test. In each iteration of the applied approach, a training set and a test set were created from scratch and a proper algorithm was run. Both of these procedures were performed k times. The aforementioned process can be defined as sampling with replacement. One record could have been drawn for the test set several times. The k parameter was assigned the value of 10 in the calculation, because 10-fold cross-validation is often used in the literature.

Each result in each iteration was saved and their average score was treated as the final result (Table 14). In terms of the value of AUC, the best results were obtained using the

CatBoost method, whereas the worst outcome was produced using the built decision tree. Interestingly, all the boosting-based algorithms proved to be more efficient than classic, easily interpretable methods such as logistic regression or the decision tree. The same was true of Cohen's kappa indicator. Again, CatBoost was the best, and the decision tree the worst. Once again, the superiority of boosting over other methods was confirmed.

Table 14. Mean value of AUC and Cohen's kappa coefficients for the applied supervised learning methods after a 10-fold division of the sample into a training set and a test set and running the specified method.

Method	AUC	Cohen's Kappa
Logistic regression	0.6522	0.1903
Decision tree	0.5913	0.1767
XGBoost	0.7159	0.2754
Gradient boosting	0.7100	0.2925
LightGBM	0.7178	0.2716
CatBoost	0.7321	0.3027

In order to analyze the stability of individual methods, standard deviations were calculated for the results obtained in ten iterations. They are presented in Table 15. This time, the most stable method with respect to AUC was gradient boosting, whereas the least stable outcomes were produced by the decision tree. The results were similar in the case of Cohen's kappa index values. Once again, a boosting-based algorithm proved to be the best, but in this instance, it was XGBoost. The worst results were again produced by the decision tree. However, it is worth noting that in terms of the Cohen's kappa indicator, the second least stable method was LightGBM. This might stem from this technique's poor performance when there are small datasets, as was the case in this research. The stability of this method was worse than that of logistic regression. In the case of the AUC values, LightGBM also produced the least stable results in comparison with the other three boosting-based algorithms and was the only one to exceed the value of 0.02.

Table 15. Standard deviation value of AUC and Cohen's kappa coefficients for the applied supervised learning methods after a 10-fold splitting of the sample into a training set and a test set and running the specified method.

Method	Standard Deviation of AUC	Standard Deviation of Cohen's Kappa
Logistic regression	0.0244	0.0425
Decision tree	0.0347	0.0657
XGBoost	0.0194	0.0162
Gradient boosting	0.0181	0.0319
LightGBM	0.0210	0.0566
CatBoost	0.0195	0.0393

Furthermore, the SHAP values, based on the Shapley value, of each iteration were averaged for each of the variables and a special ranking was created (Figure 5). Detailed charts of the SHAP approach for each method and for each iteration were analyzed. Figure 6 shows an example of this approach for one method (based on gradient boosting) for one of the iterations. These graphs can help to determine which variables affect the dependent variable and in what manner.

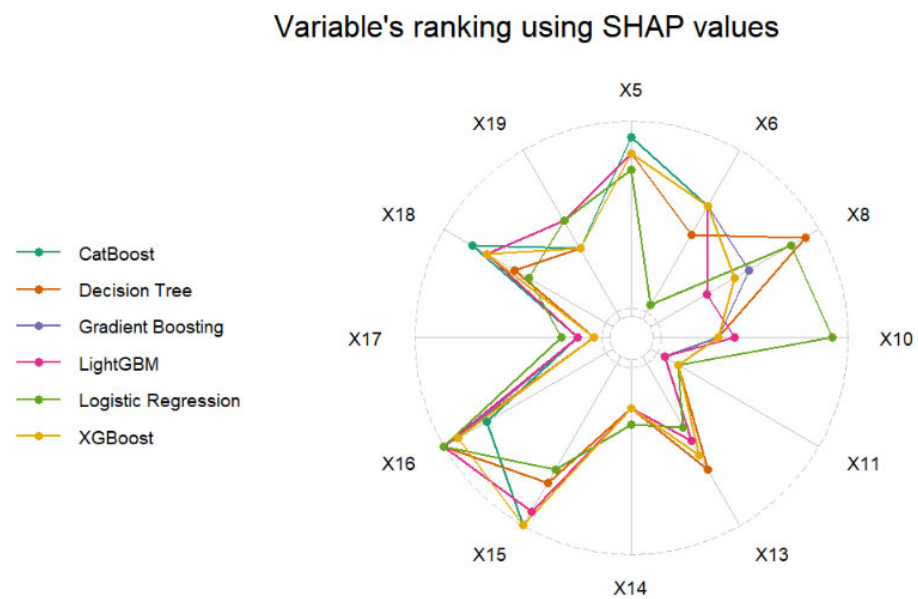


Figure 5. Feature importance ranking by SHAP value per supervised learning algorithms used.

