Explainability of Machine Learning Models for Bankruptcy Prediction

MIN SUE PARK\(^1\), HWIJAE SON\(^2\), CHONGSEOK HYUN\(^3\), AND HYUNG JU HWANG\(^1,4\)

\(^1\)Department of Mathematics, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea
\(^2\)Stochastic Analysis and Application Research Center, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
\(^3\)BNK Financial Group Inc., Busan 48400, Republic of Korea
\(^4\)Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea

Corresponding author: Hyung Ju Hwang (hjhwang@postech.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grants through Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2017R1E1A1A03070105 and Grant NRF-2019R1A5A1028324, in part by the Institute for the Information and Communications Technology Promotion (IITP) Grant through Korean Government [Ministry of Science, ICT and Future Planning (MSIP)] [Artificial Intelligence Graduate School Program (Pohang University of Science and Technology (POSTECH))] under Grant 2019-0-01906, and in part by the Information Technology Research Center (ITRC) Support Program under Grant IITP-2020-0-01441.

\section*{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 interpretability, 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.

\section*{INDEX TERMS}
Bankruptcy prediction, machine learning, explainable AI, feature importance.

\section*{I. INTRODUCTION}
Owing to the importance in measuring corporate solvency, bankruptcy prediction has been a widely studied topic in the field of finance and economics [1], [2]. The bankruptcy prediction model, which predicts whether a company will go bankrupt, 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 a key aspect in determining whether the model is usable in the industry.

During the early stage, researchers mainly focused on a small number of features and the statistical models. For instance, Altman [4] and Altman \textit{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 independent variables and a simple model were required for a clear interpretation of the model. As a consequence, many studies attempting to select the most relevant features and model the bankruptcy based upon the selected features and a simple statistical model have been reported [7]–[11]. Another way to deal with large numbers of features is to apply machine learning algorithms [3], [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 chosen 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 complex to be clearly interpreted. Recently, Son et al. [3] suggested 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 extremely complicated bankruptcy prediction model, but their result remains a model-wise interpretation.

There is one clear limitation of the model-wise interpretation. It is impossible to track the important features company by company in a model-wise interpretation scenario. Therefore, as a good alternative, instance-wise interpretation has been spotlighted in the machine learning community. Although there are many studies regarding the interpretability of machine learning algorithms (e.g., [18]–[20]), we focus on an instance-wise local interpretation method. In the previous work of Ribeiro et al. [21], the authors proposed a local interpretation method called local interpretable model-agnostic explanations (LIME). LIME can generate an instance-wise explainable prediction of any classifier by learning a locally interpretable model. Compared to general sensitivity analysis explaining the models themselves, LIME has an advantage in that it gives an explanation for each data point.

In this study by leveraging the advantage of LIME, we propose a novel, highly accurate, and instance-wise interpretable bankruptcy prediction model. The proposed model meets the two aforementioned requirements of high accuracy and interpretability. The experiment results show that the instance-wise interpretation of a LightGBM (or XGBoost) based bankruptcy prediction model is mostly consistent with the model-wise interpretation, which implies that the instance-wise interpretation is reliable. We also empirically show that instance-wise feature importance is more robust along with the predicted probability when equipped with the LightGBM-based model than with the XGBoost-based approach. Moreover, the experiments show that the important feature distribution is similar in the training and testing data, which implies that our instance-wise interpretation is robust to a random splitting of the data.

The rest of this paper is organized as follows. In section II, we provide information regarding the data we used. In section III and IV, we briefly introduce the tools we used in our experiment, including LIME. In section V, we present the methodology how we preprocessed our data. In section VI, we present results and a comparison between instance and model-wise interpretations. In section VII, we present the some concluding remarks regarding this research.

The main contribution of our work comprises the following items.

1) For bankruptcy prediction problem, it is important to provide a reason for the judgment. By demonstrating that the method by which tree-based models measure feature importance in a model-wise manner can be sufficiently reproduced using LIME on bankruptcy dataset, we showed the possibility that the feature importance can be meaningfully extracted by using LIME on other models that do not have the ability to measure feature importance themselves but perform better.

2) Since credit regulatory systems require the provision of appropriate information on credit rating standards, we empirically showed that a model with a relatively high consistency in the selection of feature importance can be chosen by applying the LIME method to black-box models such as XGB and LightGBM.

II. DATA

A. 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 individual 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 training 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 currency) in sales. Details are given in Table 1.

