# PREDICTING THE FINANCIAL STABILITY OF A COMPANY

**OMPRAKASH .S.**
*M.Sc (Decision and Computing Sciences) – IV$^{th}$ year*
*Coimbatore Institute of Technology*
*Coimbatore, India*
mailto:1933022mdcs@cit.edu.in

**DEEPAK RAJ.A**
*M.Sc (Decision and Computing Sciences) – IV$^{th}$ year*
*Coimbatore Institute of Technology*
*Coimbatore, India*
mailto:1933007mdcs@cit.edu.in

**DANESH DHEERTHAN .J**
*M.Sc (Decision and Computing Sciences) – IV$^{th}$ year*
*Coimbatore Institute of Technology*
*Coimbatore, India*
mailto:1933005mdcs@cit.edu.in

**Dr.V.SAVITHRI**
*Assistant Professor*
*Dept. of Computing (DCS)*
*Coimbatore institute of technology*
*Coimbatore, India*
mailto:v.savithri@cit.edu.in

*Abstract*— **Predicting the financial stability is a critical area of research in finance and accounting. The goal of this prediction is to develop models that can accurately identify financially distressed companies before they file for bankruptcy. Various financial ratios and accounting metrics are used to evaluate a company's financial health and assess its likelihood of bankruptcy. In recent years, the application of machine learning algorithms and statistical techniques has significantly improved the accuracy of bankruptcy prediction models. These models have practical applications in various domains, including corporate finance, investment analysis, and risk management. Accurate bankruptcy prediction can help stakeholders take preemptive measures to mitigate potential financial losses, thereby contributing to the stability of financial markets. We find that tree-based ensemble methods, especially R-tree, can achieve a high degree of accuracy in out-of-sample bankruptcy prediction By analyzing financial data, bankruptcy prediction models can provide early warning signals to stakeholders, enabling them to take preventive measures to mitigate potential financial losses. This abstract provides a brief overview of the importance of bankruptcy prediction and the role of predictive models in identifying financially distressed companies. While the prediction accuracy is similar to several previous models in the literature, our model is very simple to implement and represents an accurate and user-friendly tool to discriminate between bankrupt and non-bankrupt firms.**

## I. INTRODUCTION

Prediction of a financial stability , also named as corporate bankruptcy prediction or corporate failure prediction, has long been a significant topic in the field of accounting and finance , since the health of a firm is highly important to its creditors, investors, shareholders, partners, even its buyers and suppliers. This research topic can be traced back to almost 50 years ago, when discriminant analysis and logistic regression were two well-known statistical machine learning techniques used in bankruptcy prediction. Since 1990's, machine learning models have been extensively applied as tools to predict bankruptcy of firms, such as decision tree, neural networks and Support Vector Machines. Similar to the credit scoring, bankruptcy prediction is also typically a classification problem, which means it can be dealt with by classifying algorithms. In general, the task of bankruptcy prediction is to predict whether the firm will go bankrupt or not, which is a binary classification problem. To accurately conduct the prediction, we have to use algorithms to train the datasets, such as the financial data from the firm's financial statements .

This training process is where machine learning and deep learning techniques are applied. Through the training process of dataset, we can obtain a classifier with good classification accuracy, which can be used to do the bankruptcy prediction. This is the basic principle of bankruptcy prediction using machine learning. Historically, bankruptcy prediction models have relied on various financial ratios and accounting metrics to evaluate a company's financial health. However, the accuracy of these models was limited by the subjectivity of financial ratios and the inability to capture complex relationships among variables. Recent advancements in machine learning algorithms and statistical techniques have significantly improved the accuracy of bankruptcy prediction models.
These models can analyse large amounts of financial data and identify patterns that are not easily detectable by humans. This paper provides an overview of the importance of bankruptcy prediction and the role of predictive models in identifying financially distressed companies. The paper first discusses the motivation behind bankruptcy prediction and the consequences of financial distress. The paper then reviews the historical development of bankruptcy prediction models and highlights the limitations of traditional models. Finally, the paper describes recent advancements in machine learning algorithms and statistical techniques and their applications in improving the accuracy of bankruptcy prediction models

