# 財務指標とマクロ経済インデックスに基づく中小企業のデフォルト予測

# Default Prediction for Small and Medium-sized Companies with Financial Indicators and Macroeconomic Indexes

大石　敬昌 [1 *]　　田中 利幸 [1 †]
Takamasa Oishi　　Toshiyuki Tanaka

京都大学大学院情報学研究科 [1]
Graduate School of Informatics, Kyoto University

**Abstract:** The goal of our research is to build default prediction models on the basis of machine learning models and to obtain useful information for corporate credit risk evaluation. The novelty of this work is twofold. The first point is on how to use time-series information of macroeconomic indexes for the default prediction model for small and medium-sized companies. Since macroeconomic indexes and financial data are different in frequency of being obtained, we considered how to combine these two kinds of data, as input of the default prediction model. In order to combine these data, we summarized time-series information of macroeconomic indexes in the form of mean, percentage change, and volatility. Regarding percentage change, some periods were adopted for the purpose of summarizing both of macrotrends and microtrends. The summarized forms and corporate financial indicators were used as input of the default prediction model in this research. As a result, the default prediction model with inputs of the financial indicators and the macroeconomic indexes outperformed the model with inputs of only financial indicators. Furthermore, the model, to the inputs of which the percentage changes in the fine periods summarizing microtrends were added, outperformed the model not considering the percentage changes in the fine periods. Therefore, considering macroeconomic indexes, especially our proposed method summarizing macrotrends and microtrends, has been found effective for default prediction. The second point is regarding which financial indicators are important in default prediction for small and medium-sized companies by industry sectors. We divided companies into eight industry sectors and investigated which financial indicators are important in each industry sector on the basis of variable importance evaluated with random forest.

## 1 Introduction

Tightness of banks with their money for small and medium-sized private companies is at issue recent years. One of the reasons for the tightness is "International convergence of capital measurement and capital standards", which was established by Basel Committee on Banking Supervision in 1988. Since 1988, the development of financial derivative products and the collapse of the bubble economy have enhanced the necessity for tightening the regulation concerning risk taking of the banks. That is why the Basel Capital Accord has been strengthened from Basel I to Basel II and Basel III, in 2004 and in 2010, respectively. Tightening the regulation for risk taking of the banks has forced the banks to avoid their risks. The necessity that the financial institutions assess credit risk of companies precisely has developed credit risk evaluation models by using machine learning.

Availability of a sufficient amount of training data is a key to success of machine learning models. Since it is mandatory for listed companies to publish their financial reports periodically, a sufficient amount of data are available to make high-performing credit risk evaluation models for listed companies to learn well. On the other hand, unlisted companies, typically small or medium-sized, do not have to publish their financial reports, making it difficult for us to collect an enough amount of data for training machine learning models for credit risk evaluation of small and medium-sized companies. Consequently, the banks had tightness with their money for small and medium-sized companies. In order to facilitate finance for small and medium-sized companies smoothly, it is urgent to collect and analyze financial data of small and medium-sized companies. In Japan, the Credit Risk Database institution (CRD) has been collecting and analyzing financial data of small and medium-sized companies, with the vision of realizing smoothing finance for small and medium-sized companies and advancing management of credit risk. Since the database of CRD contains financial data of over one million small and medium-sized companies, it is ideal for use as training data for machine learning models for credit risk evaluation of small and medium-sized companies. The goals of our research are to build default prediction models for small and medium-sized companies on the basis of machine learning models and to obtain useful information for corporate credit risk evaluation for small and medium-sized companies. To achieve the goals, we use the financial data of small and medium-sized companies provided by CRD.

The novelty of this work is twofold. The first point is on how to use time-series information of macroeconomic indexes in a default prediction model. Not only corporate financial data but also macroeconomic factors should

[*]京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町
E-mail: oishi@sys.i.kyoto-u.ac.jp

[†]Email: tt@i.kyoto-u.ac.jp

be considered for default prediction. For example, difficulty of obtaining loans differs from high-interest period to low-interest period. Existence of strong correlation between corporate credit risks and macroeconomic indexes has been reported [Bonfim 09, Ali 10, Chen 14, 大橋 03, 尾木 13]. Therefore, default prediction model may be more accurate by considering macroeconomic factors than not. However, there is one serious problem in using macroeconomic indexes for default prediction model. The problem is difficulty of combining macroeconomic indexes and financial data because macroeconomic indexes and financial data are different in frequency of being obtained. Financial data of small and medium-sized companies are typically available once a year, which is different from big companies which have to publish their financial reports quarterly. On the other hand, macroeconomic indexes can be obtained from quarterly to daily or even finer, so that most of the macroeconomic indexes can be obtained far more frequently than financial data of small and medium-sized companies. Therefore, we have to consider how to combine these two kinds of data which are different in frequency of being obtained, as input of the default prediction model. This problem has not been discussed in any existing research sufficiently. That is why our research discusses how to use the time-series information of macroeconomic indexes obtained more frequently for default prediction model.

The second point is regarding which indicators are important in default prediction for small and medium-sized companies in each industry sector. There is not much existing research which analyzed and compared properties for small and medium-sized companies by industry sectors, because difficulty of collecting financial data of small and medium-sized companies has made it impossible to analyze properties by industry sectors with sufficient amount of data. However, CRD provided us with such large-scale data, containing financial data of about 1 million companies, that we were able to analyze properties in each industry sector. In this research, we divided companies into eight industry sectors, which will be described in detail in Section 3.2.2, and investigated which financial indicators are important in each industry sector on the basis of variable importance evaluated with random forest.

