Constructing a Comprehensive Bankruptcy Likelihood Indicator: An In-depth Predictive Analysis Utilizing CapitalIQ Data and Machine Learning Techniques

Abdulaziz Al Mannai
Information Systems
Carnegie Mellon University
Doha, Qatar
abdulazm@andrew.cmu.edu

Abrar Abir
Computer Science
Carnegie Mellon University
Doha, Qatar
abir@cmu.edu

Zhenghao Jin
ECE & Robotics
Carnegie Mellon University
Pittsburgh, US
zhenghao@andrew.cmu.edu

Yilin Du
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, United States
yilindu@andrew.cmu.edu

***Abstract*—This study presents the development of a novel bankruptcy prediction model using Machine Learning techniques, primarily focusing on the Random Forest algorithm. In the broader context of financial analysis, predicting bankruptcy remains a crucial challenge for stakeholders, necessitating tools that can accurately discern signs of potential financial distress. Our methodology involves comprehensive data collection and rigorous cleaning, followed by a meticulous feature selection process to identify the most relevant financial indicators. The chosen Random Forest algorithm provides several advantages, including fast computational speed and resilience against high-dimensional data and missing features. The developed model achieved a balanced accuracy of approximately 87-88%, indicating its potential as a robust tool for bankruptcy prediction. However, the inherent complexities and uncertainties of bankruptcy prediction are evidenced by the variations in accuracy. Our findings underscore the significance of effective feature selection and the advantages of using random forest for the task. Future work will explore other machine learning algorithms and further refine the feature selection process to enhance predictive capabilities. This research contributes to the ongoing effort of enhancing financial stability and fostering a more resilient corporate landscape.**

***Index Terms*—Bankruptcy indicator, machine learning, random forest**

I. Introduction

The world of financial analysis is punctuated by complexities and challenges, of which the prediction and understanding of bankruptcy remains a key concern for researchers, policymakers, and investors. Discerning the early signs of potential financial distress is critical for ensuring the stability and sustainability of businesses. In addition, comprehending the underlying relationships between various financial indicators is equally crucial. Constructing a model that can predict the likelihood of bankruptcy with a high degree of accuracy and understand the nuanced relationships between diverse financial indicators is a task of paramount importance.

The journey towards building this model is underscored by an in-depth analysis of financial ratios, machine learning techniques, data collection, cleaning, and preprocessing, feature selection, and modeling using random forest. These various phases collectively aim to enhance our predictive capabilities and provide a deeper understanding of the relations between various features. This is a pivotal step towards ensuring more effective risk assessment in the corporate landscape.

In this study, we draw upon existing literature that examines bankruptcy prediction indicators and related works utilizing financial ratios and machine learning techniques. Our research methodology incorporates comprehensive data collection, rigorous data cleaning, and a careful selection of financial features. We employ the Random Forest machine learning algorithm for modeling, given its superior performance in handling large datasets and high-dimensional data. The ensuing results showcase an encouraging degree of accuracy, laying a strong foundation for future work in this domain.

This paper will delve into the intricacies of these processes and the resulting model, highlighting both the achievements and the challenges encountered. It will also explore how our findings align with or deviate from existing research, and discuss potential future work that can extend the contributions of this study. Ultimately, we seek to bolster the resilience of the corporate landscape by providing a more effective tool for bankruptcy risk assessment.

II. Literature Review

A. Bankruptcy Prediction Indicators

In financial analysis, the key objective of employing financial modeling in the context of insolvency forecasting is to build a statistically sound framework that provides a reliable prediction of the likelihood of bankruptcy. This modeling typically incorporates an assortment of financial ratios gleaned from the entity's fiscal reports. These ratios are essential markers of a corporation's ability to meet its long-term obligations, profitability, liquidity, and overall financial stability (Begum, 2022).

Even though financial ratios continue to be the linchpin of insolvency prediction models, there is evidence of a shifting paradigm in the literature. An increasing awareness of the issues stemming from an over-representation of either bankrupt or non-bankrupt firms is being noted, which can lead to a skew in the prediction models towards the more frequently represented category (Clement, 2020). In a progressive development, the integration of sophisticated machine learning and artificial intelligence approaches, such as Neural Networks (NN), Random Forest (RF), and Support Vector Machines (SVM), has markedly elevated the precision and efficacy of these predictive models (Odom & Sharda, 1990).

