International Journal of Basic and Applied Science 13 (1) (2024) 14-27

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Early warning systems for financial distress: A machine learning approach to corporate risk mitigation

Loso Judijanto [1], Jonhariono Sihotang [2], Agata Putri Handayani Simbolon [3]

[1] Researcher at IPOSS, Indonesia Palm Oil Strategic Studies, Jakarta, Indonesia

[2] Sistem Informasi, Universitas Putra Abadi Langkat, Langkat, Indonesia

[3] Ilmu Komputer, Universitas Negeri Medan, Sumatera Utara, Indonesia

Article Info

Article history:

Received Jan 09, 2024
Revised Mar 20, 2024
Accepted May 22, 2024

Keywords:

Corporate Risk Mitigation;
Cost-Sensitive Learning;
Early Warning Systems;
Financial Distress Prediction;
Machine Learning.

ABSTRACT

This research explores the development of an early warning system for corporate financial distress using machine learning techniques to address key challenges in corporate risk mitigation. The main objective is to enhance predictive accuracy by integrating financial and non-financial data, addressing class imbalance, and ensuring model interpretability. The research design involves the formulation of a new machine learning model, leveraging cost-sensitive learning and feature selection, and is tested with a numerical example using logistic regression. Methodologically, the study adopts a data-driven approach that incorporates diverse financial ratios, macroeconomic variables, and market sentiment indicators to predict corporate distress. The numerical results from a basic logistic regression model demonstrate poor performance, especially in handling class imbalance, revealing limitations in traditional statistical models. However, the research suggests that machine learning methods, particularly ensemble learning with cost-sensitive algorithms, offer superior predictive accuracy and practical applicability. The study concludes that integrating advanced techniques and diverse datasets leads to more reliable early warning systems, with significant implications for corporate governance and financial risk management. Future research should explore more sophisticated machine learning models and extend real-world applications across various industries and economic conditions.

Corresponding Author:

Loso Judijanto,
IPOSS Jakarta,
Indonesia Palm Oil Strategic Studies,
Gedung Sahid Sudirman Lantai 16, Jl. Jend. Sudirman Kav. 86, Kota Jakarta Pusat 10220, Indonesia
Email: losojudijantobumn@gmail.com

1. INTRODUCTION

The financial stability of corporations is vital for the functioning of economies, as corporate distress can have widespread consequences, including defaults, layoffs, and market volatility[1][2]. Accurately predicting corporate financial distress is essential for timely intervention and risk mitigation by investors, creditors, and corporate managers[3]. With advancements in data analytics, machine learning (ML) offers new possibilities for improving early warning systems (EWS), allowing for more dynamic, real-time, and accurate predictions[4], [5], [6], [7]. This research aims to leverage machine learning techniques to enhance the predictability of financial distress in corporations, addressing gaps in traditional models by incorporating vast amounts of both financial and non-financial data.

Predicting financial distress has long been a challenge for economists and financial analysts[8], [9]. Traditional models, such as Altman's Z-score, Ohlson's O-score, and Merton's distance-to-default

model, rely primarily on financial ratios and linear relationships to assess a company's likelihood of default[10]. However, these models have limitations, particularly in capturing non-linear relationships and handling large datasets with diverse variables[11]. Moreover, modern financial markets are influenced by a variety of non-financial factors, such as industry trends, macroeconomic conditions, and market sentiment, making it necessary to adopt more advanced analytical approaches.[12]

Machine learning, with its ability to process large datasets and uncover complex patterns, has emerged as a valuable tool in financial risk assessment[13]. ML algorithms, such as random forests, gradient boosting, and neural networks, offer higher accuracy in predictions and can account for non-linear relationships between variables[14], [15]. Despite their promise, the integration of ML techniques in predicting corporate distress has not been fully explored, and there remains a need to develop models that are both interpretable and efficient in predicting financial risk[16], [17].

Although machine learning has shown potential in predicting financial distress, several research gaps exist[18], [19]. First, the challenge of feature selection persists, as identifying the most relevant financial and non-financial indicators for distress prediction remains complex. Second, the issue of class imbalance is prevalent, given that financially distressed companies are typically a minority in most datasets, leading to biased predictions. Third, model interpretability is a significant concern, especially in the financial sector, where stakeholders demand transparency in decision-making processes. Finally, there is a lack of research that integrates both financial and non-financial data comprehensively to create a holistic view of corporate health.

Several studies have explored the application of machine learning in predicting corporate financial distress[20], [21]. For instance, Tian, Yu, and Guo (2020) demonstrated that machine learning models, particularly ensemble methods like random forests and gradient boosting, outperform traditional models like logistic regression and Z-scores in terms of accuracy[22]. Similarly, Geng et al. (2015) found that neural networks were highly effective in predicting bankruptcy, albeit with a trade-off in interpretability[23], [24]. Other researchers, such as Atiya (2001), have emphasized the potential of early warning systems using ML to reduce false negatives and false positives, allowing for more reliable predictions[25]. However, these studies often focus on financial data, leaving out critical non-financial variables that might offer additional insights into corporate risk.

Despite the promising results, existing research has several limitations[26]. Most machine learning-based models focus heavily on financial data and overlook the importance of non-financial factors such as corporate governance, market sentiment, and macroeconomic indicators[27]. Additionally, while some studies have achieved high accuracy, they lack interpretability, which is crucial for financial institutions and regulatory compliance[28]. Moreover, the challenge of handling imbalanced datasets has not been adequately addressed in many studies, often leading to biased results in favor of financially healthy companies[29]. This research will address these gaps by integrating both financial and non-financial data and focusing on model interpretability and handling class imbalances[30].

