doi:10.20944/preprints202110.0006.v1

Can system logs enhance the performance of credit scoring? – Evidence from an internet bank in Korea

Sunghyon Kyeong[†], Daehee Kim[‡], Jinho Shin[*]

Abstract

This study is the first to examine whether the performance of credit rating, one of the most important data-based decision-making of banks, can be improved by using banking system log data that is extensively accumulated inside the bank for system operation. This study uses the log data recorded for the mobile app system of Kakaobank, a leading internet bank used by more than 14 million people in Korea. After generating candidate variables from Kakaobank's vast log data, we develop a credit scoring model by utilizing variables with high information values. Consequently, the discrimination power of the new model compared to the credit bureau grades was significantly improved by 1.84% points based on the Kolmogorov–Smirnov statistics. Therefore, the results of this study imply that if a bank utilizes its log data that have already been extensively accumulated inside the bank, decision-making systems, including credit scoring, can be efficiently improved at a low cost.

Keywords: Credit scoring, Banking, Data mining

[†] Division of Big-data Analytics, Kakaobank. Address: 5F H-Square S, 680 Sampyung-dong Sungnam-si, Kyeonggi-do, Korea. E-mail: devyn.k@kakaobank.com

[‡] Division of Big-data Analytics, Kakaobank. Address: 5F H-Square S, 680 Sampyung-dong Sungnam-si, Kyeonggi-do, Korea. E-mail: finch.harold@kakaobank.com

[*] Corresponding author. Division of Research and Development, Kakaobank. Address: 5F H-Square S, 680 Sampyung-dong Sungnam-si, Kyeonggi-do, Korea. Tel: +82-2-6420-3333. E-mail: william.shin@lab.kakaobank.com

doi:10.20944/preprints202110.0006.v1

1. Introduction

Recently, banks are making efforts to enhance their decision-making using various unstructured data inside and outside the bank. The term digital footprints, which has emerged recently, is also derived from this trend. Digital footprints are behavioral log data, such as the various actions of customers in mobile applications and websites. To understand the characteristics of their users, many companies try to analyze these data (e.g., Berg, Burg, Gombovic, and Puri, 2020).

Banks can enhance decision-making if they use digital footprints. According to Berg, Burg, Gombovic, and Puri (2020), performance improvement can be achieved using digital footprints in credit evaluation. Specifically, they confirmed that digital footprints, such as information related to hardware or operating system, e-mail, and purchase time, could improve the performance of the credit scoring system. Additionally, Lin, Prabhala, and Viswanathan (2013) found that online friendships of borrowers could be a signal of credit quality because online friendships are related to their funding ability. Netzer, Lemaire, and Herzenstein (2019) analyzed the contents of documents of a loan application using text mining and machine learning techniques and found that the textual information contributed to predicting loan defaults. Also, Óskarsdóttir, Bravo, Sarraute, Vanthienen, and Baesens (2019) show that combining call records with traditional data significantly increases their credit scoring performance when measured in the area under the receiver operating characteristic (AUROC). Taken together, various big data and digital footprints play an additional role in improving the performance of credit rating.

Meanwhile, banks may not easily access or use digital footprints because the data are not accumulated in well-refined forms as in Berg, Burg, Gombovic, and Puri (2020). Also, most companies do not collect such data, or even if the data are collected, it may be challenging to use if the data is not large enough. In this study, as an alternative to these shortcomings in using digital footprints, we examine the possibility of improving the bank's credit rating by using log data that are continuously recorded during the operation of the banking system. These log data contain all actions on the electronic device that customers consciously and unconsciously leave behind.

The log data is mainly used for monitoring system operation and detecting system anomalies by using various methods such as pattern recognition, normalization, classification, correlation analysis, and artificial ignorance (e.g., Farzad and Gulliver, 2020). With the recent data analysis technology, there are many attempts to use log data for business compliance, security, audit, regulation, and so on. However, although many efforts are made to utilize log data in practice, academic research has rarely been conducted. One of the reasons is that log data is basically records loaded for system operation, and thus, additional processing is necessary but not easy. In this study, we processed log data to extract customer’s digital footprints in a numerical form.

