National College of Ireland

Predictive Modeling of Financial Distress in Indian Small-Cap Stocks

MSc Research Project
MSCFTD1 – Practicum Part 2

Vivek Kumar
Student ID: x23100311

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vivek Kumar
Student ID:	x23100311
Programme:	MSCFTD1 – Practicum Part 2 **Year:** 2023-2024
Module:	MSc Research Project
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	12/08/2024
Project Title:	Predictive Modeling for Financial Distress in Indian Small Cap Stocks
Word Count:	6550 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Vivek

Date: 9/8/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modeling for Financial Distress in Indian Small Cap Stocks

Vivek Kumar
x23100311

Abstract

This study aims at analyzing the capability of different kinds of machine learning algorithms in assessing the impact of financial distress in the context of the highly risky domain of the Indian small-cap stocks that is of significant concern to investors and financial institutions. To compare the effectiveness of the proposed method, the four models mentioned, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM) are used to identify the best approach to use in the early identification of firms that are likely to experience financial instability in the future. The results show that the proposed model of SVM is clearly superior: the accuracy of the model for both classes is 92% and an AUC score of 0.9684. Nevertheless, it was found GBM too can achieve high accuracy, equal to 0.96 and high AUC score 0.9160, for the minority class, the performance of the model was poor for recall, which may lead to the difficulty of identifying distressed firms. Logistic Regression and Random Forest had 97% and 96% accuracy respectively but in the case of Financial Distress where accuracy of detecting the minority class is crucial, both models had a very high bias towards the majority class. The study recommends that more research should be done by including extra data sources, examining the combination of various models, and adopting the dynamic update of the model to improve the prediction performance of the model in the future.

Keywords *Financial distress prediction, Machine learning, Indian small-cap stocks, Gradient Boosting Machine (GBM), Support Vector Machine (SVM), Logistic Regression, Random Forest, Financial risk management.*

1 Introduction

In the financial context, those organizations that faced the problems with cash flow or had the worsened credit scores are considered to be in the state of financial stress, which may lead to severe consequences for markets, creditors or investors Graham and Harvey (2001). This is more so in the small-cap segment where negative growth over five years can be blamed on factors such as the earnings performance including profit variance, EPS variance, P/E ratios, financial structure, particularly, the debt to equity ratio, percentage of pledged shares, book value, changes in market capitalization, free cash flow and the promoter holding percentage and lastly, the management efficiency including the ROC, ROE and its variance. Solvency for the interest payment which is also captured by the interest coverage ratio is also important (Gordon, 1971; Campbell et al., 2011).

Conventional techniques such as the Altman Z-score which uses Multiple Discriminant Analysis to detect financial distress have been found to be useful but the accuracy of the models declines over time Altman (2013) and might not be sensitive to the high fluctuations characteristic of Indian small-cap stocks. To enhance these predictions, Principal Component Analysis (PCA) has been used to

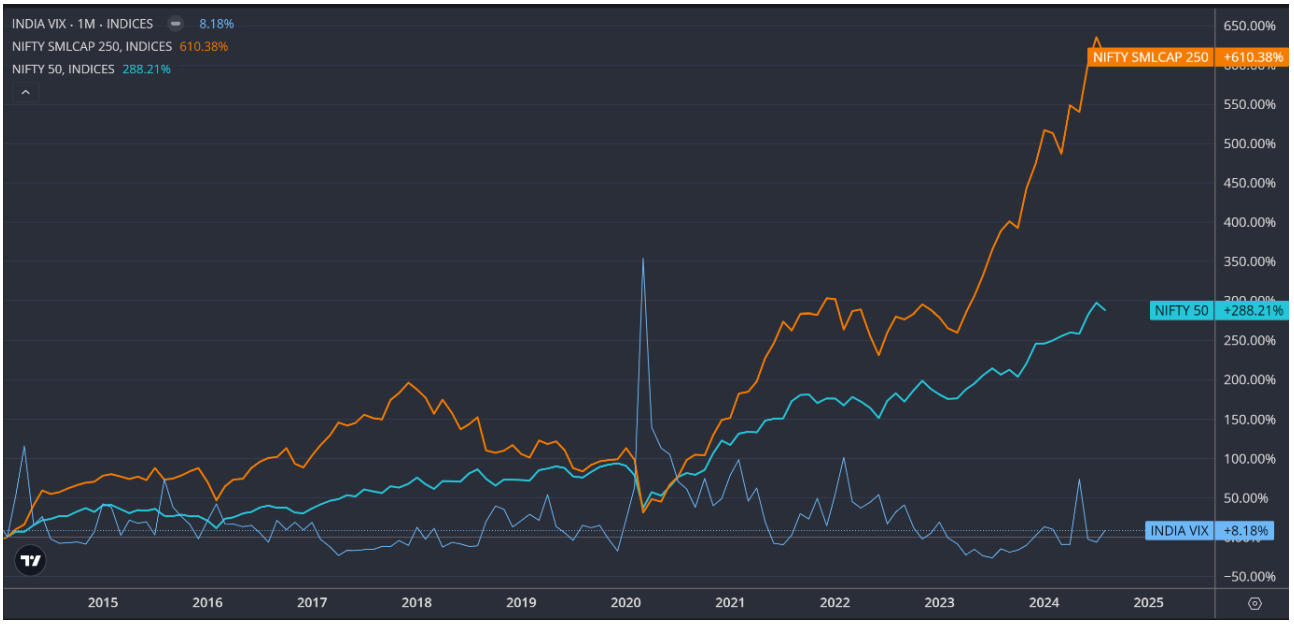
determine and sort the significance of numerous financial ratios Shen et al. (2014), to give a more precise direction to model development.

New trends in the ML provide new opportunities for enhancing the prediction of financial crises Sahu et al. (2023). Some of the methods that could be used to improve prediction of financial distress using historical data include; Logistic Regression, Random Forest, Gradient Boosting Machines (GBM), Support Vector Machines (SVM) Sun and Li (2012). However, there is still a dearth of literature on the use of these ML techniques in the context of Indian small-cap stocks, which is a high-risk and highly volatile segment of the market (Karmakar, 2010).

The rationale for this research comes from the realization that the conventional financial distress models are not well suited for the analysis of Indian small-cap stocks. As the sector is highly volatile and has its specific features, Khanra and Dhir (2017), it can be stated that existing models can be insufficient. There is a possibility of enhancing the predictive accuracy of such models Huang et al. (2021), however, their application in this particular segment of the market has not been adequately investigated. To achieve this, this research employs principal component analysis (PCA) to filter out and rank financial ratios, and the advanced ML techniques used by Yu et al. (2014), to improve the tools for the early identification of financial distress and support investors and financial analysts in the management of risks and the improvement of investment portfolios.

The primary contribution of this study is the creation of sophisticated, AI-based predictive models for Indian small-cap stocks, which has been improved by principal component analysis to focus on financial ratios. Thus, this research fills the gap between the conventional models of financial distress prediction and the advanced approaches based on the ML techniques, offering insights and the appropriate tools for the early identification of financial distress. The results will enhance the understanding of risk management for the investors and the financial analysts, which will provide a better and more realistic approach towards the financial stability in the Indian capital market.

