Using Machine Learning to Identify Factors Contributing to Firms' Bankruptcy: A Case Study of the Taiwanese Market

MSc Research Project
FinTech

Saumya Chaudhary
Student ID: x21175365

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Saumya Chaudhary
Student ID:	x21175365
Programme:	FinTech
Year:	2023
Module:	MSc Research Project
Supervisor:	Brian Byrne
Submission Due Date:	14/08/2023
Project Title:	Using Machine Learning to Identify Factors Contributing to Firms' Bankruptcy: A Case Study of the Taiwanese Market
Word Count:	5646
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Saumya Chaudhary
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Using Machine Learning to Identify Factors Contributing to Firms' Bankruptcy: A Case Study of the Taiwanese Market

Saumya Chaudhary
x21175365

Abstract

Making the right financing decisions requires effective bankruptcy prediction on the part of financial institutions. In general, the two most significant aspects determining the prediction performance are the input variables (or features), such as financial ratios, and prediction methodologies, such as statistical and machine learning approaches. Even though numerous relevant publications have suggested innovative prediction methods, only a few have examined the crucial financial ratios that influence bankruptcy prediction. This research examines various statistical and machine-learning approaches to identify the important financial factors that can contribute to a company's bankruptcy. Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector, Balanced Random Forest, and Easy Ensemble classifiers are trained to evaluate their efficacy on the real-world Taiwanese bankruptcy dataset. Sensitivity, specificity, type 1 and type 2 error rates, and receiver operating characteristics values are the metrics used to assess the models' predictability. Among the ensemble methods, Balanced Random Forest and Easy Ensemble outperformed the others, and eleven financial ratios were deemed important: ROA (C) and depreciation before interest, Degree of Financial Leverage, Borrowing dependency, Debt ratio, Non-industry income and expenditure/revenue, Equity to Liability ratio, Interest Coverage Ratio, Total income/expense ratio, Interest Expense Ratio, Net Value Per Share, and Total debt/Total net worth ratio.

1 Introduction

It is acknowledged that one of the key areas for research in the world of financial accounting is the forecast of bankruptcy. When a company experiences ongoing, significant losses or when its liabilities outweigh its assets, it experiences financial failure. Financial distress is known to have a variety of causes and symptoms, such as inadequate management, autocratic leaders, and challenges in conducting profitable business. The business community and society at large suffer significant losses when a company goes bankrupt. The economic disruption follows bankruptcy, impacting stakeholders, markets, and business continuity. In light of the potential for timely warnings to management, investors, employees, shareholders, and other interested parties who seek to minimize their losses, bankruptcy prediction is therefore crucially important to creditors, shareholders, and employees.

Due to its potential to revolutionize risk assessment, improve decision-making, and support financial stability, Financial Technology (FinTech) has a critical role to play in predicting firm bankruptcies. Leading-edge FinTech companies are pioneering the development of complex bankruptcy prediction models using advanced data analytics, machine learning, and artificial intelligence. These models can help investors, lenders, and regulators make educated decisions, allocate resources optimally, and preserve the integrity of financial markets by accurately projecting bankruptcy risks.

Financial ratios are essential for predicting bankruptcy because they offer quantitative measurements for evaluating the stability and health of a company's finances (Altman (1968)). These play a pivotal role in predicting bankruptcy due to their ability to provide a range of invaluable insights. These include serving as early warning indicators, enabling objective evaluations, facilitating comprehensive analyses, ensuring consistency over time, instilling confidence in investors and creditors, aiding regulatory compliance, and offering strategic insights for effective management decisions. These statistics make it possible for analysts, financiers, and lenders to assess a company's capacity to fulfill its financial commitments and avert insolvency. Stakeholders can spot early warning signals of financial hardship and take proactive steps to reduce potential risks by studying a mix of several financial statistics. Financial ratios provide insights into various aspects of a company's financial position, such as Liquidity Ratios, Profitability Ratios, Liquidity Ratios, Leverage Ratios, Solvency Ratios, Activity or Efficiency Ratios, Cash Flow Ratios, Coverage Ratios, Profitability Growth Ratios, Inventory Management Ratios, Activity Turnover Ratios, Working Capital Ratios, Liability-Related Ratios, and Miscellaneous Ratios.

The purpose of FinTech is to disrupt conventional financial practices; therefore, its participation in bankruptcy prediction is consistent with this objective. Advanced data analytics enable businesses to investigate a wide range of variables, producing forecasts that are more accurate than those made using traditional techniques. The studies (Van Gestel et al. (2003); Smiti and Soui (2020); Liang et al. (2016); Zhou (2013)) found that machine learning techniques exhibit higher predictive performance when integrated into bankruptcy prediction models, enabling the timely detection of enterprises in danger. Investors and creditors can manage risks more skillfully because of this precision.

This research explores the statistical method along with the machine learning approaches to predict bankruptcy in the Taiwanese dataset. The objectives are as follows:

- Examine current bankruptcy prediction studies to determine effective approaches.
- Implement statistical feature selection methods to be used with the models.
- Evaluate the models based on specificity, sensitivity, type-1 and type-2 error rate, and ROC value.
- Compare the resampling-based strategies with the algorithm-based approach to find the best contributing financial ratios for bankruptcy prediction.

2 Related Work

Predicting bankruptcy is a crucial component of evaluating financial risk for both firms and investors. In order to improve the precision and efficacy of bankruptcy prediction models, researchers have used machine-learning approaches over the years. In this section, a systematic literature review was conducted. A systematic review procedure helps

a) to encapsulate the existing documentation concerning a technology b) to identify any research gaps and suggest further investigation c) to provide a framework in order to appropriately establish novel research (Kitchenham (2004)). There are some characteristics of a systematic review:

- **Defining a review protocol** states the research question being addressed and the methods that will be implemented to accomplish the review. In this study, the protocol was to review studies from the Finance and Machine learning domain.

- **Search strategy** aims to identify as much of the relevant literature as possible. Here, the search strategy includes keywords such as Financial Distress OR Insolvency OR Bankruptcy AND Machine Learning OR Deep Learning, Bankruptcy AND Statistical Methods, and Financial Ratio AND Bankruptcy Prediction.

- **Inclusion and exclusion** are the criteria that assess each potential primary and secondary study. In this study, only papers published in English have been considered.

2.1 Early Approaches and Traditional Models

As a key factor in the economic growth and financial viability of a country for many stakeholders, bankruptcy prediction has remained in the research limelight. Traditional statistical methods including logistic regression, discriminant analysis, and ratio analysis were historically the main foundation for bankruptcy prediction models.

2.1.1 Beaver's Model

One of the first statistical models for bankruptcy prediction was Beaver's bankruptcy prediction model, created by Robert E. Beaver in 1966 (Beaver (1966)). It used financial ratios to construct a linear equation using discriminant analysis. Using a computed score, this equation might subsequently be used to divide businesses into solvent and insolvent categories.

2.1.2 Altman's Z-Score Model

Edward Altman pioneered the concept of bankruptcy prediction with the introduction of Altman's Z-Score model in 1968 (Altman (1968)). The likelihood of bankruptcy was predicted by this model using a linear combination of various financial ratios. The Z-Score divided businesses into zones according to their risk of insolvency. It was created for publicly traded manufacturing companies and had a notable track record of identifying impending financial trouble.

2.1.3 Ohlson's O-Score Model

James A. Ohlson introduced a probabilistic method for bankruptcy prediction with his 1980 proposal of Ohlson's O-Score model (Ohlson (1980)). To determine the likelihood of bankruptcy, this model used a logit structure and financial ratios. It stressed the distinction between bankrupt and solvent organizations using a determined O-Score.

2.1.4 Taffler's Model

In 1983, Richard J. Taffler developed Taffler's model, which used logistic regression and added extra financial variables to the conventional strategy (Taffler (1983)). With the help of this model, bankruptcy prediction accuracy was intended to be increased by taking more complex correlations between financial variables into consideration.

The foundation for later research in this area was laid by early bankruptcy prediction models. The intricate linkages and non-linear patterns found in financial data presented challenges for these classic models, despite the fact that they offered insightful analysis. Researchers were able to get around these restrictions and create more precise and reliable prediction models due to the development of machine learning techniques.