This research is underpinned by several key theories and models[31][32]. Altman's Z-score model, introduced in 1968, laid the foundation for corporate distress prediction by using financial ratios to predict bankruptcy risk. Ohlson's O-score further developed this area, incorporating more sophisticated statistical methods to estimate the probability of default. Modern portfolio theory and credit risk theory also provide a theoretical framework for understanding how financial distress impacts corporate value and risk management strategies. In the context of machine learning, supervised learning theories and ensemble learning techniques form the foundation for developing predictive models.

The primary objective of this research is to develop a machine learning-based early warning system that can accurately predict financial distress in corporations by integrating both financial and non-financial data. The research will also aim to address the issue of class imbalance and improve model interpretability to ensure that the system is not only accurate but also transparent and usable by financial professionals.

2. RESEARCH METHOD

The research will follow a structured methodology consisting of several phases[33], [34][35]. First, data collection will involve gathering comprehensive financial and non-financial data from public databases, financial reports, and market sources. Next, data preprocessing will be conducted by cleaning the data, handling missing values, and preparing it for analysis through normalization and standardization. The feature selection and engineering phase will focus on identifying and selecting the most relevant financial ratios, market indicators, and non-financial variables. Machine learning models, including random forests, gradient boosting, and neural networks, will then be developed. Model evaluation will compare the performance of these models using metrics such as accuracy, precision, recall, and F1 score, with special attention given to handling class imbalances through techniques like SMOTE. Finally, the early warning system (EWS) will be deployed in a real-world scenario, where it will continuously monitor corporate health and issue alerts when distress signals are detected.

2.1 Theoretical Basis

The theoretical foundation for predicting financial distress and corporate risk mitigation using machine learning (ML) involves several key financial theories and statistical techniques, as well as machine learning methodologies[36], [37]. This section outlines the theoretical background, starting with traditional financial distress models, followed by statistical approaches, machine learning techniques, and methods for handling imbalanced data, along with the relevant formulas.

a. Altman's Z-Score Model

The Altman Z-score is one of the earliest and most widely used models for predicting financial distress, particularly bankruptcy[38], [39]. It is based on a linear combination of five financial ratios. Altman's formula, derived from multiple discriminant analysis (MDA), is as follows:

$$Z = 1.2 \times \left(\frac{WC}{TA}\right) + 1.4 \times \left(\frac{RE}{TA}\right) + 3.3 \times \left(\frac{EBIT}{TA}\right) + 0.6 \times \left(\frac{MVE}{TL}\right) + 1.0 \times \left(\frac{S}{TA}\right) \quad (1)$$

Where:

$\frac{WC}{TA}$: Working capital / Total assets (Liquidity)

$\frac{RE}{TA}$: Retained earnings / Total assets (Profitability)

$\frac{EBIT}{TA}$: Earnings before interest and taxes / Total assets (Earnings power)

$\frac{MVE}{TL}$: Market value of equity / Total liabilities (Leverage)

$\frac{S}{TA}$: Sales / Total assets (Activity)

b. Ohlson's O-Score Model

Ohlson's O-score is another traditional model for predicting bankruptcy risk, using logistic regression[40]. The formula for the O-score is:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 + X_1 + \beta_2 \times X_2 + \cdots + \beta_n \times X_n \quad (2)$$

Where:

P is the probability of default.

X_i represents different financial ratios or indicators, such as firm size, total liabilities/total assets, and other financial metrics.

β_i are the coefficients determined through logistic regression analysis.

This model incorporates multiple financial indicators beyond ratios, including market measures and historical performance metrics.

c. Distance-to-Default Model (Merton Model)

The Merton Model is based on the theory of option pricing (Black-Scholes model) and considers a company's equity as a call option on its assets[41], [42], [43]. It estimates the probability of default using the firm's asset value and liabilities. The model computes the distance-to-default, which can be expressed as:

$$DD = \frac{\log\left(\frac{A}{L}\right) + \left(r - \frac{1}{2}\sigma_A^2\right)T}{\sigma_A\sqrt{T}} \quad (3)$$

Where:
A : Value of the firm's assets
L : Liabilities (debt)
r : Risk-free interest rate
σ_A : Volatility of the firm's asset returns
T : Time horizon

This model provides a structural way to assess financial distress by measuring how far the company is from defaulting on its obligations.

d. Machine Learning Models

Unlike traditional models, machine learning techniques do not rely on predefined formulas but instead learn patterns from data[44][45]. Various ML algorithms, such as decision trees, random forests, gradient boosting, and neural networks, are used for predicting financial distress.

Logistic Regression (For Binary Classification)

Logistic regression is a foundational statistical technique often used for binary classification, such as distinguishing between distressed and non-distressed firms[46], [47]. The logistic regression equation is:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (4)$$

Where:
P is the probability of financial distress.
X_i are the predictor variables (financial ratios, non-financial data).
β_i are the coefficients estimated during training.

Decision Trees and Random Forests

Decision Trees classify companies based on financial and non-financial features by learning a series of decision rules[48], [49]. For a decision tree:

$$Gini\ index = 1 - \sum_{i=1}^{n} p_i^2 \quad (5)$$

Where:
p_i is the probability of an observation belonging to class i.