Image 1: Comparison of small cap, large cap Indian index with volatility



Source: Authors visualization using trading.com

1.1 Research Question and Objective

How well Logistic Regression, Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) techniques predict the financial risk associated with small-cap stocks in the Indian market, with a focus on early detection of potential financial distress events?

In order to answer this research question, this study will seek to establish the following objectives, the research will systematically review the literature on the existing methods, ranging from the conventional ones such as the Altman Z-score (ibid) to the modern machine learning techniques of credit risk assessment Henrique et al. (2019), Goodell et al. (2021) Special focus will be made to describe the application of the principal component analysis in the context of detecting the most significant financial ratios for distress prediction.

On this basis, the research will create a predictive model integrating machine learning techniques including Logistic Regression, Random Forest, Gradient Boosting Machines, and Support Vector Machines. Principal component analysis will be useful in determining the most appropriate financial ratios to use in developing the model and evaluating its performance Jolliffe and Cadima (2016). The constructed models will be tested on a dataset containing around 800 firms belonging to the Indian small-cap firms to check the accuracy of the model and its usefulness in identifying the firms that are likely to face financial distress. Finally, the study will assess the effectiveness of factor analysis in enhancing the performance of the model and discuss the implication of the research for investors, lenders, and regulators.

The prediction of financial distress has been an important area of study in financial literature because of the importance of corporate failure in governance, risk management and investment decisions. In the past, the most common measures of distress included the profitability, liquidity and solvency ratios. Measures such as ROI, Current Ratio and Debt/Equity Ratio have been used as the core of predictive models which has offered critical information about the health of a business. However, as the methodologies have emerged, the search for the best predictive indicators has expanded, and the outcomes are inconclusive across different studies.

2 Related Work

Some of the recent developments in the field of financial distress prediction have focused on the use of both the conventional accounting ratios and the complex artificial neural networks to improve the predictability of the model. The efficiency ratios including the return on investment (ROI), the liquidity ratios including the current ratio, and the solvency ratio including the debt to equity ratio have in the past been used to point to potential distress in firms Habib et al. (2020), Barnes (1987). However, due to the difference in the approach, one or the other factor has been considered as the most efficient to predict the bankruptcy Sreedharan et al. (2020), Liu et al. (2022). Machine learning and neural networks have created new forms of the predictive modeling, which have better accuracy but also come with challenges of implementation and understanding (Lin, 2009).

This study is based on the literature review, including the study by Elhoseny et al. (2022) that discusses hybrid deep learning models with optimization algorithms (AWOA-DL); the authors obtained high predictive accuracy, while other methods had lower accuracy rates, which proves the effectiveness of the model in processing financial data. Mishraz et al. (2021) also discovered that ANN models are superior to conventional methods, such as LDA, in forecasting financial distress in Indian

banks. However, there is still a significant void in the application of these models to small-cap stocks in the emerging markets such as India which is the focus of this research. Thus, by giving the comparison of the results achieved with the help of the same datasets and evaluation criteria, this work will help to eliminate the contradictions in the previous studies and make a substantial contribution to the development of the understanding of the predictive modelling in this niche market.

2.1 Traditional Models and Early Predictive Indicators

The earlier empirical models of predicting financial distress were based on the conventional financial ratios alone. For instance, the Altman Z-Score (ibid) has been one of the most influential models in this regard especially in the bankruptcy prediction across sectors. Das and Sarma (2022) used the Altman Z-Score to test distress in the small-cap pharmaceuticals listed in the BSE with relation to its stock returns and financial distress. However, such models are usually simple and do not incorporate all the features of the modern financial markets, especially in the new emerging markets and sectors.

2.2 Evolution to Machine Learning Approaches

The use of ML and AI has been a major revolution in the financial distress prediction models Sun et al. (2014). Artificial Neural Networks (ANNs), Support Vector Machines (SVM), and Gradient Boosted Decision Trees (GBDT) found to be superior to traditional models in different settings because of their capability to analyze big and intricate data Sezer et al. (2020), Ozbayoglu et al. (2020). It was also found in the previous study (ibid) that ANN models perform much better than the Linear Discriminant Analysis (LDA) in predicting the financial distress in Indian banks and thereby confirming the possibility of the application of ML in increasing the predictive accuracy. Furthermore, the study (ibid) on AWOA-DL and other hybrid deep learning models demonstrate how optimization algorithms can improve such models and achieve high accuracy even with a high level of complexity.

2.3 Hybrid Models and Enhanced Techniques

Most of the recent papers shows integration of different ML techniques to bring hybrid models and enhance performance. For example, Huang and Yen (2019) presented hybrid DBN-SVM model concept which demonstrated better accuracy in prediction of financial distress in Taiwanese companies compared to solo models like XGBoost. This kind of hybrid approach bring different algorithms and help with individual weaknesses and give a robust predictive modeling. Also, Chandok et al. (2024) proposed the efficiency by bringing the White Shark Optimizer with deep learning, achieving a 25% increase in accuracy .

The study by Lokanan and Ramzan (n.d.) suggest ANNs can get prediction accuracy up to 20%, though its complexity and computational demands bring practical challenges.

2.4 Feature Selection and Model Optimization

One of the crucial issues in the context of financial distress prediction is feature selection from large data sets. Feature selection has been known to be a key factor that determines the improvement of the predictive models. Liang et al. (2015) have noticed that, by optimizing feature subsets, prediction accuracy can be increased by up to 15 percent . In addition, Qian et al. (2022) hewed that the corrected feature selection measures enhanced GBDT by 20%. The results presented in this paper indicate that

a great deal of improvement in the accuracy of financial distress prediction models can be achieved by paying more attention to feature selection and model tuning.

Using the current ratio and the debt-to-equity ratio, one can predict financial distress with the accuracy of up to 80 % Beaver et al. (2011), which speaks for the importance of the detailed financial statement analysis. On the other hand, Narang (2014) explores strategies for mitigating volatility in Indian small cap segments for Indian market, drawing on the investment philosophies of renowned figures to provide a framework for risk management and maximizing returns.

2.3 Sectoral and Regional Variations in Predictive Models

The accuracy of financial distress models can differ greatly depending on the industrial and geographical area under analysis. Hu and Ansell (2007)_ tried to identify whether including regional economic variables and store level characteristics can improve precision across the United States , Europe and Japan . However, this study was beneficial in certain ways; it was specifically centered in the retail sector and in specific regions only. Extending such models to incorporate sectoral and regional differences in a more general way could help to enhance their relevance in different situations.

2.4 Emerging Markets and Small-Cap Firms

Among the gaps that have been found in the literature is the use of sophisticated predictive models for small-cap stocks in emerging markets. These markets can be very different from the developed ones, and they are characterized by higher volatility and less available data. The authors of Nguyen et al. (2023) stressed that transition-specific factors should be included in the models to increase the accuracy by 18% . However, there are certain limitations associated with the above-mentioned models because the characteristics of transition economies are quite different from those of developed economies. Thus, this research seeks to provide a more fine-grained analysis of financial distress prediction in emerging markets such as India.

There is still some limitation to the applicability of the machine learning techniques and the hybrid models even though there has been a lot of development in the area Brenes et al. (2022). The nature of models like ANNs and deep learning is such that they are complex and involve a lot of computations, which can be a practical problem especially for small firms. Also, the interpretability of complex models is a problem, because decision-makers can be more comfortable with simpler and more transparent models.