2.2 Importance of Financial Ratios

Financial ratios are invaluable tools in assessing a company's financial well-being and predicting the potential risk of bankruptcy. These ratios offer a quantifiable means of evaluating a company's financial health and its susceptibility to insolvency, making them indispensable instruments for bankruptcy prediction. By scrutinizing a variety of measures that gauge liquidity, solvency, profitability, and efficiency, analysts gain a deeper understanding of the company's capacity to meet its financial obligations and navigate through economic challenges. For company management, financial ratios provide actionable insights into various aspects of operations. By identifying areas where ratios fall below industry benchmarks or historical trends, management can formulate strategies to improve performance and address potential weaknesses (Taffler (1983); Altman (1968); Beaver (1966)).

In essence, financial ratios transcend numerical calculations; they provide a language through which analysts, investors, creditors, and company management can collectively comprehend a company's financial landscape. By translating complex financial data into meaningful insights, ratios enable proactive decision-making, risk mitigation, and strategic planning, all of which are essential components in predicting and preventing bankruptcy.

2.3 Machine Learning Approaches

With more academics contributing to the literature, machine learning (ML) for bankruptcy prediction is becoming increasingly popular. Researchers have focused on both parametric, such as Multivariate Discriminant Analysis (MDA) and Logistic Regression (LR), and non-parametric, such as Neural Networks (NN), Bagging, Boosting, Ensemble, and Support Vector Machines (SVM), models. A typical ML task includes various steps such as data preparation, modeling, evaluation, and deployment. In past studies, researchers have used various statistical methods to prepare the data.

2.3.1 Feature Selection and Engineering

When the feature set under consideration is big, the methods based on the filter approach are typically computationally effective and statistically scalable (Blum and Langley (1997); Guyon and Elisseeff (2003)). Methods such as T-test (Gudmund et al. (1976)), discriminant analysis (Chen and Hsiao (2008); Klecka (1980)), and Pearson's correlation (Guyon and Elisseeff (2003)).

Wrapper approaches include using a particular machine learning algorithm's performance to direct the feature selection procedure (Kohavi and John (1997)). Cross-validation is frequently used in this method to evaluate how feature subsets affect model performance. This strategy also considers the problem of feature dependence. As a result, this approach's methods typically include a process of searching for a useful feature subset, necessitating a significant amount of computing (Liu and Yu (2005)). If exhaustive search is used, the search space becomes rapidly intractable as the total number of the feature set grows (Ngai et al. (2011)) The wrapper approach's propensity for overfitting issues is another well-known flaw.

Embedded feature selection techniques integrate feature selection into the model training procedure. These techniques make use of specific algorithms that automatically weigh the significance of features as they are being built into the model. When predicting bankruptcy, embedded approaches have the advantage of simultaneously developing the predictive model and choosing pertinent features, which could increase performance and interpretability (Valencia et al. (2019)).

Another method to select features is T-test based method that determines whether there is a statistically significant difference between the means of two groups. It aids in the identification of features that demonstrate notable variations between two classes (for example, between bankrupt and non-bankrupt enterprises) in the context of feature selection (Liang et al. (2016)).

2.3.2 Modeling

In order to predict bankruptcy one year in advance, (Barboza et al. (2017)) employed support vector machines, bagging, boosting, and random forests on data from American firms that were acquired for 28 years. They then compared their performance to that of results from neural networks, logistic regression, and discriminant analysis. (Wang et al. (2017)) implemented SVM, neural network, and autoencoder on the Qualitative Bankruptcy dataset by including industrial risk, management risk, financial flexibility, credibility, competitiveness, and operating risk as attributes.

2.3.3 Addressing Imbalanced Nature of Data

To address the class skewness, (Zhou (2013)) implemented resampling strategies such as Oversampling and Under-sampling with the USA bankruptcy dataset. Synthetic Minority Oversampling Technique and Random Oversampling increase the number of records in the minority class while Random Under-sampling reduces the number of data points from the majority class. In addition to Oversampling and Under-sampling strategies, (Veganzones and Séverin (2018)) incorporated the Easy Ensemble method to address the class imbalance with the income statement of French firms. (Smiti and Soui (2020)) implemented a deep learning-based architecture and named it BSM-SAE (Borderline Synthetic Minority oversampling technique (BSM) and Stacked AutoEncoder (SAE)) to predict bankruptcy on the Polish dataset.

2.3.4 Evaluation

(Zhou (2013)) used sensitivity, specificity, accuracy, F measure, and area under the curve value to asses the model. (Veganzones and Séverin (2018); Le (2022)) used the G-mean score along with AUC to assess the performance of the models. (Liang et al. (2016)) used type-1 and type-2 errors, average accuracy, and ROC value to show the results from the machine learning models.

Summary: In the realm of bankruptcy prediction, researchers have harnessed an array of techniques across datasets encompassing various countries. While the investigation has extended to the Taiwanese dataset, untapped potential lies in the exploration of statistical methodologies, coupled with an examination of the effectiveness of a comprehensive feature set. Addressing the challenge of imbalanced classification, a prominent concern in bankruptcy prediction, has predominantly involved resampling, such as over-sampling or under-sampling, in line with established literature. Nonetheless, it's crucial to acknowledge the noteworthy impact of algorithmic-level interventions, particularly the integration of ensemble methods, which can markedly uplift predictive performance. To the best of the knowledge of the author, very few studies have made an attempt to identify the top contributing financial ratios to a company's bankruptcy. This study aims to find the main financial ratios that can make an impact on bankruptcy.

