



Article

Can System Log Data Enhance the Performance of Credit Scoring?—Evidence from an Internet Bank in Korea

Sunghyon Kyeong ¹, Daehee Kim ¹ and Jinho Shin ^{2,*}

¹ Division of Big-Data Analytics, KakaoBank, Seongnam-si 13494, Korea; devyn.k@kakaobank.com (S.K.); finch.harold@kakaobank.com (D.K.)

² Division of Research and Development, KakaoBank, Seongnam-si 13494, Korea

* Correspondence: william.shin@lab.kakaobank.com; Tel.: +82-2-6420-3333

Abstract: The credit scoring model is one of the most important decision-making tools for the sustainability of banking systems. This study is the first to examine whether it can be improved by using system log data that are stored extensively for system operation. We used the log data recorded by the mobile application system of KakaoBank, a leading internet bank used by more than 14 million people in Korea. After generating candidate variables from KakaoBank's log data, we created a credit scoring model by utilizing variables with high information values and logistic regression, the most common method for developing credit scoring models in financial institutions. To prove our hypothesis on the improvement of credit scoring model performance, we performed an independent sample *t*-test using the simulation results of repeated model development and performance measurement based on randomly sampled data. Consequently, the discrimination power of the proposed model using logistic regression (neural network) compared to the credit bureau-based model significantly improved by 1.84 (2.22) percentage points based on the Kolmogorov–Smirnov statistics. The results of this study suggest that a bank can utilize the accumulated log data inside the bank to improve decision-making systems, including credit scoring, at a low cost.

Keywords: credit scoring model; system log data; logistic regression; data mining; machine learning; fintech



Citation: Kyeong, S.; Kim, D.; Shin, J. Can System Log Data Enhance the Performance of Credit Scoring?—Evidence from an Internet Bank in Korea. *Sustainability* **2022**, *14*, 130. <https://doi.org/10.3390/su14010130>

Academic Editors: Ariful Hoque and Thi Le

Received: 12 November 2021

Accepted: 21 December 2021

Published: 23 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As unsecured loans are the main business of commercial banks, it is important to measure the default risk of loan applicants [1,2]. Recently, banks have tried to enhance their decision-making using emerging techniques such as machine learning [3–6], and various unstructured data inside and outside the bank [7–11]. The term digital footprint, which has emerged recently, is derived from this trend. Digital footprints are behavioral log data, such as various customer actions on mobile applications and websites. Many companies try to analyze these data to understand the characteristics of their users [12].

Banks can enhance decision-making by using digital footprints. According to Berg et al. [12], a 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 systems, e-mail, and purchase time, can improve the performance of the credit scoring system. Additionally, Lin et al. [13] found that borrowers' online friendships could be a signal of credit quality because online friendships are related to their funding ability. Netzer et al. [14] analyzed the contents of documents of a loan application using text mining and machine learning techniques and found that textual information contributed to predicting loan defaults. Óskarsdóttir et al. [15] showed that combining call records with traditional data significantly improves credit scoring performance when the area under the receiver operating characteristic (AUROC) is measured. To summarize, various big data and digital footprints have roles to play in improving the performance of credit scoring.

Meanwhile, banks may not easily access or use digital footprints because the data do not accumulate in well-refined forms [12]. Additionally, most companies do not collect such data, or even if the data are collected, it may be challenging to use data that are not large enough [16]. In this study, as an alternative to these shortcomings, we examine the possibility of improving banks' credit scoring by using log data that are continuously recorded during the operation of the banking system. These log data contain all actions that customers consciously and unconsciously leave behind on electronic devices [17].

The novelty of this study is that it enriches the research area on the relationship between improving credit scoring performance and utilizing system log data. In this study, we verify whether the internal log data of a bank can enhance a credit scoring model. To the best of our knowledge, no study utilizes system logs or demonstrates their usefulness for credit scoring. For example, Berg et al. [12] found the discriminatory power of digital footprints in a German e-commerce company by using ten easily accessible variables such as the device type, operating system, and access channel. Contrary to the data in Berg et al. [12], the system log data used in this study are complex, large, and difficult to access because they have been recorded in an unstructured way for system operation. Therefore, it is practically difficult for researchers and data analytics to use system logs to improve credit scoring. From this point of view, the system log data is different from the digital footprints or big data analyzed in previous studies.

