symmetry

MDPI

Article

Spatiotemporal Integration of Mobile, Satellite, and Public Geospatial Data for Enhanced Credit Scoring

Naomi Simumba *, Suguru Okami, Akira Kodaka and Naohiko Kohtake

Graduate School of System Design and Management, Keio University, Yokohama 223-8526, Japan; yiu68252@nifty.com (S.O.); akira.kodaka@sdm.keio.ac.jp (A.K.); kohtake@sdm.keio.ac.jp (N.K.)
* Correspondence: naomi.simumba@keio.jp

Abstract: Credit scoring of financially excluded persons is challenging for financial institutions because of a lack of financial data and long physical distances, which hamper data collection. The remote collection of alternative data has the potential to overcome these challenges, enabling credit access for such individuals. Whereas alternative data sources such as mobile phones have been investigated by previous researchers, this research proposes the integration of mobile-phone, satellite, and public geospatial data to improve credit evaluations where financial data are lacking. An approach to integrating these disparate data sources involving both spatial and temporal analysis methods such as spatial aggregation was employed, resulting in various data combinations. The resulting data sets were used to train classifiers of varying complexity, from logistic regression to ensemble learning. Comparisons were based on various performance metrics, including accuracy and the area under the receiver operating-characteristic curve. The combination of all three data sources performed significantly better than mobile-phone data, with the mean classifier accuracy and F1 score improving by 18% and 0.149, respectively. It is shown how these improvements can translate to cost savings for financial institutions through a reduction in misclassification errors. Alternative data combined in this manner could enhance credit provision to financially excluded persons while managing associated risks, leading to greater financial inclusion.

Keywords: credit scoring; satellite data; machine learning; ensemble learning; financial inclusion

check for updates

Citation: Simumba, N.; Okami, S.; Kodaka, A.; Kohtake, N. Spatiotemporal Integration of Mobile, Satellite, and Public Geospatial Data for Enhanced Credit Scoring. *Symmetry* **2021**, *13*, 575. https://doi.org/10.3390/sym13040575

Academic Editor: Jan Awrejcewicz

Received: 15 February 2021
Accepted: 26 March 2021
Published: 31 March 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

Financial exclusion, which implies a lack of access to useful, affordable financial products and services, is a prevalent socioeconomic challenge affecting a large portion of the world's population. An estimated 31% of adults do not possess an account with formal monetary institutions or mobile money lenders [1], which is a key indicator of financial inclusion. Among the causes of financial exclusion are illiteracy, low income, and long distances [2], all of which affect the ability of financially excluded persons to build the financial history typically required for credit scoring. Subsequently, a lack of financial history limits one's access to credit services. To resolve this, alternative data types, such as utility-payment records and social-media data, have been proposed to replace or supplement data traditionally used in credit-risk assessments [3]. This has the potential of reducing financial exclusion by providing new ways of evaluating financially excluded borrowers [4], increasing their access to credit services and allowing them to engage in more income increasing activities such as education or business investments.

To overcome the problem of long distances currently restricting borrowers' access to financial services, remote data collection is key. As mobile phones have become increasingly ubiquitous, they have become common tools for data collection. In fact, there were over 8 billion mobile/cellular telephone subscriptions and over 5 billion active mobile broadband subscriptions as of 2018 [5]. Mobile-phone application data have been increasingly employed in various credit-decision systems [6], increasing access to financial services through digital credit in emerging markets. However, data from a single alternative data

source may have a limited number of features or predictive power. Additional information pertaining to a borrower's environment that may affect repayment ability can be obtained through satellites and public geospatial sources, both of which allow for remote data collection. Satellite data are obtained through Earth observation satellites collecting information about the Earth's surface and atmosphere. Modern Earth observation satellites have a higher revisit frequency, coverage area, and resolution. As a result, satellites provide several advantages with regard to data collection, including a wide coverage and ability to collect information that would be difficult to collect otherwise, making them an ideal source of data for a wide range of applications [7]. Geospatial data, collected and maintained by governmental and nonprofit organizations for project planning and policymaking, is often publicly available and provides information on environmental characteristics that could be relevant to credit evaluation. Examples include income levels, as well as access to roads and other infrastructure. Although merging mobile-phone, satellite, and public geospatial data sources for credit scoring may have benefits, the challenge in incorporating satellite and public geospatial data lies in the need to extract only the data that are relevant to individual borrowers.

This research proposes an approach using location and time, collected through mobile phones, as the basis of spatiotemporal analysis to extract public geospatial and satellite data features specific to individual borrowers. Because these data sources provide information on different personal and environmental aspects of the borrower, combining them could improve the performance of credit-scoring models based on alternative data. By extension, this may reduce costs for financial institutions. The feasibility of this approach is demonstrated by integrating data with geospatial analysis in the Quantum Geographic Information System software QGIS [8] and spatiotemporal analysis in Google Earth Engine [9]. Credit-scoring models are built on these data combinations, and any enhancements in their performance measures are evaluated. Further analysis shows the potential resulting cost savings for financial institutions using this method. This paper next presents related research, detailed in Section 2. The proposed approach and algorithms are given in Sections 3 and 4, respectively. In Section 5, an empirical evaluation is conducted with data collected from farmers in rural Cambodia. Section 6 discusses the results of the evaluations. Finally, the conclusion and suggestions for future research are given in Section 7.

2. Related Work

2.1. Credit-Scoring Algorithms

For analysis purposes, credit scoring is treated as a classification problem employing assorted algorithms [10]. Logistic regression (LR) is commonly employed as a basis for comparison with other, more complex methods such as support vector machines (SVMs), artificial neural networks (ANNs), and extreme learning machines [11]. Recent research trends have leaned towards the use of ensemble classifiers, which combine several classifiers for improved performance [12]. Beyond accuracy, additional considerations in the selection of classifiers include complexity and the cost of misclassification [13], because they affect the implementation and use of credit-scoring models. Further, recent research has placed greater emphasis on enhancing profitability through feature selection or profit scoring [14,15].

2.2. Alternative Data in Credit Scoring

Broadly, data used for credit scoring can be categorized as traditional or alternative. Information traditionally used for credit scoring includes demographic information and financial history such as loan inquiries. Due to the deficiency of these types of information for financially excluded persons, the use of alternative data in credit scoring has risen. In recent years, several companies operating in developing economies have begun to offer digital credit services, with prescreening of potential borrowers done based on mobile-phone-usage data [16]. Ref. [17] developed a credit-scoring system for selecting customers

to transition from prepaid to postpaid mobile-phone subscriptions, a form of digital credit. Behavioral signatures from mobile-phone usage, including frequency and duration of calls, were employed to predict defaults. It was found that these indicators performed better performance than credit-bureau information for thin-file borrowers (borrowers for whom credit bureaus hold limited information). Ref. [18] saw a significant improvement in credit-scoring-model performance, as measured by the area under the receiver operating curve (AUC), when financial history (including bank account and credit card activity) was incorporated with mobile-phone-usage data. Ref. [19] proposed the use of mobile-application-usage data for credit evaluation with alternative scoring factors for financially excluded persons. Mobile applications provide another source of alternative data, as shown by Ntwiga et al., who used data from mobile financial transactions to create a credit-evaluation process for unbanked individuals [20]. Ref. [21] proposed the development of a mobile application to collect data from social media for credit scoring to simplify data collection for financial institutions. Ref. [22] proposed a method of improving the precision of credit-scoring models by recursively incorporating client network data.