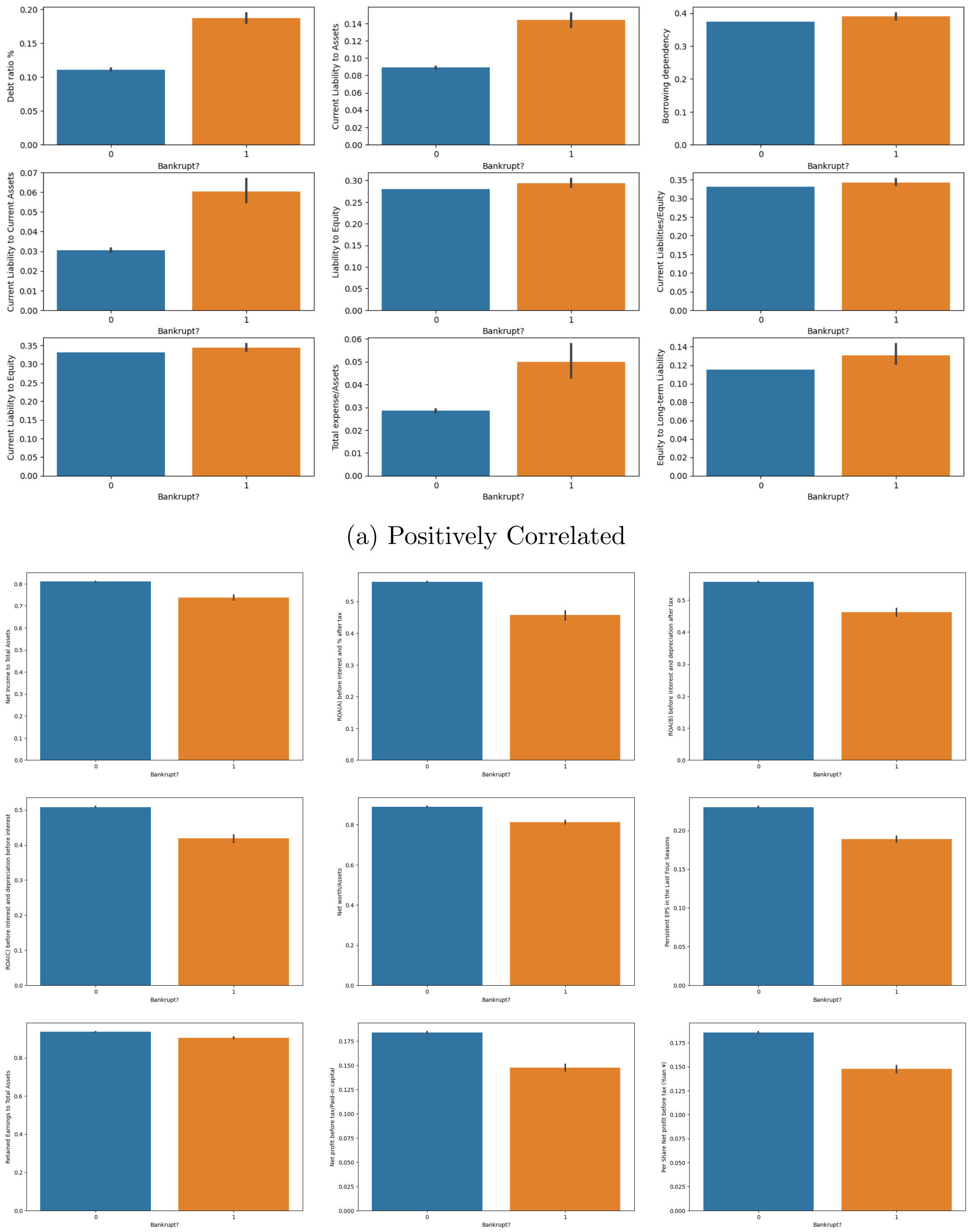
3 Methodology

This section comprehensively delineates the methodologies utilized to conduct the study. The various steps are data collection and exploration, data preparation, feature selection, and modeling.

3.1 Data Collection and Exploration

(Liang et al. (2016)) suggested the pedagogical research for this study. In order to evaluate the effectiveness of the proposed approach, the Taiwanese firms' bankruptcy dataset (*Taiwanese Bankruptcy Prediction* (2020)) is employed. Data from the Taiwan Economic Journal for the years 1999 to 2009 were gathered for this investigation. The definition of corporate bankruptcy was established using the Taiwan Stock Exchange's business rules [1]. Additionally, two criteria were applied for gathering the data samples. First, the sample companies had to have three years' worth of full public disclosure prior to the financial crisis. In order to compare the bankrupt and non-bankrupt situations, there should be a significant number of comparable businesses in the same industry that are similar in size. The resulting sample comprises businesses from the manufacturing sector, which consists of industrial and electronics businesses (346 businesses), the service sector, which consists of businesses in the travel, retail, and shipping sectors (39 businesses), and other sectors (93 businesses), but not financial businesses. The target variable is Bankruptcy which is categorical and has two classes, 0 meaning non-bankrupt events and 1 meaning bankrupt events. Evidently, the highest correlated variables positively align with label 1, representing bankrupt events, while the most negatively correlated variables correspondingly exhibit peaks for label 0, signifying non-bankrupt events.

[1] `https://twse-regulation.twse.com.tw/ENG/EN/law/DOC01.aspx?FLCODE=FL007304&FLNO=49+++`



(a) Positively Correlated

(b) Negatively Correlated

Figure 1: Top 9 Positively and Correlated Variables

3.2 Data Preparation

The output of the results depends on the quality of the data. Therefore, meticulous data preparation is required prior to modeling. In the initial phase, a comprehensive examination of each column was undertaken to identify any instances of missing values, revealing that none were present. Subsequently, the dataset was stratified into dependent and independent variables, a crucial step that paved the way for the subsequent division of the data into distinct training and testing subsets.

The continuous features were normalized using min-max scaling and standard scaling strategies, and both preprocessed data were used in the modeling phase. The min-max normalization is given in Equation 1.

$$X_{min-max} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where, X_{min} and X_{max} are the minimum and maximum values of each feature, respectively, and X is an element of the column.

Equation 2 presents the mathematical formulation of standard scaler normalization.

$$X_{std-scl} = \frac{X - X_{avg}}{StDev} \tag{2}$$

where, X_{avg} and $StDev$ are the average and standard deviation of each feature, respectively, and X is an element of the column.

3.3 Feature Selection

The dataset under consideration encompasses a substantial 95 features, a reality that exacerbates the dimensionality of the dataset. Such an increase can lead to the emergence of the well-known "Curse of Dimensionality," giving rise to a constellation of challenges encompassing computational intricacies, vulnerability to overfitting, feature redundancies, and augmented model complexities (Verleysen and François (2005)). To tackle this multifaceted issue, this study leveraged Pearson's Correlation method (Yeh and Chen (2020); Brownlee (2019); Paraschiv et al. (2021); Yang et al. (2011)). This technique, known for its effectiveness, has been used to identify important features and relieve the pressure caused by dimensionality.

Moreover, to further address dimensionality, Principal Component Analysis (PCA) was judiciously applied (Abdi and Williams (2010)). PCA, a prominent dimensionality reduction technique, was deployed to reconfigure the data's structure and trim its dimensionality.

3.4 Modeling

Multiple machine learning (ML) algorithms and statistical methods were implemented to predict bankruptcy.

3.4.1 Logistic Regression

When doing binary classification tasks, the statistical technique of logistic regression is utilized to estimate the likelihood that a given instance would belong to a certain class (Hosmer Jr et al. (2013); Wright (1995); Hauser and Booth (2011); Kim and Gu

(2006)). The logistic function, which converts the linear combination of input features into a probability value between 0 and 1, is used to model the link between one or more independent factors and a binary outcome. It establishes a decision boundary dividing the classes in the feature space by fitting the logistic regression model to the training data and estimating the coefficients. Due to the model's ease of use, readability, and capacity for making probabilistic predictions, it is frequently used in many different sectors.

3.4.2 Decision Tree Classifier

An approach to machine learning that is used for both classification and regression applications is the decision tree classifier. It builds a hierarchical tree-like structure with internal nodes that each represent a decision made in response to a particular aspect and branches that indicate potential outcomes. By choosing the feature at each node that optimizes information gain or Gini impurity reduction, the algorithm divides the data into subsets in a recursive manner. As a result, a series of binary decisions that lead to leaf nodes that represent class labels or regression values are produced. Because they are simple to understand, straightforward to visualize, and capable of capturing complicated decision boundaries, decision trees are widely used. However, they can be susceptible to overfitting and lack robustness in noisy data (Olson et al. (2012); Syed Nor et al. (2019); Safavian and Landgrebe (1991)).

3.4.3 Support Vector Classifier

For binary classification tasks, a potent machine learning technique called the Support Vector Classifier (SVC) is frequently utilized. By maximizing the margin between the two classes, it seeks to identify the ideal hyperplane for separating them. The margin shows how far each class's closest data points are from the hyperplane. The technique builds the decision boundary based on the critical instances, or support vectors, which are the data points that are closest to the hyperplane. The SVC can use kernel functions to translate the data into a higher-dimensional space, enabling non-linear decision boundaries, in situations when linear separation is not practical (Shin et al. (2005); Horak et al. (2020); Min and Lee (2005); Brereton and Lloyd (2010)).

3.4.4 Naive Bayes Classifier

For classification problems, the Naive Bayes classifier is a common probabilistic machine learning technique. By multiplying the probabilities of each of an instance's features under the assumption that it belongs to a specific class, it determines the likelihood that it does. In spite of its "naive" premise, Naive Bayes frequently outperforms expectations in the real world, especially when working with huge feature sets. It has a high processing efficiency, is particularly suited for spam filtering and text classification, and can handle large-scale data (Sun and Shenoy (2007); Aghaie and Saeedi (2009); Rish et al. (2001)).

3.4.5 Ensemble Methods

- Random Forest Classifier: The Random Forest Classifier is an ensemble learning method used in machine learning for classification applications. To build a more precise and reliable model, it mixes various distinct decision trees. The risk of overfitting is decreased because each tree is built using a random portion of the data and a random selection of characteristics. A majority vote or an average of the forecasts from each individual tree is used to determine the final categorization. It is well known that Random Forests can handle high-dimensional data, capture complicated relationships, and offer insights into the importance of particular features. In comparison to a single decision tree, they improve generalization, reduce variation, and improve predictive accuracy (Joshi et al. (2018); Kim et al. (2022)).

- Balanced Random Forest Classifier: In imbalanced classification problems, the class imbalance is addressed by the Balanced Random Forest Classifier, a method provided by the imbalanced-learn library[2]. The balanced sampling method used during tree construction, whereby the majority class is randomly under-sampled at each split and the minority class is randomly over-sampled, extends the capabilities of the Random Forest algorithm. By reducing the difficulties given by unbalanced datasets, this approach improves the model's capacity to correctly categorize both groups. The generalization and robustness of this classifier are enhanced, and it performs particularly well in situations when the class distribution is skewed (Chen et al. (2004)).