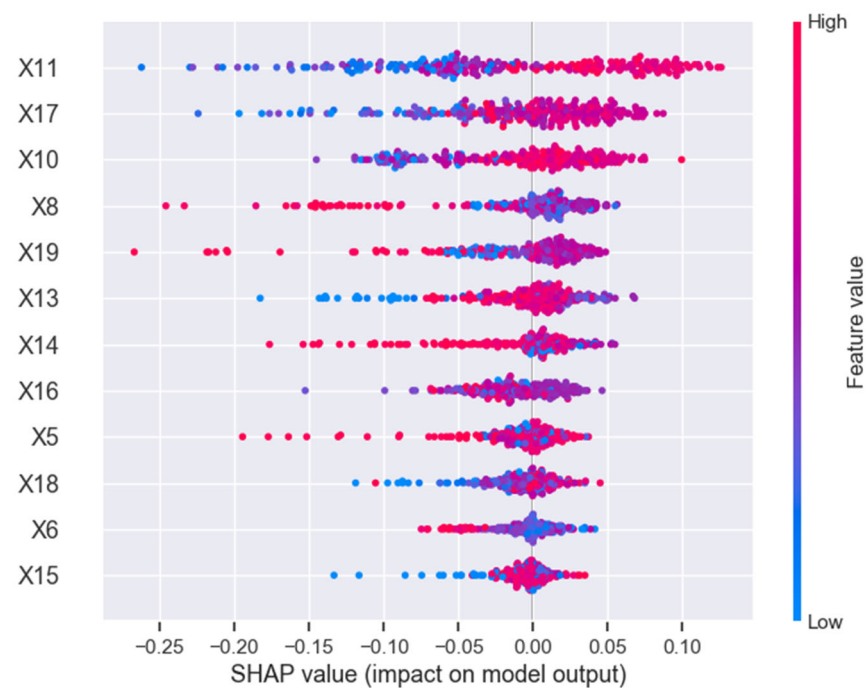


Figure 6. Example plot of SHAP values for one iteration of the model evaluation process for the CatBoost algorithm.

The results of the logistic regression are essentially different from other methods. Figure 7 shows the Spearman rank correlation coefficient values for the scores of each algorithm. The effects of logistic regression are the most correlated with the decision tree results. The value is 0.66. In the case of other techniques, they are correlated with each other at a level of at least 0.75, which is a strong, significant dependence. This should not be surprising, because the gradient boosting methods are based on decision trees.