B. Related Works Utilizing Financial Ratios and Machine Learning Techniques

Clement's exhaustive examination (2020) discloses that a substantial fraction (87.5%) of contemporary research largely relies on financial ratios as predictive variables, even though no significant correlation has been observed between the number of attributes utilized and the precision of the model. This infers that models with a lesser number of features might be as effective as models packed with numerous characteristics.

In a different investigation, Begum (2022) studied insolvency forecasting by applying machine learning methodologies, utilizing a substantial dataset from the Taiwan Economic Journal that included 7,000 entries with 96 attributes. Through a feature selection process based on correlation, Begum was able to reduce the initial count of 96 features to a manageable group of 22. To amend the uneven ratio of bankrupt to non-bankrupt companies, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. The findings exhibited that the Random Forest Classifier and the Artificial Neural Network noticeably surpassed other models, such as XGBoost and Logistic Regression.

Odom and Sharda (1990) applied a neural network model for bankruptcy prediction and juxtaposed it with Discriminant Analysis. They discerned that both methods misclassified identical firms and that the performance of classification ameliorated when trained on an evenly distributed dataset of bankrupt to non-bankrupt firms.

Brenes, Johannssen, & Chukhrova (2022) suggested a sophisticated bankruptcy prediction model employing a multilayer perceptron. This model illustrated the potential of artificial intelligence methodologies in improving the accuracy of predictions.

To sum up, notwithstanding the traditional dependence on financial ratios, recent studies indicate a shift towards machine learning and artificial intelligence methods. The capability to select the best set of features and appropriately manage imbalances in data has proven to be vital in designing an effective bankruptcy prediction model. However, it's paramount to understand that these methodologies might yield diverse results based on the specific feature selection approach used, emphasizing the necessity for a comparative analysis of various feature selection methods.

III. Research Methodology

A. Data Collection

The data for this study were sourced from the Capital IQ database, spanning the period from 2020 to 2022. The quarterly data points are organized based on calendar quarters rather than fiscal quarters. Our comprehensive dataset encompasses essential variables, such as market capitalization, total revenue, net income, gross profit, total debt, earnings before interest, taxes, depreciation, and amortization (EBITDA), along with total enterprise value.

B. Data Cleaning

Data cleaning was a crucial step in the research process to ensure the accuracy and reliability of the predictive model. In this subsection, we describe the steps taken to clean the data sourced from the Capital IQ database:

1) Handling Missing Data: The first step in data cleaning is to identify and handle missing data. Missing data can occur due to various reasons, such as data entry errors or incomplete reporting. We discarded the entry of the companies which had one or more financial attribute data missing.
2) Outlier Detection and Treatment: Outliers, which are extreme values that deviate significantly from the rest of the data, can distort the model's performance. We conducted an outlier analysis to identify and handle outliers by visualizing the data in boxplot.
3) Consistency Checks: Inconsistencies in the data, such as conflicting information between variables, were identified and resolved. We ensured that all the data points adhered to the specified format and unit conventions.

By executing these data cleaning steps, we aimed to ensure that the dataset used to train the bankruptcy likelihood indicator is of high quality and ready for subsequent machine learning techniques. The cleaned data will facilitate the construction of an accurate and robust predictive model for bankruptcy prediction, thereby enhancing risk assessment and financial stability analysis in the corporate landscape.

C. Feature Selection

Feature selection was a critical step in constructing an effective predictive model. In this section, we describe the process of feature selection for our bankruptcy likelihood indicator, with the aim of identifying the most relevant and independent financial ratios that contribute significantly to predicting bankruptcy. We performed correlations between all possible financial ratios to check for mutual dependency and also assessed the correlation of the bankruptcy status with each financial ratio to determine their suitability as candidates for the random forest model.