# 2 Machine Learning for Default Prediction

## 2.1 Default Prediction Models

In the field of default prediction, default prediction models are classified into the following three categories: traditional statistical models, Artificial Intelligence (AI) models, and theoretical models [梅谷 13, 辻 07]. It should be noted that the classification is peculiar to the field of default prediction. We review in this section existing studies which are related with the traditional statistical models and the AI models, both of which are based on statistics and/or machine learning.

Historically, before default prediction models emerged, default prediction was typically performed via calculating financial indicators to measure profitability, safety, and capital efficiency with financial data and comparing the financial indicators with thresholds. Instead of such traditional ratio analysis, Altman proposed in 1968 a default prediction model on the basis of traditional statistical models [Altman 68]. The proposed model was a multivariate discriminant analysis model, called the Z Score. It was furthermore been extended in 1977 to the ZETA$^{\mathrm{TM}}$ Score [Altman 77], which gained popularity and was often used for default prediction. These scores are pioneers of traditional statistical models for default prediction.

Subsequently, logistic model [Flagg 91, 高橋 02, 山下 03, 安道 04, 三浦 08, 森平 09, 山下 11], hazard model [山下 04], and conditional probability model [Goldberg 04] were used for default prediction in existing researches. Strong points of these traditional statistical models are a small amount of calculation and of data needed for learning and high-interpretability because these models are relatively simple. Due to the former point, the traditional statistical models were mainly used when the performance of computers was insufficient and there were no institutions which had a large amount of financial data of companies. However, the traditional statistical models are so simple that their expression capability is poor.

On the other hand, from the middle of 2000s the AI models have been reported to have its higher performance than the traditional statistical models. First, neural network models [Odom 90, Altman 94] were reported as AI models for default prediction. Performance of the neural network models was, however, lower than those of the traditional statistical models because of overfitting. However, in 2000s support vector machines so greatly outperformed the traditional statistical models and the neural network models that the AI models collected a lot of attention [Min 05, Shin 05, Chen 11]. Furthermore, in 2010s, ensemble learning, such as bagging, boosting, and random forest, spread in the field of default prediction, and the AI models, especially ensemble learning and support vector machines, were reported to have the highest performance in many researches [Zhang 10, Kim 10, Wang 11, Barboza 17]. According to the existing researches in the field of default prediction, at present, the best performing models seem ensemble learning and support vector machines. A strong point of the AI models is high expressivity, which leads to higher performance of the AI models than the traditional statistical learning models. Although learning of the AI models needs an enormous amount of calculation and of data, improvement of computing capability and accumulation of financial data about small and medium-sized companies have enabled the AI models to be applied to default prediction. However, the AI models are so complicated that most of them have low interpretability. Highly interpretable models are appropriate to obtain useful information for credit risk evaluation.

## 2.2 Combining Macroeconomics into Default Prediction

In order to use macroeconomic indexes for default prediction models, one should consider which indexes, what forms, and which terms are appropriate for summarizing time-series information of the macroeconomic indexes. In existing researches, discussion of methods summarizing time-series information of macroeconomic indexes is insufficient. For example, means and volatilities [1] of Nikkei

[1] In this paper, the term "volatility" implies historical volatility. The historical volatility $\sigma$ is defined as $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(u_i - \bar{u})^2}$, where $N$

225[2] and Yen to US dollar exchange rate over a year to the companies' account closing months are treated as inputs of the default prediction model [大橋 03, Tinoco 13, Nam 08, Alifiah 14], or year-to-year percentage changes of means and volatilities on their account month are treated as the inputs [森平 09], without any basis. In order to use the macroeconomic indexes for the default prediction model, further discussion of which indexes, what forms, and which terms are appropriate for summarizing time-series information of the macroeconomic indexes, is necessary.

First, regarding which indexes should be used for default prediction of small and medium-sized companies, Ogi and Moridaira [尾木 13] used a probit model for default prediction of small and medium-sized companies, with financial data as well as a single postulated background macroeconomic factor as the regressors. They first estimated values of the background macroeconomic factor on the basis of data collected by Tokyo Shoko Research, LTD., and then performed another regression analysis, where the dependent variable is the estimated background factor and where the independent variables are macroeconomic indexes, in order to see what macroeconomic indexes describe the background macroeconomic factor, and consequently, the default probabilities. They reported that the macroeconomic indexes which have influences on corporate default of small and medium-sized companies are short-term interest rate, long-term interest rate, exchange rate, and stock-market indexes.

Second, we should consider what forms are appropriate for summarizing time-series information of the macroeconomic indexes. The forms we considered as summaries of their time-series information are mean, volatility, and percentage change because of the following reasons. The mean of the interest rates and of exchange rate imply information of high-interest or low-interest and of conditions with a strong yen, a strong dollar, and so on, respectively. The volatility is also important because in terms of stock price or exchange rate, even though the mean is the same, if the volatility differs, the interpretation of the economy will differ significantly. The percentage change connotes trends of markets. Even though the mean and the volatility are the same respectively, whether the market is uptrend or downtrend will affect the interpretation of the economy greatly. Hence, the mean, the volatility, and the percentage change are considered appropriate summaries for macroeconomic indexes.

