Nguyen, Lan H.; Sagara, Megumi

Working Paper

Credit risk database for SME financial inclusion

ADBI Working Paper Series, No. 1111

Provided in Cooperation with:
Asian Development Bank Institute (ADBI), Tokyo

Suggested Citation: Nguyen, Lan H.; Sagara, Megumi (2020) : Credit risk database for SME financial inclusion, ADBI Working Paper Series, No. 1111, Asian Development Bank Institute (ADBI), Tokyo

This Version is available at:
https://hdl.handle.net/10419/238468

ADBI Working Paper Series

CREDIT RISK DATABASE FOR SME FINANCIAL INCLUSION

Lan H. Nguyen and Megumi Sagara

No. 1111
April 2020

Asian Development Bank Institute

Lan H. Nguyen is a senior analyst at the Credit Risk Database Association. Megumi Sagara is director of the Overseas Division at Credit Risk Database Association.

The Working Paper series is a continuation of the formerly named Discussion Paper series; the numbering of the papers continued without interruption or change. ADBI's working papers reflect initial ideas on a topic and are posted online for discussion. Some working papers may develop into other forms of publication.

In this report, "$" refers to United States dollars.

Suggested citation:

Nguyen, L. H. and M. Sagara. 2020. Credit Risk Database for SME Financial Inclusion. ADBI Working Paper 1111. Tokyo: Asian Development Bank Institute. Available: https://www.adb.org/publications/credit-risk-database-sme-financial-inclusion

Please contact the authors for information about this paper.

Email: sagara@crd-office.net, nguyen@crd-office.net

Asian Development Bank Institute
Kasumigaseki Building, 8th Floor
3-2-5 Kasumigaseki, Chiyoda-ku
Tokyo 100-6008, Japan

Tel: +81-3-3593-5500
Fax: +81-3-3593-5571
URL: www.adbi.org
E-mail: info@adbi.org

Abstract

We introduce the Credit Risk Database (CRD) and its contribution to financial inclusion efforts in Japan. By collecting financial data about small and medium-sized enterprises (SMEs), the CRD contributes to the overall understanding of the SME sector, to the adaptation of risk-based lending and to a fairer loan guarantee system. In addition to financial data, the CRD also includes alternative data, bank account transaction data, when assessing SME credit. A machine learning model is adopted to process the extremely large body of transaction data. The best performing predictors of default include cash balance and cash outflow related to repayments. The machine learning model outperforms the logistic model and is highly accurate in predicting the probability of short-term default. The alternative data and model can serve as both an enabling short-term monitoring instrument and a credit assessment tool for SMEs without financial statements.

Keywords: SME credit assessment, machine learning, transaction data

JEL Classifications: G21, G28, G32

Contents

1. INTRODUCTION

The access that small and medium-sized enterprises (SMEs) have to finance has been central to policy discussion in both advanced and developing countries. Maningo (2016) and Amornkitvikai et al. (2016) describe common situations in developing Asia where SMEs struggle with limited funding, restrictive repayment periods, high financing costs, and a complex lending process. One of the major concerns is collateral, which most SMEs cannot provide.

Efforts to loosen collateral requirements only began after the Asian financial crisis as a result of declining financial assets and property. These efforts, however, made only a small impact on closing the SME finance gap. The World Bank reports an unmet SME finance demand of about $5.3 trillion every year, the largest share of which is from Asia. Financial institutions are encouraged to further lessen dependence on collateral and to fast-track risk-based lending for SMEs, to improve their access to finance.

In this paper, we introduce the Japanese approach to SME risk-based lending. We present Japan's nationwide financial database, which plays an important role as SME infrastructure enabling SME credit risk analysis and modeling. We also introduce the three most important contributions of the database: SME statistical analysis, credit risk assessment models and the guarantee fee rating scheme. Second, we explore bank transaction data as an alternative for incorporation into risk models to tackle challenges in short-term monitoring and the lack of financial information for micro SMEs and startups. We briefly describe our model, which was developed together with the Bank of Japan (BOJ) and a Japanese Megabank, along with the main variables, the machine learning approach and its overall performance. Finally, policy implications for the improvement of SME access to finance are drawn from the Japanese experience, indicating the leadership role that the financial sector authority should take in establishing SME data infrastructure and promoting statistical analysis using both traditional financial data and alternative data.