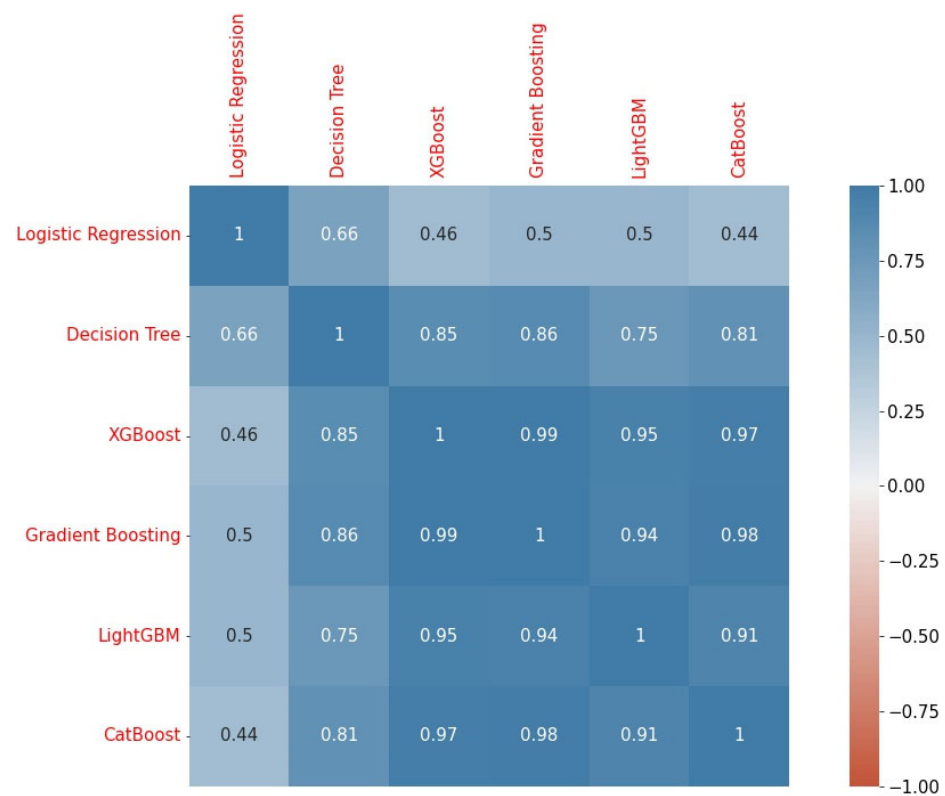


Figure 7. The values of the Spearman rank coefficient of correlation between the positions of the independent variables in the feature importance rankings based on SHAP values for different machine learning methods.

Bearing in mind the logistic regression effects, the most important characteristic which has an influence on the prediction of the examined undesirable phenomenon is the current ratio (X6). The lower the values of this feature, the more the model leans towards the “bad” class. In the case of the other methods, the results were relatively similar. An importance ranking was created which omitted the logistic regression effect. It is presented in Table 16. The most important variables that signaled a payment of penalties or compensation by a company in the following year were: the long-term debt to equity ratio (X11), the receivables to payables coverage ratio (X17) and the basic earning power ratio (X14).

Table 16. Ranking of feature importance by mean SHAP values.

Ranking	Variable
1	X11
2	X17
3	X14
4	X10
5	X19
6	X13
7	X8
8	X6
9	X18
10	X5
11	X16
12	X15

A deeper analysis of these independent variables for the applied methods showed that the classification of a business as “bad” is supported by high long-term debt to equity ratio (X11) values and high receivables to payables coverage ratio (X17) values and also

by lower basic earning power ratio (X14) values. This was illustrated in Figure 6. The higher the long-term debt to equity ratio or receivables to payables coverage ratio values, the greater the SHAP value in the positive direction, so the larger the probability that the company will pay a significant penalty or compensation the following year. Lower basic earning power ratio values mean that a company is more likely to pay a significant penalty or compensation the following year.

5. Discussion

The financial statements of companies provide a great deal of information about their financial activities. In this paper, a dependent variable was created based on information about the payment of penalties or compensation to other companies.

This is an original approach in assessing companies. Previous studies have not used the payment of penalties or compensation as a dependent variable. I decided to compare the modelling effects presented in other publications, focusing mainly on models that assess the probability of negative financial events, in particular the prediction of bankruptcy or insolvency of companies. It is worth highlighting that the analyzed research methodologies exhibited differences in the size and kind of the population, the set of variables or the period under investigation. Only those methods were compared which were used in this study. In the case of (Jabeur et al. 2021), the evaluated data was not older than three years before company failure. The closer a company was to bankruptcy, the higher the AUC was for all the created models. The best algorithm in each period was CatBoost, for which the AUC ranged from 0.764 to 0.994. For XGBoost, the AUC value was between 0.715 and 0.931, while in the case of gradient boosting, its values ranged from 0.718 to 0.951. In the case of logistic regression, its AUC values ranged from 0.744 to 0.919, and the results were better than for XGBoost or gradient boosting when determining the prediction of bankruptcy three years prior. In the study by (Pham and Ho 2021), the boosting-based algorithms were compared. The AUC value for XGBoost and gradient boosting was 1, which is the ideal state. However, we do not know if model overfitting occurred in this case. Pisula (2017) compared the results of several methods for an unbalanced and balanced sample. For example, the decision tree was used as a stand-alone classifier and a base classifier for the ensemble model. For each of these iterations, the AUC value was greater than 0.9, while Cohen's Kappa was above 0.8. This demonstrates the high level of the predictive power of the decision tree.