Third, we have to focus on which terms of the mean, the volatility, and the trend should be summarized. The mean and the volatility within a short time interval upto the financial month should be considered as features affecting corporate defaults, because the farther the mean and the volatility are from the financial month, the less influence they might have on a corporate default. On the other hand, the trend within both of a short and a long time intervals upto the financial month should affect corporate defaults, since there is general knowledge that the trend has the following two types: macrotrends and microtrends, both of which are important. Accordingly, we summarize information of the macrotrends and microtrends, and confirm whether the general knowledge is appropriate.

is sample size, $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$, $u_i = \log \frac{S_i}{S_{i-1}}$, and $S_i$ is the central rate of $i$-th day.

[2]Nikkei 225 © Nikkei Inc.

# 3 Experiments

## 3.1 Purposes

We performed two experiments. In Experiment 1, whose purpose was finding important properties for default prediction, we divided companies into eight industry sectors and investigated which financial indicators are important in each industry sector on the basis of variable importance evaluated with random forest. In Experiment 2, whose purpose was improving the performance of the default prediction, we devised models considering both the macroeconomic indexes and the financial indicators. We then compared the models considering the macroeconomic indexes and the financial indicators and the model with only the financial indicators, in order to see whether or not incorporation of macroeconomic indexes is effective in default prediction.

## 3.2 Model, Industry Sectors, Financial Indicators, and Macroeconomic Indexes Used in The Experiments

### 3.2.1 Random Forest

Random forest (see, e.g., [Hastie 09]) is one of the ensemble learning models and is reported to achieve high performance of default prediction in existing researches. The procedure of random forest is described in **Algorithm 1**. Random forest also has high interpretability. In the process of learning, random forest measures importance of each input variable, called variable importance. By using the variable importances, the present work investigates important financial indicators in each industry sector.

There are three major parameters in random forest: the number $B$ of trees in the forest, the maximum depth $d_{\max}$ of the tree, and the minimum number $n_{\min}$ of samples required to split an internal node. In this study, these three parameters were determined as 10000, 10000, and 2, respectively, on the basis of preliminary experiments.

The area under the receiver operating characteristic (ROC) curve, abbreviated as AUC, was used for performance evaluation. Given an output $\mathrm{P}(\mathbf{x})$ of the random forest algorithm, as well as test datasets consisting of financial data of default and non-default companies, AUC is estimated as

$$\mathrm{AUC} = \frac{1}{n_{\mathrm{D}} n_{\mathrm{ND}}} \sum_{i \in \mathrm{D}, j \in \mathrm{ND}} \mathrm{I}(\mathrm{P}(\mathbf{x}_i) - \mathrm{P}(\mathbf{x}_j) \geqq 0), \qquad (1)$$

where D and ND denote test datasets of default and non-default companies, respectively, where $n_{\mathrm{D}}$ and $n_{\mathrm{ND}}$ denote the numbers of data of default and non-default companies, respectively, and where $\mathrm{I}(c)$ denotes the indicator function of the condition $c$.

**Algorithm 1**: Random Forest

**for** $b = 1$ to $B$:
  Draw a bootstrap sample of size $N$ from the training data
  Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the maximum depth of the tree $d_{\max}$ or the minimum node size $n_{\min}$ is reached:
    (i) Select $m$ variables at random from the $p$ variables
    (ii) Pick the best variables/split-point among the $m$ variables
    (iii) Split the node into two daughter nodes
  $T_b$ can calculate the probability $\mathrm{P}_b(\mathbf{x})$ of classifying $\mathbf{x}$ into the positive class
Output: $\mathrm{P}(\mathbf{x}) = \{\sum_{b=1}^{B} \mathrm{P}_b(\mathbf{x})\}/B$

Table 1: The numbers of all default data and all non-default data by industry sectors in 2010–2016.

| Industry Sectors | Default Companies | Non-Default Companies |
|---|---|---|
| Construction | 2991 | 793457 |
| Manufacturing | 2868 | 723275 |
| Information and Communication | 343 | 59759 |
| Wholesale and Retail | 4229 | 901999 |
| Real Estate and Goods Leasing | 612 | 257299 |
| Accommodation and Food Service | 977 | 124278 |
| Lifestyle-related Entertainment | 319 | 58268 |
| Medical and Welfare | 199 | 91924 |

#### 3.2.2 Industry Sectors

In financial data provided by CRD, each company is labeled with the industry sector to which it belongs. In this work, we chose the following eight industry sectors: construction industry, manufacturing industry, information and communication industry, wholesale and retail industry, real estate and goods leasing industry, accommodation and food service industry, lifestyle-related entertainment industry, and medical and welfare industry.

The numbers of default companies and non-default companies in the dataset are shown in Table 1. Since there were only a few of default companies, we adjusted the number of non-default companies in order to avoid the class imbalance problem. In each industry sector, we used all the default data and 100 times as many non-default data as the default data. The non-default data used in our experiments were chosen randomly.

#### 3.2.3 Financial Indicators

In assessment for credit risk of a company, there are many financial indicators measuring profitability, safety, and capital efficiency. In this work, we chose the following 18 financial indicators: sales to total assets (総資産売上率), return on equity (自己資本当期利益率), ratio of gross profit to sales (売上高総利益率), operating profit on sales (売上高営業利益率), ratio of ordinary profit to sales (売上高経常利益率), current ratio (流動比率), quick assets ratio (当座比率), fixed ratio (固定比率), fixed assets to fixed liability ratio (固定長期適合率), debt ratio for total assets (借入金依存度), SGA ratio (売上高販管比率), inventory turnover period (棚卸資産回転日数), accounts receivable turnover period (売上債権回転日数), trade payable turnover period (買入債務回転日数), tangible fixed assets turnover rate (有形固定資産回転率), ratio of depreciation to sales (売上高減価償却率), and interest coverage ratio, ratio of interest-bearing debt to cash-flow (キャッシュフロー有利子負債比率). These financial indicators were treated as inputs of random forest.

