Bank of Japan Working Paper Series

Application of Machine Learning to a Credit Rating Classification Model: Techniques for Improving the Explainability of Machine Learning

Ryuichiro Hashimoto*
ryuuichirou.hashimoto@boj.or.jp

Kakeru Miura*
kakeru.miura@boj.or.jp

Yasunori Yoshizaki*
yasunori.yoshizaki@boj.or.jp

Bank of Japan
2-1-1 Nihonbashi-Hongokucho, Chuo-ku, Tokyo 103-0021, Japan

* Financial System and Bank Examination Department

No.23-E-6
April 2023

Application of Machine Learning to a Credit Rating Classification Model: Techniques for Improving the Explainability of Machine Learning*

Ryuichiro Hashimoto†, Kakeru Miura‡, Yasunori Yoshizaki§

April 2023

Abstract

Machine learning (ML) has been used increasingly in a wide range of operations at financial institutions. In the field of credit risk management, many financial institutions are starting to apply ML to credit scoring models and default models. In this paper we apply ML to a credit rating classification model. First, we estimate classification models based on both ML and ordinal logistic regression using the same dataset to see how model structure affects the prediction accuracy of models. In addition, we measure variable importance and decompose model predictions using so-called eXplainable AI (XAI) techniques that have been widely used in recent years. The results of our analysis are twofold. First, ML captures more accurately than ordinal logit regression the nonlinear relationships between financial indicators and credit ratings, leading to a significant improvement in prediction accuracy. Second, SHAP (Shapley Additive exPlanations) and PDP (Partial Dependence Plot) show that several financial indicators such as total revenue, total assets turnover, and ICR have a significant impact on firms' credit quality. Nonlinear relationships between financial indicators and credit rating are also observed: a decrease in ICR below about 2 lowers firms' credit quality sharply. Our analysis suggests that using XAI while understanding its underlying assumptions improves the low explainability of ML.

JEL classification: C49, C55, G32
Keywords: Credit risk management, Machine learning, Explainability, eXplainable AI (XAI)

* The authors would like to thank the staff of the Bank of Japan for their valuable comments. The views expressed in this paper are those of the authors and do not necessarily reflect the official view of the Bank of Japan or of the Financial System and Bank Examination Department.
† Financial System and Bank Examination Department, Bank of Japan (ryuuichirou.hashimoto@boj.or.jp)
‡ Financial System and Bank Examination Department, Bank of Japan (kakeru.miura@boj.or.jp)
§ Financial System and Bank Examination Department, Bank of Japan (yasunori.yoshizaki@boj.or.jp)

1 Introduction

Machine learning (ML) has been used increasingly in a wide range of operations at financial institutions. The Bank of England (2022) points out that more than 70% of UK financial institutions use ML in various business areas, including customer engagement, anti-money laundering measures, fraud detection, and risk management.[1]

In the field of credit risk management, as discussed by the European Banking Authority (2021) ("EBA (2021)"), many financial institutions are starting to use ML to construct credit scoring models and default models that predict the creditworthiness of individual borrowers or firms based on their financial data and macroeconomic variables.

Traditionally, parametric models such as logit regression have been widely used to build a model for borrowers' default rates and credit ratings. In the field of default rate models, recent years have seen a growing trend in the use of ML, showing that ML can improve the accuracy of predictions as it can capture more complex nonlinearity. For example, using data on retail loans extended by individual Spanish banks, Alonso and Carbó (2021) show that ML achieves more accurate prediction of default rates than logit regression. In addition, Miura et al. (2019) construct a default model that can be applied to non-listed firms, mainly small and medium-sized enterprises, using information on their deposit account activities, arguing that the ML-based model achieves better default prediction than logit regression.

In the field of credit rating classification models, on the other hand, while there have been several studies using parametric methods, such as Kobayashi (2001),[2] few studies have used ML. A credit rating classification model is a framework to predict firms' credit ratings based on their financial indicators and macroeconomic variables. Since ML can capture complex model structures, ML-based models are considered to have higher prediction accuracy than parametric models, especially when a nonlinear relationship exists between financial indicators (explanatory variables) and credit ratings (the dependent variable), as is often seen in practice. In those practical cases, if an ML-based credit rating classification model improves the accuracy

[1] In addition to private financial institutions, central banks have also started to use ML. Araujo et al. (2022) point out that ML has been introduced into a wide range of central bank operations, such as data collection, monetary and economic analysis, monetary policy management, and prudence operations.

[2] Using data for Japan's manufacturing industry, Kobayashi (2001) estimates ordinal and multinomial probit regressions with financial indicators as explanatory variables and corporate bond ratings as the dependent variable.

of prediction, we can expect to refine the credit risk assessment of individual firms and to increase the usefulness of scenario analysis based on the model.[3]

Despite its ability to capture complex nonlinearity, ML often faces criticism for its model complexity and associated low explainability. For example, EBA (2021) points out that because the relationship between model prediction and explanatory variables is hard to understand, ML presents a challenge to ensuring adequate understanding by management and to justifying the results to supervisors.[4]

In response to ML's low explainability, studies on techniques to address this issue – so-called eXplainable AI (XAI) – have been developing rapidly in recent years.[5] Reflecting these considerations, in this paper we first evaluate the prediction accuracy of an ML-based credit rating classification model, and we then examine the relationship between firms' financial indicators and creditworthiness using XAI. Finally, we discuss the caveats of using ML to estimate credit rating classification models.

The main contributions of this paper are twofold. First, we use ML to estimate a credit rating classification model. Compared with research on default models, where a number of empirical studies have reported improvements in prediction accuracy when applying ML, few studies have applied ML to a rating classification model. Thus, this paper constructs two credit rating classification models based on ML and ordinal logit regression using the same dataset, and examines how model structure affects the prediction accuracy of models. The results show that, as in previous literature on default models, an ML-based credit rating classification model achieves higher prediction accuracy than ordinal logit regressions.

Second, we use XAI techniques to examine the relationships between firms' financial indicators and creditworthiness. We use two XAI techniques to understand what determines firms' creditworthiness in the estimated credit rating classification model: (1) SHAP (SHapley Additive exPlanations), which represents the contribution of explanatory variables to model prediction, and (2) PDP (Partial Dependence Plot), which identifies changes in model prediction when changing explanatory variables. These XAI techniques reveal that, as discussed in

[3] The Financial System Report, published in October 2022 by the Bank of Japan, estimates the response of firms' default curves to a deterioration in ICR due to an increase in firms' funding costs using an ML-based credit rating classification model. See Bank of Japan (2022) for details.
[4] Apart from ML's low explainability and complexity, Alonso and Carbó (2022) mention that financial institutions also need to consider data security and privacy issues when implementing ML.
[5] Molnar (2019) and Morishita (2021) provide comprehensive coverage of studies on XAI.

research on default models, total revenue and ICR have high explanatory powers in the model, and ICR has a nonlinear impact on firms' credit ratings.

The remainder of the paper is organized as follows. Section 2 presents an overview of our models and data employed in the analysis. Section 3 compares the prediction accuracy of two classification models based on ML and ordinal logit regression. Section 4 examines the relationship between firms' financial indicators and creditworthiness using SHAP and PDP, and discusses a number of caveats regarding the application of ML in credit rating classification models. Section 5 concludes.

2 Data and models

2.1 Data

Overview

We use credit ratings and financial data for about 6,000 firms, excluding Japanese firms. The observation period ranges from the January-March quarter of 1991 to the April-June quarter of 2022. The data frequency is quarterly.

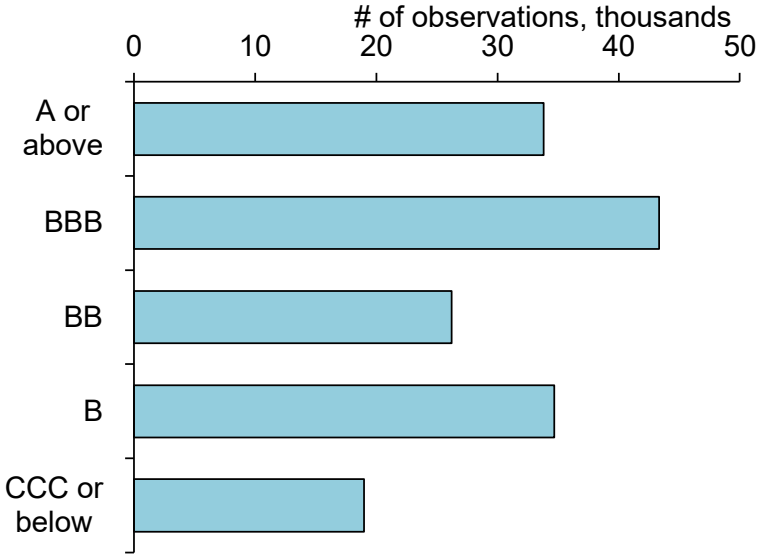
The dependent variable is the long-term issuer ratings obtained from Moody's Investors Service.[6] Figure 1 shows the distribution of the number of firms by credit rating in the dataset. It shows the existence of an imbalance in the number of firms by rating class: while there are a cumulative total of 43,000 observations rated as "BBB," only a cumulative total of 19,000 observations are rated as "CCC or below."[7] If the class imbalance is not addressed, the prediction accuracy for minority class firms may deteriorate because the classification model tends to label firms as majority class. For this reason, following Chawla et al. (2002), we use SMOTE (Synthetic Minority Over-sampling Technique) to construct a model with the same sample size for all classes.[8]

[6] In this paper, for the sake of analysis, Moody's ratings of A3 or above are classified as "A or above," Baa1 to Baa3 as "BBB," Ba1 to Ba3 as "BB," B1 to B3 as "B," and Caa1 or below as "CCC or below."

[7] An observation corresponds to a given firm in a given quarter in our dataset.

[8] In addition to SMOTE, there are other techniques to handle class imbalance, including Oversampling, which resamples instances for the minority class, and Downsampling, which removes instances for the majority class. We obtain similar results to SMOTE when using these techniques.

Figure 1: Number of observations by rating



Note: Shows the total number of firms by rating throughout the observation period

As explanatory variables, we use six types of financial indicators based on individual firms' financial data obtained from S&P Global Market Intelligence: business size, repayment capacity, profitability, financial leverage, liquidity and sector growth (Figure 2).