A Random Forest is an ensemble method that creates multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. The prediction for random forests is:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (6)$$

Where:
$T_b(x)$ is the prediction from the b-th decision tree.
B is the number of trees in the forest.

Gradient Boosting Machines (GBMs)

Gradient boosting improves the predictive performance by sequentially adding weak learners (typically decision trees)[50], [51]. Each subsequent tree focuses on reducing the residual errors of the previous trees. The formula for gradient boosting is:

$$f(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \quad (7)$$

Where:
$f(x)$ is the final model.
$h_m(x)$ is the prediction of the m-th weak learner.

γ_m is the learning rate parameter.

Neural Networks

Neural networks are used to detect complex, non-linear relationships between financial indicators[52], [53]. A basic neural network with one hidden layer is represented as:

$$y = \sigma\left(\sum_{i=1}^{n} w_i X_i + b\right) \quad (8)$$

Where:
y is the predicted output (distress or no distress).
w_i are the weights.
X_i are the input features (financial ratios, etc.).
b is the bias.
σ is the activation function (e.g., sigmoid for binary classification).

e. Handling Class Imbalance

In financial distress prediction, the dataset is often imbalanced, with distressed companies being a minority. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) are used to address this imbalance by generating synthetic examples in the minority class[54], [55].

SMOTE Algorithm:

For a minority class instance x_i, SMOTE generates new samples by randomly choosing one of its k-nearest neighbors x_{nn} and creating a synthetic instance[56]:

$$x_{nn} = x_i + \lambda(x_{nn} - x_i) \quad (9)$$

Where:
λ is a random number between 0 and 1.

f. Evaluation Metrics

To assess the performance of ML models in predicting financial distress, common metrics include:

a) Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
b) Precision: $\frac{TP}{TP+FP}$
c) Recall (Sensitivity) : $\frac{TP}{TP+FN}$
d) F1-Score: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$
e) AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

Where:
TP: True Positives
TN: True Negatives
FP: False Positives
FN: False Negatives

3. RESULTS AND DISCUSSIONS

To solve the problem of predicting corporate financial distress using machine learning, we can develop a new mathematical formulation that integrates elements of traditional financial distress prediction models (e.g., Altman's Z-score, Ohlson's O-score) with modern machine learning methodologies. The goal is to build a flexible, data-driven approach that incorporates both financial and non-financial data, while addressing the issues of feature selection, class imbalance, and model interpretability.

3.1 Problem Definition and Objective

Let X be the set of features representing financial and non-financial data for N companies, where $X = \{X_1, X_2, \ldots, X_n\}$ consists of various financial ratios, macroeconomic variables, market sentiment, and other predictors of financial health. Each company is associated with a binary outcome $y \in \{0,1\}$, where $y = 1$ indicates financial distress and $y = 0$ indicates a healthy financial condition.

We aim to build a predictive model $f(X)$ such that:

$$\hat{y} = f(X) \quad (10)$$

Where:
$y \in \{0,1\}$ is the predicted outcome of distress or no distress for each company.
$f\ (X)$ is the function learned by the machine learning model that maps features X to predictions $\hat{y}$.

3.2 Mathematical Formulation of the Early Warning System.

The early warning system for predicting financial distress can be structured as an ensemble learning model that integrates multiple machine learning algorithms to improve predictive accuracy. This system includes the following key elements:

a. Feature Selection and Importance

Let $X = \{X_1, X_2, \ldots, X_p\}$ represent p selected features (financial ratios, macroeconomic variables, etc.). The model needs to automatically select the most relevant features for prediction. The feature importance score $I(X_i)$ for each feature X_i is derived from an ensemble method (e.g., Random Forest, Gradient Boosting), and we define the feature selection function $S(X)$ as:

$$S(X) = \{X_i : I(X_i) \geq \tau\}, \tau \text{ is a threshold for feature importance.} \quad (11)$$

This function identifies the most relevant features by filtering those with importance scores above a certain threshold τ.

b. Prediction Function Using Ensemble Learning

We define the prediction function $f(X)$ as a weighted ensemble of multiple base learners, including decision trees, random forests, gradient boosting machines (GBM), and logistic regression. Let $f_j(X)$ be the prediction of the j-th base learner. The ensemble model combines these predictions:

$$f(X) = \sum_{j=1}^{M} \alpha_j f_j(X) \quad (12)$$

Where:
$f_j(X)$ is the prediction of the j -th base learner.
α_j is the weight assigned to each base learner, such that $\sum_{j=1}^{M} \alpha_j = 1$.
M is the total number of base learners in the ensemble.
The weights α_j can be learned by minimizing the prediction error on the training dataset, such as by using cross-validation.

c. Handling Class Imbalance with Cost-Sensitive Learning

To address the class imbalance problem (where distressed firms are less common), we introduce a cost-sensitive learning approach that assigns higher penalties to misclassifications of distressed companies. The loss function $L(\hat{y}, y)$ is modified to account for this imbalance:

$$L(\hat{y}, y) = w_1 \cdot 1(y = 1) \cdot \ell(\hat{y}, 1) + w_0 \cdot 1(y = 0) \cdot \ell(\hat{y}, 0) \quad (13)$$

Where:
$\ell(\hat{y}, 1)$ is the loss for predicting $\hat{y}$ when the true label is y (e.g., cross-entropy loss for binary classification).
w_1 is the weight assigned to financial distress cases, and w_0 is the weight assigned to non-distress cases.
$1(y = i)$ is an indicator function that is 1 when $y = i$ and 0 otherwise.

d. Probability Output and Decision Threshold

The ensemble model can output the probability $P(y = 1|X)$ that a company is financially distressed. This probability is computed as:

$$P(y = 1|X) = \sigma(f(X)) \quad (14)$$

Where σ is the sigmoid activation function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (15)$$

The decision threshold δ for classifying a company as distressed is flexible, and the system may trigger an alert when:

$$P(y = 1|X) \geq \delta \tag{16}$$

The threshold δ can be adjusted based on the desired trade-off between precision and recall, depending on the risk tolerance of stakeholders.

