A Project Report

On

Predictive Modeling for Financial Distress

BY

Khush Bhuta

2022A7PS1333H

Under the supervision of

Prof. Shreya Biswas

SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS OF

CS F266: STUDY PROJECT

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(February, 2025)

ACKNOWLEDGMENTS

I am extremely grateful to Prof. Shreya Biswas for entrusting me with the opportunity to pursue this project on Predictive Modeling for Financial Distress. I am grateful for her insightful feedback and encouragement that has significantly increased my understanding on this topic. I would like to thank my teammates – Srujaan, Saaketh and Piyush for guiding me, and resolving any queries that I had during the course of this project.

Lastly, I would like to thank the CS-IS and the Economics Department of BITS Hyderabad for providing me with the infrastructure and resources to be able to pursue this project.

Birla Institute of Technology and Science-Pilani,

Hyderabad Campus

Certificate

This is to certify that the project report entitled "**Predictive Modeling for Financial Distress**" submitted by Mr. Khush Bhuta (ID No. 2022A7PS1333H) in partial fulfillment of the requirements of the course CS F266, Study Project Course, embodies the work done by him under my supervision and guidance.

Date: 23rd February, 2025

(Prof. Shreya Biswas)

BITS- Pilani, Hyderabad Campus

ABSTRACT

In the modern era, Machine Learning models have become integral across various sectors worldwide. Bankruptcy prediction is a critical subject in the Finance sector. Finding an intersection between Finance and Computer Science is essential. This report presents a comprehensive study on developing a Bankruptcy Prediction Models for **Indian Companies** using Machine Learning and Deep 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 Machine Learning 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. Our target with this project is to develop useful predictive models for companies in the Indian context. This project aims to enhance financial understanding and deliver actionable insights to stakeholders.

CONTENTS

1. Introduction:

- *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.**

a. 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:

- **Formula**:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

  Where:
  - X_1 = Working Capital / Total Assets
  - X_2 = Retained Earnings / Total Assets
  - X_3 = EBIT / Total Assets
  - X_4 = Market Value of Equity / Total Liabilities
  - X_5 = Sales / Total Assets
- **Interpretation**:
  - Z > 2.99: Low risk of bankruptcy.
  - 1.81 < Z ≤ 2.99: Gray zone (moderate risk).
  - Z ≤ 1.81: High risk of bankruptcy.

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.

b. 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.

- *Comparing Machine Learning Models to Traditional Models:*

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 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.

2. Literature Review:

I have studied about Fundamentals of Finance and Accounting, Derivatives and Risk Management, Security Analysis and Portfolio Management and Financial Management as a part of my Finance Minor. I am currently pursuing Financial Risk Analytics and Management and Business Analysis and Valuation. Throughout these courses, I gained a lot of insight on how financial ratios are important in making correct investment decisions, and their implications in the sector. However, I lacked actual implementation skills, and hence am pursuing this project, to find the intersection between the Computer Science and Finance Domains through Machine Learning. To gain more insight on the research work done previously in the domain of Bankruptcy Prediction, I underwent thorough readings of the following research papers. I have listed the key insights associated with them.

- ***Deep Learning-Based Model for Financial Distress Prediction:*** This paper was based one applying the Adaptive Whale Optimization Algorithm (AWOA) along with Deep Learning to improve and optimize model parameters, select features and improve the prediction accuracy. The AWOA is used since it can identify the most relevant financial features (market ratios and indicators), reducing the noise and dimensionality. It can enhance time-series forecasting by fine-tuning parameters of sequential models like LSTMs or GRUs. The model yielded a weighted-average accuracy of 95.8% after applying this algorithm.
- ***Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction - A comprehensive study:*** This paper focuses on optimizing the feature vector, and dimensionality of features for a Bankruptcy Prediction Machine Learning Model. It lays emphasis on the use of Corporate Governance Indicators like – Board Structure, Ownership Structure, Key People retained and Cashflow rights in the model. It also shows that the best performing model was a Support Vector Machine (SVM), and it was obtained by performing Stepwise Discriminatory Analysis (SDA). Numerous models like – KNN's, Naïve-Bayes Classifiers, Multilayer Perceptron (MLPs) were trained on an open-source dataset published by the Taiwanese Economic Journal, but the SVMs gave the best performance metrics.

We did not find a lot of research on Bankruptcy Prediction for Indian Listed Companies, and decided to pursue a research project on the same.