The log data are mainly used for monitoring system operation and detecting system anomalies related to rare events, system failure, or conditions different from normal system operation, using various methods such as pattern recognition, normalization, classification, correlation analysis, and artificial ignorance [18]. There have been many attempts to use log data for business compliance with the rise of recent data analysis technologies, including security, audit, and regulation. However, although many efforts have been made to utilize log data in practice, academic research is scarce. As explained earlier, one of the reasons is that log data are basically records loaded for system operation, and thus, additional processing is necessary but not easy. This study employs a simple method to deal with log data to extract customers' digital footprints in a numerical form.

For empirical analysis, we use the system log data recorded in KakaoBank, a leading Internet bank in Korea with an overwhelming market share of 14.17 million active customers as of March 2021, a deposit balance of USD 21.57 billion, and the highest monthly active users among all financial institutions. Interestingly, each online behavior of all customers in KakaoBank is recorded as log data because KakaoBank operates only as 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.

Our study differs from previous literature as it statistically verifies whether the proposed model obtained from the empirical analysis has been improved. Most published studies argue that they improve their proposed model based on performance indices obtained by developing only one model. The conclusions drawn this way would require further robustness checks. However, our study finds the distribution of performance measures based on iteratively developing models through simulation, followed by tests to check whether the model improvement is statistically significant. This process suggests a new research method for improving credit scoring.

In short, although there have been many studies regarding the effects of social network data or digital footprints on credit scoring, there are no studies on the effect of banking system log data on credit scoring. Considering that system log data are already massively loaded in KakaoBank, we use these log data to advance the credit scoring performance in this study. In the following section, we present the hypotheses and methodology. Section 3 builds a credit scoring model using the variables derived from the system log data and measures the improvement in model performance. In Section 4, we discuss the results and their implications. Finally, Section 5 concludes the study.

2. Materials and Methods

In the present study, we hypothesize that system log data might improve credit scoring performance and develop two models to test this hypothesis. One is the baseline model, which only includes the credit grade provided by the Korea Credit Bureau (KCB). The other is the log data-based model, which includes the independent variables derived from the system log data and the KCB credit grade. Our test methodology is consistent with many previous studies in verifying that additional information contributes to the performance improvement of the credit scoring model. A baseline model and an alternative model are developed, and the AUROC or Kolmogorov-Smirnov (K-S) statistics between the two models are compared [10,12,15].

Figure 1 shows the steps of our study. First, we prepare the datasets by randomly sampling KakaoBank's unsecured loans. Second, we select features with high predictive power for the default of KakaoBank's loan. Third, we develop the baseline model using the KCB grades alone and the proposed model using the system log data. Finally, by repeating these three steps to obtain the distribution of the model performance measures, we test whether the performance improvement of credit scoring is statistically significant.

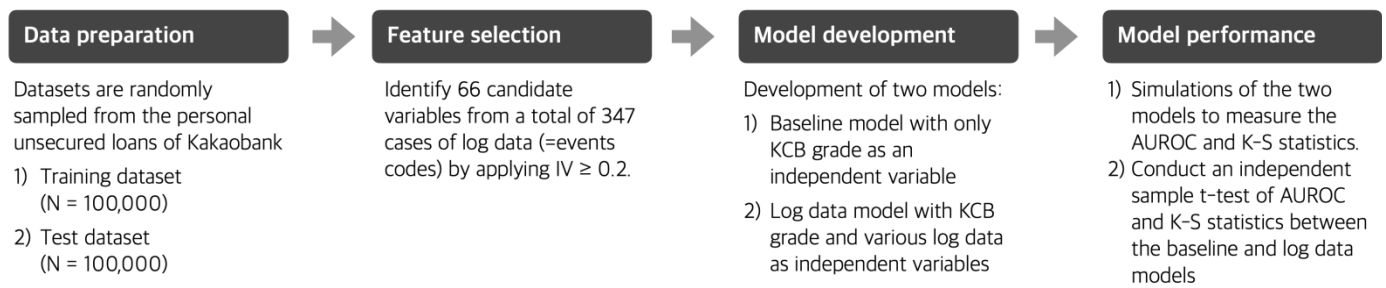


Figure 1. Schematic diagram of analysis methods and description for each analysis step.

This study uses KakaoBank's log data to construct an alternative 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, they first go through the authentication process to access the banking application. Once they enter the home screen, they perform a specific action by touching the screen. For instance, when obtaining 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. Additionally, all operations behind the mobile screen are recorded as log data. In KakaoBank's loan-related log data, the number of event codes for each process from loan inquiry to execution is 15.

