Comparative Analysis of ESG Indicators in Enhancing Debt Default Prediction with Transformer-derived Attention Models

Zilin Zheng *

Due Diligence Department, Industrial Bank Co., LTD, Beijing, 100020, China

* Corresponding author: zhengzilin@cib.com.cn

Abstract. This study explores the application of Transformer-derived algorithms, namely TabNet and Tab Transformer, in predicting corporate debt defaults and compares their performance with classical machine learning models. Concurrently, it assesses the efficacy of incorporating ESG indicators into the corporate debt default risk assessment framework. The research focuses on Chinese A-share listed companies from January 2018 to June 2023, comprising 23 companies with observed defaults and a control group of 846 companies with regular debt maturities. The findings indicate that although attention-based models like TabNet and Tab Transformer provide enhanced interpretability, their performance does not significantly surpass ensemble algorithms such as XGBoost. Attention-based models emphasize the importance of merging the advantages of deep learning with the interpretability of traditional algorithms, especially when dealing with vast, high-dimensional datasets. Additionally, the incorporation of ESG data did not yield a significant improvement in prediction outcomes. Potential reasons for the limited impact of ESG indicators on predictions, including data quality and the comprehensiveness of existing financial disclosures, are discussed. Given the limited sample size and constraints related to test data, future research directions suggest expanding the dataset and diversifying ESG data sources.

Keywords: Corporate debt default, Transformer, Machine learning, ESG, TabNet, Tab Transformer.

1. Introduction

With the establishment of global objectives such as carbon neutrality and peaking of carbon emissions, China's regulations on corporate environmental, social responsibility, and corporate governance disclosures are becoming more refined. Enterprises and capital markets are placing increasing importance on practicing green development concepts, promoting the development of Environmental, Social and Governance (ESG) in China's capital markets. However, current global theoretical research on ESG still lags behind ESG activities in the credit market. There is a strong correlation between banks' social responsibilities, environmental protection duties, profitability, and green lending [1]. However, financial institutions lack efficient ESG-related credit evaluation systems and models to control and measure corporate default risks.

Currently, many studies have demonstrated that introducing ESG development indicators into credit risk assessment systems better reflects a company's current credit level, reducing credit risks for financial institutions. Jang et al. found that ESG is a beneficial complement to credit quality assessments, suggesting credit rating agencies should incorporate ESG scores into risk assessment systems [2]. Evangelinos and Nikolaou argued that including ESG scores in corporate credit risk assessments effectively reduces the credit risks faced by banks [3]. Aslan et al. found that companies with high ESG scores have a much lower probability of credit default than those with low scores [4]. Yeo-Hwan et al. studied 1,879 companies from 2010-2012 and found that companies with higher ESG ratings displayed better credit rating results [5]. They suggested ESG ratings are negatively correlated with default risk. Henisz et al. found significant correlations between major corporate credit events and the ESG scores before the occurrence of such events. However, some researchers are skeptical about ESG evaluations revealing corporate default risks [6]. Atif and Ali found that, for some traditional companies, ESG disclosures negatively impacted corporate default risks [7].

In recent years, with the widespread development of artificial intelligence, represented by machine learning and deep learning, there have been breakthroughs in computer vision and natural language processing. More and more credit risk studies are incorporating machine learning into credit

assessment systems, achieving remarkable results. Technologies like logistic regression, XGBoost, Support Vector Machine (SVM), random forests, and neural networks have shown their value in credit risk identification. Makowski was the first to apply the decision tree to personal credit assessments, finding the decision tree offers good interpretability for predicting customer loan defaults [8]. Odom et al. found the neural network powerful in processing credit risk predictions, especially for cases with complex nonlinear relationships between features [9]. Sang applied SVM and BP neural network to assess the credit risks of small businesses in supply chain finance and found SVM suitable for small samples with missing or abnormal data, offering high data processing efficiency and stability [10]. Mohammed and Shazli applied various machine learning models to predict customer loan defaults, finding machine learning models to be more stable than shallow neural networks when handling imbalanced datasets [11]. At the same time, cutting-edge technology researchers have begun to try to use deep learning methods to build loan default models [12].