In this study, we use the digital footprints recorded in the system log of Kakaobank, which is a leading internet bank in Korea with an overwhelming market share of 14.17 million customers as of March 2021, with a deposit balance of KRW 25.39 trillion, and ranked first in Monthly Active Users among all financial institutions. Additionally, every online behavior of all customers in Kakaobank is recorded as log data because Kakaobank operates only a non-face-to-face mobile channel. Therefore, Kakaobank’s log data has a unique advantage because it captures customers' behaviors or preferences without distortion by a bank branch employee.

In short, although there have been many studies regarding the effects of social network data or digital footprints on credit scoring, there is no study regarding the effect of banking system log data on credit scoring. Considering that system log data are already massively loaded in Kakaobank, we aimed to use these log data to advance a credit scoring performance. In the present study, we hypothesized that the system log data might improve the performance of credit rating. To test this hypothesis, we developed two models. One is the baseline model, including only the credit grade provided by the Korea Credit Bureau. The other is the proposed model, including the independent variables driven from the system log data as well as the credit grade.

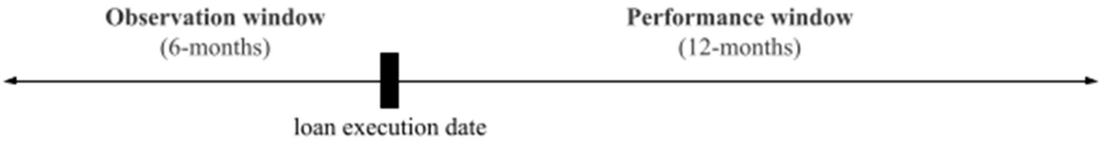
2. Analysis methodology

This study uses Kakaobank's log data to construct the credit scoring model. The log data contains all types of online activities, including customer actions and system operations. Specifically, if a customer wants to use a specific mobile banking service, he goes through the authentication process as a first step to access the banking app. Once he enters the home screen, he performs a specific action by touching the screen. For instance, to obtain an unsecured loan from Kakaobank, event logs are recorded during the entire process, from entering the screen to executing the loan. The system automatically acquires external credit information during the loan processes, and then related logs are accumulated even if there is no change displayed on the screen. All operations behind the mobile screen are also recorded as log data.[1]

2.1 Datasets

Our datasets are randomly sampled data from the personal unsecured loans of Kakaobank. We prepare training and test datasets for the development and (out-of-sample) validation of the credit scoring models. Each dataset consists of 100,000 randomly sampled cases from all unsecured loans newly booked during the third and fourth quarters of 2018. The *binary target variable* is defined as bad if the loan interest payment is overdue for more than 60 days within 12-months (performance window in **Fig. 1**) after the unsecured loan is initially booked as a good otherwise. We extract data related to all types of event logs for both datasets, including user actions and system operations. Then, we aggregate each event by users for the event logs recorded from the loan execution date to the past six months (observation window), as shown in **Fig. 1**. Finally, we construct a numeric tabular dataset that contains a user as a row and the number of actions for each event as a column. We note that the datasets used in this study received permission from Kakaobank.

Figure 1. Observation and performance window of the datasets



[1] In the case of Kakaobank's loan-related log data, the number of event codes for each process from loan inquiry to execution is 15.