B. FEATURE DESCRIPTION

There are 110 features, 6 of which are categorical features, labeled type_1, type_2, type_3, type_4, type_5, and type_6, respectively. These features have values of zero or 1, indicating whether a company’s business is of the corresponding type. Among the given features, important features used in previous studies related to bankruptcy prediction, such as [4], [5], or [6] are included. Of the 110 features used in our study, 28 use information from a study by Lee and Kim [28], which systematically arranged suitable features based on a study on the bankruptcy characteristics of Korean companies. A comparison of the relation between the features we used and the features of other previously analyzed papers is given in Table 2.
TABLE 1. Ratio of bankrupt companies by year for 2009-2015.

| Year | Medium or Large Corporation | Small Corporation | Private Enterprise | Total |
|------|-----------------------------|-------------------|-------------------|-------|
|      | number of companies bankrupt | number of companies bankrupt | number of companies bankrupt | number of companies bankrupt |
| 2009 | 29                           | 36                | 16                | 81    |
| 2010 | 230                          | 234               | 223               | 692   |
| 2011 | 553                          | 629               | 2857              | 9039  |
| 2012 | 3287                         | 3781              | 10438             | 17566 |
| 2013 | 32195                        | 61139             | 55459             | 148795|
| 2014 | 37991                        | 75027             | 63877             | 178895|
| 2015 | 41812                        | 81412             | 73387             | 196611|
|      | Total                        | 2796 (2.4%)       | 9042 (4%)         | 16467 (3%) |

TABLE 2. Summary of the features reported in previous bankruptcy prediction studies.

| Category   | Feature                  | Meaning                                | Reference |
|------------|--------------------------|----------------------------------------|-----------|
| Solvency   | X31                      | Financial expenses to sales            | Nam [22]  |
|            | X39                      | Equity to total asset                  | Kim [23]  |
|            | X50                      | Debt to total assets                   | Ohlson [6]|
|            | X24                      | Financial expenses to liabilities      | Bae and Hong [24] |
|            | X45                      | Total borrowings and bonds payable to total assets | Park and Ahn [25] |
|            | X95                      | Current liabilities to total assets    | Zmijewski [26] |
|            | X42                      | Debt to equity ratio                   | Park and Ahn [25] |
|            | X79                      | Receivables to payables               | Jun [27]  |
|            | X76                      | Current ratio                          | Zmijewski [26] |
| Size       | X92                      | Sales                                  | Lee and Kim [28] |
|            | X93                      | Total assets                           | Altman et al. [5] |
| Growth     | X5                       | Growth rate of shareholder’s equity    | Kim J [23] |
|            | X1                       | Growth rate of total assets            | Lee and Kim [28] |
|            | X6                       | Growth rate of sales                   | Lee and Kim [28] |
|            | X9                       | Growth rate of net income              | Ohlson [6] |
| Profitability | X37                 | Retained earnings to total assets      | Altman [4] |
|            | X12                      | Net income to total assets             | Ohlson [6] |
|            | X21                      | Net income to sales                    | Nam [22]  |
|            | X33                      | Net income before income tax expense to financial expenses | Altman et al. [5] |
|            | X11                      | Net income before income tax expense to total assets | Altman [4] |
| Liquidity  | X84                      | Net working capital to total assets    | Altman [4] |
| Activity   | X60                      | Total assets turnover ratio             | Altman [4] |
|            | X72                      | Payables turnover ratio                 | Kang J [29] |
|            | X62                      | Capital stock turnover ratio            | Kim J [23] |
|            | X65                      | Non-current assets turnover ratio       | Jeong [30] |

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.

A. NOTATION

**Definition 1:** Let \( z_1, \ldots, z_n \in \mathbb{R}^d \) be the inputs and \( y_1, \ldots, y_n \in \{0, 1\} \) be the corresponding targets, and define \( X := \{z_1, \ldots, z_n\} \) and \( D := \{(z_1, y_1), \ldots, (z_n, y_n)\} \). That is, we are considering a binary classification problem.

**Definition 2:** Let \( f \) be a trained black-box model for the dataset \( D \) and \( g \) be a simple and explainable model.

**Definition 3:** If an input \( z_i \) is given, we set the **proximity metric** \( \pi(z_i) \) to be a bounded metric between \( z_i \) and \( z_k \). One such candidate of bounded metric would be the Gaussian radial basis function \( e^{-\|z_i - z_k\|^2/\sigma^2} \) in which \( \sigma \) is a hyperparameter.

