mathematics

MDPI

Article

Consumer Default Risk Portrait: An Intelligent Management Framework of Online Consumer Credit Default Risk

Miao Zhu [1,2], Ben-Chang Shia [3,4], Meng Su [5,6,7] and Jialin Liu [5,6,7,*]

1 School of Statistics, Huaqiao University, Xiamen 361021, China; zhumiao@hqu.edu.cn
2 Institute of Quantitative Economics, Huaqiao University, Xiamen 361021, China
3 AI Development Center, Fu Jen Catholic University, New Taipei City 242, Taiwan; 025674@mail.fju.edu.tw
4 Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City 242, Taiwan
5 School of Medicine, Xiamen University, Xiamen 361005, China; sumeng@stu.xmu.edu.cn
6 National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China
7 Data Mining Research Center, Xiamen University, Xiamen 361005, China
* Correspondence: zyxzjs107@126.com

Abstract: Online consumer credit services play a vital role in the contemporary consumer market. To foster their sustainable development, it is essential to establish and strengthen the relevant risk management mechanism. This study proposes an intelligent management framework called the consumer default risk portrait (CDRP) to mitigate the default risks associated with online consumer loans. The CDRP framework combines traditional credit information and Internet platform data to depict the portrait of consumer default risks. It consists of four modules: addressing data imbalances, establishing relationships between user characteristics and the default risk, analyzing the influence of different variables on default, and ultimately presenting personalized consumer profiles. Empirical findings reveal that "Repayment Periods", "Loan Amount", and "Debt to Income Type" emerge as the three variables with the most significant impact on default. "Re-payment Periods" and "Debt to Income Type" demonstrate a positive correlation with default probability, while a lower "Loan Amount" corresponds to a higher likelihood of default. Additionally, our verification highlights that the significance of variables varies across different samples, thereby presenting a personalized portrait from a single sample. In conclusion, the proposed framework provides valuable suggestions and insights for financial institutions and Internet platform managers to improve the market environment of online consumer credit services.

Keywords: online consumer credit; consumer default risk management; user portrait; Shapley Additive Explanations; machine learning

MSC: 68T09

Citation: Zhu, M.; Shia, B.-C.; Su, M.; Liu, J. Consumer Default Risk Portrait: An Intelligent Management Framework of Online Consumer Credit Default Risk. *Mathematics* **2024**, *12*, 1582. https://doi.org/10.3390/math12101582

Academic Editors: Xinwei Cao, Tran Thu Ha, Dunhui Xiao, Vasilios N. Katsikis, Ameer Hamza Khan and Shuai Li

Received: 26 March 2024
Revised: 28 April 2024
Accepted: 15 May 2024
Published: 18 May 2024

1. Introduction

Consumer credit enables individuals to pay for goods or services in installments over some time, leveraging online platforms. As a crucial financial tool, online consumer credit service has revolutionized the conventional credit system by overcoming geographical and temporal constraints [1]. For merchants, the "buy-now-pay-later" scheme could attract more consumers, thereby improving sales and customer conversion. By providing online consumer credit services, e-commerce platforms can also increase user stickiness and improve platform activity and user experience. In short, the promotion of online consumer credit can expand internal demand and encourage consumption, thus contributing to economic growth [2]. Presently, leading e-commerce platforms have robustly developed and extensively promoted this service.

The role of the online consumer credit service in the consumer market is undeniable. However, it also encounters the challenge of evaluating consumer credit risks [3]. The

continued growth of online consumer credit combined with the fragility of the financial system may lead to excessive consumer debt, which is not conducive to the stability of the financial market and economic growth [4]. For example, in the year 2008, excessive credit expansion in the United States led to sharp fluctuations in the housing market, triggering the subprime crisis [5]. The crisis spread to the banking systems of various countries, with unprecedented consequences for the global economy and greatly eroded consumer confidence [6]. Online consumer credit services have characteristics such as small loan amounts, dispersed borrowers, lack of collateral, and easy market entry [7]. These factors make the risks associated with such services more pronounced. Financial institutions, including commercial banks, as well as e-commerce platforms, like Internet companies, need to establish an enhanced mechanism for effectively managing the risks associated with consumer loan defaults. It is essential for achieving sustainable and sound development.

