University of Nottingham

UK | CHINA | MALAYSIA

Bankruptcy prediction using big data techniques.

Aditya Kumar Sasmal

Student ID: 20478461

Submitted September 2023, in partial fulfilment of the conditions for the award of the degree of MSc. Computer Science with Artificial intelligence

I declare that this dissertation is all my own work, except as indicated in the text.

Abstract

Predicting corporate bankruptcy holds immense practical significance, prompting the exploration of big data techniques for this purpose. Strategic implementations of machine learning models using big data techniques utilizing publicly available datasets from American and Taiwanese companies. Leveraging financial statements from over 5,000 firms, we employ Spark-driven machine learning algorithms to ensure scalability. This effort is underpinned by the backdrop of the 2008 global housing market collapse, which precipitated widespread bankruptcy in both Taiwan and the US.

Throughout the analysis, we employ exploratory data analysis, generating various graphs including correlation plots, heat maps, and box plots. These visualisations serve as the foundation for our subsequent studies. to ready the datasets for modelling, a methodical pre-processing regimen is applied. This involves Hyperparameter tuning for best parameters estimation, SMOTE to address imbalances, and principal component analysis (PCA) for dimensionality reduction.

In tandem with our predictive modelling, I've adopted Speedups, Scaleups, and Sizeups metrics to holistically assess model performance. These metrics allow us to gauge how well our models perform under varying data volumes, ensuring their scalability and adaptability as we deal with large datasets. I've evaluated the models using the F1 score and ROC-AUC metrics. Augmenting this quantitative assessment, classification reports and informative graphs provide a detailed understanding of each model's strengths and weaknesses.

Overall, this study propels the realm of bankruptcy forecasting forward through its holistic methodology for predicting bankruptcy. Moreover, the cross-regional comparison enhances our understanding of bankruptcy prediction factors across diverse economies. This research contributes to the broader domain of financial risk assessment and informs strategic decision-making for investors and financial professionals. By meticulously analysing these datasets, we uncover patterns indicative of imminent bankruptcy strategies.

Keywords: Bankruptcy prediction; Predictive modelling; Principle component analysis (PCA); Spark framework.

GitHub Link for the research: https://github.com/Aditya2809-2000/MSc_project

Acknowledgements

First and foremost, I would like to express my deep gratitude for my hardworking supervisor, Rebecca Tickle. Her unflagging support, wisdom and professional expertise have been essential throughout this academic journey. The direction by Rebecca and enlightening feedback enhanced greatly the effectiveness of this project.

Moreover, I want to express my deep thanks to the University of Nottingham for creating an environment favourable to learning and research. The resources, facilities, and academic community all played important roles in shaping this research. I am extremely appreciative for the opportunities and experiences that the university has provided me, as well as its consistent commitment to academic success. To family members I am deeply grateful for your strong support, boundless cares and love, and endless patience during this endeavour. Your unwavering belief in my abilities serves as a constant source of inspiration.

I also want to express deepest gratitude to all friends who provided moral support, companionship, and moments of relaxation during the challenging aspects of my research work. Your presence made this journey more enjoyable and unforgettable.

Finally, I extend my thanks to all individuals, institutions, and resources that contributed to success for this research. Your support and knowledge opportunities gained along the way are really appreciated.

Table of Contents

Table of Figures

Table of Abbreviations

SMOTE	Synthetic Minority Over-Sampling Technique
PCA	Principal Component Analysis
MSO-PSO	Multi-Objective Algorithm-Particle Swarm Optimization
FFNN	Fast Forward Neural Network

Chapter 1: Introduction

In this modern era of rapidly changing economic landscapes and volatile financial markets, bankruptcy continues to be an ever-relevant and essential aspect of financial management. From the global financial crisis to more recent economic downturns, bankruptcy has been a crucial mechanism to help individuals and companies weather economic storms and find a way back to solvency.

Bankruptcy is an important concept with significant implications for both individuals and businesses. Its importance lies in its role as a legal process that addresses financial distress and insolvency.[1] Bankruptcy is a serious financial event that can have a significant impact on a company's stakeholders, including creditors, employees, and shareholders. It is a process that allows debtors to seek relief from overwhelming debt and provides a structured framework for the fair treatment of creditors[2]. Early detection of bankruptcy can help companies take steps to avoid it, such as restructuring their debt or selling assets. Bankruptcy prediction is the process of identifying companies that are at risk of going bankrupt[3]. This is a critical task for businesses, investors, and creditors, as it can help them to avoid losses. In addition to financial data, some bankruptcy prediction models also use non-financial data, such as the company's management team, industry, and economic conditions[4]. Traditionally, bankruptcy prediction has been done using statistical methods. However, in recent years, there has been a growing interest in using machine learning models for bankruptcy prediction. This is because machine learning models can learn from large amounts of data and identify patterns that would be difficult to detect using traditional methods.[5]

The primary aim of this research is to conduct a comprehensive cross-country bankruptcy prediction analysis, focusing on Taiwan and the United States, by leveraging the available Taiwan bankruptcy dataset and the American bankruptcy prediction dataset from Kaggle. Furthermore, using feature engineering , I have made a dataset of Taiwan-American combined to gain insights of 2008 crisis of cross-regional economies. Inferring bankruptcy prediction patterns in Taiwan and the United States to identify any significant differences and similarities in the underlying financial indicators and risk factors leading to bankruptcy in the two countries.

Link for American dataset:

https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset

Link for Taiwan dataset:

https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

The bankruptcy trends in Taiwan and the United States reveal significant waves of economic turmoil during specific periods. In Taiwan, corporate bankruptcies surged twice: firstly during the dotcom bubble crash in the early 2000s and later around 2008-2009 due to the global financial crisis[1]. The construction, manufacturing, and retail sectors bore the brunt of these downturns. Similarly, the US experienced major bankruptcy waves in the early 1990s, early 2000s, and the 2008-2010 period. The 2008 financial crisis saw prominent companies like Lehman Brothers, GM, Chrysler, and Washington Mutual succumb to bankruptcy[6]. These financial shocks impacted various sectors including finance, automotive, retail, and energy. While the datasets likely encompass bankruptcy cases from these tumultuous periods, the specific years for individual companies remain unspecified. Consequently, the predictive models will strive to forecast bankruptcy regardless of the precise timeframe, focusing on the overarching patterns of economic distress[7]. Financial landscape is becoming increasingly complex. This is due to factors such as globalization, technological change, and regulatory uncertainty. As a result, it is becoming more difficult to predict the financial health of companies.

The availability of big data is growing rapidly. This is due to the increasing use of sensors, social media, and other data collection technologies. Big data can provide insights into the behaviour of companies and their customers that would not be possible to obtain using traditional methods[8]. There is a growing demand for bankruptcy prediction tools.

Data science and machine learning have shown promising results in predicting bankruptcy, enabling better risk management and decision making. With large volumes of financial and business data available, advanced analytics techniques can help identify patterns and insights to develop accurate predictive models. This research aims to leverage data science approaches for improved bankruptcy prediction in the finance domain. In addition to machine learning models, big data is also playing a role in bankruptcy prediction[9]. The combination of machine learning models and big data is creating new opportunities for the study in bankruptcy prediction. By leveraging large datasets, machine learning algorithms can identify patterns, correlations, and indicators of financial distress, allowing financial institutions, investors, and policymakers to make informed decisions and take proactive measures to mitigate bankruptcy risks[10].

This study analysed bankruptcy datasets from Taiwan and USA containing financial ratios of companies. Feature selection in bankruptcy management refers to the process of selecting the most relevant and informative features (variables) from a dataset to build accurate and efficient bankruptcy prediction models. Feature selection is performed to improve the performance of the prediction models by eliminating irrelevant or redundant features and focusing on those that have the most significant impact on bankruptcy outcomes. Class disparity frequently arises in bankruptcy prediction datasets, with the count of bankrupt companies being notably fewer than their solvent counterparts. This

imbalance can lead to biased models and inaccurate predictions, as the model may favour the majority class[4]. We implemented SMOTE which can help to capture the underlying patterns and relationships within the imbalanced data. To improve model performance, key steps like handling class imbalance, feature selection using PCA, hyperparameter tuning were implemented[11].

Appropriate evaluation metrics like ROC-AUC, precision, recall and F1-score were used to assess the models. Spark was leveraged for scalable data processing to train robust models on large datasets. Score evaluation of machine learning models in bankruptcy prediction is a crucial step to assess the model's performance and its ability to classify companies as either bankrupt or non-bankrupt[12]. Evaluation of predictive performance of the developed models on their respective test datasets. Use appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the model's effectiveness in distinguishing between bankrupt and non-bankrupt companies in Taiwan and the United States.

Many existing bankruptcy prediction studies focus on traditional financial ratios and variables derived from company financial statements which lacks in accurate risk assessments and proactive decision-making. Big data techniques, such as Apache Spark, are increasingly being utilized in prediction to handle voluminous datasets with the potential for regular updates in the dataset[8]. Having large datasets enables us in scalability and train more robust and accurate bankruptcy prediction models.

It's essential to consider the specific context and objectives of bankruptcy prediction while selecting the appropriate evaluation metric. For instance, if minimizing false negatives (missing bankruptcies) is more critical, recall might be emphasized. On the other hand, if avoiding false positives (misclassifying non-bankrupt companies as bankrupt) is a priority, precision should be given more weight. The choice of evaluation metric should align with the desired risk tolerance and business needs, striking the right balance between different performance measures to create an effective bankruptcy prediction model. Regularly monitoring and evaluating the model's performance are essential to ensure its reliability and accuracy in practical applications.

In summary, this research addresses the multifaceted problem of bankruptcy prediction, extensive exploratory data analysis, integrating big data techniques, dataset diversity, model evaluation, and scalable implementation to offer insights into the intricate world of corporate finance. The key contribution of this project are:

- **Extensive Exploratory Data Analysis (EDA):** Through rigorous EDA, we uncover hidden patterns, trends, and critical factors that underlie the dynamics of bankruptcy prediction. This foundational step not only enhances our understanding but also lays the groundwork for robust predictive models.

- **Machine Learning Pipelines:** This pipeline is meticulously designed to pre-process data, engineer informative features, and employ advanced algorithms.
- **Evaluations, Big Data Scalability, and Effectiveness:** We take a holistic approach to model evaluation, going beyond traditional metrics. Alongside comprehensive evaluations, we place significant weight on the scalability of our models in big data.

In this study, we will have an extensive research on the bankruptcy predictive models tested on Taiwanese and American bankruptcy prediction using big data techniques.

Chapter 2: Background

2.1 Problem description

Predicting corporate bankruptcy stands as a critical financial analysis challenge, now approached through the lens of big data techniques using American and Taiwanese datasets available on Kaggle. This research develops robust models that accurately anticipate the likelihood of financial distress and potential bankruptcy by leveraging diverse financial attributes. It serves as a crucial tool for investors, creditors, and policymakers, enabling them to mitigate substantial financial losses and prevent economic instability.[1] While previous studies have addressed bankruptcy prediction, the integration of big data techniques alongside the amalgamation of American and Taiwanese datasets introduces a unique set of complexities to the problem.

Bankruptcy prediction poses multifaceted challenges that can be addressed through innovative big-data techniques. Integrating heterogeneous data sources requires thoughtful pre-processing for consolidation. Imbalanced classes with few bankruptcies need balanced representation. High-dimensional attributes lead to noise and multicollinearity, dimensionality reduction while retaining explanatory factors[3]. Rigorous out-of-sample validation across time ensures generalizability. Ensemble models combining algorithms improve performance[5]. Interesting big data developments include utilizing large datasets exceeding traditional limitations, leveraging unstructured data like management commentary, and employing big data techniques for enhanced scalability. Cloud computing enables scaling the data and model complexity. Overall, big data methodologies help overcome key challenges in bankruptcy prediction to develop accurate early warning systems with profound real-world impact. However, thoughtful implementation is crucial for robust and generalizable results[13].

A critical challenge lies in effectively evaluating model performance given the rarity of bankruptcy events amidst abundant solvent instances such as F1 and ROC-AUC, given that traditional accuracy measures might be deceptive. Beyond mere predictions, it's vital to extract pragmatic insights, with methods like Principal Component Analysis illuminating influential features and relationships. The algorithmic choice becomes a balancing act between model interpretability and performance, while simpler algorithms like logistic regression ensure clarity, complex models like gradient boosting may offer enhanced accuracy, albeit with potential interpretability trade-offs[14]. As the magnitude of datasets increases, platforms like Spark shine, providing scalability without incurring excessive computational overhead. Recent integrations of metrics like size-up, scale-up, and speed-up further refine this scalability assessment, revealing a model's adaptability to evolving data sizes and resources,

thus underscoring the multifaceted intricacies of efficient and insightful bankruptcy prediction in contemporary research[15].

2.2 Dataset Description

This research explores corporate bankruptcy prediction using big data techniques, utilizing datasets from both American and Taiwanese companies, accessible through public repositories. The American dataset comprises 8,262 publicly listed companies from the New York Stock Exchange and NASDAQ, spanning 1999-2018. It employs a unique labelling strategy, marking the year preceding bankruptcy as "Bankruptcy" (1) and non-bankrupt years as "Operating" (0), ensuring data integrity.

The American dataset offers 87 financial attributes, encompassing various ratios like ROA and Debt Ratios, providing insights into financial health. The Taiwanese dataset covers 1999-2009, sourced from the Taiwan Economic Journal, with 96 features tailored to Taiwan's economic landscape.

