Journal of Infrastructure, Policy and Development 2025, 9(1), 10072.
https://doi.org/10.24294/jipd10072

Article

A machine learning and ANOVA-based approach to model financial predictors for corporate failure in imbalanced dataset: The case of Taiwan

Borislava Toleva[1,*], Ivan Ivanov[1], Vincent Hooper[2]

[1] Faculty of Economics and Business Administration, Sofia University, 1113 Sofia, Bulgaria
[2] SP Jain School of Global Management, Block 5, Dubai International Academic City, Dubai 502345, UAE
*** Corresponding author:** Borislava Toleva, vrigazova@uni-sofia.bg

CITATION

Toleva B, Ivanov I, Hooper V. (2025). A machine learning and ANOVA-based approach to model financial predictors for corporate failure in imbalanced dataset: The case of Taiwan. Journal of Infrastructure, Policy and Development. 9(1): 10072. https://doi.org/10.24294/jipd10072

ARTICLE INFO

Received: 4 November 2024
Accepted: 9 December 2024
Available online: 6 January 2025

Abstract: This study meticulously explores the crucial elements precipitating corporate failures in Taiwan during the decade from 1999 to 2009. It proposes a new methodology, combining ANOVA and tuning the parameters of the classification so that its functional form describes the data best. Our analysis reveals the ten paramount factors, including Return on Capital ROA(C) before interest and depreciation, debt ratio percentage, consistent EPS across the last four seasons, Retained Earnings to Total Assets, Working Capital to Total Assets, dependency on borrowing, ratio of Current Liability to Assets, Net Value Per Share (B), the ratio of Working Capital to Equity, and the Liability-Assets Flag. This dual approach enables a more precise identification of the most instrumental variables in leading Taiwanese firms to bankruptcy based only on financial rather than including corporate governance variable. By employing a classification methodology adept at addressing class imbalance, we substantiate the significant influence these factors had on the incidence of bankruptcy among Taiwanese companies that rely solely on financial parameters. Thus, our methodology streamlines variable selection from 95 to 10 critical factors, improving bankruptcy prediction accuracy and outperforming Liang's 2016 results.

Keywords: Taiwan bankruptcy dataset; imbalanced data; ANOVA; feature selection

1. Introduction

Corporate bankruptcy prediction in Taiwan involves the use of financial analysis, statistical models, and increasingly, machine learning techniques to assess the likelihood of a company becoming insolvent. This area of study is crucial for investors, creditors, financial institutions, and regulatory bodies, as it helps in making informed decisions regarding credit risk management, investment strategies, and policy formulation (Du et al., 2020; Liang et al., 2015; Pereira et al., 2016; Tian et al., 2015; Uthayakumar et al., 2020; Zelenkov et al., 2017; Zhou et al., 2015).

The traditional approach, to bankruptcy prediction in Taiwan, as in many other countries, has relied on financial ratios derived from company balance sheets and income statements (Kou et al., 2017; Zhao et al., 2024). These ratios often pertain to liquidity, profitability, leverage, and efficiency. Models such as the Altman *Z*-score, originally developed in the United States, have been adapted and applied to Taiwanese firms to evaluate their financial health. Altman's *Z*-score model, developed by Edward I. Altman in the 1960s, is a financial formula used to predict the likelihood of bankruptcy among companies. It does not directly include a component for "model imbalance" in the way that modern machine learning models might address class imbalance in datasets.

In recent years, there has been a shift towards more sophisticated analytical techniques (Zhang, 2017; Zhao et al., 2024). Machine learning models, and deep learning models (Brenes et al., 2022), including decision trees, neural networks, and support vector machines, have been employed to improve the accuracy of bankruptcy predictions (Liang et al., 2016). These models can handle a vast amount of data and identify complex nonlinear relationships between various financial indicators and bankruptcy risk. Machine learning models have been widely used in various fields like public health, education, food science, telemedicine and the renewable energy sector (Chatterjee et al., 2024; Kooptiwoot et al., 2024a, 2024b, 2024c; Kooptiwoot and Javadi, 2022).

The development of bankruptcy prediction models in Taiwan considers unique local economic conditions, regulatory environments, and industry-specific factors that might affect a company's financial stability (Du et al., 2020; Zhao et al., 2024). For instance, the importance of the technology sector in Taiwan's economy requires models to account for the rapid innovation and fluctuating market demands that characterize this industry. The examination of bankruptcy within the realms of economics and finance has garnered significant attention, with an emphasis on identifying and understanding the multifaceted factors contributing to corporate insolvency. Academic inquiries have centred around the analysis of financial ratios derived from companies' financial statements as indicators of bankruptcy risk (Du et al., 2020; Liang et al., 2015; Liang et al., 2016; Pereira et al., 2016; Tian et al., 2015; Uthayakumar et al., 2020; Zelenkov et al., 2017; Zhou et al., 2015). This approach underscores the value of traditional financial metrics in assessing a firm's health. Concurrently, there has been a growing interest in the exploration of payment network-based variables as predictors of bankruptcy, improving the understanding of financial interconnectedness and its implications (Kou et al., 2021; Letizia and Lillo, 2019). Furthermore, the impact of non-financial factors on bankruptcy risk has been acknowledged, expanding the scope of research beyond mere financial indicators (Grunert et al., 2005).

The literature also addresses the influence of asymmetric information and conflicts of interest among creditors, pinpointing these elements as not only exacerbating the risk of bankruptcy but also contributing to the elevation of bankruptcy costs (Dou et al., 2021). These studies illustrate the complex interplay of internal and external factors in the bankruptcy process.

In response to the multifaceted nature of bankruptcy determinants, researchers have employed a diverse array of methodologies to isolate the most predictive factors. A significant portion of this research has leveraged machine learning (ML) and deep learning (DL) algorithms, alongside traditional credit scoring models based on statistical analysis. These models aim to refine the prediction of bankruptcy risk by filtering through a vast array of potential predictors to identify those with the most significant impact (Du et al., 2020; Liang et al., 2015; Pereira et al., 2016; Tian et al., 2015; Uthayakumar et al., 2020; Zelenkov et al., 2017; Zhou et al., 2015). The integration of feature selection techniques within ML and deep learning models represents a strategic approach to tackling the challenge of identifying the most relevant bankruptcy predictors amid a sea of data.

However, the heterogeneity of bankruptcy factors across different datasets, economic sectors, and geographical regions complicates the task of generalizing findings. As a result, bankruptcy models are often tailored to specific contexts, such as public companies, or are focused on countries or regions (e.g., Lohmann and Möllenhoff (2023) for public companies; Kalak et al. (2017) for U.S. companies; Yuxia et al. (2022) for Chinese companies; Mateika (2022) for Taiwanese companies). This specificity addresses the challenges posed by data complexity, including class imbalance, the scarcity of high-quality data, and variations in macroeconomic and regulatory environments (Fernández-Gámez et al., 2020; Mattos and Shasha, 2024; Shen et al., 2020).

Given these complexities, it becomes evident that research often focuses on a singular sector or country to investigate bankruptcy factors. In this vein, our research endeavours to pinpoint the ten most critical factors precipitating bankruptcy among Taiwanese companies during the period 1999–2009. Building upon a similar study that utilized the same dataset, the Taiwan corporate bankruptcy dataset, our research introduces innovative feature selection methodologies to address the issue of class imbalance, thereby enhancing the predictive accuracy of bankruptcy determinants in the Taiwanese context (Liang et al., 2016). The contributions of this paper can be summarized in several directions. First, an effective algorithm is introduced that outperforms other existing algorithms despite the class imbalance issue. This result suggests that the proposed algorithm can be further tested to solve the general issue of class imbalance in other datasets. Second, it provides better results than other researchers, which means that an in-depth analysis of the factors that drove Taiwan corporate bankruptcy between 1999 and 2009 can be performed. Such an analysis is important for economists and financiers to isolate factors that have had a temporary contribution to corporate bankruptcy from factors that would usually drive this bankruptcy. Therefore, this study can be extended to a greater period and may have important implications for economists and policy makers. Therefore, this research not only contributes to the existing body of literature but also offers practical insights for stakeholders seeking to mitigate bankruptcy risks within this specific geographical and economic milieu.