In recent years, an increasing number of studies have been dedicated to managing the default risk associated with online consumer loans [8]. However, most of them concentrate on developing traditional credit scoring models [9–12]. Risk management is approached as a straightforward classification problem [13]. Although regularly updated classification algorithms can enhance the predictive performance of the model, the emphasis on achieving higher prediction accuracy hinders model interpretability, making it challenging to gain the trust of managers [14]. Hence, the utilization of these AI programs is severely constrained, with some cases where they are not utilized at all [15]. Moreover, existing models only target specific data, and there may be conscious changes in the data due to the smart behavior of fraudsters, which affects the level of feature contribution in credit scores. Current models are not intelligent enough to adapt to the evolving credit data [16,17]. Importantly, when default prediction is solely based on the overall data, managers are unable to discern sample variations in the prediction process and receive limited constructive feedback. As a result, the core requirements of managers go unfulfilled. This paper aims to depict the default risk portrait of consumers to address the aforementioned challenges. Leveraging the user portrait theory, this study exceeds the scope of developing prediction models for specific datasets or platforms. It focuses on constructing a management framework with broader applicability and potential for dissemination. It assists financial institutions in devising personalized marketing strategies, implementing risk management measures, and making informed loan decisions to enhance user experience and attain effective risk control.

With the increasing demand for personalized services across multiple industries, user portrait technology has emerged as a prominent area of interest [18]. The concept of user portraits was initially introduced by Alan Cooper [19], considered the pioneer of interaction design. User portraits are created by leveraging extensive user behavior data to generate a comprehensive and detailed description of their characteristics, transforming the data into a valuable tool for problem-solving purposes [20]. The wide application of user portraits has also made positive contributions in intrusion detection [21], personalized recommendation [22], medical [18], public opinion analysis [23], and other fields. Conventional mainstream user profiling methods encompass collaborative filtering, content-based, and knowledge-based approaches [24]. In the domain of credit risk, the utilization of user profiling is primarily centered on machine learning methods with black-box attributes [25–28]. However, the current research is limited to the credit rating level.

The Shapley value, derived from the cooperative game theory by Shapley in 1953, is a method used to quantify the contributions of members to the cooperative benefits in cooperative games [29]. Based on Shapley values, Lundberg and Lee proposed the Shapley Additive Explanations (SHAP) method to elucidate individual contributions during the prediction process [30]. This model-independent postmortem approach considers variables as "players" in the game and can calculate the importance of these variables with theoretical support. This method measures the contribution level of each feature to the prediction outcome, achieving dimensional consistency in the feature space [31]. This eliminates dimensional discrepancies caused by varying characteristics and value ranges of features, thus avoiding deviation problems. Additionally, calculating the average marginal contribu-

tion of variables enables machine learning to capture complex nonlinear relationships and analyze data heterogeneity [32]. Researchers can gain a comprehensive understanding of dataset features and differences, offering valuable insights for subsequent data modeling, statistical inference, and decision-making processes [33,34].

This paper has developed an intelligent management framework called the consumer default risk portrait (CDRP) for controlling the risk of default in online consumer loans. The credit data exhibits a prevalent category imbalance issue, which can significantly impact subsequent analyses [35]. Thus, we propose the utilization of the near-miss undersampling method to tackle the challenge of handling highly unbalanced data. Then, we establish the relationship between user characteristics and default risk using machine learning techniques, leading to the development of a prediction model. In order to break the black box phenomenon in the application of machine learning to forecast, we utilize the SHAP method to analyze the overall contribution and local influence mechanisms. By considering various characteristics as dimensions, this approach facilitates a comprehensive demonstration of the role played by each factor in predicting default risk. Finally, we analyze the personalized portraits of individual users to examine the heterogeneity within the sample. Our approach has broad applicability to various complex machine learning prediction methods, allowing us to achieve a balance between accuracy and transparency. The CDRP framework offers managers valuable and novel insights, providing fresh perspectives for decision making.