#### 3.2.4 Macroeconomic Indexes

In view of the analysis by Ogi and Moridaira [尾木 13] mentioned in Section 2.2, which states that the macroeconomic indexes which have influences on default probabilities of small and medium-sized companies are short-term interest rate, long-term interest rate, exchange rate, and stock-market indexes, we chose in this research the following macroeconomic indexes: unsecured overnight call rate[1], 1-year Japanese Government Bond[2], 10-year Japanese Government Bond[2], Yen to US dollar exchange rate[1], Yen to Euro exchange rate[1], and Nikkei 225[3].

We summarized unsecured overnight call rate and Japanese Government Bonds in terms of the following five elements: mean of the rate in financial month and percentage changes in one/three/six/twelve months up to financial month. Therefore, there are 15 elements summarizing unsecured overnight call rate, 1-year Japanese Government Bond, and 10-year Japanese Government Bond, altogether.

We summarized exchange rates in terms of the following six elements: mean of the rate in financial month, percentage changes in one/three/six/twelve months up to financial month, and volatility in financial month. Therefore, there are 12 elements summarizing Yen to US dollar exchange rate and Yen to Euro exchange rate, altogether.

Since only monthly data are available for Nikkei 225, we summarized Nikkei 225 in terms of the following six elements: mean of the highest price and the lowest price in financial month, percentage changes in one/three/six/twelve months up to financial month, and volatility index[4] of Nikkei 225 in financial month. As above, by using the percentage changes in various periods, time-series information of macroeconomic indexes can be summarized.

In total, these 33 elements of macroeconomic indexes were treated as inputs of random forest.

### 3.3 Experiment 1: Analyzing Important Financial Indicators by Industry Sectors

#### 3.3.1 Premise

In Experiment 1, only the financial indicators were used as inputs of the default prediction model, the macroeconomic indexes not being used. For each industrial sector, random forest was learned on the basis of financial data only and we predicted whether each company would go into default or not in a year from financial month. Financial data, account closing months of which were in 2010, 2011, 2013, and 2015, were used as training data, and financial data, account closing months of which were in 2012, 2014, and 2016, were used as test data. The numbers of default and non-default companies used in Experiment 1 by industry sectors and by financial years, are shown in Table 2. On the basis of variable importance of the financial indicators evaluated with random forest, we considered influential financial indicators on corporate default by industry sectors.

#### 3.3.2 Results

The variable importances of the 18 financial indicators are shown in Figure 1 by industry sectors. In this work, we repeated the calculation 50 times. The performance of random forest will be shown in Section 3.4, where the model used in Experiment 1 is referred to as Model 1.

[1]Source of data is the website of Bank of Japan Time-Series Data Search (https://www.stat-search.boj.or.jp/). We are permitted to use data for non-commercial purposes.

[2]Source of data is the website of Ministry of Finance (https://www.mof.go.jp/jgbs/reference/interest_rate/index.htm). We are permitted to use data for non-commercial purposes.

[3]Source of data is Nikkei 225 Official Site (https://indexes.nikkei.co.jp/nkave/index). We obtained permission from Nikkei Inc. to use data of Nikkei 225.

[4]Source of data is volatility index of Nikkei 225 Official Site (https://indexes.nikkei.co.jp/nkave/index). We obtained permission from Nikkei Inc. to use data of Nikkei 225.

### 3.3.3 Discussion

In all the industry sectors except the information and communication industry, the most or second most important variable is quick assets ratio. This result corresponds to common knowledge that quick assets ratio is the more appropriate indicator than current ratio for the rigid evaluation for short-term safety. On the other hand, the reason that quick assets ratio has a not high variable importance in the information and communication industry is that information and communication business does not need so much inventory that current ratio and quick assets ratio are almost the same.

In all the industry sectors, debt ratio for total assets is also important, which is the most, second most, or third most important variable in the construction industry, the manufacturing industry, the information and communication industry, the wholesale and retail industry, and the real estate and goods leasing industry. Additionally, in other three industry sectors variable importance of the debt ratio for total assets is not low. The result that the debt ratio for total assets is important is as we expected. On the other hand, in all the industry sectors ratio of depreciation to sales has a low variable importance: the ratio of depreciation to sales has the lowest importance in the construction industry, the manufacturing industry, the wholesale and retail industry, the accommodation and food service industry, and the lifestyle-related entertainment industry and not high importance in the other three industry sectors. The ratio of depreciation to sales is difficult to use to make rules, such that the higher the indicator is, the safer the company is. Our result confirmed the properties. Consequently, the ratio of depreciation to sales should be focused on in comparison of a company with a few other companies in the same industry sector.

The ratio of interest-bearing debt to cash-flow is an indicator implying how long it takes to return interest-bearing debt in current cash flow of the company. Industry sectors where the importance of the ratio of interest-bearing debt to cash-flow is high are the manufacturing industry and the real estate and goods leasing industry. In these two industry sectors the initial investment is costly, e.g. buying large-scale equipment in the manufacturing industry and purchasing real estates. Our result implies whether the initial debt can be returned or not has a large influence on corporate default in the manufacturing industry and the real estate and goods leasing industry.

On the other hand, accounts receivable turnover period is an indicator implying how many days it will take for the accounts receivable to become cash. Only in the information and communication industry, the importance of accounts receivable turnover period is high. The reason is considered that companies in the information and communication industry have so few assets that can be cashed quickly that the companies cannot repay the debts when accounts receivable turnover period is extended.

