

VAANI PRUTHI

VaaniPruthiPlagVersion5.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:91193623

Submission Date

Apr 14, 2025, 4:35 PM GMT+5:30

Download Date

Dec 4, 2025, 9:42 PM GMT+5:30

File Name

VaaniPruthiPlagVersion5.pdf

File Size

1.1 MB

38 Pages





6,665 Words

38,685 Characters




8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **38 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 6%  Internet sources
- 4%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 38 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 6% Internet sources
- 4% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.coursehero.com	<1%
2	Internet	docshare.tips	<1%
3	Submitted works	Nottingham Trent University on 2018-09-08	<1%
4	Submitted works	King's College on 2023-03-28	<1%
5	Internet	fastercapital.com	<1%
6	Internet	vtechworks.lib.vt.edu	<1%
7	Internet	www.medrxiv.org	<1%
8	Internet	roderic.uv.es	<1%
9	Publication	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng...	<1%
10	Internet	pypi.org	<1%

11	Internet	assets-eu.researchsquare.com	<1%
12	Submitted works	Aston University on 2023-09-28	<1%
13	Internet	navi.com	<1%
14	Submitted works	Middle East Technical University on 2016-11-17	<1%
15	Publication	Keerthy Reghunandanan, V.S. Lakshmi, Rose Raj, Kasi Viswanath, Christeen Davis...	<1%
16	Submitted works	Leeds Beckett University on 2022-09-11	<1%
17	Internet	link.springer.com	<1%
18	Internet	mediatum.ub.tum.de	<1%
19	Internet	pdfs.semanticscholar.org	<1%
20	Internet	insis.vse.cz	<1%
21	Internet	pdfcoffee.com	<1%
22	Internet	repositorio-aberto.up.pt	<1%
23	Submitted works	University of Glasgow on 2023-06-26	<1%
24	Internet	irjet.net	<1%

25	Internet	www.ijjsae.org	<1%
26	Publication	Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Tutorial and critical ana...	<1%
27	Internet	www.mdpi.com	<1%
28	Internet	www.researchgate.net	<1%
29	Submitted works	Liverpool John Moores University on 2023-02-26	<1%
30	Submitted works	University of Edinburgh on 2024-02-29	<1%
31	Publication	van der Colff, Francois. "An Artificial Intelligence Model to Predict Financial Distre...	<1%
32	Publication	Ashish Mishra, Nguyen Thi Dieu Linh, Manish Bhardwaj, Carla M. A. Pinto. "Multi-...	<1%
33	Publication	Jagafa, Kanadi. "A Framework for Construction Business Recovery in Small and M...	<1%

EXECUTIVE SUMMARY

The study titled "*Machine Learning for Indian Corporate Default Prediction*" aims to explore and evaluate the effectiveness of various widely adopted techniques for predicting corporate defaults in the Indian context. Among the models examined are the traditional Altman Z-Score and its variant tailored for emerging markets. In addition, the study advocates for the use of ML classification algorithms—specifically Random Forests, logistic regression, K-Nearest Neighbors (KNN)—to forecast default events.

By analyzing a huge range of variables, including financial ratios, the study's core objective is to evaluate the probability of corporate defaults. The findings suggest that machine learning approaches offer notable advantages over conventional models. In particular, the logistic regression model demonstrates superior predictive precision compared to traditional methods. In addition, the study finds that incorporating both financial ratios and market indicators greatly improves the ability to predict logistic regression.

Contributing to the increasing pool of information in corporate finance, study provides fresh look into the application of ML techniques for predicting default in the Indian corporate landscape. Its conclusions hold practical relevance for policymakers, investors and financial institutions, while also opening new avenues for research and offering a more adaptive framework for forecasting corporate default events.

CHAPTER 1: INTRODUCTION

1.1 Background

Corporate entities universally require capital, which is typically sourced through equity issuance, debt financing, or retained earnings from past profits. However, in today's landscape—characterized by high-growth companies and startups often operating with negative cash flows—debt has taken on a more prominent role. With the use of debt comes the inherent risk of default. While the likelihood of default may appear minimal when viewed broadly, it remains a critical factor worthy of in-depth analysis, study, and potentially, prediction.

1.1.1 Defining Default

A credit event, or what is commonly referred to as a 'default,' occurs when a corporate entity fails to meet its debt obligations. This can happen when even a single instalment or tranche becomes overdue, leading the organization to be classified as a defaulter.

While India has experienced several major corporate defaults in the past, such occurrences

are not unique to its economy. Defaults are, in fact, a common feature across many free market economies and are often considered a normal byproduct of a modern capitalist system.

Corporate defaults can stem from a huge range of factors, such as:

1. Internal Factors:

- Poor financial management
- Overleveraging
- Mismanagement
- Operational inefficiencies
- Weak corporate governance

2. External Economic Conditions:

- Adverse economic cycles or downturns
- High-interest rate environments
- Overregulation
- Shallow or underdeveloped markets

3. Unforeseen Events:

- Natural disasters and pandemics
- Black-swan events like COVID-19, which are difficult to anticipate

4. Shifts in Competitive Dynamics:

- Rapid changes in consumer behavior and market demand
- Technological disruption that quickly erodes an organization's competitive edge
- In today's volatile and fast-paced global economy, these risk factors are more interconnect than ever, making the prediction and management of default events a crucial area of focus for financial analysts, investors, and policymakers alike.

1.1.2 Default in the Indian Context

The RBI (Reserve Bank of India) defines default as fail of a corporate entity to honour its debt commitments within 30 days of the due date on instruments. However, credit rating agencies in India often adopt a different perspective when assessing and classifying defaults.

Table 1: Definitions of default as used by Indian credit rating agencies

Default definition	CRISIL	ICRA	CARE	Fitch (India)
Recognition of default	A single day's worth of late payment or missing rupees ¹	An issuer's missing or delayed payment in violation of the issue's terms	On its rated instrument, any missing payments	Failure of an Obligor to Pay Principal and/or Interest in a Timely Manner as Required by the Terms of Any Financial Obligation

Source: CRISIL, KD [1]

¹ NOTE: Data for the defaulted companies was taken from CRISIL Investors Information Center, Hence the definition for default which reigns supreme is CRISIL's

1.1.3 Major Credit Events: Indian Corporates

1.1.3.1 Essar Steel (\$5.9 Billion Default)

Essar Steel was burdened with an extremely high level of debt, with a debt-to-equity ratio of 5.9 significantly above the industry average. The company experienced major delays in the implementation of its key capital expenditure project—the Gujarat steel plant—resulting in rising debt levels without a corresponding increase in revenue.[4]

1.1.3.2 Reliance Communications (\$6.1 Billion Default)

Reliance Communications (RCOM), a telecommunications firm led by Anil Ambani—brother of Mukesh Ambani of Reliance Industries—defaulted on debts amounting to ₹45,000 crore in 2018.

The company became entangled in a series of legal disputes, most notably with telecom equipment provider Ericsson, over unpaid dues. This culminated in a court-ordered settlement of \$76 million.

RCOM was also heavily impacted by sector-specific challenges, including declining Average Revenue Per User (ARPU), falling profitability, and intensified competition from well-capitalized market entrants—creating a perfect storm that ultimately led to its financial downfall.

1.1.3.3 Kingfisher Airlines (\$1 Billion Default)

Kingfisher Airlines collapsed in 2012 after declaring a moratorium on its debt repayments, leaving creditors with an exposure of nearly \$1 billion. Operating in a high-risk industry notorious for financial volatility, Kingfisher pursued aggressive expansion, which led to strained working capital and financing decisions. Notably, the **Altman Z-Score** for Kingfisher was calculated at **-8.63**—indicating **bankruptcy well before the actual default**—highlighting how early warning models could have alerted lenders to impending risk.[2]

The airline was also plagued by serious accounting irregularities, which eroded investor confidence and caused its equity value to plummet.[3]

1.2 Problem Statement

India, as a developing economy, is witnessing rapid evolution in its debt and capital markets. The Indian bond market has seen remarkable growth, expanding from \$432 billion to \$1.8 trillion over the past decade—a nearly fourfold increase (non-inflation adjusted).