3. Methodology:

Step 1: *Data Collection, Cleaning and Merging:*

By leveraging the Prowess-dx database, and the Bloomberg Lab at our campus, we gained access to relevant data. The Prowess-dx database was used to extract Raw Financials, Credit Rating Data and Equity Ownership Patterns for all listed NSE and BSE companies for a fixed timeframe (April 1st 2014 – March 31st 2024). This data was extracted in the form a .txt with '|' delimiters, which was converted to separate .csv files using Excel. Besides this data, we obtained the list of Indian Listed Companies that had filed for insolvency using the Bloomberg Terminal. This data was obtained as a .csv file.

The next step was to clean the respective files and merge the data. Since the data was extracted from different sources, the 'Firm Name' was the only common underlying field for merging the data. We applied a left join on the 'Name' field using the pandas library in python. This resulted in a merged database. The merged data was cleaned and refined using Pandas.

Using the Left Join, an additional 'Label' column was created, assigning Binary Values (*0 – Non-Bankrupt/1 - Bankrupt*) to the database. This is useful while classifying the Bankrupt v/s Non-Bankrupt companies. The final dataset was used to train Machine Learning classifiers, namely – Decision Trees and Random Forests.

	Name	Company code	Total income	Sales	Sales returns	Net sales	Sales / Net fixed assets	Change in stock	Total expenses	Profit after tax	...	Change in cash and bank balance	Change in current liab and provisions	Change in PAT net of P&E	Change in sales	Change in total income	Change in working capital	Change in working capital assets	Change in working capital liabilities	Revenue	Label
0	20 MICRONS LTD.	11	3097.0	3010.4	NaN	2888.8	204.9425	51.2	3146.9	1.3	...	26.1	403.2	-22.9	154.6	195.1	-157.7	240.9	210.4	NaN	0
1	20 MICRONS LTD.	11	3316.2	3252.7	NaN	3116.7	227.7801	21.3	3378.2	-40.7	...	100.9	-115.1	-50.3	242.3	219.2	182.9	57.9	183.6	NaN	0
2	20 MICRONS LTD.	11	3527.6	3495.1	NaN	3338.3	249.5431	8.2	3442.4	93.4	...	-80.7	1.6	157.3	242.4	211.4	25.1	46.9	-145.9	NaN	0
3	20 MICRONS LTD.	11	3741.8	3729.3	NaN	3569.9	239.0117	-42.8	3567.4	131.6	...	11.4	-5.7	31.3	234.2	214.2	-48.2	-50.3	-11.5	NaN	0
4	20 MICRONS LTD.	11	3922.0	3891.8	NaN	3845.7	237.2036	7.2	3770.4	158.8	...	-25.6	51.4	32.1	162.5	180.2	28.4	80.8	85.4	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
55072	HAMPS BIO LTD.	713053	46.5	46.5	NaN	46.5	178.1609	0.1	48.0	-1.4	...	0.2	-1.0	0.4	8.2	8.1	5.3	4.3	-1.8	NaN	0
55073	HAMPS BIO LTD.	713053	40.0	38.3	NaN	38.2	126.4026	0.5	42.7	-2.2	...	0.9	9.4	-2.5	-8.2	-6.5	-11.0	-1.6	4.2	NaN	0
55074	HAMPS BIO LTD.	713053	53.4	53.4	NaN	53.4	223.4310	-0.9	51.5	1.0	...	-0.9	-3.4	4.9	15.1	13.4	2.5	-0.9	-3.5	NaN	0
55075	HAMPS BIO LTD.	713053	55.9	55.8	NaN	55.5	217.1206	-1.6	50.6	3.7	...	-0.1	3.2	2.7	2.4	2.5	-4.6	-1.4	1.6	NaN	0
55076	HAMPS BIO LTD.	713053	64.9	64.9	NaN	64.8	257.5397	7.7	67.6	5.0	...	0.1	-7.3	1.3	9.1	9.0	18.3	10.9	-2.3	NaN	0

Step 2: *Machine Learning Models and Intepretation*

The **Decision Tree ML Algorithm** is used for Classification Tasks. It creates a tree-like structure of decisions by dividing data into smaller sub-quantities. Each internal node represents a characteristic-based decision and each branch represents an outcome of this decision. Each leaf node represents a class name (bankrupt/not bankrupt). We have used the Gini criteria instead of Entropy to split the data in each step. The process terminates when the stopping condition is satisfied. Gini critera finds the probability of incorrect classification of a random element if it was randomly labelled.

Mathematical Definition

The Gini impurity for a dataset or subset is calculated as:

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Where:

- p_i = Proportion of elements belonging to class i in the subset.
- n = Total number of classes.