For the third dataset, a hybrid approach was employed, meticulously amalgamating features from both the Taiwanese and American bankruptcy prediction datasets for the year 2008. This synthesis was achieved through a deliberate selection process informed by domain knowledge and expertise achieving 19 synthetic attributes. This resulted in a consolidated dataset that not only represents the commonalities between different geographical financial landscapes but also ensures that the features encapsulate the most relevant indicators of bankruptcy, refining the quality and potential predictive accuracy of subsequent analytical models.

This research underscores the power of data science and domain expertise in solving bankruptcy prediction challenges with applications in auditing, credit scoring, and risk analysis. Rigorous evaluation ensures model robustness, making the models universally applicable and fostering innovation in computational finance. The Kaggle data serves as a valuable resource for future research.

2.3 Literature review

2.3.1 Bankruptcy prediction

Bankruptcy prediction holds paramount significance in the realms of finance, economics, and corporate decision-making due to its profound implications for various stakeholders. The forecasting of bankruptcy is a critical real-world concern. It serves as a crucial tool for investors, creditors, and policymakers, enabling them to mitigate substantial financial losses and prevent economic instability.[1] The ability to accurately predict bankruptcy is particularly vital in financial markets, where investments and lending decisions can have far-reaching consequences. When a company faces bankruptcy, it signifies a severe deterioration in its financial health and operational viability. Investors

can suffer significant losses, creditors may face defaults, and the broader economy can be negatively impacted as jobs and economic activity are disrupted. Moreover, bankruptcy prediction goes beyond financial markets; it extends its influence to internal management decisions, auditor's assessments, and public authorities' regulatory actions. Within companies, accurate bankruptcy prediction aids management in proactive financial planning and risk management. Auditors rely on these predictions to assess the integrity of financial statements and to provide early warnings if a firm's financial health is deteriorating. Public authorities and regulators use bankruptcy prediction models to monitor the stability of financial markets and enforce regulations that safeguard the interests of investors and the overall economy. In an era marked by increased complexity in financial transactions and global economic interconnectedness, the ability to foresee financial distress, from temporary cash flow difficulties to bankruptcy, is an indispensable tool for maintaining economic stability and promoting responsible corporate behaviours.

The challenge of bankruptcy prediction in the realm of data science is a non-trivial endeavour, characterized by several formidable complexities. One of the foremost challenges is the inherent data imbalance in bankruptcy datasets, where the number of companies that go bankrupt, the minority class, is significantly dwarfed by the majority class of financially stable companies. This imbalance poses a substantial hurdle as it can lead to biased models that tend to favour the majority class, often resulting in the under-identification of bankrupt companies, which are of primary interest in this context. Moreover, the high dimensionality of financial datasets, which encompass numerous financial ratios, market-based indicators, and economic metrics, introduces additional layers of intricacy.[16] This necessitates meticulous feature selection or dimensionality reduction techniques to mitigate overfitting and select the most informative features for prediction. Furthermore, the temporal nature of financial data adds a layer of complexity, requiring models to capture and leverage temporal dependencies and trends in financial indicators, thereby enhancing predictive accuracy[17]. Additionally, the real-world implications of incorrect bankruptcy predictions are profound, where false positives may trigger unwarranted financial distress for companies, and false negatives can lead to substantial financial losses for investors and creditors. This underscores the gravity of achieving high prediction accuracy and reliability. The adoption of more complex machine learning models, while promising, introduces its own set of challenges, including hyperparameter tuning and the risk of model overfitting. The task of interpreting these complex models further compounds the complexity, as stakeholders often require models that can elucidate the rationale behind specific predictions. Moreover, the dynamic nature of economic conditions and industry-specific factors adds an additional layer of intricacy, demanding models that can adapt to evolving financial landscapes. In sum, the multifaceted nature of bankruptcy prediction necessitates a comprehensive approach, encompassing meticulous data pre-processing, expert-driven feature engineering, judicious model selection, and rigorous evaluation techniques that

go beyond standard data science practices to ensure the development of reliable and actionable bankruptcy prediction models.

2.3.2 Machine learning in bankruptcy prediction

Machine learning techniques are a significant advancement in applied mathematics and have far-reaching effects on how categorization issues are resolved. This study evaluated the precision of five machine learning models for bankruptcy prediction: support vector machine, J48 decision tree, logistic model tree, random forest, and decision forest. Predicting bankruptcy is crucial when making financial decisions. Logistic regression is commonly used as a benchmark to compare against more advanced algorithms [17]. Machine learning methods like logistic regression, neural networks, support vector machines, and ensemble methods have been applied for bankruptcy prediction. These can handle the high dimensionality and non-linear relationships in financial data. Various machine learning models such as decision tree, XGBoost, Logit analysis, Probit, MOA-PSO algorithms are implemented to improve the accuracy of predictive modelling. Support vector machines, neural networks and ensemble methods like random forests tend to provide higher accuracy.[18]

Machine learning models like neural networks, support vector machines, random forests and gradient boosting trees are being widely used for bankruptcy prediction due to their high accuracy compared to traditional statistical models. However, these machine learning models act like black boxes and lack interpretability. The authors apply a technique called Local Interpretable Model-Agnostic Explanations (LIME) to generate instance-level feature importance explanations for the models.[19] They found the feature importance rankings from LIME were largely consistent with built-in feature importance methods for tree-based models like XGBoost and LightGBM. LightGBM was found to have more stable feature importance across different predicted bankruptcy probability buckets compared to XGBoost. Discriminant analysis (DA) and Logit analysis (LA) were found to be slightly superior predictors to the Cox proportional hazard model [27]. Nevertheless, [33] argued that the Sensitive analysis approach was more natural, flexible, and appropriate and used more information in Business Failure prediction.

The outcomes demonstrate the superiority of FS-Boosting over conventional boosting techniques. The information including the firm's experiences that may cause bankruptcy in the future is also used to create a profile-based model for bankruptcy prediction.[20] The proposed model outperforms single statistical and machine learning models like multiple discriminant analysis, logistic regression, decision trees, and artificial neural networks, as well as ensemble models like boosting and bagging, according to the experimental results. When training the Fast Forward Neural Network (FFNN), the hybrid MOA-PSO (Multi-Objective Algorithm-Particle Swarm Optimization) algorithm successfully increases the effectiveness of the MOA and PSO algorithms by increasing both accuracy and training time per run. Although the discriminant analysis and linear regression model have become the most commonly used in bankruptcy prediction, their inherent drawbacks of statistical assumptions such as linearity, normality

and independence among variables have constrained both applications. Recent trends in the development of artificial intelligence have brought forth new alternatives in solving nonlinear problems[21].

Ensemble methods have emerged as a pivotal technique in bankruptcy prediction, serving to amplify the accuracy of predictions by combining multiple models to produce an optimized consensus output. Techniques like bagging, boosting, and stacking, all subsets of ensemble methods, have shown substantial promise in handling the challenges of bankruptcy prediction, such as class imbalance and feature diversity[22]. For instance, boosting algorithms like AdaBoost and Gradient Boosting Machines (GBM) iteratively adjust the weights of misclassified instances, making the model more attuned to intricate patterns associated with bankruptcy indicators. the implementation of ensemble methods in bankruptcy prediction resulted in a notable increase in prediction accuracy compared to traditional models, underscoring their importance in the domain[23].

However, it's important to note that no single machine learning algorithm consistently outperforms others across various datasets. Ensembles and hybrid models, which combine multiple algorithms, tend to deliver the best results in bankruptcy prediction. In the context of forecasting bankruptcy across different time periods, XGBoost emerged as the superior model based on accuracy scores and Area under ROC curve (ROC AUC). Furthermore, the study contextualized its findings within the COVID-19 pandemic, predicting 690 bankruptcies in 2020, reflecting a bankruptcy rate of 4.35% for all US-based companies, the highest observed in the past decade. This underscores the ever-evolving landscape of bankruptcy prediction, influenced by both financial and external factors[24].

2.3.3 Feature engineering in bankruptcy prediction

Feature selection and engineering play a crucial role in developing accurate and robust bankruptcy prediction models. Selecting the most relevant variables from the extensive set of financial ratios available can improve model performance and simplify interpretation[21]. Both filter methods like correlation analysis and wrapper methods like recursive feature elimination have been applied [2]. Domain expertise in finance is also key for engineering informative new features.[25] There are several advantages to be obtained by performing feature selection. A variety of feature selection methods have been proposed, which can be categorized into filter, wrapper, and embedded methods. Feature selection are based on filter- and wrapper-based methods.[26] Feature engineering is important to find the most weighted features for our model. New non-linear features like logarithmic transformations were also engineered based on financial insights to capture complex relationships.[27] Exploring the inclusion of relevant economic variables like interest rates, unemployment rate, GDP growth rate could be an area for future work.

The class imbalance in the datasets also needs special handling through techniques like synthetic minority oversampling [28]. Economic variables including Altman Z score, profitability, liquidity and solvency creates a healthy model for predicting bankruptcies. XGboost based approach is used for feature engineering [12]. Newly computed financial ratios were integrated, accompanied by varying sets of hidden layer nodes and hyper parameters to enhance predictions. Both RelieF and Pearson correlation, along with a LIME-based approach, were utilized to determine feature significance. The research considered two feature selection strategies: filter-based and wrapper-based methods. Additionally, the prediction methodologies employed both statistical and machine learning classification techniques.

As an outcome, the question of which type of prediction approach can offer the highest performance improvement by which sort of feature selection methods remains unresolved in bankruptcy prediction. Furthermore, our study did not concentrate on feature selection, which is a standard method in recent studies that investigate many variables. However, as is typical in bankruptcy research, only a small number of factors are available. Even though we used a large number of observations, the databases only provided a restricted number of variables. As a result, the impact of feature selection would be minimal in our study. Credit risk applications, in particular default prediction, should be researched further, especially in efforts to obtain models related to macroeconomic variables. [25]

2.3.4 Class imbalance in bankruptcy prediction

Class imbalance is addressed using balancing techniques such as SMOTE and random sampling techniques [13]. F2 measure is choose as a metric for imbalanced data, Monte Carlo resampling techniques are used to generate multiple sub samples to gain better accuracy [7]. Hyperparameter tuning were implemented and resulted that Adams algorithm is good at handling noisy and sparse gradients. Bagging, also known as ‘bootstrap aggregating’, is a technique involving independent classifiers that uses portions of the data and then combines them through model averaging, providing the most efficient results concerning a collection [29]. Bagging creates random new subsets of data through sampling, with replacement, from a given dataset, generating confidence-interval estimates[30]. The objective of bagging is to reduce the overfitting of a class within the model. Rather than using the collection to check if the model is over fitted, the training set is recombined to produce better classifiers. The boosting technique consists of the repeated use of a base prediction rule or function on different sets of the initial set. Boosting builds on other classification schemes and assigns a weight to each training set, which is then incorporated into the model[31]. The data are then reweighted. Boosting can apply the base classifier to find a model that better classifies the set, identified by a low error rate for the training set. However, the challenge of data imbalance in historical bankrupt company data remains a concern, leading to incorrect predictions and potential economic losses. Researchers have explored various

techniques, such as under-sampling and ensemble methods, to address data imbalance and enhance model performance.[5]

Bagging, short for bootstrap aggregating, is an ensemble technique that can provide competitive performance compared to more complex machine learning methods for financial distress prediction. Bagging involves creating multiple versions of a predictor model, such as a decision tree, trained on random subsets of the data sampled with replacement. The predictions from all the sampled models are combined through voting or averaging to produce the overall bagged model prediction. This technique reduces variance and helps avoid overfitting. Bagging shows reasonable performance and may provide an interesting alternative to the more computationally intense machine learning systems. Bagging stabilizes model predictions and works well with noisy financial data. Researchers found bagging effective at reducing errors for credit scoring models with inconsistent or incomplete data[32].While it is difficult to confirm a single preferred technique from the curves, bagging, boosting, and RF are the most promising candidates.[12] Adaptations to bagging such as worked examples learning may also improve distressed firm identification[33].Overall, bagging presents an interesting ensemble modelling approach for financial distress prediction, offering reasonable accuracy without high complexity.

2.3.5 Evaluation

The performance of bankruptcy prediction models is commonly evaluated using metrics adapted for the imbalanced class distribution, where solvent firms far exceed bankruptcies. The F1 score combines precision and recall into a harmonic mean, providing a balanced measure even with skew. However, it is sensitive to the decision threshold. AUC-ROC curves demonstrate model discrimination across different thresholds and are threshold-invariant[34]Higher accuracy scores and advancements in evaluation techniques such as ROC-AUC score, F1 and F2 score, TPR-FPR (True Positive Rate-False Positive Rate) lead to high reliability over these models.[8][9]

Classification reports detail additional metrics like precision, recall, and specificity for each class. However, these vary with chosen threshold and can be misleading with imbalance[35] No single metric paints a complete picture. Hence, it is recommended to evaluate models using a combination like AUC, F1, classification report, and calibration test[7]**.** The choice of evaluation metric will depend on the specific application of the bankruptcy prediction model. For example, a model that is used to make investment decisions may need to be more sensitive than a model that is used to make regulatory decisions.

2.3.6 Big data techniques in bankruptcy prediction

Traditional data approaches and systems have encountered limitations in storage and performance capacities, making it challenging to effectively analyse extensive financial datasets. Consequently, there's an imperative for a robust platform capable of conducting in-depth analytics and yielding reliable and precise outcomes for data extraction. In response to this growing demand, innovative big data platforms have emerged, offering cost-effective solutions for storage and distribution.[36]. Big data performance metrics include throughput, bandwidth, concurrency, latency, durability, and availability. Advanced technologies like Apache Spark and Hadoop have been employed to process and analyse voluminous data efficiently [14].