Our research makes several key contributions. Firstly, we primarily present the construction of a consumer default risk portrait (CDRP) intelligent framework capable of generating comprehensive risk profiles for consumers. From the technical perspective, the proposed framework could integrate advanced machine learning algorithms and interpretable models to analyze and interpret consumer data effectively. As for managers seeking to mitigate default risk among users, this framework serves a holistic view and risk portrait of consumer default risks. Secondly, we delve into elucidating the specific contributions of individual features to the risk evaluation process in consumer loan scenarios. Based on the heightened importance of individual heterogeneity in effectively managing consumer default risk, we employ the SHAP method to address the challenge of personalized user "black box" issues within default risk analysis. Although the existing literature has extensively focused on managing online consumer default risks, it has largely neglected the exploration of individual heterogeneity. Our study pays meticulous attention to each sample, uncovering the nuanced influence of every feature on the predicted results. Consequently, our research offers novel insights that equip managers with a fresh vantage point for strategic decision making in risk management practices. Thirdly, our proposed framework demonstrates a high degree of flexibility. The utilization of the SHAP method within the CDRP framework is model agnostic, ensuring seamless applicability across diverse machine learning models without encountering compatibility concerns. Moreover, this integration harmoniously interfaces with sophisticated algorithms, facilitating enhanced analytical depth and nuanced interpretations.

The rest of this study is organized as follows. Section 2 describes relevant background research on online consumer credit default risk. The data and methodology used in this study are demonstrated in Section 3. The results of our analyses are shown in Section 4. Finally, Section 5 concludes this study and provides further discussion of this topic.

2. Literature Review

Previous research on online consumer loan default risks mainly focuses on credit scoring method updates and constructing evaluation index systems as a classification task. The advent of artificial intelligence in recent years has presented numerous prospects for the financial sector [36]. Machine learning methods have significantly contributed to enhancing the accuracy of default risk prediction and validating new assessment indicators. For instance, Li et al. [37] introduced transfer learning into the consumer loan risk assessment model, demonstrating higher AUC compared to models without transfer learning.

Papouskova et al. [38] proposed a two-stage credit risk model combining unbalanced ensemble learning and regression integration, highlighting the efficacy of heterogeneous ensemble methods in modeling consumer credit risks. Costa et al. [39] utilized a logistic regression model to evaluate the default risk using credit score data from a Portuguese financial institution. Hou et al. [40] extracted characteristic data from credit data using multi-granularity modules and employed a combination of random forest and gradient boosting decision trees (GBDTs) to address variance and bias techniques. Wen et al. [41] confirmed the significance of consumption information in evaluation indicators using logistic regression and the light gradient boosting machine (LightGBM) algorithm. Additionally, researchers have integrated professional knowledge with feature selection techniques and applied machine learning methods like the genetic algorithm (GA) and K-nearest neighbors (KNN) algorithm to enhance result performance [42]. These models exhibit strong performance on specific datasets, forming a robust basis for the research conducted in this study. However, fraudsters tend to exhibit intelligence in credit data. They often employ ingenious techniques and strategies to evade detection and discovery [43]. Therefore, in addition to updating forecasting models, it is essential to develop a management tool that can be extensively and continually utilized by managers.

Despite the significant research advancements of machine learning in credit scoring, it often faces skepticism due to its "black box" nature. Regulators demand transparency and auditability in credit scoring models. However, excessively complex machine learning models pose challenges in explaining the approval process to both customers and regulators [44]. Consequently, merely enhancing classification accuracy can no longer meet the management requirements for default risks. In recent years, there has been a rise in the utilization of interpretable machine learning methods in the research on managing default risk in online consumer lending. Xia et al. [45] demonstrated the significance of the external credit rating variable in predicting default loans using interpretable machine learning techniques. Zhang et al. [46] calculated the Shapley value of the feature and came to the conclusion that the "overdue" feature contributed the most to the prediction, that is, the longer the overdue time, the more likely the borrower will not repay the loan and become a defaulter. Zhou et al. [47] employed the Shapley value to evaluate the scheme's effectiveness and discovered that an increase in the number of early morning phone calls raised the risk of default by 13% for consumers predicted to be defaulters. However, the research in this direction remains limited in depth, preventing us from comprehending the specific contribution of each characteristic to individual consumers.

