Improving Bankruptcy Prediction Using Machine Learning

Mrs. Sandhya Rani[1], Kompally Poorvika[2], Bolledu Sri Chandana[3]
Manda Samyuktha[4], Shate Krishna Shree[5]
[1]*HOD, M. Tech, (Ph. D), Department of CSE - Cyber Security, Sphoorthy Engineering College, Hyderabad, India*
[2,3,4,5]*Department of CSE - Cyber Security, Sphoorthy Engineering College, Hyderabad, India*

***Abstract*—Proper prediction of bankruptcy is a key component in creditor, investor, enterprise, and policy maker's decision-making frameworks. Such forecasting mechanisms can also be used to curb wider undesirable impacts on the economy and society by facilitating a more accurate assessment of individual risks. However, the classical bankruptcy forecasting frameworks are often limited by the major assumptions; they assume linear correlation and, therefore, perform poorly in the case of an extremely lopsided financial data set. In order to address these limitations, the current paper proposes a modern machine-learning model tested on the Taiwan Bankruptcy Prediction Dataset. To successfully reduce the problem of class imbalance and enhance the overall predictive accuracy, a set of approaches were cross-validated, with special attention given to the usage of the Synthetic Minority Over Sampling Technique (SMOTE) in combination with the Random Forest classifiers. The usefulness of the proposed model in practice is demonstrated by building an easy-to-use web-based dashboard. The application was created with Python, Pandas, Scikit-Learn, and Stream lit as it allows users to upload financial data automatically and obtain instant and probabilistic estimates of the risk of bankruptcy. In the end, this system fills the gap between complex financial trends and evidence-based and useful insights, thus providing a powerful tool that helps financial institutions and investors in decision making.**

***Index Terms*—Bankruptcy prediction, machine learning, SMOTE, random forest classification, cross validation, accuracy.**

I. INTRODUCTION

The topic of bankruptcy prediction forms a central aspect of modern financial decision-making by a wide range of participants such as creditors, investors, corporate entities, and policy-makers. Accurate prediction in this sphere is an essential requirement to reduce the negative economic and social consequences of corporate failure. However, the importance of such predictions is often compromised by old models that assume linear relationships amongst variables thus inability to capture the non-linear financial patterns that characterize the contemporary markets. One major methodological issue in this field is the lack of balanced data, as a small number of bankruptcies compared to solvent companies will have a harmful impact on the model.

In order to overcome these systematic shortcomings, the current paper suggests a sophisticated machine-learning-based bankruptcy prediction system, which is designed to detect the key financial risk variables with increased accuracy. The study attempts to outperform the predictive power of the traditional approaches by using the Taiwan Bankruptcy Prediction Dataset with the help of the Random Forest algorithm. By including the Synthetic Minority Over-Sampling Technique (SMOTE) into the procedure, the issue of class-imbalance is purposefully resolved, so that the model could achieve the right level of robustness and be implemented into real-life settings. The algorithm is embedded in Python, as well as the implementation is done using a Python-based technology stack (Scikit-Learn, Pandas, and Stream lit) to create a user-friendly dashboard. The system supports the automatic absorption of financial data, producing visual analytics, such as ROC curves, feature-importance plots, which automates business financial analysis and provides practical information to manage risks and make investment decisions. Finally, this project demonstrates the ability of modern machine-learning technologies to significantly improve the process of

bankruptcy identification at its initial stages and, thus, to help to create more stable financial conditions.

II. METHODOLOGY

This research methodology follows a structured pipeline that aims at converting raw financial data into viable bankruptcy risk judgments. It starts with the Data Upload Module whereby the users feed in financial data in CSV format. This is then followed by a strict Data Preprocessing Module which takes care of the pragmatic data problems including the treatment of missing data and transformation of categorical variables into a numerical format that can be processed by the algorithms. To deal with the acute problem of class imbalance, i.e. the large number of solvent firms compared with the number of bankrupt ones, the methodology combines the Synthetic Minority Over-sampling Technique (SMOTE). This method is used together with cross-validation to improve the accuracy of the model and to make sure that the model can accurately predict the minority group of failing firms.

The central part of the analytical engine is the Machine Learning Model Module, which uses the Random Forest Classifier. The choice of this ensemble learning algorithm is based on its capacity to model complicated and non-linear financial trends and its high performance relative to the traditional linear models. Taiwan Bankruptcy Prediction Dataset of the UCI Machine Learning Repository is used to train and validate the model. A robust Python-based technical stack is being used to implement it, that is, Pandas and NumPy to work with data, Scikit-Learn to deploy algorithms, and Matplotlib and Seaborn to visualize the statistics.