- Easy Ensemble Classifier: Class imbalance in classification tasks is addressed by the Easy Ensemble Classifier provided by the imbalanced-learn library[3]. To deal with the skewed class distribution, a combination of under-sampling and ensemble approaches is used. Using random undersampling, the algorithm first divides the majority class into many subsets, and then it trains a different classifier on each subset. These individual classifiers cast their votes, and the one with the most votes makes the final forecast. Performance is improved as a result of this strategy's emphasis on the minority class and the reduction of the effects of class inequality (Liu et al. (2008)).

4 Design Specification

This section outlines the specific requirements, criteria, and parameters for the design and implementation of this study.

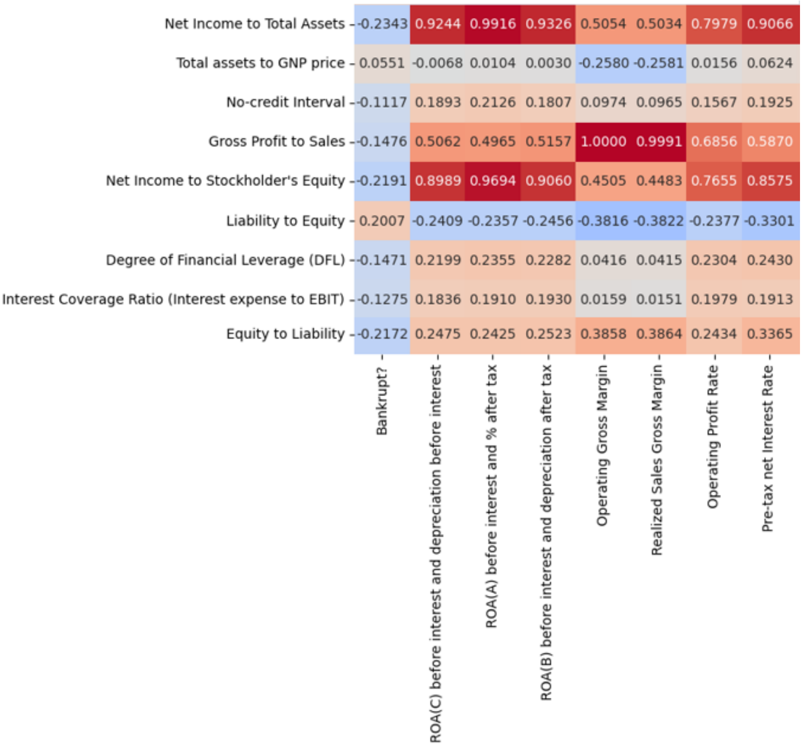
4.1 Feature Selection Technique

Multicollinearity is a statistical phenomenon when two or more independent variables have a high degree of correlation, which causes instability and makes it difficult to interpret the coefficients of the model (Alin (2010); Belsley et al. (2005)). It is challenging to separate the individual impacts of correlated factors on the dependent variable when multicollinearity is prevalent. This may produce false results and less accurate parameter estimates. To address this challenge, the variable exclusion method was proposed based on Pearson's Correlation coefficient. A threshold of $|0.7|$ was chosen, which means if two variables have more than 70% correlation, either will be selected (Dormann et al. (2013)).

[2] https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html

[3] https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.EasyEnsembleClassifier.html

Figure 2 shows some of the features' heatmap. A heatmap of correlation matrix for all the features can be found in the Notebook [4] under the Feature Exploration section.



	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate
Net Income to Total Assets	-0.2343	0.9244	0.9916	0.9326	0.5054	0.5034	0.7979	0.9066
Total assets to GNP price	0.0551	-0.0068	0.0104	0.0030	-0.2580	-0.2581	0.0156	0.0624
No-credit Interval	-0.1117	0.1893	0.2126	0.1807	0.0974	0.0965	0.1567	0.1925
Gross Profit to Sales	-0.1476	0.5062	0.4965	0.5157	1.0000	0.9991	0.6856	0.5870
Net Income to Stockholder's Equity	-0.2191	0.8989	0.9694	0.9060	0.4505	0.4483	0.7655	0.8575
Liability to Equity	0.2007	-0.2409	-0.2357	-0.2456	-0.3816	-0.3822	-0.2377	-0.3301
Degree of Financial Leverage (DFL)	-0.1471	0.2199	0.2355	0.2282	0.0416	0.0415	0.2304	0.2430
Interest Coverage Ratio (Interest expense to EBIT)	-0.1275	0.1836	0.1910	0.1930	0.0159	0.0151	0.1979	0.1913
Equity to Liability	-0.2172	0.2475	0.2425	0.2523	0.3858	0.3864	0.2434	0.3365

Figure 2: Correlation Between a Few Variables

4.2 Dimensionality Reduction

Principal Component Analysis (PCA) is a widely used technique for reducing the dimensionality of high-dimensional data while preserving as much variance as possible (Abdi and Williams (2010)). It entails converting the initial variables into a new collection of uncorrelated variables (principal components) that account for the majority of the variability in the data. In this study, the top 8 principal components accounted for 99% of the variation in the dataset.

4.3 Resampling

As the data is biased towards non-bankrupt events, the resampling strategies including Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. (2002); Smiti and Soui (2020)), Random Oversampling (ROS) (Mohammed et al. (2020)), and Random Undersampling (RUS) (Wang and Liu (2021); Liu et al. (2008)) have been implemented.

[4] https://colab.research.google.com/drive/1h3soHoASLr37PEKnUWYd0S7F1J7vUd36?usp=sharing

5 Implementation

The studies were conducted in a Google Colab environment (Bisong and Bisong (2019)) utilizing a Jupyter Notebook (Kluyver et al. (2016)) and the Python programming language(Sanner et al. (1999)). Table 1 shows the various tools used with their purpose.

Sl. No.	Tools and Technologies	Purpose
1	Microsoft Excel	*For the purpose of storing and retrieving Taiwanese bankruptcy data in CSV format*
2	Google Colaboratory	*Free, cloud-based Jupyter notebook for collaborative Python programming and experimentation*
3	Jupyter Notebook	*To implement Python Codes*
4	Python	*For carrying out a high-level coding task for machine learning*

Table 1: Various Tools and Technologies, and Purpose

5.1 Data Handling

The dataset was downloaded as CSV format which had 96 variables and 6819 instances. Pandas and NumPy are two Python modules that made it possible to read data into data frames and then use it for exploratory data analysis. The variable names needed preprocessing as there was a white space at the beginning which was removed. The dataset comprises 95 features out of which two are categorical, namely, 'Liability-Assets Flag' which has two categories, 1 if total liability exceeds total assets, 0 otherwise, and

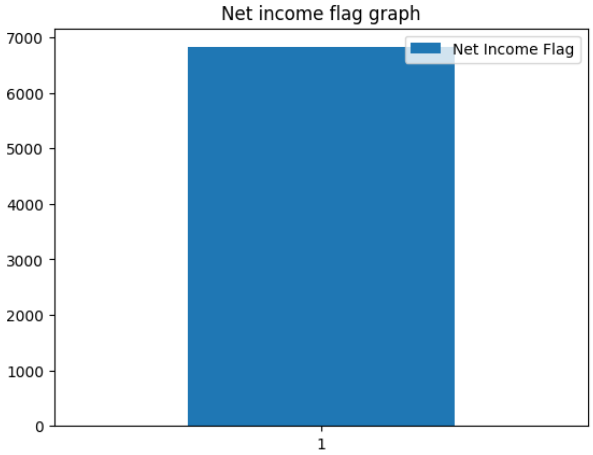
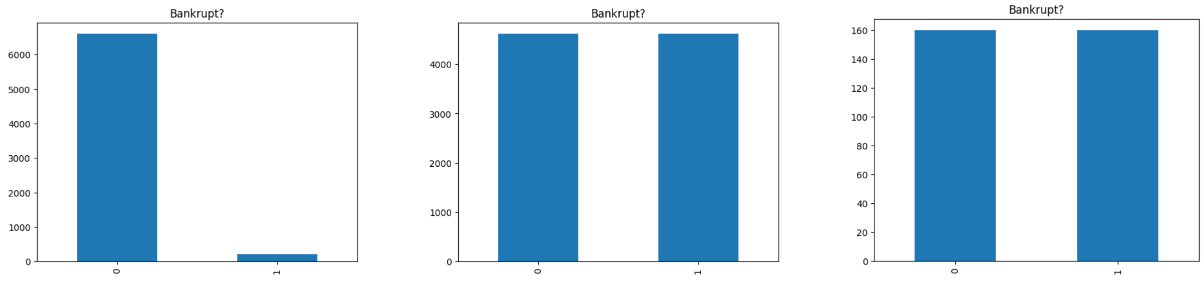