B. LOCAL APPROXIMATION

First, \( X \) is discretized into bins using a method such as quantile discretization. Let \( x_1 \in X \) be an instance we are considering, and its discretization be denoted by \( x_1' \). Then, with respect to the bin weight, \( x_2', \ldots, x_l+1 \) are sampled and undiscretized to \( x_2, \ldots, x_l+1 \) using a method such as sampling from truncated normal distributions.

Now, we create an \((l+1) \times d\) matrix \( T \) in the following way.

- Fill in the first row to be 1s, representing \( x_1' \).
- For each \( 2 \leq i \leq l+1 \) and \( 1 \leq j \leq d \), if features \( x_i' \) and \( x_1' \) of \( j \) are contained in the same bin, we set \( T_{ij} := 1 \); otherwise, we set \( T_{ij} := 0 \).

This procedure can be regarded as selecting points \( x_2, \ldots, x_l+1 \) near \( x_1 \). After creating the matrix \( T \), we train \( g \) for the data set \( \{(T_1, f(x_1)), \ldots, (T_{l+1}, f(x_{l+1}))\} \) with the sample weight \( \pi(x_i) \). This entire procedure can be regarded as locally approximating \( f \) near \( x_1 \) by \( g \), which is our desire.

C. MEASURING FEATURE IMPORTANCE

Among the various choices for \( g \), we choose to use the mixture of a lasso and ridge regression, which is the method...
applied by Ribeiro et al. [21]. First, to lower the model complexity of \( g \), a feature selection was applied. The number of selected features is called the “length of explanation.” In detail, if we set the length of explanation \( K \), we first use a lasso regression in place of \( g \) and train it to select the top \( K \) important features. Let \( T(\in R^{(l+1)\times K}) \) be the remaining features of \( T \) after eliminating the remaining features. Then, we use a ridge regression \( \tilde{g}(z) := \sum_{l=1}^{K} w_l z_{l} + \lambda \| w \|_2^2 \) to train the dataset \( \{(\tilde{T_1}, f(x_1)), \ldots, (\tilde{T_l}, f(x_1))\} \), where \( w_l \)'s are learnable parameters and \( \lambda \) is a hyperparameter. After training the model \( \tilde{g} \), we regard the higher the value \( |w_l| \) is, the more important we regard the corresponding feature.

### IV. MODEL DESCRIPTION

A tree-based gradient boosting method is a type of ensemble method, which minimizes the loss sequentially by weak learners. In detail, for a given dataset \( D = \{(x_i, y_i)\}_{i=1}^{N} \) \((x_i \in \mathbb{R}^m, y_i \in \mathbb{R})\), a tree ensemble model \( F \) uses \( K \) additive functions to predict the output.

\[
\hat{y}_i = F(x_i) = \sum_{k=1}^{K} f_k(x_i)
\] (1)

![Diagram showing the process of LIME](image)

**Algorithm 1 LIME Pseudocode**

*Require:* Classifier \( f \), Number of samples \( l \)

*Require:* Instance \( x_1 \)

*Require:* Proximity metric \( \pi_{x_1} \), Length of explanation \( K \)

\[
\mathcal{Z} \leftarrow \emptyset
\]

for \( i \in \{2, \ldots, l + 1\} \) do

\[
x'_i \leftarrow \text{sample around}(x_i)
\]

\[
\mathcal{Z} \leftarrow \mathcal{Z} \cup \{\pi_{x_i}(x'_i)\}
\]

end for

\[
w \leftarrow K\text{-}
\]

Lasso\( \mathcal{Z}, K \) \( \triangleright \) with \( x'_i \) as features, \( f(x_i) \) as target

return \( w \)

where \( f_k(x) = w_q(x) \) is a weak learner. Here, \( q : \mathbb{R}^m \rightarrow T \) represents the structure of each tree that maps a sample to the corresponding leaf index, and \( T \) denotes the number of leaves in the tree. Hence, \( f_k(x) \) represents the leaf weight of the corresponding leaf index \( q(x) \). To learn the set of functions \( f_k \), we minimize the following loss function:

\[
\text{Loss}(F) := \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k),
\] (2)

where the first summation is taken over data points and the second summation is taken over \( K \) weak learners.
and $l$ is a differentiable convex loss function and $\Omega(f_k)$ is a regularization term.

The above loss function includes functions as parameters and cannot be optimized directly using traditional optimization techniques. Instead, the model is trained in an additive manner. If we write $\hat{y}_i^{(t)}$ as the prediction of the $i$-th sample at the $t$-th iteration, we will need to add $f_t$ to minimize the following objective.

$$L^{(t)} = \sum_i L(y_i, \hat{y}_i^{(t)} + f_t(x_i)) + \Omega(f_t), \quad (3)$$

where the summation is taken over data points.