Lastly, trade payable turnover period is an indicator representing how many days it will take to return the purchase obligation on the basis of ratio of accounts and bills payable to sales. In the same way as the ratio of depreciation to sales, the trade payable turnover period is also difficult to use to make rules, such that the higher the indicator is, the safer the company is. That is why the importance of the trade payable turnover period is low in seven out of the eight industry sectors. However, in the accommodation and food service industry, the trade payable turnover period is the second most important indicator. Although we could not find any clear explanation for that, the trade payable turnover period should be focused on in default prediction for companies in the accommodation and food service industry.

## 3.4 Experiment 2 : Comparison Between Models Considering Only Financial Indicators and Considering Also Macroeconomic Indexes

### 3.4.1 Premise

In Experiment 2, not only the financial indicators but also macroeconomic indexes were used as inputs of the default prediction model. We devised the following three random forest models. Model 1 takes as input only the 18 financial indicators. Model 2 takes as input the 18 financial indicators and the following 15 elements of macroeconomic indexes: mean of the rate in financial month and percentage change in one year up to financial month, of unsecured overnight call rate and Japanese Government Bonds, mean of the rate in financial month, percentage change in one year up to financial month, and volatility in financial month, of Yen to US dollar and Yen to Euro exchange rates, and mean of the highest price and the lowest price in financial month, percentage change in one year up to financial month, and volatility index of Nikkei 225 in financial month, of Nikkei 225. Model 3 takes as input the 18 financial indicators and all the 33 elements of macroeconomic indexes listed in Section 3.2.4. Models 1, 2, and 3 predicted whether each company would go into default or not in a year from financial month. Financial data, account closing months of which were in 2010, 2011, 2013, and 2015, were used as training data, and financial data, account closing months of which were in 2012, 2014, and 2016, were used as test data. Also in Experiment 2, the models learned the training data of each industry sector, and conducted default prediction for companies in the same industry sector. The default and non-default companies used in Experiment 2 were the same as those used in Experiment 1. The numbers of default and non-default companies used in Experiment 2 by industry sectors and by financial years, are shown in Table 2.

Table 2: The numbers of default (D) and non-default (ND) companies used in Experiments 1 and 2, by industry sectors and by financial years.

| | 2010 | | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | | 2016 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Industry Sectors | D | ND | D | ND | D | ND | D | ND | D | ND | D | ND | D | ND |
| Construction | 246 | 42786 | 255 | 42754 | 226 | 42546 | 279 | 42965 | 250 | 42596 | 584 | 42780 | 1151 | 42673 |
| Manufacturing | 255 | 40897 | 206 | 40906 | 245 | 40862 | 274 | 40918 | 308 | 41221 | 584 | 40972 | 996 | 41024 |
| Information and Communication | 26 | 4925 | 29 | 4864 | 28 | 4918 | 40 | 4927 | 36 | 4941 | 79 | 4848 | 105 | 4877 |
| Wholesale and Retail | 257 | 60498 | 254 | 60376 | 305 | 60411 | 351 | 60666 | 402 | 60421 | 939 | 60320 | 1721 | 60208 |
| Real Estate and Goods Leasing | 86 | 8674 | 65 | 8692 | 54 | 8755 | 59 | 8763 | 64 | 8762 | 115 | 8794 | 169 | 8760 |
| Accommodation and Food Service | 78 | 13961 | 106 | 13918 | 97 | 13966 | 114 | 14019 | 95 | 13963 | 191 | 14004 | 296 | 13869 |
| Lifestyle-related Entertainment | 22 | 4559 | 26 | 4554 | 30 | 4591 | 31 | 4573 | 34 | 4572 | 57 | 4525 | 119 | 4526 |
| Medical and Welfare | 8 | 2839 | 10 | 2762 | 17 | 2832 | 16 | 2879 | 20 | 2923 | 43 | 2882 | 85 | 2783 |

Table 3: Comparison of model performance measured by AUC, industry by industry, in Experiment 2. The results of AUC are shown by mean and standard deviation (SD) multiplied by $10^{-3}$ through 50 calculations.

| | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| Industry Sectors | Mean | SD | Mean | SD | Mean | SD |
| Construction | 0.776 | 0.756 | 0.808 | 0.505 | 0.815 | 0.678 |
| Manufacturing | 0.823 | 0.618 | 0.847 | 0.542 | 0.850 | 0.525 |
| Information and Communication | 0.764 | 2.56 | 0.789 | 1.75 | 0.798 | 1.60 |
| Wholesale and Retail | 0.813 | 0.724 | 0.841 | 0.544 | 0.844 | 0.431 |
| Real Estate and Goods Leasing | 0.780 | 1.98 | 0.790 | 1.29 | 0.795 | 1.41 |
| Accommodation and Food Service | 0.760 | 1.67 | 0.767 | 0.990 | 0.771 | 1.15 |
| Lifestyle-related Entertainment | 0.744 | 2.67 | 0.772 | 1.38 | 0.764 | 1.75 |
| Medical and Welfare | 0.727 | 3.34 | 0.759 | 2.20 | 0.771 | 1.92 |

### 3.4.2 Results

Comparison of the performance of the models measured by AUC is shown in Table 3 and in Figure 2, where the mean and the standard deviation over 50 trials are shown. These imply that random forest learned the training data and predicted whether each company in the test data would go into default or not in a year from financial month.