Figure 3: Class Distribution of Net Income Flag



(a) Original Frequency (b) Oversampled Frequency (c) Undersampled Frequency

Figure 4: Class Frequencies for Original, Oversampled, and Undersampled Datasets

'Net Income Flag' which is 1 if the net income is negative for the last two years, 0 otherwise. The rest of other variables are continuous. Figure 3 shows the distribution of classes in 'Net Income Flag' where all the data points belong to class 1. Hence, this column has been dropped from the analysis.

The data was scaled using the min-max scaler and the standard scaler. The min-max scaler rescaled the data points in the range of 0 and 1 while the standard scaler rescaled them between -3 to 3. To match the class frequency of the target variable, the Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling (ROS), and Random Undersampling (RUS) were implemented where SMOTE and ROS upsampled the minority class, i.e., bankrupt events and RUS downsampled the majority class, i.e., non-bankrupt events. Figure 4 shows the bar plot for each type of class distribution.

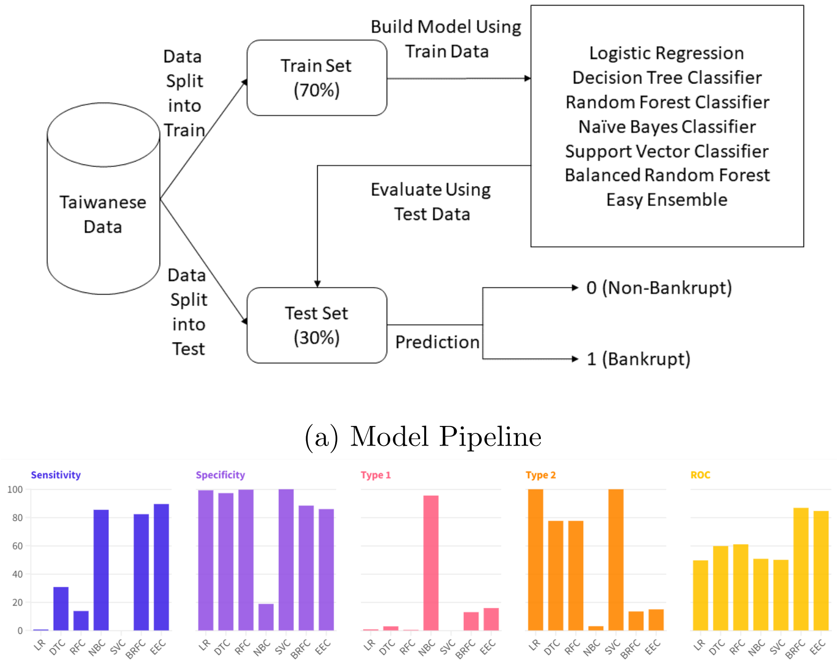
The feature selection was performed using Pearson's correlation coefficient with a threshold of $|0.7|$ and 63 out of 95 features were remaining. The dataset was split into training

5.2 Classification Models

The Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Naive Bayes Classifier (NBC), Support Vector Classifier (SVC), Balanced Random Forest Classifier (BRFC), and Easy Ensemble Classifier (EEC) were implemented on the original data, min-max and standard scaled data, principal components, and oversampled and undersampled data.

6 Evaluation

The bankruptcy prediction classification task focuses on classes, which influenced the choice of metrics capable of effectively assessing models while considering their class imbalance. Relying solely on accuracy metrics could be misleading, as they prioritize the class with more instances, often favoring the negative class in binary classification scenarios. Metrics such as Area Under the Receiver Operating Characteristic Curve (AUC-ROC), specificity and sensitivity (also known as true negative rate and true positive rate, respectively), and Type 1 error rate (false positive rate) and Type 2 error rate (false negative rate) are crucial for assessing the performance of models (He and Ma (2013); Fawcett (2006); Liang et al. (2016)). Equations 3, 4, 5, and 6 explain the formula for



(a) Model Pipeline

(b) Results

Figure 5: Model Pipeline and Results with Original Dataset

calculating sensitivity, specificity, type 1 error, and type 2 error respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

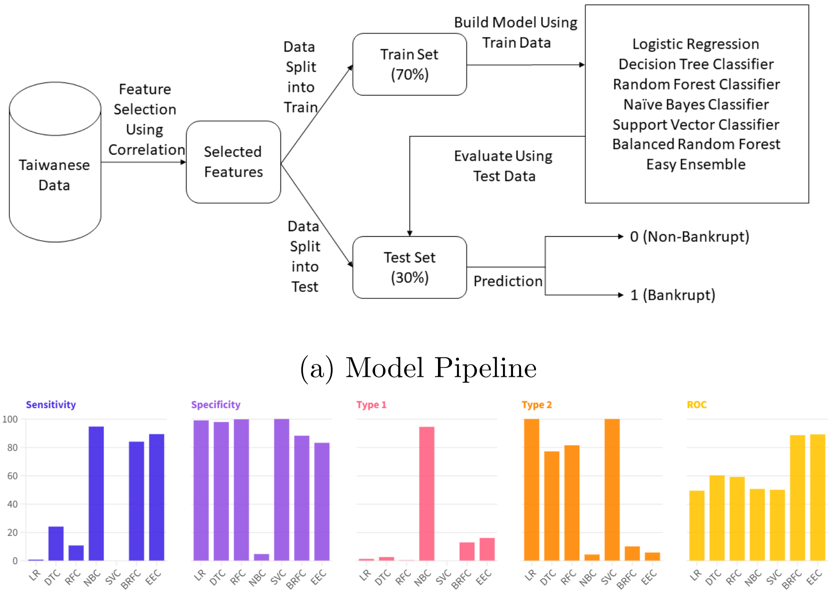
$$Type1 = \frac{FN}{FN + TP} \tag{5}$$

$$Type2 = \frac{FP}{FP + TN} \tag{6}$$

where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative respectively.

6.1 Experiment 1: Modeling with original dataset

In the context of this research study, all the models were put into practice utilizing the initial dataset. The process flow is illustrated in detail in Figure 5a, providing a visual representation of the sequence of steps followed. The outcomes of this experimental approach are visually captured in Figure 5b, effectively showcasing the results derived from the utilization of the unaltered dataset. This methodology underscores the foundation upon which the study's analyses and conclusions are based, thereby establishing a clear and transparent framework for subsequent evaluation and discussion.



(a) Model Pipeline

(b) Results

Figure 6: Model Pipeline and Results with Selected Features

6.2 Experiment 2: Modeling with selected features

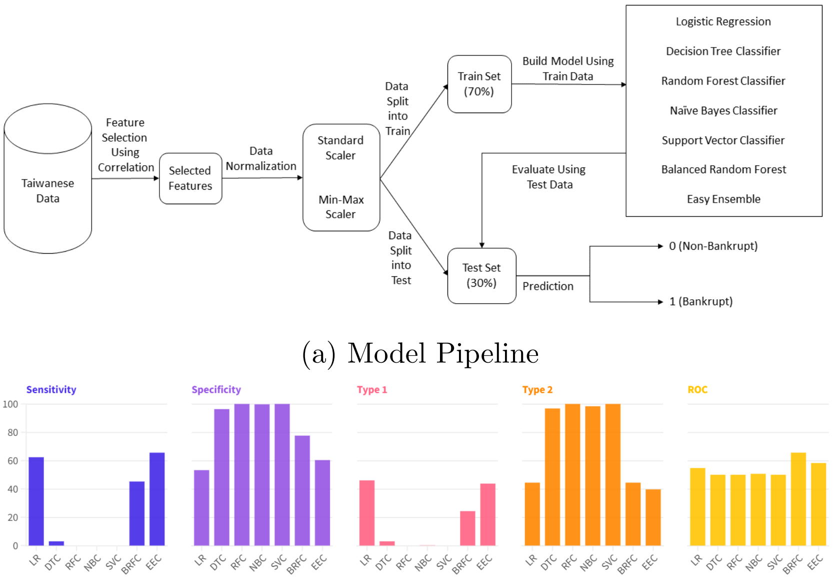
In this experiment, all of the models were instantiated using the specifically chosen features, a selection made through the utilization of Pearson's correlation coefficient. The process followed a well-defined pipeline, the visual representation of which can be observed in Figure 6a. The outcomes and findings stemming from this experimental approach are visually depicted and elaborated upon in Figure 6b. These representations collectively provide a comprehensive overview of the methodology and results of the study.