2. JAPAN'S SME FINANCIAL INFRASTRUCTURE: CREDIT RISK DATABASE

2.1 Background

The Credit Risk Database was established following the collapse of the Japanese bubble economy in the 1990s. The collapse caused a sharp decline in land price and a consequent credit crunch affecting many businesses, especially SMEs. Banks at that time needed to move away from lending basing on land-collateral to lending based on credit risk. To evaluate a business based on credit risk, it is necessary to use a statistical model with a relatively large pool of data. In 2001, in order to establish such a large pool of data, the SME Agency, an authorized body of the Ministry of Economy, Trade and Industry, and the BOJ together established the Credit Risk Database (CRD) using government funding. The CRD is managed by the CRD Association, a private not-for-profit organization with data-providing institutions as membership-paying members.

Data has been duly collected from member financial institutions since 2001. Members include 51 credit guarantee corporations, 103 private financial institutions and five government-affiliated institutions. The BOJ, Financial Service Agency (FSA) and SME Agency are also members. As of November 2019, the database had collected

2.5 million SMEs and 1.2 million sole proprietors. There are 21 million financial statements (data points) for SMEs and 5.6 million financial statements for sole proprietors.

2.2 Impact of the Credit Risk Database on the Financial Inclusion of SMEs

The database contributes to SME financial inclusion in several ways. First, it helps mitigate the information asymmetry problem by providing information and benchmark statistics for SME sectors. Information such as the business performance of a typical SME in a specific industry in a specific region is now available. Policy makers such as the SME Agency, BOJ, and FSA are frequent users of such overall statistical data and use it for policy analysis. The recent SME Agency white paper (Japan SME Agency 2019) included an interesting study of how business performance could be improved through a shift in management from elderly to younger managers. A recent study of the database also provides insights into the impact of subsidies policy in an evidence-based policy making (EBPM) study sponsored by the Japan SME Agency (Deloitte Tohmatsu Consulting 2018).[1]

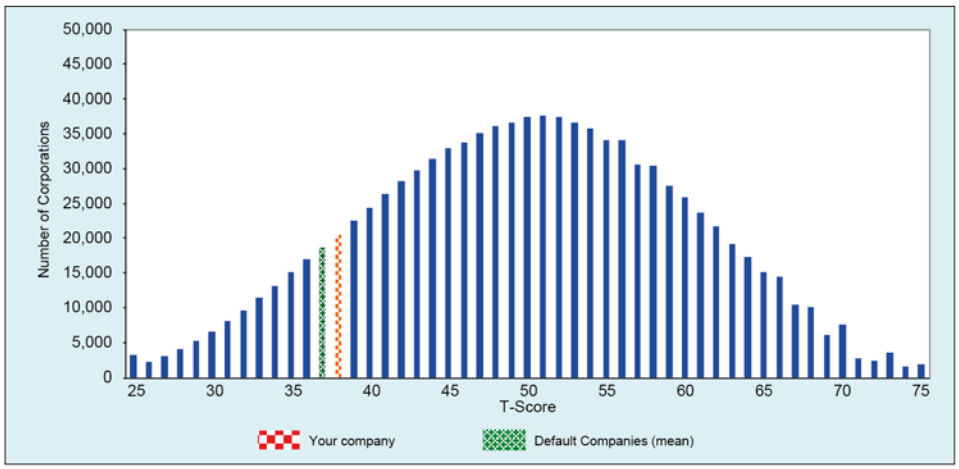
The database also makes the credit risk evaluation of SMEs possible. Credit scores are given to an SME based on the credit risk model developed by the CRD Association. All SMEs who are clients of a member bank are eligible for the scoring service. Where a member bank already has an internal risk model, the scores given by the CRD Association can be used to cross-checking or validate the internal model, a practice that is strongly recommended by the Japanese financial authority. Figure 1 presents the distribution of credit scores for a sample set of data. The horizontal axis represents the credit score range, in which 0 indicates the highest risk and 100 indicates the lowest risk. The company introduced in this example has a score of 38. To determine whether 38 is good or bad, it is useful to look at its position relative to the mean score of defaulted companies. The mean score takes a value of 37, as depicted in green. According to its score, this company's credit position is relatively close to default, it ranks 7,200th in its industry and 220,000th against similar sized sales groups. Benchmarking in the region and further breakdowns by industry are also feasible.

The third impact of CRD is its contribution to improving the guarantee policy for SMEs. A SME in Japan can apply to be guaranteed by a credit guarantee corporation if deemed necessary by the lending institution. The SME pays a guarantee fee in exchange for the guarantee service. In case of default, the credit guarantee corporation shall bear all responsibility to reimburse the lending institution. In 2006, the credit guarantee corporation changed their fee structure from charging a universal guarantee rate to charging a different rate based on credit risk measured by CRD risk evaluation model. In this fee scheme, the riskiest SME pays a guarantee fee of 2.2% of the total loan amount, while the least risky SME pays only 0.5%. This fairer fee system, considering credit quality, has helped to lower the subrogation rate and mitigate the adverse selection problem. Figure 3 describes the nine levels corresponding to the nine ranks of credit risk and their respective fee rates.