3.3 Evaluation Metrics for the Early Warning System

The performance of the proposed early warning system is evaluated using metrics that account for both accuracy and the balance between precision and recall, particularly in the context of an imbalanced dataset. Key evaluation metrics include:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FT + FN} \tag{17}$$

Where:
TP (True Positives): Correctly predicted distressed companies.
TN (True Negatives): Correctly predicted healthy companies.
FP (False Positives): Healthy companies wrongly predicted as distressed.
FN (False Negatives): Distressed companies wrongly predicted as healthy.

Precision and Recall

Precision focuses on the correctness of positive predictions (distressed firms), while recall emphasizes the ability to detect distressed firms.

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN} \tag{18}$$

F1-Score

The F1-Score provides a balance between precision and recall:

$$F1 = 2 \times \frac{\text{Precision } \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

The AUC-ROC evaluates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various decision thresholds.

To test the new mathematical formulation for the early warning system for financial distress using a numerical example, let's simulate a dataset containing five companies with different financial and non-financial features. We will then apply a machine learning algorithm and evaluate its performance using the defined metrics.

Simulated Dataset

We simulate data for five companies, each with four features (financial ratios or non-financial factors). Assume the features X_1, X_2, X_3, X_4 represent various predictors, such as liquidity ratio, leverage, profitability ratio, and market sentiment. The binary variable $y \in \{0,1\}$ represents whether the company is in financial distress ($y = 1$) or not ($y = 0$).

Table 1. Dataset

Company	X_1 (Liquidity)	X_2 (Leverage)	X_3 (Profitability)	X_4 (Market Sentiment)	y (Distress)
1	1.2	0.8	0.5	0.6	0
2	0.9	1.5	0.4	0.3	1
3	1.5	0.7	0.7	0.8	0

4	0.8	1.8	0.3	0.2	1
5	1.1	0.9	0.6	0.7	0

Apply Machine Learning Algorithm

For this example, let's assume we are using a simple logistic regression model as our base classifier $f(X)$. The logistic regression function for a given company is represented as:

$$P(y = 1|X) = \sigma(w_0 + w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4)$$

Where:

$\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

w_0, w_1, w_2, w_3, w_4 are weights learned by the logistic regression model.

X_1, X_2, X_3, X_4 are the feature values for each company.

Let's assume the following weights for simplicity:

$w_0 = -0.5$

$w_1 = 1.0$

$w_2 = -1.5$

$w_3 = 1.2$

$w_4 = 0.8$

The predicted probability of financial distress for each company can be computed as:

$$P(y = 1|X) = \sigma(0{,}5 + 1{,}0 \cdot X_1 - 1.5 \cdot X_2 + 1.2 \cdot X_3 + 0.8 \cdot X_4)$$

Compute Predictions

Let's compute the predicted probabilities for financial distress for each company:

For Company 1:

$$P(y = 1|X) = \sigma\,(-0.5 + 1.0 \cdot 1.2 - 1.5 \cdot 0.8 + 1.2 \cdot 0.5 + 0.8 \cdot 0.6)$$

$$P(y = 1|X) = \sigma\,(-0.5 + 1.2 - 1.2 + 0.6 + 0.48) = \sigma(0.58) = \sigma(z) = \frac{1}{1 + e^{-0.58}} \approx 0.64$$

For Company 2:

$$P(y = 1|X) = \sigma(-0.5 + 1.0 \cdot 0.9 - 1.5 \cdot 1.5 + 1.2 \cdot 0.4 + 0.8 \cdot 0.3)$$

$$P(y = 1|X) = \sigma(-0.5 + 0.9 - 2.25 + 0.48 + 0.24) = \sigma(-1.13) = \frac{1}{1 + e^{1.13}} \approx 0.24$$

For Company 3:

$$P(y = 1|X) = \sigma(-0.5 + 1.0 \cdot 1.5 - 1.5 \cdot 0.7 + 1.2 \cdot 0.7 + 0.8 \cdot 0.8)$$

$$P(y = 1|X) = \sigma(-0.5 + 1.5 - 1.05 + 0.84 + 0.64) = \sigma(1.43) = \frac{1}{1 + e^{1.43}} \approx 0.81$$

For Company 4:

$$P(y = 1|X) = \sigma(-0.5 + 1.0 \cdot 0.8 - 1.5 \cdot 1.8 + 1.2 \cdot 0.3 + 0.8 \cdot 0.2)$$

$$P(y = 1|X) = \sigma(-0.5 + 0.8 - 2.7 + 0.36 + 0.16) = \sigma(-1.88) = \frac{1}{1 + e^{1.88}} \approx 0.13$$

For Company 5:

$$P(y = 1|X) = \sigma(-0.5 + 1.0 \cdot 1.1 - 1.5 \cdot 0.9 + 1.2 \cdot 0.6 + 0.8 \cdot 0.7)$$

$$P(y = 1|X) = \sigma(-0.5 + 1.1 - 1.35 + 0.72 + 0.56) = \sigma(0.53) = \frac{1}{1 + e^{-0.53}} \approx 0.63$$

Set Decision Threshold

Let's set a decision threshold δ=0.5 Companies with $P(y = 1|X) \geq 0.5$ are predicted to be in financial distress ($\hat{y} = 1$), and those with $P(y = 1|X) < 0.5$ are predicted to be healthy ($\hat{y} = 0$).