2.1. Datasets

Our datasets were randomly sampled from the personal unsecured loans of KakaoBank. We prepared training and test datasets for the development and out-of-sample validation of the credit scoring models. Each dataset consisted of 100,000 randomly sampled cases from all unsecured loans booked during the third and fourth quarters of 2018. To define our target variable, a performance window of at least 12 months after the reference time point is required. Therefore, we set the reference time point to 12 months before the start of the study. The binary target variable was considered bad if the loan interest payment was overdue for more than 60 days within 12 months (performance window in Figure 2) of the unsecured loan booking, and was otherwise considered good.



Figure 2. Observation and performance window of the datasets.

As for the independent variables, we used the raw event log data accumulated for every user action. The total number of unique event codes observed in the raw event log data was 347. More specifically, we obtained event logs during user actions such as the registration processes, custom setting of a banking application, menu or tab clicks, user authentication, transaction, management of account, selection of card types and options, logging in and out, responses to the recommendation, and optical character recognition (OCR) processes. We counted event logs for each event code for both datasets, including user actions and system operations. Then, we aggregated each event by users for the event logs recorded from the loan execution date to the past six months (observation window), as shown in Figure 2. Finally, we constructed a numerical tabular dataset that contained a user as a row and the number of actions for each event as a column.

Additionally, we used the credit grade provided by the KCB as an independent variable. According to the KCB, credit grade was calculated by fully utilizing all customers' financial transaction information. For example, it included 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 excellent discrimination power. We received appropriate permission from Kakaobank to use the datasets in this study.

2.2. Selection of Candidate Variables

To select candidate variables as inputs to the credit scoring model, we compute the weights of evidence (WOE) and information value (IV) for each variable [19]. 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 default risk. However, a high negative value refers to high default risk. To select statistically significant variables, we computed IV as follows:

$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \cdot WOE_i, \quad (1)$$

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

$$WOE_i = \ln \left(\frac{\% \text{ of Goods}}{\% \text{ of Bads}} \right)_i \quad (2)$$

As the difference between the proportions of Goods and Bads for each bin becomes similar, IV approaches zero. However, the larger the distribution difference, the higher is the IV. The higher the variable's discriminative power, the greater the number of bins with a large difference between the proportions of Goods and Bads; thus, IV increases. Note that the sign of $(\% \text{ of Goods} - \% \text{ of Bads})$ follows the WOE, and thus, IV is always positive.

Credit scoring textbooks provided the following IV rules of thumb to evaluate the predictive power of an explanatory variable in logistic regression for practical purposes: less than 0.02, uninformative; 0.02 to 0.1, weak predictor; 0.1 to 0.3, medium predictor; greater than 0.3, strong predictor [20,21]. We choose variables with $IV \geq 0.02$, which means that the variable has weak predictive power at the very least. We then have 66 candidate variables in a converted form of WOE as inputs for a credit scoring model after excluding the variables directly related to loan application events, such as confirmation of loan inquiry results and the number of rejected loans.

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

2.3. Logistic Regression

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. This study uses 66 candidate variables and the KCB credit grade as input variables to develop a log data-based model. Ten significant variables remained in the log data-based model with backward selection as a variable selection method.

2.4. Model Performance Evaluation

To compare model performances between the log data-based model and the baseline model that uses only the KCB credit grade as an input variable, we use the K-S statistics and the AUROC. The AUROC and K-S statistics is a very widely used performance measure of the models for classification problems. In studies on the credit scoring model which is one of the classification models, it is desirable to use the AUROC and K-S statistics as performance indicators of the model [20–22]. Briefly, the K-S statistic measures the maximum difference between the two cumulative distributions of Goods and Bads. The larger K-S statistics indicate a better performance of the credit scoring model [20–22]. Additionally, the AUROC is an important measure for evaluating the discriminatory power of a credit scoring model, which can be interpreted as the probability that the Goods receive better scores than the Bads [20–22]. To test whether the K-S statistics and the AUROC of the log data-based model are greater than those of the baseline model, we conducted simulations according to the following steps: (1) created two random samples from the training and test datasets with a size of $n = 20,000$ from the entire dataset; (2) estimated the model parameters for the baseline and log data 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 log data models, respectively; (5) repeated steps 1–4 20 times and computed the mean of K-S statistics and the AUROC; (6) repeated step 5 200 times to create the sampled mean distribution; (7) and finally, conducted an independent sample *t*-test to compare the differences in the sampled mean distribution of K-S statistics and the AUROC between the two models. Notably, according to the central limit theorem [23], the sampled mean distribution follows a normal distribution.