2.2 Selection of candidate variables

The total number of unique events observed in both datasets is 347, including user actions such as login/logout, authentication, account open/close, touch menu, view account, touch notification tab, touch recommendation tab, and so on. To select the candidate variables as inputs of the credit scoring model, we compute the weights of evidence (WOE) and information value (IV) for each variable (Zeng, 2014). The WOE for each variable is defined as the logarithm of the proportion of "Goods" over the population of "Bads" indicating that high positive values refer to low risk, whereas high negative values refer to high risk. To select statistically significant variables, we compute IV as follows:

$$\mathrm{IV} = \sum_{\mathrm{i}} (\%\text{ of Goods} - \%\text{ of Bads}) \cdot \mathrm{WOE_i},$$

where WOE for *i*-th binning of each variable is defined as:

$$\mathrm{WOE_i} = \ln\left(\frac{\%\text{ of Goods}}{\%\text{ of Bads}}\right)_{\mathrm{i}}.$$

Among all events, we choose variables with IV ≥ 0.02, which means that the variable has at least a weak predictive power. We then have 66 candidate variables in a converted form of WOE as inputs for a credit scoring model after excluding the variables related to loan application events, such as confirmation of loan inquiry results, the number of rejected loans, etc.

The candidate variables can be categorized into ten types of actions: 1) registration category includes nine variables such as actions related to registration, account open, account closure, etc; 2) custom setting category consists of change of account name, account color, and registration of bookmark of a specific action; 3) menu/tab category includes eight variables like touch menu, touch home tab, touch guide tab, etc.; 4) authentication category has five variables regarding to various types of user authentications; 5) transaction category includes six variables such as touch transaction button, completion of transaction, sharing the transaction results, etc.; 6) account category includes nine variables such as balance view of my account, touch my account button, etc.; 7) card category consists of touch card application, select card type, and completion of card application process; 8) login category includes login, logout, and run application; 9) recommendation category has eleven variables such as touch pop-up, alarm, app-push, touch recommendation tab, etc.; 10) Optical Character Recognition (OCR) category has ten variables such as camera execution, taking photo of personal ID card, completion of taking photo, etc.

2.3 Logistic Regression

The logistic regression has been widely used in building a scoring model because Goods/Bads odds ratios in logistic regression are easy to calculate and interpret in a binary dependent variable. In this study, we develop two logistic regression models. First, we use the credit grade provided by the Korea Credit Bureau (KCB) as an input variable for the baseline model. According to the KCB, their credit grade is developed using all customer's financial transaction information. For example, it includes credit card transaction details, credit loan repayment details, loans and credit card holding information, account opening information, and delinquency history for all financial transactions. Therefore, the KCB grade has an excellent discrimination power. Second, we use 66 candidate variables and the CB credit grade as input variables to develop the log data-based model. Ten significant variables remain in the log data-based model by backward selection as a variable selection method.

2.4 Model performance evaluation

To compare model performances between the baseline and proposed models, we use the Kolmogorov-Smirnov (K-S) statistics and the AUROC. Briefly, the K-S statistics measure the maximum difference between the two cumulative distributions of Goods and Bads. The larger K-S statistics indicate the better performance of the credit scoring model (Chi and Hsu, 2012). The AUROC is also an important measure to evaluate the discriminatory power of a credit scoring model, which can be interpreted as the probability that the Goods receive better scores than the Bads (Chi and Hsu, 2012). To test whether the K-S statistics of the proposed model are greater than that of the baseline model, we conducted simulations according to the following steps: 1) We created two random samples from the training and test datasets with a size of n=20,000 from the whole datasets; 2) Estimated the model parameters for the baseline and proposed models using the sampled training dataset; 3) Predicted the model output for the sampled test dataset using the fitted parameters acquired in the second step; 4) Computed K-S statistics and AUROC for the baseline and proposed models, respectively; 5) Repeated steps 1-4 for 20 times and computed the mean of K-S statistics and AUROC; 6) Repeated the step 5 for 200 times to create the sampled mean distribution; 7) Finally, we conducted the independent sample t-test to compare the differences in the sampled mean distribution of K-S statistics and AUROC between the two models. Note that the sampled mean distribution follows the normal distribution according to the central limit theorem (Gendenko and Kolmogorov, 1954).

doi:10.20944/preprints202110.0006.v1

3. Empirical Analysis

3.1 Default ratio

The default ratios of the training and test datasets are 1.28% and 1.24% (**Table 1**). However, we note that these default ratios are not representative of credit borrows in the Kakaobank because we use a relatively small fraction of datasets sampled from a particular period.