To sum up, the current research on online consumer loans remains inadequate. Firstly, most of them are limited to selecting credit evaluation indicators and forecasting methods without considering the diverse nature of credit data in different consumer loan scenarios. However, due to the variability of credit data, adapting to different consumer loan data becomes challenging. Hence, there is a pressing need for an intelligent management tool for online consumer loans that can be widely adopted by managers. Secondly, an opaque machine learning model fails to meet the requirements for credit scoring, resulting in hesitation from both managers and consumers regarding its reliability. Lastly, the current focus of interpretability research primarily revolves around evaluating macroscopic features, disregarding the unique performance of different features in diverse samples. Therefore, the significance of constructing personalized consumer portraits cannot be overlooked. Leveraging these insights, this study introduces a consumer default risk portrait (CDRP) framework designed for versatile application across diverse datasets in online consumer lending contexts, offering both interpretability and personalized features.

3. Data and Methodology

3.1. Data

The data utilized in this study are sourced from a collaborative loan venture established between a domestic commercial bank and an e-commerce platform. Users on the platform can browse and purchase products offered by different dealers. Moreover, they

are provided with the option to make installment payments and submit their application information online at the time of order settlement. Initially, the e-commerce platform employs its internally developed risk control system to assess and select user applications. Subsequently, the approved application form is forwarded to the commercial bank for a secondary review. After the loan application is approved, the commercial bank disburses funds to the dealer according to a predefined proportion agreed upon with the platform. During the post-loan management phase, the platform undertakes the task of initiating communication with users, issuing reminders to customers for timely deposits of monthly payments into the designated account, and transmitting payment instructions to the bank. If users are unable to fully repay by the agreed-upon date, additional daily penalty interest will be charged until both the principal amount and interest are settled.

3.2. Variables Description

In this study, a total of 23 independent variables and 1 dependent variable were included, as shown in Table 1. The dependent variable's value is determined based on whether the user has experienced overdue payments. A value of 1 indicates the user has experienced overdue payments, whereas a value of 0 signifies the user has not experienced overdue payments. The dataset comprises 79,656 samples, with 6.5% exhibiting overdue payments. Based on the recommendations provided by experts in consumer credit risk management at banks, we gathered 23 variables categorized into three distinct groups: basic information, traditional credit investigation, and platform information.

Table 1. Descriptive statistics of variable.

Variable	Definition	Mean	SD	Min	Max
Dependent Variable					
Overdue	Whether users have exceeded the repayment date for when repaying	0.0654	0.2472	0	1
Basic information					
Gender	Gender of the user (0 = female, 1 = male)	0.6003	0.4898	0	1
Age	Age of the user	30.0244	5.4145	23	55
Marital status	Marriage probability rating of the user	1.6602	0.6981	1	3
Working industry	Working industry of the user	7.1753	2.5073	1	9
Traditional credit investigation					
Revolving line	The average value of cc quota	1.5645	0.7032	1	5
Revolving line utilization	The average level of cc quota utilization rate	1.8724	1.0709	1	4
Delinquency history	The number of uncleared LN	2.2005	0.8442	1	4
Past loan number	Number of LN approval cause queries in the last month	1.1020	0.3236	1	3
Lending time	Maximum mob from card issuance (all credit cards) time to report generation time	2.9178	1.2729	1	4
Lending organization number	The number of card issuers of CC	1.8186	0.7623	1	5
Lending number	Number of CC cards	2.5105	1.1099	1	5
Credit grade	whether the Bank's credit score is greater than 700	0.0062	0.0784	0	1
Repayment periods	Duration of loan	209.8841	133.8044	29	396
Loan amount	The amount of borrowing	3277.0772	4769.4677	99	40,000
Debt to income type	Day rate	0.0154	0.0031	0.0083	0.0195
platform information					
Social networking	Number of cities where users have logged in in the past 90 days	1.8111	0.7328	1	3
Number of Online Search Scenarios	The number of channels that users access in the platform.	2.3746	0.7181	1	3
Number of Days in Online Search	The number of days to access the platform	2.7432	0.4938	1	3
Number of Days in Online Takeaway Search	The number of days the user has accessed the Takeaway channel	1.9409	0.8823	1	3
Number of Days in Online Groupon Search	The number of days a user has visited a Groupon channel	2.3569	0.7270	1	3
Amount of Online transactions	The number of successful transactions made by the user	2.7925	0.5134	1	3
Number of Online Transaction Scenarios	The service types provided by the platform	1.8694	0.3369	1	2
Number of Online Transaction	The frequency of users' consumption on the platform in the past 365 days	1.8773	0.3281	1	2

