Bankruptcy prediction with machine learning algorithms

Sarkis Chichkoyan

Abstract—**In an era characterised by unpredictable economic shifts, the ability to foresee corporate bankruptcy has emerged as a pivotal concern for stakeholders across the financial spectrum. This study embarks on an analytical journey to enhance the accuracy of bankruptcy predictions by leveraging a comprehensive dataset with 96 financial indicators, including a wide array of financial ratios and performance metrics. Through an exploration and preprocessing of this dataset, we ensure the highest data quality by addressing null values, duplicates, and standardising features. In this research we will use machine learning models specifically, logistic regression, support vector machines (SVM) and random forest classifiers to identify the most effective tool for predicting corporate bankruptcy. Among these, the random forest classifier emerged as the best model, demonstrating a remarkable predictive accuracy of 96.2%. The results garnered from this study hold implications for investors, regulatory bodies, and companies alike, offering a more nuanced understanding of financial health and paving the way for the development of sophisticated early warning systems.**

I.INTRODUCTION

In the progressively volatile global economic landscape, the ability to accurately predict corporate bankruptcy stands at the forefront of risk management and investment decision-making processes. The ramifications of corporate bankruptcies are extensive, affecting stakeholders across the spectrum, from employees and investors to creditors and consumers. Thus, the development of reliable predictive models, utilising data analytics techniques, has gained significant attention within the financial analytics community. This study contributes to this field by leveraging machine learning algorithms to enhance the accuracy of bankruptcy predictions.

The advent of big data and advancements in machine learning provide a lot of opportunities to improve upon traditional models of bankruptcy prediction. Despite these advancements, the challenge of achieving high predictive accuracy remains, due to the complexity and variability of financial data. This research aims to address this challenge by applying a range of machine learning techniques to a comprehensive dataset, thereby identifying the most effective methods for predicting corporate bankruptcy.

The significance of this study lies not only in its potential to improve financial risk assessment but also in its contribution to the broader understanding of financial health indicators. This research provides insights into the characteristics that are most predictive of bankruptcy. This can inform the development of more nuanced and effective financial risk management strategies.

Moreover, this research has implications beyond the immediate context of bankruptcy prediction. By demonstrating the applicability of machine learning techniques in financial analytics, it opens the door to further explorations into other areas of financial risk and opportunity assessment, including but not limited to credit risk analysis, fraud detection, and financial market predictions.

In summary, this study sets out to achieve two primary objectives: firstly, to identify the most effective machine learning techniques for predicting corporate bankruptcy; and secondly, to contribute to the broader discourse on financial risk assessment and management.

II. RELATED WORKS

In the sphere of corporate bankruptcy prediction, a significant body of research has been dedicated to enhancing prediction accuracy through various machine learning models. Sarojini Devi and Radhika (2018) offer a comprehensive review of both statistical and machine learning techniques in bankruptcy prediction, highlighting the superior predictive accuracy of machine learning techniques over traditional statistical methods for smaller datasets and the potential for further accuracy improvements through optimisation techniques like genetic algorithms and particle swarm optimisation when dealing with larger datasets. Very similar to that, Talha Mahboob Alam and colleagues (2020) explored the efficacy of machine learning models, including support vector machine (SVM), J48 decision tree, and random forest, among others, in predicting corporate bankruptcy. Their study emphasised the significance of data balancing techniques, with the decision forest model achieving a remarkable 99% accuracy, therefore outperforming other models and indicating the critical role of data preparation in predictive modelling.

These studies collectively underscore the evolving landscape of bankruptcy prediction, where the integration of diverse machine learning algorithms and data preprocessing techniques opens new opportunities for more accurate and timely predictions. The consensus among researchers about the necessity of sophisticated models and the handling of imbalanced datasets illustrates the field's move towards more refined and practical solutions for financial risk assessment.

III. DATASET DESCRIPTION

The dataset that we use for this research comprises a big collection of financial indicators from companies, structured across 96 features with one target variable indicating bankruptcy status (1 for bankrupt companies and 0 for solvent companies). This data encompasses a broad spectrum of financial metrics, including Return on Assets (ROA), Gross Margin, Operating Profit Rate, and various leverage ratios, therefore offering a multidimensional view of corporate financial health.

Sourced from Kaggle a publicly available website with several datasets, this dataset was meticulously curated to facilitate the study of bankruptcy prediction. It represents a diverse array of companies, ensuring a wide applicability of the research findings. Each record in the dataset corresponds to a unique company's financial standing over a specified fiscal year, allowing for the analysis to capture both cross-sectional and temporal dimensions of financial health.

