Bankruptcy Prediction Model Using Machine Learning

Abstract

In the modern era, Machine learning models have become integral across various sectors worldwide. Bankruptcy prediction is a critical subject that should be discussed. This report presents a comprehensive study on developing a Bankruptcy Prediction Model using machine learning techniques. Leveraging data from data sources such as Prowess-dx and Bloomberg Terminal we obtained the superset of all the companies in the National Stock Exchange with their financials as data points. The report also dives into various ML models which can be used to predict bankruptcy. It outlines the data collection process, feature engineering, model selection, and evaluation metrics used to build a predictive framework. This project aims to enhance financial understanding and deliver actionable insights to stakeholders.

# **Introduction**

## The Importance of Predicting Corporate Bankruptcy

Bankruptcy forecasting plays a critical role in financial risk management, benefiting investors, creditors, regulators, and businesses. An accurate model helps stakeholders assess their creditworthiness, avoid losses, and ensure stability. Companies monitor systemic risks while using them as early warnings to take corrective actions. Bankruptcy can disrupt the economy, cause job losses, and reduce investor’s trust.

While traditional methods such as Altman’s Z-Score have limitations, Machine Learning offers a more robust solution by analyzing complex data patterns. This project uses ML to improve bankruptcy forecasting and provide implementable insights into financial stability and decision-making.

## 

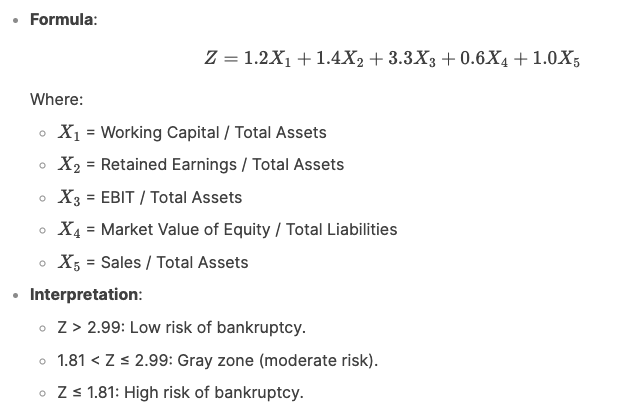
## 

## Traditional Models Used to Predict Bankruptcy

After doing some literature review, bankruptcy prediction models that are being used over a long time are **Altman Z-score** and the **Logistic Regression Model.**

### Altman Z-Score

Altman's Z-score is an extensive financial model for predicting the likelihood of corporate bankruptcy. Developed by Edward Altman in 1968, the Z-score combines several financial situations into a single value to assess the financial health of a company. The Z-score equation is:



Z score gives a single value that can be used to decipher the possibility of bankruptcy. But at the same time, this score is based on very few financial metrics. Also, it is mainly developed for manufacturing companies and does not work in other industries. This model is based on historical financial data and reduces adaptability to dynamic economic situations and sudden obstacles. Additionally, the market value of capital is required, limiting its applicability to listed companies and the exclusion of private companies. Z-scores also use only five financial parameters, allowing you to simplify the complex factors affecting your financial burden.

### Logistic Regression Model.

It estimates the probability of binary outcomes (in this case bankruptcy or non-bankruptcy) based on many financial predictors such as debt rates, profitability, and liquidity metrics. In contrast to Altman's Z-score, logistics regression can handle a wider area of ​​variables and is not limited to a particular industry.

However, it assumes a linear relationship between predictors and log bankruptcy data. This may not record non-linear and complex patterns in financial data. Despite this limitation, it remains a general basic model for assessing financial burdens.

## 3. Why are Machine Learning Models better than Traditional ones?

Machine learning models (ML) are considered to be superior to traditional bankruptcy forecasts for several important reasons.

1. **Handling Complex, Non-Linear Relationships**: ML models, such as decision trees, random forests, and neural networks, can capture intricate, non-linear patterns in financial data that traditional models like Altman's Z-score or logistic regression may miss.

2. **Scalability with large data records**: ML models are published in the processing of large data volumes containing structured (financial status) and unstructured data (text, market sentiment, etc.). This leads to a more comprehensive analysis.

3. **Functional Engineering**: ML models can automatically identify and use important features from raw data.

4. **Improved Accuracy**: Advanced ML techniques such as ensemble methods and deep learning often achieve higher predictability and robustness compared to traditional statistical models.