2.3. Data Combination

Integration of data from various sources has been used in several applications to model or improve data accuracy in the credit evaluation process. Ref. [23] combined sociodemographic, e-mail usage, and psychometric data for credit scoring. The psychometric factors encompassed whether the borrower identified as a team player or an individualist. Examples of e-mail usage data included the number of e-mails sent and what fractions of those e-mails were sent on different days. Sociodemographic data comprised age, gender, and number of dependents, among other factors. Combining data from all three sources was found to increase training AUC. Public geospatial data such as macroeconomic indicators have proved useful in credit scoring [24]. Ref. [25] found that the inclusion of macroeconomic factors; for example, unemployment and mortgage rates, led to significant improvement in accuracy of multistate credit-delinquency models. Ref. [26] determined that the inclusion of a spatial risk parameter enhanced the performance of LR models used for scoring the credit of small and medium enterprises (SMEs). However, efforts to include alternative data in credit scoring have mainly considered one source of alternative data and have not focused on remote data collection. There is limited research describing the process, challenges, and value of integrating several alternative data sources, particularly the integration of mobile-phone data with public geospatial and satellite data, which is a gap addressed by this research.

3. Proposed Approach

This research proposes the integration of three alternative data sources; namely, mobile-phone data, satellite data, and public geospatial data, to maximize the performance of credit-scoring models built on alternative data. It is theorized that the inclusion of data from multiple sources can provide valuable information on different aspects of the potential borrower and their environment, thus improving the accuracy of credit-scoring models. Further, as all three sources can be collected remotely, their use overcomes the problem of long distance commonly associated with financially excluded persons. A credit-scoring system based on this approach would operate as shown in Figure 1. Industry experts may be consulted to determine which factors from each data source could be relevant in each scenario. It is crucial to note that because of their nature, public geospatial and satellite data cannot be used to collect borrowers' personal details.

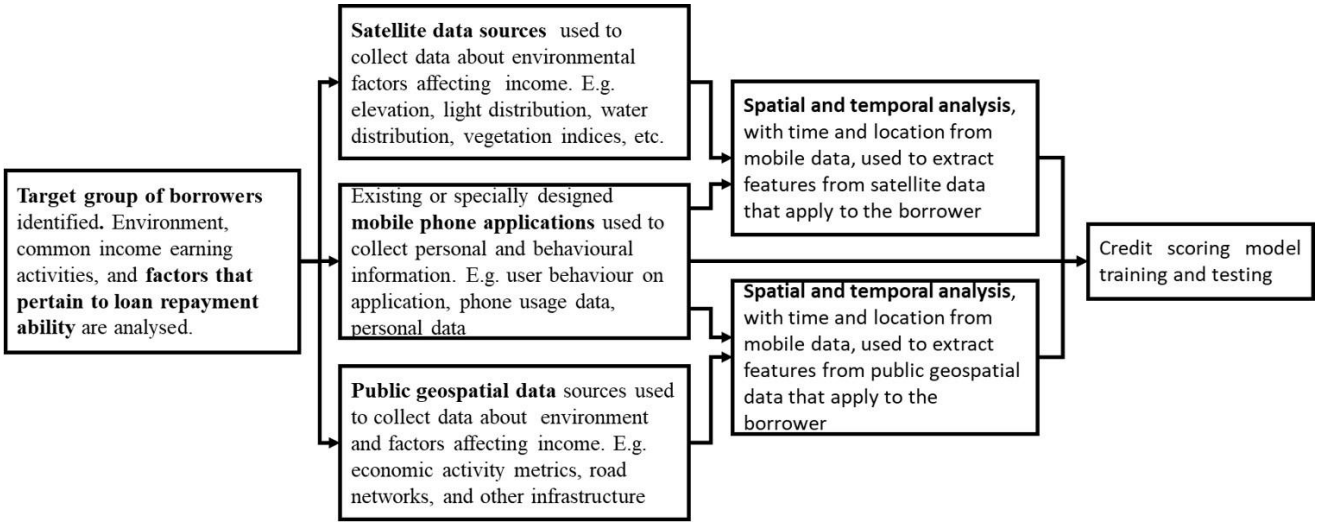

Figure 1. Proposed approach.

Because this information is general, it must be combined with personal and behavioral information of the borrowers through geospatial and temporal analysis. Importantly, to integrate data from mobile phones with data from the other sources, location information must be collected through the mobile phone. Data may be integrated using various geospatial- and temporal-analysis tools, resulting in an integrated dataset for the development of a credit-scoring model. Credit-decision models may then be trained on these data with methods that have been proven useful for credit scoring.

4. Methods

Credit decisions were historically based on the lender's knowledge of the borrower [27]. Recently, statistical and machine-learning approaches have taken precedence. The goal of these approaches is to assess the risk of lending to a prospective borrower, thus distinguishing between "good" and "bad" borrowers. This section introduces several classifiers that are applied in this domain.

4.1. Multiple Logistic Regression

The probability, Y, that a borrower belongs to a given class is found using associated predictors, X, by (1); where bn denotes the coefficient of Xn. A probability threshold is selected to determine the borrower's class, where a probability above the threshold would signify a "good" classification.

$$ln\frac{Y}{1-Y} = b_0 + b_1X_1 + b_2X_2 + \ldots \quad (1)$$

4.2. Support Vector Machines (SVM)

Here, a hyperplane is defined to separate the data points into classes while maximizing the margins around the hyperplane. Data points lying on the margins are referred to as support vectors, and soft margins are used to determine the impact of any points that fall into the wrong classification. A cost parameter, C, controls the soft margin, and thus the cost of errors. SVMs fit both linear and nonlinear data due to the use of the kernel trick, which maps features into a higher dimensional space when necessary [28]. Two types of kernels are used here.