Xgboost (XGB) and LightGBM (LGBM) are examples of tree-based gradient boosting models, although they are slightly different in the way they grow trees for weak learners.

As described in Figure 2, whereas XGB chooses the level-wise tree growth algorithm to learn weak learners, LGBM chooses the leaf-wise tree growth algorithm.

![Figure 2. (a) Level-wise growth strategy and (b) leaf-wise growth strategy.](image)

The level-wise tree growth method searches the best possible node to split, and we split it one level down. This will result in symmetric trees and trees will be grown horizontally.

The leaf-wise tree growth method searches the leaves, which will reduce the loss the most, and split this leaf without bothering the rest of the leaves at the same level. Following this method, the tree will be grown vertically.

The leaf-wise tree growth method tends to achieve a lower loss as compared to the level-wise tree growth method. However, it tends to be more likely to overfit than the level-wise tree growth method.

## V. METHODOLOGY

### A. PREPROCESSING

Raw financial data is usually incomplete and the data distribution is complex. As is generally well-known through experiments [31], [32], data must be standardized to make our model stabler and more accurate. Because our data shared the same problem, we needed a suitable preprocessing process, and we used the following preprocessing methods.

As indicated in other data analysis studies, raw financial data are incomplete with missing values and complex data distributions [3]. Our data also had some missing values. Among the various methods used for filling in missing values, we applied a Pearson’s correlation between features and medians. To simplify the data distribution, a Box-Cox transformation was used.

1) MISSING VALUES

Out of 110 features, 59 features had missing values. We used the following methods to fill in these features.

When the Pearson’s correlation $\rho_{XY}$ between two random vectors $X$ and $Y$ is $\pm 1$, almost surely $Y = aX + b$. Using this fact, for when a feature $f_1$ had a missing value, there was a feature $f_2$ without a missing value and the Pearson’s correlation between these two features was $\approx \pm 1$, we created a linear regression model to learn coefficients $a$, $b$ satisfying $f_1 = af_2 + b$, and then filled in the missing values of $f_1$ using this learned model and $f_2$. Specifically, we used this filling method for the case when $|\rho_{XY}| \geq 0.9$. Eight features with missing values were filled using this method. The rest of the features with missing values were filled in using their medians for simplicity.

2) STANDARDIZATION

Although there are various ways to standardize the data, we exploited the Box-Cox transformations [33] for the method of normalization because as shown in the previous work by Son et al. [3], this method greatly reduces the skew-ness of the data and thus enables the machine learning models to perform well. Because a Box-Cox transformation requires inputs to be positive, and some features of our data have negative values, we shifted each feature by its minimum value such that every value becomes positive, and we then applied a Box-Cox transformation.

### B. MODELS

Because the purpose of this study is to emphasize the scalability and consistency of an instance-wise feature importance measurement method LIME, we choose black-box models that are widely used for measuring the model-wise feature importance in the machine learning community and compare these model-predicted feature importances with our results achieved using LIME. Specifically, we used XGBoost and LightGBM because they are likely the most commonly used models one uses for measuring the model-wise feature importance in the community.
TABLE 3. ROC-AUC scores of XGB and LightGBM.

| Model | Dataset                     | AUC(Fold1) | AUC(Fold2) | AUC(Fold3) | AUC(Fold4) | AUC(Fold5) |
|-------|-----------------------------|------------|------------|------------|------------|------------|
| XGB   | Medium or Large Corporation | 0.926      | 0.941      | 0.923      | 0.932      | 0.918      |
|       | Small Corporation           | 0.879      | 0.888      | 0.875      | 0.874      | 0.876      |
|       | Private Enterprise          | 0.888      | 0.897      | 0.877      | 0.872      | 0.877      |
| LightGBM | Medium or Large Corporation | 0.928      | 0.981      | 0.926      | 0.931      | 0.919      |
|       | Small Corporation           | 0.882      | 0.910      | 0.881      | 0.881      | 0.882      |
|       | Private Enterprise          | 0.892      | 0.934      | 0.883      | 0.877      | 0.883      |

importance and achieve a state of the art performance, particularly for classification problems [16], [17].

For tuning the hyperparameters, we used a Bayesian optimization method [34] for XGBoost and a grid-search cross-validation [35] for LightGBM. Both methods have their own advantages and disadvantages; however this is not the focus of our paper, and thus we do not go into details of this herein.