[1] The study investigates firm performance before and after receiving subsidies from the government. Firms are categorized by their level of creditworthiness measured by default probability, and by growth potential measured by growth scores. Both default probability and growth scores are build-in parameters calculated using CRD models. For further details, refer to the Japan SME Agency sponsored paper (in Japanese): https://www.meti.go.jp/meti_lib/report/H30FY/000316.pdf.

As part of the SME inclusion efforts, the credit grantee system for SMEs is being discussed and designed in many developing nations, such as the Philippines, Myanmar, Thailand, and so on. Maningo (2016) pointed out that the lack of SME information is a challenge for the Philippines' Credit Surety Fund, in order to correctly assess and manage default risk, as the past due ratio increased to 1.6% in 2014 compared to the 2013 level of 0.6%. A database such as the CRD could both provide SME information for guarantee decision-making and create an enabling instrument to strategically decide guarantee fees.

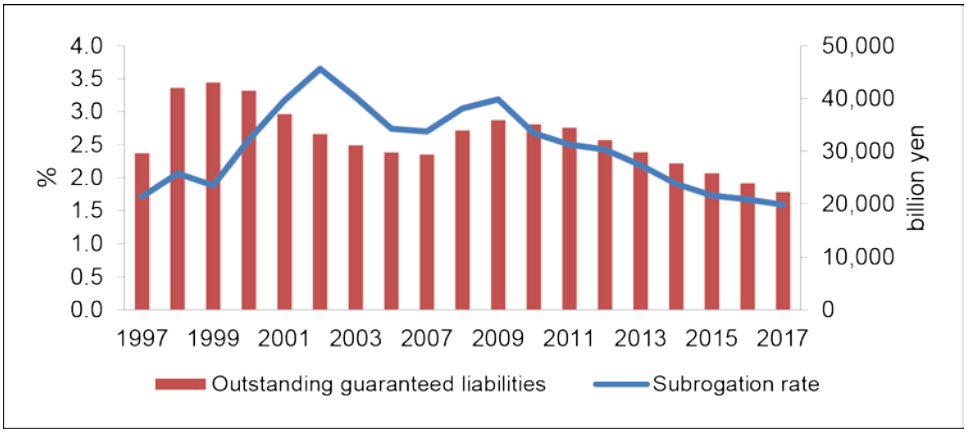
Figure 1: Credit Score and Its Uses



Fiscal Term	C R D Rank (A~E)	Your company T-Score	Default Companies T-Score (mean)	Ranking in the "Manufacture of transportation equipment" Group	Ranking in Tokyo	Ranking in the Same Scale of Sales	
						Category	Ranking
March, 2014	D	38	37	7,200 (in 8,000)	120,000 (in 140,000)	\100 to \300 million	220,000 (in 250,000)
March, 2013	C	45	37	6,000 (in 8,000)	91,000 (in 140,000)	\100 to \300 million	175,000 (in 250,000)
March, 2012	C	46	37	5,700 (in 8,000)	85,000 (in 140,000)	\100 to \300 million	165,000 (in 250,000)

Source: CRD Association.

Figure 2: Subrogation Rate and Outstanding Guaranteed Liabilities



* Subrogation rate = Subrogation Amount/Outstanding Guarantee Liabilities.

Source: Japan Federation of Credit Guarantee Corporation.

Figure 3: Credit Guarantee Fees Scheme before and after 2006

<Until March, 2006>

Classification	No Classification
Credit Guarantee fee rate	All 1.35

<Since April, 2006>

Classification	1	2	3	4	5	6	7	8	9
Credit Guarantee fee rate	2.20	2.00	1.80	1.60	1.35	1.10	0.90	0.70	0.50

Source: Japan Federation of Credit Guarantee Corporation.

2.3 Credit Scoring Model for Credit Risk Management

The conventional credit scoring model normally applies logistic regression to predict default, which often takes a binary value: default or not default. A common approach of the various default definitions is to follow the monetary authority's borrower classification definition. Borrowers in Japan are typically categorized into six main categories: Normal, Needs Attention, Special Attention, In Danger of Bankruptcy, De facto Bankrupt and Bankrupt. In addition to borrower classification, loan repayment records can be complimentarily used as a default definition. Such records are conventionally updated monthly by most lending institutions.