Figure 2: List of explanatory variables

Financial Indicators	Variables
Business Size	Total Revenue (Log-scale)
Repayment Capacity	ICR (x)
	Net DER (x)
Profitability	EBITDA Margin (%)
	Net Income Growth (%)
	Total Assets Turnover (x)
Financial Leverage	Leverage (%)
Liquidity	Current Ratio (x)
Sector Growth	Sector-level Revenue Growth (%)

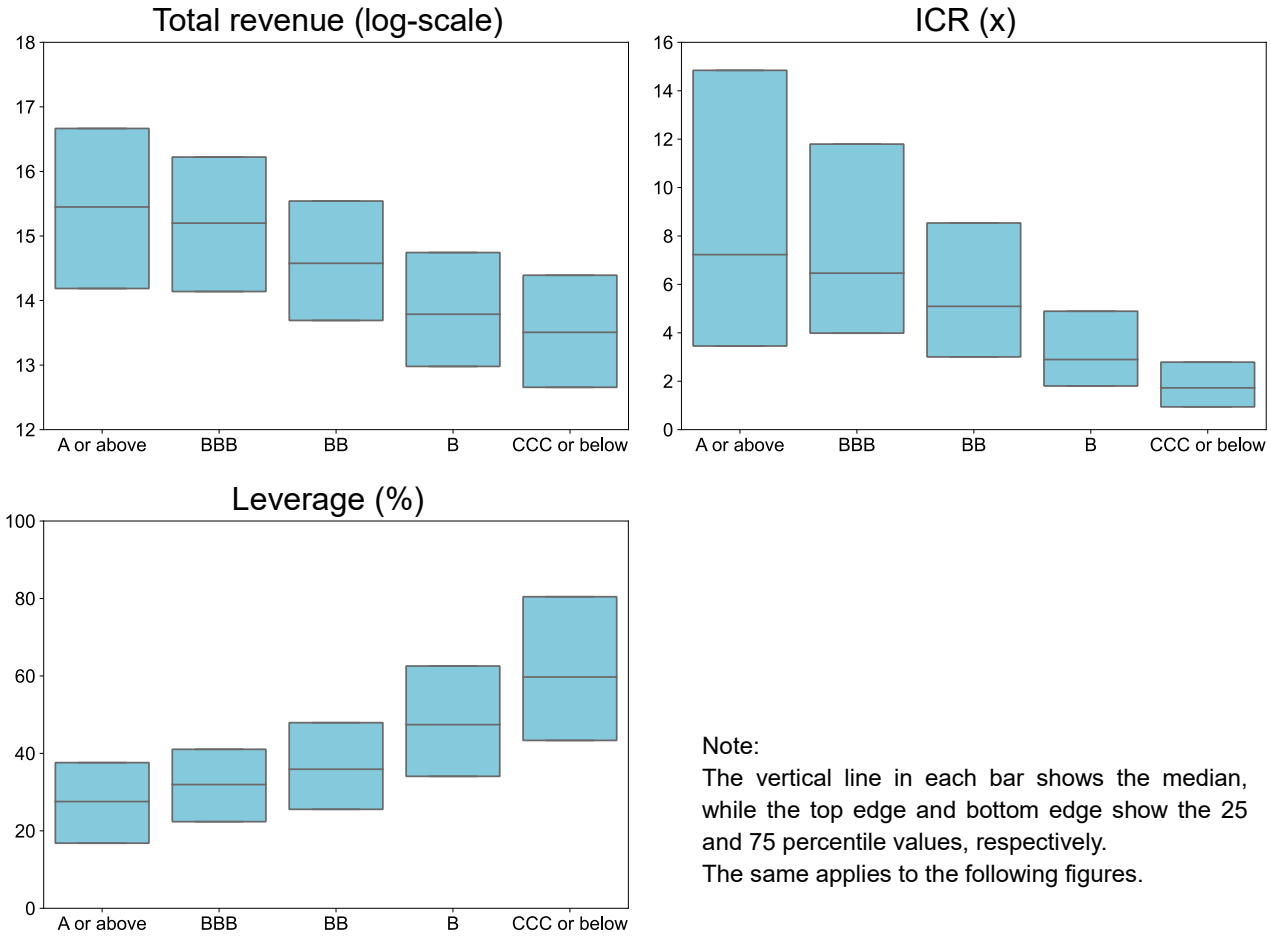
Specifically, the following nine explanatory variables are used; (1) business size: Total Revenue (log-scale); (2) repayment capacity: ICR (EBITDA / Cash Interest) and Net DER ((Interest-bearing Debt – Cash and Cash Equivalents) / Total Equity); (3) profitability: EBITDA margin (EBITDA / Total Revenue), Net Income Growth, and Total Assets Turnover (Total Revenue / Total Assets); (4) financial leverage: Leverage (Interest-bearing Debt / Total Assets);

(5) liquidity: Current Ratio (Current Assets / Current Liabilities); and (6) sector growth: Sector-level Revenue Growth.[9] These indicators are roughly consistent with the credit rating methodologies employed in Moody's Investors Service (2021) and S&P Global Market Intelligence (2020).

Relationship between credit ratings and financial indicators

Figures 3 through 6 show the distributions by rating of the financial indicators used as explanatory variables. These distributions show several patterns in the relationship between ratings and financial indicators.

Figure 3: Financial indicators by rating (1)



Note:
The vertical line in each bar shows the median, while the top edge and bottom edge show the 25 and 75 percentile values, respectively.
The same applies to the following figures.

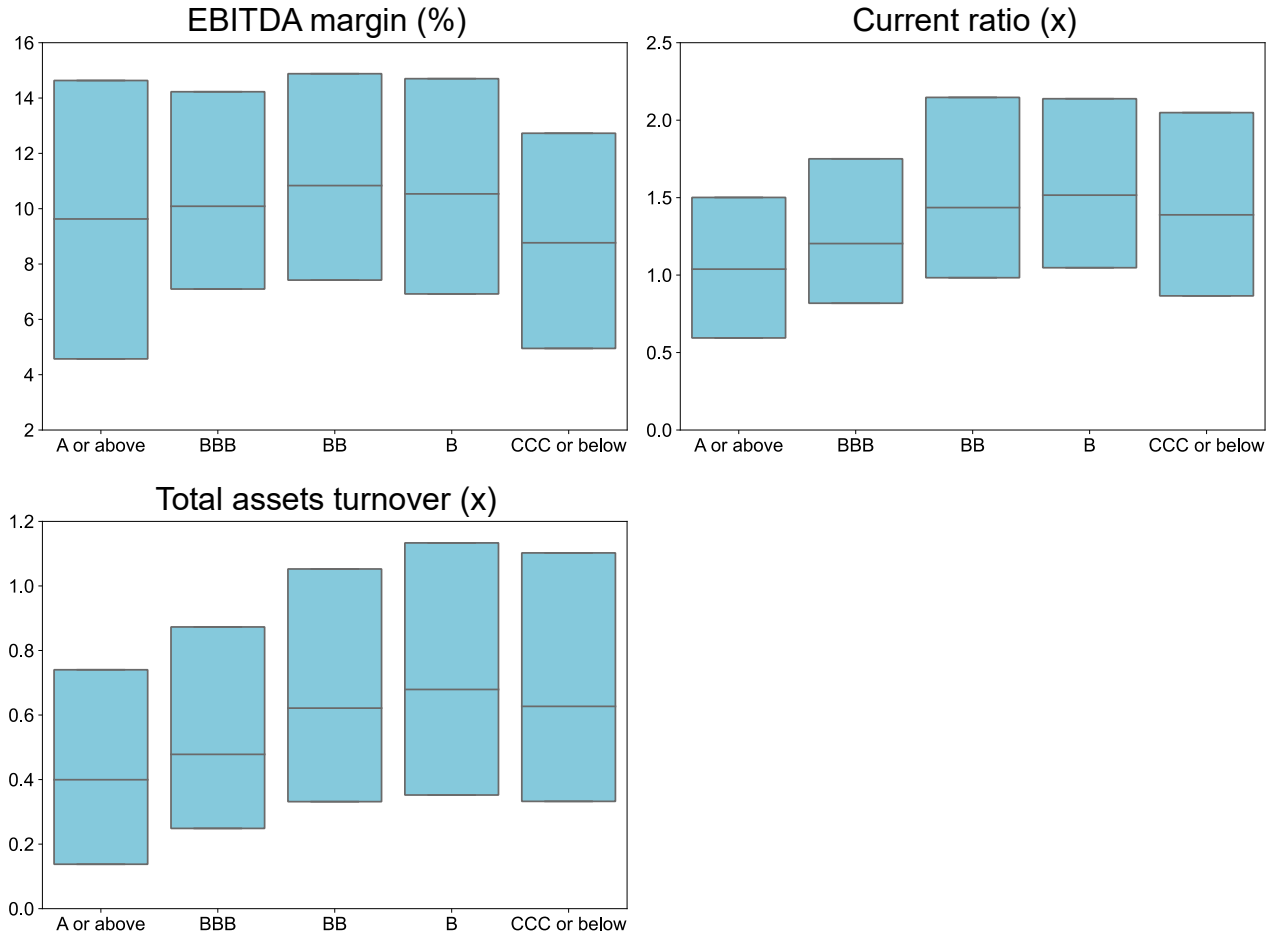
First, we observe the monotonicity between ratings and financial indicators. The median values of total revenue and ICR in Figure 3 decrease monotonically and linearly as the rating worsens. In contrast, the median values of leverage increase monotonically and nonlinearly

[9] Sectors are the 24 industry groups based on the S&P Global Industry Classification Standard. See https://www.spglobal.com/spdji/en/landing/topic/gics/ for details.

with rating downgrades. The change in median values of leverage from "A or above" to "BB" is relatively limited, while that for lower rating classes is more pronounced.

On the other hand, some indicators do not behave monotonically as credit ratings change. First, the median value of financial indicators in Figure 4 shows an upward convex shape. The EBITDA margin first increases in the rating changes from "A or above" to "BB," reaching its peak at "BB", then gradually decreases thereafter. Current ratio and total assets turnover also show an upward convex shape, except that they peak at "B."