3. Results

3.1. Default Ratio

The default ratios of the training and test datasets were 1.28% and 1.24%, respectively (Table 1). However, we note that these default ratios do not represent credit borrowings in KakaoBank because we use a relatively small fraction of datasets sampled from a particular period. Previous literature recommends that the number of bad is about 1500 for building a very high-quality and robust credit scoring model and at least 500 observations for a model with good predictive power [20,21]. Therefore, the sample size in this study would

be appropriate because the numbers of bads in the training and test datasets were 1276 and 1238, respectively.

Table 1. Descriptions of training and test datasets.

| | Training Dataset | Test Dataset |
|-----------------|---------------------------|----------------------------|
| New book period | the third quarter of 2018 | the fourth quarter of 2018 |
| Samples | 100,000 | 100,000 |
| Number of Bads | 1276 | 1238 |
| Default ratio | 1.28% | 1.24% |

3.2. Baseline and Log Data-based Credit Scoring Model

We develop a baseline model that includes only the KCB credit grades as explanatory variables. The log data-based model includes derivative variables from the log data and KCB grades as explanatory variables. Table 2 shows the model fit results. According to the backward selection approach, the log data-based model includes various distinct variables related to users' online activity logs, 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. Although all variables in the log data-based model appear to have positive signs due to the WOE transformation, each variable lowers the 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 imply that the more transactions a customer makes, the less likely they are to default.

Table 2. Statistical properties and model fit results of the baseline and log data models.

| Variables | Basic Statistical Properties | | | Model Fit Results | | |
|-------------------------------|------------------------------|----------|----------|-------------------|---------|---------|
| | Mean (SD) | Skewness | Kurtosis | Coefficient | Z-Value | p-Value |
| Baseline model | | | | | | |
| Constant | - | - | - | -4.35 | -136.51 | <0.001 |
| KCB grade | -0.50 (1.02) | 0.24 | -1.09 | 1.00 | 31.39 | <0.001 |
| Log data model | | | | | | |
| Constant | - | - | - | -4.35 | -134.90 | <0.001 |
| KCB grade | -0.50 (1.02) | 0.24 | -1.09 | 0.98 | 30.60 | <0.001 |
| Registration | -0.01 (0.12) | 2.95 | 6.71 | 0.94 | 4.40 | <0.001 |
| Production cancellation | -0.01 (0.17) | -1.99 | 2.09 | 0.85 | 4.12 | <0.001 |
| Touch profile tab | 0.00 (0.04) | 2.08 | 2.33 | 1.89 | 2.91 | 0.004 |
| Product information inquiry | -0.01 (0.17) | 0.91 | -0.69 | 0.76 | 4.43 | <0.001 |
| Touch transfer request button | -0.03 (0.25) | 0.16 | -1.11 | 0.79 | 6.82 | <0.001 |
| Confirmation of personal ID | -0.02 (0.17) | 1.35 | 1.38 | 0.51 | 3.16 | 0.002 |
| Typed customer information | 0.00 (0.08) | 1.92 | 1.69 | 1.25 | 3.66 | <0.001 |

Abbreviations: ID, identification; KCB, Korean Credit Bureau; SD, standard deviation.

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

Table 3. Correlation among variables in the log data model.

| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|--|-------|-------|-------|-------|-------|------|-------|------|
| KCB 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 |

To test the goodness of fit, we conducted the Hosmer-Lemeshow test for the log data-based model. We obtained a very small Chi-squared statistic and a very large p -value (Chi-squared statistic = 0.01; p -value = 0.995), indicating that our log data model fits well [24]. We also conducted the Hosmer-Lemeshow test for the test dataset and obtained similar values (chi-squared statistic = 0.05; p -value = 0.953), indicating that our log data-based model was well fitted for the training and test datasets. Additionally, pseudo R-squared and log-likelihood values are 0.086 and −6251 (0.097 and −6178) for the baseline model (the log data model).

As the number of system log records increases with the increase in customers' activities on bank applications, it is reasonable to think that the multicollinearity between log data would be high. Therefore, it is necessary to carefully examine multicollinearity in the model. To address this issue, we first check the correlation coefficients between the variables in Table 3. The highest correlation coefficient is 0.36 between product information inquiry (5) and confirmation of personality identification (7), and the second-highest correlation is 0.18 between product information inquiry (5) and typed customer information (8), so the correlation coefficients between variables are overall quite low.