The rest of the paper is organized as follows:

Section 2, ‘Literature review’, demonstrates the history of bankruptcy models and explains some of the current most popular algorithms. Section 3, ‘Materials and methods’, details the innovative approach combining feature selection and a machine learning classification algorithm to analyses the financial indicators leading to corporate failures in Taiwan. Section 4, ‘Results’, presents the outcomes of the analysis, highlighting the ten critical financial factors identified as predictors of bankruptcy. This is a substantial reduction in the variables used in previous studies using the same database without losing accuracy or including corporate governance variables. Finally, section 5, ‘Discussion’, synthesizes the study’s findings, emphasizing the significant impact of these financial variables on corporate bankruptcy risk and suggesting implications for future research and practice in financial risk management. More specifically our approach narrows down from 95 to 10 key predictors, enhancing bankruptcy prediction and surpassing previous benchmarks and finds no need to include corporate governance variables.

2. Literature review

The earliest research of corporate bankruptcy dates to 1932 when Fitzpatrick analysed financial ratios of failed companies. This was one of the first research to provide a system to predict the success of the company. The research was applied on thirteen financial ratios. Accounting combined with univariate analysis was the first tool to model corporate failures. Since then, various tools have been used, with them becoming increasingly interdisciplinary. Zhao et al. (2024) provides a thorough review of the development of tools for corporate failure. He summarizes existing research from 1932 till now in several fields. He defines six fields in which the research of corporate bankruptcy has developed in the past century:

- Defining corporate failure and economic distress and identifying them,
- Defining prediction models for corporate bankruptcy,
- Inclusion of new factors in existing prediction models and building new models,
- Using feature selection to define factors for corporate bankruptcy,
- Designing methodologies for evaluation of model performance,
- Analysis of issues affecting model performance.

Zhao et al. (2024), and Veganzones and Severin's (2021) research provides a summary of the evolution of general model building and performance. However, based on his research, a summary of the evolution of corporate failure models in terms of the model type can be made:

- Univariate analysis—where accounting knowledge is combined with statistics to distinguish between successful and failing companies (Fitzpatrick, 1932). Since then, more accounting ratios have been used with the univariate analysis, examining their independent effects on corporate failure. In this analysis, only one variable at a time has been explored to model corporate failure.
- Altman's (1968) *z*-score model—this model marked a transition to multivariate analysis based on a multivariate discriminant analysis. The simultaneous influence of several variables on company bankruptcy could now be explored, which was revolutionary.
- Stochastic modelling has been used since 1973 (Black and Scholes, 1973). Stochastic models have allowed considering time variance an important factor for corporate failure. Also, dynamic models could be built. Black and Scholes' model is the famous stochastic model that continues to be used today.
- Machine learning. In 1980 classification models started being used (Ohlson, 1980). Since then, various versions of regression models have been introduced and modified to meet the requirements of the corporate failure environment.
- Deep learning (Odom and Sharda, 1990). In 1990 artificial neural networks were applied to corporate bankruptcy for the first time allowing hidden relations among various factors to be examined. Also, how changes in factors can affect a company's bankruptcy. This line of research allowed for modelling much more complex data, often without a prior structure.
- Bayes classification—Sarkar and Sriram (2001) were the first ones to apply Bayes classification in the field of corporate failures.

Since 2001 academic research has been focused on several lines of research in corporate bankruptcy. The first line is testing the significance of certain economic variables for corporate bankrupt by adding them to the econometrics models of their economies, e.g., Noga and Adamowicz (2021), Salehi and Pour (2016), Almamy et al. (2016), Ékes and Koloszár (2014), Karas et al. (2023), Hwang et al. (2009), Ko et al. (2017), Ruxanda et al. (2018). Other authors tailor the corporate bankruptcy prediction models to specific contexts such as:

- public companies by specific regions (e.g., Lohmann and Möllenhoff (2023) for public companies; Kalak et al. (2017) for U.S. companies; Yuxia et al. (2022) for Chinese companies; Mateika (2022) for Taiwanese companies).
- Modelling issues like data complexity, including class imbalance, the scarcity of high-quality data, and variations in macroeconomic and regulatory environments (Fernández-Gámez et al., 2020; Mattos and Shasha, 2024; Shen et al., 2020; Voda et al., 2021).

Despite the various directions and models for corporate prediction, machine learning and deep learning models have proved to be the most flexible ones for corporate failure prediction. A recent comprehensive overview of the most often used machine learning (ML) and deep learning (DL) models for corporate failure prediction can be found in (Ahmed et al., 2022; Dasilas and Rigani, 2024; Jones et al., 2023; Qu et al., 2019). Due to their ability to capture hidden relations in complex datasets and overcome the lack of flexibility of traditional credit score and econometric models, ML and DL models have become a key group of bankruptcy prediction models (Bock et al., 2020; Borchert et al., 2023; Bragoli et al., 2022; Jabeur et al., 2023).

Despite their flexibility, ML and DL models demonstrate a further level of complexity that traditional credit scores and econometric models do not. The algorithm we propose aims to fill this gap. In this research, we propose a simple but effective ML algorithm for predicting corporate bankruptcy given class imbalance in the target variable. The advantages of the proposed algorithm are simplicity, fast calculation, easy parameter tuning and wide applications to various bankruptcy datasets as it has built-in algorithm for identifying the most important variables for bankruptcy.

In this research we built our methodology on the Taiwan Bankruptcy dataset (Taiwanese Bankruptcy Prediction—UCI Machine Learning Repository). This dataset is suitable for the aim of this paper due to three reasons. The first one is that it is one of the most used datasets for developing bankruptcy models for companies, which provides ground for comparison with existing research (Brenes et al., 2022; Liang et al., 2016). The second reason is that valuable insights into Taiwan corporate bankruptcy can be uncovered that can be useful for two purposes. First, get a deeper understanding of the Taiwan corporate environment. Second, understand the drivers behind Taiwan corporate bankruptcy and make a similar analysis in other countries to understand whether some drivers of the corporate bankruptcy are similar among various countries. Therefore, our methodology provides a broader field for research of corporate bankruptcy. That is providing a tool for comparing bankruptcy factors among countries without introducing bias coming from applying different models to countries. Lastly, as research on this dataset is scarce and the dataset suffers from class

imbalance, the Taiwan corporate bankruptcy dataset becomes a solid ground for developing ML corporate bankruptcy models that perform well in the presence of class imbalance.

3. Materials and methods

The methodology's robustness lies in its systematic approach to data preparation, feature selection, and model evaluation, tailored to address the specific challenges of class imbalance with respect to bankruptcy prediction. This section sets the stage for the results and analysis, where the effectiveness of the proposed algorithm and the implications of the identified bankruptcy factors will be thoroughly explored.

Figure 1 shows a diagram of the methodology proposed.

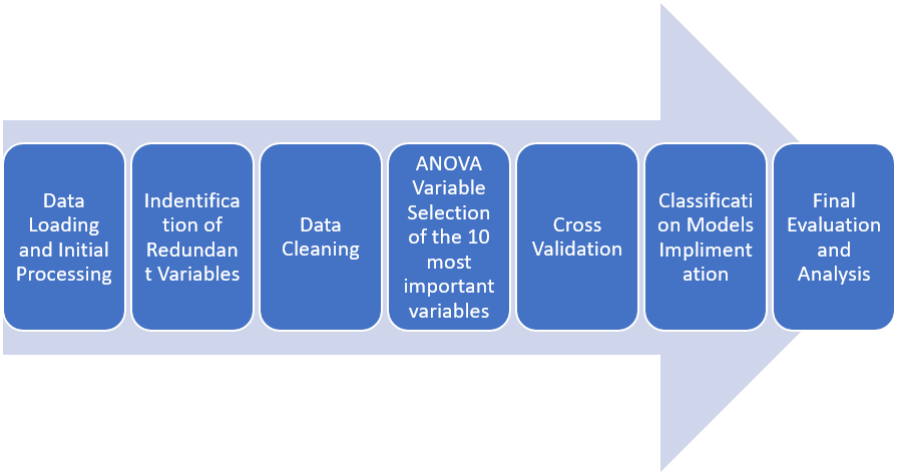

Figure 1. Proposed methodology.