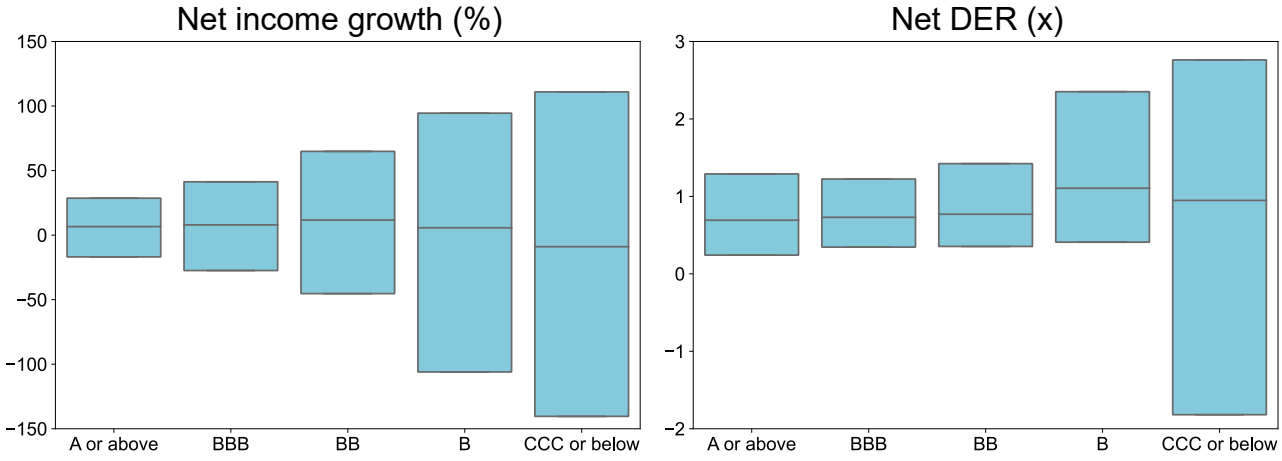
Figure 4: Financial indicators by rating (2)



Second, for several financial indicators such as ICR, leverage, net income growth, and net DER, the variance among firms is correlated to the level of credit ratings. The 25% – 75% interval of ICR expands as the rating gets higher: the interval is narrow around 0.9 - 2.8 at "CCC or below", while it is much wider around 3.4 - 14.6 at "A or above". In contrast to ICR, the 25% – 75% interval of leverage expands as the rating lowers: it remains the same around 20%

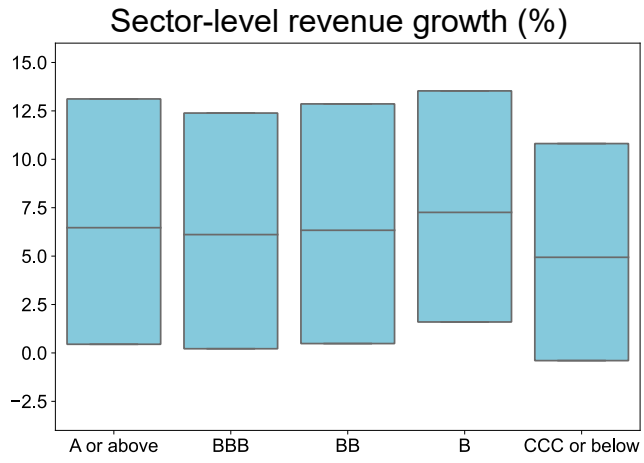
- 40% at “A or above” and “BBB,” while jumping to 44% - 81% at “CCC or below”. The absolute values of net income growth and net DER shown in Figure 5 tend to be larger for lower-rated firms: while their median values stay almost the same across credit ratings, the wider the 25-75% interval expands as the rating lowers, and the more likely it is to be an outlier. One possible reason for this is the smaller business size of lower-rated firms. So, the denominators of the net income growth rate and net DER, i.e., net income and equity capital, tend to be small for low-rated companies, and the absolute values of the net income growth rate and net DER may vary significantly among firms.

Figure 5: Financial indicators by rating (3)



Finally, sector-level revenue growth shown in Figure 6 seems to bear no clear relationship with credit rating. Its median value decreases from “A or above” to “BBB,” but then rebounds to “B,” and turns again to decrease at “CCC or below.”

Figure 6: Financial indicators by rating (4)

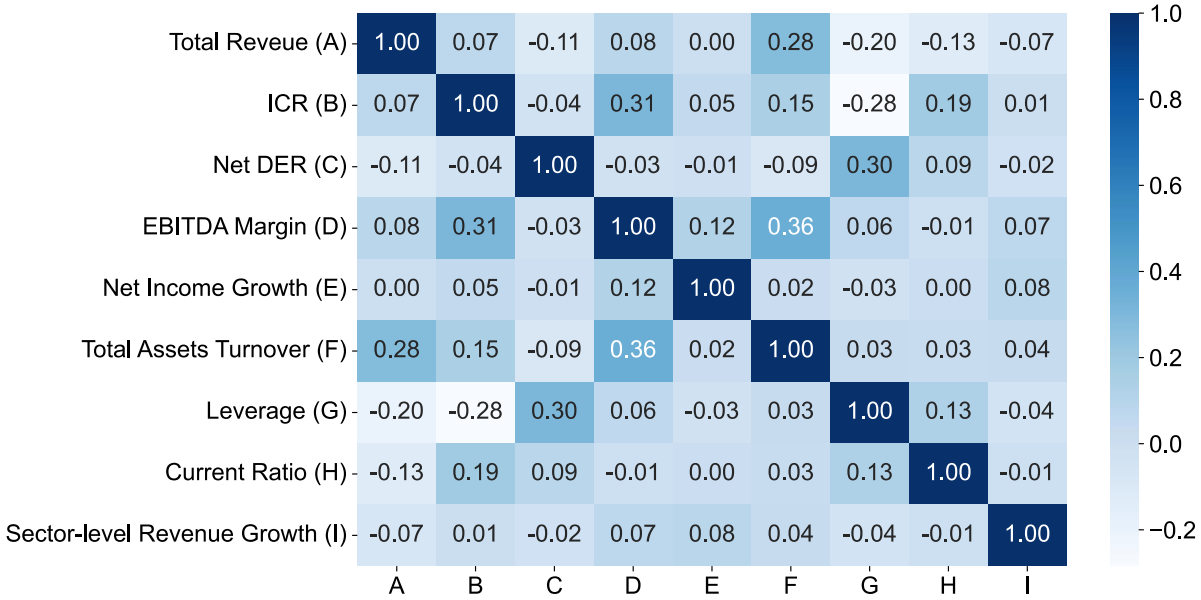


Given the shape of the distributions above, we can safely say that there is not necessarily a monotonous linear relationship between financial indicators and credit ratings. This implies that, in order to estimate a credit rating classification model with high prediction accuracy, it is essential to incorporate the nonlinear relationship between financial indicators and credit ratings and information on variances, as well as the average level of indicators. This also implies that an ML-based model, which can capture complex relationships between the dependent and explanatory variables, is likely to incorporate such nonlinear relationships between financial indicators and credit ratings, whereas an ordinal logit regression leads to relatively low prediction accuracy as it only evaluates the average level of indicators.

Correlation matrix of financial indicators

It is widely known that multicollinearity reduces the stability of estimated parameters in traditional parametric models such as logit regression. In addition, Morishita (2021) points out that ML-based models may face difficulty in evaluating the model when multicollinearity exists. However, the correlation matrix of financial indicators in Figure 7 shows that any pair-wise correlations are relatively low in our dataset, the highest of which is limited to 0.36 between EBITDA margin and total assets turnover. Thus, we consider the impact of multicollinearity on parameter stability and model evaluation to be small in our model.

Figure 7: Correlation matrix of financial indicators



Note: Each cell is colored according to the magnitude of the correlation.

2.2 Models

We estimate two credit rating classification models, based on both ordinal logit regression and ML, which target five rating classes, namely "A or above," "BBB," "BB," "B," "CCC or below." For our ML model, we choose a gradient boosting tree (LightGBM, henceforth LGBM)[10] as it is capable of both high prediction accuracy and fast computation time.[11] For model estimation and evaluation, we divided our dataset into three sub-sets: training, validation, and test sets. We estimate our model using the training and validation sets and evaluate the prediction accuracy based on the test set.

In order to simplify computation and interpretation of SHAP and PDP, which will be presented in Section 4,[12] we also estimate a model that classifies firms into two classes: "Investment Grade" ("IG"), and "Non-investment Grade" ("Non-IG"). Since the number of firms for both classes is almost the same (IG: 77,000, Non-IG: 80,000), we did not employ SMOTE for the 2-class model.

Ordinal logit regression

Yamashita and Miura (2011) argue that ordinal logit regression is often employed when estimating a classification model for ordered discrete data such as credit ratings. Ordinal logit regression assumes that the ordered discrete data Y with J classes is determined by the latent variable y^* and thresholds as follows.[13] Here, $t_1, \dots, t_{J-1}$ are the thresholds for determining classes, based on which observation is labeled as a respective class depending on the level of the latent variable y^*.

$$Y = \begin{cases} 1 & -\infty \leq y^* \leq t_1 \\ 2 & t_1 \leq y^* \leq t_1 \\ \vdots & \\ J & t_{J-1} \leq y^* \leq \infty \end{cases}$$

[10] LightGBM (Light Gradient Boosting Machine) is a gradient boosting tree developed by Microsoft Research in 2016. As the name "Light" suggests, it reduces the estimation time required to tune parameters, achieving high prediction accuracy; see Ke et al. (2017) for more information on LGBM.

[11] Although some minor differences are observed in model prediction accuracy, other typical ML-based models (Random Forest and XGBoost) show roughly similar results to LGBM.

[12] SHAP values and PDPs in multi-class classification problems are calculated as many classes to be classified. For example, in the case of SHAP values, it is necessary to calculate the "SHAP value of an explanatory variable for the probability of being a certain credit rating" for each rating class, which takes more computation time and is difficult to interpret.

[13] The theoretical background of ordinal logit regression can be found in McCullagh (1980) and others.

The latent variable y^* is defined as the weighted linear combination of each component x_i $(1,2,\dots,M)$ of M explanatory variables with the respective weight w_i. Note that the probability of being $Y = j$ relates to the probability that y^* falls between two thresholds that determine the interval of class j.

$$y^* = \sum_{i=1}^{M} w_i x_i = \boldsymbol{w^T x}$$

$$P(Y = j|\boldsymbol{x}) = P\left(t_{j-1} < y^* \leq t_j|\boldsymbol{x}\right)$$

As shown in the equation above, ordinal logit regression is one of the generalized linear models, in which the latent variable is a linear combination of explanatory variables.[14] The key characteristic of ordinal logit regression is that it assumes the weights w_i are independent from classes. This assumption, which is known as the equal slopes assumption, is strong in the sense that the importance of explanatory variables in classification remains the same across all classes. In other words, when the importance of explanatory variables differs significantly across classes, it is inappropriate to use ordinal logit regression, and the model prediction accuracy is likely to decline.