## .II. METHODOLOGY

### *A. DATASET DESCRIPTION*

The dataset contains a collection of various attributes in the banking and finance fields . The data includes 20 factors that may be useful for predicting the financial stability including sum, guarantees, credits, credit_report , marital_status,etc .It helps to create machine learning models to predict the financial stability of a company. Generally, the financial stability depends on these parameters. These parameters employ a very important role in the predictive analysis of the company's financial performance. The dataset consists of about 22 columns and about 800 rows. The information availed in the columns are

*sum:* This states the total sum of amount

*reason:* states the reason for the performance

*credit_report:* defines the state of the company

*marital_status:* states the marital status of the people

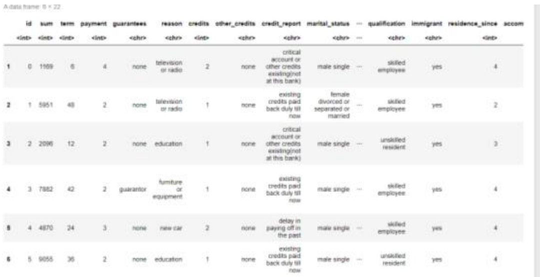
*Savings:* Total savings of the people

*Credit report:* It states the critical bank or other credits

*Qualification :* It states whether the employess are qualified or not

*Immigrant :* It says whether immigrant or not.

*Bankruptcy :* It says whether it is bankrupt or not.



A data frame: 6 × 22

| | id | sum | term | payment | guarantees | reason | credits | other_credits | credit_report | marital_status | ⋯ | qualification | immigrant | residence_since | accom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <int> | <int> | <int> | <int> | <chr> | <chr> | <int> | <chr> | <chr> | <chr> | ⋯ | <chr> | <chr> | <int> | |
| 1 | 0 | 1169 | 6 | 4 | none | television or radio | 2 | none | critical account or other credits existing(not at this bank) | male single | ⋯ | skilled employee | yes | 4 | |
| 2 | 1 | 5951 | 48 | 2 | none | television or radio | 1 | none | existing credits paid back duly till now | female divorced or separated or married | ⋯ | skilled employee | yes | 2 | |
| 3 | 2 | 2096 | 12 | 2 | none | education | 1 | none | critical account or other credits existing(not at this bank) | male single | ⋯ | unskilled resident | yes | 3 | |
| 4 | 3 | 7882 | 42 | 2 | guarantor | furniture or equipment | 1 | none | existing credits paid back duly till now | male single | ⋯ | skilled employee | yes | 4 | |
| 5 | 4 | 4870 | 24 | 3 | none | new car | 2 | none | delay in paying off in the past | male single | ⋯ | skilled employee | yes | 4 | |
| 6 | 5 | 9055 | 36 | 2 | none | education | 1 | none | existing credits paid back duly till now | male single | ⋯ | unskilled resident | yes | 4 | |

*Figure 2.1 Dataset description*

### *B. PROCESSING STEP*

Data cleaning is a very important step in preparing the data for machine learning process and model creation.. Some of the steps are mentioned below for cleaning data in financial stability prediction:

1. *Handling missing data:* If there are any missing values in the dataset, they need to be handled with precsion. An approach is to remove the samples with missing values and the other approach is to remove the samples with missing values.

2. *Handling outliers:* Outliers are data points that lie far from the majority of the data points and can adversely affect the model's performance. They need to be identified and handled appropriately. One common technique is to remove the outliers or transform the data using techniques such as log transformation.

3. *Handling categorical data:* Machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

4. Feature *Scaling :* Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable

5.*Contingency tables*: A contingency table displays frequencies for combinations of two categorical variables. Analysts also refer to contingency tables as crosstabulation and two-way tables. Contingency tables classify outcomes for one variable in rows and the other in columns. The values at the row and column intersections are frequencies for each unique combination of the two variables..