These results confirmed what was stated before in the literature related to algorithm accuracy. Ensemble classifiers are more accurate than logistic regression or an individual decision tree. It would be preferable for future studies to concentrate on these algorithms. For example, it might be worth trying to use AdaBoost, which is an algorithm which is also based on boosting. Future research should concentrate on selecting hyperparameters whose optimization may help to increase the values of model evaluation metrics.

The results of this research indicate which financial measurements could signal the future occurrence of negative events that exacerbate the financial situation of a business entity. This is a very important subject in business management, as it allows managers to focus on those measures which reduce the enterprise's rating. Such information is valuable in the context of a company analysis or financial statement evaluation; it makes the management of a company more effective and helps to minimize the risk of mistakes resulting in financial losses. According to the results of this study, the most important variables in the analysis of a business are the long-term debt to equity ratio, the receivables to payables coverage ratio and the basic earning power ratio, as determined by ensemble classifiers, as well as the current ratio, as shown by logistic regression effects. This information could also be helpful for investors interested in buying company shares. First of all, they could concentrate on these specific indicators instead of conducting a comprehensive fundamental analysis, saving time, which is crucial in decision-making processes on the stock market. However, they should not limit their analysis to the indicated ratios, as this could obscure the overall picture of a company's financial condition. This model could also help investors to reduce

risk in their investment portfolio. The model's predictions pertain to companies listed on the stock exchange, which is why it could be helpful in building a portfolio of assets.

High long-term debt to equity ratio or the receivables to payables coverage ratio values, as well as lower basic earning power ratio or the current ratio values make a company riskier. Reduced financial liquidity, as measured by the current ratio, means that a given company cannot cope with the repayment of its current liabilities. In addition, growing long-term debt causes a company to become over-indebted, which means it may fail to fulfill signed contracts. This situation can lead to the imposition of penalties or compensation. In the end, it may even lead to a company's bankruptcy. At first sight, high receivables to payables coverage ratio values may appear to be a good thing in the context of a given company's financial standing, but in the long term, it may raise doubts. Companies with many debtors may, after a certain period of time, have problems obtaining these receivables due to the financial problems of their debtors. Receivables which the customers have not paid are known as bad debts. This can cause problems for companies in paying their liabilities due to the lack of financial resources they were expected to receive from their counterparties. Secondly, not obtaining receivables causes less profit for the company.

It is important to evaluate many aspects of a business. The developed model can be used as a part of one comprehensive scoring algorithm. So far, no publications have been found that include a dependent variable based on the information about significant penalties or compensation paid to other companies and therefore it is difficult to directly compare this model with similarly devised ones. In general, this type of algorithm can be classified as one that assesses the activities of businesses. This category of model mainly focuses on the prediction of bankruptcy. In such algorithms, the dependent variable is based on information about companies that have failed. A value of one is assigned to those businesses that went bankrupt during the analyzed period and a value of zero to other businesses. Such models have a higher accuracy than those proposed in this publication, i.e., those with a dependent variable based on information about the payment of a significant penalty or compensation to another company.

This study can serve as a prelude to future research. However, consideration should be given to increasing the sample size and expanding the criteria for selecting data to include capital companies and business entities with a different legal form. Moreover, other machine learning methods could be incorporated or the information about penalties or compensation may be combined with another variable and included as a new dependent variable. Such information could be regarded as complementary when calculating the score of a company. A model based on a function which determines the payment of a penalty or compensation should be one of the components of a scoring algorithm. It is also essential to take into account other independent variables which are based on other financial indicators.