In this paper, we select seven variables out of nine financial indicators and the sector dummy as explanatory variables based on AIC and estimate the ordinal logit regression to classify credit ratings. Note that the ordinal logit becomes the normal binomial logit under the 2-class classification problem (IG/Non-IG). Details of the estimation results are provided in Appendix Figure 1.

LGBM

LGBM is one of the most widely used models in the field of ML. Gradient boosting trees, including LGBM, have a model structure in which many decision trees are sequentially connected. Each successive decision tree gradually reduces the residuals between actual data and model prediction produced using the previous trees.[15]

Each decision tree in the gradient boosting tree splits observations in the training set. At each

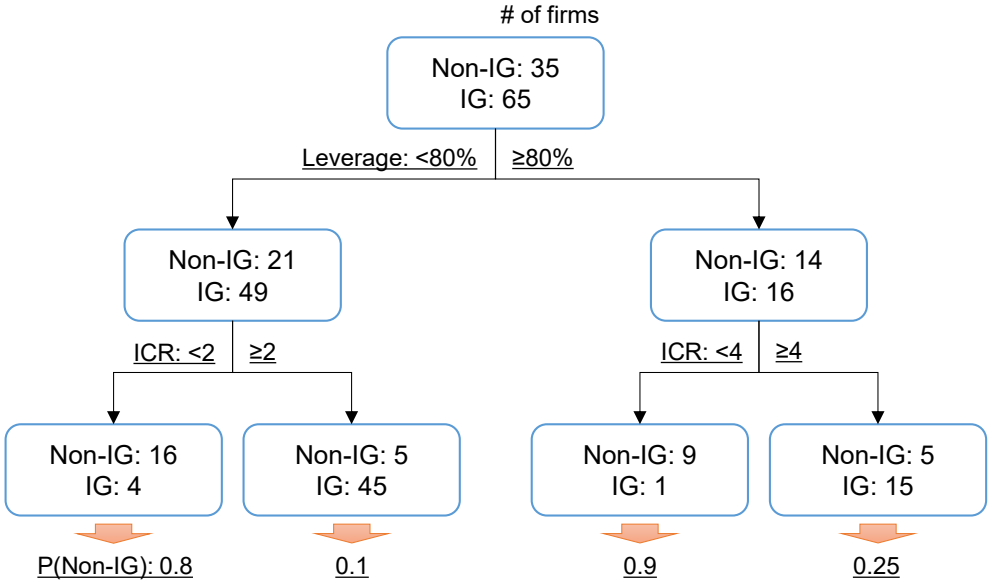
[14] In general, the weights of explanatory variables w_i and thresholds for classes t_j are estimated using the maximum likelihood method as in binomial logit regression.

[15] Tree-based models include models using a method called boosting, in which decision trees are sequentially connected, and those using a method called bagging, in which trees are trained in parallel, as in the case of Random Forests.

split node, a split condition is set using an explanatory variable, such as ICR less than 2. This decision tree structure allows LGBM to capture nonlinear relationships between explanatory and dependent variables.

Figure 8 shows the splits in a decision tree. At each split node, observations that satisfy the condition move on the left side of the node, otherwise on the right, moving to the next node. At the terminal nodes, where all observations arrive after all split nodes, prediction values are assigned, such as 80% probability of being Non-IG. In each decision tree, the choice of explanatory variables for splits, their thresholds, and the prediction values of the terminal nodes are determined to minimize the discrepancy between the actual ratings in the training set and the model prediction, given predetermined parameters such as the maximum depth of the tree.[16] It should be noted that in case the model structure becomes too complex by splitting the observations more than necessary, the model tends to over-fit the training set (so-called overfitting). This results in extremely poor out-of-sample prediction accuracy and takes much more time for the estimation.

Figure 8: Diagram of splits in a decision tree



LGBM introduces several innovations to addresses one of the key challenges in modeling a gradient boosting tree, the slowness of estimation and parameter tuning. These innovations include improving the method of splitting and threshold selection of a decision tree and attempting to minimize errors by focusing on observations with large errors in the previous tree.

[16] The maximum depth of tree is a parameter to determine the number of splits in the training set. A larger number of splits will further reduce the error between the training set and the model prediction, but it may also deteriorate the generalizability, namely worsening the fit to data not used in training, e.g. the test set.

As a result, LGBM often achieves high prediction accuracy in a relatively short estimation time.

Considering these advantages, we use LGBM for our ML-based model. We estimate LGBM with the nine financial indicators listed above and sector (a categorical variable) as explanatory variables, credit ratings as the dependent variable, and log-loss as the loss function.[17]

Interpretation and caveats of estimation results

Generally, in addition to quantitative information such as a firm's financials and macroeconomic indicators, credit rating agencies also use qualitative information in their rating process, such as the firm's business environment and risks for the outlook.[18] However, qualitative information is sometimes not available for every firm in the dataset. Thus, we only use firms' financial indicators to estimate credit rating classification models, as in the previous literature such as Kobayashi (2001). In addition, we group the dependent variable, credit ratings, into five (A or above/BBB/BB/B/CCC or below) or two (IG/Non-IG) classes.

It should be noted that our model cannot perfectly predict credit ratings of rating agencies as we only use financial indicators. This may explain why, despite LGBM's high prediction accuracy, there is still a discrepancy between the actual ratings and the model's predictions, as we will see in Section 3. However, our analysis still provides valuable insights into how accurately the ratings can be estimated using only financial indicators and how the accuracy of ordinal logit regression and LGBM can differ given the same dataset.

3 Prediction accuracy of models

This section compares the prediction accuracy of two models using accuracy score and AUC (Area Under the Curve), which are metrics widely used to evaluate the prediction accuracy of classification models.

[17] In estimation, we employ early-stopping to reduce the risk of overfitting and use a grid search for parameter tuning. Early-stopping is a technique that stops the estimation when no decrease in the loss function is observed in the validation set.

[18] Moody's Investors Service (2021) and S&P Global Market Intelligence (2020) note that in their rating process, agencies consider qualitative information gathered by their analysts, e.g., the firm's business profile (its position in the sector, regulations faced by the sector, the commitment to ESG) and financial/capital policies (whether it has captive financial subsidiaries that enables smooth financing, the commitment of its management to maintaining healthy financials and capital structure), in addition to quantitative information such as the firm's financials, risk scores and inflation of the country where it resides.

3.1 Evaluation using accuracy score

First, we evaluate the accuracy score of both models. Figure 9 shows the confusion matrices, with the predicted ratings on the horizontal axis and the actual ratings on the vertical axis. The accuracy score is the percentage of observations for which the model correctly labels the rating. Here, the accuracy score for the total observations is defined as the sum of the diagonal components of the confusion matrix divided by the total number of observations.

Figure 9: Confusion matrices

(1) 5-class classification

Ordinal logit regression

		Prediction					
		A or above	BBB	BB	B	CCC or below	Accuracy (%)
Actual	A or above	4,049	1,716	710	183	105	59.9
	BBB	2,593	2,953	2,240	743	138	34.1
	BB	626	1,015	1,947	1,308	347	37.1
	B	260	556	1,473	2,450	2,195	35.3
	CCC or below	57	127	376	1,108	2,131	56.1
						Total Accuracy:	43.1

LGBM

		Prediction					
		A or above	BBB	BB	B	CCC or below	Accuracy (%)
Actual	A or above	5,992	569	139	48	15	88.6
	BBB	681	6,965	723	223	75	80.4
	BB	132	528	3,796	628	159	72.4
	B	53	234	645	5,171	831	74.6
	CCC or below	7	49	129	590	3,024	79.6
						Total Accuracy:	79.4

(2) 2-class classification

Ordinal logit regression

		Prediction		Accuracy (%)
		IG	Non-IG	
Actual	IG	12,268	3,162	79.5
	Non-IG	3,832	12,144	76.0
		Total Accuracy:		77.7

LGBM

		Prediction		Accuracy (%)
		IG	Non-IG	
Actual	IG	13,966	1,464	90.5
	Non-IG	1,817	14,159	88.6
		Total Accuracy:		89.6

Note: Based on the test set. Each cell is colored according to the number of firms.

The accuracy score for the total observations in the 5-class classification problem is 43.1% for ordinal logit regression and 79.4% for LGBM. Although both models show deviations from the actual ratings, the accuracy score for LGBM is much higher. The confusion matrix for ordinal logit regression shows that while the accuracy score for both the highest and lowest classes ("A or above" and "CCC or below") is relatively high at around 55 - 60%, that for other classes is relatively low, in the 30% range. This may be because ordinal logit regression is unsuited to capturing complex relationships such as nonlinearity between financial indicators and ratings and the correlation between the variance of financial indicators and ratings. The accuracy score for the 2-class classification problem is 77.7% for ordinal logit and 89.6% for LGBM, with LGBM showing the higher prediction accuracy, as in the 5-class classification problem.