The last stage entails the Prediction and Visualization Modules, which are implemented through a web-based dashboard through the Streamlit framework. This interface gives users the estimated likelihood of bankruptcy of particular firms, and also creates visual representations including confusion matrices, ROC curves, precision recall curves, and a list of feature importance. This holistic solution is appropriate to make sure that the system is not just technically correct but also offers an interpretable solution that is needed in the banking, investment and corporate auditing industries.

III. MODELLING AND ANALYSIS

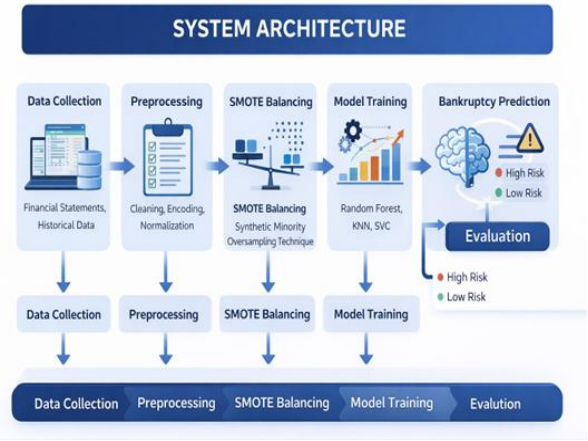
This investigation involves modelling and analysis phase that aims to work out how raw financial data can be transformed into a predictive framework that will detect corporate distress with a high degree of precision. The key to this process is the creation of a machine learning model using the Random Forest Classifier, which is an ensemble learning approach due to its ability to handle complex, non-linear relationships, which are often not well-modeled by more traditional linear-based approaches to machine learning. In order to achieve statistical strength, the methodology incorporates the Synthetic Minority Over-Sampling Technique (SMOTE) to level the difference between classes, and is performed with cross-validation to increase overall predictive performance. It uses the Taiwan Bankruptcy Prediction Dataset provided by the UCI Machine Learning Repository as the basis of its analysis, which gives a solid foundation of the historical financial values to train the model.

The implementation is modular whereby data is first processed to fill missing values and then convert categorical variables into numerical ones. The model is then trained using the Random Forest algorithm, and the most salient financial risk factors are emphasized. The final analytical product is visualized as a broad ensemble of measures such as confusion matrices, ROC curves, and precision-recall curves, which makes it easy to provide a fine-tuning evaluation of model performance. Besides, it includes a feature importance analysis in its process to highlight the particular financial indicators that are most predictive of the bankruptcy risk. Using these models as part of a Streamlit-powered dashboard will translate the difficult statistical analysis into insights anyone can use to help investors and financial organizations automate the process of risk assessment and make decisions based on actual data.

IV. ARCHITECTURE DESCRIPTION

The design of the bankruptcy prediction system is developed as a pipeline with sequential transformation that converts unprocessed financial information to a robust predictive system. It follows a highly organized path starting with data gathering, all the way to

balanced model training, and, thus, improves not only the predictive performance but also the practical usefulness.



4.1 Dataset

The first stage is the acquisition of financial and functional data, namely the use of Taiwan Bankruptcy Prediction Dataset received at the UCI Machine Learning Repository. This data has a wide range of features and financial indicators that define the financial health of the firm, such as such key variables as revenue, debt ratios, and a variety of financial ratios.

4.2 Data Preprocessing

Since raw financial data are often noisy, contain gaps, or outliers, a specific preprocessing module is used. This step methodically purifies the information by filling in the non-existent ones, balancing variables and converting categorical data into numerical values, and thus making the information ready to undergo further machine-learning processing.

4.3 Prepared Dataset

After preprocessing is complete, the dataset is officially declared a prepared dataset and it is strategically divided to two separate sets namely the training set which is used to build and refine the predictive model and the testing set which is used to test the model on unseen data to ensure that it possesses predictive competence.

4.4 K-fold Cross-Validation

In order to achieve efficiency in training and strength, K-fold cross-validation is used. The set of training will be separated into several folds, where the model is being trained on all but one-fold and tested on the remaining fold in a cyclical manner. This is an iterative process that ensures that the model generalizes well to novel data, and also alleviates over-fitting.