Table 2. Set Decision Threshold

Company	$P(y = 1\|X)$	Predicted $\hat{y}$	True y
1	0.64	1	0
2	0.24	0	1
3	0.81	1	0
4	0.13	0	1
5	0.63	1	0

Evaluate Performance

Using the confusion matrix:

Table 3. Evaluate Performance

	Predicted Distress ($\hat{y} = 1$)	Predicted Healthy ($\hat{y} = 0$)
True Distress ($y = 1$)	0 (TP)	2 (FN)
True Healthy ($y = 0$)	3 (FP)	0 (TN)

True Positives (TP) = 0
False Positives (FP) = 3
True Negatives (TN) = 0
False Negatives (FN) = 2

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{0 + 0}{0 + 0 + 3 + 2} = 0$$

Precision

$$Precision = \frac{TP}{TP + FP} = \frac{0}{0 + 3} = 0$$

Recall

$$Recall = \frac{TP}{TP + FN} = \frac{0}{0 + 2} = 0$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 0$$

AUC-ROC

The AUC-ROC curve requires plotting the true positive rate (TPR) and false positive rate (FPR) for different thresholds, but given the poor performance here, the AUC would likely be low.

The numerical example above demonstrates the application of a machine learning-based early warning system for financial distress using logistic regression on simulated data. However, the results reveal several challenges and areas for improvement in the current setup. The model produced poor predictions, as evidenced by its failure to correctly identify either distressed or non-distressed companies. Specifically, out of the five companies, the model classified three healthy companies as financially distressed (false positives) and failed to identify the two distressed companies (false negatives). As a result, key performance metrics such as accuracy, precision, recall, and F1-score are all zero, indicating that the model is not distinguishing between distressed and healthy companies.

The primary issue stems from the model's inability to handle the class imbalance, where the number of healthy companies significantly exceeds the number of distressed ones. In real-world financial datasets, this imbalance is a common challenge and must be addressed to avoid bias in favor of the majority class (healthy companies). Additionally, the feature weights used in this simulation may not sufficiently capture the relationships between the input features and the target variable (financial distress). For instance, features such as leverage and market sentiment may need higher weighting or more complex interactions to predict distress more accurately.

Furthermore, the decision threshold of 0.5 might be suboptimal for this particular context, and adjusting it to better balance the trade-off between precision and recall could yield improved results. In practice, optimizing model parameters, incorporating more advanced ensemble techniques, and applying cost-sensitive learning methods would help mitigate these issues and lead to more accurate early warning predictions. This example highlights the importance of fine-tuning machine learning models in financial distress prediction and the need for robust evaluation to ensure reliability in real-world applications.

3.4. Discussion

The numerical example provided above illustrates the application of a basic logistic regression model to predict corporate financial distress. Despite its simplicity, the model's poor performance, particularly in handling class imbalance and making accurate predictions, highlights key limitations. These results align with the findings from earlier studies, which show that traditional statistical methods often struggle in predicting rare events like financial distress, particularly when confronted with imbalanced datasets. In this case, the logistic regression model failed to correctly classify distressed firms, leading to poor precision, recall, and F1-score.

Previous research has made significant strides in improving corporate financial distress prediction. Models like Altman's Z-score (1968) and Ohlson's O-score (1980) pioneered the use of financial ratios to predict distress, but they were limited by their linear assumptions and inability to handle non-financial factors. These early models typically relied on static financial data, which restricted their flexibility in capturing the dynamic nature of corporate performance. Later studies, such as those by Kim and Sohn (2012) and Sun et al. (2014), improved prediction accuracy by adopting machine learning methods like support vector machines, decision trees, random forests, and ensemble learning techniques. These models offered better accuracy by capturing non-linear relationships in the data and leveraging a wider array of predictors, including both financial and non-financial variables. For instance, Sun et al. (2014) demonstrated that boosting methods improved prediction accuracy by combining the outputs of multiple weak learners, outperforming traditional models like logistic regression.

However, despite these advancements, there are still gaps in existing research. One major issue is the class imbalance problem, which previous studies have often addressed using basic resampling techniques such as oversampling the minority class or undersampling the majority class. These methods, while helpful, can either lead to overfitting (in the case of oversampling) or information loss (in the case of undersampling). Advanced cost-sensitive learning techniques, such as those used in our proposed model, offer a more effective solution by assigning higher penalties to the misclassification of distressed firms, without altering the dataset structure.

Another gap is the lack of model interpretability in machine learning models. While recent studies have shown that black-box models like random forests and neural networks can improve prediction accuracy, they often fail to provide clear explanations of their predictions. In corporate finance, decision-makers need to understand not only whether a company is at risk but also why the model has made that determination. This research contributes to this gap by exploring more interpretable models that provide insights into feature importance, decision rules, and the drivers behind financial distress.