Furthermore, we obtained the condition number to statistically test multicollinearity. The condition number measures the effect of the change in the explanatory variable on the regression coefficient. If there is no multicollinearity, it becomes zero, and the condition number increases as the multicollinearity increases. A rule of thumb for the decision of multicollinearity is that if the condition number is approximately 15, multicollinearity can be a concern; if it is greater than 30, multicollinearity is very serious [25]. In this study, the condition number of our log data model appears to be 1.62, which means little multicollinearity in the model.

3.3. Model Performance

The empirical results in Table 4 show that the K-S statistics and the AUROC values of the log data 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 conduct an independent sample t -test to verify whether this improvement is statistically significant. Consequently, the log data model showed significantly higher K-S statistics (p -value < 0.0001) and AUROC values (p -value < 0.0001) compared with those of the baseline model.

Table 4. The performance of the credit scoring models.

| | Baseline Model | Log Data 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 |

Numbers in parentheses represent standard deviations.

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

3.4. Robustness of the Results

First, we test whether our results are affected by specific modeling methodologies. We employed logistic regression, the most widely used method for building a credit scoring model. To check the robustness of our results, we further employed machine learning algorithms such as random forest and neural networks, which generally perform well when applied to build a credit scoring model [3–6]. Table 5 reports that both random forest and neural networks provide better performance than the log data model using logistic regression as well as the baseline model. The improvements of K-S statistics and the AUROC values are 0.60 and 2.08 percentage points (%p) for the random forest model, and 1.18%p and 2.22%p for the neural network model with five hidden layers, respectively. Therefore, all of the modeling techniques enhance the performance of credit scoring models using system log data, and the neural network model yields the best results.

Table 5. Results of the robustness test.

| | | | Baseline Model | Log Data Model |
|---------------------------------|----------------|----------------|----------------|----------------|
| Alternative Modeling Techniques | Random Forest | K-S statistics | 39.76% | 41.84% |
| | | AUROC | 76.01% | 76.61% |
| | Neural Network | K-S statistics | 39.75% | 41.97% |
| | | AUROC | 76.39% | 77.57% |
| Larger sample size | | K-S statistics | 39.40% | 40.31% |
| | | AUROC | 75.81% | 76.67% |

Second, we checked whether our results are affected by sample size, which should be large enough to develop a robust and high-quality model [20,21]. After combining all the existing training and test datasets, we performed random sampling again to prepare a larger training dataset. Consequently, the numbers of training and test datasets for the robustness test are 150,000 and 50,000, respectively. When a larger training dataset is used, Table 5 reports that the improvements in K-S statistics and the AUROC values are 0.86%p and 0.91%p respectively, and thus, the performance of the models using a larger sample appears to be very similar to those of the proposed model in Table 4.

4. Discussion

Our study contributes to scientific research by being the first empirical analysis to examine whether banking system log data improve the performance of a credit scoring model. In addition, this study proposes a new method to statistically verify the performance improvement of the credit scoring model using iterative simulations. Although many previous studies use various big data, including digital footprints for credit scoring [12], to the best of our knowledge there is no study that focuses on system log data, which is recorded every moment whenever banking systems are operated by customers. This study found that credit scoring performance was improved by utilizing the variables generated by simply counting each event code of log data.

4.1. Performance of the Log Data-Based Model

We developed two models for empirical analysis. The baseline model includes only the credit grade provided by the Korea Credit Bureau, whereas the log data-based model includes variables from the log data and credit grade. Compared to the baseline model, the log data-based model using logistic regression showed significant improvements in credit scoring performance, with the K-S statistics improving by 1.84%p, and the AUROC improving by 0.42%p. Additionally, we tested the feasibility of log data to improve the credit scoring model. For example, the random forest (neural network) model with log data and credit bureau (CB) grade as inputs showed an improvement in K-S statistics and the AUROC values of 2.08%p and 0.60%p (2.22%p and 1.18%p) respectively, compared to the random forest (neural network) model with CB data used only as input.