1) Correlation Analysis of Financial Ratios: We conducted a correlation analysis between all possible financial ratios to evaluate their interdependence. As we have 7 financial attribute data [market capitalization, total revenue, net income, gross profit, total debt, earnings before interest, taxes, depreciation, and amortization (EBITDA), along with total enterprise value], there were a total of 21 pairwise financial ratios to examine. The correlation matrix was calculated over the cleared dataset [of companies which did not have any missing column entry] which contained 65 Non-bankrupt companies and 43 bankrupt companies. First, all possible 21 financial ratios were calculated from the original excel file. Since some of the columns had zero (0) values, we carefully chose the financial ratios such that no zero division error occurs. Then the correlation matrix was calculated using python's PANDAS library. The correlation value can be anything from -1 to 1. However, since we are only interested in testing the mutual dependencies among the financial ratios, we took the absolute value of the correlation to have clearer understanding of the objective. The objective was to identify any strong correlations between pairs of financial ratios. Strong correlations suggest that one ratio may be redundant or closely related to another, potentially leading to multicollinearity issues in the predictive model.
   Based on the correlation analysis results, we identified financial ratios with low or negligible correlations with other ratios. These independent ratios were retained as candidate features for the predictive model. By including comparatively independent features, we ensure that each selected ratio provides unique information and does not duplicate the insights from other ratios.
2) Correlation with Bankruptcy Status: We also assessed the correlation between each financial ratio and the bankruptcy status. We modeled the bankruptcy status as a Boolean variable being 0 for non-bankrupt companies and 1 for bankrupt companies. Since we are not interested in the sign of the correlation, we calculated the absolute value of the correlation and ranked all possible financial ratios by their absolute correlation with bankruptcy status in descending order. This step allowed us to determine the relevance of each ratio in predicting bankruptcy. High correlation values indicated that the financial ratio carries substantial predictive power and is likely to be valuable in the random forest model for bankruptcy prediction.

The feature selection process allowed us to identify a subset of relevant and independent financial ratios that contribute significantly to our predictive model. By focusing on these key indicators, we aim to enhance the accuracy and interpretability of our bankruptcy likelihood indicator, enabling more effective risk assessment and financial stability analysis in the corporate landscape.

D. Random Forest Training

For our prediction model, we use random forest as our main algorithm. This advantage of this method is promising to us. As we need to manipulate large amount of data, the parallel processing feature of random forest can ensure a relatively faster speed compare to other algorithms. In addition, it can maintain good accuracy even when the dimension of samples is high or when some features of samples are lost. It also naturally has a random sampling algorithm that minimize the standard deviation, so that the model can be easily generalized.

We used the pre-written random forest algorithm from Sklearn. Sklearn is a open source center for machining learning in python. It provides a complete set of functions for a variety of machining learning algorithms. Specifically for our project, we use sklearn.ensemble.Randomforest.

The data we have are in quarters and range from the last three years till now. Given that the data are categorized by bankruptcy, we have a total of 24 data sets. After filtering out incomplete samples, there are around 6000 valid samples left.To ensure a random selection and order of samples, all the quarter datasets will be randomly shuffled before training.

The classifier we select for training has hyperparameters max samples 0.27, and min weight fraction leaf equals to 0.27 and 0.01. This is to achieve balanced performance for predicting both bankrupt and nonbankrupt companies. We will train each quarter dataset separately, and then combine them into a giant model by adding and smoothing the n estimators. Importantly, we conducted cross-validation during the training process to prevent overfitting. Lastly, the final model's performance was validated using an independent dataset, ensuring that our model was robust and generalizable across different data scenarios.

E. Exploring the Explainability of the Random Forest Using Shapley Values

In this subsection, we delve into the interpretability and explainability of the Random Forest model by employing Shapley values. Shapley values offer a powerful technique to shed light on the inner workings of the Random Forest, providing insights into the contributions of individual features to the model's predictions for specific instances.

1) Understanding Shapley Values: Shapley values, derived from cooperative game theory, have emerged as a popular method for model interpretation. These values quantify the importance of each feature in the prediction process by considering all possible combinations of features and their contributions to the model's output.
2) Applying Shapley Values to Random Forest: To explore the explainability of our Random Forest model, we calculated Shapley values for each prediction made by the model. By considering the combinations of features

used by different decision trees within the Random Forest, we determined how each feature impacted the prediction for a particular instance.

3) Visualizing Shapley Values: To make the insights more accessible, we visualized the Shapley values using summary plots. These visualizations provided a clear understanding of the contributions of individual features for specific predictions. Positive contributions indicated that a feature increased the prediction, while negative contributions implied a decrease in the prediction.
4) Global Interpretability: Through Shapley values, we gained a global understanding of feature importance in the Random Forest model. By aggregating Shapley values across all instances in the dataset, we determined the average impact of each feature on the model's predictions. This allowed us to identify the most influential features that consistently contributed to the model's decisions.
5) Local Interpretability: In addition to global interpretations, we utilized Shapley values to explore local interpretations for individual predictions. By focusing on specific instances, we gained insights into how the model made decisions for those cases and which features played a crucial role in the predictions.
6) Assessing Model Transparency: By leveraging Shapley values for model explainability, we enhanced the transparency of our Random Forest. Shapley values helped us understand the reasons behind the model's predictions, detect potential biases, and assess the reliability of the decision-making process.