6. Conclusions

The assessment of the activities of a business is extremely important nowadays. With the huge growth in available data, new opportunities are emerging that make the construction of advanced algorithms possible. In this study, the research hypothesis formulated in the introduction was confirmed. The values of financial indicators signal, one year in advance, the occurrence of a negative event—penalties or compensation payments, reflected in the financial situation of the business entity. An example of such a negative effect is the Polish company Elektrobudowa SA described in Section 1. In this case, such penalties and compensation exacerbated the company's problems with other counterparties and shareholders⁵. This information is valuable, for example, for stock exchange investors who could make a decision to buy or sell a company's shares based on the values of such measures as the long-term debt to equity ratio, the receivables to payables coverage ratio, the basic earning power ratio or the current ratio. The same applies to people who manage business entities. It is important that such individuals examine a large number of indicators more broadly, but it would be reasonable for them to focus on information about those metrics that could reveal the likelihood of negative developments. In terms of businesses,

such information may accelerate decision-making processes in the company. This paper shows that ensemble classifiers based on decision trees produced better results in terms of accuracy and stability than the use of a single decision tree. The combination of weaker classifiers had a greater effect than one weak classifier. Also, compared to logistic regression, boosting-based methods produced better final scores. The logistic regression differed from the other methods in terms of the importance of variables. In this case, the current ratio was the most important feature that signaled the occurrence of a penalty or compensation paid to another company in the following year. In the case of the other methods, these variables were the long-term debt to equity ratio, the receivables to payables coverage ratio and the basic earning power ratio. It is recommended to researchers continue to use machine learning methods and compare them, while also taking into account other independent variables. Managers or investors who would like to implement obtained results should firstly analyze companies' financial standings using indicators highlighted as significant in the context of predicting penalties or compensation payments, and only after that should they focus on other indicators.

Funding: The project is financed by the Ministry of Science and Higher Education in Poland under the programme “Regional Initiative of Excellence” 2019–2022 project number 015/RID/2018/19, with a total funding amount of 10 721 040,00 PLN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request from the corresponding author.

Conflicts of Interest: The author declares no conflict of interest.

Notes

- ¹ Elektrobudowa SA financial situation (in Polish) is available on: <https://www.rynekelektryczny.pl/spor-elektrobudowy-i-orlenu-dotyczacy-metatezy/> (accessed on 21 January 2022).
- ² Documentations of these libraries are available on the official websites dedicated to these packages: <https://scikit-learn.org/stable/index.html>, https://xgboost.readthedocs.io/en/latest/python/python_intro.html, <https://catboost.ai/en/docs/>, <https://lightgbm.readthedocs.io/en/latest/index.html> (accessed on 21 January 2022).
- ³ This information was available on the website: <https://lightgbm.readthedocs.io/en/latest/Features.html#references> (accessed on 21 January 2022).
- ⁴ Information about the payment of penalties by companies is available on: https://www.knf.gov.pl/o_nas/Kary_nalozone_przez_KNF (accessed on 21 January 2022).
- ⁵ The explanation of Elektrobudowa SA situation (in Polish) is available on: <https://wysokienapiecie.pl/39649-elektrobudowa-idzie-pod-mlotek/> (accessed on 21 January 2022).

References

- Ala'raj, Maher, and Maysam F. Abbod. 2016. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems* 104: 89–105. [\[CrossRef\]](#)
- Al-Hashedi, Khaled Gubran, and Pritheega Magalingam. 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review* 40: 100402. [\[CrossRef\]](#)
- Almamy, Jeehan, John Aston, and Leonard N. Ngwa. 2016. 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* 36: 278–85. [\[CrossRef\]](#)
- Altman, Edward I. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance* 23: 589–609. [\[CrossRef\]](#)
- Awad, Mariette, and Rahul Khanna. 2015. *Efficient Learning Machines. Theories, Concepts, and Applications for Engineers and System Designers*. New York: Apress Open.
- Bakouregui, Abdoulaye Sanni, Hamdy M. Mohamed, Ammar Yahia, and Brahim Benmokrane. 2021. Explainable extreme gradient boosting tree-based prediction of load-carrying capacity of FRP-RC columns. *Engineering Structures* 245: 112836. [\[CrossRef\]](#)
- Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. Machine learning models and bankruptcy prediction. *Expert Systems with Applications* 83: 405–17. [\[CrossRef\]](#)
- Beaver, William H. 1966. Financial Ratios As Predictors of Failure. *Journal of Accounting Research* 4: 71–111. [\[CrossRef\]](#)