### 3.4.3 Discussion

By comparison of Model 1 and Model 2 in all the industry sectors, we confirmed that considering macroeconomic indexes improves the default prediction model for small and medium-sized companies.

In terms of comparison of Model 2 and Model 3, Model 3 outperformed Model 2 in seven out of the eight industry sectors except the lifestyle-related entertainment industry. Additionally, there was a significant difference between AUC of Model 2 and Model 3 on the basis of Mann Whitney U test, all p values of which were lower than $1.0 \times 10^{-17}$ in all the seven industry sectors. From this result, considering both of microtrends and macrotrends improves the default prediction model for small and medium-sized companies in the seven industry sectors. Although we could not find any clear explanation for the lower performance of Model 3 than Model 2 in the lifestyle-related entertainment industry, we considered that the reason would be the insufficient amount of data in the lifestyle-related entertainment industry. In the lifestyle-related entertainment industry, there are a wider variety of companies than in any other industry sectors, such as barber shops, tour businesses, movie theaters, and fitness clubs. Therefore, it should be different from a company to another how the macroeconomic factors affect on the default of the company. In order to classify the wide variety of companies, the sufficient amount of data is essential. It is true that there were a large amount of financial data provided by CRD. However, focusing on only the number of companies in the lifestyle-related entertainment industry, there were only a few amount of data, as shown in Table 2. With a sufficient amount of data in the lifestyle-related entertainment industry, the models will learn appropriately and we can improve the default prediction model for small and medium-sized companies by considering both of microtrends and macrotrends, in the lifestyle-related entertainment industry.

## 4 Conclusion

We have studied improvement of the default prediction model for small and medium-sized companies by considering not only their financial data but also macroeconomic indexes and regarding which financial indicators are important by industry sectors in default prediction for small and medium-sized companies. As a result, the default prediction model with inputs of the financial indicators and the macroeconomic indexes outperformed the model with inputs of only financial indicators. Furthermore, the model, to the inputs of which the percentage changes in the fine periods summarizing microtrends added, outperformed the model not considering the percentage changes in the fine periods. Therefore, considering macroeconomic indexes, especially our proposed method summarizing both macrotrends and microtrends, has been found effective for default prediction. Furthermore, we investigated which financial indicators are important by industry sectors.

Although the results of investigating which financial indicators are important by industry sectors presented in Section 3.3 gave us useful insights for default prediction for small and medium-sized companies, the knowledge obtained from the results does not include which macroeconomic indexes are important by industry sectors. The reason of this is that on the basis of variable importance in random forest, the variable importances of macroeconomic indexes turned out to be so lower than those of financial indicators that we were not able to consider which macroeconomic indexes are important by industry sectors. In order to consider which macroeconomic indexes are important by industry sectors, appropriate approaches for this purpose should be considered.

On the other hand, in terms of macroeconomic indexes, another interesting question would be as to which periods one should take in summarizing them into the mean, the volatility, and the trend. Although this paper considered some periods within one year, wider or finer periods, or both, should be considered. Moreover, this research set the base period of macroeconomic indexes as financial month. However the base period should be discussed in more detail. We conjecture that the appropriate base period of macroeconomic indexes would be different by industry sectors. Discovery of the appropriate base period by industry sectors will improve the default prediction for small and medium-sized companies further.

## Acknowledgment

We would like to thank Professor Hideaki Aoyama, Graduate School of Science, Kyoto University, Kyoto, Japan, for his valuable advice and managing financial data, and the CRD institution for providing their data.

## References

[Ali 10] Ali, A. and Daly, K.: Macroeconomic determinants of credit risk: Recent evidence from a cross country study, *International Review of Financial Analysis*, Vol. 19, No. 3, pp. 165–171 (2010)

[Alifiah 14] Alifiah, M. N.: Prediction of financial distress companies in the trading and services sector in Malaysia using macroeconomic variables, *Procedia-Social and Behavioral Sciences*, Vol. 129, pp. 90–98 (2014)

[Altman 68] Altman, E. I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, Vol. 23, No. 4, pp. 589–609 (1968)

[Altman 77] Altman, E. I., Haldeman, R. G., and Narayanan, P.: ZETA$^{TM}$ analysis: A new model to identify bankruptcy risk of corporationcs, *Journal of Banking & Finance*, Vol. 1, No. 1, pp. 29–54 (1977)

[Altman 94] Altman, E. I., Marco, G., and Varetto, F.: Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience), *Journal of Banking & Finance*, Vol. 18, No. 3, pp. 505–529 (1994)

[Barboza 17] Barboza, F., Kimura, H., and Altman, E.: Machine learning models and bankruptcy prediction, *Expert Systems with Applications*, Vol. 83, pp. 405–417 (2017)

[Bonfim 09] Bonfim, D.: Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics, *Journal of Banking & Finance*, Vol. 33, No. 2, pp. 281–299 (2009)

[Chen 11] Chen, S., Härdle, W. K., and Moro, R. A.: Modeling default risk with support vector machines, *Quantitative Finance*, Vol. 11, No. 1, pp. 135–154 (2011)

[Chen 14] Chen, P. and Wu, C.: Default prediction with dynamic sectoral and macroeconomic frailties, *Journal of Banking & Finance*, Vol. 40, pp. 211–226 (2014)

[Flagg 91] Flagg, J. C., Giroux, G. A., and Jr, C. E. W.: Predicting corporate bankruptcy using failing firms, *Review of Financial Economics*, Vol. 1, No. 1, pp. 67–78 (1991)