Source: the authors.

The methodology section of this research outlines a comprehensive approach to identifying the most significant factors influencing bankruptcy among Taiwanese companies, leveraging a dataset comprising 6819 entries and 95 independent variables (Liang et al., 2016). The dataset's dependent variable, termed 'bankrupt,' exhibits a pronounced class imbalance, with most cases belonging to the non-bankrupt class. This study introduces a novel algorithm designed to mitigate the challenges posed by class imbalance and isolate the ten variables most critical to bankruptcy prediction. The methodology is executed using Python 3.11 on a Windows 11 system, powered by an Intel Core i3 processor, underscoring the study's technological and computational foundation. The database can be found in (Liang et al., 2016).

3.1. Description of ML models

In this section we present only classification models that performed well with the proposed methodology. The classification models we used are support vector machines (SVM), decision trees (DT), logistic regression (LR) and random forest (RF). We selected these models as they are widely applied to model corporate bankruptcy (Brenes et al., 2022; Liang et al., 2016; Mateika et al., 2022). Therefore, comparability of results is possible. Also, improving the output from widely applied

ML models would have greater practical implications when uncovering reasons behind corporate failure.

These machine learning algorithms are very flexible; however, each has different parameters that account for data characteristics. The model's parameters must be tuned properly to provide a good fit for the data. For instance, support vector machines divide all observations into two or more hyperplanes depending on the distance among observations and the number of classes. SVM models have two key parameters for model's performance—the parameter C and the kernel. The C parameter is a shrinkage parameter, controlling the trade-off between bias and variance, while the kernel adjusts the model to reflect the data curve as accurately as possible. These attributes make the SVM applicable in a wide range of real-life problems. Therefore, it is often used for corporate bankruptcy as well (Zhao et al., 2024).

Logistic regression is often used in tasks with two classes. It models the probability of one observation belonging to a class. It is less flexible than the SVM. Despite this, LR is often used for bankruptcy predictions as it captures the probability of going bankrupt. Logistic regression is often used as the base of a traditional credit scoring model (Zhao et al., 2024). But LR can also be used independently to predict corporate bankruptcy (Zhao et al., 2024). In this paper, we use it independently.

Decision trees and random forests are often used in datasets where class imbalance is present and there are complex data connections. They can be also used as a base for a credit scoring system predicting corporate failure. Like LR, DT and RF can also be used independently to model corporate failure (Zhao et al., 2024). In this paper we use them as two separate classification models to test them with our methodology. They are suitable for the Taiwan corporate failure dataset as the dataset suffers from heavy class imbalances (James et al., 2021).

As the next two subsections demonstrate, prior to classification, we employed ANOVA (Analysis of Variance) to select the ten most important variables (Ross, 2019). Although various variable selection methods exist, we use ANOVA due to its ability to rate the importance of variables prior to and independently of classification. Using ANOVA, a list of all variables rated by their contribution to corporate bankruptcy is provided. Then, the researcher can either review it manually to notice some hidden connections or they can adjust the ANOVA model to select automatically the most k important variables and use them in classification. Also, ANOVA achieves this aim in a simple way without further complicating the interpretation of the important variables unlike other similar methods (e.g., Recursive Feature Elimination and Extra Tree Classifier) (Mohtasham et al., 2024).

Next two subsection detail the proposed methodology.

3.2. Data preparation and feature selection

Step 1: Data loading and initial processing.

- The initial step involves loading the dataset and delineating the independent (X) and dependent (y) variables. An inspection for high correlation among the X variables is conducted to ensure no redundancy affects the model's performance. The y variable is transformed into categorical labels to facilitate analysis.

Step 2: Identification of redundant variables.

- Variables exhibiting a correlation coefficient above 0.8 are deemed highly correlated and are subsequently removed from the dataset. This reduction process leaves 70 variables from the original 95.

Step 3: Data cleaning.

- This step focuses on eliminating variables with substantial missing values or erroneous data entries. An emphasis is placed on removing implausible rate variables, applying a threshold whereby rates must not exceed one. This criterion further narrows the field to 46 variables.

Step 4: Variable selection via ANOVA.

An ANOVA ranking is employed to discern the ten most influential independent variables, utilizing the SelectKBest function with f_classif as the scoring mechanism. This process identifies a subset of variables that significantly contribute to the predictive model. Variables with higher ANOVA scores contribute to corporate bankruptcy to a higher extent. The commands for this step are shown below:

test = SelectKBest(score_func = f_classif, *k* = 10),
fit = test.fit(clean, *y*), set_printoptions(precision = 3)
print(fit.scores_)
features = fit.transform(clean)

The dataframe called features contains the first 10 most important independent variables for explaining the bankrupt (*y*) based on f-classification. The 10 most important variables are selected by the SelectKBest function by taking the variables with the highest ANOVA scores. The matrix called 'features' contains the 10 most important variables. The variables included are shown in the Results section.

It is important to note that the researcher may perform classification with a smaller or larger number of important variables. In this case, the parameter k should be changed to the necessary value. Also, the parameter 'score_fun' defines how the ANOVA scores are calculated. They can be calculated either based on f-classification or based on chi-square tests. However, chi-square works only with categorical data. As some of the independent variables are numerical, we applied ANOVA scoring through f-classification.

3.3. Model evaluation and validation

Step 5: Stratified cross-validation setup.

- To address the dataset's class imbalance, stratified k-fold cross-validation is employed with the number of splits set to ten (Szeghalmy and Fazekas, 2023). We have used this number as it is a standard in machine learning theory (James et al., 2021). The option for random shuffling is set to True so that each training set would contain random observations that are different from the previous training set. This is necessary to build a greater set of existing examples for the model to predict better the class of a new observation. The random seed is set to 7. This number is selected based on our empirical results. Theory has no strict recommendation what number for seed to be used. This approach ensures a balanced representation of classes across each fold, enhancing the model's generalizability and reliability. This step is given by the commands:

Set seed = 7

skf = StratifiedKFold(n_splits = 10, shuffle = True, random_state = seed).

The random seed may be any number. It guarantees the reproducibility of the experiments and defines the randomness of training and test splits. We used 7 as this number provided the most suitable training and test sets for our experiments. The number of splits is set to ten, similarly to other papers (Zhao et al., 2024). Setting the parameter 'shuffle' to True guarantees that at each iteration of the cross validation different observations will be included in the training and test sets. Therefore, in-sample and out-of-sample forecasting is performed.

Step 6: Classification model implementation.

- Four classification models, logistic regression (LR), support vector machines (SVM), decision tree (DT) and Random Forest (RF) are chosen for their ability to handle binary outcomes effectively. All models are adjusted to account for class imbalance by setting the class_weight parameter to 'balanced' (Python documentation, 2023). This is a novel application of the class_weight' parameter in Python for corporate bankruptcy prediction. Although this parameter is available in Python by default, studies about its efficiency in balancing classes for bankruptcy prediction have not been performed. Adjusting the rest of the parameters following best practices recommended in the literature. A note should be made that the values shown for the parameters are based on our empirical tests. Therefore, other combination of parameters that provides satisfactory results may exist. We aim to demonstrate how proper tuning of the model's parameter may eliminate the need for a more complicated classification model. The models we propose have their parameters adjusted as follows:
- LR—LogisticRegression(class_weight = 'balanced')
- SVM—SVC(*C* = 11, kernel = 'poly', gamma = 'auto', class_weight = "balanced")
- RF—RandomForestClassifier(n_estimators = 35, max_depth = 3, random_state = 33, class_weight = 'balanced')
- DT—DecisionTreeClassifier(max_depth = 3, class_weight = 'balanced'). In this model, however, the stratified cross validation uses random_state of 645.