With this surge, investor exposure to corporate debt has also grown significantly. This includes not only retail investors but also institutional players such as banks, asset management companies (AMCs), mutual funds, and sovereign wealth funds. This heightened exposure underscores the critical need for effective mechanisms to predict and mitigate corporate defaults.

Moreover, in the Indian context, the stakes are even higher due to the significant presence of public sector government owned banks, and, by extension, taxpayers. As such, preventing corporate defaults is a topic of public interest and fiscal responsibility. In this report, we aim to examine both traditional and modern approaches used to assess financial health and predict corporate default. Our analysis will cover established models like Altman Z-Score, Modified Z-Score for Emerging Markets, and Merton Model. Additionally, we will evaluate the performance of supervised classification algos—such as K-Means, K-Nearest Neighbours (KNN), as well as more advanced models like Logistic Regression (LR), Random Forests and Neural Networks (NN)—using a dataset comprising 79 Indian corporate entities.

Corporate default prediction is inherently complex and subjective, owing to the multitude of influencing variables and the presence of significant multicollinearity among them. We will use metrics like the confusion matrix and AUC to measure the accuracy and effectiveness of our models.

1.3 Need for the study

Default prediction is gaining critical importance in the India financial landscape. RBI has become increasingly proactive in urging banks to adopt more robust credit risk management frameworks and predictive models.

As part of its regulatory oversight, the RBI mandates that banks maintain specific liquidity and solvency ratios, including the Capital Adequacy Ratio (CAR) and proper handling of Non-Performing Assets (NPAs). NPAs refer to loans that have remained overdue for more than 90 days and are a key indicator of a bank's asset quality.

According to Basel III norms, which India adheres to, the RBI mandates banks to have at least CAR of 11.5%, with at least 9% allocated as Tier 1 capital. These measures are designed to ensure financial stability, and predictive models for defaults play a crucial role in helping banks meet these regulatory benchmarks while minimizing risk exposure.[5]

These regulatory requirements significantly influence the profitability of banks, primarily by limiting their lending capacity. However, with accurate and timely corporate default prediction, there is potential for these norms to be calibrated more effectively. Enhanced predictive accuracy could lead to more confident lending decisions, thereby improving both the lending ability and overall profitability of banks.[6]

Accurate default prediction models not only strengthen banks' financial performance but also contribute to broader market efficiency. When banks can better assess credit risk, they can afford to take on higher levels of calculated risk, thereby optimizing the use of their Loss Given Default (LGD) models and capital reserves.

Moreover, such predictive tools are valuable for a wide range of institutional investors—including banks, asset management companies (AMCs), mutual funds, and sovereign wealth funds—who are increasingly exposed to corporate debt. From a macroeconomic standpoint, improved default prediction is also in the best interest of the taxpayer and the sovereign exchequer.

1.4 Scope of the Study

This study seeks to analyze and compare a range of corporate default prediction models within the Indian financial ecosystem. Additionally, it aims to implement advanced classification algos—like Logistic Regression (LR), Random Forests (RFs), K-Nearest Neighbours (KNNs) and Artificial Neural Networks (ANNs)—to estimate likelihood of corporate default.

1.5 Limitations of the Study

Corporate default is a multifaceted phenomenon, making it challenging to capture all relevant factors within a single predictive model. This study is based on a limited dataset comprising 100 publicly listed entities on India's major stock exchanges—BSE and NSE. Due to inconsistencies in the raw data, significant cleaning was required, which resulted in the loss of some data points. Moreover, the scope of this report is constrained, and not all potential prediction models and methodologies could be explored in depth.

CHAPTER 2: LITERATURE REVIEW

In this part of the report, we'll dive into key terms that will guide our analysis throughout the report. Our discussion will span a wide range of concepts related to predictive modeling, including theoretical frameworks and evaluation metrics like confusion matrices, AUC curves, & various accuracy measures. These elements are vital to understand and interpreting performance of the model discussed in the report.

2.1 Credit Event

According to CRISIL, a credit event refers to any occurrence that renders a borrower or issuer incapable of meeting their debt obligations. In contrast, RBI defines default as a situation where a borrower is 30 days overdue on its payments before being officially classified as a defaulter. However, independent credit rating agencies typically do not offer such a grace period and may classify a default when payment is not made.

2.2 Regulatory Framework Governing Corporate Defaults in India

The RBI (Reserve Bank of India) and the SEBI (Securities and Exchange Board of India) are the primary regulators supervising bankruptcy and insolvency processes in India, while the NCLT (National Company Law Tribunal) is responsible for handling legal proceedings that follow.

The IBC (Insolvency and Bankruptcy Code), launched in 2016, made to simplify & speed up the resolution of credit events. Main features of the IBC are:

- 1. Time-Bound Resolution:** It is mandated to be completed within 330 days, ensuring swift action.
- 2. Creditor-Driven Approach:** The IBC empowers creditors by incorporating their input into the resolution process, making it more inclusive and participatory.
- 3. Legal Hierarchy:** The code establishes a clear judicial structure, with the NCLAT and, Supreme Court of India (SCI) serving as appellate authorities.

2.3 Financial Ratios as Indicators of Corporate Failure

According to financial theory, signs of corporate insolvency often include a decline in asset values or a lack of liquidity—specifically, the inability to generate or access sufficient cash to fund operations or investments. Consequently, it is expected that financial ratios reflecting these aspects would vary significantly between defaulting and solvent companies, particularly in terms of cash flow dynamics and changes in the market value of assets.[9]

But, it's crucial to remember that financial ratios exhibit high levels of multicollinearity, which can complicate model interpretation and reduce predictive reliability. Despite this, models incorporating cash flow-related variables—such as those based on the LOGIT framework—have demonstrated predictive accuracy rates of up to 83% as early as one year prior to a credit event.[10]

Table 2: Overview of Potential Downfall in Highly Leveraged Firms

Explanation of Performance Decline	Explanation	Explanation Predicts Loss of Sales Revenue?	Explanation Predicts Decline in Firm Value?	Other Predictions
Customer driven	Customers and stakeholders abandon the firm	Yes	Yes	Performance decline worse for firms with specialized products
Competitor driven	Competitors reduce prices to gain market share	Yes	Yes	Performance decline worse in concentrated industries
Manager driven	Managers efficiently downsize by cutting poorly performing assets	Yes	No	Performance decline may be related to firm size

Source: Opler et. Al

2.4 Idiosyncratic Factors

In industries that are heavily consolidated and where firms invest significantly in R&D, the link between high leverage and declining performance is often more pronounced. These findings suggest that sales downturns in such firms are driven, at least in part, by external factors—such as customer behavior and competitive pressures—rather than solely by internal cost-cutting measures like strategic workforce reductions by management.

2.4.1 Impact of Financial Distress on Employment, Asset Sales & Investment

A variety of recent studies have found the tendency of financially distressed firms to cut back on employment and investment, as well as to sell off assets. These actions are often associated with what is termed management-driven sales losses. While these studies generally establish a connection between such strategic decisions and the financial condition of the firms, their research designs often fall short of fully distinguishing whether these changes are purely responses to financial distress or also influenced by underlying operational or structural weaknesses within the firms.[10]

1. Ineffective Financial and Operational Management

- **Excessive Leverage:** Overdependence on debt can strain cash flows and increase vulnerability to financial shocks.
- **Weak Management Practices:** Poor leadership and strategic missteps can lead to unsustainable business decisions.
- **Operational Inefficiencies:** Ineffective internal processes can reduce productivity and increase costs.
- **Lack of Robust Corporate Governance:** Weak oversight mechanisms often result in poor financial discipline and higher risk exposure.