With the rapid iteration of artificial intelligence, models based on the self-attention mechanism, represented by Transformers, have replaced traditional machine learning and deep learning algorithms in NLP and CV. However, in the credit default assessment system, research and applications of such model lag behind other fields, hindering the construction of intelligent credit risk systems for financial institutions. Huang et al. introduced the TabTransformer, a Transformer-based model architecture for tabular data that uses contextual embeddings to capture complex relationships between features, achieving significant performance improvements on multiple tabular datasets [13]. Arik et al. introduced TabNet, which while retaining the end-to-end and representational learning features of DNNs, also offers the interpretability and sparse feature selection advantages of decision tree models, achieving high accuracy in classification prediction tasks when tested on real datasets [14].

This project compares the performance of self-attention algorithms, represented by Tab Transformer and TabNet, with other classic machine learning algorithms in the field of corporate debt default prediction. Simultaneously, the study compares the corporate credit evaluation system incorporating ESG indicators with the traditional model only containing financial data, assessing the impact of ESG indicators in corporate debt default prediction.

2. Methodology

2.1. Data Source and Description

Given the emphasis on data authenticity and availability, the study comprises a sample of 23 Chinese A-share listed companies that defaulted between January 2018 and June 2023. Additionally, 846 companies, which have honored their bond obligations in the duration, serve as control samples. Since the companies under review are publicly-listed A-share entities, their financial and operational data, sourced from voluntary disclosures, ensure transparency and veracity.

Table 1. Eleven Categories of Financial Attributes

Index	Aspect	Number of Indicators
1	Ratio Structure	34
2	Solvency	26
3	Growth Potential	30
4	Risk Level	7
5	Dividend Distribution	8
6	Operating Ability	32
7	Per-share Indicators	38
8	Disclosed Financial Indicators	9
9	Cash Flow Analysis	28
10	Relative Value Indicators	14
11	Profitability	54

The study leverages 425 financial attributes across 11 categories to holistically represent the non-ESG operational landscape of these companies (Table 1). Despite the array of ESG indicators available, no unified standards exist. For research consistency, the study employs ESG data for China's A-share listed companies from Bloomberg. Bloomberg's ESG evaluations cover environment, society, and governance aspects, culminating in an aggregated score. It is pivotal to note that Bloomberg's ESG dataset is processed, with the original data undisclosed. Bond default data, as well as normal bond maturity information, were sourced from the Wind financial client, which also provided the Bloomberg ESG data and other financial metrics.

2.2. Introduction to Methods

Previous research has delved into applying ESG indicators for debt default prediction using diverse machine learning algorithms. Recently, transformer-derived self-attention algorithms, particularly prevalent in the domains of Natural Language Processing (NLP) and Computer Vision (CV), have yet to be explored in enterprise debt default prediction. While Transformer models were primarily conceived for sequential data and achieved notable success in NLP, the flexibility of their architecture and the efficacy of the self-attention mechanism render them promising for tabular data. This adaptability has led to the application of Transformer's core concepts, especially the self-attention mechanism, across diverse data types, including images, time series, and tables. Recent advancements, such as Tab Transformer and TabNet, are tailored specifically for tabular data [13, 14]. This study endeavors to implement Tab Transformer and TabNet in corporate debt default prediction and compare their performance with six other established machine learning algorithms: XGBoost, Random Forest, Support Vector Machine (SVM), Adaboost, Deep Neural Network (DNN), and Catboost. Beyond a comparative analysis of algorithmic performance, this study elucidates the potential enhancement in debt default prediction by incorporating ESG indicators. ESG remains a contentious indicator. While certain researchers champion the revelatory capacity of ESG indicators in discerning corporate credit risks and elevating operational standards, some express reservations. Initially, this research employs the mentioned eight models to predict debt defaults solely using corporate finance and operational data. Subsequently, a holistic prediction integrates ESG data with corporate finance and operational data. Ensuring sample veracity, debt default encompasses both substantial default and extensions.