C. ACCURACY

Because our data shows that our classification problem is imbalanced (only 3% are bankrupt companies overall), instead of a typical 0-1 loss, we drew the receiver operating characteristic (ROC) curve, and measured the area under the curve (AUC) as a metric indicating whether our black-box model is trained correctly or not.

VI. 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 scores were sufficiently high, and thus we concluded that our models were trained well. In our experiment, the performances of the models 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 measuring the feature importance. The fixed fold data distribution is briefly described as follows. For medium or large corporations, among the training set of size 90613, 2266 companies went bankrupt and among the test set of size 23220, 530 companies went bankrupt. For small corporations, among the training set of size 177806, 7207 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. For private enterprises, among the training set of size 44452, 1835 companies went bankrupt. Consequently, it does not matter which training set and test set we choose for measuring LIME-Global-Importance. Indeed, we tested this for the XGB model, and we obtain affirmative results (Figure 3).

In practice, the training set is large relative to the test set, and thus it would take much more time to measure LIME-Global-Importance for the training set than for the test set. When this algorithm is implemented for business purposes, it is recommended to use the test set for measuring LIME-Global-Importance, which is also supported by our experiment results.

2) COMPARISON

Because Model-Global-Importance is calculated during the training, only the training set affects its the value. Hence, although it may seem reasonable to measure LIME-Global-Importance on the training set for comparison with Model-Global-Importance, following the justification we made earlier (VI-A1), we measured LIME-Global-Importance on the test set. If \( I \) is the set of top-10 most important features of LIME-Global-Importance and \( J \) is the set of top-10 most important features of Model-Global-Importance, we define the intersection ratio as follows:

\[
\text{intersection ratio} = \frac{|I \cap J|}{10} \times 100(\%)
\]
FIGURE 3. The graphs of the left are lime-global-importances measured on the training set, and the graphs of the right are LIME-global-importances measured on the test set. For each histogram, the top-10 features occur when sorted in order of high LIME-global-importance. It shows that the top-10 most important features are identical.

FIGURE 4. Global-importance comparison for XGB models. The graphs on the left are model-global-importance measured on the training set, and graphs on the right are LIME-global-importance measured on the test set. In each histogram on the left side, the top-10 features with the highest model-global-importance are listed. In each histogram on the right side, the top-10 features with the highest LIME-global-importance are listed.

In our experiment, the intersection ratio ranged from 30% to 70%, as shown in Figures 4 and 5. Of the 110 features, those selected as the top 10 by two different metrics are consistent with each other, which indicates a significantly high correlation between two metrics. In conclusion, we can state that the method for measuring the global feature importance using LIME is sort of a generalization of customary model-wise feature importance measuring methods.

B. INSTANCE-WISE FEATURE IMPORTANCE

The LIME algorithm can approximate the feature importance of any given models in addition to tree-based models such as XGBoost and LightGBM. Using this property, we propose a method for verifying the consistency of the feature selection in the bankruptcy prediction problem.

Given a trained machine learning model estimating the bankruptcy probability, we analyze the change in feature importance derived by the LIME according to each section.
of the bankruptcy probability given by the machine learning model as follows:

**Step 1:** Apply the LIME algorithm on the trained machine learning model at each data point.

**Step 2:** Collect the feature importance measured by the LIME and the predicted bankruptcy probability using the trained machine learning model at each data point.

**Step 3:** Divide the results in segments according to the predicted bankruptcy probability, and analyze the feature importance of data points belonging to a segment for each segment.

In this paper, we choose the top-20 important features for each data and divide the results into 10 segments according to the predicted bankruptcy probability. In each segment, the ratio of a given feature $f_i$ selected as the important feature is defined by the number of data in a segment having $f_i$ as one of the top-20 important features divided by the number of data in a segment.

In Figure 6, the ratios of the features selected as the important features are plotted when LIME is applied on the trained models XGB and LightGBM for corporation data. In the two graphs, the points that rise sharply indicate important features, and both models achieve similar results in terms of important features such as X78 (cash ratio), X88 (cash and short-term investments of the current asset), X104 (growth rate of enrollment), and type_3 (construction industry). By contrast, in the case of XGB, compared to LightGBM, it seems inconsistent in that it shows a characteristic in which the important features change frequently according to the predicted bankruptcy probability.
Similarly, Figure 7 shows the ratios of the features selected as important for private enterprise data, and indicates that both models XGB and LightGBM have similar results in terms of such features as X6 (growth rate of sale), X69 (raw materials turnover), X86 (current debt obligation to current asset) and X104 (growth rate of enrollment) as important. Moreover, it is also similar in that XGB, compared to LightGBM, has a characteristic in which the important features change frequently according to the predicted bankruptcy probability.