2. Economic Challenges

- **High Interest Rate Environment:** Rising borrowing costs can burden highly leveraged companies.
- **Stringent Regulatory Frameworks:** Overregulation can stifle innovation and operational flexibility.
- **Shallow Market Depth:** Limited liquidity and investor diversity can exacerbate financial instability during downturns.

3. Pandemics and Natural Disasters

- **Unpredictable Black Swan Events:** Crises like COVID-19 are difficult to forecast and prepare for, yet they have far-reaching economic and operational consequences.

4. Shifting Competitive Landscape

- **Rapid Consumer Behavior Changes:** Evolving preferences and demands can render products or services obsolete quickly.
- **Technological Disruptions:** Innovations can swiftly erode existing competitive

advantages, requiring constant adaptation.

2.5 Default Prediction Models

The absence of a unified theory for predicting corporate failure has led to a surge in diverse empirical approaches across different markets. As a result, financial institutions, particularly banks, increasingly rely on statistical models to find the chance of default. The models serve as early warning systems, enabling banks proactively find the credit risk of their corporate clients take timely risk mitigation measures.

2.5.1 Altman Z Score Model

The Altman Z Score model, stands as the most influential pioneering frameworks for default prediction. Notably, it has demonstrated the ability to forecast corporate defaults as early as five years prior to the actual credit event, making it a valuable tool for early risk assessment.[12]

The model formulates a Z- score as shown below.

$$Z\text{-Score} = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E$$

Where:

A = Working capital/total assets

B = Retained earnings/total assets

C = EBIT/total assets

D = Market value of equity/BV of total liabilities

E = Sales/total assets

These financial ratios can be sourced through publicly available APIs such as Yahoo Finance and similar platforms.

The original threshold values established by Altman in 1968 were as follows:

Table 3: Cut-off Values for Altman Z -Score approach and Interpretations²

Bands	Zones	Interpretation
Z-Score > 2.99	Safe Zone	Company is in a safe position with low probability of default
1.81 < Z-Score < 2.99	Grey Zone	Company may have certain financial difficulties but is not yet at risk of default
Z-Score < 1.81	Distress Zone	Company is in a dangerous position with high probability of default

Source: Self-Compiled

2.5.1.1 API Implementation

Figure 1: Yahoo Finance API Implementation

```
# Download the financial data from Yahoo Finance
data = yf.download(ticker, start=start_date, end=end_date)

# Calculate the parameters required for the Altman Z-Score
working_capital = data["Total Current Assets"] - data["Total Current Liabilities"]
total_assets = data["Total Assets"]
retained_earnings = data["Retained Earnings"]
ebit = data["EBIT"]
market_value_of_equity = data["Market Cap"]
book_value_of_total_liabilities = data["Total Liab"]
sales = data["Total Revenue"]
```

Source: Self-Compiled

The Altman Z score employs multivariate linear discriminant analysis model to forecast corporate defaults.

² NOTE: These cut-off values are the original values as defined by Altman in 1968, they have since been optimized to fit into their applicable markets. The changes have been of a similar nature for coefficients of original Z score equation.

2.5.1.2 Altman Z-Score Performance in Indian Context

A research done by Satyendra et al. applied Altman Z score model to 183 firms to estimate default probabilities. The model successfully predicted 81% of defaults with a lead time of up to 3 years.[13] Similarly, a 2017 study by Duggal and Prakash evaluated the model's effectiveness within the Indian pharmaceutical sector, where it got accuracy rate of 85% for a five-yr period.[14]

These findings indicate that, despite its age, the Altman Z score model remains valuable method in Indian context. But, it's vital to acknowledge the newer models have been specifically optimized for emerging market conditions and may offer improved performance within their targeted domains. These contemporary models are explored further.

2.6 Z-Score Optimized for Emerging Markets

Emerging Market Z-Score (EMZ) model was introduced in 1998 as an evolution of actual Z Score model. It incorporates key adjustments for better reflect the distinct characteristics and economic conditions of companies operating in emerging markets. These include heightened volatility, increased dependence on external funding, and a less standardized regulatory and creditor protection environment.

Emerging markets are typically economies in transition—nations experiencing growth and integration into the global financial system. While they share some traits with developed markets, such as improving infrastructure and expanding capital markets, they often fall short in areas like regulatory maturity and investor protections.

Developed markets are defined by sustained economic growth, rising per capita income, deep and liquid capital markets, greater openness to foreign investment, and comprehensive regulatory frameworks. The EMZ model, tailored for the more complex and less stable dynamics of emerging economies, consists of:

1. Working Capital / Total Assets

A liquidity indicator measuring the firm's capacity to fulfill near-term debts

2. Retained Earnings / Total Assets

Shows level of internal financing and firm's financial history.

3. EBIT / Total Assets

Gauges operational profitability, showing how efficiently assets generate earnings.

4. **Market Value of Equity / Total Liabilities**

Captures market sentiment regarding the firm's financial health and growth outlook.

$$\text{EM Score} = 6.56(X_1) + 3.26(X_2) + 6.72(X_3) + 1.05(X_4) + 3.25$$

where $X_1 = \text{WC/Total Assets}$

$X_2 = \text{retained earnings/total assets}$;

$X_3 = \text{operating income/total assets}$;

$X_4 = \text{book value of equity/total liabilities}$.

Several studies have found that the EMZ model offers improved predictive accuracy relative to actual **Altman Z Score model** when used for organizations in emerging markets.

Table 4: Z-Score and appropriate credit rating

	Z"-Score			Rating	Z"-Score			Rating	
Safe zone	{	>	8.15	AAA	5.65	–	5.85	BBB-	Grey zone
			7.60 – 8.15	AA+	5.25	–	5.65	BB+	
			7.30 – 7.60	AA	4.95	–	5.25	BB	
			7.00 – 7.30	AA-	4.75	–	4.95	BB-	
			6.85 – 7.00	A+	4.50	–	4.75	B+	
			6.65 – 6.85	A	4.15	–	4.50	B	Distress zone
			6.40 – 6.65	A-	3.75	–	4.15	B-	
			6.25 – 6.40	BBB+	3.20	–	3.75	CCC+	
			5.85 – 6.25	BBB	2.50	–	3.20	CCC	
					1.75	–	2.50	CCC-	
					< 1.75			D	

Source: Altman et.al 1998, Emerging Markets Review

Note: The API implementation for the EMZ model closely mirrors actual Altman Z Score model, with adjustments made for specific parameters. Classifications for Safe Zone, Distress Zone and Grey Zone also align with those defined in the original Z-Score framework.

2.6.1 EMZ Performance in the Indian Context

In a research done by Karagiannis Apte, 137 Indian companies was analyzed over the period from 2008 to 2017. The findings revealed that the EMZ model outperformed the actual Altman Z Score in predicting financial distress. EMZ model got accuracy of 89%, compared to 80% for the Altman Z-Score. [16]

2.7 Merton's Model

The Merton Model operates on a set of simplifying assumptions regarding company's financial framework. It evaluates probability of credit event based on market value of the company's assets relative to debt obligations. A default is deemed to occur when a firm's asset value declines beneath a critical threshold, often known as the default point. Model estimates probability of default using following equation:

$$PD = N(d_2)$$

Where PD = probability of default,

N = standard normal cumulative distribution function

$$d_2 \text{ is } d_2 = [\ln(V/E) + (r + \sigma^2/2)T] / (\sigma\sqrt{T})$$

Where V = value of firm's assets

E = value of its liabilities

r = risk-free interest rate

σ = volatility of the firm's assets

T = time to maturity of firm's liabilities.