6. **Flexibility**: ML models are applied to industry and can be adapted to specific applications, as opposed to traditional models that are limited to specific sectors and assumptions.

In summary, ML models offer greater flexibility, accuracy, and adaptability, making them more effective for modern bankruptcy prediction tasks.

# **Methodology: Data Collection, Machine Learning Models, and Implementation Framework**

## Data Collection, Cleaning and Merging

By the use of Prowess-dx and the Bloomberg terminal we were able to get relevant data.

Prowess was used to obtain the superset of all National Stock exchange companies and all of the financial and basic data of these companies. However, bankruptcy of a company is not specified in Prowess.

Therefore with the use of Bloomberg terminal, the companies with Most recent Bankruptcy Insolvency Date available, were filtered out. Filtering out companies with this date gave us the list of all the bankrupt companies from 2014 to 2024.

After cleaning and merging both of the datasets using Python coding (using the reference of company names) a final dataset with financial data of all the bankrupt companies and a training set with all the companies which will be used to test various models is obtained.

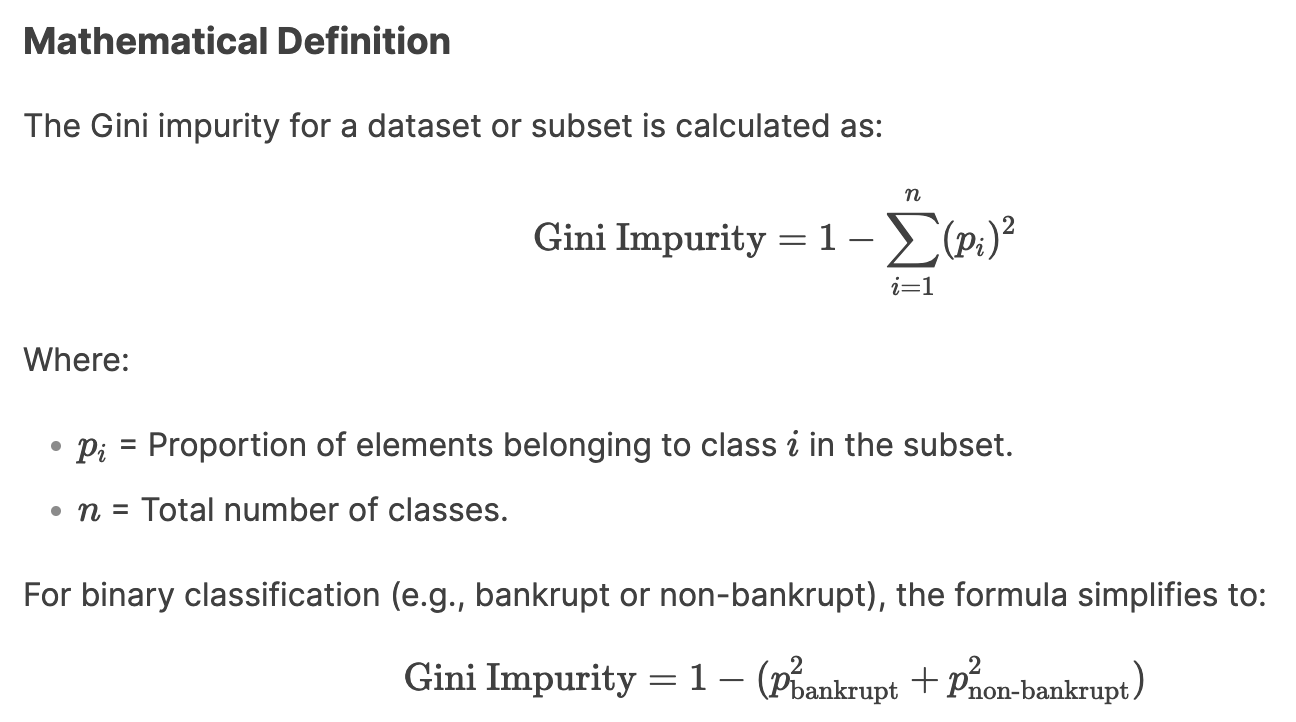
In the final data set comprised of around 128 bankrupt companies out of which 80 were mapped. The models were trained on 82 parameters which is way more than what was used in the traditional models.