The basic information comprises "Gender", "Age", "Marital status", and "Working industry". "Gender" is categorized as either male or female. "Age" is derived from the birth year indicated on the ID card. "Marital status" is classified as either married or unmarried. "Working industry" is categorized into 10 specific types, including agriculture, forestry, fishery, animal husbandry, manufacturing, social service industry, and so on.

The information of the traditional credit investigation group comes from the user history credit investigation issued by the credit investigation system of the People's Bank of China, which contains 11 variables. "Revolving Line" means that the maximum account limit is set at the same card issuer with the same status deadline, and the average is calculated. "Revolving Line Utilization" refers to the ratio between the total amount used and the total maximum credit line for a credit card under normal conditions. "Delinquency History" is the number of loans that go unpaid. "Past Research Number" refers to the number of inquiries made by non-platform companies for loan approval in the last month. "Lending time" is the maximum mob of all credit card cut-off data generation time. "Organization Number" indicates the number of credit card issuers. "Card Number" indicates the number of credit cards held by the user. "Credit Grade" is a categorical indicator that determines whether the Bank's credit score is greater than 700. "Repayment Periods" are calculated by the number of days from the borrowing date to the maturity date. The "Loan Amount" is the total amount borrowed by a consumer in the last 30 days. "Debt to Income Type" refers to the ratio of a consumer's income to its debt.

The platform information group is provided by the platform and comes from the user's activity process such as commodity browsing and trading on the platform in the past 90 days, which contains a total of eight variables. "Social Networking" refers to the number of active cities, active means the day in the city login behavior. "Number of Online Search Scenarios" is the number of channels that users access in the platform. "Number of Days in Online Search" indicates the number of days to access the platform. "Number of Days in Online Takeaway Search" indicates the number of days the user has accessed the Takeaway channel. "Number of Days in Online Groupon Search" indicates the number of days a user has visited a Groupon channel. "Amount of Online transactions" refers to the number of successful transactions made by the user. According to the service types provided by the platform, "Number of Online Transaction Scenarios" is divided into four categories, covered by user consumption include dining, take-out, travel, and others. based on the frequency of users' consumption on the platform in the past 365 days, "Number of Online Transaction" divided into five levels.

3.3. Methodology

The development of the CDRP revolves around online consumer default risk portrait, as illustrated in Figure 1. The framework relies on traditional credit information data, basic user information, and Internet platform data as fundamental components for managing consumer credit risks. It comprises four distinct modules. The first module, data processing, tackles the challenge of data imbalance. Numerous methods can be utilized in this module to address the corresponding issue of unbalanced data. This section focuses on elucidating the implementation of the near-miss method as an example. In the following module, model-building machine-learning techniques are employed to establish the relationship between user characteristics and default risks, thereby aiding in the development of predictive models. In the third module, feature analysis, the SHAP method is used to assess both the global and local contributions of different features. Finally, the user portrait module allows for the creation of personalized profiles for individual users, facilitating the analysis of sample heterogeneity.