[Goldberg 04] Goldberg, L. R. and Giesecke, K.: Forecasting default in the face of uncertainty, *The Journal of Derivatives*, Vol. 12, No. 1, pp. 14–25 (2004)

[Hastie 09] Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning*, Springer, 2nd edition (2009)

[Kim 10] Kim, M.-J. and Kang, D.-K.: Ensemble with neural networks for bankruptcy prediction, *Expert Systems with Applications*, Vol. 37, pp. 3373–3379 (2010)

[Min 05] Min, J. H. and Lee, Y.-C.: Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert Systems with Applications*, Vol. 28, pp. 603–614 (2005)

[Nam 08] Nam, C. W., Kim, T. S., Park, N. J., and Lee, H. K.: Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies, *Journal of Forecasting*, Vol. 27, No. 6, pp. 493–506 (2008)

[Odom 90] Odom, M. D. and Sharda, R.: A neural network model for bankruptcy prediction, *IEEE International Joint Conference on Neural Networks*, pp. 163–168 (1990)

[Shin 05] Shin, K.-S., Lee, T. S., and Kim, jung H.: An application of support vector machines in bankruptcy prediction model, *Expert Systems with Applications*, Vol. 28, pp. 127–135 (2005)

[Tinoco 13] Tinoco, M. H. and Wilson, N.: Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables, *International Review of Financial Analysis*, Vol. 30, pp. 394–419 (2013)

[Wang 11] Wang, G. and Ma, J.: Study of corporate credit risk prediction based on integrating boosting and random subspace, *Expert Systems with Applications*, Vol. 38, pp. 13871–13878 (2011)

[Zhang 10] Zhang, J. L. and Hardle, W. K.: The Bayesian additive classification tree applied to credit risk modelling, *Computational Statistics and Data Analysis*, Vol. 54, pp. 1197–1205 (2010)

[安道 04] 安道 知寛, 山下 智志 : 財務指標の時間依存を考慮した信用リスク評価モデルーデフォルト予測への応用, 金融庁ディスカッションペーパー (2004)

[高橋 02] 高橋 久尚, 山下 智志 : 大規模データによるデフォルト確率の推定ー中小企業信用リスク情報データベースを用いてー, 統計数理, Vol. 50, No. 2, pp. 241–258 (2002)

[三浦 08] 三浦 翔, 山下 智志, 江口 真透 : 信用リスクスコアリングにおける AUC と AR 値最大化法, 金融庁ディスカッションペーパー (2008)

[山下 03] 山下 智志, 川口 昇 : 大規模データベースを用いた信用リスクの計測の問題点と対策（変数選択とデータ量の関係）, 金融庁ディスカッションペーパー (2003)

[山下 04] 山下 智志, 安道 知寛 : 時間依存共変量を用いたハザードモデルによるデフォルト確率期間構造の推計手法, 統計数理, Vol. 54, No. 1, pp. 23–38 (2004)

[山下 11] 山下 智志 : 与信判断が確率変動するときの倒産企業の信用リスク値分布のモデル, 統計数理, Vol. 59, No. 1, pp. 3–23 (2011)

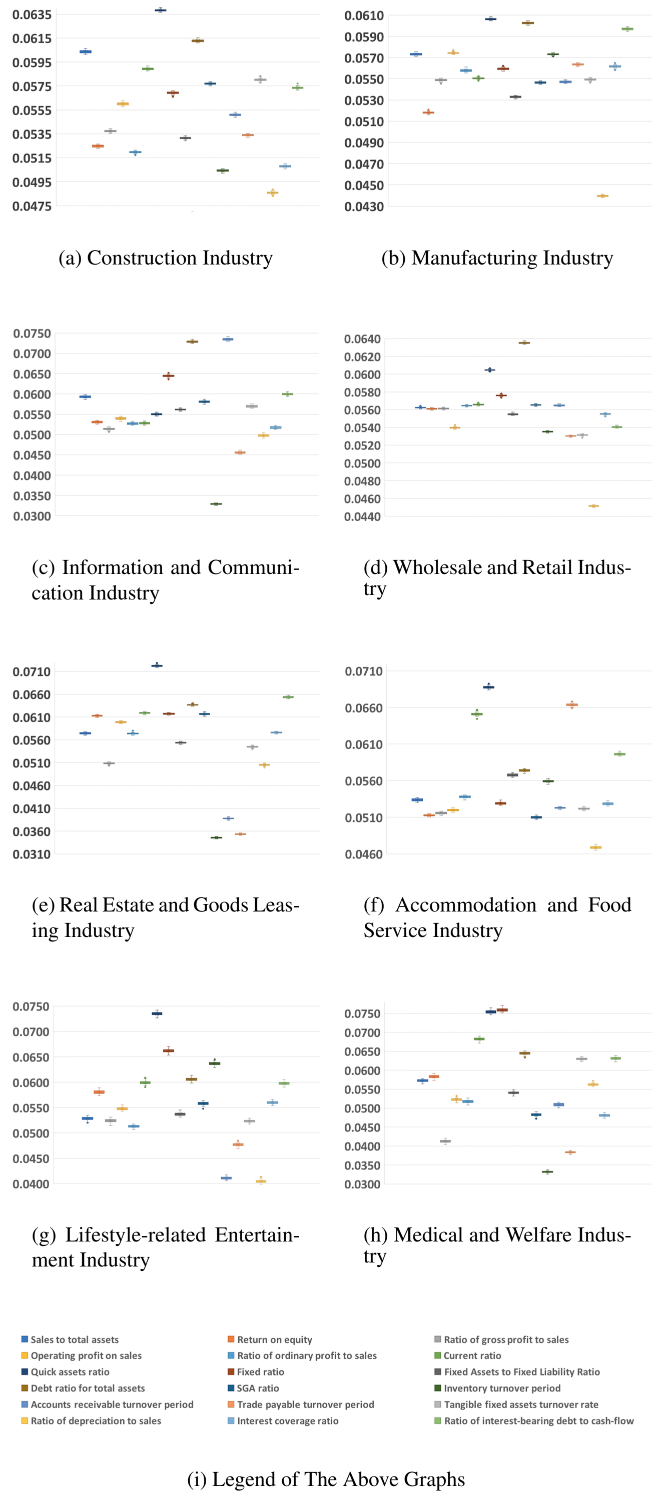
[森平 09] 森平 爽一郎, 岡崎 貫治 : マクロ経済効果を考慮したデフォルト確率の期間構造推定, 日本ファイナンス学会第１７回大会予稿集, pp. 103–112 (2009)