- Bequé, Artem, and Stefan Lessmann. 2017. Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications* 86: 42–53. [\[CrossRef\]](#)
- Betz, Frank, Silviu Oprică, Tuomas A. Peltonen, and Peter Sarlin. 2014. Predicting distress in European banks. *Journal of Banking and Finance* 45: 225–41. [\[CrossRef\]](#)
- Chang, Yung-Chia, Kuei-Hu Chang, and Guan-Jhih Wu. 2018. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing* 73: 914–20. [\[CrossRef\]](#)
- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16: 321–57. [\[CrossRef\]](#)
- Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; New York: ACM, pp. 785–94. [\[CrossRef\]](#)
- Chollet, Francois. 2018. *Deep Learning. Praca z Językiem Python i Biblioteką Keras*. Gliwice: Helion.
- de Roux, Daniel, Boris Perez, Andrés Moreno, Maria del Pilar Villamil, and César Figueroa. 2018. Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. Paper presented at the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, August 19–23; New York: ACM, pp. 215–22. [\[CrossRef\]](#)
- Dietterich, Thomas G. 1997. Machine-Learning Research. *AI Magazine* 18: 97–136. [\[CrossRef\]](#)
- Dumitrescu, Elena-Ivona, Sullivan Hué, Christophe Hurlin, and Sessi Tokpavi. 2020. Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. *SSRN Electronic*. [\[CrossRef\]](#)
- Fawcett, Tom, and Foster Provost. 2014. *Analiza Danych w Biznesie. Sztuka Podejmowania Skutecznych Decyzji*. Gliwice: Helion.
- Feng, Changyong, Hongyue Wang, Naiji Lu, Tian Chen, Hua He, Ying Lu, and Xin M. Tu. 2014. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry* 26: 105–9. [\[CrossRef\]](#)
- Friedman, Jerome H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38: 367–78. [\[CrossRef\]](#)
- Futagami, Katsuya, Yusuke Fukazawa, Nakul Kapoor, and Tomomi Kito. 2021. Pairwise acquisition prediction with SHAP value interpretation. *The Journal of Finance and Data Science* 7: 22–44. [\[CrossRef\]](#)
- Geng, Ruibin, Indranil Bose, and Xi Chen. 2015. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research* 241: 236–47. [\[CrossRef\]](#)
- Harris, Terry. 2015. Credit scoring using the clustered support vector machine. *Expert Systems with Applications* 42: 741–50. [\[CrossRef\]](#)
- Jabeur, Sami Ben, Cheima Gharib, Salma Meftah-Wali, and Wissal Ben Arfi. 2021. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change* 166: 120658. [\[CrossRef\]](#)
- Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. Paper presented at the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, December 4–9; Red Hook: Curran Associates Inc., pp. 3149–57. [\[CrossRef\]](#)
- Le, Hong Hanh, and Jean-Laurent Viviani. 2018. Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in International Business and Finance* 44: 16–25. [\[CrossRef\]](#)
- Lundberg, Scott M., and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Paper presented at the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, December 4–9; Red Hook: Curran Associates Inc., pp. 4768–77. [\[CrossRef\]](#)
- Maldonado, Sebastián, Julio López, and Carla Vairetti. 2019. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing* 76: 380–89. [\[CrossRef\]](#)
- Mangalathu, Sujith, Seong-Hoon Hwang, and Jong-Su Jeon. 2020. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures* 219: 110927. [\[CrossRef\]](#)
- Marqués, A. I., V. García, and J. S. Sánchez. 2012. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications* 39: 10244–50. [\[CrossRef\]](#)
- Matthews, Spencer, and Brian Hartman. 2021. mSHAP: SHAP Values for Two-Part Models. *Risks* 10: 3. [\[CrossRef\]](#)
- Monedero, Iñigo, Félix Biscarri, Carlos León, Juan I. Guerrero, Jesús Biscarri, and Rocío Millán. 2012. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. *International Journal of Electrical Power & Energy Systems* 34: 90–98. [\[CrossRef\]](#)
- Mselmi, Nada, Amine Lahiani, and Taher Hamza. 2017. Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis* 50: 67–80. [\[CrossRef\]](#)
- Ng, Wing W.Y., Zhengxi Liu, Jianjun Zhang, and Witold Pedrycz. 2021. Maximizing minority accuracy for imbalanced pattern classification problems using cost-sensitive Localized Generalization Error Model. *Applied Soft Computing* 104: 107178. [\[CrossRef\]](#)
- Ohlson, James A. 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18: 109–31. [\[CrossRef\]](#)
- Park, Sunghwa, Hyunsok Kim, Janghan Kwon, and Taeil Kim. 2021. Empirics of Korean Shipping Companies' Default Predictions. *Risks* 9: 159. [\[CrossRef\]](#)
- Patel, Harsh H., and Purvi Prajapati. 2018. Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering* 6: 74–78. [\[CrossRef\]](#)