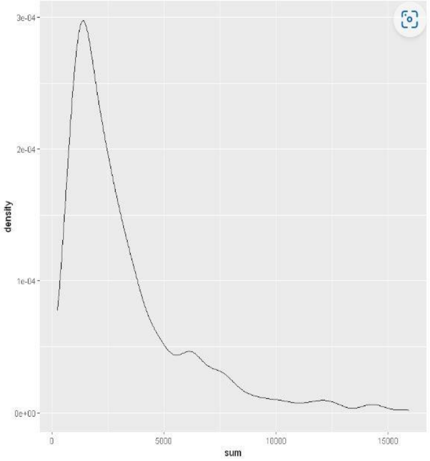


Figure 2.2 ggplot for density and sum.

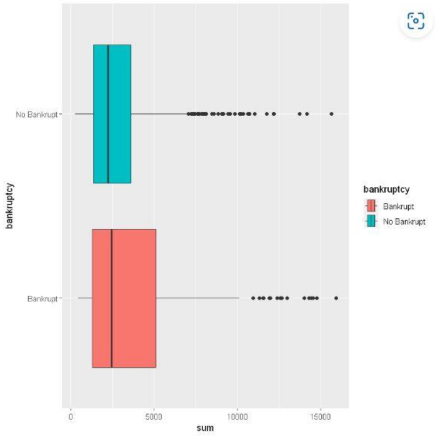


Figure 2.3 plot for sum and bankruptcy.

### *C. MODEL BUILDING STAGE*

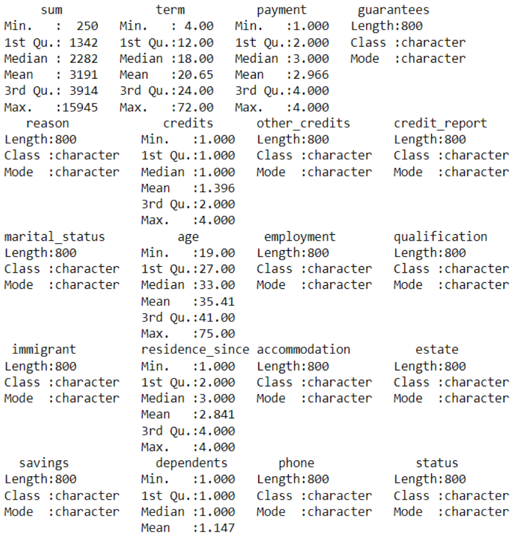
After performing data cleaning, the next step is to build the machine learning model for physical activity prediction. In this case, random forest algorithm is used to build the model. The following are the steps for building the model:

1. *Split the dataset into training and testing sets:* The dataset needs to be divided into two parts: one for training the model and the other for testing the model's performance. Typically, 70-80% of the data is used for training, and the rest is used for testing.

2. *Feature selection*: It is essential to select the most relevant features for the model to reduce the dimensionality of the data and improve the model's performance. This can be done using techniques such as correlation analysis or feature importance analysis.

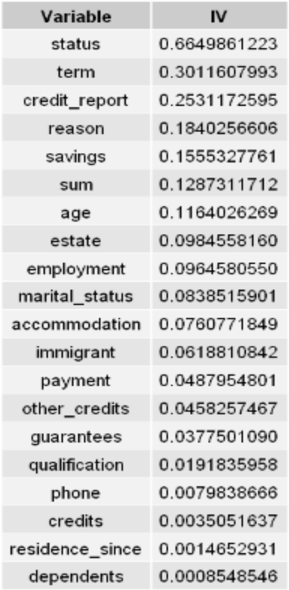
3. *Building the model:* After feature selection, build the random forest model using the training data. The model will learn to predict physical activity based on the selected features.