The traditional model mainly examines the correlation between default and financial ratios derived from a financial statement. At the time of the loan screening, three-consecutive-year of financial statement or more are often requested for credit checking. These statements are then verified by lending officers who examine the statement accountability. By collecting financial statements from the lending institutions, we are likely to receive information with high accountability. In fact, collecting financial statements from lending institutions has been the main CRD approach.

Building the model ideally involves a wide range of financial ratios to capture profitability, efficiency and stability, as well as the growth potential of the business. Given the volume of the CRD database, we have included financial ratios representing all above categories without having to resort to using non-financial data to obtain a model with high accuracy.[2]

The output of the model is default probability within 1 to 3 years which conveniently reflects a business's chance of bankruptcy within the respective time horizon. This time horizon is a common time frame that lenders can look at: 1 year for a credit line and 2–3 years for a term loan. The current model could not provide such analysis if lenders need to look at default probability within a few months, due to the lack of input variables, because interim financial statements are often unavailable. The bank transaction model which outputs default probability within 3–6 months could resolve this problem. Details of the model are discussed in Section 3.

To ensure the model's adaptation to the latest portfolio, it is validated yearly under the supervision of an independent committee of scholars, financial sector practitioners and policy makers. Committee members should understand the technical aspects of the model, as well as validation procedures. Model accuracy is tested against an out-of-

[2] 35–85 financial items in financial statements, approximately 174 financial ratios, are calculated and become candidates for model variables.

sample dataset segmented into industries and regions. If the model doesn't fit to the out-of-sample dataset and accuracy falls, adjustments to the model shall be advised by the committee. The results of validation are available yearly to all member institutions. Validation framework is described in more details by Kuwahara et al. (2019).

3. CREDIT SCORING USING BANK TRANSACTION DATA WITH THE MACHINE LEARNING APPROACH

This section summarizes the CRD's published research, jointly undertaken with the Bank of Japan and a Japanese megabank to adapt alternative data for credit risk evaluation.[3] Miura et al. (2019) address the need for short-term monitoring by using bank transaction data to predict short-term default probability. The approach involves applying machine learning to high-frequency transaction data to study its patterns and correlation with defaults.

Apart from its monitoring function, the model innovatively reveals new financing opportunities for micro SMEs and startups who are bank account owners but have not prepared financial statements. The presence of such micro SMEs is especially noticeable in developing nations where the number of personal bank accounts is rapidly increasing but access to finance is still limited.

Bank transaction data is abundant and relatively easy for a deposit-taking bank to gather, as such data is recorded automatically. Almost 90%–100% of adults in developed countries have one or more accounts. In developing Asia, the number is lower, at approximately 78% in Thailand, 36% in Indonesia, 31% of adults in the Philippines, and 31% in Viet Nam (World Bank 2017). With increasing financial inclusion efforts in developing nations, the number of bankable individuals is expected to increase soon. Notable efforts to increase account ownership include full-service transaction accounts for government cash grants and the promotion of private sector wage payments by the government of the Philippines (Banko Sentral ng Pilipinas 2019). The abundance of available data allows machine learning to perform its best, and to provide reliable predictions without much human intervention.

Another merit of transaction data is its reliability and accountability. Large data is often associated with errors, but full automation in bank transaction recording, means that transaction data is typically free from missing values and nearly impossible to manipulate.

3.1 Dataset

The transaction dataset consists of data ranging from October 2014 to May 2018 (44 months).[4] The original frequency of the transaction data was measured in seconds but later transformed into monthly frequency for analysis purposes. The two main reasons for this transformation are that default information is often available in monthly frequency; and that some major transaction data, such as fixed revenues and fixed costs, are often monthly payments.

Table 1 describes the three categories of data in our dataset: cash-inflow, cash-outflow and cash balance. The *cash-inflow* subcategory mainly consists of cash flow from

[3] The original study is published in Bank of Japan Working Paper series (Japanese): https://www.boj.or.jp/research/wps_rev/wps_2019/wp19j04.htm/.

[4] This dataset is prepared for research purpose only.

various types of revenue, investing activities and financing activities. *Cash-outflow* comprises cash-out flow for payments for cost of goods and services, variable cost, fixed cost, investing activities and financing activities. Finally, *cash balance* is calculated on the daily average of a monthly balance instead of the end-of-month balance for a better reflection of cash movements.