3.2 Evaluation using AUC

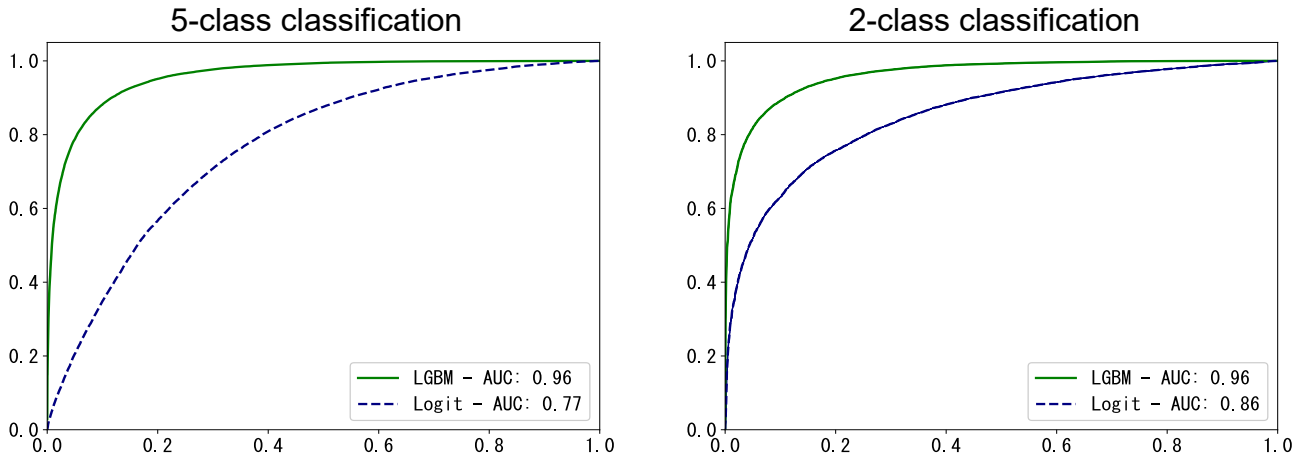
Next, following Alonso and Carbó (2021), we compare the prediction accuracy of two models using AUC, which is one of the typical metrics used to measure the prediction performance of classification models. AUC represents the area under the ROC (Receiver Operating Characteristic) curve,[19] which plots the TPR (True Positive Rate)[20] on the vertical axis and the FPR (False Positive Rate)[21] on the horizontal axis, with a larger AUC indicating a higher prediction accuracy of the classification model.[22]

Figure 10 shows the AUCs calculated for both LGBM and ordinal logit regression. The left and right charts show the AUCs for the 5-class and 2-class classification problems. The figure shows that although both models deviate from the actual ratings, the AUC for the LGBM has a larger area under the discrimination curve and higher prediction accuracy for both the 5-class and the 2-class classification, as in the accuracy score.[23] It should also be noted that the AUC

[19] The ROC curve is the curve plotting how TPR and FPR change when the class criterion (threshold) changes in the classification model.
[20] TPR corresponds to the percentage of correct labeling of Non-IG firms as Non-IG in a 2-class classification problem (IG/Non-IG) where Non-IG is considered as positive. In a 5-class problem, TPR is calculated based on the definition of OvR (One-versus-Rest); defining a certain rating (e.g., "A or higher") as positive and others as negative, TPR refers to the percentage of correct predictions of the positive rating as positive.
[21] FPR corresponds to the percentage of mistaken labeling of IG (negative) firms as Non-IG (positive) in the 2-class problem above. In the 5-class, the TPR is calculated based on the definition of OvR.
[22] AUC should be equal to 1 for a model with perfect prediction, while it is close to 0.5 for a model with random prediction.
[23] The 5-class ROC curve in Figure 10 is a simple arithmetic average of the ROCs calculated based on the OvR definition for each class.

for ordinal logit regression shows a reasonable prediction accuracy of 0.77 for the 5-class and 0.86 for the 2-class classification.[24]

Figure 10: AUC



Note: Based on the test set.

3.3 Background to differences in prediction accuracy

As seen in the previous sections, although both ordinal logit regression and LGBM achieve reasonable prediction accuracy, deviations from the actual ratings still remain. This may be because the models in this paper differ from the methodologies taken by rating agencies in that we do not take into account qualitative information such as firms' business profiles or risks.

In the sections above, we see that LGBM achieves higher prediction accuracy than ordinal logit regression both in the accuracy score and AUC. Possible reasons for higher predictive performance in LGBM are that ordinal logit regression fails to capture the nonlinear relationship between the explanatory variables and the dependent variable, as well as the correlation between the variance of the explanatory variables and the dependent variable, as discussed in Section 2. For example, a 2 decrease in ICR from 10 to 8 and that from 2 to zero would have a different impact on a firm's creditworthiness: the latter would deteriorate its credit quality significantly compared with the former. This nonlinear impact on creditworthiness cannot be captured in ordinal logit regression due to its formulation.[25] In addition, a parametric

[24] Ogi (2017) argues that, as a rough approximation level, a reasonable prediction accuracy corresponds to the AR value of about 0.8 (about 0.9 for AUC) for the default model using financial information for large firms.

[25] It is possible to incorporate such nonlinearity even in ordinal logit regression through variable transformation, such as employing additional power terms, interaction terms, and/or Z-values standardized by the mean and variance of the explanatory variables. However, choosing the proper variable transformation

model that assumes the distribution and the functional form a priori fails to capture the relationship where the average level remains almost the same across ratings but the variance increases as ratings change, as seen in net income growth. In contrast, the structure of a tree-based ML model such as LGBM allows us to capture those nonlinear relationships, leading to higher prediction accuracy.

In Section 4, we will discuss the relationship between firms' financial indicators and creditworthiness using techniques to improve the ML explainability that have been studied in recent years. This includes how the ML-based model actually captures the nonlinear structure between variables, which parametric models hardly capture.

4 Relationship between firms' financial indicators and creditworthiness

4.1 Techniques to improve ML explainability (XAI)

Parametric models assume a priori a functional form and distribution. Thus, one can infer the impact of explanatory variables on model prediction by the estimated regression coefficients. In contrast, an ML-based model has a complex model structure that makes such inferences difficult. This low explainability is often pointed out as a challenge for ML.

In practice, it is vital to understand which explanatory variables contribute to the model prediction. To meet this practical need, a growing number of studies on techniques to improve the explainability of ML-based models (XAI) have been conducted. Figure 11 shows an overview of the main XAI techniques. XAI techniques are categorized depending on the purpose of the analysis, such as whether researchers want to interpret the model as a whole (global) or understand the determinants of the model prediction on an observation basis (local). In most cases, researchers can calculate XAI independently of the model structure (model-agnostic). This characteristic makes XAI applicable to a wider variety of ML-based models,[26] and also allows for comparison of the results derived by XAI techniques across different models.[27]

methods requires expertise in this area. In addition, using these transformed terms would undermine the model explainability, which is one of the benefits of parametric approaches.

[26] Kaneda et al. (2022) visualize the fluctuations of crude oil prices using SHAP and argue that, in addition to supply, demand, and market factors, monetary policy factors (measured by the balance sheet size of the Federal Reserve) have contributed to recent price fluctuations.

[27] These techniques are based on several assumptions that require caution in their interpretation. For example, PDP calculates the impact on model prediction without taking into account dependencies among explanatory

Figure 11: Overview of the main XAI techniques

	Model-level (global)	Observation-level (local)
Overview/ usage	✓ Which explanatory variables are important (learned) by the model? ✓ Which variables have what impact on the predicted values?	✓ How much does each explanatory variable contribute to the prediction result of the observation?
Metrics[28]	✓ Variable Importance (PFI, Permutation Feature Importance) ✓ Partial Dependence Plot (PDP)	✓ LIME (Local Interpretable Model-agnostic Explanations)[29] ✓ ICE (Individual Conditional Expectation)[30]
	SHAP (both model-level and observation-level are applicable)	

In this section, we discuss the relationship between firms' financial indicators and creditworthiness in the estimated LGBM using SHAP, which evaluates at the instance level how much each explanatory variable contributes to the prediction, and PDP, which visualizes the change in the prediction when one or two explanatory variables change while others are held constant (the metrics underlined in Figure 11). Note that all of the following discussion is based on the LGBM for the 2-class classification problem. Thus, the SHAP values and PDPs presented in this section represent the probability of being Non-IG predicted by the model and the contribution to its prediction.

4.2 SHAP

SHAP (SHapley Additive exPlanations) is a technique to evaluate the contribution of each explanatory variable to the predicted value for each observation, and the contribution of each variable is called the SHAP value.[31] Specifically, SHAP additively decomposes the model

variables. However, if such dependencies exist, it may give certain weights to combinations of explanatory variables that are unlikely to occur.

[28] While there are other variable importance metrics that are specific to tree-based models, such as Gains (the amount of reduction in prediction errors when splitting) or Splits (the total number used in splitting), this table only shows model-agnostic metrics. For more information on variable importance such as Gains, see Nembrini et al. (2018).

[29] LIME is a metric that constructs a linear regression model for the data space around the instance to be examined and measures how much each variable contributes to the prediction. See Ribeiro et al. (2016a, 2016b) for details. Because LIME has a limitation in that its results easily fluctuate depending on how "around the instance" is defined, SHAP tends to be more widely used.

[30] ICE is a metric that shows how changes in a certain explanatory variable change the prediction of an observation in a single line. See Goldstein et al. (2015) for details of ICE. Note that PDP is the average of ICEs over all observations.

[31] SHAP is proposed by Lundberg and Lee (2017) and Lundberg et al. (2018) based on Shapley values in Game Theory. Shapley values in Cooperative Game Theory measure the marginal contribution by each player to the total payoffs achieved by the cooperation of all players so that the payoffs can be distributed among players in a fair manner.

prediction of an observation into the average of the predicted values of all observations and the SHAP value of each explanatory variable, as shown in the following formula.

$$\underbrace{f(x)}_{\text{Predicted values of an observation } x} = \underbrace{E_X\left[\hat{f}(X)\right]}_{\text{Average of predicted values of all observations}} + \underbrace{\sum_{j=1}^{M} \phi_j}_{\text{Sum of SHAP values } \phi_j \text{ of } M \text{ explanatory variables for an observation } x}$$

Details of the method used to calculate SHAP values can be found in Lundberg and Lee (2017), among others. Here, we illustrate the specifics of how SHAP values are calculated with a simple example model that predicts the probability of being Non-IG using two explanatory variables, ICR and leverage. First, suppose that the model predicts the probability that a firm with ICR 2 and leverage 80% is Non-IG is 0.7. Also assume that the average of the predictions for the entire sample is 0.5. In this case, the difference between them (0.2) is considered to be the sum of the contributions of ICR and leverage (SHAP values). Then the SHAP value of ICR is calculated as the change from the difference between "predicted value calculated without information on ICR" and the "predicted value calculated with information on ICR". This method follows the method used to calculate the Shapley value for the marginal contribution of a player.[32] The SHAP value of leverage can be calculated in the same manner. Next, we will discuss how SHAP can be used with specific examples.

Observation-level evaluation

In risk management practice, it is important to evaluate the credit rating classification model itself, so as to understand which explanatory variables are the most important determining factors in the rating, and it is also important to check the contribution of explanatory variables at the individual firm level. For example, when a firm is estimated to have a 90% probability of being Non-IG, it is sometimes important to know which financial indicators led to this result.