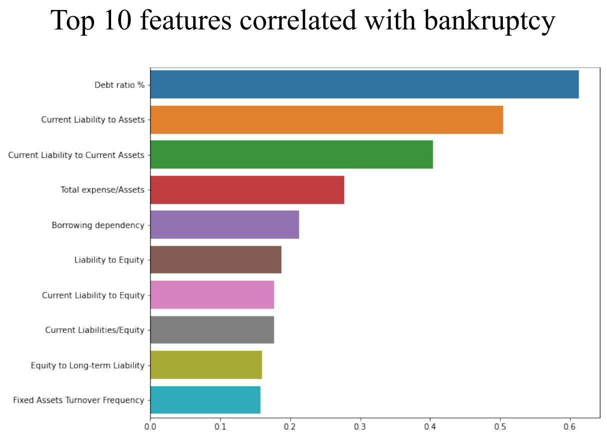

Fig. 1. Bar plot

Key Characteristics of the Dataset include:

- Size and Scope : The dataset consists of 6,819 entries, each representing a unique company's financial profile in a given year.
- Feature Diversity : The 95 variables cover a wide range of financial indicators, from liquidity and profitability ratios to operational and leverage metrics. This diversity allows for a comprehensive analysis of factors influencing bankruptcy risk.
- Data Quality : Prior to analysis, the dataset is going through rigorous preprocessing to ensure high data quality. This included the removal of duplicate entries, handling of missing values, and normalisation of numerical features to a standard scale.
- Target Variable : The binary nature of the target variable (Bankrupt?) facilitates the use of classification algorithms to predict the likelihood of bankruptcy.

The dataset's rich composition and high quality make it an exemplary basis for applying machine learning techniques to bankruptcy prediction. Through exploratory data analysis, significant features were identified and selected for modelling, laying the groundwork for a nuanced understanding of the predictors of corporate bankruptcy.

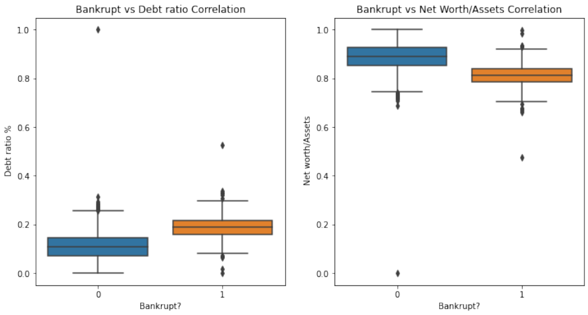

Fig. 2 Feature correlation plot

IV. METHODOLOGY

Data Preprocessing :

The foundation of our methodology lies in rigorous data preprocessing to ensure the dataset's suitability for machine learning analysis. This process began with an examination of the dataset for null values and duplicate entries. Given the critical importance of data integrity in predictive modelling, we first ensured that our dataset was devoid of any missing values and duplicate records.

Subsequently, we turned our attention to the dataset's structure, specifically its diversity of financial indicators. With 96 features encompassing a broad range of financial metrics from liquidity and profitability ratios to operational and leverage metrics standardisation was paramount. We employed a standardisation procedure to normalise the numerical features, facilitating a consistent scale across all variables. This step is crucial for models such as logistic regression and SVM, which are sensitive to the scale of input features.

Data Imbalance Handling and Application of SMOTE:

Recognising the challenge posed by the dataset's inherent imbalance where instances of bankruptcy are markedly fewer compared to non-bankruptcy cases we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate this issue.

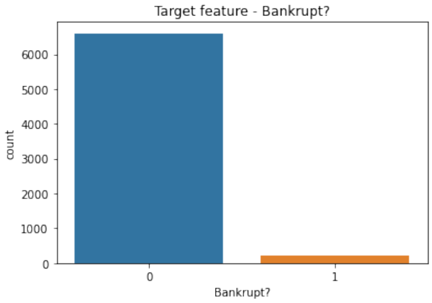

Fig . 3. Imbalanced in the target value

SMOTE synthesises new examples from the minority class to match the number of instances in the majority class, thereby balancing the dataset without losing valuable information. We applied SMOTE after the initial preprocessing steps but before model training, ensuring our machine learning algorithms logistic regression, support vector machines (SVM), and random forest classifiers benefited from training on a dataset that accurately reflects the diversity of outcomes in corporate financial health. This methodological adjustment was crucial for enhancing the models' ability to generalise and accurately predict unseen data, as evidenced by improvements in precision, recall, and F1-scores, metrics particularly sensitive to data imbalance.

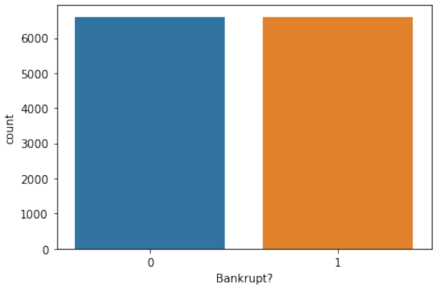

Fig . 4. Balanced with SMOTE

Model training and evaluation :

Logistic Regression, despite its name suggesting a regression algorithm, operates as a classification method, particularly effective for binary classification problems. It extends the concept of linear regression by estimating probabilities using a logistic function, often referred to as the sigmoid function. This model is working for our dataset where the dependent variable bankruptcy status is binary.