Linear kernel: a linear feature space is maintained, and features are mapped using the relationship expressed in (2) below, where x and z are two feature vectors.

$$K(x,z) = x \cdot z \tag{2}$$

Gaussian kernel (radial basis function, RBF): this method transforms the feature space into higher dimensions, making it a better candidate when there are nonlinear interactions in the data [29]. The kernel function describing the mapping relationship is given in (3) below, where x and z are two feature vectors and γ is a parameter to be optimized by tuning.

$$K(x,z) = exp\left\{-\gamma\left(\|x-z\|^2\right)\right\} \tag{3}$$

4.3. Artificial Neural Networks (ANNs)

An artificial neural network is a combination of interconnected nodes, inspired by neural networks in the brain, which can be applied to a wide range of analyses. Nodes are aggregated into layers, and each node converts its input to an output based on an activation function, with the logistic function (4) and hyperbolic tangent function (5) being two commonly used activation functions. In the case of feedforward networks, each layer's output is used as the input to the next layer until the final output is produced. Multilayer perceptron (MLP) networks consist of at least three layers: input, output, and at least one hidden layer [30]. The number of input nodes is determined by the number of inputs, and a single output node is required for binary classification, with hidden nodes determined by tuning.

$$S(x) = \frac{1}{1+e^{-x}} \tag{4}$$

$$S(x) = \tan \mathrm{h}(x) \tag{5}$$

4.4. Ensemble Methods

Ensemble classifiers grouped into bootstrap aggregation (bagging) [31], boosting [32], and stacking, combine several classifiers to improve performance. With bagging, subsets of the training data are created via bootstrap sampling. Base learners are trained on these subsets, and their outputs are aggregated. This has the advantage of decreasing variance and overfitting. Random forests bear some similarity to bagging in that they too use bootstrap sampling to subset training data. Beyond this, random forests also create subsets of the input features. The result is further error reduction and decreased variance. In boosting, base learners are combined in sequence. The weights of misclassifications are increased to give them greater significance in the training of the next learner. Thus, each subsequent classifier improves overall performance. Generally, boosting reduces bias. Adaptive boosting (AdaBoost) [33] is a popular example of this method.

4.5. Model-Evaluation Methods

Classification algorithms can be compared based on their accuracy and the area under the receiver operating curve (AUC) [34], both of which are independent of the analysis method used. The receiver operating characteristic (ROC) curves plot the true positive rate, or sensitivity (7), against the false positive rate, also given as 1—specificity (6).

$$specificity = true\ negatives / true\ negatives + false\ positives \tag{6}$$

$$sensitivity = true\ positives / true\ positives + false\ negatives \tag{7}$$

The AUC value gives the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, and AUC values closer to 1 imply a better classifier. Finally, accuracy is defined, where:

$$accuracy = true\ positives + true\ negatives/total \quad (8)$$

5. Empirical Evaluation

5.1. Study Area

For a practical evaluation of the proposed approach, credit data from rural farmers in Cambodia were used. As of 2014, 14% of the country's population lived below the national poverty line [35]. Nearly 80% of the country's 15.4 million people reside in rural areas where the major economic activity is agriculture [36]. In fact, half the population is employed in the agricultural industry [37]. Access to financial services remains a challenge, with less than 30% of adults using formal lending services [38].

5.2. Data Collection and Preparation

Credit scoring for smallholder farmers requires consideration of the data pertaining to repayment ability such as rainfall and temperature. Information about factors such as water availability, access to markets, and vegetation, can be obtained from satellite images and public geospatial data. Moreover, information about behavior and demographics can be collected through a mobile application.

5.2.1. Credit Data

The repayment status of loans provided to farmers in rural Cambodia from 2018 to 2019 was assessed. Loans matured after one year and fell into one of two categories. The "Paid" status, encoded as binary 1, was applied to matured loans that had been repaid in full at the time of loan maturity. In contrast, the "Default" status, encoded as binary 0, was applied to loans that had not been fully repaid by loan maturity.

5.2.2. Mobile-Phone Data

The mobile-phone data used in this study were collected by an agribusiness company through its mobile application. Application users collected information about farm activity from farmers in their area, who were unable to use the application directly. The data used here were anonymized and used solely for research purposes. It included the behavior of the mobile-application users, the number of farms they service, and whether they take recommended actions such as uploading photographs. Behavioral information was chosen because it can give insight into an individual's character. Importantly, the location of each farm, which is necessary for combining the mobile data with the other data types, was also collected, as shown in Figure 2a.

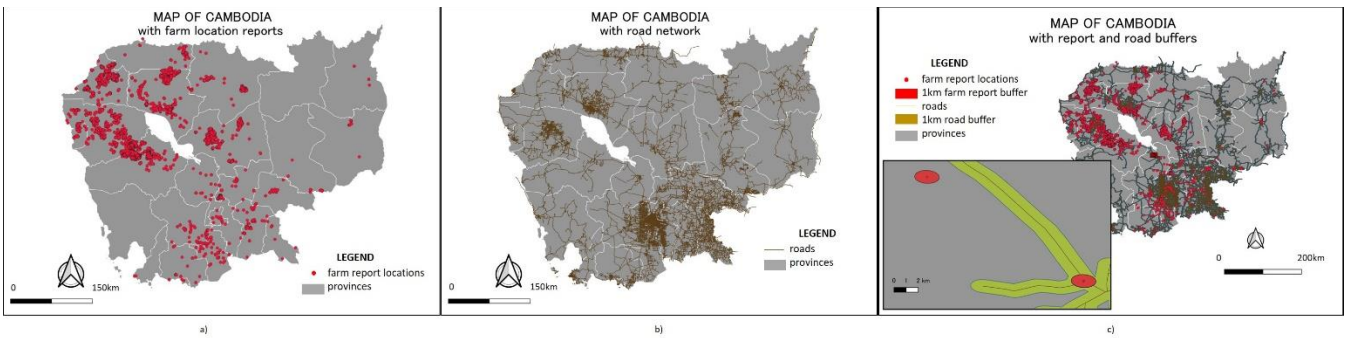

Figure 2. Preparation of public geospatial data: (**a**) locations of mobile-phone data; (**b**) road network from public geospatial data source; (**c**) spatial analysis to determine distance.

5.2.3. Public Geospatial Data

Geospatial data (data linked to a geographical location) were collected from the public data source OpenDevelopment Cambodia [39]. The collected geospatial data included data on canals, roads (Figure 2b), and rivers; these data were selected for their potential impact on revenue from farming activities and thus loan repayment.

To prepare the data for use in the scoring process, they were analyzed using the QGIS platform. Spatial analysis was conducted to find the overlap between buffers around reports and buffers around roads, rivers, and canals. For example, to determine which farms were within a kilometer of major roads, a 1 km fixed buffer was created around all roads. A separate 1 km fixed buffer was created around farm locations. Farms with a buffer that intersected the road buffer were identified as within 1 km of roads, as shown in Figure 2c. This process was repeated with buffers of varying lengths, as well as with the other data. As a result of this analysis, it was possible to identify farmers within 1 km, 2 km, 5 km, and 10 km of rivers, roads, and canals. From these, specific variables were selected based on suitability, correlation with other variables, and predictive power.