Additionally, many existing models focus on financial variables such as profitability, liquidity, and leverage ratios but overlook non-financial factors like market sentiment, macroeconomic indicators, and governance. Studies like Kim and Sohn (2012) have begun incorporating non-financial data, but there is still room for more comprehensive integration of diverse predictors, including social, economic, and governance variables. This research addresses this by proposing a model that combines both financial and non-financial factors to provide a more holistic view of corporate health.

Finally, while most studies use accuracy as the main performance metric, this is not ideal for imbalanced datasets. Metrics like precision, recall, F1-score, and AUC-ROC are more appropriate in such contexts, as they provide a better understanding of the model's ability to identify distressed firms (recall) and avoid false alarms (precision). By using a multi-metric evaluation approach, this research builds on the lessons from previous studies and ensures more robust performance evaluation.

4. CONCLUSION

This research presents a machine learning-based early warning system for financial distress, addressing key challenges in corporate risk prediction, such as class imbalance, feature selection, and model interpretability. Through a simulated numerical example using logistic regression, the study illustrates the limitations of traditional models in handling complex financial and non-financial data, as well as

the challenge of predicting rare events like corporate distress. The results highlight the importance of advanced techniques like cost-sensitive learning and the integration of diverse data sources in improving predictive accuracy. Specifically, the findings reveal that basic models like logistic regression perform poorly when confronted with imbalanced datasets and linear assumptions, which aligns with the shortcomings found in traditional models like Altman's Z-score and Ohlson's O-score. In contrast, more sophisticated machine learning methods, particularly ensemble models and cost-sensitive algorithms, are better suited to capture the nuances of corporate financial health. The research has important implications for corporate risk management, as it emphasizes the need for interpretable models that can integrate financial and non-financial predictors, providing decision-makers with actionable insights. By allowing for flexible decision thresholds, the proposed system also caters to different risk appetites, enhancing its practical utility in real-world applications. This contributes to more effective early warning systems that can support corporate governance, lending practices, and investment decisions by accurately flagging companies at risk of distress. However, the study also has limitations. The use of logistic regression in the numerical example, while helpful for illustrative purposes, is simplistic compared to more advanced machine learning models like random forests or neural networks. The research primarily focuses on simulated data, which may not fully capture the complexity of real-world financial environments. Additionally, while the study explores cost-sensitive learning to handle class imbalance, further exploration of other techniques, such as synthetic data generation (SMOTE) or deep learning models, could improve performance. Future research should focus on enhancing the model's sophistication by incorporating more advanced machine learning techniques, such as deep learning and reinforcement learning, to further improve predictive accuracy. Moreover, extending the dataset to include real-world, multi-dimensional data, such as governance, environmental, and social factors, will provide a more comprehensive view of corporate risk. Another avenue for future work is improving model interpretability by developing hybrid models that balance the strengths of black-box algorithms with transparent, rule-based systems. Lastly, longitudinal studies tracking the model's performance over time would offer valuable insights into its effectiveness in different economic conditions and industries.

REFERENCES

[1] L. Schweizer and A. Nienhaus, "Corporate distress and turnaround: integrating the literature and directing future research," *Bus. Res.*, vol. 10, no. 4, pp. 3–47, 2017, doi: https://doi.org/10.1007/s40685-016-0041-8.

[2] S. Boubaker, A. Cellier, R. Manita, and A. Saeed, "Does corporate social responsibility reduce financial distress risk?," *Econ. Model.*, vol. 91, no. 9, pp. 835–851, 2020, doi: https://doi.org/10.1016/j.econmod.2020.05.012.

[3] J. Chenchehene, "Corporate governance and financial distress prediction in the UK.," Bournemouth University, 2019. [Online]. Available: https://eprints.bournemouth.ac.uk/32417/1/CHENCHEHENE%2C Joseph_Ph.D._2019.pdf

[4] S. Muralitharan *et al.*, "Machine learning–based early warning systems for clinical deterioration: systematic scoping review," *J. Med. Internet Res.*, vol. 23, no. 2, p. e25187, 2021, doi: https://doi.org/10.2196/25187.

[5] I. E. Agbehadji, T. Mabhaudhi, J. Botai, and M. Masinde, "A systematic review of existing early warning systems' challenges and opportunities in cloud computing early warning systems," *Climate*, vol. 11, no. 9, p. 188, 2023, doi: https://doi.org/10.3390/cli11090188.

[6] I. M. Hayder *et al.*, "An intelligent early flood forecasting and prediction leveraging machine and deep learning algorithms with advanced alert system," *Processes*, vol. 11, no. 2, p. 481, 2023, doi: https://doi.org/10.3390/pr11020481.

[7] M. M. Alshater, I. Kampouris, H. Marashdeh, O. F. Atayah, and H. Banna, "Early warning system to predict energy prices: the role of artificial intelligence and machine learning," in *Annals of Operations Research*, Springer, 2022, pp. 1–37. doi: https://doi.org/10.1007/s10479-022-04908-9.

[8] M. E. Zmijewski, "Methodological issues related to the estimation of financial distress prediction models," *J. Account. Res.*, vol. 22, no. 22, pp. 59–82, 1984, doi: https://doi.org/10.2307/2490859.