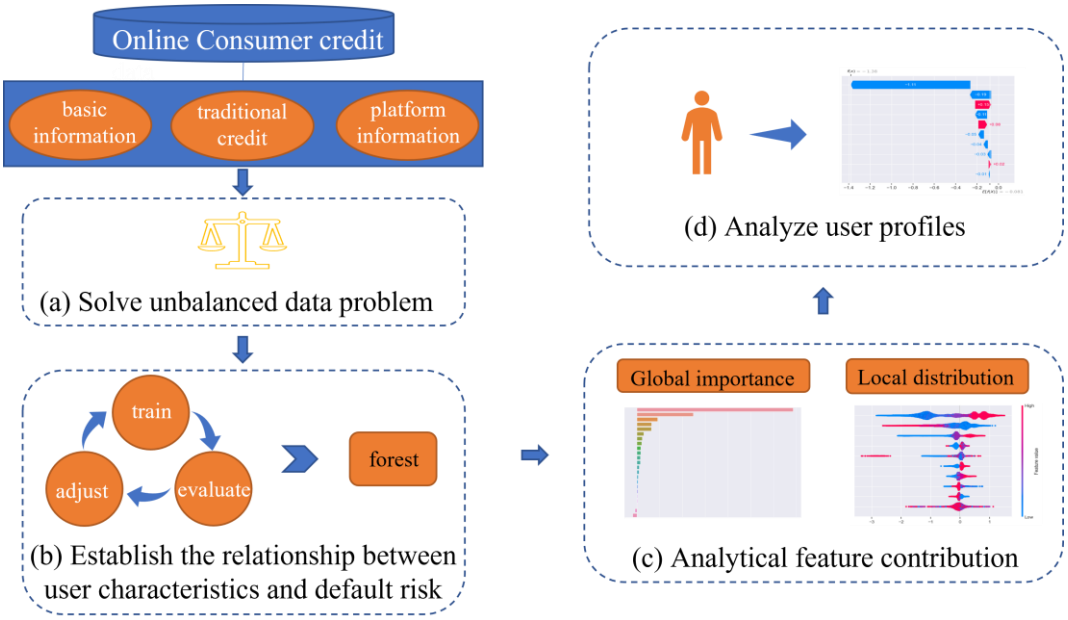

Figure 1. The CDRP framework. (**a**) Solve unbalanced data problem. We could employ various approaches within this module to tackle the challenge of imbalanced data. (**b**) Establish the relationship between user characteristics and default risk. Machine-learning techniques are employed to develop the predictive models. (**c**) Analytical feature contribution. The SHAP method is employed to evaluate both the global and local contributions of all features. (**d**) Analyze user profiles. Create user profiles based on the individual feature contributions of distinct users.

3.3.1. Data Processing

The CDRP framework employs diverse strategies to address imbalanced data based on unique data attributes. Imbalanced datasets exhibit notably larger sample sizes within one category compared to others, potentially resulting in model oversensitivity to dominant classes and inadequate categorization of minority classes. In the context of default risk prediction, imbalanced data impedes timely default identification. This research evaluates the effectiveness of the data point methodology by utilizing an undersampling approach based on the near-miss method [48]. This technique establishes an equilibrium in class distributions, thereby enhancing classification accuracy within highly imbalanced datasets [49]. The near-miss method finds extensive applications in managing default risks within unbalanced datasets [48]. Serving as a tool for addressing imbalances, near-miss concentrates on prototype selection by opting for the most representative majority class samples during training, thereby mitigating the information loss often associated with random undersampling methods [35].

Near-miss aims to choose majority class samples that are nearest to minority class samples to enhance the clarity of the decision boundary between the two classes. The distance calculation formula is as follows:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

This method computes the average distance between each majority sample and the minority sample, then selects the majority sample with the shortest distance. Thus, nea-miss reduces the quantity of majority class samples while preserving the distinctiveness of decision boundaries.

NearMiss-1 chooses the majority sample closest on average to the nearest k-minority sample, while NearMiss-2 picks the majority sample with the closest mean distance to the farthest k-minority sample. Near Algorithm 1: Miss-3 involves selecting k closest majority

samples per minority sample to enclose each minority sample within the majority samples. While NearMiss-1 and NearMiss-2 require calculating the k-nearest neighbors for every multi-class sample, this incurs substantial computational costs. However, NearMiss-1 is vulnerable to the impact of outliers. We opt for utilizing NearMiss-3 to handle online consumer credit default risk data as a foundation for further analysis.