Corporate credit default inherently presents a binary classification challenge. Amongst the plethora of machine learning and deep learning algorithms, it is essential to identify reasonable models, demanding a robust evaluation metric system pertinent to credit risk control. This paper employs the AUC, which is particularly valuable in corporate credit default prediction. Besides conventional classification metrics like AUC, F1, Recall, and Precision, it is vital to comprehensively assess model efficacy in credit risk control. Given that credit defaults are relatively infrequent, positive and negative samples are imbalanced. AUC, considering both positive and negative classifications, provides a balanced evaluation even in the face of such imbalance, as evidenced in prior research on corporate credit defaults. An AUC value lies between 0.5 and 1, with proximity to 1 indicating superior model authenticity. An AUC value of 0.5 suggests minimal authenticity, meaning the model is ineffectual. Evidently, a higher AUC signifies a more effective classifier.

2.3. Data Processing

Problems like missing or anomalous data were prevalent during data acquisition, potentially undermining model accuracy. To enhance model performance, data preprocessing - encompassing blank value management and outlier processing - becomes essential. This research, aiming to mirror real-world business scenarios, exclusively employs zero-filling for vacancies without outlier treatment.

Utilizing eight distinct algorithms, this research predicts debt defaults of Chinese A-share listed companies spanning from January 2018 to June 2023. By comparing the outcomes from the eight models - with and without the inclusion of ESG indicators - this study discerns the significance of

ESG information in recognizing corporate debt default risks and evaluate the algorithms' efficacies in this domain. The predictive model adopts indicators from the preceding year (t-1 period) to predict default risks for the current year (t period). For instance, indicators in the year of 2017 determine the 2018 default risk. This method is uniformly applied to both defaulting and non-defaulting companies. For evaluation robustness, data is partitioned randomly into training (80%) and testing (20%) sets. Given the data imbalance - with fewer defaulting companies, the ADASYN algorithm is employed to balance positive and negative sample proportions. Distinct from conventional oversampling methods, ADASYN, derived from SMOTE, generates samples based on distribution density, generating more synthetic samples in areas with lower distribution density to increase the number of minority class samples [15]. Post ADASYN optimization, hyperparameters for the eight algorithms were tuned using grid search.

3. Results and Discussion

This study systematically contrasts the eight machine learning algorithms' efficacy in predicting debt defaults of China's A-share listed companies. Incorporating ESG indicators, the research observes ESG's potential to augment machine learning performance in default prediction. AUC, a prevalent metric in machine learning, serves as the primary evaluation criterion, supplemented by accuracy, precision, recall, and F1 score. To explain the intuitive meaning of the above metrics, the concept of a confusion matrix, represented as a 2 by 2 matrix due to the binary nature of bond defaults, is introduced.

3.1. Display of Confusion Matrices

In the shown confusion matrix, '0' signifies the occurrence of debt default, whereas '1' indicates the absence of default. The x-axis represents predicted values, while the y-axis captures the actual outcomes. Figure 1 and 2 portray the confusion matrices for the eight machine learning algorithms, incorporating ESG indicators or excluding them respectively.