4.5 Oversampling Module

The financial data in this field are often highly imbalanced in class, with a small number of insolvent organizations compared to the number of solvent ones. In order to preclude prejudice against the majority group, an oversampling module using SMOTE (Synthetic Minority Over-Sampling Technique) is applied. This operation balances the proportion of classes in the training set allowing the Random Forest algorithm to more effectively identify the salient risk indicators that are linked to the minority bankruptcy class.

4.6 Training

The training step is a core element of the machine learning pipeline and it focuses on the creation of a model that is capable of predicting corporate bankruptcy with reasonable accuracy by learning the patterns of historical financial data. It starts with the usage of a carefully designed and balanced training set that serves as a starting point and on which machine learning algorithms learn these patterns. In the modelling phase, the dataset is periodically divided using 80-20 ratio of the data where 80 percent is used to train and the remaining 20 percent to test. This type of segmentation ensures that a large share of the data is allocated to model development, and at the same time, there is enough data to confirm that the model is performing well on unknown data.

4.7 Machine Learning Algorithms

Machine learning forms the basis of the analytical paradigm of the current research, shifting away not only an underlying linear model but also capturing the non-linear financial trends of a distressed corporation. Scikit-Learn makes it easier by providing the framework that facilitates the implementation of a variety of algorithms such as Random Forest, K-Nearest Neighbors, and Logistic Regression to analyze the Taiwan Bankruptcy Prediction Dataset. Since such datasets are inherently unequal based on the classes, the pipeline uses the Synthetic Minority Over-Sampling Technique (SMOTE) to artificially create examples that represent bankrupt entities, which in turn avoids the pipeline having a bias towards solvent firms.

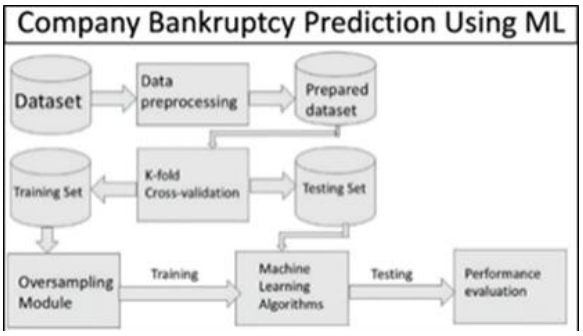
4.8 Testing

Testing phase is the final assessing component of machine learning pipeline wherein the trained model is tested on a separate untested dataset to determine its predictive fidelity in real-world conditions. This step is essential to ensure that the model has learned generalizable patterns and not just memorized the training data.

The following components define the testing process in this project as follows: Testing Set Allocation: In the process of data partitioning, around twenty percent of the set of data that has been prepared will be used in testing. This subgroup will not be affected in training and oversampling stages, thus protecting an objective evaluation.

V. PERFORMANCE EVALUATION

The algorithm used is the Random Forest to make inferences about the bankruptcy of the firms that are located on the testing set. The results are then compared with the predicted result- constituting the ground truth, to compute salient performance measures.



5.1 Importing modules

The bankruptcy prediction system is built based on a domain-specific collection of Python modules, each handling a different step in the data pipeline starting with preliminary work and concluding with deployment.

5.2 Manipulation of Data with Pandas.

Pandas is the main library of the data structures and analytical processes. It is used to read financial data that is in the form of CSV files into Data Frame objects, which allows them to be cleansed efficiently, missing values to be addressed, and complex manipulation of data to be performed.

5.3 Numico for Numerical Computing.

NumPy provides the basic array framework on which mathematical calculations are based. It is used to perform high-performance numerical operations and to operate the multi-dimensional arrays underlying the machine-learning models that are used to perform financial ratios and indicators.

5.4 Scikit-Learn Machine learning and evaluation.

Scikit-Learn is the main modelling framework that has the Random Forest Classifier and other related algorithms. Along with monitored training, it provides essential facilities to data partition, cross-validation process and compute the performance measures including accuracy and precision.

5.5 Class balancing Imbalanced-Learn (SMOTE).

It has the SMOTE (Synthetic Minority Over-Sampling Technique) module, which helps address the problem of the imbalance of the classes of bankruptcy data. SMOTE protects the model against bias towards the solvent firms by synthetically creating the samples of the minority population-bankrupt enterprises.

5.6 Statistical Visualization It Matplotlib and Seaborn.

The use of Matplotlib and Seaborn to convert complex numerical results into graphics is used. Matplotlib is used to do the basic plotting of graphs, whereas Seaborn provides a more high-level interface to creating statistically informative visualization, such as heat maps of confusion matrices and distributional plots of financial characteristics.