2.7.1 Performance of the Merton Model in the Indian Context

In their 2019 study, Sehgal and Rai found that Merton's Model demonstrated strong predictive power in identifying corporate financial distress within the Indian context. The model outperformed traditional statistical methods, delivering an overall prediction accuracy of 86.8%. Additionally, the model got an area under ROC (Receiver Operating Characteristic) curve of 0.931, indicating high level reliability in forecasting defaults.

2.8 Machine Learning Techniques: Classification Algorithm

2.8.1 Logistic Regression

Logistic regression is a statistical technique used to model probability of categorical outcome based on 1 or more independent variables. It's commonly applied to estimate binary outcomes, like if a customer will make a purchase or whether a patient will develop a particular illness. In cases involving more than 2 possible outcomes, multinomial logistic regression can be used. Logistic regression serves as an effective tool for classification tasks, where motive is to give new observations to 1 of several predefined categories.

$$p(y = 1|x) = 1 / (1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)))$$

where:

- $p(y = 1|x)$ denotes probability that dependent variable y equals 1, given set of predictor variables x_1, x_2, \dots, x_p .
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are coefficients in logistic regression model that quantify relationship between each predictor variable and probability of outcome.
- $\exp()$ refers to exponential function used to change linear combination of inputs into probability between 0 and 1.
- x_1, x_2, \dots, x_p are predictor or independent variables, which may be either continuous or categorical and are assumed to influence the probability of the outcome.

2.8.1.1 Logistic Regression for Default Prediction

The study titled "*Predicting Corporate Default in India: An Analysis Using Logistic Regression*" by R. Venkatesh and K. Murali Krishna, reported an accuracy rate of **78.46%** for the logistic regression model in predicting corporate defaults in India. The research analyzed a dataset of **134 non-financial companies** listed on **National Stock Exchange (NSE)** between **2007 and 2012**. The model utilized **financial ratios, market indicators, & macroeconomic variables** as predictors. Key variables included stock returns and volatility (market factors), **debt-to-equity ratio, interest coverage ratio, and current ratio** (financial indicators), along with GDP growth (a macroeconomic variable).[16]

2.8.2 K Nearest Neighbours and K Means Clustering

K-Nearest Neighbours (KNN) is a **supervised machine learning algorithm** used for both **classification and regression tasks**, though it's applied to classification and prediction issues in practice. KNN is best described by the following core characteristics:

Proximity-Based Classification: KNN classifies a data point by assigning it to class most common amongst its nearest neighbors, enabling effective predictive modelling through similarity measures.

Non-Parametric Nature: It doesn't rely on assumptions about underlying data distribution, which allows for adaptability across diverse datasets.

KNN is also considered **lazy learning** algorithm, meaning it doesn't involve dedicated training phase. Instead, it uses the entire dataset during prediction. Additionally, it is a **centroid-based** technique, where each cluster is associated with a centroid. The algorithm aims to **minimize the total distance** between data points and their respective clusters to enhance classification accuracy.

2.8.3 Ratio Rationale

1. **Market Capitalization:** In a 2017 study by Altman et al., market capitalization emerged as a key indicator of default risk in U.S.A corporate bond market. The research suggested that larger firms typically exhibit lower chances of default.
2. **Consolidated EPS:** Elyasiani et al. (2017) identified earnings per share (EPS) as a critical financial ratio in forecasting corporate default. Their analysis of U.S. industrial firms showed that lower EPS figures are linked with a greater likelihood of default.
3. **P/E Ratio:** Research by Scholtens et al. (2017) highlighted the price-to-earnings ratio as an important default predictor in emerging markets. Their study across 13 emerging economies concluded that firms with lower P/E ratios generally face higher default risks.
4. **P/B Ratio:** A study by Norden and Weber (2004) found the price-to-book ratio to be significant determinant of risk of default in European corporate bond markets. Companies with higher P/B ratios were shown to have a reduced probability of default.
5. **Turnover and Total Income from Continuing Operations:** Bhandari and Biswal (2019) discovered that both turnover and total income from ongoing operations are strong indicators of default risk in the Indian banking sector. Larger institutions with higher values in these metrics tend to exhibit lower default probabilities.
6. **EV/PBIDTA:** Singh and Sharma (2016) reported that the enterprise value to PBIDTA ratio is a useful measure of corporate default risk within India's corporate bond market. Firms with higher EV/PBIDTA ratios were found to be less likely to default.

7. **Enterprise Value (EV):** In a long-term analysis from 1985 to 2015, Keasey et al. (2018) found that enterprise value is a meaningful predictor of risk of default in UK corporate bond market. Companies with higher EVs were observed to be at a lower risk of default.



CHAPTER 3: RESEARCH METHODOLOGY



3.1 Introduction



Motive of project is to assess performance of various machine learning algorithms using customized set of financial ratios. We aim to compare these algorithms based on key accuracy metrics, including Accuracy, F1 Score, Precision, Recall, and Area Under ROC Curve (AUC).

Our goal is to demonstrate capability of these models to complement existing, well-established default prediction frameworks, thereby enhancing overall predictive accuracy.

For this analysis, we have compiled a dataset comprising 79 companies — 39 of which have previously defaulted on their debt obligations, while the remaining 40 have maintained a clean credit history.

3.2 Research Questions

1. Which machine learning algorithms demonstrate the highest effectiveness in predicting corporate defaults?
2. What is the level of accuracy achieved by these algorithms in forecasting default events?

3.3 Methodology

3.3.1 Data Collection

Dataset for report was compiled several publicly available financial databases, including CMIE Prowess, CRISIL, and NCLAT. CMIE Prowess provided access to the financial ratios and performance metrics of the selected firms, while CRISIL was instrumental in identifying companies that had defaulted on their debt obligations, along with the specific dates of those defaults—an essential detail for enabling accurate classification by the models. The dataset was designed to follow a balanced structure, with 50% of the companies having defaulted and the other 50% being financially healthy. For the defaulting firms, financial ratios were sourced from CMIE Prowess, taken retrospectively one year prior to the default event, ensuring an effective prediction window.

3.3.2 Data Pre-processing

Exploratory Data Analysis (EDA) employed gain an initial understanding of the dataset and to uncover common data quality issues, including missing values and outliers. To address gaps in the data, missing values were supplemented using alternative financial sources like Money control and Screener. Additionally, approximately 21 fields with persistent missing data were removed from the dataset to maintain overall data integrity.

3.3.3 Model Selection

Model selection played a vital role in this project and was carried out following a comprehensive literature review and in-depth research on classification techniques. Our

objective was to evaluate a range of models—starting from simpler ones like K-Nearest Neighbours and K-Means, to intermediate approaches like Logistic Regression and Gradient Boosting, and finally showcasing the potential of newer techniques such as Decision Trees.

The implementation was done in Python, utilizing a variety of libraries including Scikit-learn (SKlearn), Matplotlib, NumPy, Pandas, and Seaborn. SKlearn, in particular, served as the core library, providing access to most of the machine learning algorithms applied in this analysis.[27]

3.3.4 Model Evaluation

The models were assessed using a randomized data split, meaning the division between training and testing sets was done independently of the data's original order in the spreadsheet. This approach was adopted to ensure model's generalizability & improve its capability to handle new, unseen data. Additionally, cross-validation was performed alongside the standard training-testing split to even more validate model's performance and reliability.