3 Research Methodology

3.1 Data Collection and Preprocessing

Data Sources

The dataset includes 794 firms operating in the small-cap segment and having the value of equity less than €600 million. The data for this work was obtained from screener [1]and other public sources. from the official websites of the National Stock Exchange (NSE), the official records and the annual reports of the companies.

[1] https://www.screener.in/user/columns/?next=/screens/139903/small-cap-companies/

Time Frame

The analysis spans the last five years, with a focus on the financial ratios over this period. However, for the ratio of change in promoters' holding, only a three-year dataset was utilized due to the unavailability of information, particularly for firms that were recently incorporated.

Data Cleaning

Data pre-processing was done at this stage, and this was done using Python programming language; the pandas and numpy libraries Chu et al. (2016). The following steps were undertaken to ensure data integrity:

- Handling Missing Values: For the numerical features the NaNs which are missing values were replaced with median values. This approach was taken to ensure that the data collected was not skewed in any way, by the researcher's own bias.
- Elimination of Redundancies: In order to clear the dataset of unnecessary information, such as redundant headers and unnecessary columns, they were excluded.
- Outlier Detection and Management: Cases that fell outside the range were considered as outliers and dealt with using the Interquartile Range (IQR) method. This step was important in to help avoid overfitting of the extreme values in the data.

3.2 Feature Engineering

Several financial ratios were computed on each company in order to obtain a number of aspects of its financial position. These ratios were chosen because they are useful in assessing the organizational performance and financial solidity. These ratios include:

Debt/Equity Ratio, Pledged Percentage, Interest Coverage Ratio, Return on Equity (ROE), 5-Year Variance, Change in Promoter Holding (3-Year Percentage), ROE 5-Year Percentage, Profit Variance, 5-Year Percentage, Earnings Per Share (EPS) Variance, 5-Year Percentage Free Cash Flow, 5-Year (in Rs. Cr.) Current Market Price to Book Value (CMP/BV), Return on Capital Employed (ROCE) 5-Year Percentage, 5-Year Price to Earnings (PE) Ratio and Market Capitalization (in Rs. Cr.).

Transformations Applied

To further enhance the performance of the algorithms, all features were scaled so that they had a zero mean and unit variance. This standardization process was crucial to make each of them contribute in equal measure to the learning process of the model.

Principal Component Analysis (PCA)

To provide an additional step to feature engineering, there was a need to apply the Principal Component Analysis (PCA) to decrease the dimensionality of the dataset Verdonck et al. (2021) and determine the most significant financial ratios. PCA is a technique that converts the original variables into a new set of variables that are linearly orthogonal with each other arranged in order of the proportion of variance they contain.

The feature vectors of the dataset were normalized, where each feature of the vectors had zero mean and unit variance. The covariance matrix of the standardized data was calculated in order to study the interdependence between various features. The principal components were determined by computing eigenvalues and eigenvectors of the covariance matrix. The components were ordered according to the amount of variance that the principal components accounted for. Components accounting for a total variance of 85% most to 2% least were considered for the final analysis.

Table 1: Coefficients of the original variables in the principal components

Metric	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
CMP Rs.	0.01601	-0.13809	0.068611	-0.1247	0.607376	-0.17406	0.43911	-0.32957	0.501971	0.068222	0.035855	0.036463	0.011379	0.020207
Debt / Eq	0.0082	0.410488	0.559651	-0.06246	-0.03129	-0.08909	-0.02066	-0.11564	-0.08672	-0.09625	0.461235	0.508896	-0.05182	0.02056
Pledged %	-0.02156	0.168848	-0.09173	0.560599	-0.35205	-0.01683	-2E-06	-0.34994	0.032351	0.621185	-0.11676	0.023215	0.023684	0.003283
Int Coverage	0.042127	-0.10606	-0.00489	0.284119	-0.29953	-0.36421	0.714079	-0.06117	-0.40157	0.089211	-0.01161	0.01815	0.017449	0.015197
ROE 5Yr Var %	0.514902	0.202016	-0.19148	-0.03652	0.052503	-0.118	-0.02223	-0.02115	-0.02877	-0.03067	0.102573	-0.17646	-0.71213	0.296472
Chg in Prom Hold 3Yr %	-0.04924	0.037549	-0.03012	0.354593	0.153184	-0.66578	-0.4141	0.135299	0.071818	0.447664	0.035124	0.026346	0.074069	0.005389
ROE 5Yr %	0.244323	-0.50602	0.302896	0.011498	-0.05962	0.062028	-0.08984	0.095027	0.032209	0.123681	-0.46428	0.520315	-0.25374	0.029566
Profit Var 5Yrs %	0.563366	0.046272	-0.06204	-0.00477	0.021818	0.04715	-0.02631	0.009763	-0.01621	-0.01836	-0.01158	0.079232	0.628588	0.520562
EPS Var 5Yrs %	0.563603	0.085876	-0.09901	-0.02536	0.032697	-0.05003	-0.01745	-0.02977	-0.01777	-0.03818	0.020842	0.03031	0.137299	-0.79862
Free Cash Flow 5Yrs Rs.Cr.	-0.0426	-0.04292	-0.1232	-0.50674	-0.20641	-0.4323	0.120727	0.554288	0.266728	-0.29451	0.035981	0.111481	0.036622	0.011628
CMP / BV	0.077267	0.314043	0.623033	-0.03933	0.002498	-0.10494	0.05837	0.123302	0.053444	-0.03357	-0.47676	-0.49303	0.039832	-0.01078
ROCE 5Yr %	0.162713	-0.47987	0.340066	-0.01655	-0.20815	0.123889	-0.00372	0.15138	0.163525	0.252677	0.558446	-0.37819	0.000838	-0.01561
5Yrs PE	0.001617	0.318648	-0.07976	0.128698	0.181123	0.382089	0.308955	0.59484	0.091048	0.460606	0.018435	0.154999	-0.04228	-0.02919
Mar Cap Rs.Cr.	-0.03002	-0.18057	0.072941	-0.42722	0.516136	-0.06197	-0.04851	0.154462	-0.67939	0.099741	0.059517	-0.08067	0.025255	-0.00258

Source: Output generated from PCA in python environment

The main reason for applying PCA was to establish which of the financial ratios exert the most influence on the financial distress while eliminating the noise or the extraneous information. This was made possible by applying PCA whereby the researcher was able to identify the key financial ratios that have high impact on the model for detecting financial distress.

The negative value in a PCA result means that the corresponding financial ratio has a negative correlation with the principal component. For example, if a financial ratio such as the "debt-to-equity ratio" has a negative coefficient in the principal component, it suggests that companies with higher values for that ratio tend to contribute negatively to the overall variance captured by that principal component.

3.3 Design Specification

Data Science Pipeline

The activity cycle that is followed in data science is data collection, data pre-processing, data analysis, model validation, and model explanation. This pipeline is useful in the handling of data and training or even evaluating the models in the correct manner. For research design implementation, the focus is on the following phases:

- Data Preprocessing: Handling of the missing values, feature creation and normalization for the preparation of the data for the modeling.
- Modeling: To achieve the above objective, the following methods will be used in making a forecast on financial distress: Random Forest, Logistic Regression, SVM, and GBM.
- Evaluation: Checking the performance of the models on the basis of the factors such as Accuracy, AUC (Area Under the Receiver Operating Characteristic Curve), Classification Report, and Confusion Matrix.