Banks typically assign a code for each type of transaction, to identify its purpose for internal monitoring. Each transaction is classifiable to the three large categories noted, and to various subcategories by referring to this code. It should be noted, however, that such classification is based on the authors' experience in the banking industry and is not a universal approach. Furthermore, internal codes and the way a bank monitors transaction purposes could also be very different from bank to bank. To achieve a more standardized approach to classifying transactions, it is necessary to discuss them with other financial institutions and investigate their practices.

Table 2 describes the data transformations of the data given in Table 1. Level data, first difference data, frequency data and standardized data is constructed. The number of cash flow data points, including both money amounts and frequency, totals 984. The standardized cash-flow data using sales, total assets and outstanding loans data as the basis adds another 2,952 variables. The growth in cash-flow data in terms of year-on-year and month-on-month data adds another 1,435 variables. Finally, the period total and volatility of cash in-flows and out-flows data adds another 615 variables. In total there are approximately 6000 variables.

3.2 Default Definition

Default is defined in this study as borrowers classified as "Needing Special Attention" including those who have been late with repayments for 3 months or more. Records of default status are updated on monthly basis, as a common practice in the banking sector. Our transaction data will be matched with defaults observed within 3 months to produce a prediction of default probability within 3 months. In practice, we also work with other time horizons and calculate default probability within 1, 2, 6, 9, and 12 months. The dataset includes the data of 7,000 anonymous borrowers, of whom 1,400 are associated with default.

The matching process for default information and transaction data is straightforward. For example, the January 2016 transaction data will be matched with default information from February to April 2016 if our time horizon is 3 months and will be matched with default data from February to July 2016 if our time horizon is 6 months.

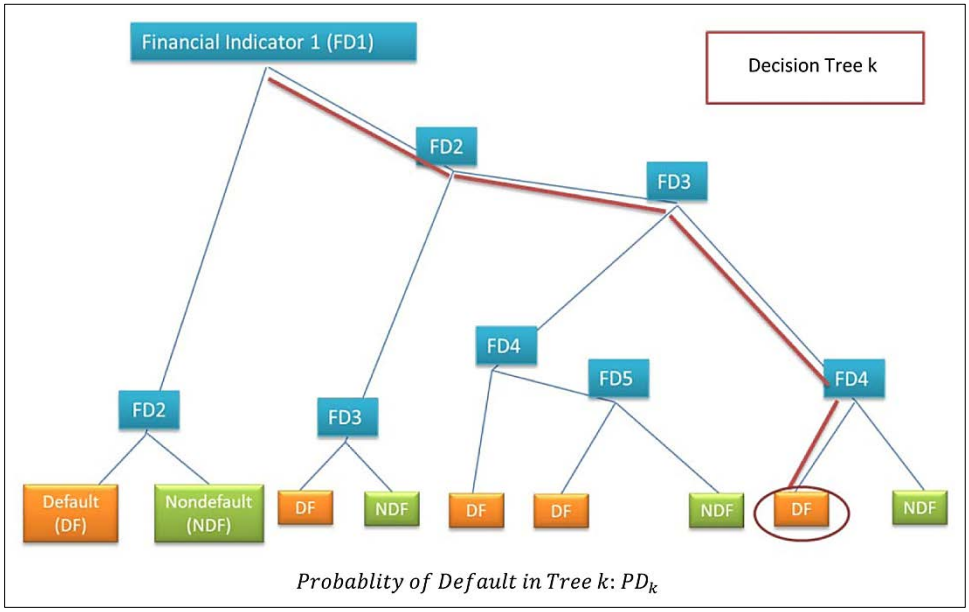
Table 1: Categories of Transaction Data

Cash-inflow	Revenue	Bank Transfers for Revenue, FX…
	Investing Activities	Dividend, Proceeds from Trust Fund…
	Financing Activities	Borrowing (Loans)…
	Other Cash-inflow	Cash, Credit Card, Interest…
Cash-outflow	Cost of Goods and Services	Bank Transfers…
	Variable Cost	Credit Card, Insurance, Tax…
	Fixed Cost	Utility (Electric, Gas, Water), Cable…
	Investment Activities	Investment Trust Fund…
	Financing Activities	Loan Repayment, Interest Payment, Guarantee Fees…
	Other Cash-outflow	Penalties…
Cash Balance	Cash Balance	Average Monthly Cash Balance…

Table 2: Summary of Variables

Variables	Data Type	Data Source
Cash-inflows, Cash-outflows and Cash Balance	Level	Transaction Data
Money Amount Frequency		
Sales	Level	Financial Data
Total Assets	Level	Financial Data
Outstanding Loan Amount	Level	Loan Data
Standardized Transaction Data by Sales, Total Assets and Outstanding Loan Amount.	Standardized	
YoY of Cash-inflows	First Difference	
MoM of Cash-outflows	First Difference	
Period Total of Cash-inflows, Cash-outflows and Cash Balance	Level	
Standard Deviation of Cash-inflows, Cash-outflows and Cash Balance		
YoY Period Total of Cash-inflows, Cash-outflows and Cash Balance	First Difference	