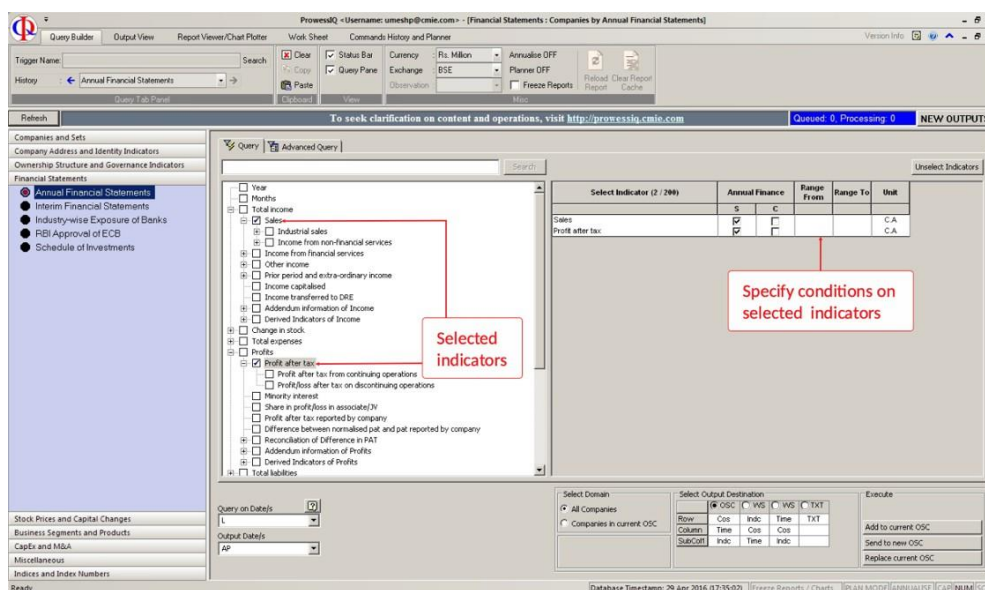
3.3.5 Model Interpretation

The performance of the models was assessed by examining main features and metrics, including Accuracy, Precision, F1 Score, and Recall. Additionally, the Area Under the ROC Curve (AUC-ROC) was utilized to draw meaningful conclusions about the effectiveness of each model.

CHAPTER 4: DATA COLLECTION

Data for the research was primarily collected from the CMIE ProwessIQ and CRISIL databases, focusing on sample of 79 companies listed on Indian stock market. Key financial ratios—including market capitalization, consolidated EPS, P/E ratio, P/B ratio, turnover, total income from continuing operations, EV/PBIDTA, and EV—were sourced from ProwessIQ for analysis.

Figure 2 : CMIE ProwessIQ UI Screen Capture



Source: Prowess IQ Brochure Document

Default dates were gathered from multiple reliable online financial sources, including Economic Times, Money control, and the CRISIL corporate bond rating database. The specific quarters and years of default have been detailed in the annexure.

The remaining half of the dataset, comprising non-defaulting companies, was randomly selected from listings on the NSE and BSE.

4.1 Collection Rationale

The reasoning behind the selection of financial ratios has been discussed in detail in the literature review section of this report.

4.2 Sample Data Structure

Table 5: Sample Data Formatting

Company Name	Market Capitalisation	Consolidated EPS	P/E	P/B	Turnover	Total income from continuing operations	EV/PBIDTA	EV	Default
Alok Industries Ltd.	37.39	-1.09	-3.04	0.25	3308.05	-1717	0	11312.05	1
Altico Capital India Ltd.	3.27	-6.22	-0.31	0.02	0	180.3	0	2577.71	1
Amtek Auto Ltd.	0.13	-14.6	-0.02	0.01	119.39	-234.36	0.01	1143.04	1
Athena Energy Ventures Pvt. Ltd.	0.03	-0.25	-0.12	0.11	0	-1.97	0	-0.01	1
Ballarpur Industries Ltd.	0.19	-30.17	-0.01	0.01	1367.67	-1023.14	0	2045.16	1
Bharati Defence and Infrastructure	0.09	-32.56	-0.02	0.01	0.07	-31.71	0.01	1014.47	1
Bhartiya Rail Bijlee Company Ltd.	0.44	-6.22	-0.08	0.01	0	-20.56	0	6935.55	1
Bhushan Power and Steel Ltd.	3.61	-25.51	-0.06	0.02	1541.71	-1300.18	0	8628.18	1
Bhushan Steel Ltd.	0.67	-46.07	-0.01	0.01	1574.47	-2243.62	0	3348.9	1

Source: Self Compiled

CHAPTER 5: DATA ANALYSIS

5.1 Tools Used

Excel, Jupyter-Notebooks, GitHub, Python, scikit-learn, sklearn

5.1 Exploratory Data Analysis

The dataset's descriptive statistics are presented below.

Figure 3: Descriptive Statistics

	Market Capitalisation	Consolidated EPS	P/E	P/B
count	7.900000e+01	79.000000	79.000000	79.000000
mean	1.577534e+06	37.776618	39.698636	9.094490
std	2.550293e+06	72.319505	31.916762	10.175946
min	1.395600e+02	-193.540000	0.060000	0.000000
25%	2.825460e+03	-11.000000	29.095000	3.360000
50%	9.089949e+05	34.320000	39.698636	9.094490
75%	1.711029e+06	55.115000	39.698636	9.094490
max	1.521018e+07	313.250000	275.460000	74.280000

	Turnover	Total income from continuing operations	EV/PBIDTA
count	79.000000	7.900000e+01	79.000000
mean	2200.004167	1.313428e+05	23.555417
std	4572.846259	1.788317e+05	47.438039
min	0.010000	9.974000e+02	-4.760000
25%	3.490000	2.982575e+04	1.665000
50%	784.390000	1.265897e+05	12.890000
75%	2200.004167	1.313428e+05	23.555417
max	30925.810000	1.155900e+06	343.300000

Source: Self Compiled

The dataset comprises 1 dependent variable & 8 independent variables. Independent variables include Market Capitalization, Consolidated EPS, P/E Ratio, P/B Ratio, Turnover, Total Income from Continuing Operations, Enterprise Value (EV), and the EV/PBIDTA ratio (Enterprise Value to Earnings Before Interest, Taxes, Depreciation, and Amortization). The dependent variable, labeled "Default," is binary and indicates whether company is likely to default.

Market Capitalization in the dataset ranges from ₹1.3956 billion to ₹1.4821 billion, while Turnover spans from ₹0.01 billion to ₹2200.004167 billion. The average value of Total Income from Continuing Operations is ₹1313.43 crore.