4. *Testing the model:* Finally, test the model's performance using the testing dataset. Evaluate the model's performance using metrics such as accuracy, Confusion matrix, precision, recall, and F1 score.



```
      sum             term          payment       guarantees
 Min.   :  250   Min.   : 4.00   Min.   :1.000   Length:800
 1st Qu.: 1342   1st Qu.:12.00   1st Qu.:2.000   Class :character
 Median : 2282   Median :18.00   Median :3.000   Mode  :character
 Mean   : 3191   Mean   :20.65   Mean   :2.966
 3rd Qu.: 3914   3rd Qu.:24.00   3rd Qu.:4.000
 Max.   :15945   Max.   :72.00   Max.   :4.000
    reason             credits      other_credits      credit_report
 Length:800         Min.   :1.000   Length:800         Length:800
 Class :character   1st Qu.:1.000   Class :character   Class :character
 Mode  :character   Median :1.000   Mode  :character   Mode  :character
                    Mean   :1.396
                    3rd Qu.:2.000
                    Max.   :4.000
 marital_status          age          employment         qualification
 Length:800         Min.   :19.00   Length:800         Length:800
 Class :character   1st Qu.:27.00   Class :character   Class :character
 Mode  :character   Median :33.00   Mode  :character   Mode  :character
                    Mean   :35.41
                    3rd Qu.:41.00
                    Max.   :75.00
  immigrant         residence_since accommodation         estate
 Length:800         Min.   :1.000   Length:800         Length:800
 Class :character   1st Qu.:2.000   Class :character   Class :character
 Mode  :character   Median :3.000   Mode  :character   Mode  :character
                    Mean   :2.841
                    3rd Qu.:4.000
                    Max.   :4.000
   savings            dependents        phone              status
 Length:800         Min.   :1.000   Length:800         Length:800
 Class :character   1st Qu.:1.000   Class :character   Class :character
 Mode  :character   Median :1.000   Mode  :character   Mode  :character
                    Mean   :1.147
```

*Figure 2.4 Summary statistics*



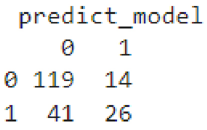
| Variable | IV |
|---|---|
| status | 0.6649861223 |
| term | 0.3011607993 |
| credit_report | 0.2531172595 |
| reason | 0.1840256606 |
| savings | 0.1555327761 |
| sum | 0.1287311712 |
| age | 0.1164026269 |
| estate | 0.0984558160 |
| employment | 0.0964580550 |
| marital_status | 0.0838515901 |
| accommodation | 0.0760771849 |
| immigrant | 0.0618810842 |
| payment | 0.0487954801 |
| other_credits | 0.0458257467 |
| guarantees | 0.0377501090 |
| qualification | 0.0191835958 |
| phone | 0.0079838666 |
| credits | 0.0035051637 |
| residence_since | 0.0014652931 |
| dependents | 0.0008548546 |

*Figure 2.4 Feature importance*

## III. ALGORITHMS USED

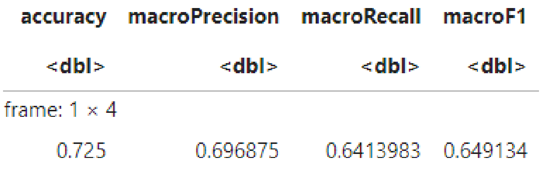
### *A. RANDOM FOREST CLASSIFIER*

Random forest classifier is a machine learning algorithm that can be used for prediction of financial stability . It is a type of ensemble learning method that uses multiple decision trees to make predictions. Each decision tree is undergone training on a random subset of the data, and the final prediction is based on the majority of all the decision trees. Overall, a random forest classifier can be a useful tool for physical activity prediction, as it can handle complex datasets with many features and can provide accurate predictions with relatively low computational cost.



```
   predict_model
      0   1
  0 119  14
  1  41  26
```

Figure 5.1 confusion matrix.