5.2.4. Satellite Data

The use of satellite data is relevant because it has found applications in many different domains. For instance, vegetation indices, calculated by combining spectral reflectance values, are used to monitor vegetation health and predict crop yields [40]. The satellite data used in this analysis were prepared on the Google Earth Engine platform. Data were selected based on their potential to predict the loan repayment given the income-earning activities of the target group of borrowers. Shuttle Radar Topography Mission (SRTM) digital elevation data [41] were used to determine the elevation and slope of each farm location. Normalized difference vegetation index (NDVI) values were assessed using data from the Moderate Resolution Imaging Spectrometer (MODIS) 16-day NDVI composite data set [42], whereas the Terra Land Surface Temperature and Emissivity Daily Global 1 km dataset [43] was used to calculate temperature, and the Terra Surface Reflectance Daily L2G Global 1 km and 500 m data set [44] was used for Normalized difference water index (NDWI) assessment. As the evaluation required aggregation over an area, the lowest political boundary, known as a commune, was chosen. For a farmer Y living in commune X, the process outlined in Figure 3 was used to prepare the satellite data using known locations from mobile data.

This process was repeated for all farmers to determine the average NDVI, NDWI, and temperature values for each commune for each month in the three years prior to the loan disbursement. Variables were selected based on suitability, correlation with other variables, and predictive power. Table 1 details the resulting features, chosen to keep the feature set small. Variable importance was assessed by conducting receiver operating characteristic analysis. To achieve this, a logistic regression model was trained with each input variable in turn, and the ROC curve was plotted using sensitivity and specificity. The AUC values of the resulting ROC curves were used to measure variable importance. Essentially, this showed how well each individual variable predicts the dependent value.

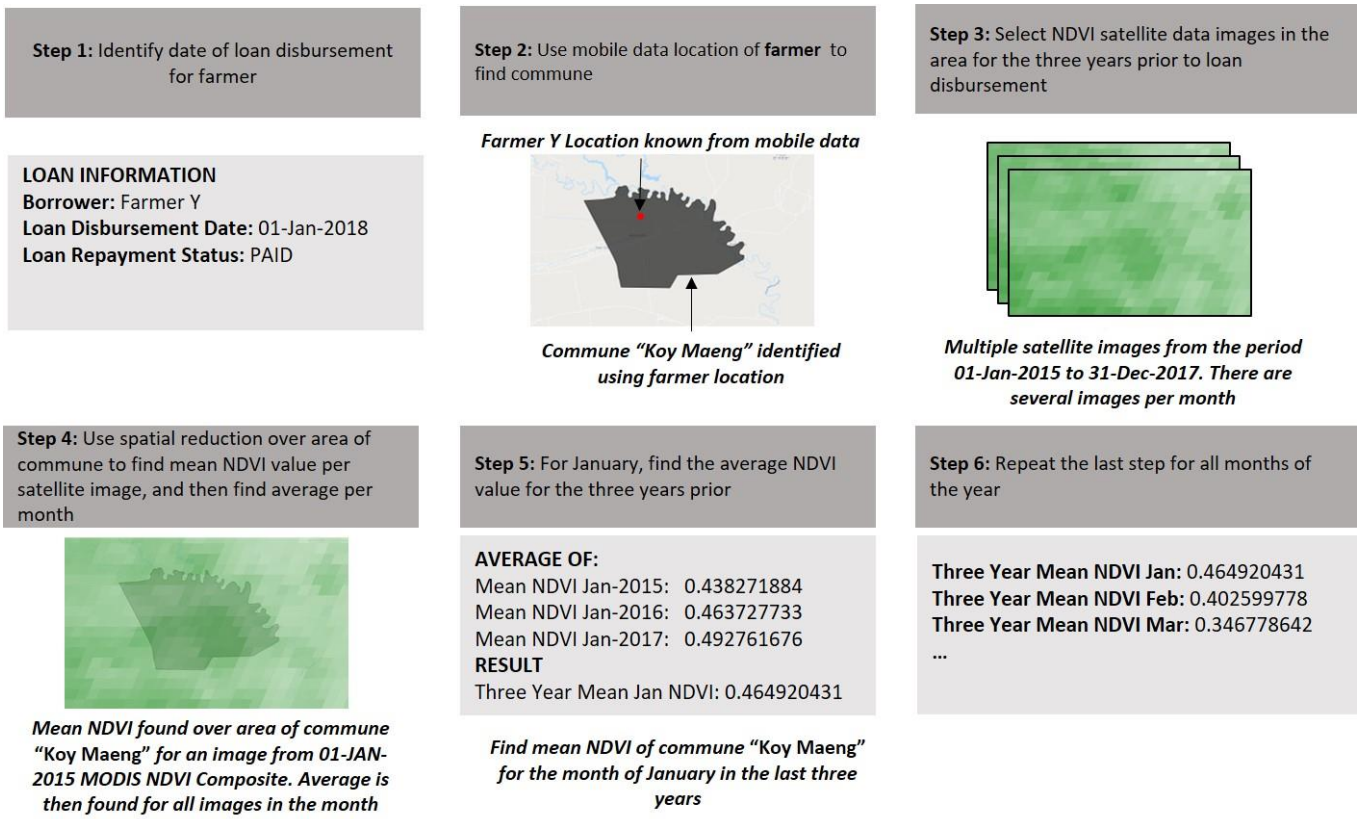

Figure 3. Preparation of satellite data.

Table 1. Details of variables.

Data Source	Variables	Type	Importance
Mobile-phone data	Has uploaded a photo of the family	Nominal	0.518
	Has uploaded a photo of the house	Nominal	0.508
	Has uploaded a photo of the land document	Nominal	0.514
	Has uploaded a photo of the ID card	Nominal	0.508
	Number of farms registered	Numeric	0.614
Satellite data	Mean July temperature	Numeric	0.734
	Mean November NDVI	Numeric	0.697
	Mean September NDWI	Numeric	**0.778**
	Farm elevation	Numeric	0.661
	Farm slope	Numeric	0.603
Public geospatial data	Farm is within 1 km of a canal	Nominal	0.642
	Farm is within 2 km of a canal	Nominal	0.629
	Farm is within 1 km of a road	Nominal	0.569
	Farm is within 2 km of a road	Nominal	0.568
	Farm is within 5 km of a river	Nominal	0.558

6. Analysis

Four data sets were created by combining variables according to their source: mobile-phone data only, mobile and satellite data, mobile and geospatial data, and finally mobile, satellite, and geospatial data. This combination process is shown in Figure 4. The pre-processing referred to in Figure 4 involved imputing data using mean values and rescaling using the max-min method for numeric variables. For nominal variables with r nominal values were replaced with $r - 1$ binary columns. There were 245:43 entries, resulting in an imbalance ratio of 1 to 5.7. The synthetic minority oversampling technique (SMOTE), which synthesizes new instances of the minority class, was employed to obtain class balance.