Figure 4: Example of a Decision Tree



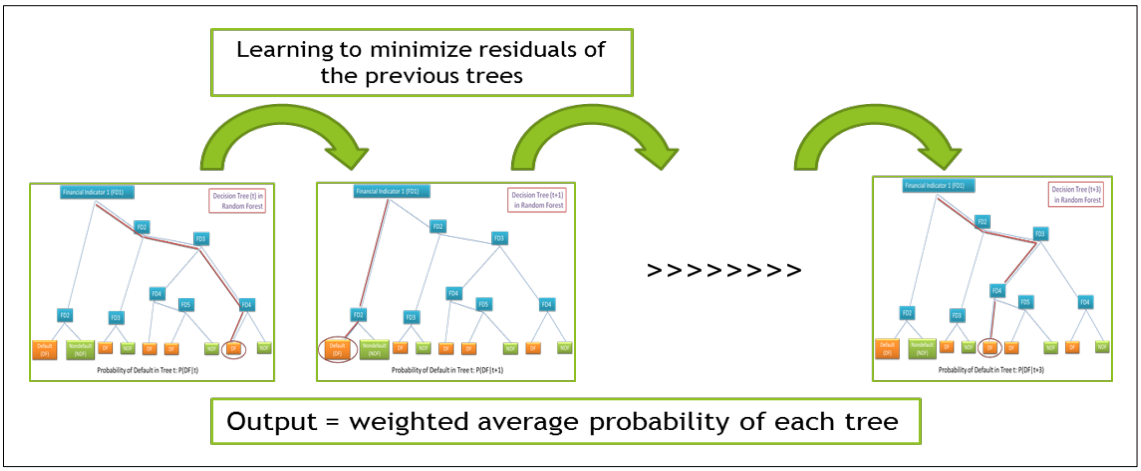
Source: Author.

3.3 Modelling Approach

Given the volume and the quality of the data, a machine learning model is a good choice compared to a traditional logistic model. Machine learning is powerful in dealing with a large volume and information-rich dataset to find certain patterns, and specifically a pattern of default in our case. It has the capacity to process many variables and the flexibility to accommodate numerous data patterns. The model uses Random Forest, a popular classification model in machine learning, which consists of a large number of uncorrelated decision trees (models). Each tree in the forest produces a prediction, and the average probability becomes the model's final prediction. The prediction errors of many uncorrelated models cancel out to output the best prediction.

The study further employs XGBoost, a more advanced version of Random Forest, which can customize the creation of each decision tree. In XGBoost, new trees are added to correct or minimize the errors of sequential trees. The best model is created when no further improvement can be made. XGBoost is easier to overlearn than Random Forest and so the regularization and cross-validation of the model is necessary to prevent such overlearning.

Figure 5: XG Boost Model



Source: Authors.

Table 3: Parameter Tuning Result for Random Forest and XGBoost (Out-of-sample Dataset)

			Random Forest		XGBoost	
Transaction Data	**Default Definition**	**Default Horizon**	**Hyper Parameter**	**AR**	**Hyper Parameter**	**AR**
Monthly	Needs Special Attention	3 months	(150, 300)*	0.699	(3, 0.01)**	0.7266
Monthly	Needs Special Attention	3 months	(150, 350)*	0.707	(4, 0.01)**	0.7329
Monthly	Needs Special Attention	3 months	(150, 400)*	0.702	(5, 0.01)**	0.7181

Note: * (number of trees, number of information features), ** (tree's max depth, learning rate).

Table 4: Out-of-sample Testing Results

Dataset	**Model**	**Default Horizon: 3 Months**
Testing Dataset (parameter tuning)	Random Forest	0.7070
	XGBoost	0.7329
	Logistic	0.7113
Back test	Random Forest	0.7482
	XGBoost	0.7728
	Logistic	0.7174

One drawback of machine learning in general is overlearning or overfitting to the dataset. Parameter tuning and out-of-sample cross validation help tackle this problem. For random forest, parameter tuning is undertaken for a number of trees and a number of information features (financial variables to be used), and in XGBoost, the hyper parameters are the number of information features, maximum depth of trees and the learning rate.

Given the abundance of data, cross-validation in the form of back-testing both the random forest model and XG Boost model is possible. Back-test datasets involve twelve out-of-sample datasets, each of which comprises monthly data from June 2017 to May 2018.