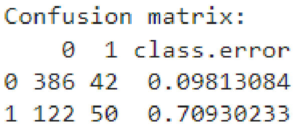


```
accuracy  macroPrecision  macroRecall  macroF1
   <dbl>           <dbl>        <dbl>    <dbl>
frame: 1 × 4
   0.725        0.696875    0.6413983  0.649134
```

Figure 5.2 precision and recall

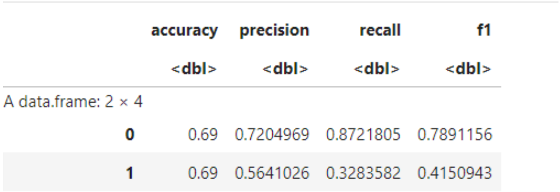
### *B. DECISION TREE CLASSIFIER*

A decision tree classifier is a machine learning algorithm that can be used for company bankruptcy prediction. It works by recursively dividing the data into smaller subsets based on the values of different features, until the subsets are pure or nearly pure in terms of their target variable. One advantage of decision tree classifiers is that they are easy to interpret, as the resulting tree can be visualized and used to identify the most important features for predicting physical quantity. They are very powerful algorithms, capable of fitting complex datasets However, decision trees can also be prone to overfitting if the tree is too deep or if there are too many features, so it is important to tune the hyperparameters carefully to avoid overfitting.



```
Confusion matrix:
    0  1 class.error
0 386 42  0.09813084
1 122 50  0.70930233
```

*Figure 5.2.1 Confusion Matrix*



| | accuracy | precision | recall | f1 |
|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> |
| A data.frame: 2 × 4 | | | | |
| **0** | 0.69 | 0.7204969 | 0.8721805 | 0.7891156 |
| **1** | 0.69 | 0.5641026 | 0.3283582 | 0.4150943 |

Figure 5.2.2 Precision Recall.

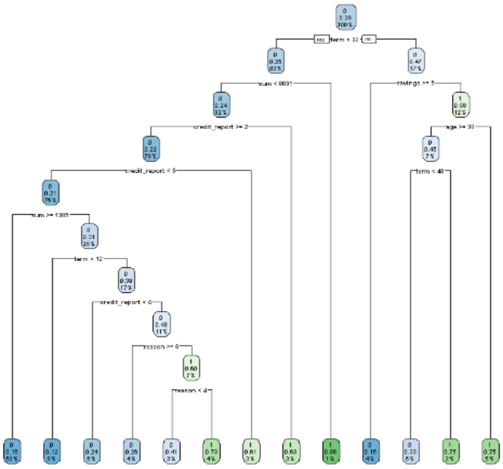


*Figure 3.1 Decision Tree*

## V. RESULT ANALYSIS

Accuracy is the percentage of correctly classified instances out of the total instances in the test set. Precision is the percentage of correctly classified positive instances out of all instances classified as positive. Recall is the percentage of correctly classified positive instances out of all actual positive instances. F1 score is the harmonic mean of precision and recall. Overall, analyzing the results of physical activity prediction using a decision tree classifier in R involves using performance metrics to evaluate the model's accuracy and effectiveness, identifying patterns and trends in the data that may be affecting the model's performance, and experimenting with different hyperparameters to optimize the model's performance.

## VI. CONCLUSION

In conclusion, our study provides insights into the use of machine learning techniques for bankruptcy prediction. The Decision tree classifier model achieved the highest accuracy of 72.5 % in our analysis. The current ratio, quick ratio, gross profit margin, and return on assets were identified as the most significant variables in predicting bankruptcy. Our findings can be useful for investors, creditors, and regulatory agencies in making informed decisions about the financial health of companies. Future research can explore the use of alternative data sources, such as social media and news articles, in bankruptcy prediction

## VII. REFERENCES

[1] Altman EI (1968) The Prediction of Corporate Bankruptcy: A Discriminant Analysis. J Finance 23:193–194.

[2] Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. Machine learning models and bankruptcy prediction.

[3] Brédart, Xavier. 2014. Bankruptcy Prediction model using Neural networks. Accounting and Finance Research 3: 124–28.