## Machine Learning Models for Bankruptcy Prediction: Overview and Implementation

### Decision Tree Classifier

It is a monitored machine-learning algorithm used for classification tasks. Based on the properties, it divides data into smaller sub-quantities, creating a tree-like structure of decisions. Each internal node represents a characteristic-based decision, each branch represents an outcome of this decision, and each leaf node represents a class name (in this case not bankrupt and bankrupt).

We have used the Gini criteria to split the data at each step. The process continues until a stopping condition is satisfied.

Gini criteria finds the probability of incorrect classification of a random element if it was randomly labelled. 

In this case the Gini criteria is used instead of Entropy criteria as it is faster and generally produces similar results compared to the more detailed ones by entropy criteria.

Gini criteria finds out the most significant financial features and by minimising its impurity at each split, the model classifies the companies based on their financial health.

### Random Forest Model

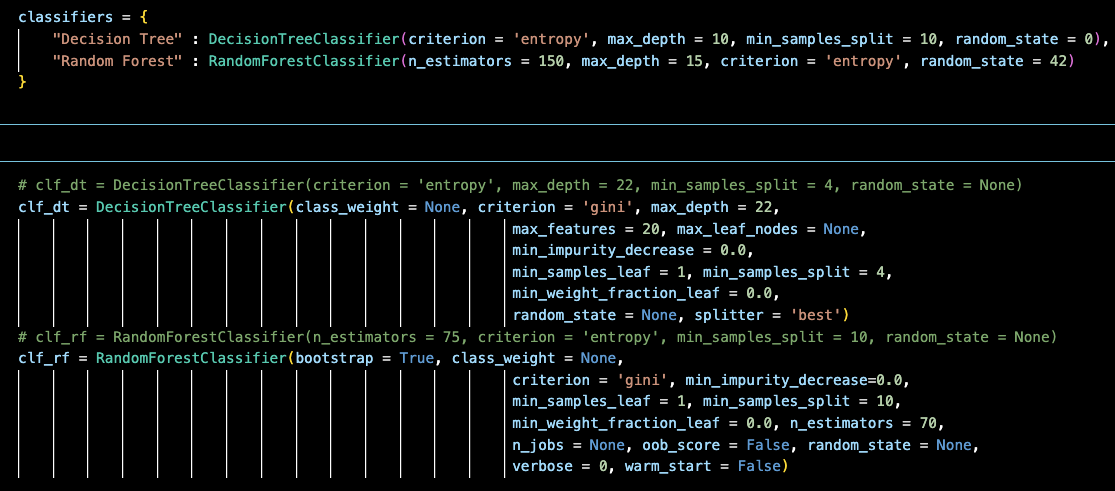
Random Forest is an ensemble learning method used for both classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. It is known for its high accuracy, robustness, and ability to process large-scale data records.

How Random Forest Works

1. **Bootstrapping**:
   * Random subsets of the training data are created using bootstrapping (sampling with replacement).
   * Each subset is used to train a separate decision tree.
2. **Feature Randomness**:
   * At each split in a decision tree, a random subset of features is considered.
   * This ensures that the trees are diverse and reduces the risk of overfitting.
3. **Aggregation**:
   * For classification tasks, the final prediction is determined by majority voting across all trees.
   * For regression tasks, the final prediction is the average of all tree predictions.

This model needs specification of some parameters, such as,

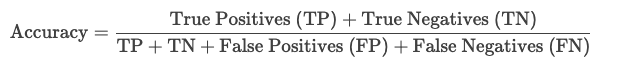
1. n\_estimators (Number of Trees) - This parameter specifies the number of decision trees to be built in the Random Forest.
2. max\_depth (Maximum Depth of Trees) - This parameter controls the maximum depth of each decision tree in the forest.
3. max\_features (Number of Features to Consider at Each Split) - This parameter determines the number of features to consider when looking for the best split at each node



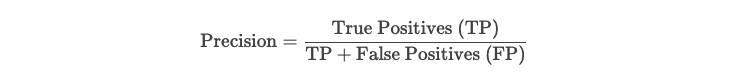
# 

## Evaluation Metrics

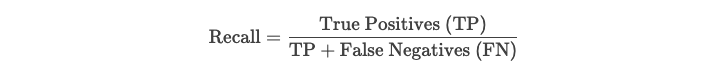
1. **Accuracy** - It is a simple representation of the correct prediction a model can make, in this case, of both bankrupt and non-bankrupt companies. This tells you overall, the proportion of times the model makes correct decisions.



1. **Precision** - It is the proportion of correctly guess bankrupt companies out of all the companies in the dataset.