Data were split into training (80%) and holdout (20%) sets. The models used in these experiments were multiple LR (LR), SVM with a linear kernel (SVM-linear), SVM with an RBF kernel (SVM-RBF), ANN, random forest, bagging classification and regression trees (bagging CART), and AdaBoost. Tenfold cross-validation was repeated three times to train the models. The best model from this analysis was selected according to its ROC curve performance. The tuned parameters for the final models are given in Appendix A for parameters that required tuning. For each model, the average ROC curve from each iteration of the cross-validation was plotted. The resulting models were validated on the remaining 20% of the original data and evaluated for accuracy. Trained models were evaluated by comparing their specificity, sensitivity, accuracy, ROC curves, and the area under the ROC curve (AUC). These metrics were chosen for their lack of dependence on classification method. Training time taken for each model was calculated as the difference between the system time before and after model training on a computer with an i5 2.5 GHz processor and 8 GB RAM.

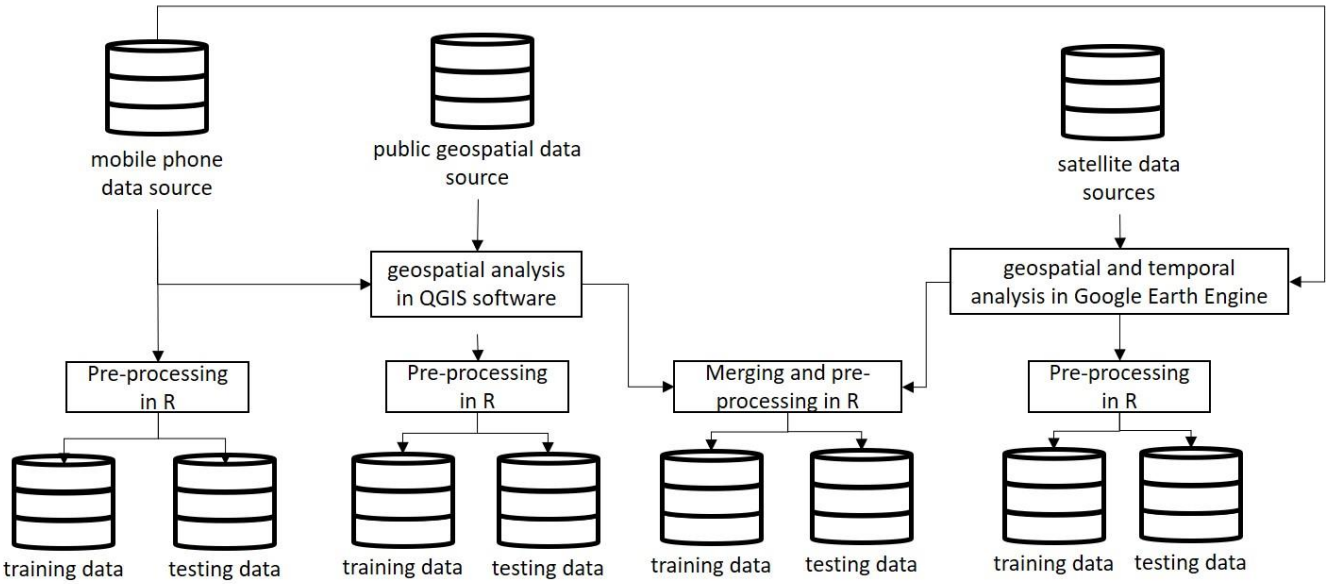

Figure 4. Data-combination process.

7. Results

For each data combination, model training was conducted as detailed in the previous section for each of the seven algorithms. Thus, 28 models were trained in total. The ROC curves obtained during training were evaluated with the corresponding AUC values given in Table 2, which shows how well the models performed on training data. The trained models were then evaluated on the remaining data. The accuracies, as defined by Equation (8), of the trained models on the 20% holdout data are given in Table 3 for each data set. Additionally, the true positive rate (TPR), given by Equation (7), for each model on the holdout data is presented in Table 4 The true negative rate (TNR) is similarly given in Table 5. Finally, the F1 values of the models on the holdout data are shown in Table 6. In Figure 5, clustered columns are used to graphically show the improvement in performance as measured by accuracy and F1. Table 7 gives the training time in seconds required to train each model with each data combination. Mean and median accuracy of all the models for each data combination is displayed in Figure 5a, while the mean and median F1 values are shown in Figure 5b.

Table 2. Training AUC values.

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.582	0.679	0.814	0.824
SVM—linear	0.484	0.637	0.800	0.821
SVM—RBF	0.725	0.846	0.874	0.902
ANN	0.688	0.805	0.839	0.861
Random Forest	0.814	0.922	0.93	**0.938**
Bagging CART	0.863	0.917	0.917	0.919
AdaBoost	0.687	0.771	0.831	0.828

Table 3. Accuracy.

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.511	0.609	0.685	0.761
SVM—linear	0.522	0.577	0.653	0.783
SVM—RBF	0.707	0.772	0.816	0.859
ANN	0.609	0.696	0.837	0.827
Random Forest	0.772	0.870	0.848	**0.881**
Bagging CART	0.761	0.870	0.859	0.859
AdaBoost	0.707	0.892	0.837	**0.881**

Table 4. True Positive Rate (TPR, Sensitivity).

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.756	0.552	0.633	0.633
SVM—linear	0.858	0.449	0.572	0.674
SVM—RBF	0.654	0.654	0.776	0.858
ANN	0.572	0.449	0.837	0.776
Random Forest	0.735	0.858	0.817	**0.898**
Bagging CART	0.735	0.878	0.837	0.878
AdaBoost	0.572	0.858	0.796	0.878

Table 5. True Negative Rate (TNR, Specificity).

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.233	0.675	0.745	**0.907**
SVM—linear	0.140	0.721	0.745	**0.907**
SVM—RBF	0.768	0.907	0.861	0.861
ANN	0.652	0.977	0.838	0.884
Random Forest	0.814	0.884	0.884	0.861
Bagging CART	0.791	0.861	0.884	0.838
AdaBoost	0.861	0.931	0.884	0.884

Table 6. F1.

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.622	0.600	0.682	0.739
SVM—linear	0.657	0.531	0.637	0.768
SVM—RBF	0.704	0.753	0.818	0.866
ANN	0.609	0.612	0.846	0.827
Random Forest	0.775	0.875	0.852	**0.889**
Bagging CART	0.766	0.878	0.864	0.869
AdaBoost	0.675	0.894	0.839	0.887

Table 7. Training time in seconds.