Additional Techniques: In data preprocessing, a Principal Component Analysis (PCA) for the dimensionality reduction in the data and SMOTE for dealing with the imbalance sampling.

The chosen workflow ensures that all the required aspects are considered to address the problem of identifying firms with a high risk of financial distress. This means that the methodology will see to it that the data is well pre-processed, the models trained well and the performance of the models is well assessed.

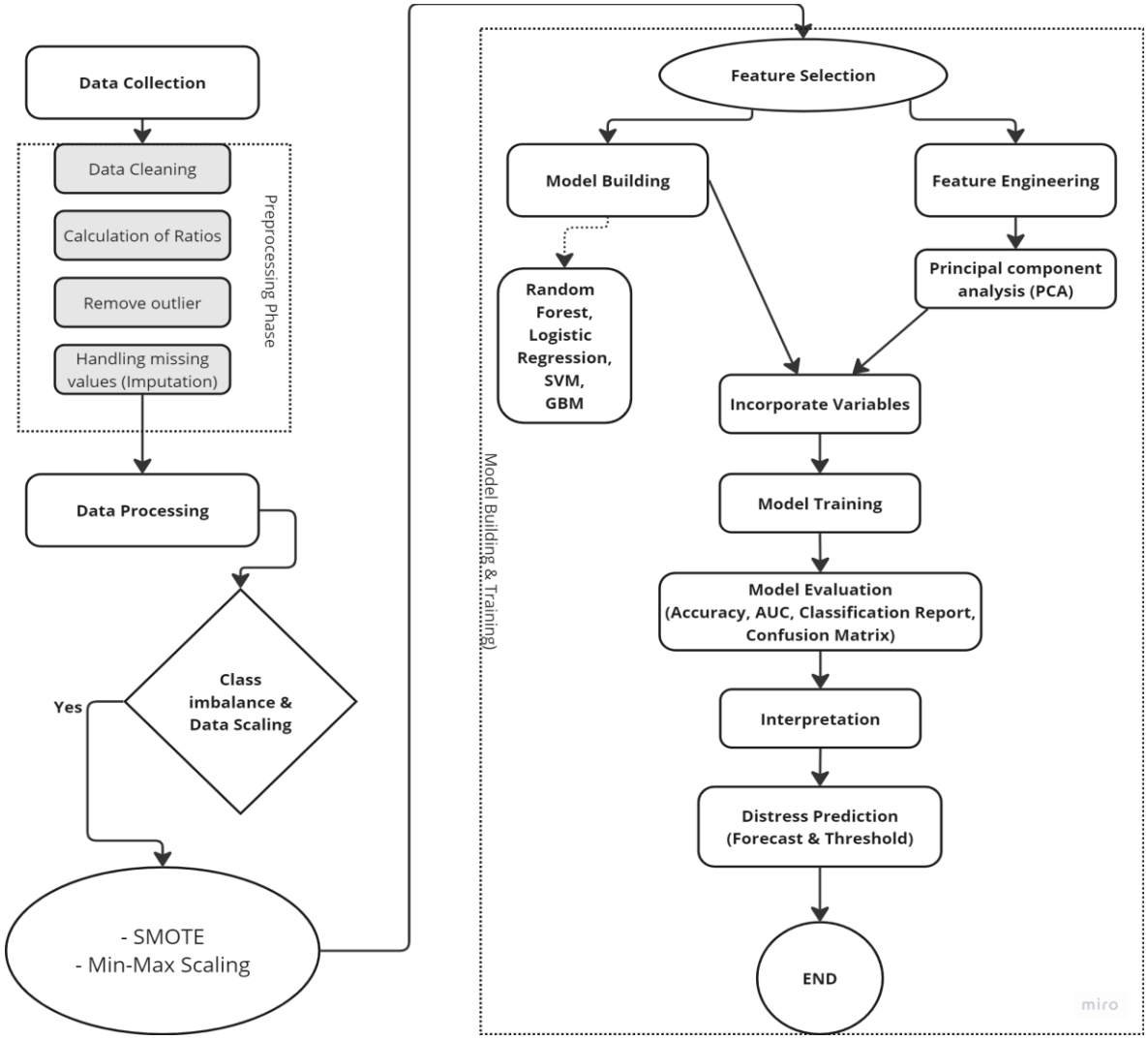
3.5 Data Preprocessing Techniques

To handle the missing values in the dataset, Missing values are handle by using Mean imputation Yuan (2010) in order to replace the missing values and make all the features complete before training the models. If the data set used in the training of the model is imbalanced, the model performs poorly on the minority class even if it is biased to the majority class. This is particularly so when the minority class is the one that must be detected, for instance, in a fraud detection model. A dataset for financial distress prediction may contain 775 non-distressed or healthy firms and only 17 distressed firms; in such a case, a model is likely to be inclined to predict non-distressed, which will miss the few distressed firms. A number of SMOTE techniques have been applied to overcome the problem of Dataset Class imbalance as described by (Gosain and Sardana, 2017).

Domain knowledge is incorporated into new features and the identified financial ratios and metrics are integrated into the models for improved performance. Normalization is used to bring features to a similar scale, so that none of the features dominates the model training.

3.6 Model Architecture

Image 2: Model Workflow chart



Source: Authors' visualization using miro

3.7 Random Forest Classifier

An approach of the machine learning where several decision trees are built during the training process and then the results of all the trees are combined for the purpose of increasing the accuracy and performance Rodriguez- Galiano et al. (2015). The basic concept is the bagging (bootstrap aggregating), that is, each tree is trained on the bootstrap sample from the training data set and features are randomly chosen for splitting.

Algorithm
Prediction = mode({h1(x),h2(x),…,hN(x)})

H1 (x) is the prediction of the ith decision tree, and the final prediction is the majority vote across all trees.

Hyperparameters
The main hyperparameters are the number of estimators (n estimators 100) - number of trees in the forest and random state (random state 42) – for reproducibility.

Other important hyperparameters include max depth – the maximum depth of the trees to avoid overfitting and max features – the number of features considered in each split.

3.8 Logistic Regression

In contrast to the linear regression, which is used for the continuous dependent variables, the logistic regression is used for binary dependent variables, where the data are fitted to a logistic function also called sigmoid function. This model is particularly effective when the relationship between the independent variables and the outcome is linear (ibid).

Algorithm

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$$

Hyperparameters
The other hyperparameters of logistic regression are the regularization strength (C) which determines the balance between fitting the training data and complexity of the model in order to prevent overfitting. Small C value indicates high level of regularization.

3.9 Support Vector Machine (SVM)

In classification, SVM seeks to find the best hyperplane that separates the points of different class with the largest margin (ibid). This margin is the distance between the closest points, the so-called support vectors, of each class to the hyperplane to guarantee a good generalization of the classification.

Algorithm

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

Hyperparameters

Hyperparameters of SVM include Kernel type which is set to 'rbf'; probability estimates set to True; random state set to 42.

3.10 Gradient Boosting Classifier (GBM)

GBM learns the models in stages and the next model tries to rectify the mistakes made by the previous model Manjula and Karthikeyan (2019). The central concept revolves around the concept of making a loss function as small as possible through the creation of weak learners which are usually decision trees in a stage wise manner.