For binary classification (e.g., bankrupt or non-bankrupt), the formula simplifies to:

$$\text{Gini Impurity} = 1 - (p_{\text{bankrupt}}^2 + p_{\text{non-bankrupt}}^2)$$

The entropy criteria is a measure of impurity. We start with all examples at the root node. Calculate the **information gain** for all possible features and pick the one with the highest information gain. Next, we split the dataset according the selected feature and create left and right branches of the tree. Gini criteria is preferred over Entropy since it is faster. This evaluation criteria finds out the most significant financial features by minimizing its impurity at each split. The classifier parameters for the Decision Tree are – max_depth and min_samples_split.

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=42)
✓ 0.0s
```

The Train-Test Split used for the classifiers is in the 80-20 ratio.

The Random Forest Classification Algorithm is a **tree ensemble**. Trees are highly sensitive to small changes in data. Hence, we train an ensemble/collection of Decision Trees. The majority vote of the trees in the ensemble can be used to vote and the final prediction will be the majority.

Bootstrapping (sampling with replacement): is used for creating an ensemble dataset. Each subset is used to train an individual Decision Tree.

Feature Randomness: At each split in the Decision Tree, a random subset of features is considered instead of all the features. This reduces the chances of overfitting the tree. The algorithm states that – At each node, when choosing a feature to use to split, if n-features are available, pick a random subset of k<n features and allow the algorithm to choose from that subset of features only.

Aggregation: For our classification task, the final prediction is determined by majority voting across all trees.

Some important parameters needed by the model and their applications are:

- n_estimators – Number of Decision Trees in the Random Forest
- max_depth – Depth of each Decision Tree
- max_features – Number of features to be considered at each split (k<n)

```
classifiers = {
    "Decision Tree" : DecisionTreeClassifier(criterion = 'gini', max_depth = 10, min_samples_split = 10, random_state = 0),
    "Random Forest" : RandomForestClassifier(n_estimators = 150, max_depth = 15, criterion = 'gini', random_state = 42)
}
```

✓ 0.0s

```
clf_dt = classifiers["Decision Tree"]
clf_rf = classifiers["Random Forest"]
```

✓ 0.0s

The image above shows the parameters used for each classifier. We have named the two classifiers as clf_dt (Decision Tree) and clf_rf (Random Forest).

4. Evaluation and Observations:

The evaluation metrics used for our classifier are – Accuracy, Precision, Recall and F1-Score. The implications of each metrics are mentioned below:

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.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}}$$

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

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Positives (FP)}}$$

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

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Negatives (FN)}}$$

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

Our Decision Tree and Random Forest Classifiers yielded the following results:

```
Metrics for the Decision Tree Classifier:
Accuracy: 0.9863
Precision: 0.7545
Recall: 0.6282
F1_Score: 0.6697
```

```
Metrics for the Random Forest Classifier:
Accuracy: 0.9875
Precision: 0.9937
Recall: 0.5519
F1_Score: 0.5910
```

Since our dataset is highly skewed (~75 Bankrupt Companies v/s ~5500 Non-Bankrupt Companies) the model has achieved high accuracy by predicting the majority class only. Low Precision in DT indicates that the model has made a significant number of false positives. However, high Precision in the RF classifier indicates that the model is conservatively predicting anomalies, minimizing the false positives, but missing out on a lot of anomalies. Low Recall and F1-Score are common in both the classifiers.

Further improvements planned for the model:

- The merged dataset we are currently using has over 55,000 rows (~10 rows per company) sorted according to the year. On creating a single row entry for each company in the classifier, we will be able to achieve better results since the model will be able to map relationships between the same features across different years. This will significantly reduce the dataset size, but also reduce the misclassification of anomalies.
- We can iteratively find the best parameters by running a loop on the max_depth, min_samples_split and the n_estimators parameters of the classifiers.
- We can use advanced models like LightGBM or XGBoost and other learning algorithms like Neural Networks for anomaly detection.
- Further, we can finetune the hyperparameters, perform feature engineering to reduce the feature vector size, adjust the decision threshold and use resampling to reduce class imbalance.

5. Conclusion:

This report explores the topic – "Predictive Modeling for Financial Distress". By leveraging the Bloomberg Terminal and the Prowess DB, we were able to successfully create a database that never existed previously in the Indian context. We further plan to create numerous Machine Learning and Deep Learning models for the same. We have currently trained the classifiers on 82 parameters, and we plan to optimize the feature vector size and bring it down to 20.

6. References:

- Liang, Deron, et al. "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study." *European journal of operational research* 252.2 (2016): 561-572.
- Zhensong Chen, Wei Chen, Yong Shi, Ensemble learning with label proportions for bankruptcy prediction.
- Elhoseny, Mohamed, et al. "Deep learning-based model for financial distress prediction." *Annals of operations research* (2022): 1-23.
- Scikit-learn Documentation.
- Prowess-Dx, Bloomberg Terminal.