[大橋 03] 大橋 亨 : 企業倒産とマクロ経済要因ー企業倒産についての重回帰モデルの構築ー, 現代社会文化研究, Vol. 28, pp. 47–67 (2003)

[辻 07] 辻 正雄, ヴャチェスラフジュロフ : 企業倒産予知モデルの構築と比較分析, 産業経理, Vol. 67, No. 1, pp. 4–17 (2007)

[梅谷 13] 梅谷 幸平 : 経営管理目的からの倒産予測モデル研究の有用性と課題: 経営管理のための安全性指標の観点から, 大阪大学経済学会, Vol. 62, No. 4, pp. 63–83 (2013)

[尾木 13] 尾木 研三, 森平 爽一郎 : 中小企業のデフォルト率に影響を与えるマクロ経済要因ー１ファクターモデルを用いたアプローチー, 日本政策金融公庫論集, No. 20, pp. 71–89 (2013)




(a) Construction Industry

(b) Manufacturing Industry

(c) Information and Communication Industry

(d) Wholesale and Retail Industry

(e) Real Estate and Goods Leasing Industry

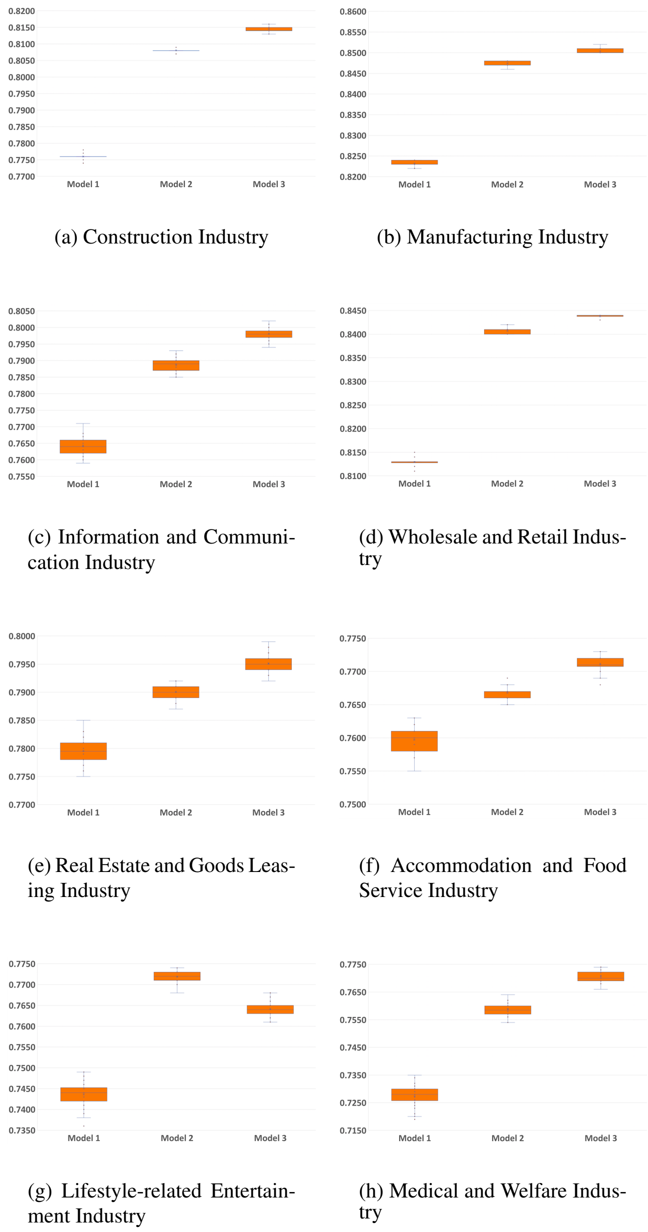
(f) Accommodation and Food Service Industry

(g) Lifestyle-related Entertainment Industry

(h) Medical and Welfare Industry

(i) Legend of The Above Graphs

Figure 1: Comparison of variable importances of 18 financial indicators calculated by random forest in Experiment 1. Each box plot shows the distribution of variable importance through 50 calculations.




(a) Construction Industry

(b) Manufacturing Industry

(c) Information and Communication Industry

(d) Wholesale and Retail Industry

(e) Real Estate and Goods Leasing Industry

(f) Accommodation and Food Service Industry

(g) Lifestyle-related Entertainment Industry

(h) Medical and Welfare Industry

Figure 2: Comparison of random forest performance by industry sectors in Experiment 2. Each box plot shows the distribution of AUC through 50 calculations.