In this step the parameters of the models are adjusted. This means that the selected parameters resulted from experimental trials. Therefore, many combinations of parameters that result in high accuracy and predict all classes accurately exist. However, discovering one or two models with proper parameters adjustment may be enough for the researcher. We performed parameters' adjustments or 'tuning' to demonstrate that in some cases proper adjustment of the classification model may be a much simpler and more reliable alternative to complex ML and DL models. Therefore, performing experiments to demonstrate some of the most suitable combinations of parameters fulfils the purpose of this paper.

Step 7: Analysis and interpretation.

- The last step involves running the selected classification models using the stratified cross-validation framework established in Step 6. Moreover, we apply the cross_val_predict function to evaluate used models. The cross_val_predict function provides the prediction that was obtained for each element in the input when it was in the test set. Cross-validation strategies assign all elements to a test set exactly once. As a result, we obtained the prediction of the full dataset. Then,

the confusion matrix is obtained for the dataset. Entries of the confusion matrix are true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Model performance is evaluated through various metrics, including the confusion matrix, accuracy, precision, recall, *F*1-score, and type I error, and type II error. The computations are done via well-known formulas for accuracy, precision, recall, sensitivity, *f*1-score, Type I Error and Type II Error (James et al., 2021):

$$\text{Accuracy} = (TP + TN) \div (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP \div (TP + FP) \quad (2)$$

$$\text{Specificity} = TN \div (TN + FP) \quad (3)$$

$$\text{Sensitivity (Recall)} = TP \div (TP + FN) \quad (4)$$

$$F1 \text{ score} = 2 \times (\text{Precision} \times \text{Recall}) \div (\text{Precision} + \text{Recall}) \quad (5)$$

$$\text{Type I Error Rate} = FPR = FP \div (FP + TN) = 1 - \text{Specificity} \quad (6)$$

$$\text{Type II Error Rate} = FNR = FN \div (FN + TP) = 1 - \text{Sensitivity} \quad (7)$$

Accuracy is a measure of correct predictions among all observations in the dataset. Precision measures the percentage of correct positive cases among all positive cases predicted by the model. Recall defines the percentage of correct positive cases among all positive cases in the actual data. F1 score represents the harmonic mean of precision and recall and can be used as an accuracy measure in imbalanced classifications. Types I and II errors reflect the cases when true positive cases are rejected and when a false case is accepted as true. Classification metrics given in Equations (1)–(7) cannot be interpreted alone. They need to be interpreted together and with the elements of the confusion matrix, which provides information about the accuracy of prediction each class. Together these metrics can define a model that is a good fit for the data (James et al., 2021).

4. Results

In contrast to most algorithms evaluated on the Taiwan bankruptcy dataset, as documented by Liang et al. (2016), our approach is distinct in its objective to not only identify but also to interpret the top ten factors that have contributed to the insolvency of Taiwanese companies during the specified period. The conventional methodology adopted by researchers, including Wang and Liu (2021), relies on under sampling coupled with feature selection to ascertain the optimal amalgamation of factors for the most precise classification outcomes. Also, traditional approach examines the influence of various groups of financial variables on Taiwan bankruptcy by applying accounting logic, which may not be an efficient approach. An example is Liang et al.'s (2016) research, who used the same Taiwan bankruptcy dataset but unlike us, he selected combinations of variables manually based on the accounting and business logic. He grouped them in 12 new datasets, which he tested on Taiwan corporate bankruptcy. Such a manual approach is not recommended as some indicators may have

intrinsic relationships that can be unobserved by accounting and business logic. Therefore, an automatic approach that is simple, easy for application, and is fit for uncovering hidden relationships among financial variables may be a better approach. Such an approach is presented by the proposed methodology.

To achieve our unique research objectives, we selected to employ a two-phased feature selection process, initially implementing a filtering method based on correlation analysis, followed by the application of ANOVA (Analysis of Variance) with f-score for the selection of the ten most pivotal factors. This methodological choice is predicated on the principle that ANOVA, unlike other feature selection techniques, computes scores of importance for each variable through f-classification, as detailed in the scikit-learn 1.4.0 documentation on sklearn.feature_selection.f_classif.

A notable advantage of this approach is the consistency of the importance scores across different classification models, thereby ensuring the reliability of the feature ranking.

The significance of each feature is quantified by its ANOVA score, with a higher score denoting greater importance. Consequently, a lower ANOVA score indicates a diminished significance of the feature in question. This ranking mechanism is crucial for our analysis, as it directly informs the selection of features for subsequent phases of our research. As elucidated in the Methodology section, the ANOVA Python function incorporates a parameter that allows researchers to specify the number of top features to be considered for further analysis. In alignment with our research focus, this parameter was set to ten, thereby facilitating the identification of the ten most critical factors influencing company bankruptcy in Taiwan, as illustrated in **Table 1**. These factors are identified and ranked based on the methodological steps 1–4, with each factor's classification and significance contextualized by findings from Liang et al. (2016).

By adopting this approach to feature selection, our research endeavours to shed light on the complex interplay of factors contributing to corporate bankruptcies in Taiwan. Through the strategic application of ANOVA and classification algorithms adept at managing class imbalance, we aspire to contribute a novel perspective to the discourse on bankruptcy prediction, emphasizing the interpretative value of identifying key influencing factors.

As delineated in the Methodology section, the implementation of steps 1 through 4 yielded a refined dataset encapsulating the ten variables deemed most critical in assessing the bankruptcy risk among Taiwanese companies from 1999 to 2009. This subset of variables constitutes the foundation for subsequent analyses. Notably, the variable positioned as the eleventh in terms of importance—Total Expense/Assets, with an ANOVA score of −134.40—was found to be nearly as influential as the 'Liability-Asset Flag' variable listed within the top ten.

To evaluate the impact of including this eleventh variable, we conducted a redundancy test by incorporating it into the classification models previously established in section 2. When the eleventh most important feature is added to the LR model in **Table 2**, the accuracy becomes 85.6% vs LR accuracy with ten features of 85.57%. The SVM model in **Table 2** and the eleven most important variables results in accuracy of 84.8% compared to accuracy of 85.28% when the ten most important

variables are used. Comparable results are valid for the random forest and the decision tree classifier. Running the RF model from **Table 2** with the eleven most important variables results in accuracy of 87.4% compared to accuracy of 87.68% when the ten most important features are used. The decision tree classifier from **Table 2** results in accuracy of 84.79% when fitted with the eleven most important features compared to accuracy of 84.5% of the same model with the ten most important features. Therefore, this redundancy analysis revealed negligible variations in classification scores and metrics, leading to the inference that the inclusion of the eleventh variable does not significantly enhance the explanatory power regarding corporate bankruptcy in Taiwan during the specified period. This observation gains further support when considering the increased disparity in importance scores beyond the eleventh feature, such as the twelfth feature 'CFO to Assets' scoring −91.98, suggesting a diminishing return on model performance with the addition of subsequent variables. Conversely, model testing with fewer than ten variables indicated a decrement in predictive accuracy, substantiating the selection of the top ten features as optimal for explaining bankruptcy within the Taiwanese context during the study period.

Table 1. Ten most important factors for Taiwan company bankrupt based on the proposed methodology.

Feature Name	**ANOVA *F*-Score**
ROA(C) before interest and depreciation before interest	497.54
Debt ratio %	455.09
Persistent EPS in the Last Four Seasons	345.27
Retained Earnings to Total Assets (X63, FRs)	339.41
Working Capital to Total Assets (X12, FRs)	263.99
Borrowing dependency	219.30
Current Liability to Current Assets (X26, FRs)	206.10
Net Value Per Share (B)	191.74
Working Capital/Equity (X23, FRs)	151.03
Liability-Assets Flag	134.72

Source: authors' calculations.

In pursuit of affirming the sufficiency of these ten variables for elucidating the dynamics of corporate bankruptcy in Taiwan, our research adopts a focused approach divergent from prior studies that might aim for maximal predictive accuracy across all variables (Liang, Lu et al., 2016). Our objective is not to surpass these studies in predictive performance but to isolate a subset of factors that most significantly contribute to bankruptcy risk, thereby offering a targeted analysis that prioritizes relevance over comprehensiveness.