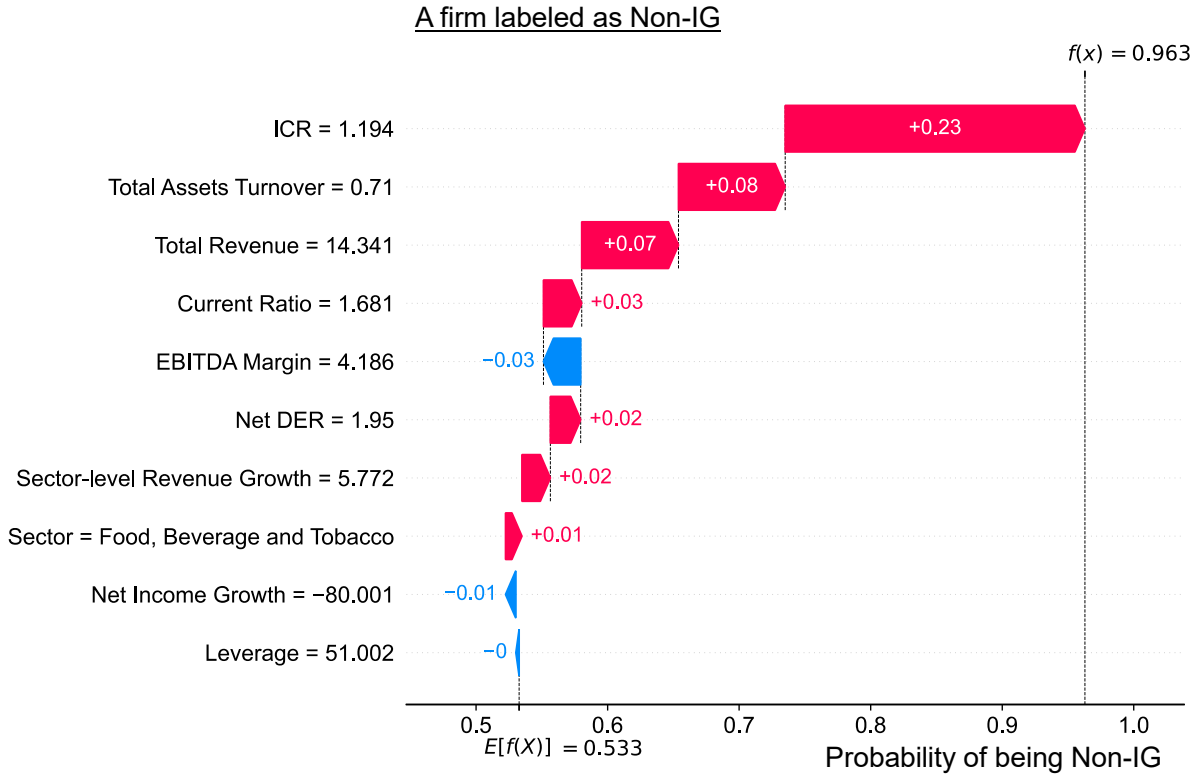
SHAP can address this issue. Figures 12 and 13 show the contribution of the explanatory

[32] Since the marginal contribution of a variable varies depending on the order in which the explanatory variables are added, a precise calculation of the Shapley values requires averaging the contributions for all possible combinations, which leads to an exponential increase in computational costs with the number of explanatory variables. Lundberg et al. (2018) develop the TreeSHAP algorithm, based on the idea of conditional expectation, which reduces the computational costs of SHAP values significantly in tree-based ML models. The algorithm allows the researcher to compute SHAP values more efficiently while maintaining the characteristics of Shapley values, which express the predicted value of an observation as the sum of contributions of explanatory variables.

variables using SHAP for two firms chosen randomly from the dataset. Figure 12 represents a firm labeled as Non-IG, while Figure 13 shows a firm labeled as IG. In both cases, the horizontal axis shows the predictions of the model ($f(x)$, the probability of being Non-IG), and the vertical axis lists the explanatory variables. The contribution of each explanatory variable to the model predictions (SHAP values, ϕ_j) is shown as the difference between the predicted value for the firm and the average of predicted values of all firms ($E[f(x)] = 0.533$).

The predicted value of the firm shown in Figure 12 is $f(x) = 0.963$, labeling as Non-IG. The contribution of each financial indicator shows that low ICR (1.194) boosts the probability of being Non-IG by 0.23 (23%pt), while low total assets turnover (0.71) and low total revenue (14.341) also contribute to the probability by 0.08 and 0.07 respectively.

Figure 12: Observation-level evaluation by SHAP (1)

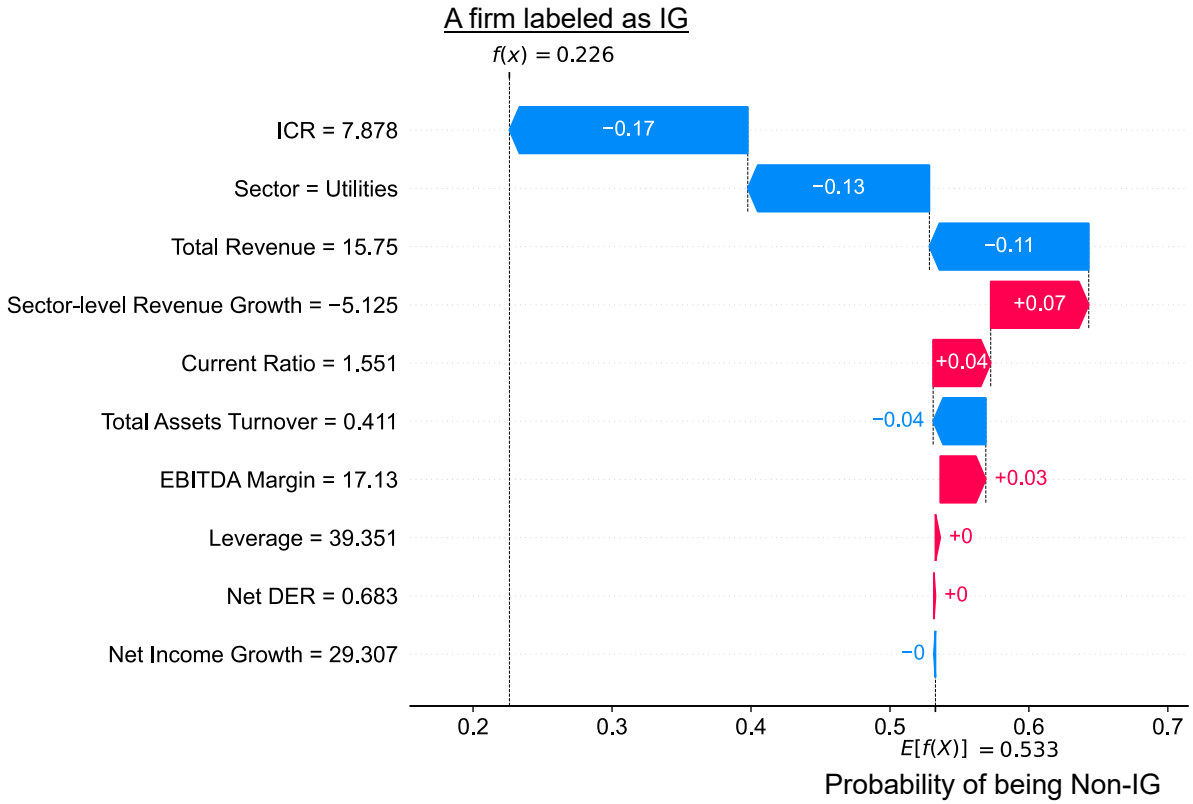


Note:
SHAP values are calculated for a firm randomly chosen from the test set.
$E[f(x)]$ in the figure shows the average of predicted values of all firms (the probability of being Non-IG).
The same applies to the following figure.

On the other hand, for the firm in Figure 13, the predicted value is low at $f(x) = 0.226$, labeling as IG. SHAP values show that high ICR (7.878) and total revenue (15.75) reduce the

probability of being Non-IG by -0.17 and -0.11 respectively, and the sector (Utilities) by -0.13.[33]

Figure 13: Observation-level evaluation by SHAP (2)



Model-level evaluation

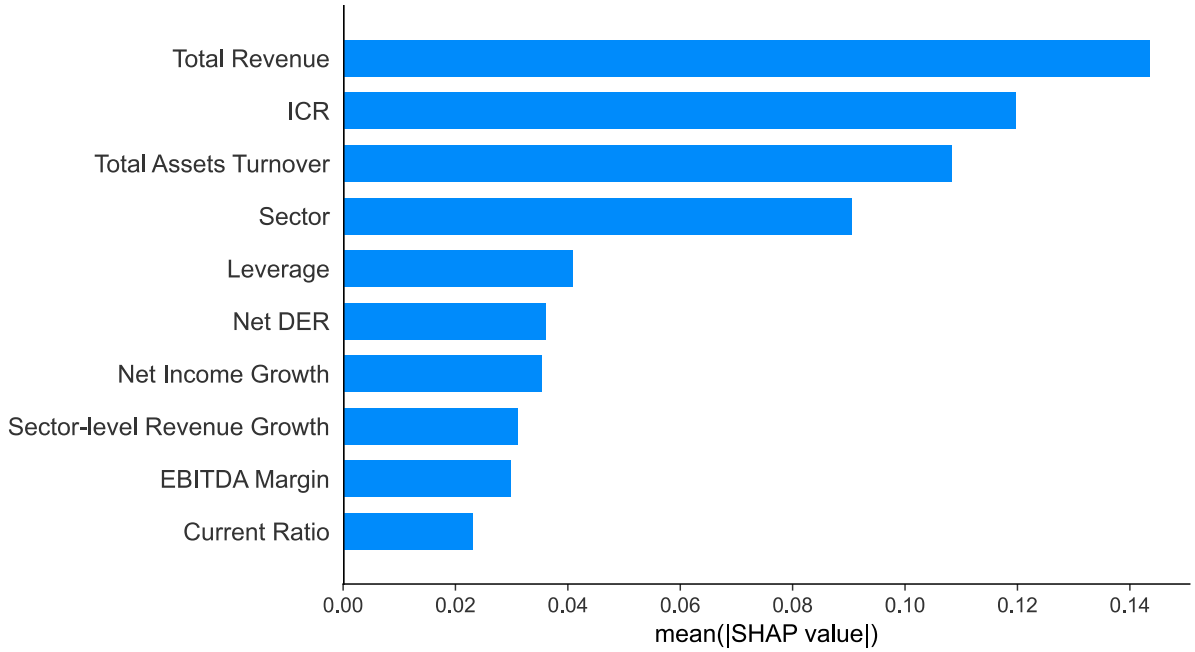
Next, we introduce the SHAP variable importance and SHAP dependence plot, which are popular model-level evaluation techniques using SHAP.