Throughput measures the amount of data that can be processed per unit of time. Bandwidth measures maximum data transfer capacity. Concurrency tracks the number of operations executing simultaneously. Latency measures delays in data transfers or query execution. Durability tracks data persistence and resistance to loss. High availability maximizes system uptime. Careful benchmarking, load testing, and monitoring of these metrics guides big data architecture choices to meet performance objectives[37]. These technologies enable distributed computing and parallel processing, allowing for faster and more scalable analysis of large-scale datasets [15]. Big data and machine learning can be used to build an updatable and appendable robust failure prediction system for the industry of any country with large records of data [16]. However, converting data to the key-value pair format of a big data Analytics database server (e.g. Hadoop) might pose some data integration challenges such as schema and semantic heterogeneity.[15] Optimizing data pipelines, using in-memory processing, and leveraging parallel computing can improve speed.[38]

Evaluation of spark machine learning models for accuracy, speedup, scaleup, and size up of these algorithms on different datasets and clusters. Big data systems must be able to handle large volumes of data, scale to meet increased demands, and process data quickly. These performance requirements are often referred to as the "three S's" - size up, scale up, and speed up.[39] The results show that both algorithms have high accuracy, but the increase in speedup is not linear. The scaleup of random forest reaches its peak when the number of nodes is 2, and after that, it decreases with the increase of the number of nodes. The size up of random forest increases sharply with the increase of the number of nodes. This study provides valuable insights into the performance of classification algorithms for big data and can be useful for researchers and practitioners working in this field.[39]

2.3.7 Key objectives

After conducting an extensive literature review of the research papers based on bankruptcy prediction using big data using various machine learning techniques, our central objectives revolve around the development of machine learning models for bankruptcy prediction using big data techniques, with a primary focus on achieving accurate predictions to empower investors, creditors, and policymakers in

making informed financial decisions and managing risks effectively. Scalability is a core concern, particularly in handling extensive financial datasets, and we address the challenge of data imbalance to ensure robust predictions for both bankrupt and non-bankrupt cases. We investigate the impact of feature engineering, specifically Principal Component Analysis (PCA), resampling techniques (SMOTE) on predictive performance, while comparative analyses between the three datasets which will provide insights into model capabilities and economic influences. Model interpretability is paramount, aiming to offer actionable insights to stakeholders, and scalability metrics are employed to assess model efficiency across varying dataset sizes. Our research showcases the real-world applicability of these models, enhancing financial decision-making and risk management in both markets, all while harnessing big data technologies like Spark for efficient processing of large-scale financial data.

Chapter 3. Methodology

The prediction of corporate bankruptcy is an important research area with significant real-world applications. This paper explores the use of big data techniques to develop models for predicting company failure on American and Taiwanese company's datasets available on Kaggle. Using financial data from more than 5,000 companies, Spark implementations for executing machine learning algorithms are implemented for scalability and noting observations of bankruptcy in Taiwan and American datasets due to global bankruptcy due to the Collapse of the Housing Market in 2008[40]. The models developed demonstrate high accuracy and have the potential to provide better results and scalability of bankruptcy prediction models.

The workflow of the methodology is as follows:

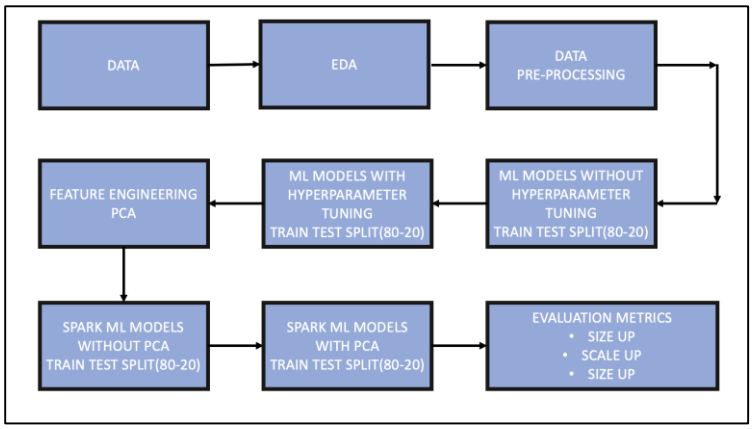

Figure 1. Workflow

3.1 Data Collection

The study utilizes publicly available datasets of American and Taiwanese companies, obtained from Kaggle.

In this research, two distinct datasets were employed to investigate bankruptcy prediction for American and Taiwanese companies. The Taiwanese dataset, spanning from 1999 to 2009, was sourced from the Taiwan Economic Journal. It includes a comprehensive set of financial attributes, encompassing measures of profitability, liquidity, solvency, and operational efficiency. The target variable, "Bankrupt?" classifies firms as either bankrupt (1) or non-bankrupt (0) based on Taiwan Stock Exchange regulations. The dataset underwent meticulous pre-processing, including data cleaning, categorical encoding, outlier handling, and addressing class imbalance using SMOTE and IQR. This dataset served as the basis for developing bankruptcy prediction models specifically tailored to the Taiwanese market.

In addition to the Taiwanese dataset, a novel dataset pertaining to American public companies listed on the New York Stock Exchange and NASDAQ was incorporated into the research. This dataset spanned from 1999 to 2018 and featured a total of 8,262 distinct companies. Bankruptcy classification in the American context was defined based on the legal provisions of Chapter 11 and Chapter 7 of the Bankruptcy Code, as monitored by the Security Exchange Commission (SEC). The dataset includes a wide array of financial and operational variables, including net income, total assets, market value, EBITDA, and more. The dataset was partitioned into training, validation, and test sets to facilitate model development and evaluation. This comprehensive dataset enabled the exploration of bankruptcy prediction for American companies within the context of financial and accounting attributes, paving the way for a thorough comparative analysis with the Taiwanese dataset.

The third dataset, meticulously crafted for our research, is a blend of American and Taiwanese bankruptcy data from the year 2008. This innovative amalgamation utilizes synthetic features, specifically designed to bridge the nuances of both regional datasets. The intent behind such a fusion is to explore the intricacies of bankruptcy within cross-regional boundaries, potentially uncovering patterns and insights that are otherwise obscured when examining the datasets in isolation. This unique approach underscores the importance of holistic data analysis in a globalized economy, emphasizing the interconnected nature of financial markets across continents.

3.2 Exploratory Data Analysis (EDA)

In the exploratory data analysis (EDA) conducted on all three bankruptcy datasets, several critical steps were taken to gain insights into the dataset's characteristics. Firstly, we performed data validation by checking for null values and duplicates, ensuring the dataset's cleanliness and integrity. Subsequently, we employed the describe function to examine the spread of data, providing valuable statistical summaries of the dataset's numerical attributes. We then investigated the class distribution to ascertain the number of bankruptcy and non-bankruptcy companies present, crucial for understanding the dataset's balance.

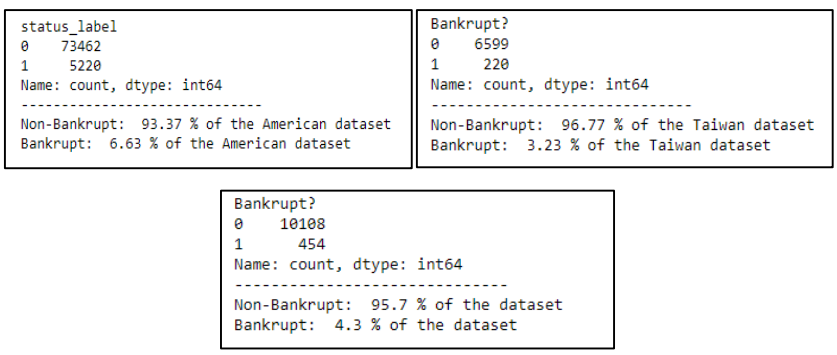

Figure 2. Number of bankrupt and non-bankrupt companies in Taiwan, American and combined dataset

Further, a distinction between categorical and numerical features needs to be established, enabling us to tailor our analysis appropriately. To visualize the data distribution, histograms were plotted for all features, offering a visual representation of the underlying data patterns. Additionally, a Spearman correlation heat map was generated to uncover potential relationships between variables, aiding in feature selection and understanding multicollinearity. In the quest to identify outliers, box plots were employed for each feature, revealing potential data points that deviate significantly from the normal. Distplots provided additional insights into feature distributions.

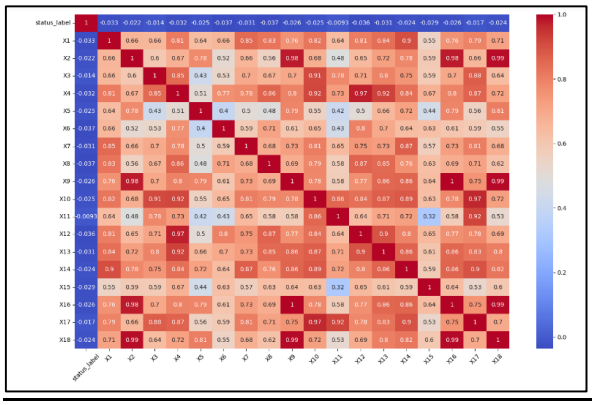

Figure 3. Heatmap of American bankruptcy dataset

Lastly, in the context of bankruptcy prediction, box plots were specifically created for features related to the target variable. This allowed us to visualize how these features vary with respect to bankruptcy status, facilitating the identification of potential discriminatory attributes. Collectively, these EDA steps provided a comprehensive understanding of the bankruptcy datasets, laying the groundwork for subsequent model development and analysis.