Table 1. Descriptions of training and test datasets

	Training dataset	**Test dataset**
New book period	third quarter of 2018	fourth quarter of 2018
Samples	100,000	100,000
Bads	1,276	1,238
Bad ratio	1.28%	1.24%

3.2 Baseline and Log Data-based Credit Scoring Model

We develop a baseline model including only CB credit grades as an explanatory variable and a log data-based model including CB grades and derivative variables from the log data as explanatory variables. **Table 2** shows the model fit results. In the log data-based model, various distinct variables related to user online activity logs are included according to the backward selection approach, and these variables appear to be statistically significant. The variables related to user actions within the last six months from the loan execution date are registration, production cancellation, touch profile tab, product information inquiry, touch transfer request button, confirmation of personal identification, and typed customer information. All variables in the log data-based model appear to have positive signs, which means the variables lower default probability. Basically, the larger the log data-based variables, the more transactions or activities related to the variables are made by the customer.

Therefore, these results indicate that the more various transactions are made, the lower the customer's default probability.

Table 2. Model fit results of the baseline and log data models

Variables	Coefficient	Z-stat.	P-value
Baseline model			
Constant	-4.35	-136.51	<0.001
CB grade	1.00	31.39	<0.001
Log data model			
Constant	-4.35	-134.90	<0.001
CB grade	0.98	30.60	<0.001
Registration	0.94	4.40	<0.001
Production cancellation	0.85	4.12	<0.001
Touch profile tab	1.89	2.91	0.004
Product information inquiry	0.76	4.43	<0.001
Touch transfer request button	0.79	6.82	<0.001
Confirmation of personal identification	0.51	3.16	0.002
Typed customer information	1.25	3.66	<0.001

We compute the correlation coefficients among eight variables in the log data model, as reported in **Table 3**. The correlation between the variables and CB grade was found to be low. Additionally, the correlations among the data-derived log variables are found to be low overall. This means that the variables derived from log data have a unique explanatory power that CB grades cannot explain.

Table 3. Correlation among variables in the log data model

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CB grade (1)	1.00							
Registration (2)	0.02	1.00						

Product cancellation (3)	0.01	0.09	1.00					
Touch profile tab (4)	0.01	-0.01	-0.17	1.00				
Product information inquiry (5)	0.02	-0.12	0.08	-0.09	1.00			
Touch transfer request button (6)	0.05	0.03	0.02	-0.04	0.13	1.00		
Confirmation of personality identification (7)	0.07	-0.11	-0.02	0.00	0.36	0.07	1.00	
Typed customer information (8)	-0.02	0.14	0.13	-0.04	-0.18	0.03	-0.15	1.00

3.3 Model performance

The empirical results in **Table 4** show that the K-S and AUROC values of the proposed model are 42.26% and 76.81%, respectively. We introduce the log data reflecting the various user activities to improve the credit scoring system, and then, we conduct t-test to verify whether this improvement is statistically significant. Consequently, the proposed model showed the significantly higher the K-S ($P < 0.0001$) and AUROC statistics ($P < 0.0001$) compared with that of the baseline model.

As explained in the previous section, it would not be easy to improve the credit scoring performance more than the external CB grade using whole financial transaction data in Korea. Nonetheless, we found that the improvement using the log data is meaningful regardless of its degree because it does not cost anything to use the log data already loaded inside the bank. Since Korean CB companies like KCB receive all transaction information from all financial institutions located in Korea, the performance of the CB grades using their data is inevitably excellent. However, in countries where all financial transaction data are not concentrated in CB companies, it would be effective to improve credit scoring performance by using the log data loaded inside the bank.