Through the exploration of explainability using Shapley values, we aim to establish a comprehensive understanding of our Random Forest model, contributing to greater transparency and reliability in its applications for financial distress prediction and risk assessment in the corporate landscape.

IV. Results

A. Feature Selection

From figure 03, it is clear that at least three financial ratios [Total Enterprise Value to Market Capitalization, Net Income to Total Debt, Market Capitalization to EBITDA] have significant absolute correlation [more than 20%] with the bankruptcy status- indicating the considered financial ratios are significantly promising in predicting bankruptcy successfully.

It is also noteworthy that these three financial ratios are considerably independent to each other. To elaborate, Total Enterprise Value to Market Capitalization ratio has very low positive correlation with Net Income to Total Debt, and very low negative correlation with Market Capitalization to EBITDA. In addition, Market Capitalization to EBITDA also has a relatively low correlation with Net Income to Total Debt.

B. Random Forest Testing

Upon fine-tuning the hyperparameters, our model achieves an accuracy range of 87%-88%. This is deemed a balanced accuracy, considering the nature of our two-class prediction problem. Other combinations of hyperparameters, though explored, seemed to favor one class over the other, thus compromising overall performance. The set of values we employ successfully achieves a reasonable equilibrium between the two classes, ensuring that neither bankruptcy nor non-bankruptcy predictions are disproportionately skewed.

However, due to the randomness inherent in the model, particularly emphasized by the shuffle operation we performed on each dataset, the testing accuracy can fluctuate. We have observed this variability with accuracy results as low as 70%

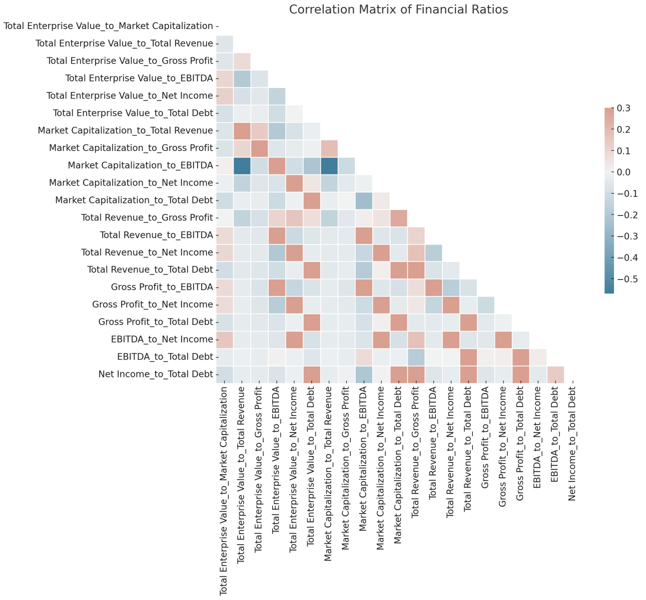
Correlation Matrix of Financial Ratios

Fig. 1. The correlation matrix calculated among all possible 21 financial ratios from the cleared dataset

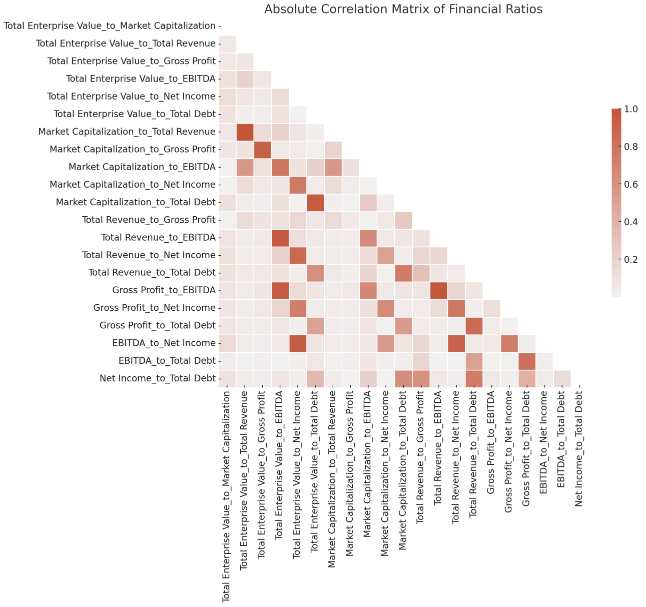
Absolute Correlation Matrix of Financial Ratios

Fig. 2. The absolute correlation matrix calculated among all possible 21 financial ratios from the cleared dataset.