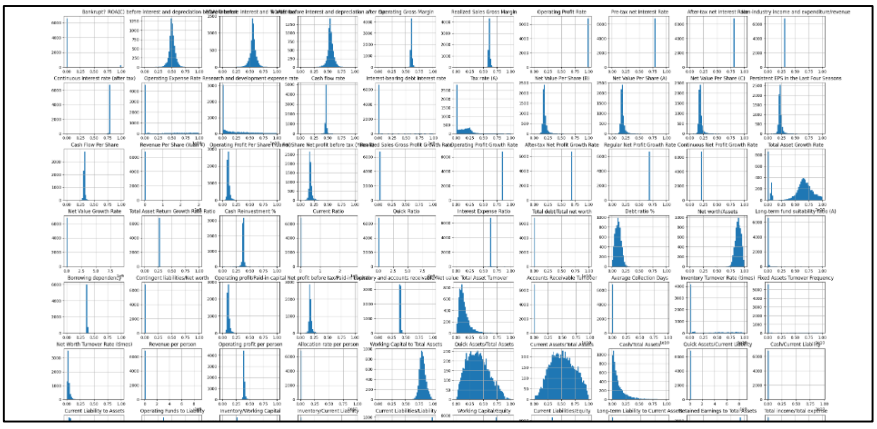

Figure 4. Histplot of Taiwan dataset

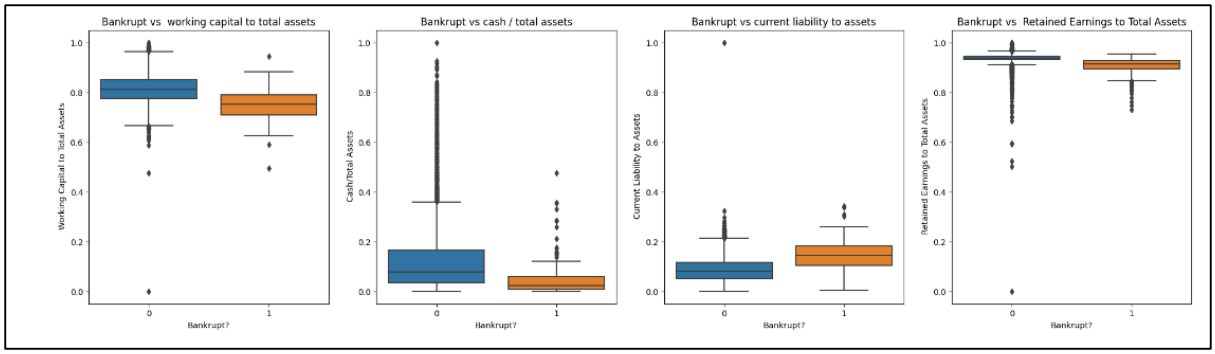

Figure 5. Boxplot of Taiwan dataset

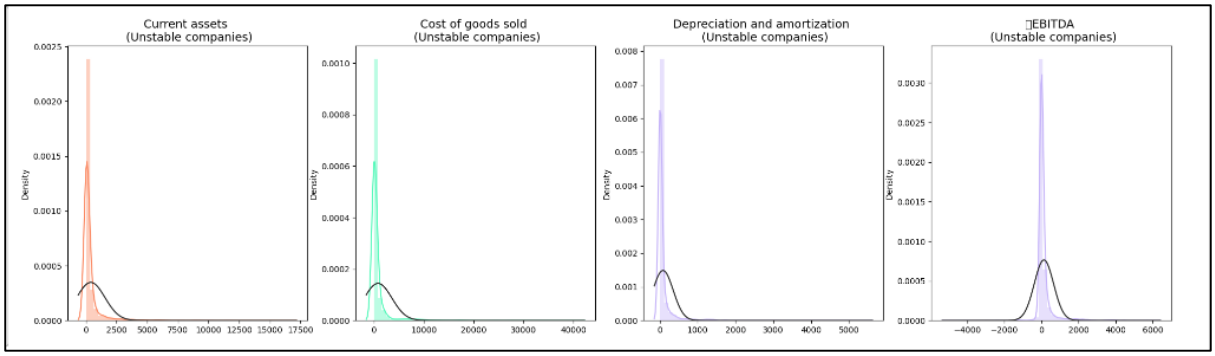

Figure 6. Distplot of American dataset

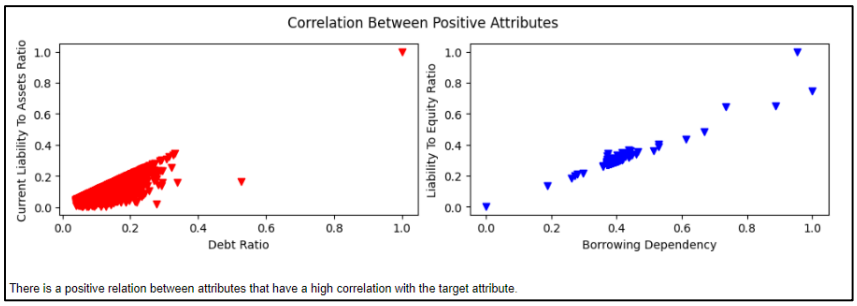

Figure 7. Scatterplot of positive attributes of Taiwan dataset

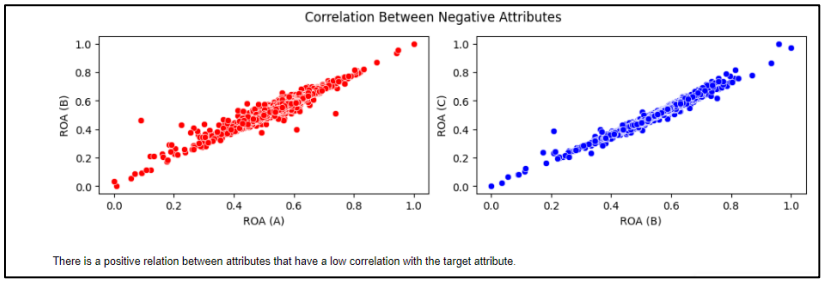

Figure 8 Scatterplot of negative attributes of Taiwan dataset

3.3 Key observations

A detailed analysis of the Taiwan dataset revealed intricate dynamics in bankruptcy prediction. While organizations registering losses consecutively for two years often have negative net income, only a small fraction declare bankruptcy, challenging the notion of using negative income as a sole predictor. Key financial attributes like "Debt Ratio %," "Current Liability to Assets," and "Current Liability To Current Assets" closely correlate with bankruptcy, emphasizing their pivotal role in financial health assessments. The contrasting interplay between various financial metrics, especially the inverse relationship between "Net Worth/Assets" and "Debt Ratio %," accentuates the nuanced nature of bankruptcy risk, underscoring the need for a holistic approach to risk management and prediction.

During the exploratory data analysis of Taiwan-American datasets, it was evident that the distribution of the features was non-normal. The datasets encompassed several financial attributes, shedding light on the companies' fiscal health and operations. Key variables included Current Assets/Total Assets, a ratio capturing short-term liquidity, and Current Liabilities/Equity, which reflects the company's leverage. Similarly, metrics like Net Income and EBITDA offered insights into a company's profitability, whereas Market Value, indicative of a firm's perception in the stock market, played a pivotal role. Moreover, attributes such as Total Receivables highlighted the amount owed by customers, while Gross Profit gave a snapshot of the company's profit after direct manufacturing and service costs.

In the American dataset, critical examination of certain financial indicators, namely Net Income (X6), which delineates the overall profitability after subtracting all expenses; Total Receivables (X7), the outstanding balance from goods or services provided; Market Value (X8), signifying the market capitalization in the stock context; EBIT (X12), representing earnings before interest and taxes; Gross Profit (X13), highlighting profitability post-direct manufacturing and sales costs; and Total Liabilities (X17), capturing the cumulative debts and obligations, illuminated their negative correlation with bankruptcy. Specifically, as these metrics trended negatively, the likelihood of a firm declaring bankruptcy increased.

3.4 Data Pre-processing

3.4.1 Data pre-processing for the Taiwan dataset

For the study on the Taiwanese bankruptcy dataset, we undertook several critical data processing steps to prepare the data for predictive modelling. Firstly, we conducted a thorough examination to identify positive and negative correlation features among the numerous financial attributes. Subsequently, we meticulously removed irrelevant features to streamline the dataset, ensuring that only the most informative variables were retained for analysis. Recognizing the challenge of data imbalance, we employed the Synthetic Minority Over-Sampling Technique (SMOTE) as part of our data pre-

processing strategy[27]. SMOTE generated synthetic instances of the minority class enhancing the representation of this critical class and addressing potential bias in our predictive models.

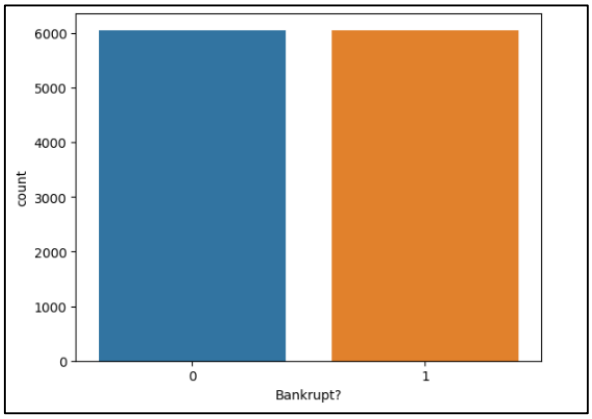

Figure 9 SMOTE graph of Taiwan dataset

Additionally, we implemented an outlier removal technique to enhance data quality, utilizing the Interquartile Range (IQR) method to systematically identify and eliminate outliers, with a particular focus on cases proximate to bankruptcy. This method systematically identified and eliminated outliers from each feature in the dataset, focusing particularly on cases close to bankruptcy. The result was a more robust dataset with reduced outlier presence, enhancing the overall quality of the data.

```
---------------------------------------------------------------------
Quartile 25: 0.565158395757604 | Quartile 75: 0.565724709506105
iqr: 0.0005663137485010239
Cut Off: 0.0008494706227515358
 Interest Coverage Ratio (Interest expense to EBIT) Lower: 0.5643089251348524
 Interest Coverage Ratio (Interest expense to EBIT) Upper: 0.5665741801288565
 Interest Coverage Ratio (Interest expense to EBIT) outliers for close to bankruptcy cases: 1421
---------------------------------------------------------------------
Quartile 25: 0.024476693570910098 | Quartile 75: 0.052837817459331596
iqr: 0.028361123888421498
Cut Off: 0.04254168583263225
 Equity to Liability Lower: -0.018064992261722153
 Equity to Liability Upper: 0.09537950329196385
 Equity to Liability outliers for close to bankruptcy cases: 549
---------------------------------------------------------------------
```

Figure 10. IQR of Taiwan bankruptcy dataset

Furthermore, we standardized the features to bring them to a consistent scale, allowing for fair comparisons and preventing features with larger magnitudes from dominating the modelling process. These comprehensive data processing steps collectively laid the foundation for our robust bankruptcy prediction analysis on the Taiwanese dataset.

3.4.2 Data pre-processing for American dataset

Data pre-processing for the American bankruptcy prediction dataset involved a comprehensive series of steps aimed at preparing the data for subsequent machine learning modelling. The dataset presented a significant challenge due to a substantial class imbalance, with 73,462 non-bankrupt samples and only 5,220 bankrupt samples. To address this issue and ensure that the machine learning models would not

be biased towards the majority class, Synthetic Minority Over-Sampling Technique (SMOTE) was employed. SMOTE generated synthetic samples for the minority class, effectively balancing the dataset.

Another critical aspect of data pre-processing was dealing with the skewness of the data. To enable machine learning algorithms to perform optimally, it was essential to standardize the features. Standardization transformed the data, giving it a mean of zero and a standard deviation of one. This process helped the algorithms converge efficiently and achieve better overall performance.

Outliers in the dataset posed a potential source of noise that could negatively impact model training. To mitigate this, an outlier removal technique was applied using the Interquartile Range (IQR) method.

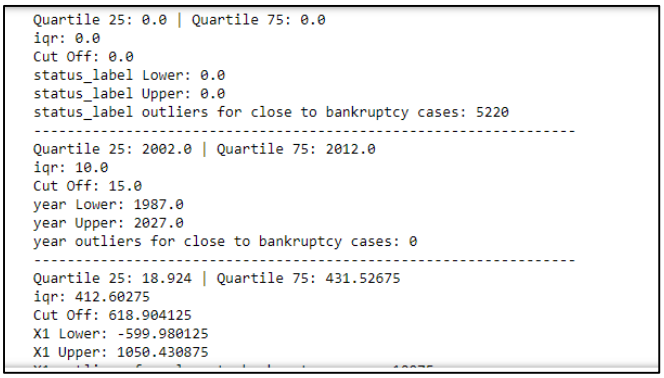

Figure 11. IQR of American dataset

Multicollinearity, a common issue in datasets with highly correlated features, was another concern addressed during data pre-processing. Several pairs of features exhibited strong correlations, leading to redundancy in the dataset. To mitigate this, feature engineering was employed. For example, it was observed that X2 "Cost of goods sold" and X9 "Net sales" were strongly correlated, and X13 "Gross profit" could be derived from these two features. Consequently, X13 was removed to reduce multicollinearity. Similar steps were taken to manage other correlated features, such as X9 "Net sales" and X16 "Total revenue," as well as X2 "Cost of goods sold" and X18 "Total Operating Expenses."

These efforts ensured that the models would be capable of making accurate bankruptcy predictions, benefiting stakeholders in the realm of finance and economics.

3.4.3 Data Pre-processing for Taiwan and American combined dataset

The data pre-processing steps employed for the combined dataset, which comprises both the American and Taiwanese bankruptcy prediction datasets, closely mirrored those used for the Taiwanese dataset. These critical steps aim to ensure data quality, enhance feature relevance, and address data imbalances. Initially, we conducted an extensive examination to identify positive and negative correlation features within the combined dataset, providing valuable insights into the interrelationships among the myriad financial attributes. Following this, we systematically eliminated irrelevant features to streamline the dataset, ensuring that only the most informative variables were retained for further analysis.

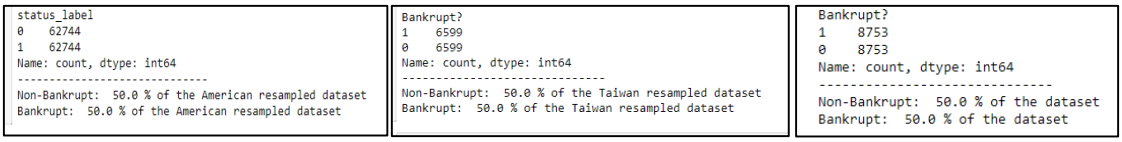

Figure 12. Resampled data of all 3 datasets

3.5 Model Selection

In our study, the selection of machine learning models for bankruptcy prediction is a pivotal decision guided by several key considerations. Firstly, **Logistic regression** is chosen for its simplicity and interpretability. This model provides valuable insights into the influence of individual financial features on bankruptcy prediction, facilitating a clear understanding of contributing factors.

1. **K Nearest Neighbours (KNN)** incorporated to capture local patterns within the data. This non-parametric algorithm has the potential to identify similar financial profiles among companies, thereby exploring the concept of proximity-based classification in the context of bankruptcy risk assessment.
2. **Support Vector Machine (SVM)** selected for its effectiveness in handling complex boundaries in high-dimensional feature spaces. By considering intricate relationships in financial data, SVM complements other models and may uncover patterns that contribute to bankruptcy prediction.
3. **Decision Trees** were included for their intuitive representation of complex decision-making processes. They have the capacity to identify critical decision points and create interpretable rules, enhancing transparency in our bankruptcy prediction models.
4. **Random Forest** as an ensemble method was employed to reduce overfitting and enhance generalization. Its combination of multiple decision trees contributes to improved predictive performance while maintaining model interpretability.
5. Lastly, **Gradient Boosting**, another ensemble technique, was introduced to sequentially refine weak learners into robust predictors. This addition aims to investigate whether boosting can further enhance the predictive capabilities of our models, particularly emphasizing challenging cases.

By incorporating this diverse array of machine learning models, our research facilitates a comprehensive exploration of the dataset's patterns and relationships. Each model's unique strengths and weaknesses allow us to compare their performance objectively, providing insights into the most effective algorithms for bankruptcy prediction within the specific context of our study. Random Forest and Gradient Boosting, mitigate overfitting concerns and enhance the generalizability of our models, a critical aspect when dealing with financial datasets.

3.6 Machine learning model Implementations

3.6.1 Model Implementation without hyperparameter tuning

I have assessed the performance of a diverse array of machine learning models across all three financial datasets American, Taiwanese, and the combined dataset of both regions from the year 2008. It's worth noting that this evaluation was carried out without any hyperparameter tuning, emphasizing the baseline performance of each model where the train test data split was 80-20 and only for the American dataset I have made the training data for American with data till 2015 and test above 2015. By applying these machine learning models to bankruptcy prediction, we aimed to comprehensively explore their intrinsic strengths and weaknesses in dealing with distinct financial environments and datasets. This unadulterated evaluation allowed us to establish a solid foundation for further refinement and optimization of the models, ensuring that the subsequent phases of our research build upon a robust understanding of their initial predictive capabilities across diverse financial landscapes.