[9] H. D. Platt and M. B. Platt, "Predicting corporate financial distress: Reflections on choice-based sample bias," *J. Econ. Financ.*, vol. 26, no. 2, pp. 184–199, 2002, doi: https://doi.org/10.1007/BF02755985.

[10] H. Fu and R. Chen, "Application of the Merton Model and the Altman Z-score Model in Credit Risk Assessment," Lund University, 2023. [Online]. Available: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9118955&fileOId=9118972

[11] N. Lawrence and A. Hyvärinen, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models.," *J. Mach. Learn. Res.*, vol. 6, no. 11, pp. 1783–1816, 2005, [Online]. Available: https://www.jmlr.org/papers/volume6/lawrence05a/lawrence05a.pdf

[12] M. Hussain and A. Gunasekaran, "An institutional perspective of non-financial management accounting measures: a review of the financial services industry," *Manag. Audit. J.*, vol. 17, no. 9, pp. 518–536, 2002, doi: https://doi.org/10.1108/02686900210447524.

[13] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine learning for financial risk management: a survey," in *Ieee Access*, IEEE, 2020, pp. 203203–203223. doi: https://doi.org/10.1109/ACCESS.2020.3036322.

[14] A. Callens, D. Morichon, S. Abadie, M. Delpey, and B. Liquet, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Appl. Ocean Res.*, vol. 104, no. 11, p. 102339, 2020, doi: https://doi.org/10.1016/j.apor.2020.102339.

[15] S. Nawar and A. M. Mouazen, "Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon," *Sensors*, vol. 17, no. 10, p. 2428, 2017, doi: https://doi.org/10.3390/s171024428.

[16] S. Islam and S. H. Amin, "Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques," *J. Big Data*, vol. 7, no. 1, p. 65, 2020, doi: https://doi.org/10.1186/s40537-020-00345-2.

[17] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, no. 12, pp. 804–818, 2015, doi: https://doi.org/10.1016/j.oregeorev.2015.01.001.

[18] E. B. Mallinguh and Z. Zéman, "Financial distress, prediction, and strategies by firms: A systematic review of literature," *Period. Polytech. Soc. Manag. Sci.*, vol. 28, no. 2, pp. 162–176, 2020, doi: https://doi.org/10.3311/PPso.13204.

[19] J. Sun, H. Li, Q.-H. Huang, and K.-Y. He, "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches," *Knowledge-Based Syst.*, vol. 57, no. 2, pp. 41–56, 2014, doi: https://doi.org/10.1016/j.knosys.2013.12.006.

[20] J. Bonello, X. BrÉdart, and V. Vella, "Machine learning models for predicting financial distress," *J. Res. Econ.*, vol. 2, no. 2, pp. 174–185, 2018, doi: https://doi.org/10.24954/JORE.2018.22.

[21] M. Elhoseny, N. Metawa, G. Sztano, and I. M. El-Hasnony, "Deep learning-based model for financial distress prediction," in *Annals of Operations Research*, Springer, 2022, pp. 1–23. doi: https://doi.org/10.1007/s10479-022-04766-5.

[22] N. Zhang *et al.*, "Forest height mapping using feature selection and machine learning by integrating multi-source satellite data in Baoding City, North China," *Remote Sens.*, vol. 14, no. 18, p. 4434, 2022, doi: https://doi.org/10.3390/rs14184434.

[23] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *Eur. J. Oper. Res.*, vol. 274, no. 2, pp. 743–758, 2019, doi: https://doi.org/10.1016/j.ejor.2018.10.024.

[24] Y. Cao, X. Liu, J. Zhai, and S. Hua, "A two-stage Bayesian network model for corporate bankruptcy prediction," *Int. J. Financ. Econ.*, vol. 27, no. 1, pp. 455–472, 2022, doi: https://doi.org/10.1002/ijfe.2162.

[25] M. F. Boyraz, "An empirical study on early warning systems for banking sector," Middle East Technical University, 2012. [Online]. Available: https://open.metu.edu.tr/handle/11511/21467

[26] Z. Hussain, S. Khan, M. Imran, M. Sohail, S. W. A. Shah, and M. de Matas, "PEGylation: a promising strategy to overcome challenges to cancer-targeted nanomedicines: a review of challenges to clinical transition and promising resolution," *Drug Deliv. Transl. Res.*, vol. 9, no. 11, pp. 721–734, 2019, doi: https://doi.org/10.1007/s13346-019-00631-4.

[27] P. Hajek and M. Munk, "Corporate financial distress prediction using the risk-related information content of annual reports," *Inf. Process. Manag.*, vol. 61, no. 5, p. 103820, 2024, doi: https://doi.org/10.1016/j.ipm.2024.103820.

[28] X. Li *et al.*, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowl. Inf. Syst.*, vol. 64, no. 12, pp. 3197–3234, 2022, doi: https://doi.org/10.1007/s10115-022-01756-8.

[29] H. Aljawazneh, A. M. Mora, P. García-Sánchez, and P. A. Castillo-Valdivieso, "Comparing the performance of deep learning methods to predict companies' financial failure," in *IEEE Access*, IEEE, 2021, pp. 97010–97038. doi: https://doi.org/10.1109/ACCESS.2021.3093461.