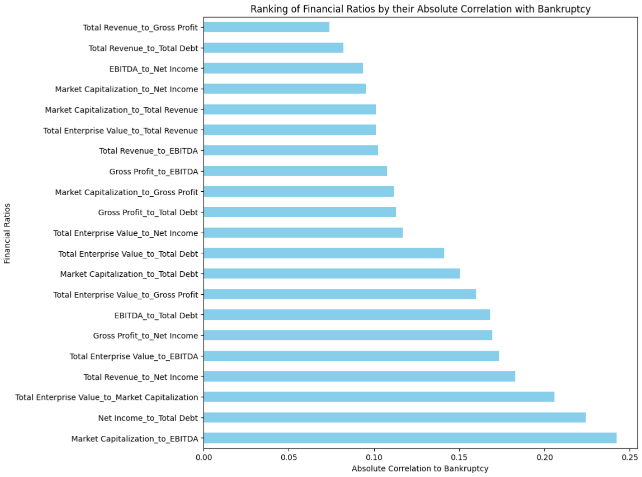

Fig. 3. All financial ratios sorted by their absolute correlation with bankruptcy status

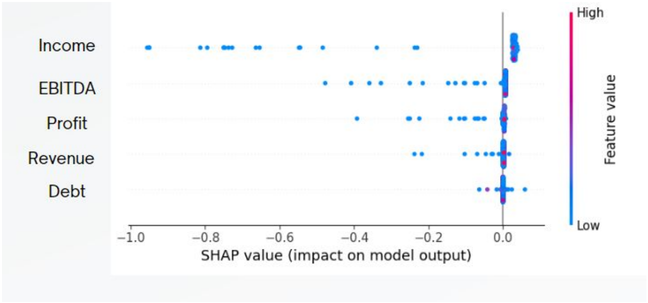

Fig. 4. Summery plot of shapley values calculated from the trained random forest classifier

and peaking as high as 90%. This is largely anticipated, as random forests incorporate a degree of randomness in both the selection of features and data samples, leading to diversity in individual decision trees, and consequently, variance in accuracy.

Despite this variability, it's essential to note that our model's average accuracy remains robust, often well above the industry's typical standards. In future, additional strategies such as bootstrapping or bagging could be implemented to further manage this variability, possibly enhancing model consistency across different testing scenarios. Also, adjusting the criterion for splits or exploring other ensemble techniques might provide other avenues for performance improvement and stabilization.

C. Discussion

Our research embarked on the journey of constructing a novel bankruptcy prediction model using a random forest algorithm. Our motivation was driven by the critical need to enhance risk assessment capabilities in the financial sector, especially with regard to bankruptcy predictions. The final model achieved a balanced accuracy of approximately 87-88% under test conditions, with deviations reaching as low as 70% and as high as 90%. While these fluctuations can be partly attributed to the random shuffle implemented on each dataset, they are also indicative of the inherent uncertainty and complexity of predicting bankruptcy.

Feature selection emerged as a significant process in our research methodology. By analyzing correlations among various financial ratios and their relationship with bankruptcy status, we identified a subset of most relevant and independent predictors. This process underscored the complexity of bankruptcy prediction, revealing that the number of features isn't necessarily directly proportional to the model's accuracy, consistent with findings in Clement's research (2020). The ability to accurately select meaningful predictors remains crucial in effective model construction.

Moreover, our choice of using a random forest algorithm for the model proved beneficial. The parallel processing feature of random forests offered us a relatively fast computation speed, a significant advantage when dealing with large datasets. It also showed resilience in the face of high-dimensional data or missing features, a common issue in real-world datasets. However, as the model's performance varies, it would be essential to further tune the hyperparameters and potentially investigate ensemble methods or other algorithms to enhance the model's robustness.