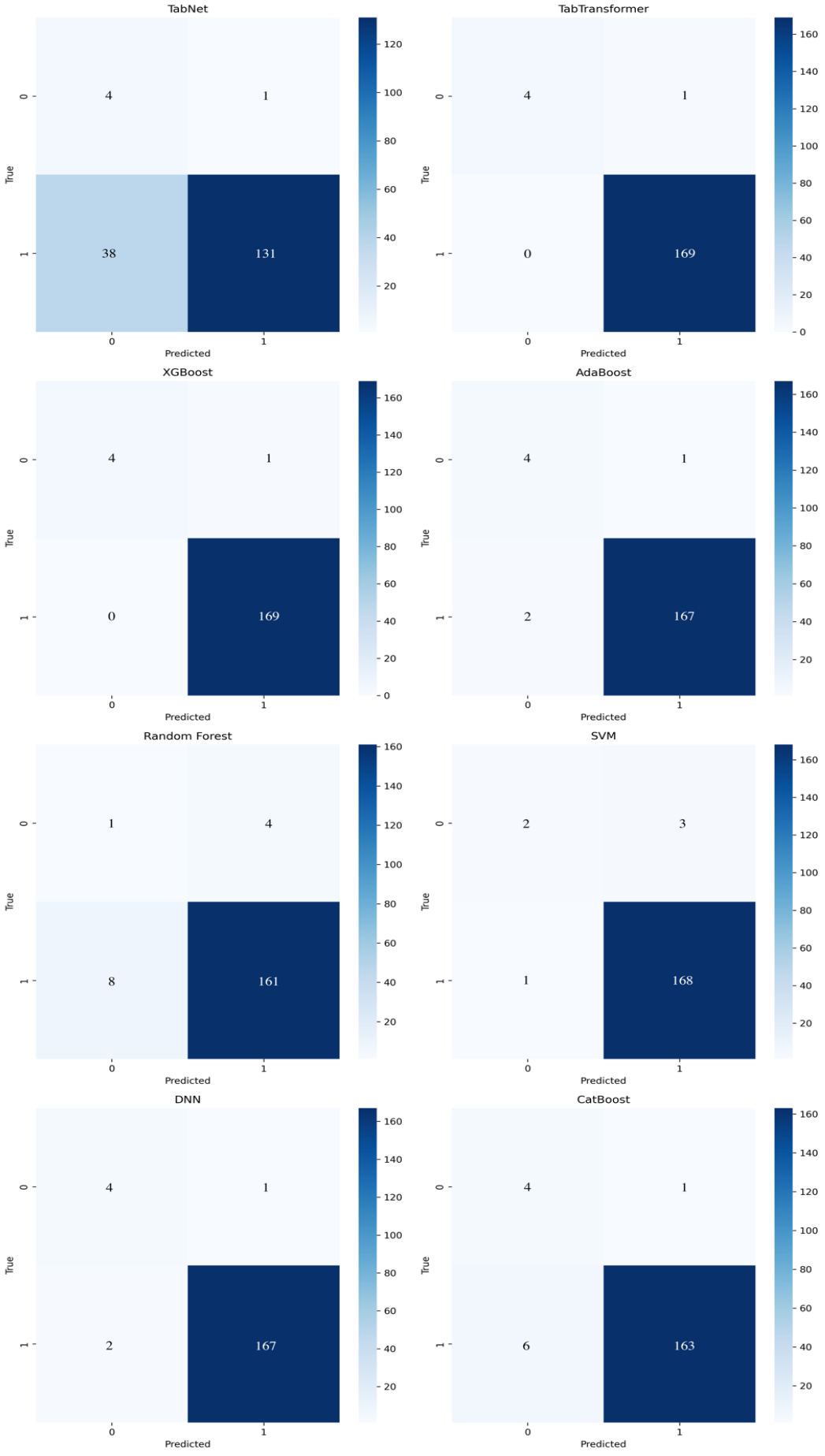

Figure 1. The confusion matrices of data without ESG

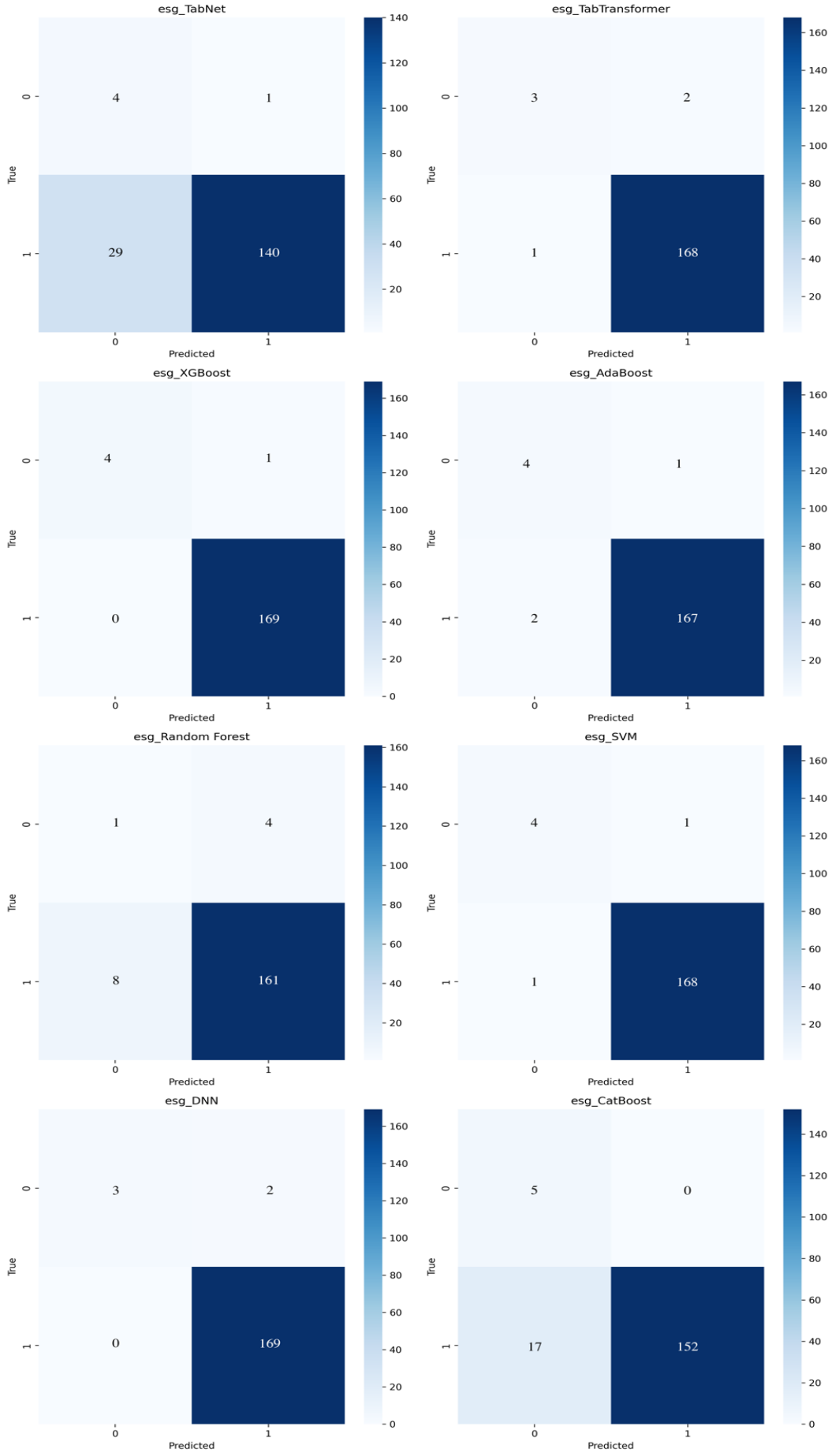

Figure 2. The confusion matrices of data with ESG

3.2. Comparison of Various Algorithms

Table 2 delineates the performance metrics of the eight algorithms. Without ESG indicators, both TabTransformer and XGBoost significantly outperform other algorithms, with Adaboost, DNN, CatBoost, and TabNet following closely. Interestingly, models such as TabNet, TabTransformer, XGBoost, Adaboost, DNN, and CatBoost demonstrate comparable performance, with AUC values ranging between 0.865 and 0.90 in AUC. The self-attention models, TabNet and TabTransformer, exhibit comparable results in the overall performance to traditional models like XGBoost.

Table 2. Performance Metrics of Algorithms

Model	Accuracy	Precision	Recall	F1	AUC
TabNet	0.9253	0.9937	0.9290	0.9602	0.8645
TabTransformer	0.9943	0.9941	1.0000	0.9971	0.9000
XGBoost	0.9943	0.9941	1.0000	0.9971	0.9000
AdaBoost	0.9828	0.9940	0.9882	0.9911	0.8941
Random Forest	0.9310	0.9758	0.9527	0.9641	0.5763
SVM	0.9770	0.9825	0.9941	0.9882	0.6970
DNN	0.9828	0.9940	0.9882	0.9911	0.8941
CatBoost	0.9598	0.9939	0.9645	0.9790	0.8822

While TabNet and TabTransformer did not yield significant performance growth compared to traditional ensemble learning algorithms (e.g., XGBoost) or deep learning algorithms (DNN), they did display enhanced interpretability. This provides an edge, especially when counterposed against algorithms like XGBoost or DNN. The inherent nature of TabNet, with its "decision steps," facilitates visualization and explanation of the decision-making process. The Transformer architecture, although primarily designed for sequence data, offers a modicum of interpretability through its self-attention mechanism. Particularly, when dealing with extensive and high-dimensional data, TabNet and TabTransformer hold a distinctive advantage over traditional machine learning, synthesizing the benefits of deep learning with the interpretable capabilities of traditional algorithms.