Table 4. The performance of the credit scoring models

	Baseline model	Proposed model	T-test results	
			T-statistics	P-values
K-S statistics	40.42 (±0.52) %	42.26 (±0.52) %	35.27	< 0.0001
AUROC	76.39 (±0.28) %	76.81 (±0.28) %	15.05	< 0.0001

* The numbers in the parenthesis represent the standard deviation.

4. Concluding Remarks

Our study is the first empirical analysis to examine whether the banking system log data improve the performance of a credit scoring model. We developed two models for the empirical analysis. The baseline model includes only the credit grade provided by the Korea Credit Bureau, whereas the proposed model includes the variables driven from the log data and the credit grade. Compared with the baseline model, the proposed model showed significant improvements of credit scoring performance such as K-S statistics by 1.84 % points and AUROC by 0.42% points, respectively. Therefore, if banks use the system log data in credit scoring, they can cost-effectively advance decision-making if the log data is appropriately processed.

Previous literature regarding credit rating improvements mainly focuses on the effects of big data or digital footprints that intuitively contain customer behavior characteristics, such as social network services (e.g., Lin, Prabhala, and Viswanathan, 2013). However, no studies are analyzing the effectiveness of credit rating improvement using system log data, which is a large amount of unstructured data that has already been accumulated inside the bank.

Although log data is accumulated on a very large scale in banks, it is difficult for data scientists to analyze it because it is recorded for system operation purposes, not for analysis. However, log data has several distinct advantages. First, the log data is large in scale and at meager utilization costs. Second, the log data captures the potential customers' behaviors directly because their actions in the banking applications are recorded with no intervention by bank clerks. By taking these advantages into account, the improvement of discrimination against CB's credit grades would be more significant in countries where information concentrated on CB companies is limited. As shown in this study, if massive log data is processed appropriately and actively used for banking business, it will be possible to efficiently improve not only credit scoring but also overall data-based decision making at a meager cost.

doi:10.20944/preprints202110.0006.v1

References

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *Review of Financial Studies, 33(7), 2845–2897*. https://doi.org/10.1093/rfs/hhz099.

Chi, B. & Hsu, C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications, 39, 2650-2661*. https://doi.org/10.1016/j.eswa.2011.08.120.

Chun, J., Ahn, J., Kim, Y., & Lee, S. (2020). Using Deep Learning to Develop a Stock Price Prediction Model Based on Individual Investor Emotions. *Journal of Behavioral Finance Published online: 18 Sep 2020*. https://doi.org/10.1080/15427560.2020.1821686.

Farzad, A. & Gulliver, T. A. (2020). Unsupervised log message anomaly detection. *ICT Express, 6(3), 229-237*. https://doi.org/10.1016/j.icte.2020.06.003.

Gendenko, B.V., Kolmogorov, A.N. (1954). Limit distributions for sums of independent random variables (*Addison-Wesley*).

Hiemstra, D. (2000). A probabilistic justification for using tf×idf term weighting in information retrieval. *International Journal on Digital Libraries, 3(2), 131–139*. https://doi.org/10.1007/s007999900025.

Lin, M., Prabhala, N., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59(1), 17–35*. https://doi.org/10.1287/mnsc.1120.1560.

María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, and Bart Baesens. (2019). The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics, *Applied Soft Computing*, 74, January 2019, 26-39.

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *Journal of Marketing Research, 56(6), 960-980*. https://doi.org/10.1177/0022243719852959.

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., Seliya, N., & Wald, R. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, Volume, 2, Article 1*. https://doi.org/10.1186/s40537-014-0007-7.

Zeng, G. (2014). A Necessary Condition for a Good Binning Algorithm in Credit Scoring. *Applied Mathematical Sciences, 8(65), 3229-3242*. https://doi.org/10.12988/ams.2014.44300.