Figure 8 describes the LightGBM results using the bar graph for each segment of the predicted bankruptcy probability in proportion to the importance of the features for corporation and private enterprise data. For corporation data, it can be seen that features such as X55 (additional paid-in capital and retained earnings to common stocks), X78, X88, and X98 (income before income taxes per capita), X102 (days after establishment), X104, type_3, and type_4 (wholesale and retail industry) are consistently important features across the entire segments. By contrast, it can also be seen that the importance of the features X3 (growth rate of current assets) and X5 (growth rate of shareholder equity) increase in the segments $P(0.8 < x \leq 0.9)$ or $P(0.9 < x)$.

Consequently, we can conclude that, even if a black-box model is given, the algorithm LIME can be used to interpret how the model estimates the feature importance. In the case of problems related to bankruptcy prediction, it has been found that the LightGBM model is more suitable than the XGB model for consistently calculating the feature importance for predicted bankruptcy probabilities.

We try to analyze why the feature importance along the predicted probabilities appears differently depending on the models. Assume that data with two features and their classification targets are given, and a smooth model $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ predicting probability is trained on these data. At a fixed point $(a, b)$, applying LIME is similar to finding the tangent plane of the surface $z = f(x, y)$ and measuring its coefficients. Because the tangent plane is given by $z = \frac{\partial f}{\partial x}(a, b) \cdot x + \frac{\partial f}{\partial y}(a, b) \cdot y + \text{constant}$, $\frac{\partial f}{\partial x}$ corresponds to the feature importance of the feature $x$. Hence, for a given predicted probability $p$, the average feature importance of $x$ will be given by the following:

$$\text{Feature importance of } x = \frac{1}{l} \int_{p=f(x,y)} \frac{\partial f}{\partial x} ds,$$
where \( l \) is the length of the curve \( p = f(x, y) \) and the integral is the line integral of the scalar field \( \frac{\partial f}{\partial x} \).

Consequently, we can state that the feature importance of \( x \) of the given predicted probability depend on the shape of the level curve and partial derivatives of the model. Hence, when desiring the robustness of the feature importance along the predicted probabilities, the best scenario is the case when the level curves all coincide together, which will be the case when the model is steep at the decision boundary. Because LightGBM is equipped using a leaf-wise tree growth method, it searches the leaves that will reduce the loss the most, and split that leaf without bothering the leaves at the same level. This may result in a narrower decision boundary than models with a level-wise tree growth method such as XGB. We compared the two models using various hyperparameters to solve the problem of fitting the function \( 1_{\|x\|_1 < 1} \) which has the value 1 on the region \( \|x\|_1 < 1 \) and 0 otherwise, and we checked that this is indeed the case. One of these experiments is given in Figure 9, and we can see that the level curves of LGBM overlap better than the level curves of XGB.

**VII. 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-wise manner 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 by using LIME on models that performs better.

Moreover, by comparing the results obtained by applying LIME on XGB and LightGBM based on the predicted bankruptcy probabilities of the model, we showed that LightGBM is more suitable than XGB for consistently estimating the feature importance for the predicted bankruptcy probabilities. 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.

Even though we did not seriously get into the regression model, it is a fundamental component of the proposed model. Instead of a linear regression model, we can employ a linear neural network to take advantage of the expressive power of a neural network. However, this may cause slow training and high computational cost since one needs to train a linear model for each data point. To address this issue, one can consider a recently proposed non-iterative training algorithm. Neural Network with Random Weights (NNRW) is an algorithm for training a neural network in a non-iterative way that results in much faster training. We think NNRW can be combined with our method to build a scalable model for bankruptcy prediction with model-agnostic explanations.

We leave this as a future work. We refer to the readers two review papers regarding NNRW [36], [37]. Moreover, instead of sampling from the entire dataset when constructing linear regression models, we could use Kullback-Leibler random sample partition [38] to improve performance and solve the memory constraints of big data analysis.

When a model is applied to two data points \( x_1 \) and \( x_2 \), there are two cases in which an equity controversy arises. First, there is a case in which \( x_1 \) and \( x_2 \) are not similar but their predicted probabilities \( f(x_1) \) and \( f(x_2) \) are, and second, there is a case in which \( x_1 \) and \( x_2 \) are similar but their predicted probabilities \( f(x_1) \) and \( f(x_2) \) are somewhat different. For the first case, by comparing the values of the important features selected in the corresponding segment, including \( f(x_1) \) and \( f(x_2) \), it would be possible to analyze which factor drives the difference between \( f(x_1) \) and \( f(x_2) \). For the second case, it will be possible to analyze the important features common to the segment containing \( f(x_1) \), the segment containing \( f(x_2) \), and the other features separately. To summarize, it can be stated that a model with high consistency in the selection of important features is highly likely to be applied to areas where bankruptcy prediction is used.