6.3 Experiment 3: Modeling with principal components

In the conducted study, the utilization of principal components was a central aspect in the implementation of all models under investigation. A visual representation of the procedural flow can be observed in the schematic presented in Figure 7a. To gauge the outcomes of this experimental approach, the findings are effectively showcased through the graphical depiction provided in Figure 7b. This comprehensive approach underscores the significance of principal components in the models and their subsequent evaluation, as evidenced by the amalgamation of the pipeline illustration and the results visualization.

6.4 Experiment 4: Modeling with normalized dataset

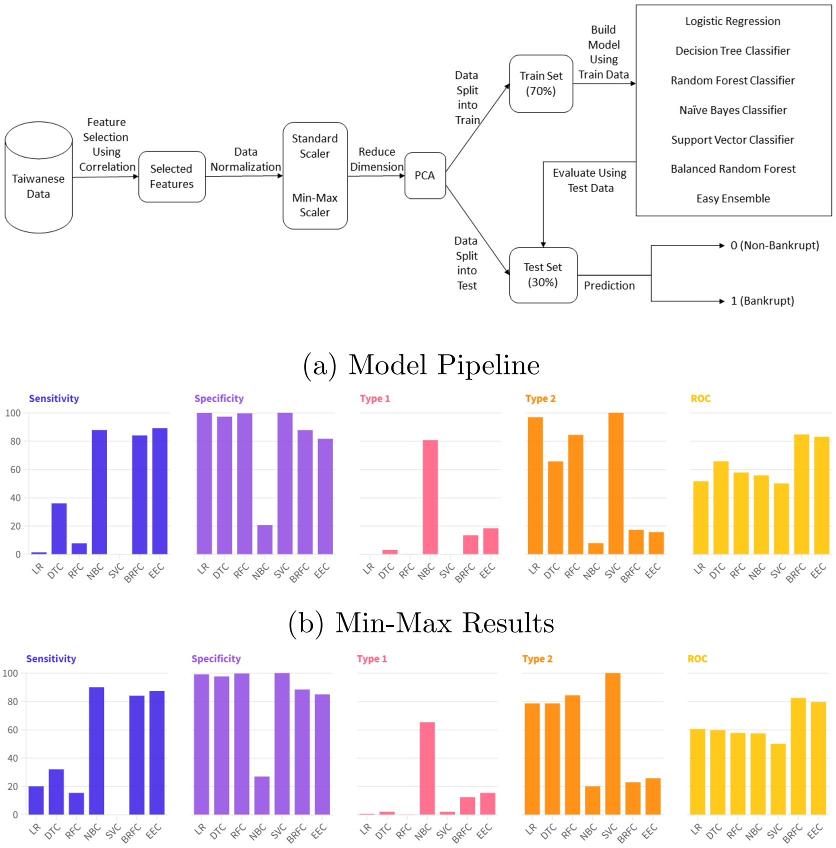
In the context of this study, it's important to highlight that all the models were developed and evaluated using normalized datasets, which enhances the comparability and fairness of the results. The process flow can be better understood by referring to the visual representation provided in Figure 8a, where a clear pipeline is depicted. For further insight into the outcomes, it's worth examining the results obtained from both min-max scaled and standard scaled datasets, presented in Figures 8b and 8c respectively. These figures shed light on the performance and behavior of the models under different scaling approaches, contributing to a comprehensive understanding of the study's findings.



(a) Model Pipeline

(b) Results

Figure 7: Model Pipeline and Results with Principal Components

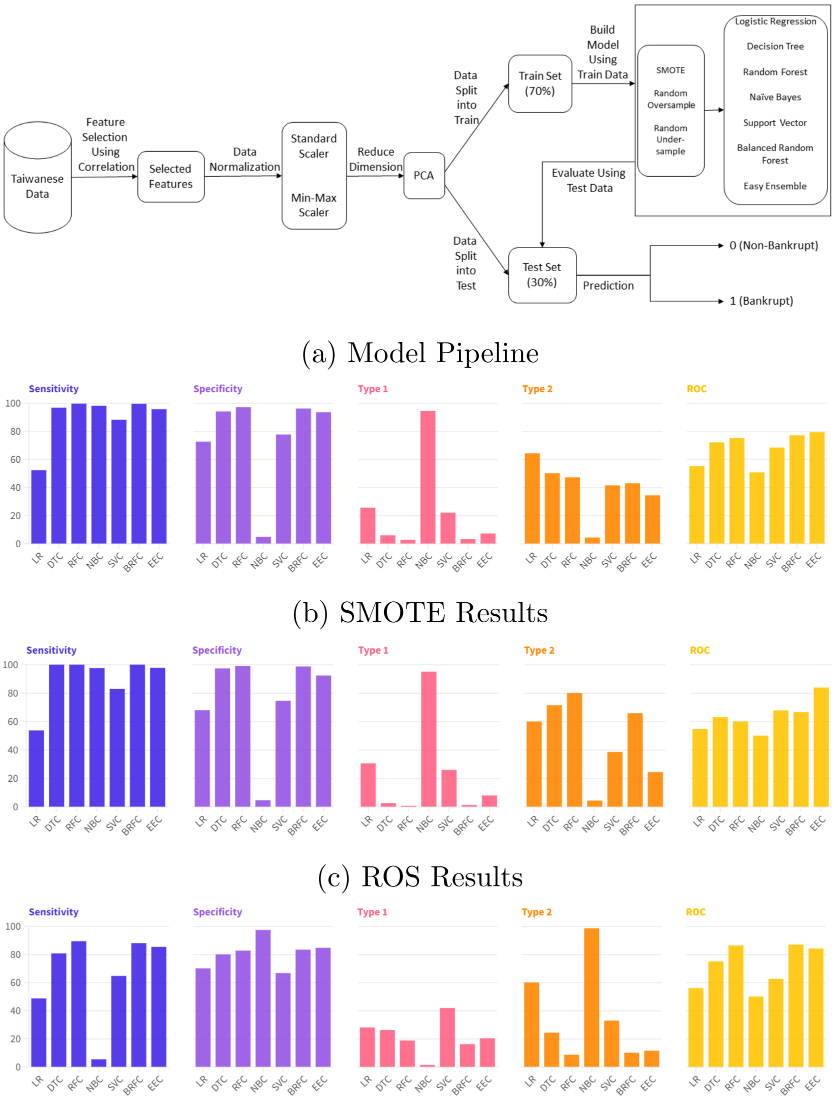


(a) Model Pipeline

(b) Min-Max Results

(c) Standard Scaled Results

Figure 8: Model Pipeline and Results with Normalized Datasets



(a) Model Pipeline

(b) SMOTE Results

(c) ROS Results

(d) RUS Results

Figure 9: Model Pipeline and Results with Resampled Datasets

6.5 Experiment 5: Modeling with resampled dataset

In the course of this experiment, all of the models were executed using the resampled datasets, which were specifically prepared for this purpose. The process and flow of operations are visually presented in the schematic depicted in Figure 9a. To comprehensively demonstrate the outcomes and effectiveness of various resampling techniques, we have included graphical representations of the results obtained from Synthetic Minority Oversampling Technique (SMOTE) in Figure 9b, outcomes stemming from the application of Random Oversampled (ROS) datasets in Figure 9c, and findings derived from utilizing Random Undersampled (RUS) datasets in Figure 9d. These visual representations serve to provide a clear understanding of the impacts and implications of each resampling approach.