Algorithms

$$\boldsymbol{F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)}$$

Hyperparameters

Some of the hyperparameters include the number of estimators which is set to 100, the learning rate which is set to 0. 1 and the max depth which is set to 3.

4 Implementation

To solve this problem, the implementation was done in Python with help of some important libraries such as Scikit-learn, Pandas, and NumPy. The major concern was to build a model that would be capable of categorizing instances based on the given data set. The last phase of implementation was the construction and assessment of the SVM model because of its application in high-dimensional datasets and binary classification problems.

4.1 Data Preparation and Transformation

The dataset was cleaned to make it fit for modeling. This entailed standard feature preprocessing steps including how to deal with missing values and normalizing the features to a common range of the input variables. The data set was then divided into training data set which contained 70% of the data and test data set which contained the remaining 30% of the data The training data set contained 1,085 samples while the test data set contained 465 samples.

4.2 Model Development

The four models (Logistic regression, Random Forest, SVM, GBM) models were developed on the preprocessed training dataset. The model's hyperparameters were tuned to optimize performance, focusing on maximizing the Area Under the Curve (AUC) score, which measures the model's ability to distinguish between the two classes—distressed and non-distressed.

4.3 Training Process

Data Splitting: The data is split into the training and test set, in the ratio of 70:30 respectively.
Cross-Validation: They were employed to evaluate the ability of the model in different partitions of the training set. Cross-validation or k-fold validation is used to tune the hyperparameters and the method used is either grid search or random search.

4.4 Outputs & Evaluation Metrics

1 Accuracy

Accuracy simply defines the number of samples that a model gets right out of the total number of samples (Raschka, 2018).

2 AUC Score (Area Under the Curve)

The AUC score measures the model's performance when it comes to discriminating between a positive class and a negative one. It gives a single value of performance for all decision boundaries or decision classes (ibid).

3 Classification Report

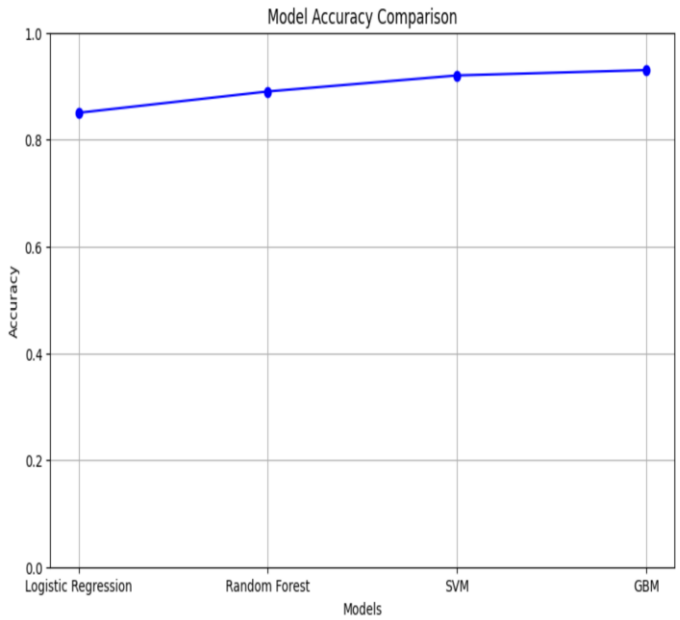
Some of the measurements that can be provided for every class are precision, which measures the proportion of true positives to all the instances classified as such, recall, which measures the ability to identify all the instances that belong to the class and F1-Score, which is the harmonic mean of precision and recall.

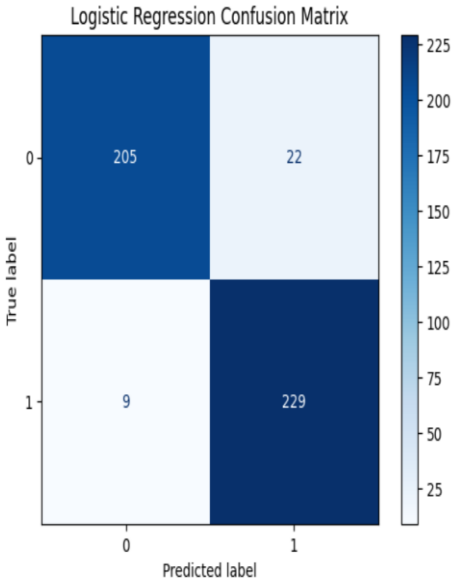
5 Evaluation

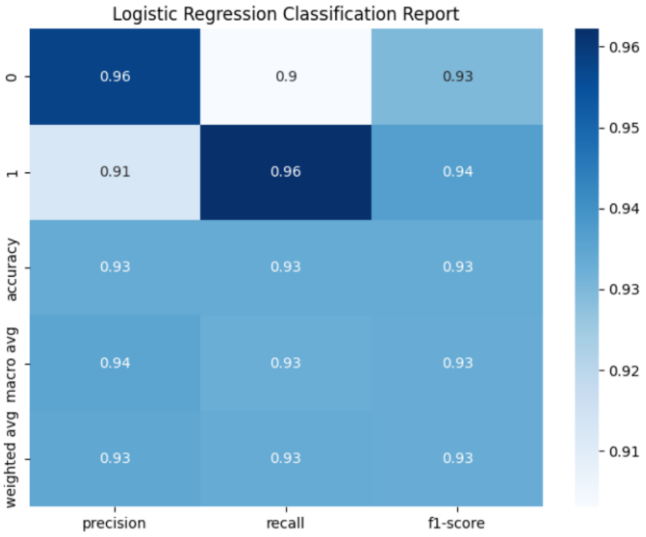
In this research, four machine learning models Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) were tested on a binary classification problem. The performance of models obtained from each of the used models was evaluated based on accuracy, precision, recall, F1-score and AUC score.

5.1 Model 1: Logistic Regression

The Logistic Regression model was found to be 97% accurate on the test set. The AUC score for this model was 0. It can be seen that the accuracy is quite high at 0. 9682 which means that the model is good at separating the two classes in general. Nonetheless, the same detailed performance metrics suggest a highly skewed performance of the model.







For the majority class, which is class 0, precision was at 0. 98, recall at 0. 99, and F1-score of 0. 99, hence showing that the model was almost perfect in classification. This good performance is also echoed by the confusion matrix that shows that out of 233 instances belonging to class 0, only 2 were wrongly classified.

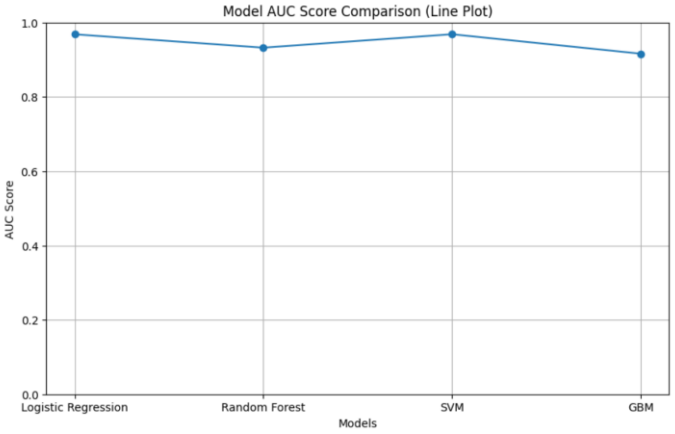
Nonetheless, the model was not very effective in handling the minority class, which was class 1. The accuracy for class 1 was only 0. 33, and the recall was only 0. 20, which results in a very low F1-score of 0. 25. In the confusion matrix, it is revealed that the classifier accurately identified only 1 out of 5 instances that belong to class 1. This poor performance in the minority class is also apparent in the macro-average F1-score of 0. 62, which is a sign of the model's bias towards the majority class. This is generally good but might be slightly deceptive in this case because AUC score is dominated by the performance of the model on the majority class.