This strategic selection aligns with our goal to identify a core set of influential factors that, while not exhaustive, sufficiently capture the essence of bankruptcy risk within the Taiwanese corporate landscape. The efficacy of this approach is validated through the deployment of two classification models, which corroborated the adequacy of the top ten variables in achieving satisfactory accuracy, precision, recall, and confusion matrix outcomes. Such results affirm the premise that a concentrated

array of critical factors can effectively illuminate the underpinnings of corporate bankruptcy in Taiwan, thus fulfilling the research aim.

Further bolstering our findings, comparisons with broader analyses, such as those by Liang et al. (2016), which integrate financial ratios (FRs) and corporate governance indicators (CGIs) for bankruptcy prediction, serve to contextualize our study's scope and emphasis. While these comprehensive assessments offer valuable insights into predictive dynamics, our investigation distinctively prioritizes the identification of a concise yet potent combination of variables that efficiently captures bankruptcy risk, thereby contributing a focused lens through which to understand this complex phenomenon.

Expanding on the initial analysis, **Table 2** provides a comprehensive overview of the performance metrics for four distinct classification models: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Decision Tree (DT), as derived from steps 1–7 from section 2. These metrics illuminate the models' efficacy in distinguishing between the bankrupt (class 1) and non-bankrupt (class 0) Taiwanese companies during the study period. Precision, a key metric, evaluates the accuracy of each model when it identifies an instance as belonging to the target class.

Table 2. Output from the proposed methodology (%).

Model	Class	Accuracy	Precision	Recall	*F*1-Score	Support	Errors	Value
LogisticRegression(class_weight = 'balanced')								
LR	0	85.57	99.47	85.54	91.98	6599	Type I	13.63
	1		16.61	86.36	27.86	220	Type II	14.46
SVC(C = 11, kernel = 'poly', gamma = 'auto', class_weight = "balanced")								
SVM	0	85.28	99.5	85.21	91.8	6599	Type I	12.73
	1		16.44	84.55	27.67	220	Type II	14.79
RandomForestClassifier(n_estimators = 35, max_depth = 3, random_state = 33, class_weight = 'balanced')								
RF	0	87.68	99.42	87.78	93.24	6599	Type I	15.45
	1		18.75	85.44	30.69	220	Type II	12.22
skf = StratifiedKFold(n_splits = 10, shuffle = True, random_state = 645) DecisionTreeClassifier(max_depth = 3,class_weight = 'balanced')								
DT	0	84.79	99.43	84.77	91.52	6599	Type I	14.55
	1		15.76	85.45	26.61	220	Type II	15.23

Source: authors' calculations.

For the LR model, an impressive precision rate of 99.5% was achieved for class 0, indicating that when the model predicted a company as non-bankrupt, it was correct 99.5% of the time. Conversely, the model's precision for class 1 stood at 16.6%, meaning that when predicting bankruptcy, the prediction was accurate in 16.6% of instances. This disparity underscores the challenge of accurately predicting less frequent outcomes, such as bankruptcy, within the dataset.

The SVM model demonstrated a balanced precision rate across both classes, suggesting a consistent performance in identifying both bankrupt and non-bankrupt companies. This balance is crucial for models applied in environments where the cost of misclassification can be significant across either class.

The RF model showcased a precision of 99.4% for class 0, slightly lower than that of the LR model but still indicating a high level of accuracy for non-bankrupt predictions. For class 1, the RF model improved upon the LR's performance, correctly identifying bankruptcy in 18.75% of cases. This increase, although modest, indicates the RF model's enhanced ability to capture the complexity associated with bankruptcy indicators.

These findings corroborate the initial hypothesis posited during the feature selection phase, which identified the ten most critical features for predicting company bankruptcy in Taiwan during the period 1999–2009. These features, detailed in **Table 1** of the study, play a significant role in the models' ability to distinguish between bankrupt and non-bankrupt companies with high accuracy. The classification metrics and the confusion matrix further validate this assertion, reinforcing the reliability of the chosen features in forecasting company bankruptcy.

To further elucidate these findings, it is essential to examine the models' recall and *F*1-score metrics, which complement precision by measuring the models' ability to identify all relevant instances of the target class and balance between precision and recall, respectively. A high recall value indicates that the model can capture a large proportion of actual positive (bankrupt) cases, while a high *F*1-score suggests a balanced trade-off between precision and recall, which is particularly important in imbalanced datasets like the one under study.

Recall is a crucial metric in classification tasks, indicating the model's capability to identify all instances of a specified target class accurately. As per the data presented in **Table 2**, our Logistic Regression (LR) model demonstrates a notable proficiency in detecting 85.5% of all instances belonging to class 0 and 86.3% of instances within class 1. Conversely, the Support Vector Machine (SVM) algorithm exhibits a slightly lower recall, successfully identifying 85.21% of class 0 instances and 84.6% of class 1 instances. Furthermore, the Random Forest (RF) model achieves superior recall rates of 87.78% for class 0 and 85.44% for class 1, highlighting its effectiveness in recognizing target class instances.

Another pivotal metric in classification analysis is the F1-score, which represents the harmonic mean of precision and recall. This measure is particularly valuable in the context of classification data, offering a balanced view of both the model's precision and its ability to recall instances of the target class (as detailed in the scikit-learn 1.4.0 documentation on sklearn.metrics.*f*1_score). While the *F*1-score is widely applied in scenarios involving imbalanced datasets, its utility is not universally applicable. This is because the dominance of the majority class can sometimes yield misleading interpretations of the model's performance, as discussed in literature on the application of the *F*1 Score in machine learning (*F*1 score for imbalanced data, *F*1 Score in Machine Learning Explained | Encord). In our analysis, both the LR and SVM models report comparable *F*1-scores of 91.98% and 91.8% for class 0. However, for the minority class (class 1), the scores are markedly lower, at 27.9% and 27.7%, respectively. The RF model stands out with *F*1-scores of 93.2% for class 0 and 30.7% for class 1, indicating a slight improvement in balancing precision and recall for both classes.

Additionally, analysing the models' performance through the lens of the confusion matrix provides deeper insights into their predictive capabilities and

potential areas for improvement. The confusion matrix outlines the models' true positives, true negatives, false positives, and false negatives, offering a granular view of where each model excels and where it may falter.

Figures 2–5 display the confusion matrices from the LR, SVM, RF and DT models.

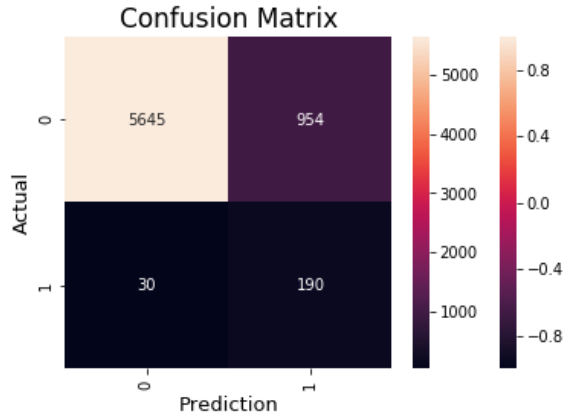

Figure 2. Confusion matrix from the LR model.

Source: authors' calculations.

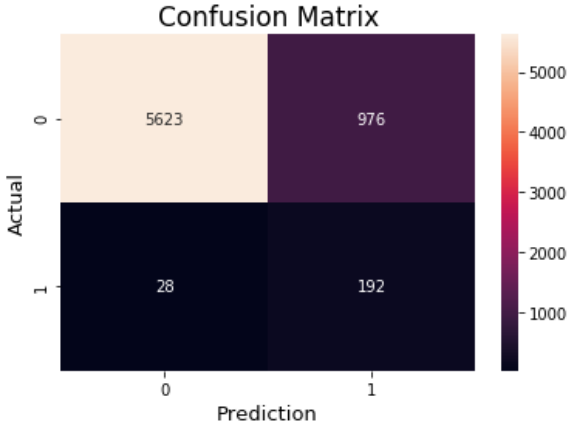

Figure 3. Confusion matrix from the SVM model.