6.6 Discussion

In this comprehensive study, an array of diverse machine learning (ML) techniques has been seamlessly integrated to formulate precise predictions regarding bankruptcy utilizing a dataset sourced from Taiwan. The ensuing outcomes have been meticulously juxtaposed with the seminal findings presented by Liang et al. in their notable work (Liang et al. (2016)). Should a model exhibit superior performance in contrast to the methodologies expounded upon in the aforementioned primary study, such advancements are conscientiously documented and highlighted. This intricate research endeavor not only seeks to enhance our comprehension of bankruptcy prediction but also serves as a testament to the evolving prowess of ML techniques in tackling intricate financial scenarios. By synergizing an expansive spectrum of ML approaches and leveraging a uniquely sourced dataset, this study contributes to the ever-growing body of knowledge, potentially paving the way for more robust and accurate predictive models in the realm of financial analysis.

6.6.1 Model Comparison

In the analysis presented, Figure 10 serves as a visual representation of the outcomes derived from three distinct models: the Random Forest Classifier (RFC), the Balanced RFC (BRFC), and the Easy Ensemble Classifier (EEC). Delving into the results, it becomes evident that the efficacy of these models varies across different datasets and scenarios. Upon scrutiny of the results, it is apparent that the Random Forest Classifier (RFC) exhibits noteworthy performance only when confronted with the RUS dataset. However, this efficacy diminishes when applied to the remaining datasets, as it fails to yield any statistically significant outcomes. This underscores the model's sensitivity to specific data configurations and its limitations when generalized to diverse datasets.

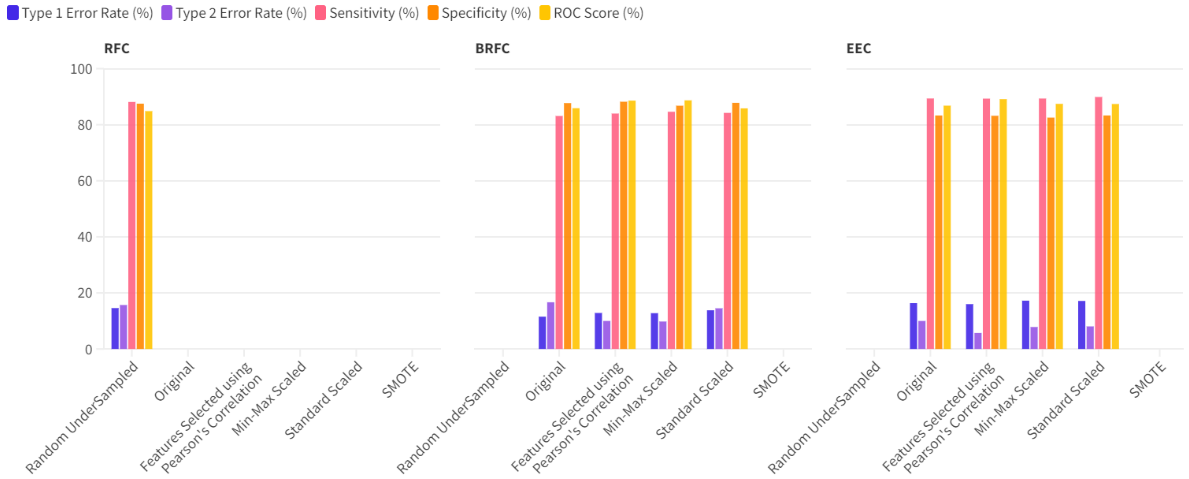


Figure 10: Results of the Reported Models

In contrast, both the Balanced RFC (BRFC) and the Easy Ensemble Classifier (EEC) showcase commendable performances across a broader spectrum of scenarios. These models emerge as robust contenders when evaluated against the original dataset, features selected through Pearson's Correlation, as well as datasets subjected to min-max scaling and standard scaling. The consistent effectiveness of BRFC and EEC across these diverse settings highlights their adaptability and underscores their potential as reliable tools for classification tasks.

It is noteworthy that the oversampled and SMOTEd datasets, which were expected to enhance the model's performance by addressing the class imbalance, did not yield the desired outcomes. Despite the conventional assumption that oversampling would contribute to improved results, the findings indicate that these datasets failed to surpass the performance of other configurations. This suggests that oversampling might not always be a panacea for class imbalance and calls for a nuanced consideration of its implementation in different contexts.

6.6.2 Important Financial Ratios

After finding the top contributing features from each of the models mentioned in 6.6.1, it is apparent that the following financial ratios play a critical role in predicting a company's likelihood of bankruptcy.

- ROA (C) before interest and depreciation before interest: Return on Assets (ROA) measures how well a company extracts profits from its assets. A low ROA could be a symptom of operational inefficiencies and decreased profitability, which could be a sign of potential bankruptcy risk.

- Degree of Financial Leverage (DFL): DFL quantifies how sensitive a company's earnings are to variations in operating income. High levels of leverage can compound financial risk, making the business more susceptible to downturns in the economy and raising the possibility of bankruptcy.

- Borrowing dependency: This refers to how heavily a business depends on borrowed money. When a firm borrows much, interest costs and financial obligations rise, which could have an adverse effect on cash flow and even result in bankruptcy if the company is unable to make its debt payments.

- Debt ratio%: A company's entire debt is compared to all of its assets in a debt ratio. In particular, if the company's cash flows are insufficient to satisfy debt payments, a high debt ratio might increase the risk of bankruptcy because it shows a heavy reliance on debt financing.

- Non-industry income and expenditure/revenue: A high percentage of non-industry-related income and expenses may signify erratic revenue flows, which could impact the company's capacity to pay its debts and raise the danger of bankruptcy.

- Equity to Liability: When contrasted to liabilities, this ratio shows what percentage of a company's assets are financed by shareholders' equity. Low equity to liabilities means excessive financial leverage and a smaller financial safety net, increasing the risk of insolvency for the company.

- Interest Coverage Ratio (Interest expense to EBIT): This ratio evaluates a company's capacity to use its profits before interest and taxes (EBIT) to cover its interest costs. A low-interest coverage ratio suggests that the business may find it difficult to pay interest payments, increasing the risk of bankruptcy.

- Total income/Total expense: This ratio provides insights into a company's overall profitability. A declining ratio could indicate weakening financial performance, making it harder for the company to manage its expenses and avoid bankruptcy.
- Interest Expense Ratio: This ratio evaluates a business's interest costs in relation to its operating revenue. High-interest costs relative to income might make it difficult to maintain a stable financial situation and can increase the risk of bankruptcy.
- Net Value Per Share (B): The company's net worth per outstanding share is represented by its net value per share. A declining net worth per share could be an indication of worsening financial condition and a larger chance of bankruptcy.
- Total debt/Total net worth: This ratio measures how much debt a corporation has in relation to its net worth. A high ratio suggests significant financial leverage, which could increase the danger of bankruptcy during recessions.

7 Conclusion and Future Work

This research centers on investigating the discriminative capability achieved by merging various sets of financial ratios (FRs) for the purpose of predicting bankruptcy. Specifically, the study evaluates thirteen distinct categories of FRs, encompassing liquidity, profitability, leverage, operational efficiency, cash flow, solvency, coverage, growth in profitability, management of inventory, turnover of operations, liabilities-related measures, and miscellaneous aspects.

The study also aims to discern the pivotal financial ratios contributing to a firm's susceptibility to bankruptcy. In pursuit of these factors, an array of machine learning (ML) experiments were meticulously executed and assessed through metrics including specificity, sensitivity, type 1 and type 2 error rates, as well as the receiver operating characteristic (ROC) value.

Despite its widespread use, the chosen statistical approach, namely Logistic Regression, yielded insignificant outcomes. This could be attributed to its fundamental assumption of a linear relationship between the target variable and predictors, which might not hold true in this context. Furthermore, the intricate interactions between financial ratios, challenging for logistic regression to capture, could be another reason behind its lackluster performance.