The model improvement in this study is similar to those of previous studies on credit scoring enhancement using big data. First, our baseline model using the KCB information showed better performance (AUROC of 76.39%) compared to credit scoring models using German CB information (AUROC of 66.5~68.3%) and US CB information (AUROC of 59.8~62.5%) [12]. The credit transaction details of all financial institutions may be concentrated in CB companies in Korea. Berg et al. [12] explained that a German credit bureau provided more discriminatory scores because of using richer information than the credit bureau in the U.S. under more strict regulations. In this study, the Korea Credit Bureau score showed higher performance than those in Germany and U.S. with an AUROC of 76.39%, because credit transaction details of all financial institutions are concentrated in the credit bureaus according to the Act on the Use and Protection of Credit Information in Korea. Second, the credit scoring model using both CB information and log data increased the AUROC by 0.4%p (5.3%p) compared to using CB information alone according to our model (other models described elsewhere [12]). In addition, by adding call-detail records, the credit scoring model showed that the AUROC increased by 2.3%p compared to the baseline model [15]. Taken together, the addition of log data might advance the performance of the credit scoring model.

Considering that the discriminating power of the KCB score used as the baseline model was much higher than those of the U.S. and Germany, the improvement of the credit scoring model obtained in our study would be an important achievement. If banks use the system log data in credit scoring, they can cost-effectively enhance decision-making regarding credit risk if the log data are appropriately processed.

4.2. Social Benefits

The improved credit scoring model can be beneficial to customers in two ways. First, the results of this study can be beneficial for active customers having good relationships with banks, even if the amount of financial transactions such as deposits is small. Considering that various relationships, such as non-credit relationships, and the depth and intensity of relationships can reduce default rates [26], the relationships between banks and customers can be considered an essential factor for the credit scoring model. However, it is difficult to obtain information on these relationships. For example, whether a customer thinks it is the main bank, how often they use the banking application and the money transfer relationship with other customers who use the bank. Meanwhile, system log data that records all activities performed by a customer while using a banking application can provide a wide variety of behavioral information about how a customer uses the banking system. Compared to credit scoring evaluated using only CB information, banks provide a more accurate evaluation on customer's creditworthiness through additional system log data capturing the customer's relationships with the bank or behaviors such as how they interact with the banking system. Consequently, it is possible to provide more financial benefits such as credit line increases or interest rate discounts for customers who make many transactions with their bank by evaluating their creditworthiness in a better manner based on system logs [15].

Second, the improved credit scoring model using log data would lower financial inequality and make credit accessibility easier for people with low credit creditworthiness or thin credit histories. Surprisingly, 1.7 billion adults worldwide (31% of adults) do not have any basic transaction account, as reported by the World Bank [27]. However, if commercial banks utilize alternative data such as digital footprints, they can facilitate credit and other services to applicants with thin credit profiles, such as young people or foreigners, thus lowering financial inequality [12,15]. Similarly, even young people with no credit history could be evaluated more superiorly using system logs recorded during non-credit transactions, such as deposits and money transfers. Consequently, it may be easier to grant new unsecured loans to these people. These effects work more effectively in countries where information sharing is difficult, such as in countries without CB. This is because banks have no choice but to evaluate their customers from a conservative point of view as it is difficult to evaluate the creditworthiness of customers, and fraudulent loans frequently occur when relevant information is insufficient [28,29]. In the case of countries where information sharing is difficult, if the system log data that is already widely stored inside the financial institution is utilized, customers can be evaluated more accurately based on the log data, even in countries without credit bureaus.

4.3. Advantages and Disadvantages of System Log Data

To the best of our knowledge, this study is the first to apply system log data recorded inside an Internet bank to a credit scoring model, which has not been paid attention to by previous studies, especially in the context of utilizing various big data. Previous literature on credit scoring improvements mainly focuses on the effects of big data or digital footprints that intuitively contain customer behavior characteristics, such as social network services [13]. However, no studies have analyzed the effectiveness of credit scoring improvement using system log data, which is a large amount of unstructured data that has already been accumulated inside the bank.

Although enormous amounts of log data are accumulated in banks, it is difficult for data scientists to analyze because it is recorded for system operation purposes, not for analysis. However, log data have several distinct advantages. First, they are extensive in scale because they are recorded the entire time customers touch the banking application. Thus, there is a high possibility of creating discriminatory variables with various data mining techniques. Second, the log data captures the potential customers' behaviors directly because their actions while using banking applications are recorded without any intervention by bank clerks. Third, there is no cost to pay for using the system logs because customer log data are already stored in the bank. Generally, it is necessary to pay a fee for data acquisition to obtain alternative data from an external data source. For example, in Korea, if a bank wants to use mobile call information for credit scoring, the bank must pay the telecommunication company on a per-inquiry basis to obtain information about the customers' mobile calls, such as average call time, monthly communication fees, and communication fee overdue history. However, if the bank has the means to extract system log data containing the default-related behavioral characteristics of customers, there is no need to spend any money on utilizing the log data. Therefore, the results of this study imply that using log data is a cost-effective way to improve credit scoring performance.