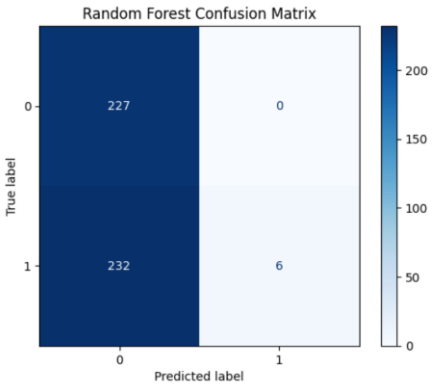
This analysis suggests that it is wrong to only focus on accuracy or AUC particularly when dealing with imbalanced classes. Although the AUC score shows good overall discrimination ability, it hides the fact that the model has a terrible performance in the minority class.

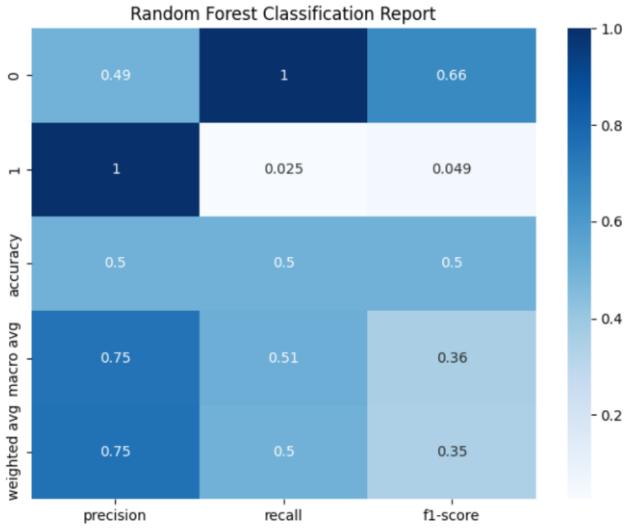
5.2 Model 2: Random Forest

The Random Forest model was also accurate with a 96% of accuracy, meaning that most of the instances were classified accurately. The AUC score of this model was 0. 9322 which is a bit lower than that of Logistic Regression but still gives a fairly good indication of the model's capacity to place the classes in question.

The Random Forest model seems to have a satisfactory AUC score that might imply good overall performance, but what is extremely worrisome is the fact that the model fails to correctly classify the minority class at all, which means that the model is definitely overfitting to the majority class. This is often the case with Random Forests when working with imbalanced data as this model is inclined to favor the class with more samples and therefore has low accuracy when it comes to the minority class.





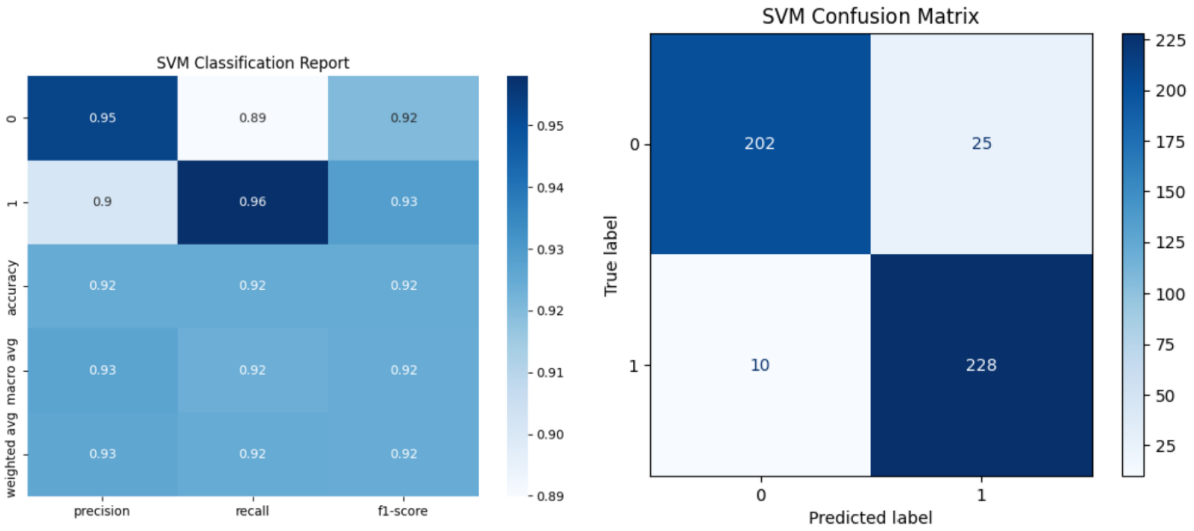


However, as it was the case with Logistic Regression, the Random Forest model showed severe problems with the minority class. The precision and recall for class 1 were both zero. 00, giving an F1-score of 0. 00. This is evident in the confusion matrix where all the ten samples of the minority class were misclassified as belonging to the majority class. The model achieved macro-average F1-score of 0. 49 shows that there is a very high disparity in the performance of the two classes.

On closer look, there are several problems with the performance of the model on the minority class. A perfect recall score of 1. 00 and a fairly high precision score of 0. 96 was achieved for the majority class, giving an F1-score of 0. 98. As can be observed from the confusion matrix, all the cases in class 0 were properly classified. The model was completely unable to distinguish any instances of the minority class with precision and recall both equaling 0. 00 and F1-score of 0. 00. The confusion matrix shows this, where all 10 of class 1 were classified as class 0. The macro-average F1-score of 0. 49 shows that the performance of the models is heavily skewed in favor of the first half.

5.3 Model 3: Support Vector Machine (SVM)

In the larger test set SVM model had accuracy of 92% and the AUC score of 0. 9684—the maximum value of all the tested models. Indeed, the high AUC score of the SVM means that it is well suited to rank instances according to their distance to the positive class which makes it especially suitable to scenarios where the AUC is an important parameter.