Douzone Bizon ERP service data are managed for the filing of tax returns or checking the internal business status of a company, and not for credit rating purposes. These include data on small corporations or private enterprises that are difficult to apply by credit rating companies that target corporations with significant assets or sales. Moreover, prior researches related to bankruptcy prediction of these type of companies have been also insufficient. The advantage of
using a machine learning methodology is that it is possible to construct a bankruptcy prediction model with high accuracy even for new observation data. As the amount of data increases, there is an increasingly higher demand for applying machine learning methodology to bankruptcy prediction because there is a high possibility that the assumptions in the existing economic analysis model are not satisfied or it will be difficult to establish a new analysis model. By contrast, credit rating regulatory systems such as Equal Credit Opportunity Act, Fair Credit Reporting ACT, or European General Data Protection Regulation require the provision of appropriate information on credit rating standards. In this paper, we empirically showed that a model with a relatively high consistency in the selection of feature importance can be chosen by applying the LIME method to black-box models such as XGB and LightGBM. We expect that our research give some useful insights in selecting a reliable and explainable machine learning models for bankruptcy prediction.

Moreover, we believe that corporate governance indicators in relation to ESG(Environment, Social and Governance), corporate governance indicators have become very important features in the financial industry. In the previous work of Liang et al. [39], the authors assert that the effect of the corporate governance indicators on bankruptcy prediction varies from country to country. Hence, it is very meaningful to conduct related research on Korean companies. For instance, Kim [40], recently, finds the evidence from Korea using a panel dataset for the period of 1991-2001 that largest shareholder ownership (i.e., ownership concentration) is likely to act as a corporate governance mechanism in reducing bankruptcy risk. Since the dataset we have experimented on does not have any corporate governance indicator feature, we decide to leave further analysis on the combined dataset of financial ratios and corporate governance indicators as a future work.

**APPENDIX**

**A. EXPERIMENTS ON RANDOM FOREST MODELS**

We present extra experiments on Random Forest model [41]. The classification results of random forest model on each training dataset using a 5-fold cross validation are given in Table 4. Like XGB and LightGBM models, the performances of random forest model 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 measuring the feature importance.

The intersection ratio between Model-Global-Importance and LIME-Global-Importance ranged from 50% to 60%, as shown in Figure 10. Hence, this also indicates a high correlation between two metrics and the scalability of LIME is additionally supported by this experiment. Even though the prediction scores of Random Forest models were slightly

| Model           | Dataset                  | AUC(Fold1) | AUC(Fold2) | AUC(Fold3) | AUC(Fold4) | AUC(Fold5) |
|-----------------|--------------------------|------------|------------|------------|------------|------------|
| Random Forest   | Medium or Large Corporation | 0.905      | 0.961      | 0.908      | 0.914      | 0.898      |
|                 | Small Corporation        | 0.862      | 0.906      | 0.858      | 0.859      | 0.861      |
|                 | Private Enterprise       | 0.871      | 0.901      | 0.855      | 0.852      | 0.855      |

**FIGURE 10.** Global-importance comparison for random forest models. The graphs on the left are model-global-importance measured on the training set, and graphs on the right are LIME-global-importance measured on the test set. In each histogram on the left side, the top-10 features with the highest model-global-importance are listed. In each histogram on the right side, the top-10 features with the highest LIME-global-importance are listed.
lower than those of XGB and LightGBM, there was no significant difference in terms of LIIME in calculating the global feature importance. However, it has been found that the LightGBM model is more suitable than the Random Forest model for consistently calculating the feature importance for predicted bankruptcy probabilities as shown in Figure 11.

REFERENCES

[1] J. L. Bellovary, D. E. Giaconino, 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. Akimade, 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.

[4] F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” Expert Syst. Appl., vol. 58, pp. 93–101, Oct. 2016.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[6] H. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.

[7] H. Son, C. Hyun, D. Phan, and H. J. Hwang, “Data analytic approach for bankruptcy prediction,” Expert Syst. Appl., vol. 138, pp. 785–794, doi: 10.1145/2939672.2939785.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and Y.-T. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3146–3154.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and Y.-T. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3146–3154.