[30] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review," in *Ieee Access*, IEEE, 2021, pp. 72504–72525. doi: https://doi.org/10.1109/ACCESS.2021.3096799.
[31] R. Hammad, Z. Khan, F. Safieddine, and A. Ahmed, "A review of learning theories and models underpinning technology-enhanced learning artefacts," *World J. Sci. Technol. Sustain. Dev.*, vol. 17, no. 4, pp. 341–354, 2020, doi: https://doi.org/10.1108/WJSTSD-06-2020-0062.
[32] T. Iyamu, "Underpinning theories: Order-of-use in information systems research," *J. Syst. Inf. Technol.*, vol. 15, no. 3, pp. 224–238, 2013, doi: https://doi.org/10.1108/JSIT-11-2012-0064.
[33] H. Kallio, A. Pietilä, M. Johnson, and M. Kangasniemi, "Systematic methodological review: developing a framework for a qualitative semi-structured interview guide," *J. Adv. Nurs.*, vol. 72, no. 12, pp. 2954–2965, 2016, doi: https://doi.org/10.1111/jan.13031.
[34] A. Van den Berg and M. Struwig, "Guidelines for researchers using an adapted consensual qualitative research approach in management research," *Electron. J. Bus. Res. Methods*, vol. 15, no. 2, pp. pp109-119, 2017, [Online]. Available: https://academic-publishing.org/index.php/ejbrm/article/view/1361
[35] A. Laurent *et al.*, "Methodological review and detailed guidance for the life cycle interpretation phase," *J. Ind. Ecol.*, vol. 24, no. 5, pp. 986–1003, 2020, doi: https://doi.org/10.1111/jiec.13012.
[36] Y.-P. Huang and M.-F. Yen, "A new perspective of performance comparison among machine learning algorithms for financial distress prediction," *Appl. Soft Comput.*, vol. 83, no. 10, p. 105663, 2019, doi: https://doi.org/10.1016/j.asoc.2019.105663.
[37] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, no. 15, pp. 405–417, 2017, doi: https://doi.org/10.1016/j.eswa.2017.04.006.
[38] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, and A. Suvas, "Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model," *J. Int. Financ. Manag. Account.*, vol. 28, no. 2, pp. 131–171, 2017, doi: https://doi.org/10.1111/jifm.12053.
[39] M. Boďa and V. Úradníček, "The portability of Altman's Z-score model to predicting corporate financial distress of Slovak companies," *Technol. Econ. Dev. Econ.*, vol. 22, no. 4, pp. 532–553, 2016, doi: https://doi.org/10.3846/20294913.2016.1197165.
[40] A. Pramudita, "The Application of Altman Revised Z-Score Four Variables and Ohlson O-Score as A Bankruptcy Prediction Tool in Small and Medium Enterprise Segments in Indonesia," in *5th Global Conference on Business, Management and Entrepreneurship (GCBME 2020)*, Atlantis Press, 2021, pp. 132–135. doi: https://doi.org/10.2991/aebmr.k.210831.027.
[41] P. Morales-Bañuelos, N. Muriel, and G. Fernández-Anaya, "A modified Black-Scholes-Merton model for option pricing," *Mathematics*, vol. 10, no. 9, p. 1492, 2022, doi: https://doi.org/10.3390/math10091492.
[42] N. Coelen, "Black-Scholes Option Pricing Model," *Recuper. http//ramanujan. math. trinity. edu/tumath/research/studpapers/s11. pdf*, vol. 6, no. 6, pp. 1–19, 2002.
[43] A. K. Karagozoglu, "Option Pricing Models: From Black-Scholes-Merton to Present.," *J. Deriv.*, vol. 29, no. 4, pp. 61–73, 2022, doi: 10.3905/jod.2022.1.158.
[44] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
[45] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.
[46] J. Harrell Frank E and F. E. Harrell, "Binary logistic regression," in *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, Springer, 2015, pp. 219–274. doi: https://doi.org/10.1007/978-3-319-19425-7_10.
[47] M. Tranmer and M. Elliot, "Binary logistic regression," in *Cathie Marsh for census and survey research, paper*, vol. 20, 2008, pp. 90033–90039.
[48] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, p. 272, 2012.
[49] M.-Y. Chen, "Predicting corporate financial distress based on integration of decision tree classification and logistic regression," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11261–11272, 2011, doi: https://doi.org/10.1016/j.eswa.2011.02.173.
[50] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. 12, p. 21, 2013, doi: https://doi.org/10.3389/fnbot.2013.00021.
[51] W. Liu, H. Fan, and M. Xia, "Credit scoring based on tree-enhanced gradient boosting decision trees," *Expert Syst. Appl.*, vol. 189, no. 1, p. 116034, 2022, doi: https://doi.org/10.1016/j.eswa.2021.116034.
[52] O. Bajo-Rubio, S. Sosvilla-Rivero, and F. Fernndez-Rodrguez, "Non-linear forecasting methods: Some applications to the analysis of financial series," in *Progress in Economics Research*, Nova Publishers, 2002, p. 77.
[53] A. Beltratti, S. Margarita, and P. Terna, *Neural networks for economic and financial modelling*.

International Thomson Computer Press London, UK, 1996. [Online]. Available: https://jasss.soc.surrey.ac.uk/2/2/review3.html

[54] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002, doi: https://doi.org/10.1613/jair.953.

[55] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13*, Springer, 2009, pp. 475–482. doi: https://doi.org/10.1007/978-3-642-01307-2_43.

[56] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding," in *2006 8th international Conference on Signal Processing*, IEEE, 2006, pp. 23–32. doi: https://doi.org/10.1109/ICOSP.2006.345752.