As described above, SHAP values can be used to evaluate the contribution of explanatory variables on an observation basis. They can also be used for model-level evaluation by accumulating the SHAP values of explanatory variables across the entire dataset (SHAP variable importance). SHAP variable importance represents the contribution of each explanatory variable for the entire dataset. This is calculated by taking the average of the absolute SHAP values for all observations. Figure 14 shows the SHAP variable importance for the estimated rating classification model. It shows that several financial indicators, such as total

[33] The SHAP value for a categorical variable such as "Sector" is interpreted as a decrease in the probability that a firm with such high ICR and total revenue is Non-IG, considering the information on the sector to which the firm belongs. In fact, as shown in Appendix Figure 2, the share of high credit rating in Utilities is higher than that of other sectors.

revenue, ICR, total assets turnover, have a significant effect on the prediction of IG or Non-IG.[34]

Figure 14: SHAP variable importance



Note: Calculated using the test set.

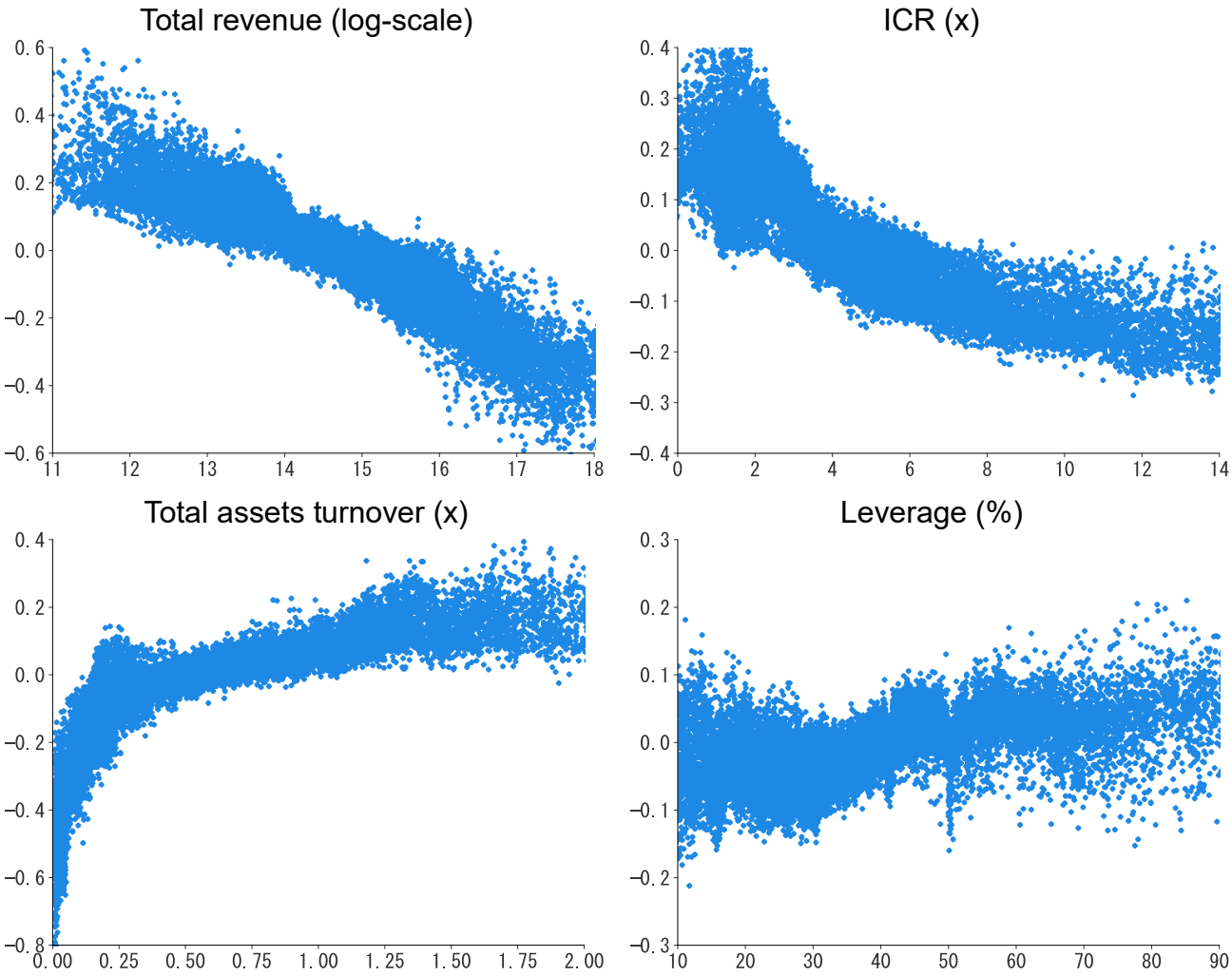
The SHAP dependence plot shows the change in SHAP values and their variance according to the level of explanatory variables. The horizontal axis of the SHAP dependence plot generally represents the values of explanatory variables, while the vertical axis represents the corresponding SHAP values. Figure 15 shows the SHAP dependence plots for explanatory variables with high SHAP variable importance: total revenue, total assets turnover, ICR, and leverage.

Figure 15 shows that total revenue has a negative relationship with the probability of being Non-IG: the probability of being Non-IG decreases as sales increase. ICR also has a negative relationship with the SHAP values, but the SHAP values increase nonlinearly when ICR drops below 2. In addition, the SHAP values gradually increase as leverage increases, and the variance of the SHAP values becomes large at both ends of the PDP. This suggests that the interaction effect with other explanatory variables likely has a more significant impact on the prediction of

[34] In addition to SHAP variable importance, which accumulates SHAP values for each observation, there are other types of variable importance metrics; Permutation Feature Importance evaluates how random replacing of an explanatory variable decreases the prediction error of the model at model-level (paradoxically, which explanatory variables contribute the most to reducing the prediction error). The results using this method are similar to those shown in Figure 14.

IG or Non-IG classification when leverage takes extremely high or low values. These results are consistent with the characteristics of the dataset described Section 2. Note that total SHAP values increase (i.e., the probability of being Non-IG increases) as total assets turnover becomes higher. One possible explanation behind this counterintuitive result is that many of the Non-IG firms in this dataset have small total assets due to their small business size, which is the denominator of the ratio, leading to high total assets turnover.[35]

Figure 15: SHAP dependence plots



Note:
Based on the test set. The horizontal and vertical axes show the level of explanatory variables and the SHAP values (contributions to the probability of being Non-IG).

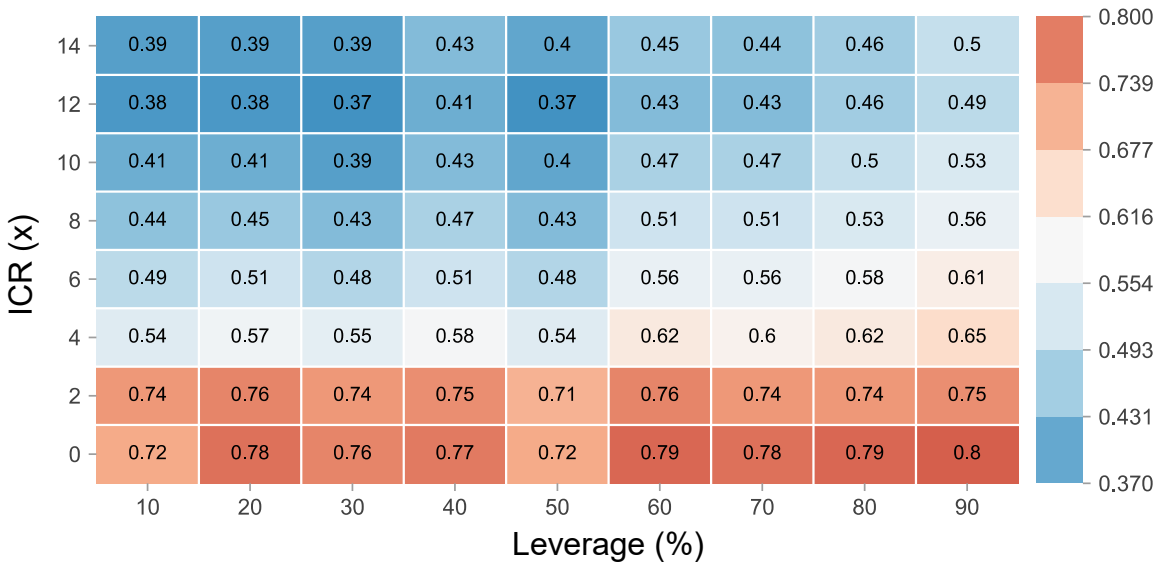
4.3 PDP

Partial dependence plot (PDP) is a typical model-level evaluation technique that visualizes

[35] Figure 4 also confirms this trend that the lower the rating, the higher total assets turnover.

the average relationship between explanatory variables and predicted values, showing the marginal effects of explanatory variables on a predicted value.[36] Specifically, PDP is calculated by changing the level of one or two explanatory variables of interest for the entire dataset, while keeping other variables constant, and then taking the arithmetic mean of the model predictions derived for each level of the variables of interest.

Figure 16: Partial dependence plot (two variables)



Note:
Based on the test set. The horizontal and vertical axes show the level of explanatory variables of interest, while the figure in each cell shows the average of model predictions for each combination of ICR and leverage.

While the SHAP dependence plots in the previous section show the relationship between a single explanatory variable and its contribution to the probability of being Non-IG, PDP can visualize the impact on the probability of being Non-IG when multiple explanatory variables change simultaneously.[37] Figure 16 shows the probability of being Non-IG for each combination of ICR and leverage under the estimated rating classification model using PDP.

The figure shows that, as in Figure 15, the probability of being Non-IG jumps sharply when ICR (vertical axis) drops below 2. As for leverage (horizontal axis), the relationship with the predicted values is weaker than that of ICR, but the probability of being Non-IG is higher for firms with higher leverage.

As seen above, a two-variable PDP visualizes the impact of changes in ICR and leverage on

[36] See Friedman (2001) for the theoretical background on PDPs.
[37] The single variable PDPs are omitted from this analysis as their results are similar to the SHAP dependence plots.

the predicted values. Depending on the combination of explanatory variables, there may be cases where the impact of the interaction of the variables cannot be captured by simply observing the behavior of one variable. A two-variable PDP can be used to understand the impact in such cases, increasing the explainability of ML-based models.