Model: Random Forest

	precision	recall	f1-score	support
Not Bankrupt	0.97	0.92	0.94	1773
Bankrupt	0.92	0.97	0.95	1729
accuracy			0.95	3502
macro avg	0.95	0.95	0.95	3502
weighted avg	0.95	0.95	0.95	3502

Model: XGBoost

	precision	recall	f1-score	support
Not Bankrupt	0.97	0.90	0.93	1773
Bankrupt	0.91	0.97	0.94	1729
accuracy			0.94	3502
macro avg	0.94	0.94	0.94	3502
weighted avg	0.94	0.94	0.94	3502

Figure 13. Classification report of combined dataset without hyper parameter tuning

3.6.2 Model Implementation with hyperparameter tuning

To ensure optimal model performance, we've employed hyperparameter tuning for each of the following machine-learning algorithms. This comprehensive evaluation allowed us to fine-tune the models and optimize their parameters, ensuring they were well-suited to the unique characteristics and intricacies of each dataset. Here in all the models, there is 80-20 train test split. By subjecting our models to this extensive testing and optimization process, we aimed to deliver highly accurate and reliable bankruptcy prediction models that could effectively operate in diverse financial environments, providing valuable insights for stakeholders and decision-makers.

```
Model: Decision Tree
              precision    recall  f1-score   support

Not Bankrupt       0.96      0.90      0.93      1773
    Bankrupt       0.90      0.97      0.93      1729

    accuracy                           0.93      3502
   macro avg       0.93      0.93      0.93      3502
weighted avg       0.93      0.93      0.93      3502

Model: SVM
              precision    recall  f1-score   support

Not Bankrupt       0.96      0.90      0.93      1773
    Bankrupt       0.90      0.97      0.93      1729

    accuracy                           0.93      3502
   macro avg       0.93      0.93      0.93      3502
weighted avg       0.93      0.93      0.93      3502
```

Figure 14. classification report with hyper-parameter tuning of combined dataset

3.6.3 Model Implementation without Feature Engineering and big data techniques

The implementation of our selected machine learning models with default partitions in Apache Spark was a pivotal aspect of our research, enhancing both scalability and efficiency in handling the American, Taiwanese, and combined American-Taiwanese financial datasets. The choice of Spark was deliberate, given its distributed computing capabilities, which allowed us to seamlessly process large-scale datasets with ease. These machine learning models were meticulously chosen for their unique strengths, including simplicity, local pattern recognition, effective handling of complex decision boundaries, rule capture, ensemble-based reduction of overfitting, and sequential refinement of weak learners, all contributing to robust predictive capabilities. By conducting the initial model assessments without feature engineering, our aim was to establish a fundamental benchmark for their predictive performance. This approach provided valuable insights into their adaptability and applicability to bankruptcy prediction, not only across different financial contexts but also within the high-performance computing environment of Spark. These foundational assessments laid the groundwork for subsequent feature engineering and model refinement efforts, underlining the significance of leveraging Spark in enhancing the overall efficiency and scalability of our machine learning implementations.

3.6.4 Model implementation with feature engineering and Big data techniques

I have implemented Principal Component Analysis (PCA) as a feature engineering technique within my Spark execution. This choice was driven by the potential of PCA to enhance model performance through dimensionality reduction while preserving essential data variance and mitigating multicollinearity and noise. In our comprehensive evaluation, we assessed the predictive capabilities of six diverse machine learning models. These models were systematically implemented and benchmarked on all three datasets, allowing us to gain a nuanced understanding of their strengths and weaknesses in the context of bankruptcy prediction. Our research thus aimed to provide a thorough assessment of the impact of feature engineering, particularly PCA, on the performance and scalability of these machine learning models, contributing to a robust and informed approach to bankruptcy prediction.

```
Top 10 Important Features:
   Operating Gross Margin  Pre-tax net Interest Rate
0                0.247628                   0.308710  \
1                0.450948                   0.418438
2                0.197288                   0.169685
3                0.117820                   0.087023
4                0.019479                   0.008808
5                0.000103                   0.001278
6                0.009646                   0.005323
7                0.026056                   0.009577
8                0.073995                   0.028108
9                0.234276                   0.079159

   After-tax net Interest Rate  Operating Expense Rate  Current Ratio
0                     0.310153                0.136573       0.003619  \
1                     0.414078                0.142752       0.002278
2                     0.171710                0.013295       0.003154
3                     0.072590                0.079253       0.019670
4                     0.002708                0.387537       0.019754
5                     0.000114                0.097862       0.849120
6                     0.004551                0.110376       0.524176
7                     0.000267                0.168290       0.027794
8                     0.019274                0.601244       0.006956
9                     0.101210                0.019661       0.000776

   Operating profit/Paid-in capital  Net profit before tax/Paid-in capital
0                          0.009499                                0.152802  \
1                          0.050646                                0.194189
2                          0.268276                                0.021345
3                          0.038330                                0.009800
4                          0.100389                                0.049495
5                          0.014747                                0.017137
6                          0.002419                                0.017000
7                          0.779652                                0.355470
8                          0.336026                                0.169497
9                          0.394820                                0.855608
```

Figure 15. Top10 features of combined dataset using PCA

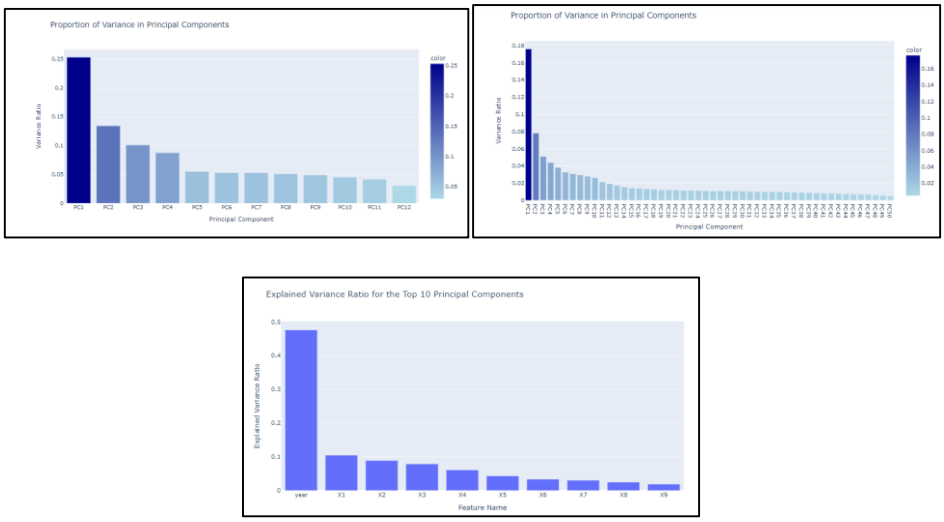

Figure 16. Important features of all three datasets

3.7 Evaluation Metrics

In our comprehensive evaluation of bankruptcy prediction models, we employed a diverse set of machine learning algorithms, the evaluation process encompassed multiple performance metrics, including the F1 score, ROC curves, and classification reports. The F1 score provided a holistic assessment of a model's precision and recall, crucial for gauging its ability to correctly identify both bankrupt and non-bankrupt cases. Furthermore, ROC curves visually represented the trade-off between a model's true positive rate and false positive rate, aiding in model selection and comparison. Lastly, the classification reports offered detailed insights into a model's performance, including metrics such as precision, recall, and accuracy. This multifaceted evaluation approach ensured a comprehensive understanding of each model's predictive capabilities across diverse financial datasets, facilitating data-driven decision-making in the realm of bankruptcy prediction.

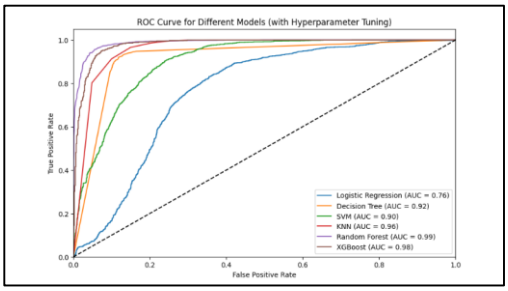

Figure 17.ROC graph of American dataset with hyperparameter tuning

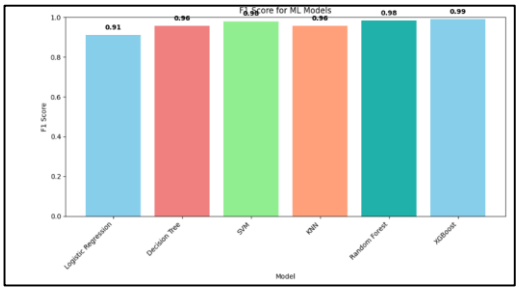

Figure 18. F1 score graph of Taiwan dataset with hyper parameter tuning

```
Decision Tree Speedups: [1.1893599354733868, 1.1220891851387949, 1.0463078870230225, 1.0491457668755966, 1.0765846653325033, 1.
032074089849005, 1.0320930115214435, 1.0150223238633456, 1.0351773677999636]
Decision Tree Scaleups: [1.1893599354733868, 0.5610445925693974, 0.34876929567434084, 0.26228644171889914, 0.21531693306650065,
0.1720123483081675, 0.14744185878877764, 0.1268777904829182, 0.1150197075333293]
Decision Tree Sizeups: [1.2380226481478152, 1.3903740344383186, 1.3944316648527832, 1.3830380856655635, 1.3180651772583623, 1.4
725403862863216, 1.4183629558836681, 1.46141593431425, 1.414962608830177]

Random Forest Speedups: [1.6190012423994857, 1.5556340454560407, 1.4614774536524147, 1.4608135873932688, 1.4663781668139588, 1.
4314532116509022, 1.4314806666265973, 1.4208656083886575, 1.4495770820703426]
Random Forest Scaleups: [1.6190012423994857, 0.7778170227280203, 0.48715915121747155, 0.3652033968483172, 0.2932756333627918,
0.23857553527515038, 0.2044972380895139, 0.1776082010485822, 0.16106412023003808]
Random Forest Sizeups: [1.2179976949692617, 1.6983274886950586, 1.723414734109577, 1.664328575648503, 1.6055901999142503, 1.765
9033999821772, 1.7564584099022744, 1.6363647635092164, 1.62996237098062]

Xgb Speedups: [3.4273698405360595, 3.229509127295366, 3.3273989774381936, 3.2643572827291134, 3.2767933973504215, 3.20589970326
36545, 3.1707182543186554, 3.211849641712567, 3.1765268367440576]
Xgb Scaleups: [3.4273698405360595, 1.614754563647683, 1.1091329924793978, 0.8160893206822784, 0.6553586794700843, 0.53431661721
06091, 0.45295975061695076, 0.4014812052140709, 0.3529474263048953]
Xgb Forest Sizeups: [0.9209978505885899, 0.9176368178821366, 0.933432969755231, 0.9504664894397351, 0.9470845890446942, 0.96324
83831487059, 0.934067222931054, 0.9506795288603782, 0.9475783365570599]
```

Figure 19. Speed up, Size up and Scale up estimations of Taiwan dataset

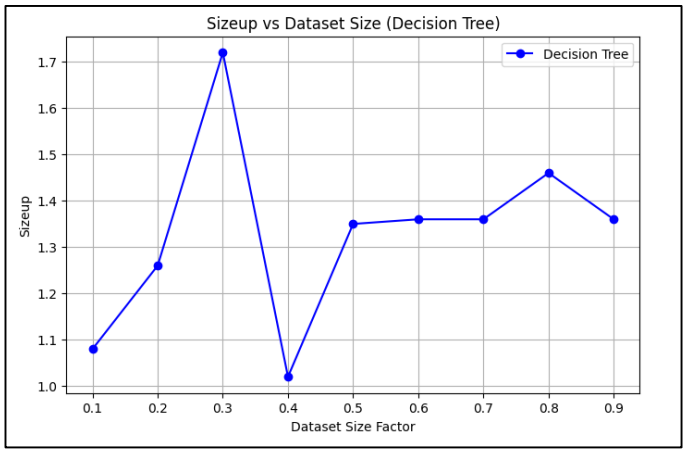

Figure 20. Decision tree size up graph of Taiwan dataset

3.8 Performance Metrics

The evaluation of our machine learning models for bankruptcy prediction encompassed a comprehensive analysis across three distinct financial datasets: American, Taiwanese, and a combined dataset representing both markets. In the context of scalability and size-up metrics, we examined the performance of three prominent models: Random Forest, XGBoost, and Support Vector Machine (SVM).

These metrics are particularly relevant in big data processing frameworks like Spark, where the goal is to efficiently process large-scale datasets across a distributed cluster of machines. When evaluating Spark applications, you want to ensure that they exhibit good speedup (processing speed improves as you add more nodes or CPUs), scaleup (the system maintains its performance with additional resources), and size up (performance remains steady as the dataset grows). These metrics help identify bottlenecks, optimize resource allocation, and ensure that your Spark applications can handle the challenges of big data processing effectively.