5.7 Streamlit Web Application Deployment.

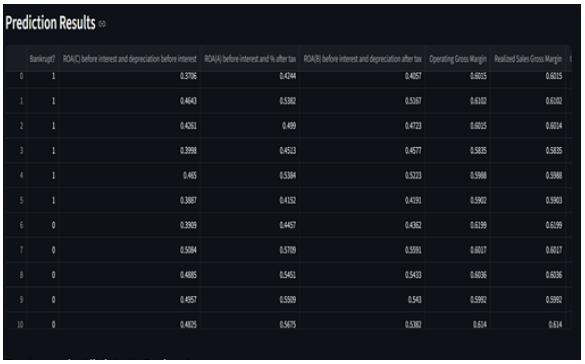
Streamlit is useful in translating the Python scripts that make up the underlying system into a web-based interactive dashboard. It supports the data upload module and the visualization module and provides an easy-to-use interface, which allows the stakeholders to obtain real-time predictions of the bankruptcy and the risk rating.

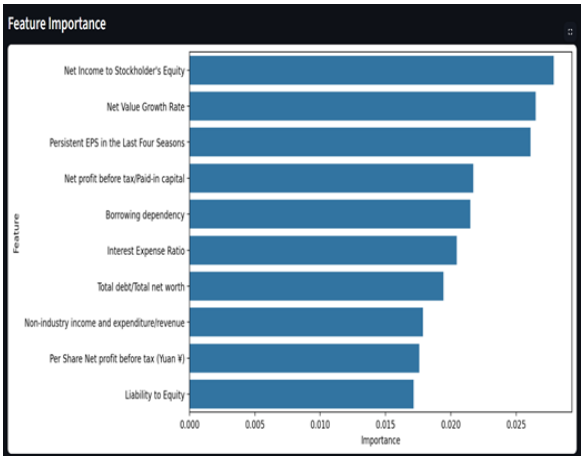
5.8 Warnings:

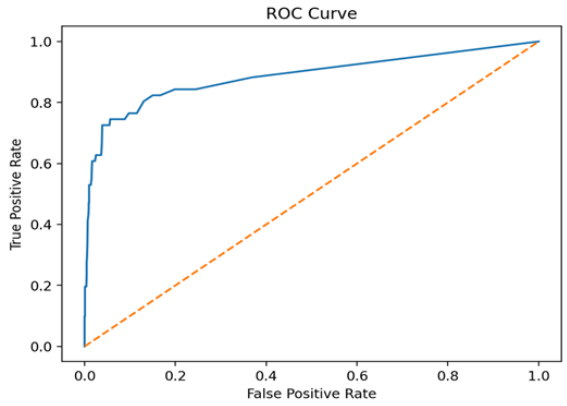
The Warnings module is used to regulate and suppress unnecessary or non-important alerts created in running and training the machine-learning model. The system can sift through all these warnings, leaving the end result and the Streamlit dashboard clean and thus enhancing user experience.

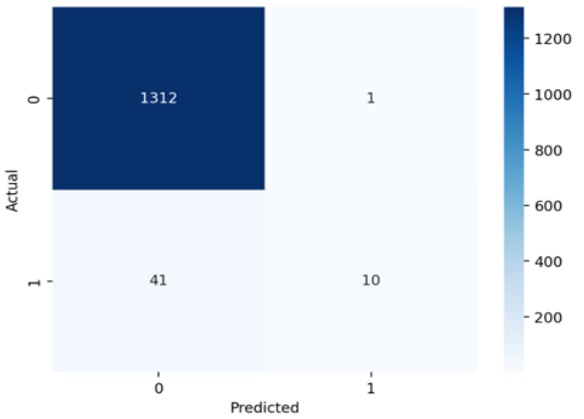
5.9 Output of Working Model

The exploratory data analysis used the visualization techniques to determine prominent trends which showed that some of the firms went into bankrupt ship even though they had more endowment than liabilities. More specifically, the Debt Ratio, Current Liabilities to Assets Ratio, and Current Liabilities to Current Assets Ratio showed strong positive relationships with bankruptcy. On the other hand, companies that had higher assets and revenue had significantly lower risk of bankruptcy as reflected by negative values of these characteristics. In order to question these trends, the various machine learning algorithms were tested to predict corporate failures such as Random Forest, K-nearest Neighbours (KNN) machine learning, Support Vector Classifier (SVC), Logistic Regression, Decision trees. The most accurate model was the Random Forest Classifier with the accuracy of 98.9% and the F1 score of 0.86 thus confirming the fact that its ensemble strategy is the most effective when it comes to predicting bankruptcy in the current dataset. K-Nearest Neighbours algorithm also presented the best results, achieving a precision of 98.02 and F1 of 0.78, compared to other models like SVC, Logistic Regression and Decision Trees with lower accuracy of 94.43, 88.86 and 84.68 respectively. Despite the easy interpretability of simpler models such as the Logistic Regression, the complexity of the Random Forest algorithm makes it more appropriate in terms of the nature of this data. These results highlight the strength of high-accuracy predictive modeling as the support mechanism to take critical decisions in the financial processes and to assess corporate health.