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
LR	0.927	1.026	1.516	1.180
SVM—linear	1.437	1.561	2.218	2.075
SVM—RBF	15.081	15.082	18.172	16.090
ANN	10.201	11.652	12.185	12.563
Random Forest	12.409	27.090	26.879	44.580
Bagging CART	4.470	5.162	5.373	6.069
AdaBoost	2.404	3.426	3.623	4.768

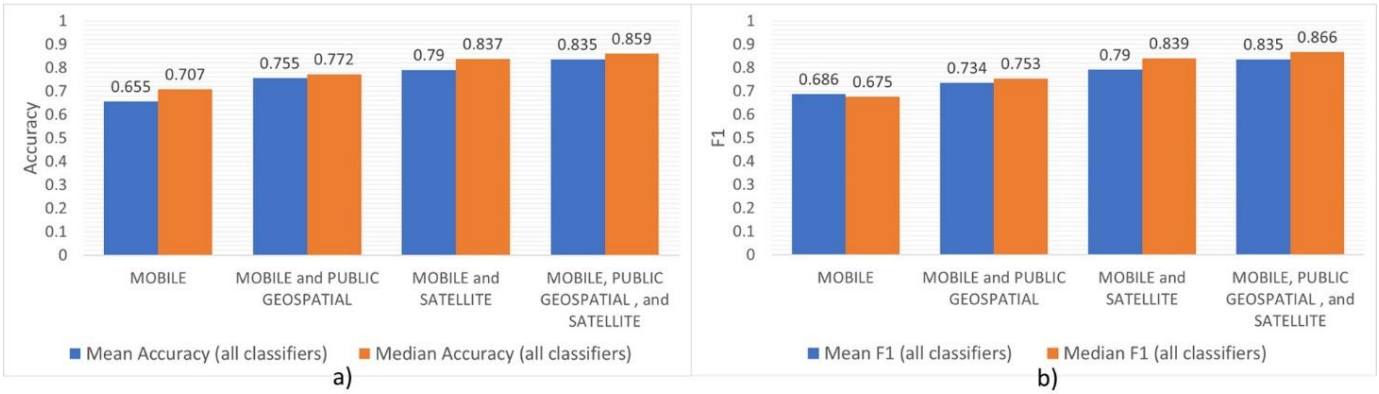

Figure 5. Mean and median performance measures: (**a**) Mean and median accuracy; (**b**) mean and median F1.

8. Discussion

8.1. Data Combination

Models trained solely on mobile-phone data performed poorly, with the lowest AUC values (Table 2) and accuracy (Table 3). This was in line with expectations, given that variables from mobile-phone data had the lowest variable importance values (Table 1). Clear improvements were observed in the AUC, accuracy, and F1 values (Table 6) when data from new sources were incorporated, supporting the hypothesis that combining data from various sources improves model performance. Overall, the grouping of mobile-phone and satellite data increased model performance, as measured by AUC, accuracy, and F1 values. Similarly, combining mobile-phone and public geospatial data largely increased these measures. Notably, variables from satellite data had higher variable importance values than those from mobile-phone and public geospatial data. Model training results

reflected this, with AUC values from the combination of mobile-phone and satellite data being superior to those from the combination of mobile-phone and public geospatial data. Further improvement was noted in the AUC values obtained using mobile-phone, public geospatial, and satellite data. Additionally, this combination outperformed mobile-phone data with respect to the mean and median accuracy as well as F1 of all classifiers (Figure 5), with mean classifier accuracy and F1 measure improving by 18% and 0.149, respectively. Similar trends were observed in the TPR (Table 4) and TNR (Table 5) values. When integrating data from varied sources, care must be taken to ensure that the resulting models are not overfitted to the training data. Thus, the performance of the models on the holdout data is crucial. Higher holdout accuracy, TPR, TNR, and F1 of the models signify that the improved performance can be obtained on data not in the training data set. However, given the small size of the holdout data used, the process should be repeated on larger sets of out-of-sample data if possible.

For further evaluation, the Friedman test [45] was adopted to compare holdout accuracy values obtained by the models on the different data sets. After rejecting the null hypothesis based on the p value, pairwise comparisons were made using the Nemenyi test [46], as shown in Table 8. The mobile-phone data set led to the poorest metrics. Meanwhile, the data set combining mobile, satellite, and public geospatial data produced the best accuracy and performed significantly better than the mobile-phone dataset.

Table 8. Friedman test.

Data	Average Ranks	Nemenyi Post Hoc Test p-Values		
		Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data
Mobile-phone data	4			
Mobile and public geospatial data	2.286	0.062		
Mobile and satellite data	2.214	0.048	1	
Mobile, satellite, and public geospatial data	1.500	**0.002**	0.666	0.729
Friedman statistic		14.391		
Friedman p value		0.002		

8.2. Alternative Data

Several considerations must be made regarding the use of alternative data in credit scoring. Public geospatial data may suffer problems with reliability and incompleteness. For instance, although several different types of public geospatial data were available, including the locations of rivers, roads, schools, railways, and canals, some of the data could not be used because of their incompleteness. Additionally, mobile and satellite data tend to have a greater update frequency than public geospatial data, meaning the data are more likely to represent the current circumstances of the borrower. Spatial resolution, frequency of collection, and reliability of satellite data similarly require consideration. Higher spatial resolutions and frequency of collection can provide information that is more consistent with the situation on the ground. Data scarcity and missingness often lead to challenges in creating credit scores for the financially excluded, making the inclusion of data from multiple sources even more valuable. Obtaining mobile data for evaluation may require collaboration between financial institutions and providers of mobile-phone services or mobile applications. Alternatively, these mobile-phone service providers may extend their services to include lending. Other data sources may be used depending on their capacity to predict loan repayment for the selected group of borrowers. Crucially, ethical selection of variables and ensuring privacy of borrowers throughout the process is essential, not only for legal compliance, but also to engender trust in the system. Ethics committees should

be used to ensure this. Special care must be taken to ensure that variables are selected fairly and with clear justification to prevent discrimination. As mobile-phone users sign up for loans, their permission and clear understanding is required to employ mobile-phone data for the credit-evaluation process. Data security should be preserved. Additionally, repeated validation of the credit-scoring models on larger data sets is important to ensure acceptance by the banking industry. As with all systems, the risk of unscrupulous users exists. However, detection of unscrupulous borrowers based on their behavior may reduce the degree to which this occurs. Updates to the data and evaluation methods may also mitigate this risk.

8.3. Classification Methods

Seven classification methods of varying complexity were applied to the credit-scoring process. Although no algorithm outperformed the rest on all metrics, the random forest method was consistently among the best. However, selecting the most suitable method for the application requires consideration of factors beyond accuracy.

Training time, ease of explanation, computational expense, and complexity also deserve consideration [11]. To consider the training time required for each model, the difference between the system time before and after training of each model on each data combination was measured in seconds (Table 7). As expected, raining time increased as more data were added. Although the random forest models performed well in terms of accuracy and training AUC, they also proved to have a greater computational cost than the remaining methods. One the other hand, simpler methods such as logistic regression and linear support vector machines have shorter training times. Viewed in this light, it becomes apparent that although the more complex methods such as random forest, support vector machines with an RBF kernel, and neural networks performed very well. The long training time that would be needed if the credit-decision system were to be scaled up to larger data sets may make a simpler classifier, such as logistic regression, preferable. It is also important that the prediction time be short as the models are implemented. Further, the coefficients of the logistic regression model allow for an intuitive understanding, making it easier for stakeholders to grasp the impact of model inputs.