- Petropoulos, Anastasios, Vasilis Siakoulis, Evangelos Stavroulakis, and Nikolaos E. Vlachogiannakis. 2020. Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting* 36: 1092–1113. [\[CrossRef\]](#)
- Pham, Xuan T. T., and Tin H. Ho. 2021. Using boosting algorithms to predict bank failure: An untold story. *International Review of Economics & Finance* 76: 40–54. [\[CrossRef\]](#)
- Pisula, Tomasz. 2017. Zastosowanie ensemble klasyfikatorów do oceny ryzyka upadłości przedsiębiorstw na przykładzie firm sektora produkcyjnego działających na Podkarpaciu. *Zarządzanie i Finanse* 15: 279–93.
- Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: Unbiased boosting with categorical features. Paper presented at the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Montréal, QC, Canada, December 3–8; Red Hook: Curran Associates Inc., pp. 6639–49. [\[CrossRef\]](#)
- Rachakonda, Aditya Ramana, and Ayush Bhatnagar. 2021. A^{Ratio}: Extending area under the ROC curve for probabilistic labels. *Pattern Recognition Letters* 150: 265–71. [\[CrossRef\]](#)
- Sagi, Omer, and Lior Rokach. 2021. Approximating XGBoost with an interpretable decision tree. *Information Sciences* 572: 522–42. [\[CrossRef\]](#)
- Sahin, Yusuf, Serol Bulkan, and Ekrem Duman. 2013. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications* 40: 5916–23. [\[CrossRef\]](#)
- Sesmero, M. Paz, José Antonio Iglesias, Elena Magán, Agapito Ledezma, and Araceli Sanchis. 2021. Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles. *Applied Soft Computing* 111: 107689. [\[CrossRef\]](#)
- Severino, Matheus Kempa, and Yaohao Peng. 2021. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications* 5: 100074. [\[CrossRef\]](#)
- Shrivastav, Santosh Kumar, and P. Janaki Ramudu. 2020. Bankruptcy Prediction and Stress Quantification Using Support Vector Machine: Evidence from Indian Banks. *Risks* 8: 52. [\[CrossRef\]](#)
- Sun, Jie, Hui Li, Hamido Fujita, Binbin Fu, and Wenguo Ai. 2020. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion* 54: 128–44. [\[CrossRef\]](#)
- Tian, Zhenya, Jialiang Xiao, Haonan Feng, and Yutian Wei. 2020. Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science* 174: 150–60. [\[CrossRef\]](#)
- Tsai, C, and J Wu. 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications* 34: 2639–49. [\[CrossRef\]](#)
- Wang, Shutao, Shiyu Liu, Jingkun Zhang, Xiangge Che, Yuanyuan Yuan, Zhifang Wang, and Deming Kong. 2020. A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning. *Fuel* 282: 118848. [\[CrossRef\]](#)
- Xia, Yufei, Chuanzhe Liu, Bowen Da, and Fangming Xie. 2018. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications* 93: 182–99. [\[CrossRef\]](#)
- Zhao, Huimin, Atish P. Sinha, and Wei Ge. 2009. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications* 36: 2633–44. [\[CrossRef\]](#)
- Zhou, Ligang. 2013. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems* 41: 16–25. [\[CrossRef\]](#)
- Zizi, Youssef, Amine Jamali-Alaoui, Badreddine El Goumi, Mohamed Oudgou, and Abdeslam El Moudden. 2021. An Optimal Model of Financial Distress Prediction: A Comparative Study between Neural Networks and Logistic Regression. *Risks* 9: 200. [\[CrossRef\]](#)
- Zizi, Youssef, Mohamed Oudgou, and Abdeslam El Moudden. 2020. Determinants and Predictors of SMEs' Financial Failure: A Logistic Regression Approach. *Risks* 8: 107. [\[CrossRef\]](#)
- Zmijewski, Mark E. 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* 22: 59–82. [\[CrossRef\]](#)