VI. CONCLUSION

The current study confirms that machine learning models present a strong and very precise framework to predict the bankruptcy of a corporation, and it significantly surpasses the limitations of the traditional manual financial analysis. With the assistance of Taiwan Bankruptcy Prediction Dataset and strict preprocessing steps along with the class-balancing techniques, e.g., SMOTE, the research demonstrates

the ability of fully automated systems to spot subtle features of financial distress that otherwise may be overlooked in traditional analysis. The results of the experiment show that many algorithms such as K-Nearest Neighbors, Support Vector Classifiers, and Logistic Regression offer some useful information but the Random Forest Classifier is the most outstanding tool, which achieved an accuracy score of 98.9. This extreme degree of accuracy and recall emphasizes the appropriateness of ensemble-based learning to the non-linear and complex features of financial data. Finally, integrating these models into an accessible Streamlit dashboard will provide a practical resource to stakeholders, investors, and banking institutions, as well as to make decisions based on data, reduce risks, and strengthen the overall economic environment.

VII. FUTURE ENHANCEMENTS

There are also major areas of enhancement and expansion in this project. The strength of the analysis could be enhanced using the multi-year financial data or by adding information on more jurisdictions. Further, the research can be refined by adding other types of data such as market trend indicators and corporate governance indicators that might provide more substantive results on bankruptcy prediction. Such technical improvements can include the use of methods like Principal Component Analysis (PCA) or sophisticated feature-engineering processes to reveal latent trends, and make the data more accessible. The project could also be expanded to an online real-time application, where the user would be able to input their financial data and receive real-time predictive results. The model is supposed to be designed to allow dynamic updates as additional information is received to ensure that the model retains fidelity and utility in a dynamic real-world environment.

It would be feasible to use more advanced algorithms, like XGBoost or LightGBM, as well as ensemble methods, which combine several models, which would significantly increase predictive accuracy. Moreover, making the interpretability frameworks such as SHAP or LIME would make the projections of the model significantly more understandable and clearer to the final users.

REFERENCES

[1] D. Wu, "Business Intelligence and Financial Analysis: Advances," J. Financial Eng., vol. 12, pp. 45–58, 2020.

[2] M. S. Devi, "Comparison of Machine Learning Models to Assess Financial Risk," Int. J. Data, vol. 8, pp. 112–125, 2019.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[4] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[7] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[8] S. R. Dubey, "A survey on deep learning to financial mathematics," Digit. Technol. Appl., vol. 2, pp. 101–115, 2021.

[9] V. Kotu and B. Deshpande, Data Science: Concepts and Practice, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2018.

[10] N. M. Krishna, "A novel approach to effectual emotion recognition with free of double truncated Gaussian mixture model and EEG," Int. J. Intell. Syst. Appl., vol. 6, pp. 33–42, 2017.

[11] N. M. Krishna, "Object detection and tracking with YOLO," in Proc. 3rd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), 2021, pp. 1234–1240.

[12] T. S. L. Prasad, K. B. Manikandan, and J. Vinoj, "Shielding NLP systems: In-depth survey on high-level AI-based methods of detecting adversarial attack in cyber security," in Proc. 3rd Int. Conf. Autom. Comput. Renewable Syst. (ICACRS), 2024, pp. 1–8.

[13] S. Sowjanya et al., "Bioacoustics signal authentication to e-medical records with blockchain," in Proc. Int. Conf. Knowl. Eng. Commun. Syst. (ICKECS), vol. 1, 2024, pp. 1–6.

[14] N. V. N. Sowjanya, G. Swetha, and T. S. L. Prasad, "Improved and AI-based vehicle

detection and classification in patterns with deep learning," in Disruptive Technologies in Computing and Communication Systems. Boca Raton, FL, USA: CRC Press, 2024.

[15] C. V. P. Krishna and T. S. L. Prasad, "Progress in predictive analytics in corporate governance," in Proc. Int. Conf. AI Finance, 2024, pp. 45–52.