3.8.1 Speedups

Speedup is often expressed as a ratio, where a higher speedup value indicates better performance scalability. It's a measure of parallelization efficiency.

$$Speedup(m) = \frac{T_1}{T_m}$$

Figure 21. Speed up formula

In the context of speed-up metrics, we investigated how parallel processing and distributed computing impacted the efficiency of our machine learning models. By executing Random Forest, XGBoost, and SVM on a Spark platform, we assessed the speed at which these models' processed data and generated predictions. These metrics were instrumental in understanding the computational efficiency and responsiveness of our models, which is crucial for real-time bankruptcy prediction and decision-making. In summary, our evaluation encompassed a thorough analysis of model scalability, adaptability to growing data sizes, and computational efficiency, shedding light on their performance across American, Taiwanese, and combined financial datasets.

3.8.2 Scaleups

It quantifies how well a system maintains its performance as you increase its capacity. Scaleup is essential for ensuring that a system can handle larger datasets without a proportional decrease in performance.

$$Scaleup(m) = \frac{Speedup(m)}{m}$$

Figure 22. Scale up formula

In the scale-up assessment, we aimed to evaluate the models' scalability concerning growing data sizes. We progressively increased the dataset dimensions while monitoring the models' performance. This allowed us to discern their capacity to handle augmented datasets without compromising predictive accuracy. The scale-up analysis provided vital information for anticipating how the models would perform as financial datasets continued to expand over time.

3.8.3 Sizeups

Size up examines how the system's performance changes as the dataset size increases. It measures the system's ability to maintain performance as the data volume grows.

$$Sizeup(m) = \frac{T_{mDB}}{T_{DB}}$$

Figure 23. Size up formula

For the size-up evaluation, our aim was to assess how well these models adapted to larger datasets. We systematically increased the dataset size to gauge their efficiency and effectiveness. The results revealed valuable insights into the models' performance as data volume expanded. These insights were pivotal for understanding how our models would scale in real-world scenarios involving extensive financial data.

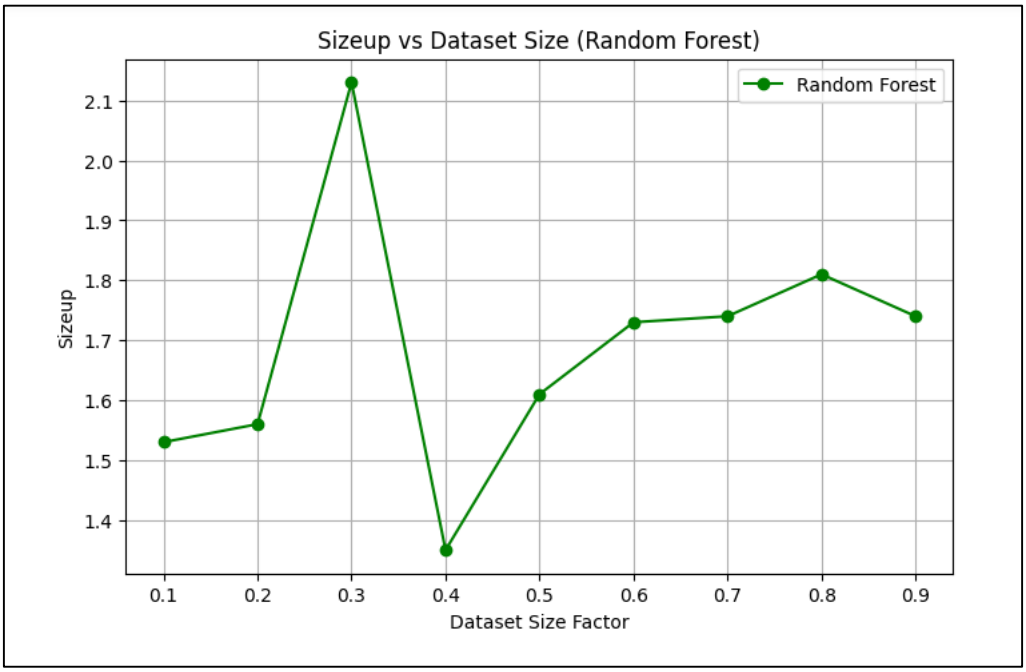

Figure 24. Random forest size up graph of Taiwan dataset

```
Decision Tree Speedups: [0.9088626289573527, 0.8262409972244911, 0.8342435513978419, 0.8291344212074737, 0.7961844097489507, 0.
8109055758936179, 0.8527229172420412, 0.795326097814654, 0.9572377712095304]
Decision Tree Scaleups: [0.9088626289573527, 0.41312049861224553, 0.27808118379928065, 0.20728360530186843, 0.1592368819497901
3, 0.135150929315603, 0.12181755960600589, 0.09941576222683175, 0.10635975235661449]
Decision Tree Sizeups: [0.9216607025283731, 1.0151058742587098, 0.9840739627723554, 1.0005645730416373, 1.0079967462218817, 1.0
495447119791734, 0.9900349912874516, 1.0078837091626163, 1.010169755814119]

Random Forest Speedups: [1.435998566215573, 1.4021536147233207, 1.3649606198963053, 1.3650828441425402, 1.3913038624037044, 1.3
711258929323438, 1.3598415002242736, 1.3600640110315216, 1.4245106623026342]
Random Forest Scaleups: [1.435998566215573, 0.7010768073616603, 0.45498687329876847, 0.34127071103563505, 0.2782607724807409,
0.22852098215539063, 0.19426307146061053, 0.1700080013789402, 0.15827896247807047]
Random Forest Sizeups: [1.3116825299443422, 1.3804784576690616, 1.3146946665187254, 1.3524703144050978, 1.4628344246616107, 1.4
414202393941338, 1.3570725323364197, 1.3659911177449349, 1.3477093189380696]

Xgb Speedups: [3.7857855552095145, 3.7638090872943586, 3.644887050301937, 3.682423800821488, 3.6131916002170263, 3.667346539036
348, 3.7368041885309, 3.7645387112001676, 3.716845595911755 7]
Xgb Scaleups: [3.7857855552095145, 1.8819045436471793, 1.2149623501006457, 0.920605950205372, 0.7226383200434052, 0.61122442317
27247, 0.5338291697901286, 0.47056733890002095, 0.4129828439901951]
Xgb Forest Sizeups: [1.0007328991457038, 1.014446552562663, 0.9964773479328517, 1.0234741370311338, 1.0128094682817321, 0.95001
91734716471, 0.9935284158158924, 1.003081842309508, 1.021286576929425]
```

Figure 25. Speed up, Size up and Scale up estimations of combined dataset

Chapter 4. Experimental study

4.1 Experimental setup

This study leveraged Apache Spark v3.4.1 for large-scale distributed data processing on cloud infrastructure. Spark's machine learning library MLlib provided scalable implementations of classifiers and feature engineering techniques. The experiments were implemented in Python 3.11.

Hyperparameter tuning for models like XGBoost, SVM and Random Forest were done using Python libraries like GridSearchCV and Hyperopt. PCA is used for feature engineering and dimensionality reduction. Functions made for Speedup and scaleup analysis determined cluster sizing and Parallelism to optimize job run times and resource usage for the large datasets. The processed data was persisted to CSV for ease of iterative modelling. The evaluation was done using AUC, accuracy, precision, recall and F1-score metrics. The end-to-end pipeline enabled rigorous experimentation to determine the optimal modelling approach.

The American and Taiwanese bankruptcy datasets were loaded into Spark Data Frames for in-depth analysis with default partitions. A synthesized dataset was crafted for the 2008 global crisis by harnessing common financial ratios between the datasets and incorporating features derived from domain expertise. For the American dataset, the model was trained using data up to the year 2015 with and without hyper parameter tuning. For subsequent modelling on the combined data, an 80:20 split was employed to separate the training and test sets. To address class imbalance, SMOTE was utilized to oversample the underrepresented bankruptcy class. To optimize the metrics for size-up, scale-up, and speed-up, the Spark partitions were configured to four. This configuration was chosen to enhance performance and ensure efficient data processing during the experiments.

The metrics analysis informs the configuration, resource allocation and parallelism settings to enable building and evaluating ML models on large bankruptcy prediction datasets. The tuned Spark jobs take advantage of distributed computing for big data analytics.

4.2 Results

During this research, we delved deep into bankruptcy prediction within both the American and Taiwanese and combined data domains. It became evident that certain machine learning algorithms stood out in their efficacy. Specifically, Random Forest and Gradient Boosting emerged as frontrunners in predictive accuracy. These algorithms, known for their ability to handle large datasets and complex interrelations between variables, showcased commendable predictive performance across various metrics. This underscores the significance of employing sophisticated ensemble methods when dissecting intricate financial datasets for bankruptcy prediction. The results from our analysis serve as

a testament to the potential of these algorithms in enhancing the accuracy and reliability of bankruptcy predictions.

4.2.1 Model evaluation of Taiwan bankruptcy dataset

The models underwent evaluation using several metrics, including the F1 score and ROC score, in addition to a detailed classification report. Additionally, performance metrics like size-up, scale-up, and speed-up were also employed to assess their efficiency and scalability.

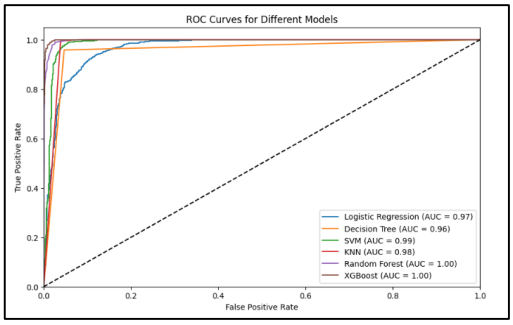

Figure 26. ROC graph of Taiwan dataset without hyper parameter tuning

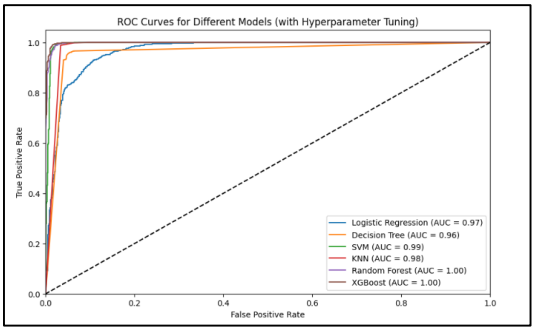

Figure 27. ROC graph of Taiwan dataset with hyper parameter tuning

The evaluations of Taiwan bankruptcy dataset using various machine learning models. Without hyper parameter tuning, XGBoost and Random Forest achieved perfect AUC scores of 1.0, followed by KNN at 0.98 AUC. Tuning improved SVM's AUC to 1.0 as well. For Spark ML models without feature engineering, XGBoost again had the best F1-score of 0.96, with other models around 0.91 F1. With principal component analysis with 85 features, XGBoost still maintained the top F1-score of 0.95, while other models dropped slightly in performance. Overall, XGBoost consistently emerged as the top performer for bankruptcy prediction on this American dataset, both before and after feature engineering using PCA.

The results exemplify how powerful ensembles like XGBoost can effectively leverage large financial datasets to create robust bankruptcy forecasting models

XGBoost outperformed both Random Forest and Decision Tree in the evaluation metrics of size up, scale up, and speed up.

4.2.2 Model evaluation American bankruptcy dataset

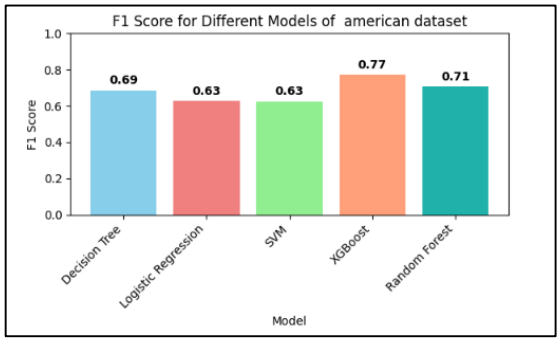

Figure 28. F1 score graphs of American dataset without feature engineering

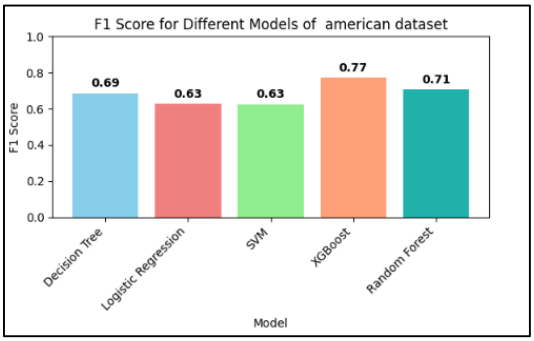

Figure 29. F1 score graphs of American dataset with feature engineering

In the analysis of the American dataset, we observed significant improvements in model performance when applying hyper parameter tuning. When using the training data from years prior to 2016, the F1 scores for the tuned models were as follows: Decision Tree achieved an F1 score of 0.86, Random Forest outperformed with an F1 score of 0.94, and XGBoost also achieved an F1 score of 0.86. Logistic Regression, without tuning, had a lower F1 score of 0.68. However, after hyper parameter tuning, XGBoost exhibited notable improvement, reaching an F1 score of 0.87 with an impressive ROC score of 0.94. When considering models without feature engineering, Decision Tree achieved an F1 score of 0.67, Random Forest scored 0.70, XGBoost performed the best with an F1 score of 0.77, and Logistic Regression had an F1 score of 0.63. With the inclusion of feature engineering, slight reductions in F1 scores were observed across the board: Decision Tree 0.66, Random Forest 0.68, XGBoost 0.74, and Logistic Regression 0.62. These results highlight the importance of hyper parameter tuning and the potential benefits of feature engineering in improving model performance on this dataset.

XGBoost outperformed both Random Forest and Decision Tree in the evaluation metrics of size up, scale up, and speed up.