5. Conclusions

Previously, studies handling log data focused on finding a way to monitor system operations and identify operational anomalies because of the primary purpose of storing log data. As a result, they found notable analysis methods including pattern recognition [18]. Additionally, other research related to improving credit scoring using alternative data have utilized well-structured data with well-explained data descriptions [12]. This study fills the gap between these two research areas by investigating whether credit scoring improvement is possible by applying the system log data to credit scoring for the first time. Given that the improved credit scoring model can contribute to advancing the soundness of unsecured

loans by lowering the default rate [2], improving credit scoring performance may be a necessary condition for the sustainability of financial institutions. This is why our research results are important. Consequently, the increased bank profits due to the enhanced credit quality can be returned to customers as benefits such as interest rate discounts or credit line increases [15].

In this study, we confirmed that system log data could contribute to the performance improvement of the credit scoring model. Since KakaoBank does not have a branch channel and only operates a mobile application channel, the log data of KakaoBank can capture the behavioral characteristics of customers without interference from branch employees.

The performance of our log data model using the system logs was found to be a K-S statistic of 42.26% and an AUROC of 76.81%. It was improved by 1.84%p and 0.42%p, respectively, compared to the baseline model using the credit bureau grade alone, and this improvement was statistically significant. To confirm the robustness of the results, we employed machine learning techniques that had recently attracted attention for their high discriminatory power. As a result, there was a further increase in credit scoring improvement. Specifically, compared to the baseline models, the K-S statistics and the AUROC were improved by 2.08%p and 0.60%p for the random forest model, and 2.22%p and 1.18%p for the neural network model, respectively. Considering that the Korea Credit Bureau score performance is much higher than that of the US or Germany, it may not be easy to improve the performance of the credit scoring model using alternative data. Therefore, this performance improvement obtained using the log data already stored inside the bank has important implications for policymakers of commercial banks.

This study had several limitations. First, only the process of counting each log event code was used. Whether credit scoring performance can be improved by using various data mining techniques can be explored in future research. Second, only logistic regression and two machine learning techniques were used to develop the credit models in this study. However, other machine learning algorithms may be more suitable for utilizing unstructured big data [30]. Therefore, it may be necessary to apply various machine learning algorithms and logistic regression to investigate the contribution of log data to credit scoring.

This study shows that if massive log data are processed appropriately and used actively for the banking business, it will be possible to efficiently improve credit scoring and overall data-based decision-making at a small cost. These results imply that policymakers of financial institutions should pay attention to utilizing system log data and the latest technology or big data to improve credit scoring performance.