The results suggest that both attention-based and other algorithms have approached an optimal performance level with AUC values between 0.86 and 0.9. It can be inferred that, given the limited test data, the performance gap between algorithms might not be discernible. The lack of test data may restrain the elucidation of performance differences between the various algorithms. This study may need more data to test the performances of various algorithms.

3.3. Effectiveness of ESG Indicators

To elucidate the value of ESG indicators, the study compares algorithmic performance across two systems: one inclusive of ESG data and another exclusive of it. Models inclusive of ESG data are labeled 'esg_algorithm' (e.g., esg_XGBoost), while their counterparts without ESG data retain their original names (e.g., XGBoost). The comparative analysis is illustrated in Table 3.

Of the eight models evaluated, the AUC values for two models showed an increase, two evidenced a decrease, and the remaining four remained static. Notably, only the Catboost and SVM models display an uplift in AUC integrating ESG indicators, increasing from 0.8822 to 0.9479 and from 0.6970 to 0.8970, respectively. While the esg_Catboost yielded the highest AUC value in these experiments, its accuracy, recall, and F1 score demonstrated an apparent drop compared with Catboost. The confusion matrices for Catboost and esg_Catboost reveal that esg_Catboost impeccably classified all negative samples associated with debt defaults. However, it also misclassified a more significant number of positive samples (those without debt defaults) in comparison to the Catboost, which, in turn, leads to its superior AUC yet inferior performance on other metrics.

Table 3. Comparison of Algorithm Performance Incorporating ESG Indicators

Model	Accuracy	Precision	Recall	F1	AUC
TabNet	0.9253	0.9937	0.9290	0.9602	0.8645
TabTransformer	0.9943	0.9941	1.0000	0.9971	0.9000
XGBoost	0.9943	0.9941	1.0000	0.9971	0.9000
AdaBoost	0.9828	0.9940	0.9882	0.9911	0.8941
Random Forest	0.9310	0.9758	0.9527	0.9641	0.5763
SVM	0.9770	0.9825	0.9941	0.9882	0.6970
DNN	0.9828	0.9940	0.9882	0.9911	0.8941
CatBoost	0.9598	0.9939	0.9645	0.9790	0.8822
esg_TabNet	0.8276	0.9929	0.8284	0.9032	0.8142
esg_TabTransformer	0.9828	0.9882	0.9941	0.9912	0.7970
esg_XGBoost	0.9943	0.9941	1.0000	0.9971	0.9000
esg_AdaBoost	0.9828	0.9940	0.9882	0.9911	0.8941
esg_Random Forest	0.9310	0.9758	0.9527	0.9641	0.5763
esg_SVM	0.9885	0.9941	0.9941	0.9941	0.8970
esg_DNN	0.9828	0.9940	0.9882	0.9911	0.8941
esg_CatBoost	0.9023	1.0000	0.8994	0.9470	0.9497

It assumes that the set of AUC without ESG indicators is a set of data, and the set of AUC that contains ESG indicators is another set of paired data. To assess the significance of the divergence between AUC sets, this study conducted a normality test. Results from the Shapiro-Wilk test suggested that neither group of data conformed to the assumption of normal distribution ($p1 = 0.00137$, $p2 = 0.01358$). Given the paired nature of the data and its deviation from a normal distribution, the Wilcoxon signed-rank test was chosen, owing to its aptitude for handling paired datasets without the constraints of normality. The Wilcoxon signed-rank test revealed no significant difference between the two paired groups ($W = 4.0$, $p = 0.715 > 0.05$). Based on these evaluations, it is concluded that ESG indicators do not markedly enhance the model's predictive capability within the context of the eight models tested.

Meanwhile, there are several potential reasons for the delicate impact of ESG indicators on debt default predictions. Firstly, a limited dataset might lead to large fluctuations in models. The incremental enhancements offered by ESG might not compensate for this fluctuation. Secondly, existing financial and operational data disclosures by listed companies might be comprehensive enough to assess corporate debt default risks, making ESG data redundant to some extent. In fact, ESG indicators also include some financial and operational-related information in some degree. Third, concerns related to the quality of Bloomberg ESG data might compromise the model's forecasting ability. Finally, ESG indicators might be risk-neutral when predicting corporate credit defaults. It is readily apparent that the level of ESG indicators can be influenced by the specific industry in which a company operates. While certain sectors are characterized by inherently low ESG scores, companies within these sectors often demonstrate robust operating conditions and significant performance capabilities. This observation is particularly prevalent among traditional heavy industries. Conversely, in industries or companies with inherently high ESG scores, the elevated ESG attributes may not directly correlate with their financial performance.

