International Journal of Advanced Technology in Engineering and Science
Vol. No. 11, Issue No. 04, April 2023
www.ijates.com
ijates
ISSN 2348 - 7550

Predictive and Probabilistic Approach for Loan Prediction

G.Rajani Priya, Ch. Rajesh,B. Mani Meghana,B. Mohana Krishna

*(Under the guidance of **Mrs. J. Lakshmi**, Assistant Professor,*
Department of Computer Science and Engineering, Tirumala Engineering College)

ABSTRACT

As the amount of data increases, it is more likely that the assumptions in the existing economic analysis model are unsatisfied or make it difficult to establish a new analysis model. Therefore, there has been increased demand for applying the machine learning methodology to bankruptcy prediction due to its high performance. By contrast, machine learning models usually operate as black-boxes but credit rating regulatory systems require the provisioning of appropriate information regarding credit rating standards. If machine learning models have sufficient interpretablility, they would have the potential to be used as effective analytical models in bankruptcy prediction. From this aspect, we study the explainability of machine learning models for bankruptcy prediction by applying the Local Interpretable Model-Agnostic Explanations (LIME) algorithm, which measures the feature importance for each data point. To compare how the feature importance measured through LIME differs from that of models themselves, we first applied this algorithm to typical tree-based models that have ability to measure the feature importance of the models themselves. We showed that the feature importance measured through LIME could be a consistent generalization of the feature importance measured by tree-based models themselves. Moreover, we study the consistency of the feature importance through the model's predicted bankruptcy probability, which suggests the possibility that observations of important features can be used as a basis for the fair treatment of loan eligibility requirements.

I. INTRODUCTION

Owing to the importance in measuring corporate solvency, bankruptcy prediction has been a widely studied topic inthe field of finance and economics [1], [2]. The bankruptcy prediction model, which predicts whether a company will gobankrupt, must meet two main requirements, high accuracy, and interpretability [3]. Because it is important to creditors, investors, and banks, a clear interpretation of the results is akey aspect in determining whether the model is usable in theindustry.

During the early stage, researchers mainly focused on a small number of features and the statistical models. For instance, Altman [4] and Altman *et al.* [5] used a multiple discriminant analysis, and Ohlson [6]

created a model based on a logistic approach. With an increase in the number of available features (e.g., financial ratios), a clear interpretation issue has arisen. In general, a small number of indepen- dent variables and a simple model were required for a clear interpretation of the model. As a consequence, many studies attempting to select the most [12]–[15]. These two branches, namely, feature selection based approach and machine learning based approach both have their own pros and cons.

Feature selection based methods are easily interpretable because they use a few number of variables that are cho- sen as relevant to a bankruptcy prediction. Feature selection based methods usually rely on a simple predictive model, such as a simple multivariate function. However, compared to the machine-learning based models, the accuracy is much lower. By contrast, although the machine-learning based methods attain a higher accuracy, such models are too com-plex to be clearly interpreted. Recently, Son *et al.* [3] sug- gested a way to overcome the lack of interpretability of the machine-learning based approaches by leveraging feature importance techniques for boosting tree models [16], [17]. This study enables one to interpret the results of an extremelycomplicated bankruptcy prediction model, but their result remains a model-wise interpretation.

II. DATA

DATA DESCRIPTION

In this study, we used data on Korean companies ranging from 2009 to 2015, provided by the Douzone Bizon ICT Group, which services enterprise resource planning (ERP) and accounting service tools. The data to be analyzed include accounting information of not only corporate but also indi- vidual businesses. As for the composition ratio, corporations account for 61.9% and private enterprises account for 38.1%.The number of data increased from 81 in 2009 to 196,611 in 2015, which is a result of the increase in the number of customers using the Douzone Bizon ERP service. We use the financial ratios gathered from the Douzone data for the features. In this paper, we classified our data into two groups, namely, corporations and private enterprises, but, when train-ing our models, we divided the data on the corporations into two sub-groups, namely, medium or large corporations, and small corporations to achieve a high performance. The medium or large corporations and small corporations were segmented into increments of 2 billion won (Korean cur- rency) in sales. Details are given in Table 1.

III. LIME

LIME is a method for trying to interpret a given black-box model locally through linearization. As the basic idea here, if we need a trained model f to be explained at an instance x, we approximate this model f within the region near x by another relatively simple and explainable model g. We describe this method briefly in this section, the general procedure of which is drawn in Figure 1.

IV. EMPIRICAL RESULTS

We trained two black-box models XGB and LightGBM on three different datasets (Medium or Large Corporation, Small Corporation, and Private Enterprise). The classification results of each model on each training dataset using a 5-fold cross validation are given in Table 3. The AUC socres were sufficiently high, and thus we concluded that our models were trained well. In our experiment, the performances of the mod- els in each fold were

similar. Hence, we fixed one fold and trained our models on that fold to compare its ability to select the feature importance using the LIME approach to measur-ing the feature importance. The fixed fold data distribution is briefly described as follows. For medium or large corpora-tions, among the training set of size 90613, 2266 companies went bankrupt and among the test set of size 23220, 530 com- panies went bankrupt. For small corporations, among the training set of size 177806, 7207 companies went bankrupt and among the test set of size 44452, 1835 companies went bankrupt. For private enterprises, among the training set of size 166609, 3692 companies went bankrupt and among the test set of size 41653, 937 companies went bankrupt. Having these trained black-box models, we set the length of expla- nation *K* to 20 in our experiment. The higher *K* we choose, the lower the interpretability of models. We heuristically chose *K* 20 believing that this is a compromise between these two.

V. CONCLUSION AND DISCUSSION

By experimenting with representative tree-based models, XGB and LightGBM, it has been shown that the method tree-based models measuring feature importance model-wisemanner can be sufficiently reproduced using LIME. Because LIME is applicable to any model even if the model does not have the ability to measure feature importance itself, our experiment shows that a feature importance can be meaningfully extracted from models such as a neural net.Based on this, not limited to tree-based models, we expect that the feature importance can be meaningfully extracted byusing LIME on models that performs better.

Moreover, by comparing the results obtained by apply- ing LIME on XGB and LightGBM based on the predicted bankruptcy probabilities of the model, we showed that Light- GBM is more suitable than XGB for consistently estimatingthe feature importance for the predicted bankruptcy proba- bilities. We believe this result will be useful in practice. For example, if credit rating results are an important factor in deciding whether to approve a loan, the observed values of the important features will be used as the basis for fair treatment of loan eligibility requirements.

REFERENCES

[1] J. L. Bellovary, D. E. Giacomino, and M. D. Akers, ''A review of bankruptcy prediction studies: 1930 to present,'' *J. Financial Educ.*, vol. 33, pp. 1–42, Dec. 2007.

[2] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akinade, and M. Bilal, ''Systematic review of bankruptcy prediction models: Towards a framework for tool selection,'' *Expert Syst. Appl.*, vol. 94, pp. 164–184, Mar. 2018.

[3] H. Son, C. Hyun, D. Phan, and H. J. Hwang, ''Data analytic approach for bankruptcy prediction,'' *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112816.