1. **Recall or Sensitivity** - Proportion of total bankrupt companies the model identifies.



1. **F1 score** - The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

# 

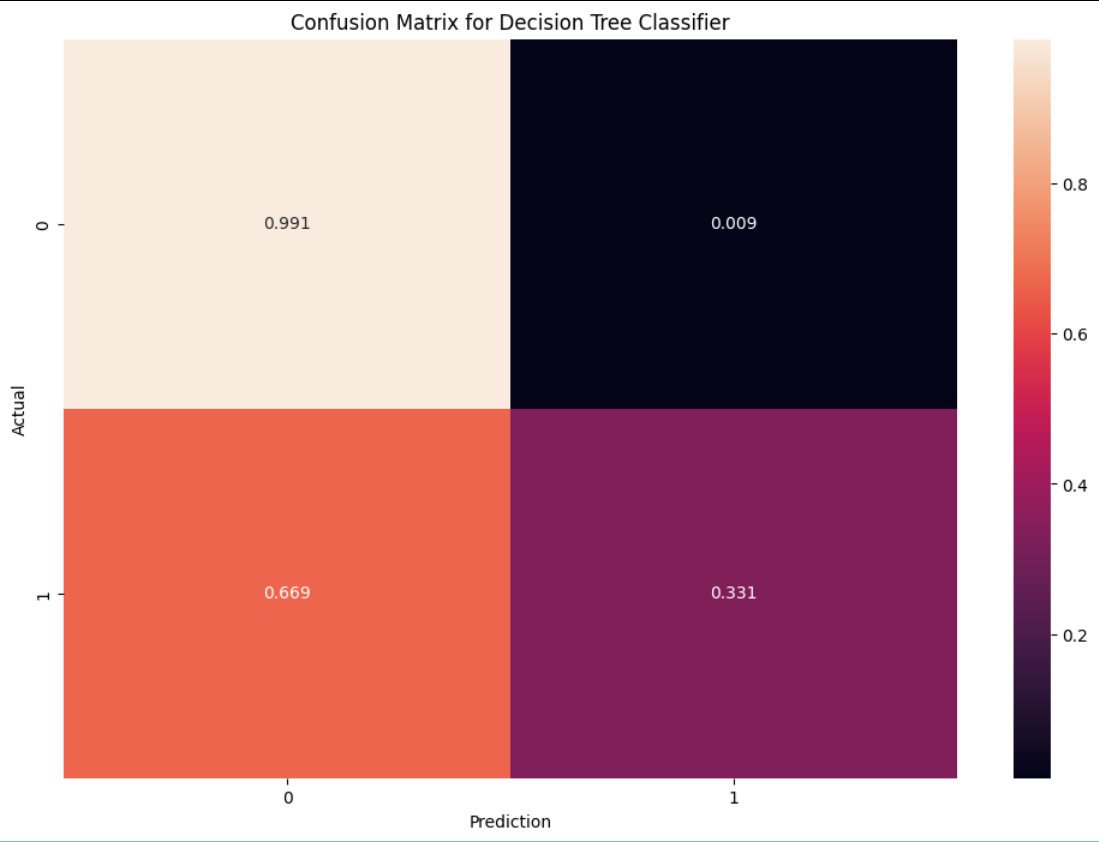
# **Observations**

After running the code for both the models, we get the following outputs from judgement parameters,

**Metrics for the Decision Tree:**

* Accuracy: 0.9822
* Precision: 0.6723
* Recall: 0.6613
* F1\_Score: 0.6666

The model identifies 66.13% of the actual bankrupt companies. This means that about 33.87% of bankrupt companies are missed (false negatives), which could be risky for stakeholders.