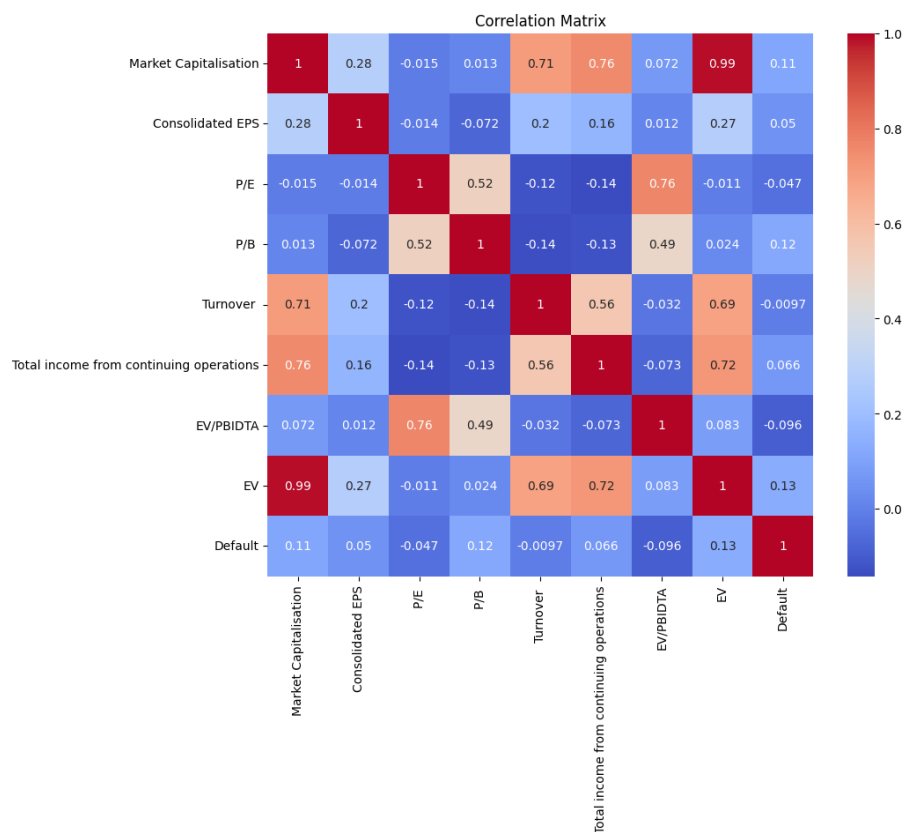
Enterprise Value (EV) shows a particularly wide range, with a mean of ₹6,14,494.63 crore and a relatively high standard deviation, indicating substantial variability across firms.

The EV/PBIDTA variable also exhibits a high degree of variability, with average value of 54.81 and substantial standard deviation of 76.48, meaning significant dispersion across the dataset.

5.2.0 Correlation Matrix

A correlation matrix is N x N table, where N represents the no. of variables being analyzed. It displays the correlation coefficients—in this case, Spearman Rank Correlation—between every possible pair of variables. This matrix is powerful and intuitive method for summarizing huge datasets & identifying patterns or relationships within data. Each variable is listed along both the rows and columns, with the correlation coefficients populating the individual cells of the matrix, often visualized through a correlogram.

Figure 4: Correlation Matrix or Correlogram



Source: Self Compiled

- Firms with larger market capitalizations report greater turnover and total income, as evidenced by the strong positive correlations between market capitalization and total income from continuing operations (0.759), as well as turnover (0.709).
- There is strong positive correlation between P/E and EV/PBIDTA ($r = 0.765$), suggesting companies with greater P/E generally show higher EV/PBIDTA multiples.
- The notable positive relationship between turnover and enterprise value (0.690) implies that companies with higher turnover typically possess greater enterprise value.
- A strong positive correlation between total income from continuing operations and enterprise value (0.722) indicates companies with greater income levels are probably to have greater enterprise values.
- The negative correlation between EV/PBIDTA and default (-0.096) shows companies with

greater EV/PBIDTA ratios might be less likely to default.

5.3 Classification Algorithms

5.3.1 Confusion Matrix

It is a widely used tools visualizing performance of classification algorithms. Represented as an N*N matrix for a classification problem with N classes, it outlines all possible combinations of predicted & actual class labels.

Matrix serves as fundamental basis for calculating key performance metrics such as AUC-ROC, precision, recall, specificity, overall accuracy.

Figure 5: Confusion Matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Source: Towards DS Repository

5.3.1 Measures

Consider a binary classification problem involving 2 classes: positive (P) and negative (N). In evaluating the model's performance, we define the following:

True Positive (TP): The no. of correctly predicted positive instances.

False Positive (FP): The no. of negative instances incorrectly predicted as positive.

False Negative (FN): The no. of positive instances incorrectly predicted as negative.

True Negative (TN): The no. of correctly predicted negative instances.

Based on these definitions, we calculate key performance metrics:

Precision: Indicates proportion of predicted positive cases which are actually positive.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall (Sensitivity): Measures proportion of actual positives which were rightly identified.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1 Score: Harmonic mean of precision and recall, providing a balanced measure which accounts for false positives and false negatives.

$$F1 \text{ Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Accuracy: Represents overall model's predictive accuracy.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

5.3.3 AUC-ROC Curve

The **Area Under the Curve (AUC)** of the **Receiver Operating Characteristic (ROC)** curve is key indicator used to find effectiveness of classification model. It is derived from a plot of **True Positive Rate (TPR)** vs **False Positive Rate (FPR)**. Diagonal line on ROC curve represents model with no discriminatory power—where the TPR equals the FPR—indicating performance equivalent to random guessing, with no meaningful distinction between class 0 and class 1.

While ROC curves typically use TPR and FPR to interpret the model's performance based on confusion matrix, other metrics—like Precision—can also be used in place of FPR in alternative evaluation plots like Precision-Recall curves.

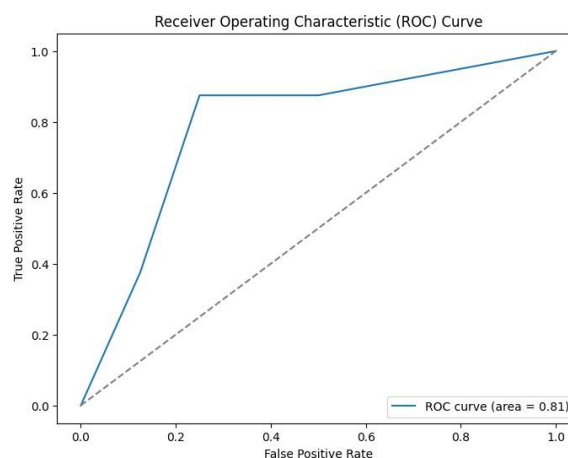
5.3.4 K Nearest Neighbours

KNN (K-Nearest Neighbours) algorithm was implemented to our dataset.