Similarly, alternative ML models including Decision Tree Classifier (DTC), Naive Bayes Classifier (NBC), and Support Vector Classifier (SVC) faced similar limitations in demonstrating significant results. Despite prior literature suggesting SVC's promise in bankruptcy prediction, its effectiveness did not materialize in this study. It is worth noting that previous studies often emphasized resampling strategies, but in this investigation, ensemble methods emerged as more effective tools. Notably, among various strategies, only random undersampling demonstrated satisfactory results when applied alongside the Random Forest Classifier (RFC).

Drawing upon the models' predictive efficacy, certain financial ratios carry significant importance in protecting a company from the brink of bankruptcy. These include ROA(C) before interest and depreciation before interest, the Degree of Financial Leverage (DFL), the level of Borrowing dependency, Debt ratio%, Non-industry income and expenditure/revenue, Equity to Liability ratio, Interest Coverage Ratio (measuring interest expense relative to EBIT), Total income/Total expense ratio, Interest Expense

Ratio, Net Value Per Share (B), and Total debt/Total net worth ratio. These metrics collectively hold the potential to play a critical role in shielding companies from the peril of bankruptcy.

A prospective avenue for future research involves an examination of bankruptcy datasets from diverse countries to ascertain the significant determinants that might contribute to a company's insolvency. This cross-country analysis could shed light on the universal and region-specific factors influencing bankruptcy risk. Additionally, an exploration into the integration of multiple feature selection techniques presents another promising direction. Investigating various methodologies for selecting pertinent features can enhance the robustness and reliability of bankruptcy prediction models.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis, *Wiley interdisciplinary reviews: computational statistics* **2**(4): 433–459.

Aghaie, A. and Saeedi, A. (2009). Using bayesian networks for bankruptcy prediction: Empirical evidence from iranian companies, *2009 International Conference on Information Management and Engineering*, IEEE, pp. 450–455.

Alin, A. (2010). Multicollinearity, *Wiley interdisciplinary reviews: computational statistics* **2**(3): 370–374.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The journal of finance* **23**(4): 589–609.

Barboza, F., Kimura, H. and Altman, E. (2017). Machine learning models and bankruptcy prediction, *Expert Systems with Applications* **83**: 405–417.

Beaver, W. H. (1966). Financial ratios as predictors of failure, *Journal of accounting research* pp. 71–111.

Belsley, D. A., Kuh, E. and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, John Wiley & Sons.

Bisong, E. and Bisong, E. (2019). Google colaboratory, *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners* pp. 59–64.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning, *Artificial intelligence* **97**(1-2): 245–271.

Brereton, R. G. and Lloyd, G. R. (2010). Support vector machines for classification and regression, *Analyst* **135**(2): 230–267.

Brownlee, J. (2019). How to choose a feature selection method for machine learning, *Machine Learning Mastery* **10**.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Chen, C., Liaw, A., Breiman, L. et al. (2004). Using random forest to learn imbalanced data, *University of California, Berkeley* **110**(1-12): 24.

Chen, L.-H. and Hsiao, H.-D. (2008). Feature selection to diagnose a business crisis by using a real ga-based support vector machine: An empirical study, *Expert systems with applications* **35**(3): 1145–1155.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography* **36**(1): 27–46.

Fawcett, T. (2006). An introduction to roc analysis, *Pattern recognition letters* **27**(8): 861–874.

Gudmund, R., Norpoth, H. et al. (1976). Analysis of variance, *Beverly Hill: Sage Publications* .

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of machine learning research* **3**(Mar): 1157–1182.

Hauser, R. P. and Booth, D. (2011). Predicting bankruptcy with robust logistic regression, *Journal of Data Science* **9**(4): 565–584.

He, H. and Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications.

Horak, J., Vrbka, J. and Suler, P. (2020). Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison, *Journal of Risk and Financial Management* **13**(3): 60.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013). *Applied logistic regression*, Vol. 398, John Wiley & Sons.

Joshi, S., Ramesh, R. and Tahsildar, S. (2018). A bankruptcy prediction model using random forest, *2018 second international conference on intelligent computing and control systems (ICICCS)*, IEEE, pp. 1–6.

Kim, H., Cho, H. and Ryu, D. (2022). Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data, *Computational Economics* **59**(3): 1231–1249.

Kim, H. and Gu, Z. (2006). A logistic regression analysis for predicting bankruptcy in the hospitality industry, *The Journal of Hospitality Financial Management* **14**(1): 17–34.

Kitchenham, B. (2004). Procedures for performing systematic reviews, *Keele, UK, Keele Univ.* **33**.

Klecka, W. R. (1980). *Discriminant analysis*, Vol. 19, Sage.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S. et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows., *Elpub* **2016**: 87–90.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection, *Artificial intelligence* **97**(1-2): 273–324.

Le, T. (2022). A comprehensive survey of imbalanced learning methods for bankruptcy prediction, *IET Communications* **16**(5): 433–441.

Liang, D., Lu, C.-C., Tsai, C.-F. and Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study, *European journal of operational research* **252**(2): 561–572.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering, *IEEE Transactions on knowledge and data engineering* **17**(4): 491–502.

Liu, X.-Y., Wu, J. and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(2): 539–550.

Min, J. H. and Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert systems with applications* **28**(4): 603–614.

Mohammed, R., Rawashdeh, J. and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results, *2020 11th international conference on information and communication systems (ICICS)*, IEEE, pp. 243–248.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision support systems* **50**(3): 559–569.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy, *Journal of accounting research* pp. 109–131.

Olson, D. L., Delen, D. and Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction, *Decision Support Systems* **52**(2): 464–473.

Paraschiv, F., Schmid, M. and Wahlstrøm, R. R. (2021). Bankruptcy prediction of privately held smes using feature selection methods, *SSRN Electronic Journal* .

Rish, I. et al. (2001). An empirical study of the naive bayes classifier, *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3, pp. 41–46.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics* **21**(3): 660–674.

Sanner, M. F. et al. (1999). Python: a programming language for software integration and development, *J Mol Graph Model* **17**(1): 57–61.

Shin, K.-S., Lee, T. S. and Kim, H.-j. (2005). An application of support vector machines in bankruptcy prediction model, *Expert systems with applications* **28**(1): 127–135.

Smiti, S. and Soui, M. (2020). Bankruptcy prediction using deep learning approach based on borderline smote, *Information Systems Frontiers* **22**: 1067–1083.

Sun, L. and Shenoy, P. P. (2007). Using bayesian networks for bankruptcy prediction: Some methodological issues, *European Journal of Operational Research* **180**(2): 738–753.

Syed Nor, S. H., Ismail, S. and Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model, *Journal of Economics, Finance and Administrative Science* **24**(47): 157–170.

Taffler, R. J. (1983). The assessment of company solvency and performance using a statistical model, *Accounting and Business Research* **13**(52): 295–308.

Taiwanese Bankruptcy Prediction (2020). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5004D.

Valencia, C., Cabrales, S., Garcia, L., Ramirez, J. and Calderona, D. (2019). Generalized additive model with embedded variable selection for bankruptcy prediction: Prediction versus interpretation, *Cogent Economics & Finance* **7**(1): 1597956.

Van Gestel, T., Baesens, B., Suykens, J., Espinoza, M., Baestaens, D.-E., Vanthienen, J. and De Moor, B. (2003). Bankruptcy prediction with least squares support vector machine classifiers, *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.*, IEEE, pp. 1–8.

Veganzones, D. and Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets, *Decision Support Systems* **112**: 111–124.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction, *International work-conference on artificial neural networks*, Springer, pp. 758–770.

Wang, H. and Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data, *Plos one* **16**(7): e0254030.

Wang, N. et al. (2017). Bankruptcy prediction using machine learning, *Journal of Mathematical Finance* **7**(04): 908.

Wright, R. E. (1995). Logistic regression.

Yang, Z., You, W. and Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction, *Expert Systems with Applications* **38**(7): 8336–8342.

Yeh, J.-Y. and Chen, C.-H. (2020). A machine learning approach to predict the success of crowdfunding fintech project, *Journal of Enterprise Information Management* (ahead-of-print).

Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods, *Knowledge-Based Systems* **41**: 16–25.