Source: authors' calculations.

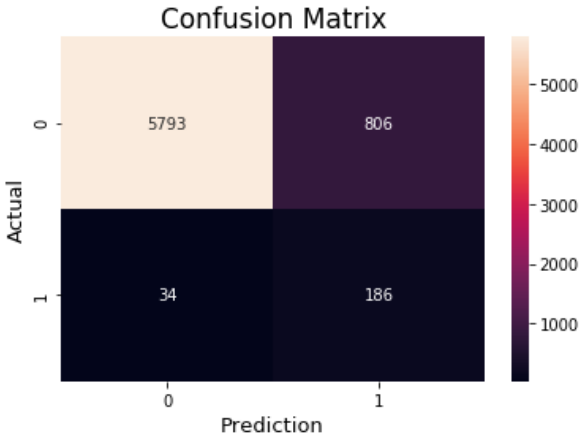

Figure 4. Confusion matrix from the RF model.

Source: authors' calculations.

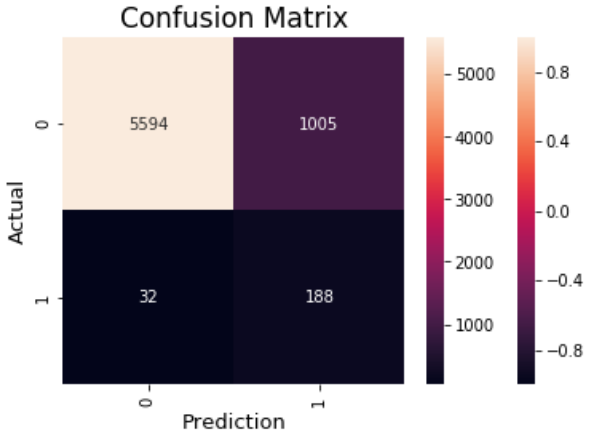

Figure 5. Confusion matrix from the DT model.

Source: authors' calculations.

In summary, the performance of the LR, SVM, RF, and DT models, as detailed in **Table 2**, highlights their respective strengths and limitations in predicting corporate bankruptcy in Taiwan. While precision rates are high for predicting non-bankrupt companies, the challenge remains in accurately identifying bankrupt cases, a common hurdle in imbalanced datasets. Further refinement of these models, through advanced feature engineering, more sophisticated algorithms, or hybrid approaches, could enhance their predictive accuracy, particularly for the minority class. This iterative process of model evaluation and enhancement is crucial for developing robust predictive tools that can effectively inform decision-making in the financial domain.

5. Discussions

Accuracy is another metric under consideration. This metric provides a general indication of the model's overall performance across both classes. **Table 3** shows the accuracy from another research. As **Table 3** shows, Liang et al. (2016) runs SVM, KNN and Cart models with different sets of variables. In the first case they include Financial Ratios (FR) and in the other—Corporate Government Indicators (CGIs). The highest accuracy they achieve is 79.1% using SVM with a set of Financial Ratios (FR). This accuracy is lower than our results from the four proposed models. Brenes et al. (2022) also performs experiments on the same dataset using Multilayer Perceptron and achieving accuracy of 86.06%. The algorithms we propose achieve better results. For instance, the LR model achieves an accuracy of 85.6%, the SVM model is close behind at 85.3%, and the RF model leading with an accuracy of 87.7%. The accuracy from our models is better than the ones Liang et al. (2016) and Brenes et al. (2022) achieves.

Table 3. Accuracy obtained by another research (Brenes et al., 2022; Liang et al., 2016).

	SVM(FRs/CGIs)	**k-NN (FRs/CGIs)**	**CART (FRs/CGIs)**	**Multilayer Perceptron (MLP)**
Accuracy	79.1%/67.9%	76.5%/60.6%	78.4%/60.2%	86.06%
Type I error	20.2%/27.7%	22.5%/30.7%	23.3%/37%	0.03%
Type II error	21.6%/36.5%	24.5%/48.1%	19.9%/42.5%	10.9%

The Logistic Regression model (table 2), with an overall predictive accuracy of 85.6%, slightly outperforms the Support Vector Machine model, which has an overall

accuracy of 85.3%. This slight discrepancy in performance highlights the uniqueness of each model's approach to the classification task, with LR potentially benefiting from its simplicity and direct approach to probability estimation. The RF model outperforms both the LR and SVM with an overall accuracy of 87.7%. However, it is a more complicated model that is slower. Nevertheless, the three models present viable options for stakeholders interested in assessing the bankruptcy risk of Taiwanese companies within the specified period, showcasing the value of rigorous feature selection and model evaluation in developing predictive tools for financial analysis.

Type I and II errors are other important metrics. Type I error denotes rejecting correctly classified observations, while Type II error means considering incorrect prediction to be correct. Therefore, they are another metric for the quality of class predictions. Because of this they need to be as low as possible. A trade-off between Type I and Type II error exists—when the first increases, the latter deceases and vice versa (James et al., 2021). **Tables 2** and **3** demonstrate that the proposed methodology results in lower Types I and II errors than Liang et al.'s (2016) results. However, this is not the case with the multilayer perceptron that Brenes (2022) presents. Brenes' Type I error is 0.03%, while we achieve the lowest Type I error by the SVM model, and it is 12.73%. Liang's Type I error is larger. This result is expected as Brenes uses a deep learning model, which is much more flexible than the SVM and other ML models. Brenes' model misclassifies correctly predicted observations in only 0.03% of the cases, while in the proposed methodology this happens in between 12.73% and 15.45% of the cases depending on the model used. This percentage is above 20% in Liang's models and in some cases above 37%. Therefore, the proposed methodology results in much lower Type I error compared to traditional ML models.

Similar findings can be found in **Tables 2** and **3**, looking at the Type II error. The type II error from the proposed methodology varies between 12.22% and 15.23% as seen in **Table 2**. Liang's Type II error varies between 36.5% and 48.1% (**Table 3**), which is much higher than the proposed methodology. Brenes' Type II error is the lowest—10.9%, which is close to some of the results in **Table 2**. The proposed methodology seems to result in a Type II error that is similar to the multilayer perceptron of Brenes. Therefore, the flexible multilayer perceptron, which is a deep learning model, accepts as correct false cases with a frequency similar to the proposed methodology. Liang's methodology, however, results in a much higher Type II error, which in combination with the remaining metrics proves that the proposed methodology outperforms classical ML models.

As seen in **Tables 2** and **3**, the proposed methodology outperforms Liang's ML models, while being competitive to more complex structures like deep learning models. The proposed methodology succeeds in this without further complicating the algorithm but simply by applying ANOVA and adjusting the parameters of the model so that they can suit the data characteristics better.

Another advantage of our methodology is that it performs well despite the presence of class imbalance, which simplifies further the interpretation of connections among related and non-related variables and their influence on corporate bankruptcy. Also, we show that setting the parameter 'class_weight' to 'balanced' in Python can be a simple tool to handle class imbalance in corporate bankruptcy prediction effectively. This is a novel application to handle class imbalance in corporate

bankruptcy prediction. Expanding upon our research, the methodology we have developed can be adapted and applied to other datasets characterized by imbalances, facilitating a comparative analysis across different countries and sectors. This approach holds the potential to uncover universal factors contributing to corporate bankruptcies, thereby enriching our understanding of bankruptcy predictors on a global scale. Moreover, the applicability of this research methodology extends beyond bankruptcy data, offering a framework for empirically investigating the performance of machine learning algorithms across diverse datasets marked by imbalance. Applying our methodology to other bankruptcy datasets with class imbalance opens another line of research—whether the output of our methodology can be improved given we use a technique for handling class imbalance. However, the proposed methodology works outperform existing research without additional algorithms for class imbalance, which makes it a tool with diverse applications for managerial, policy making and practical purposes.