4.4 Caveats for the ML application for credit rating classification

This section discusses several points that warrant attention when using ML for a credit rating classification model, taking into account the results of previous analyses.

First, the interpretation of model prediction using XAI relies heavily on several assumptions. For example, PDP is calculated without considering dependencies among explanatory variables. Using XAI for ML-based models without these assumptions in mind may lead to inappropriate conclusions. Thus, when using XAI, it is vital to consider whether its assumptions are reasonable for the situation to which it is being applied.

Second, the use of XAI to the classification problem of three or more classes requires several considerations. In this section, we focus on the 2-class classification problem, in which the calculation and interpretation of XAI is relatively simple. It should be noted that in a multi-class problem, SHAP and PDP are calculated for the number of classes, making them more difficult to calculate and interpret; in a five-class problem, five SHAP values are calculated for a particular explanatory variable in a given observation. In contrast, a traditional parametric model, such as logit regression, explicitly assumes a functional form and distribution, ensuring a high explainability in that the impact of each explanatory variable on the predicted values can be easily understood from the regression coefficients. Considering these points, it is useful to use both ML and parametric models to complement each other in prediction accuracy and explainability, especially in cases where the interpretation of an ML-based model using XAI seems to be challenging.

5 Conclusion

In this paper, we apply ML to estimate a credit rating classification model and compare its prediction accuracy with ordinal logit regression, which has been widely used in this field. Since ML-based models can incorporate complex nonlinearities between the explanatory and the dependent variables, our ML-based model achieved higher prediction accuracy compared with parametric models. Our results are consistent with those in previous literature where ML was

used to estimate default models.

We also examine the relationship between firms' financial indicators and creditworthiness using XAI techniques such as SHAP and PDP. These XAI techniques reveal the existence of nonlinearities in the relationship between financial indicators and creditworthiness, such as ICR and leverage. Using XAI makes it possible to address ML's low explainability to a certain extent, which has often been regarded as one of the key challenges for ML.

Many studies have been conducted on XAI in recent years and new techniques will continue to be developed. In line with these developments, the use of ML in the operations of financial institutions is becoming more widespread and the importance and usefulness of these techniques will continue to increase.

References

[1] Alonso, R., J. M. Carbó. (2021) "Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation," Banco de España Working Paper Series, No. 2015.

[2] Alonso, R., J. M. Carbó. (2022) "Measuring the Model Risk-adjusted Performance of Machine Learning Algorithms in Credit Default Prediction," *Financial Innovation*, 8(1), 1-35. https://doi.org/10.1186/s40854-022-00366-1

[3] Araujo, D., G. Bruno, J. Marcucci, R. Schmidt, B. Tissot. (2022) "Machine Learning Applications in Central Banking," IFC Bulletin, 57. Bank for International Settlements.

[4] Bank of England. (2022) "Machine Learning in UK Financial Services," Bank of England and FCA Joint Report.

[5] Bank of Japan. (2022) "Financial System Report (October 2022)," October 2022.

[6] Chawla, N. V., K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. (2002) "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321-357. https://doi.org/10.1613/jair.953

[7] European Banking Authority. (2021) "EBA Discussion Paper on Machine Learning for IRB Models," EBA Discussion Paper, No.4.

[8] Friedman, J. H. (2001) "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[9] Goldstein, A., A. Kapelner, J. Bleich, E. Pitkin. (2015) "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics,* 24(1), 44-65. https://doi.org/10.1080/10618600.2014.907095

[10] Kaneda, N., T. Kimata, K. Hiraki, T. Matsue. (2022) “Interpretation of a Machine Learning Model using SHAP: An Analysis of Factors Affecting Crude Oil Price Fluctuations,” Finance Workshop “Application of Machine Learning in Financial Analysis,” Institute for Monetary and Economic Studies, Bank of Japan, November 2022. [In Japanese]

[11] Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu. (2017) "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems, 30.

[12] Kobayashi, M. (2001) “Tests of the Ordinal Probit Model and its Application to Corporate Bond Rating Data,” *Journal of Financial Research (Kinyu Kenkyu)*, 20(1), 1-18. [In Japanese]

[13] Lundberg, S. M., S. Lee. (2017) "A Unified Approach to Interpreting Model Predictions," Proceedings of the 31st International Conference on Neural Information Processing Systems, December 2017, 4768-4777.

[14] Lundberg, S. M., G. G. Erion, S. Lee. (2018) “Consistent Individualized Feature Attribution for Tree Ensembles,” arXiv. https://doi.org/10.48550/arXiv.1802.03888

[15] McCullagh, P. (1980) "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society*, 42(2), 109-142. https://www.jstor.org/stable/2984952

[16] Miura, K., Y. Ijitsu, M. Takekawa. (2019) “Credit Risk Assessment using Information on Deposit Account Activities: Empirical Analysis based on Machine Learning,” Bank of Japan Working Paper Series, No. 19-J-4. [In Japanese]

[17] Molnar, C., (2019) "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," 2nd edition.

[18] Moody's Investors Service. (2021) "Rating Methodology: Manufacturing."

[19] Morishita, K. (2021) “Technologies to Interpret Machine Learning: Practical Techniques to Reconcile Predictive and Explanatory Power,”

[20] Nembrini, S., I. R. König, M. N. Wright. (2018) “The Revival of the Gini Importance?” *Bioimformatics*, 34(21), 3711-3718. https://doi.org/10.1093/bioinformatics/bty373

[21] Ogi, K. (2017) “Fundamentals of Scoring Models: How to Interpret and Use them in SME Financing,” Kinzai Institute for Financial Affairs. [In Japanese]

[22] Ribeiro, M. T., S. Singh, C. Guestrin (2016a) "Model-Agnostic Interpretability of Machine Learning," Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning, July 2016. https://doi.org/10.48550/arXiv.1606.05386

[23] Ribeiro, M. T., S. Singh, C. Guestrin (2016b) "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[24] Shapley, L. S. (1953) "A Value for N-person Games," *Contributions to the Theory of Games*, 2, 307-317.

[25] S&P Global Market Intelligence. (2020) "Machine Learning and Credit Risk Modelling," S&P Global Inc.

[26] Yamashita, S., K. Miura. (2011) "Prediction Accuracy of Credit Risk Models: AR Values and Evaluation Metrics," Asakura Publishing Co., Ltd. [In Japanese]

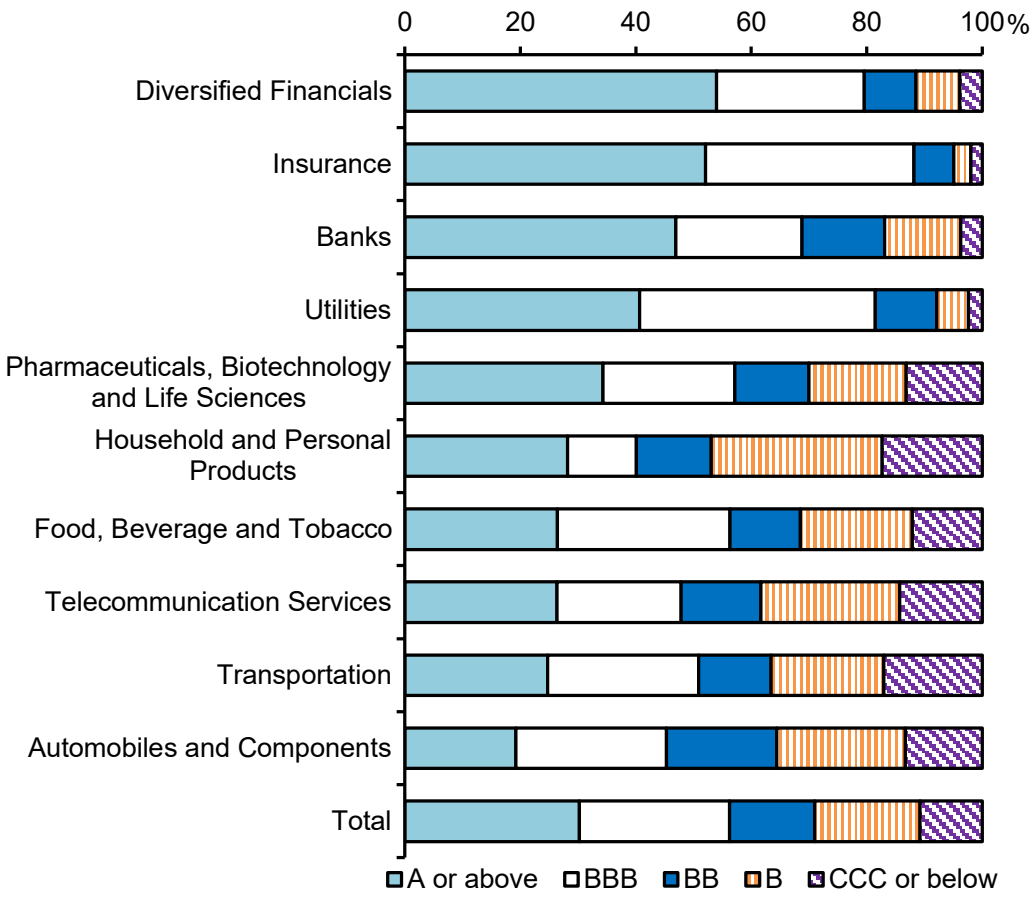
Appendix Figure 1: Regression results of ordinal logit regression

	Coef	Std. Error	Z-value	P-value
Total Revenue (Log-scale)	-0.638	0.004	-150.080	0.000
ICR (x)	-0.008	0.000	-23.925	0.000
EBITDA Margin (%)	-0.061	0.001	-74.840	0.000
Total Assets Turnover (x)	0.858	0.012	69.970	0.000
Leverage (%)	0.035	0.000	117.832	0.000
Current Ratio (x)	0.027	0.004	6.049	0.000
Sector-level Revenue Growth (%)	-0.020	0.000	-43.967	0.000

	Thresholds
A or above - BBB	-11.09
BBB - BB	-9.54
BB - B	-8.14
B - CCC or below	-6.54

Note:
The coefficient table and thresholds between classes for ordinal logit regression.
In estimating the model, net income growth and net DER are excluded based on AIC, while sector fixed effects are included.

Appendix Figure 2: Credit rating composition by sector



Note:
Out of 24 industry groups in the S&P Global Industry Classification Standard, the rating compositions of the top 10 groups with the highest percentage of firms with “A or above” are shown, together with that of all industries.