5.3.4.1 Measures

- **AUC ROC Curve**

Graph 1: ROC Curve for KNN

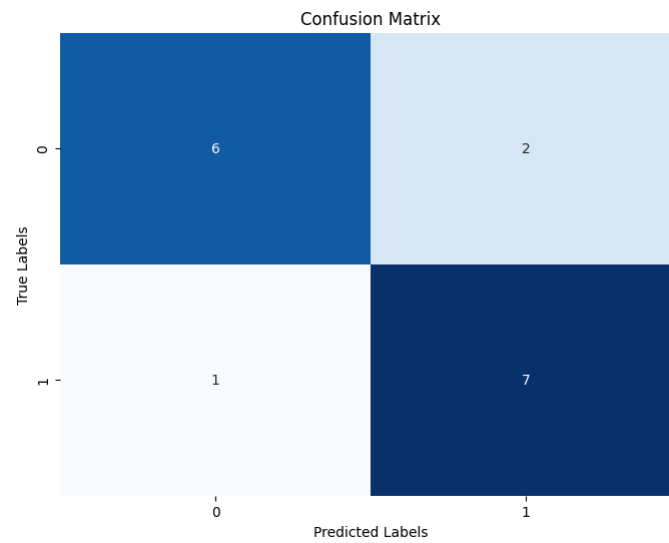


Source: Self Compiled

- **Confusion Matrix**

16

Figure 6: Confusion Matrix for KNN



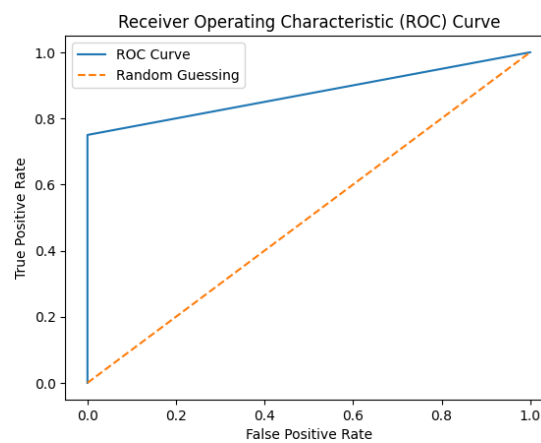
Source: Self Compiled

5.3.5 LOGIT Model

5.3.5.1 Measures

- ROC Curve

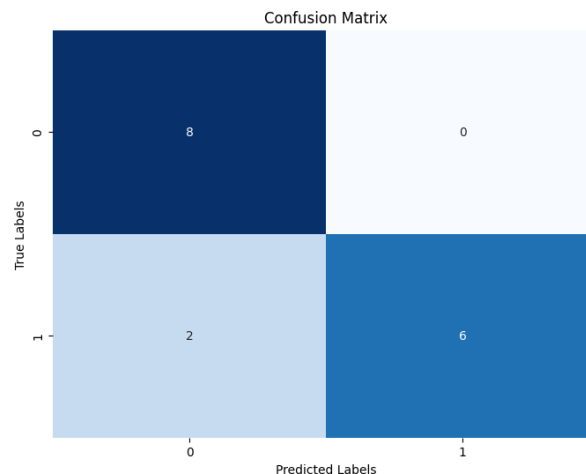
Graph 2 : ROC Curve for LOGIT



Source: Self Compiled

- Confusion Matrix

Figure 7: Confusion Matrix for LOGIT



Source: Self Compiled

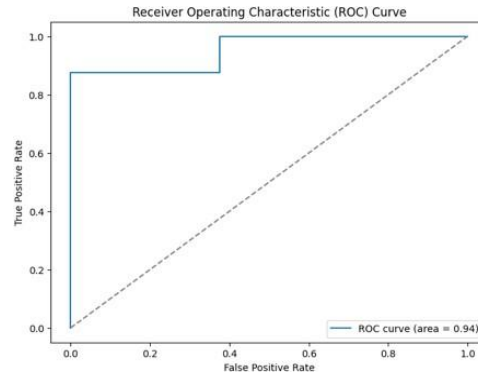
5.3.6 Gradient Boosting Classifier (GBC) Model

This machine learning algorithm belongs to the family of ensemble methods, specifically tailored for classification tasks using a variant of the gradient boosting technique. The Gradient Boost Classifier functions by iteratively forming a series of decision trees, in which every subsequent tree is designed to rectify mistakes made in past. During training process, model adjusts feature weights to minimize the loss function, optimizing its parameters through gradient descent.

5.3.6.1 Measures

- **ROC Curve**

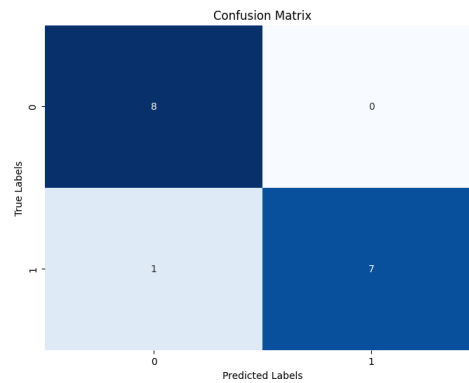
Graph 3 : ROC Curve for GBC



Source: Self Compiled

- **Confusion Matrix**

Figure 8: Confusion Matrix for GBC



Source: Self Compiled

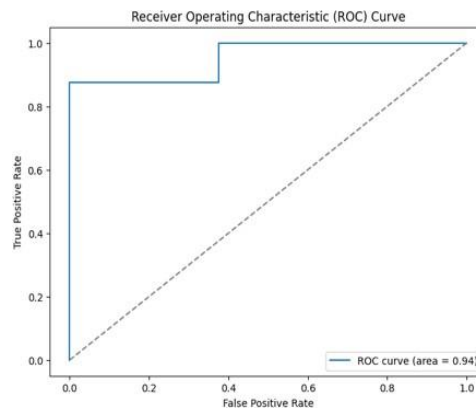
5.3.7 Linear Discriminant Analysis Model

Linear Discriminant Analysis (LDA) is supervised ML method widely made for classification problems which operates by identifying the ideal linear combination of i/p features which separates different classes. LDA performs dimensionality reduction by putting data into lower-dimensional space which maximizes class separation. This is accomplished by finding linear discriminants that increase the ratio of b/w-class variance to within-class variance, thus pinpointing directions in feature space that provide the highest degree of class distinction.

5.3.7.1 Measures

- **ROC Curve**

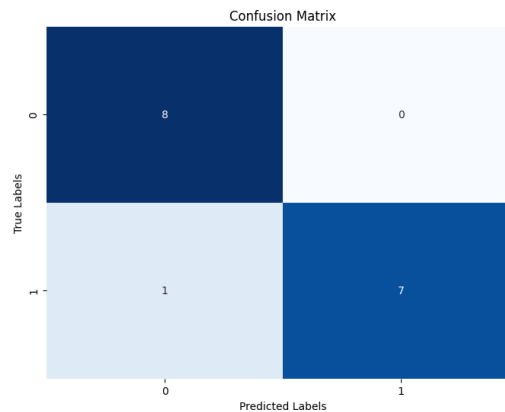
Graph 4 : ROC Curve for LDA



Source: Self Compiled

- **Confusion Matrix**

Figure 9: Confusion Matrix for LDA



Source: Self Compiled

5.3.8 Decision Tree

Decision trees offer highly intuitive & easy-to-understand approach to solving classification problems and are sometimes referred to as regression trees. They are structured like flowcharts, with every internal node representing decision or test on predictor variable.

In our model, we construct decision tree using Gini impurity, which is found out:

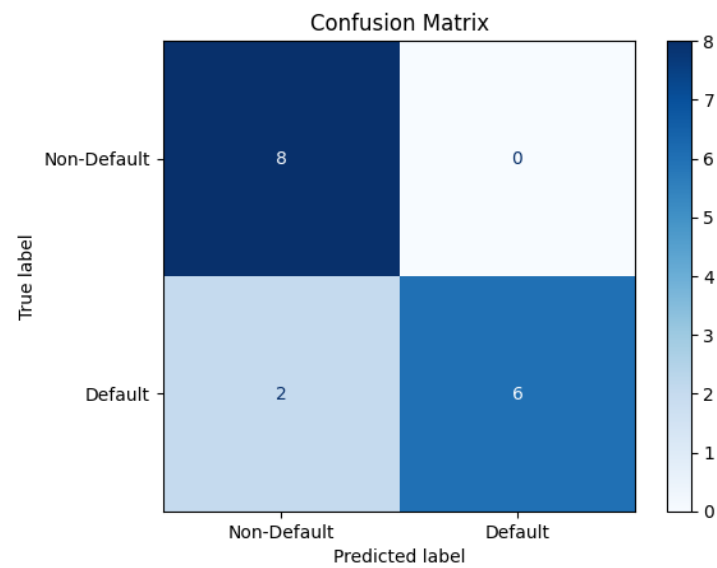
$$\text{Gini Impurity of a leaf} = 1 - (P(1))^2 - (P(0))^2$$

Every terminal node, or "leaf," depicts possible classification outcome. Branches signify decision paths, and they can be further split if additional decisions lead to more outcomes. Each branch also incorporates associated costs or criteria, guiding the decision-making process.