6. Study limitations

Our findings carry significant practical implications, as the financial variables (with no corporate governance variables) highlighted in this study. The proposed methodology offers a comprehensive basis for further examination to understand the intricate dynamics that led to corporate bankruptcies in Taiwan during the specified period and future determinants. By delving deeper into the interrelations among these variables and their interactions with other economic and financial indicators, we can gain valuable insights that may aid in the future prevention of corporate bankruptcies.

It is essential to conduct further research to ascertain whether the identified ten factors are uniquely critical for corporate bankruptcies in Taiwan during this time frame, or if they signify a broader trend wherein certain factors recurrently emerge as pivotal in bankruptcy scenarios. Therefore, extending the time frame of the research would provide better insights into the drivers behind Taiwan corporate bankruptcy. If there are factors outlined by a more recent time frame are different from this research, then reasons behind differences can provide hidden insights, dependent on single events or regular events throughout the year.

Making comparison between the 1999–2009 dataset and a dataset between 1999—current would also have important policy implications as some break points may be discovered. Break points can be high-impact political, economic or financial events that affect the variables in the Taiwan bankruptcy datasets. However, due to their indirect influence on those variables, break points may not be captured in other ways than comparison. Our research is also limited in the types of factors that might be related to corporate bankruptcy in Taiwan. As Liang's dataset contains only accounting ratios, we cannot capture external influences like economic situation, political risks, etc. Therefore, the managerial, political and practical applications from our research are valid in the context of the set of variables contained in the dataset.

Despite this limitation, our research provides a tool for managers to explore the relationship among various accounting ratios with bankruptcy of companies. They can increase the number of corporate indicators. Also, variables for various social, economic, political, financial and other factors can be added to the dataset. Then, our

methodology can be replicated on the extended dataset to provide insights based on which managers can make decisions on a company, market and country level. Therefore, policymakers can also apply the proposed methodology for the same purposes.

7. Conclusion

In conclusion, our methodology, implemented using Python, leverages a suite of machine learning models—Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Decision Trees (DT)—to analyse the data. The effectiveness of this approach is evidenced by the results presented in **Table 3**, which demonstrate its capability to accurately identify bankruptcy predictors within the Taiwanese context with much fewer variables than ever before! This not only underscores the robustness of our methodology but also highlights its potential applicability in broader research endeavours aimed at understanding and preventing corporate financial distress within the context of imbalanced datasets in accounting and finance research.

Policymakers and managers can analyse how a variable or a set of variables affect Taiwan bankruptcy of companies by adjusting the number of variables similarly to Liang. Then, using our methodology they can identify conditions when one factor affects Taiwan corporate bankruptcy. Also, they can check the conditions when the same variable does not affect the corporate outcome. A time-saving advantage of the proposed methodology is the fact that policymakers and managers can identify influencing factors using a very large dataset without preliminary knowledge of the connections among variables. The reason for this is that we introduce a methodological approach that effectively reduces the number of variables to ten indicators, focusing on the most impactful factors for bankruptcy prediction. By utilizing these 10 factors, our classification methodology enhances the accuracy of bankruptcy forecasts, making them easy for interpretation.

Other important managerial, policy and practical implications are the easy adjustment of our algorithm to mark the number of significant variables that the manager/policymaker needs to examine. For instance, we outlined the ten most important variables, but, if necessary, our methodology may provide a bigger number of important variables and their significance. The algorithm can also be used simply for rating the importance of each variable for Taiwan bankruptcy to make conclusions of influences from other variables.

Author contributions: Conceptualization, BT and II; methodology, BT and II; software, BT and II; validation, BT, II and VH; formal analysis, BT, II and VH; investigation, II and VH; resources, BT, II and VH; data curation, BT; writing—original draft preparation, BT, II and VH; writing—review and editing, BT and VH; visualization, BT; supervision, BT, II and VH; project administration, BT; funding acquisition, II and VH. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This manuscript is supported by Sofia University 2024 Research Programme.

Data availability statement: The dataset can be downloaded at Taiwanese Bankruptcy Prediction—UCI Machine Learning Repository.

Conflict of interest: The authors declare no conflict of interest.

References

Ahmed Sh., Alshater M., Ammari A., Hammami H., Artificial intelligence and machine learning in finance: A bibliometric review, Research in International Business and Finance, 61, 2022, https://doi.org/10.1016/j.ribaf.2022.101646.

Almamy, J., Aston, J., & Ngwa, L., An evaluation of Altman's Z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: Evidence from the UK. Journal of Corporate Finance, 36, 2016, pp.278–285. https://doi.org/10.1016/j.jcorpfin.201

Altman, E., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance, 23, 4, 1968, pp. 589-609

ANOVA f-classification, https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

Black F., Scholes M., The pricing of options and corporate liabilities, Journal of Political Economy, 81, 3, 1973, pp. 637-654

Bock K., Coussement K., Lessmann S., Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach, European Journal of Operational Research, 285 (20), 2020, pp. 612-630, https://doi.org/10.1016/j.ejor.2020.01.052.

Borchert P., Coussement K., Caigny A., Weerdt J., Extending business failure prediction models with textual website content using deep learning, European Journal of Operational Research, 306, 2023, pp. 348-357

Bragoli D., Ferretti C., Ganugi P. & Marseguerra G., Mezzogori D., Zammori F., "Machine-learning models for bankruptcy prediction: do industrial variables matter?," Spatial Economic Analysis, Taylor & Francis Journals, 17 (2), 2022, pp. 156-177

Brenes R. F., Johannssen A., Chukhrova N., An intelligent bankruptcy prediction model using a multilayer perceptron, Intelligent Systems with Applications, 16, 2022, https://doi.org/10.1016/j.iswa.2022.200136.

Chatterjee S., Khan P., Byun Y., Recent advances and applications of machine learning in the variable renewable energy sector, Energy Reports, 12, 2024, pp. 5044-5065

Dasilas A., Rigani A., Machine learning techniques in bankruptcy prediction: A systematic literature review, Expert Systems with Applications, 255, Part C, 2024, https://doi.org/10.1016/j.eswa.2024.124761.

Dou W., Taylor L., Wang W., Wang W., Dissecting bankruptcy frictions, Journal of Financial Economics, 142 (3), 2021, pp. 975-1000, https://doi.org/10.1016/j.jfineco.2021.06.014.

Du X., Li W., Ruan S., Li L., CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection, Applied Soft Computing, 97, 2020, https://doi.org/ 10.1016/j.asoc.2020.106758.

Ékes, K. S., & Koloszár, L., The efficiency of bankruptcy forecast models in the Hungarian SME Sector. Journal of Competitiveness, 6(2), 2014, pp. 56–73. https://doi.org/10.7441/joc.2014.02.05

F1 score for imbalanced data, F1 Score in Machine Learning Explained | Encord

F1-score, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

Fernández-Gámez M.Á., Soria J., Santos J., Alaminos D., European country heterogeneity in financial distress prediction: An empirical analysis with macroeconomic and regulatory factors, Economic Modelling, 88, 2020, pp. 398-407

Fitzpatrick P., A comparison of the ratios of successful industrial enterprises with those of failed companies, The Certified Public Accountant, 12, 1932, pp. 727-731, 598-605, 656-662 respectively

Grunert J., Norden L., Weber M., The role of non-financial factors in internal credit ratings, J. Bank. Financ., 29 (2), 2005, pp. 509-531, https://doi.org/10.1016/j.jbankfin.2004.05.017

Hwang, R. C., Cheng, K. F., & Lee, C. F., On multiple-class prediction of issuer credit ratings. Applied Stochastic Models in Business and Industry, 25(5), 2009, pp.535–550, https://doi.org/10.1002/asmb.735

Jabeur S., Stef N., Carmona P., 2023, Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering," Computational Economics, Springer; Society for Computational Economics, vol. 61(2), pp. 715-741

James G., Witten D., Hastie T., Tibshirani R.: "An Introduction to Statistical Learning: with Applications in R", Springer, Second Edition 2021, https://www.statlearning.com/

Jones, S., "A literature survey of corporate failure prediction models", Journal of Accounting Literature, 45(2), 2023, pp. 364-405, https://doi.org/10.1108/JAL-08-2022-0086

Kalak I., Azevedo A., Hudson R., Karim M., Stock liquidity and SMEs' likelihood of bankruptcy: Evidence from the US market, Research in International Business and Finance, 42, 2017, pp. 1383-1393, https://doi.org/10.1016/j.ribaf.2017.07.077.