4. Conclusion

This study discusses corporate debt default prediction, drawing upon algorithm models derived from the Transformer. It counterposes these with classic machine learning models and attempts to integrate ESG (Environmental, Social, and Governance) information into the corporate debt default prediction framework. Based on data availability, the study focused on China's A-share listed companies from January 2018 to June 2023. Within this timeframe, it is observed 23 listed companies with substantial defaults and extensions, constituting our primary research group. Concurrently, a

control group of 846 Chinese A-share listed companies that saw normal debt maturities during the same period was established. Due to the imbalanced nature of the sample distribution, the study employed the ADASYN algorithm to optimize the training set, ensuring an even ratio between positive and negative samples. Beyond the comparative analysis of the performance of eight machine learning algorithms, this study also contrasts models with and without ESG information. The aim is to identify the efficacy of ESG indicators in predicting debt defaults. Bloomberg ESG data served as the source of ESG indicates for China's A-share listed companies in the study.

Key findings from the study revealed that attention-based models like TabNet and TabTransformer did not exhibit any significant performance enhancements in corporate debt default prediction when compared with ensemble learning algorithms typified by XGBoost. Additionally, the Wilcoxon test illustrated that the incorporation of Bloomberg ESG data did not significantly promote corporate debt default predictions based on machine learning.

Given the scope of this study, centered around China's A-share listed companies, the sample size remains insufficient, especially for negative instances. While the ADASYN algorithm aids in optimizing the training set, the test set remains constrained. Future research could benefit from a larger dataset to refine the accuracy of model evaluation metrics. Moreover, as ESG evaluation indicators lack standardization, subsequent research might benefit from integrating diverse ESG data sources, evaluating their impact on corporate debt default predictions.

References

[1] Carraro C, Favero A, Massetti E. Investments and public finance in a green, low carbon, economy. Energy Economics, 2012, 34: S15 - S28.

[2] Jang G, Kang H, Lee J, Bae K. ESG scores and the credit market. Sustainability, 2020, 12 (8): 3456.

[3] Evangelinos K, Nikolaou I. Environmental accounting and the banking sector: a framework for measuring environmental-financial risks. International Journal of Services Sciences, 2009, 2 (3-4): 366 - 380.

[4] Aslan A, Poppe L, Posch P. Are sustainable companies more likely to default? Evidence from the dynamics between credit and ESG ratings. Sustainability, 2021, 13 (15): 8568.

[5] Kim Y, Kim M. The Impact of Default Risk on Corporate Social Responsibility: Evidence from Korean Firms. Journal of Industrial Economics and Business, 2014, 27 (5): 2103 - 2115.

[6] Henisz W, McGlinch J. ESG, material credit events, and credit risk. Journal of Applied Corporate Finance, 2019, 31 (2): 105 - 117.

[7] Atif M, Ali S. Environmental, social and governance disclosure and default risk. Business Strategy and the Environment, 2021, 30 (8): 3937 - 3959.

[8] Makowski P. Credit scoring branches out. Credit World, 1985, 75 (1): 30 - 37.

[9] Odom M, Sharda R. A neural network model for bankruptcy prediction. 1990 IJCNN International Joint Conference on neural networks. IEEE, 1990, 163 - 168.

[10] Sang B. Application of genetic algorithm and BP neural network in supply chain finance under information sharing. Journal of Computational and Applied Mathematics, 2021, 384: 113170.

[11] Azhan M, Meraj S. Credit card fraud detection using machine learning and deep learning techniques. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2020: 514 - 518.

[12] Freedman S, Jin G. The information value of online social networks: Lessons from peer-to-peer lending. International Journal of Industrial Organization, 2017, 51: 185 - 222.

[13] Huang X, Khetan A, Cvitkovic M, et al. Tab transformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv: 2012.06678, 2020.

[14] Arik S, Pfister T. Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI conference on artificial intelligence, 2021, 35 (8): 6679 - 6687.

[15] He H, Bai Y, Garcia E, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks, 2008, 1322 - 1328.