5.3.8.1 Measures

- **Confusion Matrix**

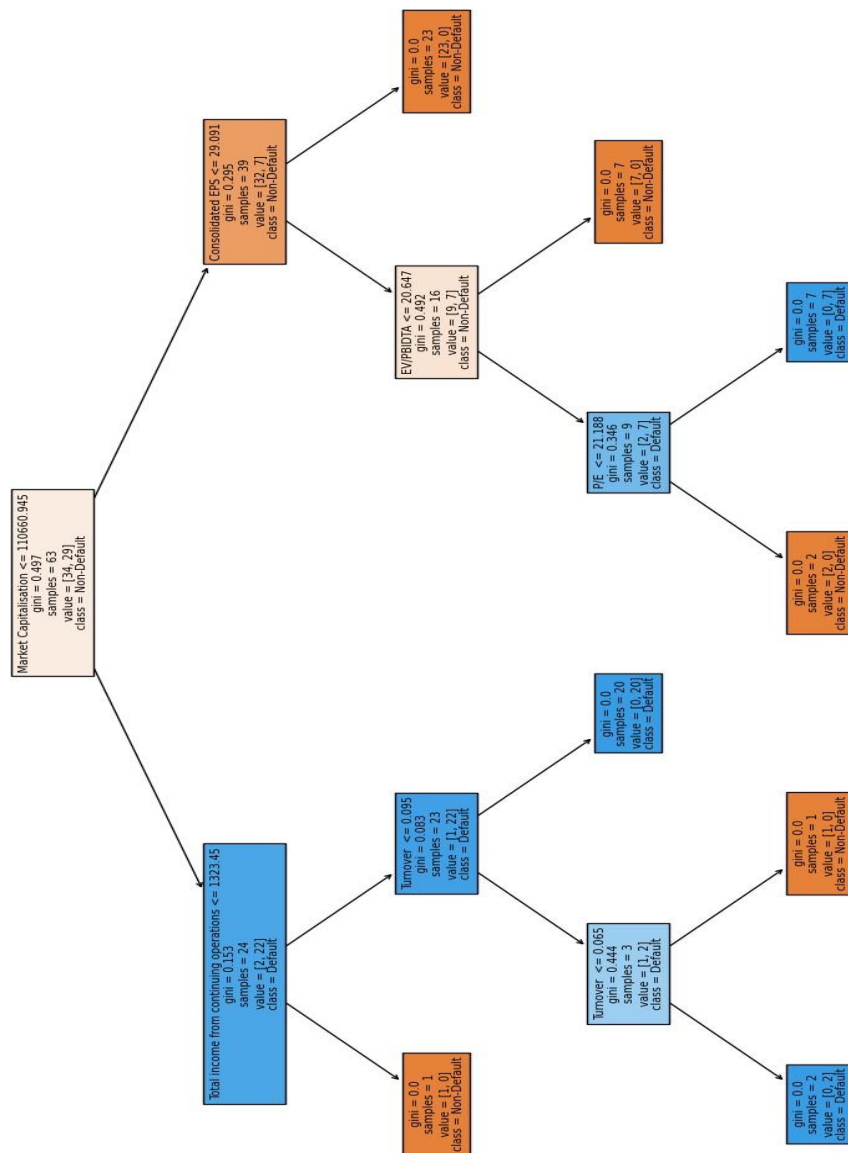
Figure 10: Confusion Matrix for DT



Source: Self Compiled

- **Decision Tree**

Figure 11: Decision Tree



Source: Self Compiled

CHAPTER 6: FINDINGS AND RECOMMENDATIONS

Based on the performance metrics and the compiled results presented below, we can draw conclusions.

Table 6: Consolidated Model Performance Outcomes

Measures	LOGIT	KNN	Gradient Boost	LDA	DT
Accuracy	0.875	0.8125	0.9375	0.9375	0.875
Precision	1	0.777	1	1	1
Recall	0.75	0.875	0.875	0.875	0.75
F1 Score	0.857	0.8235	0.933	0.933	0.857
AUC-ROC Score	0.875	0.8125	0.9375	0.9375	0.875

Source: Self Compiled

- Logistic regression (Logit)** delivered strong values from accuracy point of view (0.875), precision (1), and F1 score (0.857). But, it lagged in recall (0.75), which was lower relative to some other models. Despite this, Logit outstands for its simplicity and easiness of interpretation.
- KNN (k-Nearest Neighbours)** showed weaker performance overall, with an accuracy of 0.8125, a recall of 0.875, and an F1 score of 0.8235. Its precision score of 0.777 was reasonable, but not outstanding. While KNN can handle non-linear relationships effectively, it may not be the most suitable model for predicting corporate defaults.
- Gradient Boosting** emerged as the top performer across all evaluation metrics—achieving perfect scores for AUC-ROC (1) and precision (0.9375), and high values for F1 score (0.9375), recall (0.875), and accuracy (0.933). Despite its superior performance, it comes with increased complexity and higher computational requirements compared to simpler models.
- LDA (Linear Discriminant Analysis)** performed almost on par with Gradient Boost, posting excellent scores: 0.9375 for accuracy, 1 for precision, 0.875 for recall, 0.933 for F1 score, and 0.9375 for AUC-ROC. It offers a good balance between performance and simplicity, and like Logit, it is interpretable and computationally efficient.

- **Decision Trees (DT)** showed competitive results in terms of accuracy (0.875) and F1 score (0.857). It also achieved a perfect precision score (1), though its recall (0.75) was on the lower side. DTs are well-suited for non-linear classification problems and offer a high degree of interpretability.

Overall, Gradient Boosting and LDA are the standout models for predicting corporate default, consistently delivering strong performance across key metrics like F1 score, AUC-ROC, sensitivity and classification accuracy.

Conversely, KNN was least effective of the five models assessed. Ultimately, the ideal selection of model relies on the particular requirements of credit institution—whether priority lies in performance, computational efficiency, or interpretability.

CHAPTER 7: LIMITATIONS OF THE STUDY

The intent behind this report was to deliver a thorough and comprehensive analysis. However, several limitations were encountered during the course of the study:

Limited Sample Size:

The dataset comprised only 79 companies, which may not fully encapsulate the complexities and nuances of the Indian corporate landscape. This relatively small sample may restrict the generalizability of the findings.

Restricted Model Scope:

The study evaluated only five machine learning models, leaving room for further exploration of more advanced or unconventional tools like ANNs (Artificial Neural Networks) and broader AI-based approaches.

Model Complexity vs. Accessibility:

Some of the algorithms employed, while powerful, are complex and may not be easily understood by a non-technical audience. In some cases, the marginal performance gains may not justify the trade-off in interpretability.

Lack of Industry-Specific Focus:

Although the study's general approach enhances its applicability across sectors, the absence of an industry-specific lens may reduce predictive accuracy. A sector-focused analysis could yield more precise insights and represents a promising avenue for future investigation.

No Use of Feature Engineering:

The study did not incorporate advanced feature engineering techniques that could refine input variables and potentially improve model performance. Integrating these methods could lead to more robust predictive outcomes.

Assumption Simplification:

Certain assumptions inherent to the models may have been intentionally simplified or overlooked to enable consistent implementation across algorithms. This may have resulted in anomalies or inaccuracies in specific outcomes.

CHAPTER 8: CONCLUSION

This report presents a comprehensive review of default prediction models currently utilized

across the global financial landscape. It is increasingly clear that both financial institutions and regulatory bodies are raising the bar for more robust and secure credit assessment standards, intensifying the demand for highly accurate default prediction mechanisms. Advances in computational power and machine learning have enabled the development of sophisticated models that build upon traditional frameworks, such as Altman's Z-Score, enhancing predictive precision.

Furthermore, the report highlights the value of adopting an ensemble modelling approach, emphasizing the benefits of leveraging multiple algorithms and predictive strategies to make more informed and resilient credit decisions.

Machine learning models have proven to be not only accurate but also highly adaptable, making them practical tools in the domain of default prediction. Techniques such as Linear Discriminant Analysis (LDA), Decision Trees, and others have demonstrated considerable effectiveness as classification tools within this study's context.

In conclusion, this report underscores the growing relevance and effectiveness of machine learning in credit risk modelling. It also encourages ongoing exploration and research in this area to drive further improvements in predictive accuracy including the reliability of decision-making.