3.4 Model Performance

We first examine the prediction power of each variable by looking at the accuracy of a single-variable model. Accuracy ratio (AR) is a commonly used indicator for a model's performance among Japanese banks. It is closely related to, and is calculated from, the area under the curve of the ROC curve.[5] We first observe that monthly average cash balance has strong predicting power for defaults. Since the monthly average cash balance plays such an important role, we further examine whether the transformation of such data could be as powerful. Level data, period sum data and standardized data all produce a high accuracy ratio, some higher than 0.5.

In addition to cash balance, cash outflow for loans repayment, cash inflows from revenue, cash outflows for variable costs, cash outflow for cost of goods and services and cash inflows from loans, all demonstrate relatively strong correlations with defaulting by producing high AR.

We observe the same set of best performing variables when applying a different time horizon for default information, although their performance tends to deteriorate as the time horizon widens. Specifically, the cash balance AR drops from 0.52 for a one-month horizon to 0.46 for a twelve-month horizon while loan payment cash flow AR drops from 30th place to 88th place in the ranking of best performing variables.

Table 3 reports the results of parameter tuning, a grid-search procedure to configure the optimal combination of the hyper parameters producing highest model accuracy. Specifically, for random forest, when the default horizon is 3 months, AR peaks at 0.707 with the number of trees equals 350, the maximum number of features (financial variables) equals 150. Cash balance and cash loan payment have the highest feature importance calculated in terms of the mean decrease gini,[6] consistent with their high single-model AR. Of 150 features, the first high-performing 84 features are variables related to cash balance and cash loan payment. For XGBoost, AR peaks at 0.733 when the time horizon equals 3 months, the tree's maximum depth equals four nodes, and the learning rate, a rate indicating how fast the model moves towards to minimum error, equals 0.01.

[5] The closer that AR is to one, the better the model. A random model's AR equals 0 and a perfect model's AR equals 1. See Annex 3 of "The Study for the Introduction of Credit Risk Database (CRD) in the Philippines: Data Quality Examination Report" for technical details.

[6] Feature importance helps indicate a variable's predictive power and which variables are the top features contributing to the prediction of the model. Feature importance is calculated based on the training model, not testing models, most commonly by the Gini coefficient. Refer to Hastie et al. (2014) for technical details.

Table 4 reports the model performance of the traditional logistic model and machine learning model. Logistic regression that employs some of the variables with the highest feature importance performs slightly inferior to machine learning models on out-of-sample data. The logistic model provides an AR of approximately 0.71 for both testing data and back test data, lower than XGBoost in both tests and lower than random forest in back testing.

4. CONCLUSIONS AND POLICY IMPLICATIONS

We introduced the Japan Credit Risk Database (CRD) as part of the Japanese government's SME financial inclusion efforts. To date, few databases exclusively for SMEs have achieved a nationwide coverage and the volume required for robust credit risk modeling such as the Japan CRD. The impact of CRD is threefold. First, CRD contributes to the overall understanding of the SME sector by providing general statistics for benchmarking, policy analysis, and evidence-based policy making (EBPM) evaluation. Second, to encourage the adaptation of risk-based lending, CRD statistical scoring models make SME credit risk assessment achievable and easily accessible. Third, SME credit guarantee becomes less biased by applying nine risk-based fees schemes, allowing SMEs to pay a guarantee fee according to their risk level.

The CRD framework is being introduced to other Asian countries with the support of Japan International Cooperation Agency (JICA). The JICA-sponsored feasibility study consulted by CRD Association confirms that interested parties in the Philippines have the capacity required to establish the Philippines version of CRD. The Technical Cooperation project begins between the two governments in the first half of 2020.

Traditionally, a CRD scoring model incorporates financial ratios derived from financial statements into a logistic model. The output of the model is a prediction of defaults within one to three years. While the traditional model functions well for decision making purposes, it lacks the necessary input for analysis in cases when SMEs do not have financial statements. To tackle this challenge, alternative data, specifically bank transaction data, is adopted to create a new model. Generally, bank transaction data is abundant in quantity and credible in quality. It is also relatively easy to retrieve, as the recording of transaction data is fully automated.