Karas, M., Reznakova, M., Bartos, V., & Zinecker, M., Possibilities for the application of the Altman model within the Czech Republic, In Recent Researches in Law Science and Finances, 2023, pp. 203–207, http://www.wseas.us/e-library/conferences/2013/Chania/ICFA/ICFA-30.pdf

Ko, Y. C., Fujita, H., & Li, T., An evidential analysis of Altman Z-score for financial predictions: Case study on solar energy companies, Applied Soft Computing, 52, 2017, 748–759, https://doi.org/10.1016/j.asoc.2016.09.050

Kooptiwoot, S. & Javadi, B., Development of Decision Support System Platform for Daily Dietary Plan, Current Nutrition & Food Science, 18, 2022, https://doi.org/10.2174/1573401318666220318102124.

Kooptiwoot, S. & Kooptiwoot, S. & Javadi, B., Application of regression decision tree and machine learning algorithms to examine students' online learning preferences during COVID-19 pandemic. International Journal of Education and Practice. 12, 2024, pp. 82-94, https://doi.org/10.18488/61.v12i1.3619. (a)

Kooptiwoot, S. & Tharasawatpipat, Ch. & Choo-In, S. & Kayee, P. & Javadi, B., AI-driven telemedicine: Optimizing daily dietary recommendations amidst the COVID-19 pandemic. Journal of Infrastructure, Policy and Development. 8, 2024, https://doi/org/10.24294/jipd.v8i11.8908. (b)

Kooptiwoot,S. & Tharasawatpipat, Ch.& Choo-in, S. & Kayee, P. & Meethongjan, K. & Sangsuwon, Ch. & Javadi, B., Deciphering the complexity of COVID-19 transmission: Unveiling precision through robust vaccination policies and advanced predictive modeling with random forest regression, Journal of Infrastructure, Policy and Development, 8, 2024, https://doi.org/10.24294/jipd.v8i8.5321. (c)

Kou G., Xu Y., Yi Peng, Shen F., Chen Y., Chang K., Kou S., Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection, Decision Support Systems, 140, 2021, https://doi.org/10.1016/j.dss.2020.113429.

Letizia E., Lillo F., Corporate payments networks and credit risk rating, EPJ Data Sci., 8, 2019, pp. 8-21, https://doi.org/10.1140/epjds/s13688-019-0197-5

Liang D., Lu C., Tsai C., Shih G., Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study, Eur. J. Oper. Res., 252 (2), 2016, pp. 561-572, https://doi.org/10.1016/j.ejor.2016.01.012

Liang D., Tsai C., Wu H., The effect of feature selection on financial distress prediction, Knowledge-Based Systems, 73, 2015, pp. 289.

Logistic Regression in Python, sklearn.linear_model.LogisticRegression — scikit-learn 1.3.2 documentation

Lohmann Ch., Möllenhoff S., How do bankruptcy risk estimations change in time? Empirical evidence from listed US companies, Finance Research Letters, Volume 58, Part B, 2023, https://doi.org/10.1016/j.frl.2023.104389.

Mateika, H., Jia, J., Lillard, L., Cronbaugh, N., & Shin, W., Fallen angel bonds investment and bankruptcy predictions using manual models and automated machine learning, 2022, arXiv preprint arXiv:2212.03454.

Mattos E., Dennis S., Bankruptcy prediction with low-quality financial information, Expert Systems with Applications, 237, 2024, https://doi.org/10.1016/j.eswa.2023.121418.

Mohtasham, F., Pourhoseingholi, M., Hashemi Nazari, S.S. et al. Comparative analysis of feature selection techniques for COVID-19 dataset. Sci Rep 14, 2024, https://doi.org/10.1038/s41598-024-69209-6

Noga, T., & Adamowicz, K., Forecasting bankruptcy in the wood industry. European Journal of Wood Products, 79, 2021, pp. 735–743, https://doi.org/10.1007/s00107-020-01620-y

Odom M., Sharda R., A neural network model for bankruptcy prediction, Proceedings of the IJCNN International Joint Conference on Neural Networks, IEEE, 1990, pp. 163-168

Ohlson J., Financial ratios and the probabilistic prediction of bankruptcy, Journal of Accounting Research, 18 (1), 1980, pp. 109-131

Pereira J.M., Basto M., Silva A., The logistic lasso and ridge regression in predicting corporate failure, Procedia Economics and Finance, 39, 2016, pp. 634.

Qu Y., Quan P., Lei M., Shi Y., Review of bankruptcy prediction using machine learning and deep learning techniques, Procedia Computer Science, 162, 2019, pp. 895-899, https://doi.org/10.1016/j.procs.2019.12.065.

Ross BC. Mutual information between discrete and continuous data sets. PLoS One., 19, 2019, https://doi.org/10.1371/journal.pone.0087357.

Ruxanda, G., Zamfir, C., & Muraru, A., Predicting financial distress for Romanian companies. Technological and Economic Development Economy, 24(6), 2018, 2318–2337. https://doi.org/10.3846/tede.2018.6736

Salehi, M., & Pour, M. D., Bankruptcy prediction of listed companies on the Tehran Stock Exchange. International Journal of Law and Management, 58(5), 2016, 545–561. https://doi.org/10.1108/IJLMA-05-2015-0023

Sarkar S., Sriram R., Bayesian models for early warning of bank failures, Management Science, 47 (11), 2001, pp. 1457-1475

Shen F., Liu Y., Wang R., Zhou W., A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment, Knowledge-Based Systems, 192, 2020, https://doi.org/10.1016/j.knosys.2019.105365.

SVM In Python, sklearn.svm.SVC — scikit-learn 1.3.2 documentation

Szeghalmy S, Fazekas A. A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. Sensors (Basel)., 23(4), 2023, https://doi.org/10.3390/s23042333.

Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K., Financial crisis prediction model using ant colony optimization. International Journal of Information Management, 50, 2020, pp.538-556.

Veganzones, D. and Severin, E., "Corporate failure prediction models in the twenty-first century: a review", European Business Review, 33 (2), 2021, pp. 204-226. https://doi.org/10.1108/EBR-12-2018-0209

Voda, A. D., Dobrotă, G., Țîrcă, D. M., Dumitrașcu, D. D., & Dobrotă, D., Corporate bankruptcy and insolvency prediction model . Technological and Economic Development of Economy, 27(5), 2021, pp. 1039-1056, https://doi.org/10.3846/tede.2021.15106

Wang H, Liu X, Undersampling bankruptcy prediction: Taiwan bankruptcy data. PLoS ONE 16(7), 2021, https://doi.org/10.1371/journal.pone.0254030

Yuxia S., Congyuan Y., Zhiya L., Yanting T., Initiative for China to establish a dual model of mixed corporate governance on bankruptcy reorganization: An empirical analysis based on 93 listed companies, Heliyon, 8(12), 2022, https://doi.org/10.1016/j.heliyon.2022.e12007.

Zelenkov Y., Fedorova E., Chekrizov D., Two-step classification method based on genetic algorithm for bankruptcy forecasting, Expert Systems with Applications, 88, 2017, pp. 393.

Zhang, W., Machine Learning Approaches to Predicting Company Bankruptcy. Journal of Financial Risk Management, 6, 2017, pp. 364-374, https://doi.org/10.4236/jfrm.2017.64026.

Zhao J., Ouenniche J., Smedt J., Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction, Machine Learning with Applications, 15, 2024, https://doi.org/10.1016/j.mlwa.2024.100527.

Zhou L., Lu D., Fujita H., The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches, Knowledge-Based Systems, 85, 2015, pp. 52-61.