The Logistic Regression model computes the probability that a given feature set belongs to a certain class (bankrupt or not bankrupt). If the sigmoid function's output is greater than 0.5, the data point is classified into class 1 (bankrupt); otherwise, it falls into class 0 (not bankrupt). This default decision boundary at 0.5 offers a straightforward yet powerful way to classify companies based on their likelihood of facing bankruptcy.

$$sigmoid(x) = \frac{1}{1 + e^{x}}$$

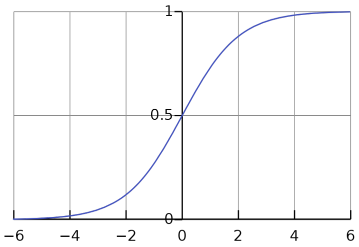

Fig. 5 . Sigmoid Function

Support vector machines is a powerful, non-linear classification technique that finds the optimal hyperplane which maximally separates the data points of different classes in the feature space. For linearly separable data, SVM looks for the hyperplane with the largest margin between the closest points (support vectors) of the classes. In cases where the data is not linearly separable, SVM uses kernel functions to transform the input space into a higher-dimensional space where a separation hyperplane can be found.

The model is robust and effective, especially for datasets where the boundary between classes is not immediately apparent or the data dimensionality is high.

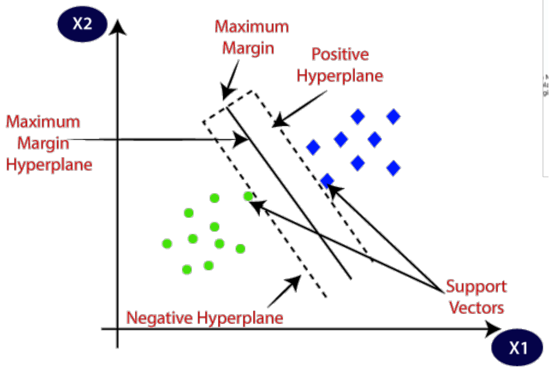

Fig. 6. SVM figure

Random Forest Classifier is an ensemble learning method that operates by constructing a multitude of decision trees during the training phase and outputting the class that is the mode of the classes (classification) of the individual trees. Random forests correct for decision trees habit of overfitting to their training set by introducing randomness in the tree building process. Each tree in the forest is built from a sample drawn with replacement (bootstrap sample) from the training set. Additionally, when splitting a node during the construction of a tree, the selected split is no longer the best split among all features but the best split among a random subset of the features. This strategy of combining diverse trees helps to achieve lower variance and better prediction accuracy.

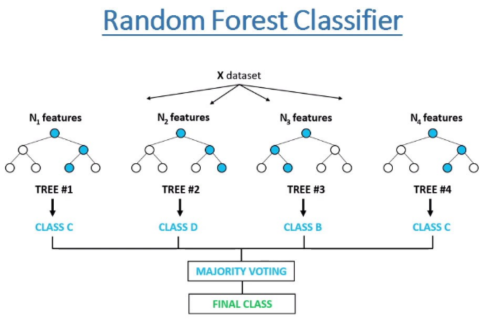

Fig. 7 RF trees

In our pursuit of a rigorous evaluation of the predictive models, we implemented k-fold cross-validation, with k set to 5. This approach allowed us to assess the models performance across different subsets of the dataset, ensuring a comprehensive validation process.

Cross-validation not only enhances the reliability of our performance metrics but also mitigates the risk of model overfitting, a critical consideration given the complexity of financial data.

V. RESULTS AND EVALUATIONS

Accuracy represents the proportion of correct predictions made by our model. In the context of binary classification, accuracy encompasses true positives and true negatives. Mathematically, it is represented as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Our models achieved the following accuracies:

– *Logistic Regression* Accuracy: 90%

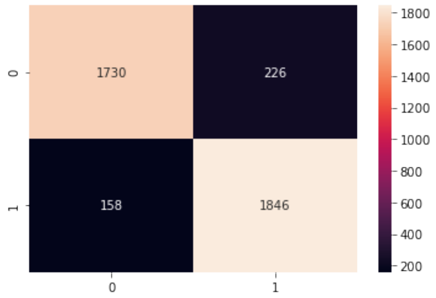

Fig .8. Logistic Regression confusion matrix

– *SVM* Accuracy: 94.8%

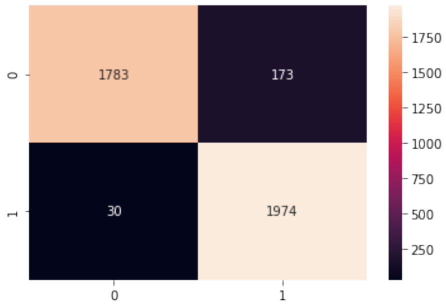

Fig .9. SVM confusion matrix

Random forest classifier Accuracy: 96.5%

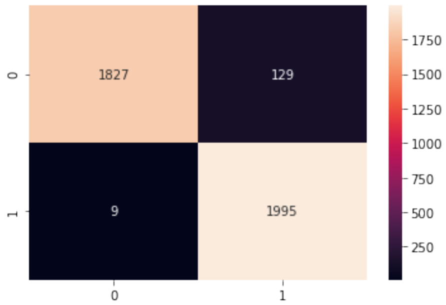

Fig .10. RF confusion matrix

Cross Validation Scores with Random forest classifier:

In evaluating the predictive performance of the Random Forest Classifier, we employed a 5-fold cross-validation technique to ensure the robustness and reliability of our model across different subsets of the dataset.

The cross-validation process had the following accuracy scores for each fold: 94.81%, 94.36%, 97.23%, 97.08%, and 97.76%.

The mean accuracy achieved through cross-validation was 96.2%, indicating a high level of consistency in the model's predictive capability across the various partitions of the data.

Notably, the standard deviation among the cross-validation scores was 1.3%, underscoring the stability of the Random Forest Classifier's performance across different subsets of the dataset. These results highlight the model's robustness and its potential utility in accurately predicting corporate bankruptcy.

VI. CONCLUSION

This study embarked on an analytical journey to explore the efficacy of machine learning algorithms in enhancing the accuracy of bankruptcy predictions. Through the preprocessing of a dataset encompassing 96 financial indicators, we ensured the integrity and standardisation of our data, setting a solid foundation for the application of advanced predictive models. A crucial step in our methodology was the implementation of the Synthetic Minority Over-sampling Technique (SMOTE), which addressed the inherent data imbalance issue, therefore enhancing the robustness of our models' predictions.

The comparative analysis of logistic regression, support vector machines (SVM), and random forest classifiers revealed that the random forest model outperformed its counterparts, achieving an impressive accuracy of 96.2%. This model's success underscores the potential of machine learning techniques in financial analytics, particularly in the prediction of corporate bankruptcy.

The adoption of cross-validation in our methodology represents a methodological advancement that significantly contributed to the reliability of our findings. This approach ensured that our models were not only accurate but also robust and generalisable across different data scenarios. Future studies could explore further refinements to the cross-validation process, such as employing stratified k-fold cross-validation to maintain the proportion of classes across folds, therefore further enhancing the fidelity of the evaluation process in imbalanced datasets like ours.

Looking ahead, we envisage several avenues for future research. Exploring the integration of macroeconomic indicators into the predictive models could offer a more holistic view of bankruptcy risk. Additionally, the application of deep learning techniques may uncover complex nonlinear patterns that traditional models may overlook. Finally, considering the dynamic nature of financial markets, continual refinement and validation of predictive models are imperative to maintain their accuracy and relevance.

In conclusion, as we advance, the fusion of machine learning techniques with financial analytics promises to enhance our foresight and resilience in the face of economic challenges.

REFERENCES

[1] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. 2019. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2):743–758.

[2] Sarojini Devi, B., and K. Radhika. 2018. Machine Learning Algorithms for Bankruptcy Prediction: A Comparative Study. *Journal of Computational and Theoretical Nanoscience*, 15(6-7), 2132-2141.

[3] Talha Mahboob Alam, Mian Saqib Mehmood, Shafaq Salam. 2020. A Comparison of Machine Learning Algorithms for Bankruptcy Prediction. *International Journal of Advanced Computer Science and Applications*, 11(3).

[4] Deron Liang and Chih-Fong Tsai. Company Bankruptcy Prediction - Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009. Uploaded to Kaggle 3 years ago. Data obtained from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction. National Central University, Taiwan.

[5] Scikit-learn.org. 2024. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation. [online] Available at: https://scikit-learn.org/stable/

[6] Seaborn.pydata.org. 2024. seaborn: statistical data visualization — seaborn 0.11.2 documentation. [online] Available at: https://seaborn.pydata.org/

[7] Numpy.org. 2024. NumPy. https://numpy.org/ [online]

[8] ”pandas - Python Data Analysis Library”, Pandas.pydata.org, 2024. [Online]. Available: https://pandas.pydata.org/.

[9] ”Matplotlib — Visualization with Python”, Matplotlib.org, 2024. [Online]. Available: https://matplotlib.org/.