Author Contributions: Conceptualization, S.K. and J.S.; methodology, S.K. and J.S.; software, S.K. and D.K.; validation, S.K. and D.K.; formal analysis, S.K.; investigation, J.S.; resources, S.K. and J.S.; data curation, S.K. and D.K.; writing—original draft preparation, S.K. and J.S.; writing—review and editing, S.K. and J.S.; visualization, S.K.; supervision, J.S.; project administration, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We would like to thank the reviewers and editors for the recommendations made to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khashei, M.; Mirahmadi, A. A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification. *Int. J. Financ. Stud.* **2015**, *3*, 411–422. [CrossRef]
2. Rasa, R. The Effects of Credit Risk on the Profitability of Commercial Banks in Afghanistan. *J. Asian Financ. Econ. Bus.* **2021**, *8*, 477–489.
3. Munkhdalai, L.; Munkhdalai, T.; Namsrai, O.-E.; Lee, J.Y.; Ryu, K.H. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. *Sustainability* **2019**, *11*, 699. [CrossRef]
4. Niu, B.; Ren, J.; Li, X. Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. *Information* **2019**, *10*, 397. [CrossRef]
5. Dumitrescu, E.; Hué, S.; Hurlin, C.; Tokpavi, S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *Eur. J. Oper. Res.* **2022**, *297*, 1178–1192. [CrossRef]
6. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable Machine Learning in Credit Risk Management. *Comput. Econ.* **2021**, *57*, 203–216. [CrossRef]
7. Shema, A. Effective credit scoring using limited mobile phone data. In Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, Ahmedabad, India, 4–7 January 2019.
8. Luo, Z.; Hsu, P.; Xu, N. SME Default Prediction Framework with the Effective Use of External Public Credit Data. *Sustainability* **2020**, *12*, 7575. [CrossRef]
9. Ali, Q.; Salman, A.; Yaacob, H.; Zaini, Z.; Abdullah, R. Does Big Data Analytics Enhance Sustainability and Financial Performance? The Case of ASEAN Banks. *J. Asian Financ. Econ. Bus.* **2020**, *7*, 1–13. [CrossRef]
10. Djeundje, V.B.; Crook, J.; Calabrese, R.; Hamid, M. Enhancing credit scoring with alternative data. *Expert Syst. Appl.* **2021**, *163*, 113766. [CrossRef]
11. Hoang, V.H.N.P.M.; Luu, T.M.N.; Vu, T.M.H. Determinants of Intention to Borrow Consumer Credit in Vietnam: Application and Extension of Technology Acceptance Model. *J. Asian Financ. Econ. Bus.* **2021**, *8*, 885–895.
12. Berg, T.; Burg, V.; Gombović, A.; Puri, M. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *Rev. Financ. Stud.* **2020**, *33*, 2845–2897. [CrossRef]
13. Lin, M.; Prabhala, N.R.; Viswanathan, S. Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Manag. Sci.* **2012**, *59*, 17–35. [CrossRef]
14. Netzer, O.; Lemaire, A.; Herzenstein, M. When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *J. Mark. Res.* **2019**, *56*, 960–980. [CrossRef]
15. Óskarsdóttir, M.; Bravo, C.; Sarraute, C.; Vanthienen, J.; Baesens, B. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Appl. Soft Comput.* **2019**, *74*, 26–39. [CrossRef]
16. Kennedy, K. Credit Scoring Using Machine Learning. Ph.D. Thesis, Technological University Dublin, Dublin, Ireland, 2013.
17. Kopka, M.; Kudělka, M. Analysis of SAP Log Data Based on Network Community Decomposition. *Information* **2019**, *10*, 92. [CrossRef]
18. Farzad, A.; Gulliver, T.A. Unsupervised log message anomaly detection. *ICT Express* **2020**, *6*, 229–237. [CrossRef]
19. Zeng, G. A necessary condition for a good binning algorithm in credit scoring. *Appl. Math. Sci.* **2014**, *8*, 3229–3242. [CrossRef]
20. Finlay, S. *Credit Scoring, Response Modelling and Insurance Rating*; Palgrave Macmillan: London, UK, 2010.
21. Siddiqi, N. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
22. Chi, B.-W.; Hsu, C.-C. A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Syst. Appl.* **2012**, *39*, 2650–2661. [CrossRef]
23. Kolmogorov, B.V.G.A.N. *Limit Distributions for Sums of Independent Random Variables*; Addison-Wesley Publishing Company: Cambridge, MA, USA, 1954.
24. Hosmer, D.W.; Lemeshow, S. Goodness of fit tests for the multiple logistic regression model. *Commun. Stat.-Theor. Method* **1980**, *9*, 1043–1069. [CrossRef]
25. Senaviratna, N.A.M.R.; Cooray, T.M.J.A. Diagnosing Multicollinearity of Logistic Regression Model. *Asian J. Probab. Stat.* **2019**, *5*, 1–9. [CrossRef]
26. Puri, M.; Rocholl, J.; Steffen, S. What do a million observations have to say about loan defaults? Opening the black box of relationships. *J. Financ. Intermed.* **2017**, *31*, 1–15. [CrossRef]
27. World_Bank. UFA2020 Overview: Universal Financial Access by 2020. World Bank Group 2018. Available online: <https://www.worldbank.org/en/topic/financialinclusion/brief/achieving-universal-financial-access-by-2020>. (accessed on 1 October 2018).
28. Barth, J.R.; Lin, C.; Lin, P.; Song, F.M. Corruption in bank lending to firms: Cross-country micro evidence on the beneficial role of competition and information sharing. *J. Financ. Econ.* **2009**, *91*, 361–388. [CrossRef]
29. Pagano, M.; Jappelli, T. Information Sharing in Credit Markets. *J. Financ.* **1993**, *48*, 1693–1718. [CrossRef]
30. Hou, R.; Kong, Y.; Cai, B.; Liu, H. Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning. *Neural Comput. Appl.* **2020**, *32*, 5399–5407. [CrossRef]