Algorithm 1: NearMiss-3

Input: non-defaulted samples set, default sample set, number of neighbors k.
1. Initializes an empty set of undersampled samples named undersampled_samples.
2. For each default sample:
    2.1 Calculate the distance between this defaulted sample and all non-defaulted samples, then select k nearest majority class samples, named nearest_majority_samples.
    2.2 Add the default sample to undersampled_samples.
    2.3 Add samples from nearest_majority_samples to undersampled_samples.
3. Return undersampled_samples as the sample set after undersampling.

3.3.2. Model Building

The pre-processed data can be used to build predictive models using any complex machine learning method to achieve the desired accuracy. This section utilized the LightGBM (LGB) algorithm as an example. LGB algorithm is a fast, distributed, and high-performance gradient lifting framework-based decision tree algorithm [50]. It is widely used in sorting, classification, regression, and many other machine learning tasks. LGB is mainly developed based on histogram algorithms. The basic idea of histogram algorithms is to first discretize continuous floating-point eigenvalues into k integers to build a histogram with a width of k. k integers as an index accumulate the statistics of the entire data in the histogram. After traversing all the data, the optimal segmentation point is found according to the discrete value of the histogram. LGB is the optimization of error acceleration on this basis. The histogram of the node can be obtained by the difference between the father node and the brother node, which can greatly improve the running speed. For example, the leaf node with a small amount of computation is calculated first, and the node with a large histogram is obtained by using the histogram difference so that the calculation cost of each node histogram can be reduced. In general, the LGB algorithm can improve the training speed and reduce the memory consumption while ensuring the accuracy by optimizing the growth strategy of decision tree and applying histogram algorithm, so that it can process large-scale data and realize fast training.

In particular, let S represent the training set space. The variance gains of splitting feature j at point d for a fixed node is defined as [51]

$$V_{j/S}(d) = \frac{1}{n_S}\left(\frac{\left(\sum_{\{x_i \in S : x_{ij} \le d\}} g_i\right)^2}{n^j_{l/S}(d)} + \frac{\left(\sum_{\{x_i \in S : x_{ij} > d\}} g_i\right)^2}{n^j_{r/S}(d)}\right), \tag{2}$$

where $n_S = \sum I[x_i \in S]$, $n^j_{l/S}(d) = \sum I\left[x_i \in S : x_{ij} \le d\right]$ and $n^j_{r/S}(d) = \sum I\left[x_i \in S : x_{ij} > d\right]$.

Let a be the sampling ratio of large gradient data and b be the sampling ratio of small gradient data. Based on this, we can determine the frequency at which each sample is sampled during the training process. Divide the instance according to the following equation:

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n^j_r(d)}\right) \tag{3}$$

where $A_l = \{x_i \in A : x_{ij} \leq d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \leq d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$. LightGBM significantly reduces computational expenses by utilizing the estimated $\widetilde{V}_j(d)$ over a smaller instance subset to determine the split point.

3.3.3. Feature Analysis

Following the prediction results generated by a suitable model, our framework employs the concept of Shapley values to address the "black box problem" inherent in machine learning operations, enabling a meticulous examination of default risk factors. The Shapley value is a technique developed by Shapley in 1953, rooted in cooperative game theory, designed to measure individual contributions to group benefits within cooperative games [29]. Lundberg and Lee proposed the Shapley Additive Explanations (SHAP) method on the basis of the Shapley values to explain the individual contribution to the prediction process. This model-independent postmortem approach treats variables as "players" in the game and can calculate the importance of these variables with theoretical support [30]. The SHAP method measures the contribution degree of each feature to the prediction result and achieves dimensional consistency in the feature space. This eliminates dimensional differences due to different ranges of features and eigenvalues, thus avoiding bias problems. In addition, calculating the average marginal contribution of variables enables machine learning to capture complex non-linear relationships and analyze data heterogeneity.

Suppose there are p features in the online consumer credit default risk prediction model, and for the $j-$th feature, its contribution to predicting default is a weighted sum over all possible combinations of feature values:

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right) \tag{4}$$

where S refers to all subsets of features in the prediction model, x is the vector that needs to interpret the eigenvalues of the sample, and *val* refers to the prediction results of the model under the eigenvalues in S.