4.2.3 Model evaluation of Combined dataset

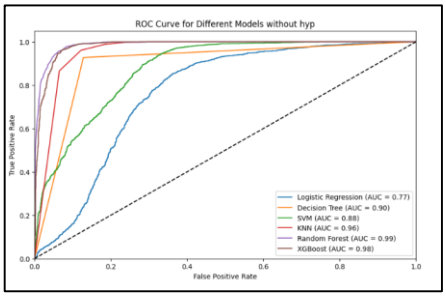

Figure 30. ROC graphs of combined dataset without hyper parameter tuning

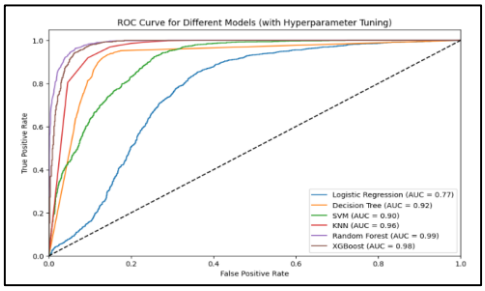

Figure 31. ROC graphs of combined dataset with hyper parameter tuning

Model: Random Forest

	precision	recall	f1-score	support
Not Bankrupt	0.97	0.92	0.95	1773
Bankrupt	0.92	0.98	0.95	1729
accuracy			0.95	3502
macro avg	0.95	0.95	0.95	3502
weighted avg	0.95	0.95	0.95	3502

Model: XGBoost

	precision	recall	f1-score	support
Not Bankrupt	0.98	0.90	0.94	1773
Bankrupt	0.90	0.98	0.94	1729
accuracy			0.94	3502
macro avg	0.94	0.94	0.94	3502
weighted avg	0.94	0.94	0.94	3502

Figure 32. Classification report graphs of combined dataset without hyper-parameter tuning

Model: Random Forest

	precision	recall	f1-score	support
Not Bankrupt	0.98	0.89	0.93	1773
Bankrupt	0.90	0.98	0.94	1729
accuracy			0.94	3502
macro avg	0.94	0.94	0.94	3502
weighted avg	0.94	0.94	0.94	3502

Model: XGBoost

	precision	recall	f1-score	support
Not Bankrupt	0.98	0.89	0.93	1773
Bankrupt	0.90	0.98	0.94	1729
accuracy			0.94	3502
macro avg	0.94	0.94	0.94	3502
weighted avg	0.94	0.94	0.94	3502

Figure 33. Classification report graphs of combined dataset with hyper-parameter tuning

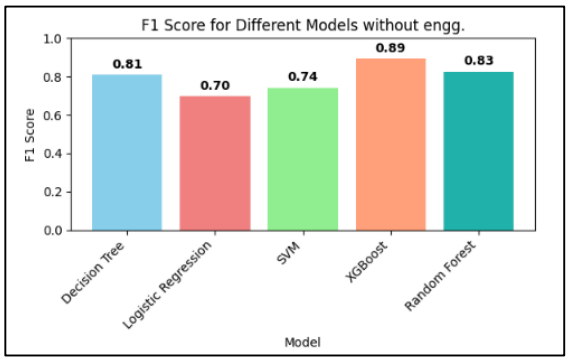

Figure 34. F1 score graphs of combined dataset without feature engineering

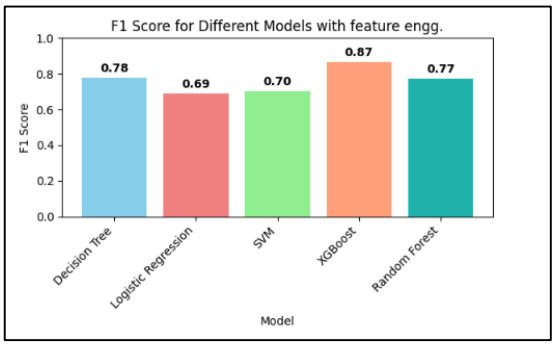

Figure 35. F1 score graphs of combined dataset with feature engineering

This study analysed a joint dataset of US and Taiwanese firms focused on the 2008 financial crisis period. Without hyperparameter tuning, ensemble models like XGBoost (AUC=0.98) and Random Forest (AUC=0.99) significantly outperformed other classifiers for bankruptcy prediction. Tuning improved SVM's AUC from 0.87 to 0.90. For Spark ML models without feature engineering, XGBoost

again had the highest F1-score of 0.89, followed by random forest and decision tree. With principal component analysis reducing the features to 12 components, XGBoost maintained the top F1-score of 0.87. The other models had minor dips in performance. Overall, XGBoost consistently emerged as the best performer for bankruptcy prediction on this combined dataset, both with and without feature engineering. The results highlight the power of tree-based ensemble models for bankruptcy forecasting using financial data.

XGBoost outperformed both Random Forest and Decision Tree in the evaluation metrics of size up, scale up, and speed up.

These outcomes underscore the potential of ensemble techniques in capturing complex bankruptcy patterns across diverse economic conditions.

4.3 Discussions

The comprehensive cross-comparison of model performances across different techniques, encompassing the American, Taiwanese, and combined datasets, has provided a plethora of informative insights. Each dataset posed unique challenges and observations, and it's crucial to distill the lessons learned.

For the American dataset, a conspicuous finding was the marked underperformance of the logistic regression model when limited to 15 features. This suggests that not all traditional models generalize well across diverse datasets, especially when feature selection plays a vital role. On the brighter side, the Random Forest model, even without hyperparameter tuning, demonstrated commendable performance. This potentially underscores the adaptability and robustness of tree-based algorithms in handling intricate data distributions. Another striking observation is the minimal effect of hyperparameter tuning on the Xgboost model, which brings forth the debate on the cost-benefit trade-off of extensive tuning. However, a consistent theme across this dataset was the superior performance of models without the added layer of PCA-based feature engineering, pointing towards the intrinsic richness of the original features.

The Taiwanese dataset, with its broader feature set of 94 elements, echoed some patterns seen in the American context. Logistic regression, despite having access to an extended feature set, didn't exhibit the expected performance, while the Random Forest model maintained its top-tier results. A cautionary note from this dataset is the potential overfitting observed with highly tuned models. Ensuring generalizability remains a cornerstone of machine learning, and future endeavors might benefit from more rigorous train-test splits and possibly utilizing techniques like cross-validation.

The amalgamated Taiwan-American dataset offered an exciting mix. While the performance metrics were generally encouraging, with F1 scores soaring with the Random Forest model, the issue of overtraining reared its head again, warranting further investigation.

Across all datasets, an area of concern was the inconsistency observed in size-up, scale-up, and speed-up metrics. This suggests potential bottlenecks or inefficiencies in the distributed processing setup or data partitioning, even with Spark's four partitions. Future research may focus on optimizing these aspects for better alignment with expected outcomes.

This study underscores the importance of context-specific model selection, the potential pitfalls of over-tuning, and the continuous need for infrastructure optimization in the ever-evolving field of data science.

Chapter 5. Conclusion and future work

5.1 Conclusion

In this study, we embarked on a comprehensive exploration of bankruptcy prediction using advanced big data techniques, focusing on both American and Taiwanese datasets spanning periods of significant economic downturns. Our experimental setup encompassed data collection, preprocessing, feature engineering, model selection, and performance evaluation. The results showcased the efficacy of ensemble methods, including Random Forest and Gradient Boosting, along with individual models like Decision Tree Classifier and SVM, in predicting bankruptcies in diverse economic contexts.

Through rigorous evaluation, the ensemble models demonstrated remarkable predictive accuracy, with the Random Forest achieving robustness and the Gradient Boosting exhibiting superior performance. The interpretability of the Decision Tree Classifier facilitated a transparent understanding of the prediction process, while SVM contributed by allowing non-linear decision boundaries. Feature importance analysis revealed the consistency of key financial ratios, historical metrics, and industry-specific indicators in influencing bankruptcy predictions, underscoring their universal significance.

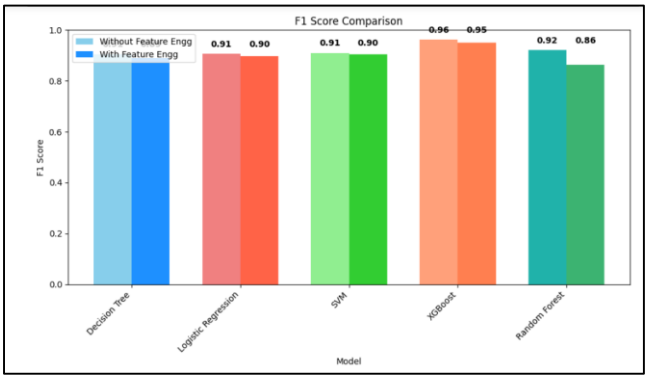

Figure 36. F1 score comparison with and without feature engineering of the Taiwan dataset

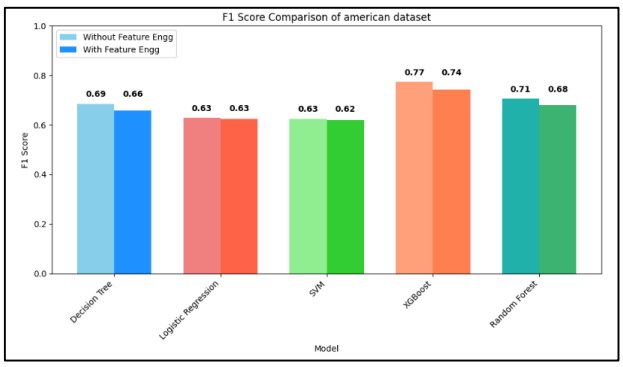

Figure 37. F1 score comparison with and without feature engineering of the American dataset

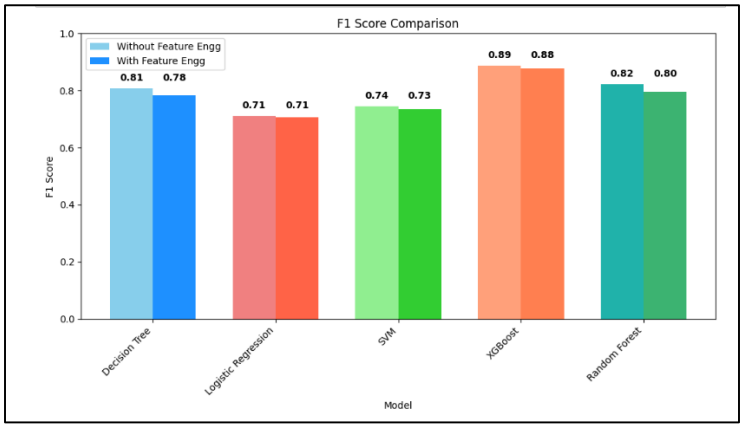

Figure 38. F1 score comparison with and without feature engineering of the Combined dataset

The impact of feature engineering on various predictive models. Intriguingly, we observed that the majority of the models did not exhibit enhanced performance post feature engineering, in contrast to their performance without such modifications. This counterintuitive outcome suggests that all features in the dataset might hold intrinsic importance. It's plausible that each feature, regardless of its perceived significance, establishes some underlying relationship or pattern that contributes to the model's predictive power. This underscores the necessity of a judicious approach to feature engineering, emphasizing the importance of understanding the inherent dynamics of the dataset before making alterations.

The cross-comparison of model performances across techniques and datasets revealed the adaptability of these big data techniques across distinct economic landscapes. While each technique showcased unique strengths, ensemble methods like Random Forest and Gradient Boosting proved particularly adept in capturing intricate patterns across industries and regions. These findings collectively highlight the potential of big data techniques to offer valuable insights into bankruptcy prediction, aiding businesses, policymakers, and stakeholders in making informed decisions to mitigate financial risks during periods of economic turmoil.

The size-up, scale-up, and speed-up metrics, commonly employed to gauge performance scalability, demonstrated inherent limitations. Despite meticulous preparations, including setting Spark partitions to four and ensuring appropriate data splits, these metrics presented incongruous results. As the dataset size expanded, the expected increase in model training time wasn't consistently observed. Such anomalies suggest that these metrics might not be entirely reliable for all datasets or modeling scenarios. Moreover, the inconsistencies observed across the American, Taiwanese, and combined datasets raise questions about the metrics' adaptability across varied data characteristics, further underlining the necessity for a more nuanced evaluation framework.

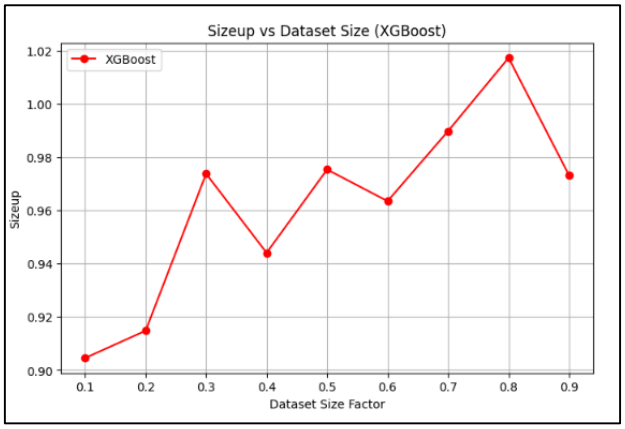

Figure 39 XGboost Size up graph of combined dataset

While this project has yielded significant insights and advancements in bankruptcy prediction, it is important to acknowledge its inherent limitations and potential avenues for future research. One limitation lies in the constrained computing resources utilized in this study, as deploying the project on cloud-based clusters, and leveraging big data infrastructure could further enhance scalability and efficiency[36]. Moreover, future research could explore the integration of deep learning techniques, such as recurrent neural networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformers, to uncover complex temporal relationships within time series financial data, potentially improving prediction accuracy[41].