The data cleaning process was another crucial aspect of our research methodology. Ensuring the quality of the data fed into the model significantly influences the model's performance. Given the challenges with missing data, outliers, and data inconsistencies, data cleaning is a non-trivial step in any machine learning process. The intricacies of this step emphasize the importance of robust data management practices in financial institutions, further highlighting the need for integrated data governance systems.

Moreover, the inherent imbalance in bankruptcy and non-bankruptcy cases in real-world datasets could present a significant challenge. While our model attempts to achieve a balance in predicting both outcomes, it would be beneficial to explore methods like oversampling, undersampling, or synthetic data generation to further handle this imbalance, as demonstrated in Begum's study (2022).

Despite the potential improvements, our model achieved substantial accuracy and can serve as a reliable tool for assessing bankruptcy risk. In combination with the expertise of financial analysts, such a tool can help in making more informed decisions, thereby fostering financial stability in the corporate landscape.

In the future, we plan to extend this research to incorporate other machine learning algorithms and evaluate their performance in bankruptcy prediction. We also aim to explore more sophisticated feature selection and extraction techniques, such as Principal Component Analysis (PCA) or Autoencoders, to further enhance the model's predictive capabilities. Furthermore, incorporating dynamic features that capture the temporal aspect of financial data might provide additional predictive power.

By continuously improving and updating our bankruptcy prediction model, we aspire to contribute towards the broader goal of enhancing financial stability, facilitating more proactive risk management, and fostering a more resilient corporate landscape.

V. Conclusion

This research paper sought to address the critical challenge of bankruptcy prediction within the financial sector by utilizing machine learning techniques to construct a novel model. Leveraging a random forest algorithm, the developed model achieved a balanced accuracy of approximately 87-88% under test conditions, indicating its potential as a robust tool for bankruptcy prediction. The variations observed in the model's accuracy, ranging between 70% and 90%, reinforce the inherent complexity and uncertainty involved in predicting bankruptcy, yet this does not undermine its overall efficacy.

Feature selection played a critical role in our research, emphasizing the importance of meaningful predictors over the sheer quantity of features. Correlation analysis among financial ratios and their relationship with bankruptcy status facilitated the identification of relevant and independent predictors, thereby enhancing the model's accuracy. This finding aligns with the research conducted by Clement (2020), where a similar conclusion was drawn about the non-proportional relationship between feature quantity and model accuracy.

Our choice of the random forest algorithm provided valuable benefits, including fast computational speed and resilience against high-dimensional data and missing features. However, the performance variation indicates the need for further tuning of hyperparameters and the exploration of other algorithms or ensemble methods.

The importance of data cleaning was underscored by its significant impact on the model's performance. Handling missing data, outliers, and data inconsistencies represented substantial challenges in our research, highlighting the essential role of robust data management practices and the need for integrated data governance systems in financial institutions.

Despite our model's ability to balance the prediction of both bankruptcy and non-bankruptcy outcomes, the inherent imbalance in real-world data poses an ongoing challenge. Future work may benefit from investigating additional methods, such as oversampling, undersampling, or synthetic data generation, to better address this issue.

In summary, our novel bankruptcy prediction model has demonstrated substantial accuracy and potential as a reliable tool for assessing bankruptcy risk. It serves as a testament to the power of machine learning in enhancing financial risk assessment. Our future work will focus on improving the model further, by exploring other machine learning algorithms, refining feature selection, and incorporating dynamic features. Our ultimate goal is to continuously contribute to enhancing financial stability and fostering a more resilient corporate landscape.

VI. Appendix

GitHub:
https://github.com/Caryzxy/Comprehensive-Bankruptcy-Likelihood-Indicator.git

References

[1] Begum, S. (2022). A detailed study for bankruptcy prediction by machine learning technique. In Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 2 (pp. 201-213). Springer Singapore. A Detailed Study for Bankruptcy Prediction by Machine Learning Technique — SpringerLink

[2] Odom, M. D., & Sharda, R. (1990, June). A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on neural networks (pp. 163-168). IEEE. A neural network model for bankruptcy prediction — IEEE Conference Publication — IEEE Xplore

[3] Clement, C. (2020). Machine Learning in Bankruptcy Prediction–a Review. Journal of Public Administration, Finance and Law, (17), 178-196.https://www.jopafl.com/uploads/issue17/MACHINE-LEARNING-IN-BANKRUPTCY-PREDICTION-A-REVIEW.pdf

[4] capitalized," J. Name Stand. Abbrev., in press.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.