The transaction model introduced in the paper was jointly developed by CRD Association, BOJ and a Japanese megabank. The machine learning model is adopted to process the extremely large volume of data. The model outperforms logistic models and is highly accurate in predicting short-term default probability. It serves as an enabling short-term monitoring instrument and a credit assessment tool for SMEs without financial statements. Resona Bank, a Japanese megabank, has recently introduced a credit line that only requires SMEs to have bank accounts at the bank for a certain period for loan screening. The credit line does not require collateral, guarantee or the submission of financial statements.[7]

Approximately 6,000 variables are employed in the transaction model, applying two popular classification modelling approaches in machine learning: random forest and XGBoost. We find that that monthly average cash balance best explains default, followed

[7] The screening for the new credit line will be done by a machine learning model, analyzing the cash movements of the applicant's account. Financial cost ranges from 3% to 9%. The lending amount is capped at 10 million yen, approximately $100,000 (1USD=100Yen). News release of Resona Bank (in Japanese): https://www.resona-gr.co.jp/holdings/news/hd_c/download_c/files/20200110_1a.pdf?_ga=2.47077284.2056143362.1579136287-646537328.1579049610.

by cash inflows from revenue, cash outflows for variable costs, cash outflow for cost of goods and services, and cash inflows from loans.

Model accuracy measured by accuracy ratio peaks at 0.707 for a random forest with 350 decision trees and 150 features, and at 0.733 for XGBoost where the tree's maximum depth equals four nodes and learning rate equals 0.01. Both models reach accuracy levels general approved by the Japanese banking sector.

This paper indicates that SME financial inclusion efforts can incorporate both traditional and alternative data. Traditional data mainly assists lending decisions, while alternative data supports monitoring as well as credit assessment for SMEs without financial statements. Since the establishment of a nationwide database involves many interested parties and complex procedures, the leadership of the financial sector authority is an important factor that helps synchronize the efforts of all parties, and directs them towards a common goal.

REFERENCES

Amornkitvikai, Y., and C. Harvie. 2016. The Impact of Finance on the Performance of Thai Manufacturing Small and Medium-Sized Enterprises. ADBI Working Paper 576. Tokyo: Asian Development Bank Institute. Date of access: 2020/04/09. Available: http://www.adb.org/publications/impactfinance-performance-thai-manufacturing-small-and-medium-sized-enterprises/.

Banko Sentral ng Pilipinas. 2019. Financial Inclusion Initiatives. Manila. Date of access: 2020/04/09. Available: http://www.bsp.gov.ph/downloads/Publications/2019/microfinance_2019.pdf.

Deloitte Tohmatsu Consulting. 2018. Data utilization of SMEs and small businesses and the optimal information website. Japan SME Agency Report. Date of access: 2020/04/09. Available: https://www.meti.go.jp/meti_lib/report/H30FY/000316.pdf.

Hastie T., Tibshirani R. and Friedman J. 2014. The Elements of Statistical Learning. Springer. California.

Japan International Cooperation Agency (JICA). 2019. The Study for the Introduction of Credit Risk Database (CRD) in the Philippines: Final Report. Date of access: 2020/04/09. Available: http://open_jicareport.jica.go.jp/pdf/12344552.pdf.

Japan SME Agency, Ministry of Economy, Trade and Industry. 2019. White Paper on Small and Medium Enterprises in Japan. Date of access: 2020/04/09. Available: https://www.meti.go.jp/english/press/2019/pdf/0426_010a.pdf.

Japan Federation of Credit Guarantee Corporation, Annual Report 2019. Date of access: 2020/04/09. Available: https://www.zenshinhoren.or.jp/english/anual2019.pdf.

Kuwahara, S., N. Yoshino, M. Sagara, and F. Taghizadeh-Hesary. 2019. Establishment of the Credit Risk Database: Concrete Use to Evaluate the Creditworthiness of SMEs. ADBI Working Paper 924. Tokyo: Asian Development Bank Institute. Date of access: 2020/04/09. Available: https://www.adb.org/publications/establishment-credit-risk-database-evaluatecreditworthiness-smes.

Maningo, G. V. 2016. Credit Surety Fund: A Credit Innovation for Micro, Small, and Medium-Sized Enterprises in the Philippines. ADBI Working Paper 589. Tokyo: Asian Development Bank Institute. Date of access: 2020/04/09. Available: http://www.adb.org/publications/credit-surety-fundcredit-innovation-micro-small-and-medium-sized-enterprises-philippines/.

Miura S., Ijitsu Y., Takekawa M. 2019. Credit Risk Assessment using Transaction Information: An Empirical Study with AI Approach. Tokyo: Bank of Japan Working Paper Series. Date of access: 2020/04/09. Available: https://www.boj.or.jp/research/wps_rev/wps_2019/wp19j04.htm/.

World Bank. 2017. Global Findex Database. Date of access: 2020/04/09. Available: https://globalfindex.worldbank.org/basic-page-overview.