8.4. Cost of Misclassification

The cost of misclassifying borrowers depends on the type of classification error made. For type 2 errors, where bad borrowers are given a good rating and given loans, due to defaults is a complex task affected by several factors, such as the installments paid before default, costs of loan collection, and time taken to recoup the loan. Additionally, high default rates may damage the reputation of financial institutions in the community. The probability of default, loss given default, and exposure to default are multiplied to calculate the expected loss on a loan, which is necessary for capital assessments according to international regulations issued by the Basel Committee on Banking Supervision. With type 1 errors, where good borrowers given a poor rating and denied loans, the financial institution loses the profit that could have been made on the loan, as well as any additional financial products that could have been extended to the borrower. It is clear the costs of type 1 and type 2 errors are not the same, making the assessment of misclassification cost challenging. Lacking information on the loss given default and exposure at default for this data set, the true cost of misclassification cannot be calculated.

For a simpler approach, the cost can be equated to the weighted sum of the false positive and false negative rates of the following equation. Here, C1 and C2 represent the cost of classifying a good applicant as bad and cost of classifying a bad borrower as good, respectively. Additionally, E1 and E2 are the associated probabilities of misclassification.

$$cost = (C_1 * E_1) + (C_2 * E_2) \quad (9)$$

Using the TPR and TNR results shown in Tables 4 and 5, the cost was calculated for several cost ratios (C1:C2). Figure 6 shows the reduction in costs as data were integrated. This implies that there is a potential for cost reduction for financial institutions by using combined data for credit scoring.

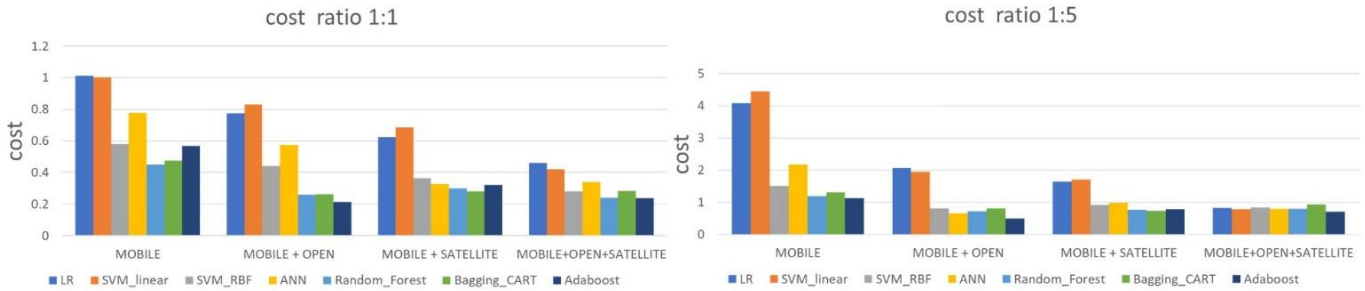

Figure 6. Cost of misclassification for varying cost ratios.

8.5. General Implementation

Although the application demonstrated in this paper focused on farmers, a generalized approach may be applied for evaluation of borrowers earning income from other economic activities. Crucial to implementing the proposed system with a different group of borrowers would be the collection of mobile data, as well as factors in the borrowers' environment that affect their economic activities. For instance, to implement the system for evaluation of small-shop owners who are financially excluded, one may consider their mobile-phone behavior, the size of their clientele base, how accessible their shop is to customers, and the distance to the point where they purchase goods. The system is most useful in evaluation of financially excluded persons who reside in rural areas, where data collection is difficult.

9. Conclusions

This paper proposed a method of combining data from three alternative data sources (mobile-phone, public geospatial, and satellite data) by spatial and temporal analysis for credit scoring of financially excluded persons with the aim of improving performance. Experimental evaluation conducted with data from a mobile application for rural farmers, as well as public geospatial and satellite data, showed that integrating the data sources improved performance, as measured by accuracy, F1, and AUC values. As a result of the reduced misclassification errors, it was demonstrated that costs for financial institutions could be reduced through this data integration. Although the empirical evaluation of this paper focused on credit systems for rural farmers, the proposed approach could be used to make credit decisions for other groups of borrowers. Such an application may require the collection of other variables that better relate to the factors affecting loan repayment among the new group of borrowers.

10. Future Work

Further evaluation may be undertaken by incorporating other data sources such as statistical data, as well as other relevant types of satellite data. The method proposed here could form the foundation of a credit-decision system using different sources and types of alternative data. Credit-scoring systems operating on alternative data collected remotely could result in greater convenience and cost savings for financial institutions lending to financially excluded persons, thus increasing financial inclusion.

Author Contributions: Conceptualization, N.S., S.O., and N.K.; methodology, N.S. and S.O.; formal analysis, validation, visualization, writing original draft, N.S.; review and editing, S.O., A.K., and N.K.; supervision, N.K.; project administration and funding acquisition, A.K. and N.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a JSPS KAKENHI Grant (Grant Number JP19H04100).

Institutional Review Board Statement: This was waived because data was anonymized by the data provider before it was provided to authors for research purposes.

Informed Consent Statement: This was waived because data was anonymized by the data provider before it was provided to authors for research purposes.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Agribuddy Ltd and are available with the permission of Agribuddy Ltd.

Acknowledgments: The authors would like to thank Agribuddy Ltd. (www.agribuddy.com) for their kind assistance in providing data.

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

Appendix A

Table A1 below gives the tuned parameters for the final models.

Table A1. Final model parameters.

Classifier	Mobile-Phone Data	Mobile and Public Geospatial Data	Mobile and Satellite Data	Mobile, Satellite, and Public Geospatial Data
SVM—linear	cost = 1	cost = 1	cost = 1	cost = 1
SVM—RBF	sigma = 8.915, cost = 64	sigma = 0.134, cost = 32	sigma = 0.343, cost = 16	sigma = 0.074, cost = 128
ANN	size of hidden layer = 5, decay = 0.0001	size of hidden layer = 5, decay = 0.1	size of hidden layer = 5, decay = 0.0001	size of hidden layer = 5, decay = 0.1
Random Forest	Number of variables randomly collected to be sampled at each split time (mtry) = 5	Number of variables randomly collected to be sampled at each split time (mtry) = 9	Number of variables randomly collected to be sampled at each split time (mtry) = 3	Number of variables randomly collected to be sampled at each split time (mtry) = 3
AdaBoost	method = AdaBoost.M1, iterations = 100	method = AdaBoost.M1, iterations = 50	method = AdaBoost.M1, iterations = 50	method = AdaBoost.M1, iterations = 50