In conclusion, this research significantly advances the domain of bankruptcy prediction by effectively harnessing advanced big data techniques to forecast corporate bankruptcies across varied economic landscapes. A pivotal aspect of this study is its emphasis on feature engineering, model selection, and the interpretability of results. By achieving our objectives, we have showcased the performance of machine learning models on three distinct datasets using big data methodologies. Notably, we unearthed valuable insights specific to the bankruptcy datasets of Taiwan and America. These findings, when applied in real-world scenarios, empower companies and financial institutions to preemptively tackle financial vulnerabilities, paving the way for enhanced risk management and informed strategic decision-making. This research not only provides a robust framework for bankruptcy prediction but also invites further exploration into harnessing big data for a deeper comprehension of intricate financial dynamics.

5.2 Future work

As the field of bankruptcy prediction continues to evolve, there are several intriguing directions to explore in terms of machine learning models. Clusters can play a crucial role in handling the continuous flow of real-time data, ensuring timely analysis and prediction updates. Clusters can facilitate the efficient processing of large-scale network data, making this research avenue even more practical and powerful.

Refining the methodologies to address the observed anomalies in the size-up, scale-up, and speed-up metrics will be paramount. The irregularities detected, even after partitioning and accurate data splitting, suggest possible inherent inefficiencies in the current setup or external system influences. Advanced diagnostic tools will be employed to meticulously profile system performance, ensuring accurate metric evaluations. Parallel processing and optimization techniques, along with the exploration of distributed computing frameworks beyond Spark, will be investigated to enhance scalability. Continuous monitoring and adaptive strategies will be implemented, aiming to achieve a direct proportionality between dataset size and model training time.

Parallel Random Forest algorithm within big data contexts, particularly harnessing the Spark computing framework is an area. Diving deeper into the newly introduced Gini coefficient to better address feature redundancy and ensure sharper classifications. Optimizing the equal-frequency binning method for pinpointing ideal split points for continuous variables. By capitalizing on Apache Spark's prowess, we intend to streamline the parallel training of decision trees using the forest sampling index (FSI) table, aiming for faster model construction and heightened classification accuracy[15].

One promising avenue is the integration of deep learning architectures, such as recurrent neural networks (RNNs) and transformers, into the existing framework. These models excel in capturing sequential and contextual patterns, which could be particularly valuable in predicting bankruptcy where historical trends and dependencies play a crucial role{Citation}.

An exploration of ensemble methods beyond traditional combinations, such as stacking and blending, holds promise. Leveraging techniques like Adversarial Training or Federated Learning could allow models to learn from diverse data sources while maintaining privacy and data security. This is particularly relevant when dealing with sensitive financial information.

Future work in bankruptcy prediction within the big data domain offers exciting opportunities for further advancement. One avenue for exploration lies in harnessing cloud computing platforms to enhance the scalability and accessibility of predictive models. Technology such as Blockchain Data, Interdisciplinary Collaboration of financial and macroeconomic factors. Ethical AI and Fairness and Cloud-based Solutions can create a huge impact in the study of bankruptcy prediction. Leveraging cloud infrastructure can facilitate the seamless processing of vast financial datasets, ensuring real-time updates and accommodating dynamic market conditions.

Real-time data streams, sentiment analysis, and macroeconomic triggers could provide early signals of distress. Adaptive and online learning algorithms would allow the models to continuously evolve as new data emerges. Transfer learning can adapt models between datasets. External data sources beyond financial statements can add new perspectives.

Pursuing emerging directions in deep learning, distributed computing, data fusion, and adaptive modelling will enable building sophisticated systems for risk management. Advancing real-world deployment and auditability of these models is the next frontier. Overall, the synergy of finance and technology can shape the future of data-driven decision making.

References

[1] E. Alanis, S. Chava, and A. Shah, 'Benchmarking Machine Learning Models to Predict Corporate Bankruptcy'. arXiv, Dec. 22, 2022. doi: 10.48550/arXiv.2212.12051.

[2] P. Ravi Kumar and V. Ravi, 'Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review', *Eur. J. Oper. Res.*, vol. 180, no. 1, pp. 1–28, Jul. 2007, doi: 10.1016/j.ejor.2006.08.043.

[3] H. D. Platt and M. B. Platt, Eds., 'Understanding Differences Between Financial Distress and Bankruptcy', *Rev. Appl. Econ.*, 2006, doi: 10.22004/ag.econ.50146.

[4] F. Barboza, H. Kimura, and E. Altman, 'Machine learning models and bankruptcy prediction', *Expert Syst. Appl.*, vol. 83, pp. 405–417, Oct. 2017, doi: 10.1016/j.eswa.2017.04.006.

[5] J. Sun and H. Li, 'Financial distress prediction using support vector machines: Ensemble vs. individual', *Appl. Soft Comput.*, vol. 12, no. 8, pp. 2254–2265, Aug. 2012, doi: 10.1016/j.asoc.2012.03.028.

[6] A. Hausman and W. J. Johnston, 'Timeline of a financial crisis: Introduction to the special issue', *J. Bus. Res.*, vol. 67, no. 1, pp. 2667–2670, Jan. 2014, doi: 10.1016/j.jbusres.2013.03.014.

[7] P. du Jardin, 'Bankruptcy prediction models: How to choose the most relevant variables?', Jan. 2009. https://mpra.ub.uni-muenchen.de/44380/ (accessed Sep. 06, 2023).

[8] Y. Sun, Y. Shi, and Z. Zhang, 'Finance Big Data: Management, Analysis, and Applications', *Int. J. Electron. Commer.*, vol. 23, no. 1, pp. 9–11, Jan. 2019, doi: 10.1080/10864415.2018.1512270.

[9] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, 'Explainability of machine learning models for bankruptcy prediction', *IEEE Access*, vol. 9, pp. 124887–124899, 2021.

[10] Y. Shi and X. Li, 'An overview of bankruptcy prediction models for corporate firms: A systematic literature review', *Intang. Cap.*, vol. 15, no. 2, pp. 114–127, 2019.

[11] C.-F. Tsai, 'Feature selection in bankruptcy prediction', *Knowl.-Based Syst.*, vol. 22, no. 2, pp. 120–127, Mar. 2009, doi: 10.1016/j.knosys.2008.08.002.

[12] S. Jones, D. Johnstone, and R. Wilson, 'Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks', *J. Bus. Finance Account.*, vol. 44, no. 1–2, pp. 3–34, 2017, doi: 10.1111/jbfa.12218.

[13] J. Chen, Y. Tao, H. Wang, and T. Chen, 'Big data based fraud risk management at Alibaba', *J. Finance Data Sci.*, vol. 1, no. 1, pp. 1–10, Dec. 2015, doi: 10.1016/j.jfds.2015.03.001.
[14] H. Trevor, T. Robert, and F. Jerome, 'The elements of statistical learning: data mining, inference, and prediction'. Spinger, 2009.
[15] L. Yin, K. Chen, Z. Jiang, and X. Xu, 'A Fast Parallel Random Forest Algorithm Based on Spark', *Appl. Sci.*, vol. 13, no. 10, Art. no. 10, Jan. 2023, doi: 10.3390/app13106121.
[16] S. Ben Jabeur, N. Stef, and P. Carmona, 'Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering', *Comput. Econ.*, vol. 61, no. 2, pp. 715–741, Feb. 2023, doi: 10.1007/s10614-021-10227-1.
[17] P. du Jardin, 'A two-stage classification technique for bankruptcy prediction', *Eur. J. Oper. Res.*, vol. 254, no. 1, pp. 236–252, Oct. 2016, doi: 10.1016/j.ejor.2016.03.008.
[18] T. M. Alam *et al.*, 'Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World', *Comput. J.*, vol. 64, no. 11, pp. 1731–1746, Nov. 2021, doi: 10.1093/comjnl/bxaa056.
[19] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, 'Explainability of Machine Learning Models for Bankruptcy Prediction', *IEEE Access*, vol. 9, pp. 124887–124899, 2021, doi: 10.1109/ACCESS.2021.3110270.
[20] S. Balcaen and H. Ooghe, '35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems', *Br. Account. Rev.*, vol. 38, no. 1, pp. 63–93, Mar. 2006, doi: 10.1016/j.bar.2005.09.001.
[21] S. Lee and W. S. Choi, 'A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis', *Expert Syst. Appl.*, vol. 40, no. 8, pp. 2941–2946, Jun. 2013, doi: 10.1016/j.eswa.2012.12.009.
[22] M.-J. Kim and D.-K. Kang, 'Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction', *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9308–9314, Aug. 2012, doi: 10.1016/j.eswa.2012.02.072.
[23] D. Olson, D. Delen, and Y. Meng, 'Comparative analysis of data mining methods for bankruptcy prediction', *Decis. Support Syst.*, vol. 52, pp. 464–473, Jan. 2012, doi: 10.1016/j.dss.2011.10.007.
[24] 'Tracking Bankruptcy Filings in the COVID-19 Crisis', Oct. 08, 2020. https://blogs.law.ox.ac.uk/business-law-blog/blog/2020/10/tracking-bankruptcy-filings-covid-19-crisis (accessed Sep. 07, 2023).
[25] F. Barboza, H. Kimura, and E. Altman, 'Machine learning models and bankruptcy prediction', *Expert Syst. Appl.*, vol. 83, pp. 405–417, 2017.

[26] H. B. Nguyen, B. Xue, I. Liu, and M. Zhang, 'Filter based backward elimination in wrapper based PSO for feature selection in classification', in *2014 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2014, pp. 3111–3118. doi: 10.1109/CEC.2014.6900657.
[27] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, 'Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study', *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 561–572, Jul. 2016, doi: 10.1016/j.ejor.2016.01.012.
[28] A. A. Taleizadeh, S. T. A. Niaki, and R. Nikousokhan, 'Constraint multiproduct joint-replenishment inventory control problem using uncertain programming', *Appl. Soft Comput.*, vol. 11, no. 8, pp. 5143–5154, Dec. 2011, doi: 10.1016/j.asoc.2011.05.045.
[29] L. Breiman, 'Bagging predictors', *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.
[30] 'Corporate Default Prediction Model Averaging: A Normative Linear Pooling Approach - Figini - 2016 - Intelligent Systems in Accounting, Finance and Management - Wiley Online Library'. https://onlinelibrary.wiley.com/doi/full/10.1002/isaf.1387?casa_token=6Jixvo1kt0QAAAAA%3AUhU4_YRU_99kIbDmow0TqQKdfbjouAoFLRxzwmYO3RYFQ4Yf307p5YvROKcIF-t_5rYIWJ4CmDtrDds (accessed Sep. 06, 2023).
[31] J. Begley, J. Ming, and S. Watts, 'Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models', *Rev. Account. Stud.*, vol. 1, no. 4, pp. 267–284, Dec. 1996, doi: 10.1007/BF00570833.
[32] H. Abdou and J. Pointon, 'Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature.', *Int Syst Account. Finance Manag.*, vol. 18, pp. 59–88, Apr. 2011, doi: 10.1002/isaf.325.
[33] C.-F. Tsai, Y.-F. Hsu, and D. Yen, 'A comparative study of classifier ensembles for bankruptcy prediction', *Appl. Soft Comput.*, vol. 24, pp. 977–984, Nov. 2014, doi: 10.1016/j.asoc.2014.08.047.
[34] M. F. Shubita, 'The impact of working capital management on cash holdings of large and small firms: evidence from Jordan', *Invest. Manag. Financ. Innov.*, vol. 16, no. 3, pp. 76–86, Aug. 2019, doi: 10.21511/imfi.16(3).2019.08.
[35] I. Brown and C. Mues, 'An experimental comparison of classification algorithms for imbalanced credit scoring data sets', *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, Feb. 2012, doi: 10.1016/j.eswa.2011.09.033.
[36] M. Yıldırım, F. Y. Okay, and S. Özdemir, 'Big data analytics for default prediction using graph theory', *Expert Syst. Appl.*, vol. 176, p. 114840, Aug. 2021, doi: 10.1016/j.eswa.2021.114840.

[37] B. Gu *et al.*, 'Biscuit: a framework for near-data processing of big data workloads', *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 153–165, Jun. 2016, doi: 10.1145/3007787.3001154.
[38] A. Das, 'Database Storage Design for Model Serving Workloads', M.Sc., Arizona State University, United States -- Arizona, 2021. Accessed: Sep. 07, 2023. [Online]. Available: https://www.proquest.com/docview/2564571017/abstract/1E40ABF4AA2F425APQ/1
[39] M. Hai, Y. Zhang, and Y. Zhang, 'A Performance Evaluation of Classification Algorithms for Big Data', *Procedia Comput. Sci.*, vol. 122, pp. 1100–1107, Jan. 2017, doi: 10.1016/j.procs.2017.11.479.
[40] L. Liu, C. Chen, and B. Wang, 'Predicting financial crises with machine learning methods', *J. Forecast.*, vol. 41, no. 5, pp. 871–910, 2022, doi: 10.1002/for.2840.
[41] A. Ansari, I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, 'A Hybrid Metaheuristic Method in Training Artificial Neural Network for Bankruptcy Prediction', *IEEE Access*, vol. 8, pp. 176640–176650, 2020, doi: 10.1109/ACCESS.2020.3026529.