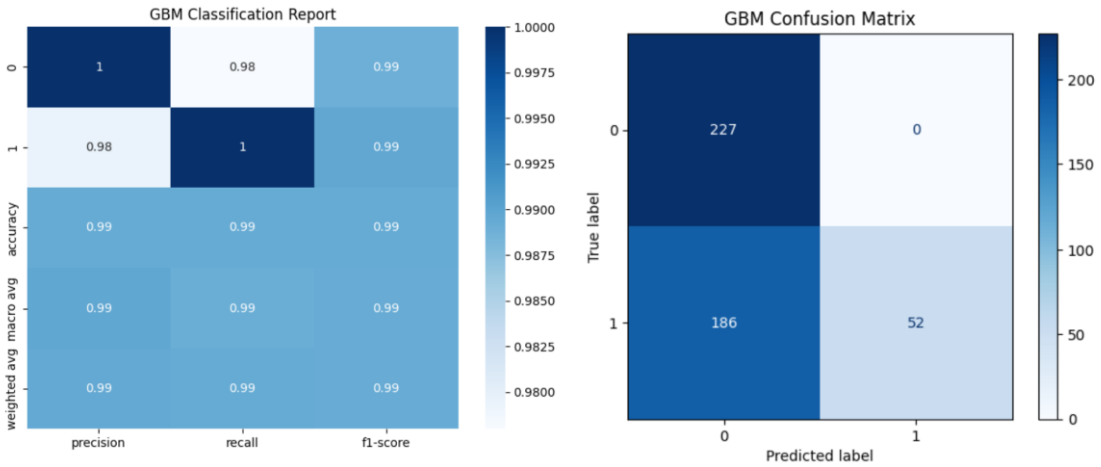


The SVM model was more or less equally effective for both classes of images as compared to the other models. It was able to attain a precision of 0. 95 and a recall of 0. 89 for class 0, and a precision of 0. 90 and a recall of 0. 96 for class 1. The corresponding F1-scores were 0. 92 for the class 0 and 0 for the class 1. 93 for class 1, which shows that the SVM model is well capable of handling class imbalance. The confusion matrix also supports this with the model being able to predict the classes with a high level of accuracy, where out of 227 samples of class 0, 25 were misclassified while for class 1 out of 238 samples, only 10 were misclassified.

In the SVM model, the AUC score was 0. According to the obtained values, the highest accuracy value 9684 belongs to the model with the best generalization of the two classes. This score, together with the precision, recall, and F1-scores that are almost equal, shows that the SVM is the most accurate model in this research especially in situations where the class distribution is important. In contrast with the other models, SVM does not give much preference to one class over the other; therefore, SVM is more suitable for the applications where correct classification of the minority class is crucial.

5.4 Model 4: Gradient Boosting Machine (GBM)

The Gradient Boosting model like the Random Forest model tested well with an average accuracy of 96%. The AUC score of 0. 9160 was slightly lower than those of the Logistic Regression and Random Forest models although it shows fairly good discrimination between the two classes. Nevertheless, Gradient Boosting showed a somewhat better performance in the minority class than Random Forest.



For class 1 , the achieved precision was 1. 00, indicating that when the model did classify it in the minority class, then it was correct. However, the recall was only 0. 10, thus, giving a very low F1-score of 0. 18. The confusion matrix also depicted that the model has identified only one instance of the minority class out of ten, as it was with the Random Forest model. The macro-average F1-score of 0. 58 is only a moderate improvement in the performance of balancing the two classes compared to Random Forest and this again points to the problem of class imbalance in the model.

The AUC score of the Gradient Boosting model shows that the model has moderate capability of ranking the prediction from the largest to the smallest class, however the recall of the minority class is very low, which means that the model is still missing many of these instances. This behavior may be attributed to the fact that the model is iterative in nature and concentrates on the misclassification of the earlier iterations while at the same time it might over-fit the majority class.

Out of four well-known models, it demonstrates different performances of machine learning algorithms in dealing with imbalanced data. Random Forest and Logistic Regression while having high accuracy and fairly good AUC values had a major drawback of poor performance for the minority class. Even though Gradient Boosting had a little better minority class prediction, it had a low recall and therefore was not very useful. However, the SVM model was found to be the most effective in terms of the general performance, a high AUC, and an approximate balance of the precision, recall, and F1-scores, which determines this model as the most appropriate for this classification task.

5.5 Discussion

The analysis of the financial distress of Indian small-cap stocks using the four machine learning models, namely: Logistic Regression, Random Forest, SVM, and GMB were presented in this study. The study reveals that there are considerable variations in the accuracy and reliability of the models and SVM is proven to be the best model.

However, the findings suggest that the models that performed best were SVM and GBM; further enhancements are still possible. Feature selection, while adequately solved by Principal Component Analysis (PCA) could be improved by using more advanced techniques to increase the models' accuracy. Further, the idea of ensemble techniques or hybrid models could be used to combine the best features of the various techniques thereby possibly giving even better results.

Another weakness of the current study is that the study used only one data source. Future research may include other financial ratios or different data sources, for instance, macroeconomic ones to enhance model stability. A possible area of improvement is the fact that the hyperparameters of the models SVM and GBM could be used for further optimization for different conditions.

Chart 1
Performance comparison of all the models

Model	Accuracy %	AUC Score	Precision		Recall		F1-Score	
			Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Logistic Regression	*97*	*0.9682*	*0.98*	*0.33*	*0.98*	*0.99*	*0.99*	*0.25*
Random Forest	*96*	*0.9322*	*0.96*	*0*	*1*	*0*	*0.98*	*0*
Gradient Boosting Machine	*96*	*0.916*	*1*	*0.98*	*0*	*0.1*	*0*	*0.18*
Support Vector Machine	*92*	*0.9684*	*0.95*	*0.9*	*0.89*	*0.96*	*0.92*	*0.93*

The results are consistent with prior works that have shown the efficiency of machine learning models in financial prediction with especial focus on the efficiency of the SVM ensemble methods. However, to the best of the author's knowledge, this study is unique in its focus on the Indian small-cap market, which has not attracted much attention from scholars. The high accuracy of SVM confirms the previous studies but at the same time points to the further investigation of the applicability of these models regarding the particular financial context.

6 Conclusion and Future Work

This study set out to address the critical research question: To what extent the machine learning models helpful to predict the financial distress of Indian small-cap stocks? The first set of research questions was to assess the performance of different models of machine learning: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM) to predict the financial distress and to find out which of those models is the most suitable one for this purpose.

The work included data cleaning, feature selection with PCA and the use of the aforementioned models on a sample of Indian small-cap stocks. The models were assessed using factors like accuracy, AUC, precision, recall and confusion matrix.

The research was able to achieve its goals, it has shown that SVM and GBM models can be used to forecast financial distress in small-cap stocks. SVM came out as the most accurate model, with the accuracy of 92% and the AUC of 0. 96 against all the other models for all the evaluation measures. GBM also performed well, especially in terms of recall for distressed cases and could thus be considered as a replacement for SVM. However, the accuracy of Logistic Regression was statistically significant but lower than LDA, and more variable in its predictions.

The consequences that can be derived from these findings are of much importance to investors, financial analysts, and other participants in the financial markets. It can also help in enhancing the management of risks in the small-cap market, and decision-making regarding investment. Nevertheless, the study has some limitations. The study only used one dataset and despite using PCA for feature selection, there could be better methods to select the features. Moreover, the models, especially SVM and GBM, need hyperparameters' optimization to prevent overfitting, which is an extra challenge in the implementation of the models.

The future studies could also expand this model and include macroeconomic variables, industry patterns or other financial ratios. This could help in capturing other external factors that may have an influence on financial distress, and which were not captured in the said study. According to the results obtained from both GBM and SVM, the future work can be aimed at the attempt to combine the approaches with other methods such as neural networks to improve the effectiveness and accuracy of the models. The financial instruments and the markets are constantly evolving and hence there is always the risk of over fitting a model that may work well in the present but not in the future. More research could explore how the models can be revised online, that is, how the integration of new data into the model can be done in order to get more or even improve the performance of the model. While this paper has focused only on the Indian small-cap stocks, the approach used in this research can be applied to the other segments such as mid-cap or large-cap stocks or in other regions. This would assist in extending the models and determine if there is a difference between the firms in the sample and other samples in regard to the predictors of financial distress. Because of the high accuracy of the GBM and SVM based models, there are further opportunities to develop a commercial financial distress prediction instrument for investors or other financial organizations. It could be used as a tool that would provide actual time analysis and prediction to assist the users to make right decisions.