References

1. Demirgüç-Kunt, A.; Klapper, L.; Singer, D.; Ansar, S.; Hess, J. *The Global Findex Database*; World Bank Group: Washington, DC, USA, 2017.
2. Grandolini, M.G. Five Challenges Prevent Financial Access for People in Developing Countries. World Bank, 15 10 2015. Available online: https://blogs.worldbank.org/voices/five-challenges-prevent-financial-access-people-developing-countries (accessed on 15 August 2017).
3. McEvoy, M.J. *Enabling Financial Inclusion through Alternative Data*; Mastercard Advisors: New York, NY, USA, 2014.
4. Turner, M.A.; Walker, P.D.; Chaudhuri, S.; Varghese, R. *New Pathway to Financial Inclusion*; Policy & Economic Research Council (PERC): Melbourne, Australia, 2012.
5. *Global and Regional ICT Estimates*; International Telecommunications Union: Geneva, Switzerland, 2018.
6. Costa, A.; Deb, A.; Kubzansky, M. Big data, small credit: The digital revolution and its impact on emerging market consumers. *Innov. Technol. Gov. Glob.* **2015**, *10*, 49–80. [CrossRef]
7. Donaldson, D.; Storeygard, A. The view from above: Applications of satellite data in economics. *J. Econ. Perspect.* **2016**, *30*, 171–198. [CrossRef]
8. QGIS Development Team. QGIS Geographic Information System. *Open Source Geospatial Foundation Project.* Available online: http://qgis.osgeo.org (accessed on 5 May 2018).
9. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
10. Thomas, L.C.; Edelman, B.D.; Crook, N.J. *Credit Scoring and Its Applications*; Society for Applied and Industrial Mathematics: Philadelphia, PA, USA, 2002.
11. Lessmann, S.; Baesens, B.; Seow, H.-V.; Thomas, L.C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **2015**, *247*, 124–136. [CrossRef]

12. Papouskova, M.; Hajek, P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decis. Support Syst.* **2019**, *118*, 33–45. [CrossRef]
13. Harris, T. Credit scoring using the clustered support vector machine. *Expert Syst. Appl.* **2015**, *42*, 741–750. [CrossRef]
14. Verbraken, T.; Bravo, C.; Weber, R.; Baesens, B. Development and application of consumer credit scoring models using profit-based classification measures. *Eur. J. Oper. Res.* **2014**, *238*, 505–513. [CrossRef]
15. Serrano-Cinca, C.; Gutiérrez-Nieto, B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decis. Support Syst.* **2016**, *89*, 113–122. [CrossRef]
16. Robinson, D.; Yu, H. *Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace* (October 2014). Available online: https://www.upturn.org/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf (accessed on 2 January 2021).
17. Bjorkegren, D.; Grissen, D. Behavior revealed in mobile phone usage predicts loan repayment. *SSRN Electron. J.* **2018**. [CrossRef]
18. 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]
19. Simumba, N.; Okami, S.; Kodaka, A.; Kohtake, N. Alternative scoring factors using non-financial data for credit decisions in agricultural microfinance. In Proceedings of the IEEE International Symposium on Systems Engineering, Rome, Italy, 1–3 October 2018.
20. Ntwiga, B.D.; Weke, P. Credit scoring for M-Shwari using hidden markov model. *Eur. Sci. J.* **2013**, *12*, 15.
21. Lohokare, J.; Dani, R.; Sontakke, S. Automated data collection for credit score calculation based on financial transactions and social media. In Proceedings of the 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), Yashada, Pune, 3–5 February 2017; pp. 134–138.
22. Li, Y.; Wang, X.; Djehiche, B.; Hu, X. Credit scoring by incorporating dynamic networked information. *Eur. J. Oper. Res.* **2020**, *286*, 1103–1112. [CrossRef]
23. Djeundje, V.B.; Crook, J.; Calabrese, R.; Hamid, M. Enhancing credit scoring with alternative data. *Expert Syst. Appl.* **2021**, *163*, 113766. [CrossRef]
24. Blanco, A.; Pino-Mejías, R.; Lara, J.; Rayo, S. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Syst. Appl.* **2013**, *40*, 356–364. [CrossRef]
25. Djeundje, V.B.; Crook, J. Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *Eur. J. Oper. Res.* **2018**, *2*, 697–709. [CrossRef]
26. Fernandes, G.B.; Artes, R. Spatial dependence in credit risk and its improvement in credit scoring. *Eur. J. Oper. Res.* **2016**, *249*, 517–524. [CrossRef]
27. Sowers, D.C.; Durand, D. Risk elements in consumer instalment financing. *J. Mark.* **1942**, *6*, 407. [CrossRef]
28. Ben-Hur, A.W.J. A.W.J. A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences*; Humana Press: Totowa, NJ, USA, 2010; pp. 223–239.
29. Haltuf, M. *Support Vector Machines for Credit Scoring*; University of Economics in Prague, Faculty of Finance: Prague, Czech Republic, 2014.
30. Zheng, R.Y. Neural Networks. In *Data Mining Techniques for the Life Sciences*; Human Press: Totowa, NJ, USA, 2010; pp. 197–222.
31. Breiman, L. Bagging predictors. *Machin. Learn.* **1996**, *24*, 123–140. [CrossRef]
32. Schapire, R. The strength of weak learnability. *Mach. Learn.* **1989**, *5*, 197–227. [CrossRef]
33. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Machine Learning: Thirteenth International Conference, Bari, Italy, 3–6 July 1996; pp. 148–156.
34. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
35. Economic Research and Regional Cooperation Department. *Basic Statistics 2017*; Asian Development Bank (ADB): Manila, Philippines, 2017.
36. National Institute of Statistics, Ministry of Planning. *Census of Agriculture in Cambodia 2013*; Ministry of Agriculture, Forestry and Fisheries: Phnom Penh, Cambodia, 2015.
37. Food and Agriculture Organisation of the United Nations. *FAO Statistical Pocketbook World Food and Agriculture*; Food and Agriculture Organisation of the United Nations: Rome, Italy, 2015.
38. World Bank. *Global Financial Inclusion Statistics*; World Bank Group: Washington, DC, USA, 2014.
39. East-West Management Institute. Open Development Cambodia. Available online: https://opendevelopmentcambodia.net/maps/ (accessed on 10 November 2018).
40. Skakun, S.; Kussul, N.; Shelestov, A.; Kussul, O. The use of satellite data for agriculture drought risk quantification in Ukraine. *Geomat. Nat. Hazards Risk* **2015**, *7*, 901–917. [CrossRef]
41. Jarvis, A.; Reuter, H.I.; Nelson, A.; Guevara, E. Hole-filled SRTM for the globe Version 4. Available from the CGIAR-CSI SRTM 90 m Database. Available online: http://srtm.csi.cgiar.org (accessed on 24 November 2019).
42. Schaaf, C.W.Z. MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global-500 m V006 [Data set]. NASA EOSDIS Land Processes DAAC. Available online: https://lpdaac.usgs.gov/products/mcd43a4v006/ (accessed on 24 November 2019).

43. Wan, Z.H.S.H.G. MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Available online: https://lpdaac.usgs.gov/products/mod11a1v006/ (accessed on 24 November 2019).
44. Vermote, E.W.R. MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1 km and 500 m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Available online: https://lpdaac.usgs.gov/products/mod09gav006/ (accessed on 24 November 2019).
45. Demsar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
46. Nemenyi, P.B. *Distribution-Free Multiple Comparisons*; Princeton University: Princeton, NJ, USA, 1963.