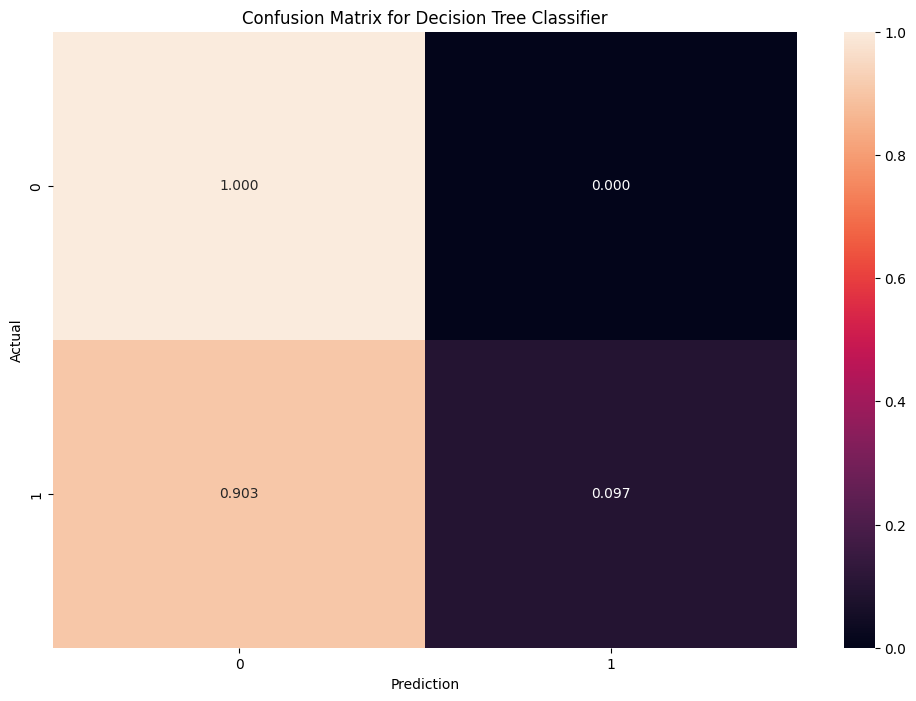


A F1 score of 0.6666 indicates a moderate balance between precision and recall, but there is room for improvement.

**Metrics for the Random Forest:**

* Accuracy: 0.9874
* Precision: 0.9937
* Recall: 0.5487
* F1\_Score: 0.5856

Extremely high precision means that almost all predicted bankruptcies are accurate, with very few false positives. If recall is considered, 45.13% of bankrupt companies are missed, which is a significant drawback.



From the above data we can say that, The Decision trees and Random Forest classifiers show very poor results, and using the heatmap we can see that the misclassification rate of bankrupt companies is very high (67% in DT and 90% in RF).

# **Conclusion**

From the data collected from various research papers, code outputs and various resources we can decipher the following conclusion.

This report explores the subject of **Bankruptcy prediction using Machine Learning Models.** By exploiting datasets such as Prowess and Bloomberg terminal, financial data of NSE companies was compiled with around 128 bankrupt companies in the years of 2014 to 2024. The models were trained on 82 parameters far exceeding traditional methods such as Altman Z score and Logistic regression, with advanced ML algorithms such as Random forest and Decision Tree classifiers, evaluating the process using parameters such as Recall, Accuracy, Precision and F1 score.

### Key Findings

1. **Machine Learning Models:**

* Decision Tree Classifier: Achieved 98.22% accuracy with moderate precision (67.23%) and recall (66.13%), demonstrating a reasonable balance for identifying bankrupt companies.
* Random Forest: Outperformed in accuracy (98.74%) and precision (99.37%) but had lower recall (54.87%), indicating high reliability in predictions but a tendency to miss nearly half of actual bankrupt cases.
* ML models excelled in handling complex, non-linear relationships and large datasets, proving superior to traditional methods.

1. **Tradeoffs**:

* Decision Tree prioritized recall, making it suitable for stakeholders focused on minimizing missed bankruptcies.
* Random Forest prioritized precision, ideal for avoiding false alarms but risking undetected bankruptcies.

### 

### 

### 

### 

### Future Directions

* Parameter Tuning - input parameters can we optimised well to fine tune the recall of the models.
* Class Imbalance techniques - Methods like Sythetic Minority Oversampling and Weight Classing can be used to make the models better.
* Models can be stacked together to balance out the problems and get an even better result.
* Real time adaptation - Continuously updatable models with new financial data to account for latest market conditions.

By providing insights, the model can be useful to investors, stakeholders and regulators in various ways.

# **References**

1. Machine learning models and bankruptcy prediction by Flavio Barbozaa, Herbert Kimurab, Edward Altmanc.
2. Predicting financial distress in TSX-listed firms using machine learning algorithms by Mark Eshwar Lokanan and Sana Ramzan
3. Financial Distress forecasting with Machine Learning Approach by Hong Hanh Ha, Ngoc Hung Dang, Manh Dung Tran.
4. Prowess Dx
5. Bloomberg Terminal
6. [www.investopedia.com](http://www.investopedia.com)