[11] M. Ancona, E. Ceolini, C. Oztireli, and M. Gross, “A unified view of gradient-based attribution methods for deep neural networks,” in Proc. NIPS Workshop Interpreting (NIPS), Zürich, Switzerland: ETH Zürich, 2017. [Online]. Available: https://scholar.google.co.kr/scholar?start=10&hl=en&as_sdt=0,5&cluster=712942280232184089

[12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2018, pp. 80–89.

[13] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[14] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[15] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.

[16] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and Y.-T. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3146–3154.

[18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2018, pp. 80–89.

[19] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[20] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.

[21] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[22] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2018, pp. 80–89.

[23] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[24] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.

[25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[26] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2018, pp. 80–89.

[27] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[28] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.

[29] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[30] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2018, pp. 80–89.

[31] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, Jul. 2019.

[32] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akimade, 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.
[30] C. Kook, G. Hong, and W. Jeong, “A comparative study on the performance of credit evaluation models,” *J. Money Finance*, vol. 11, pp. 67–104, Jul. 2006.

[31] J. Sola and J. Sevilla, “Importance of input data normalization for the application of neural networks to complex industrial problems,” *IEEE Trans. Nucl. Sci.*, vol. 44, no. 3, pp. 1464–1468, Jun. 1997.

[32] C. Su, J. Zhan, and K. Sakurai, “Importance of data standardization in privacy-preserving K-means clustering,” in *Database Systems for Advanced Applications*, L. Chen, C. Liu, Q. Liu, and K. Deng, Eds. Berlin, Germany: Springer, 2009, pp. 276–286.

[33] G. E. Box and D. R. Cox, “An analysis of transformations,” *J. Roy. Stat. Soc., B (Methodol.)*, vol. 26, no. 2, pp. 211–243, 1964.

[34] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian optimization of machine learning algorithms,” in *Proc. Adv. Neural Inf. Process. Syst.*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 2951–2959.

[35] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. IJCAI*, vol. 14. Montreal, QC, Canada, 1995, pp. 1137–1145.

[36] W. Cao, X. Wang, Z. Ming, and J. Gao, “A review on neural networks with random weights,” *Neurocomputing*, vol. 275, pp. 278–287, Jan. 2018.

[37] X. Wang and W. Cao, “Non-iterative approaches in training feed-forward neural networks and their applications,” *Soft Comput.*, vol. 23, pp. 3473–3476, Apr. 2018.

[38] C. Wei, J. Zhang, T. Valiullin, W. Cao, Q. Wang, and H. Long, “Distributed and parallel ensemble classification for big data based on Kullback–Leibler random sample partition,” in *Proc. Int. Conf. Algorithms Archit. Parallel Process.* New York, NY, USA: Springer, 2020, pp. 448–464.

[39] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, “Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,” *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 561–572, Jul. 2016.

[40] J. Kim, “Determinants of corporate bankruptcy: Evidence from chaebol and non-chaebol firms in Korea,” *Asian Econ. J.*, vol. 34, no. 3, pp. 275–300, 2020.

[41] S. Joshi, R. Ramesh, and S. Tahsildar, “A bankruptcy prediction model using random forest,” in *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2018, pp. 1–6.

**MIN SUE PARK** is currently pursuing the Ph.D. degree with the Department of Mathematics, POSTECH. His research interests include deep learning, data analysis, and partial differential equations.

**HWIJAE SON** received the Ph.D. degree in mathematics from POSTECH, in 2021. He is currently a Postdoctoral Researcher with the Department of Mathematics, KAIST. His research interests include deep learning, data analysis, and partial differential equations.

**CHONGSEOK HYUN** received the Ph.D. degree in financial engineering from Ajou University, in 2011. He is currently working with BNK Financial Group Inc. Prior to this, he was a Research Assistant Professor with Ajou University, from 2010 to 2013, and Korea Housing and Urban Guarantee Corporation, from 2018 to 2020. His research interests include risk management and structured product.

**HYUNG JU HWANG** received the Ph.D. degree in mathematics from Brown University, in 2002. She is currently a Full Professor with the Department of Mathematics, POSTECH. Prior to this, she was a Research Assistant Professor with Duke University, from 2003 to 2005, and a Postdoctoral Researcher with Max-Planck Institute, Leipzig, from 2002 to 2003. She has published more than 65 scientific articles in the fields of applied mathematics and interdisciplinary research. Her research interests include optimization, deep learning, applied mathematics, partial differential equations, and data analysis in applied fields.

* * *