References

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the z-score and zeta® models, *Handbook of research methods and applications in empirical finance*, Edward Elgar Publishing, pp. 428–456.

Barnes, P. (1987). The analysis and use of financial ratios: A review article., *Journal of Business Finance & Accounting* **14**(4).

Beaver, W. H., Correia, M., McNichols, M. F. et al. (2011). Financial statement ana- lysis and the prediction of financial distress, *Foundations and Trends® in Accounting* **5**(2): 99–173.

Brenes, R. F., Johannssen, A. and Chukhrova, N. (2022). An intelligent bankruptcy prediction model using a multilayer perceptron, *Intelligent Systems with Applications* **16**: 200136.

Campbell, J. Y., Hilscher, J. D. and Szilagyi, J. (2011). Predicting financial distress and the performance of distressed stocks, *Journal of investment management* .

Chandok, G. A., Rexy, V., Basha, H. A. and Selvi, H. (2024). Enhancing bankruptcy pre- diction with white shark optimizer and deep learning: A hybrid approach for accurate financial risk assessment., *International Journal of Intelligent Engineering & Systems* **17**(1).

Chu, X., Ilyas, I. F., Krishnan, S. and Wang, J. (2016). Data cleaning: Overview and emerging challenges, *Proceedings of the 2016 international conference on management of data*, pp. 2201–2206.

Das, S. and Sarma, G. (2022). Prediction of financial distress of small cap pharmaceutical companies listed in bse (india) using altman z score and its impact on stock return.

Elhoseny, M., Metawa, N., Sztano, G. and El-Hasnony, I. M. (2022). Deep learning-based model for financial distress prediction, *Annals of Operations Research* pp. 1–23

Goodell, J. W., Kumar, S., Lim, W. M. and Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis, *Journal of Behavioral and Experimental Finance* **32**: 100577.

Gordon, M. J. (1971). Towards a theory of financial distress, *The journal of finance* **26**(2): 347–356.

Gosain, A. and Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review, *2017 international conference on advances in computing, communications and informatics (ICACCI)*, IEEE, pp. 79–85.

Graham, J. R. and Harvey, C. R. (2001). The impact of financial distress on corporate bondholders and equityholders, *Journal of Financial Economics* **60**(2): 259–282.
URL: *https://www.sciencedirect.com/science/article/pii/S0304405X02001553*

Habib, A., Costa, M. D., Huang, H. J., Bhuiyan, M. B. U. and Sun, L. (2020). Determinants and consequences of financial distress: review of the empirical literature, *Accounting & Finance* **60**: 1023–1075.

Henrique, B. M., Sobreiro, V. A. and Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction, *Expert Systems with Ap- plications* **124**: 226–251.

Hu, Y.-C. and Ansell, J. (2007). Measuring retail company performance using credit scoring techniques, *European Journal of Operational Research* **183**(3): 1595–1606.

Huang, Y., Capretz, L. F. and Ho, D. (2021). Machine learning for stock prediction based on fundamental analysis, *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 01–10.

Huang, Y.-P. and Yen, M.-F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction, *Applied Soft Computing* **83**: 105663.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments, *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**(2065): 20150202.

Karmakar, M. (2010). Information transmission between small and large stocks in the national stock exchange in india: An empirical study, *The Quarterly Review of Economics and Finance* **50**(1): 110–120.

Khanra, S. and Dhir, S. (2017). Creating value in small-cap firms by mitigating risks of market volatility, *Vision* **21**(4): 350–355.

Liang, D., Tsai, C.-F. and Wu, H.-T. (2015). The effect of feature selection on financial distress prediction, *Knowledge-Based Systems* **73**: 289–297.

Lin, T.-H. (2009). A cross model study of corporate financial distress prediction in taiwan: Multiple discriminant analysis, logit, probit and neural networks models, *Neurocomputing* **72**(16-18): 3507–3516.

Liu, W., Fan, H., Xia, M. and Pang, C. (2022). Predicting and interpreting financial distress using a weighted boosted tree-based tree, *Engineering Applications of Artificial Intelligence* **116**: 105466.

Lokanan, M. and Ramzan, S. (n.d.). The application of machine learning and artificial neural networks algorithms to predict financial distress, *Available at SSRN 4634855* .

Manjula, K. and Karthikeyan, P. (2019). Gold price prediction using ensemble based ma- chine learning techniques, *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, pp. 1360–1364.

Mishraz, N., Ashok, S. and Tandon, D. (2021). Predicting financial distress in the indian banking sector: A comparative study between the logistic regression, lda and ann models, *Global Business Review* p. 09721509211026785.

Narang, A. (2014). *Mitigating high 'equity capital'risk exposure to 'small cap'sector in India: analysing 'key factors of success' for 'Institutional Investors' whilst Investing in small cap sector in India*, PhD thesis, Citeseer.

Nguyen, M., Nguyen, B. and Lieu, M. L. (2023). Corporate financial distress prediction in a transition economy, *Available at SSRN 4552384* .

Ozbayoglu, A. M., Gudelek, M. U. and Sezer, O. B. (2020). Deep learning for financial applications: A survey, *Applied soft computing* **93**: 106384.

Qian, H., Wang, B., Yuan, M., Gao, S. and Song, Y. (2022). Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree, *Expert Systems with Applications* **190**: 116202.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning, arXiv preprint arXiv:1811.12808.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, *Ore Geology Reviews* **71**: 804–818.

Sahu, S. K., Mokhade, A. and Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: recent progress and challenges, *Applied Sciences* **13**(3): 1956.

Sezer, O. B., Gudelek, M. U. and Ozbayoglu, A. M. (2020). Financial time series fore- casting with deep learning: A systematic literature review: 2005–2019, *Applied soft computing* **90**: 106181.

Shen, G. et al. (2014). The prediction model of financial crisis based on the combination of principle component analysis and support vector machine, *Open Journal of Social Sciences* **2**(09): 204.

Sreedharan, M., Khedr, A. M. and El Bannany, M. (2020). A multi-layer perceptron approach to financial distress prediction with genetic algorithm, *Automatic Control and Computer Sciences* **54**: 475–482.

Sun, J. and Li, H. (2012). Financial distress prediction using support vector machines: Ensemble vs. individual, *Applied Soft Computing* **12**(8): 2254–2265.

Sun, J., Li, H., Huang, Q.-H. and He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches, *Knowledge-Based Systems* **57**: 41–56.

Verdonck, T., Baesens, B., O´skarsd´ottir, M. and vanden Broucke, S. (2021). Special issue on feature engineering editorial, *Machine learning* pp. 1–12.
Yu, H., Chen, R. and Zhang, G. (2014). A svm stock selection model within pca, *Procedia computer science* **31**: 406–412.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0), *SAS Institute Inc, Rockville, MD* **49**(1-11): 1