However, online consumer credit data usually contains more features, so with the increase in the number of features, the number of feature subsets will increase exponentially. Therefore, this patent uses the approximate value of Monte Carlo sampling to approximate the contribution value of the $j-$th feature to the prediction of default by calculating the average marginal effect:

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}\left(x_{+j}^m\right) - \hat{f}\left(x_{-j}^m\right)\right) \tag{5}$$

where $\hat{f}\left(x_{+j}^m\right)$ is the prediction of x, maintaining the value of feature j, and other feature values that do not belong to S are replaced by the feature values of the random data points. $\hat{f}\left(x_{-j}^m\right)$ means that all eigenvalues that do not belong to S are replaced by eigenvalues of random data points.

To express the interpretation of Shapley values as a linear model, Equation (5) can be simplified to

$$g\left(z'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \tag{6}$$

where g is the explanatory model, $z' \in \{0,1\}^M$ is the set of features, M represents the maximum number of sets, and $\phi_j \in \mathbb{R}$ is the Shapley value of the $j-$th feature.

The SHAP method has local accuracy, missingness, and consistency.

$$f(x) = g\left(x'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j x_j' \tag{7}$$

If $\phi_0 = E_X(\hat{f}(x))$, and let all features exist, then

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_X(\hat{f}(X)) + \sum_{j=1}^{M} \phi_j \tag{8}$$

Missingness indicates the reduction to zero of the missing features.

$$x'_j = 0 \Rightarrow \phi_j = 0 \tag{9}$$

Missing features can have arbitrary Shapley values without compromising local accuracy. Let $f_x(z') = f(h_x(z'))$, and $z_{\setminus j'}$ denote $z'_j = 0$. For any f and f',

$$f'_x(z') - f'_x\left(z'_{\setminus j}\right) \geq f_x(z') - f_x\left(z'_{\setminus j}\right) \tag{10}$$

Iterate this process for each feature to calculate the Shapley values across all features. During global analysis, a bar chart aids in comprehending the contribution and impact of each feature on default prediction, facilitating the determination of feature importance rankings. In the context of local analysis, this methodology allows for the examination of how alterations in various feature values influence predicted outcomes, thereby revealing correlations between features and predictions.

The SHAP methods offer a lucid and intuitive mechanism to elucidate the impact of individual features on the model's predictions while quantitatively evaluating the significance of each feature. This affords valuable insights into the predictive mechanisms of machine learning models and furnishes dependable backing and direction for business decision-making processes.

3.3.4. User Portrait

Because in global and local analysis, we can only discuss the overall impact of features. However, it is clear that each sample has its specificity, and the average contribution of features hides the heterogeneous effects of features. Therefore, the CDRP framework paints user portraits for a single set of data through the user portrait theory. We combine user analysis with black-box-specific machine learning methods to tap into sample heterogeneity. Based on Shapley values, the CDRP framework helps to interpret the model's predictions for this user by means of a waterfall diagram, enhancing the trust and interpretability of the model.

4. Results and Analysis

4.1. Evaluation Criteria

To compare the performance of different models and select the model with the best performance for further explanation, we used five common indicators to evaluate the predictive power of the model [52]. Prior to discussing this topic, let us present the relevant concepts. The confusion matrix [53] is one of the commonly used indicators to evaluate machine-learning classification models. Based on the confusion matrix, classification results can be divided into true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP represents the number of samples correctly classified as positive, that is, the number of samples that are actually positive and are classified as positive by the classifier; FP represents the number of samples misclassified as positive, that is, the number of samples that are actually negative but are classified as positive by the classifier; FN represents the number of samples incorrectly classified as negative, that is, the number of samples that the classifier is actually positive but classified as negative; TN represents the number of samples correctly classified as negative, i.e., the number of samples that are actually negative and are classified as negative by the classifier. As a result, the five evaluation indicators used in this paper are Accuracy, Specificity, Sensitivity, G-means, and AUC.

(1) Accuracy, the prevalent evaluation metric, refers to the proportion of correctly classified samples among the total samples. Typically, in scenarios with balanced sample sizes, higher accuracy signifies superior model performance [54]. However, given the focus on unbalanced data in this study, accuracy serves merely as a point of reference.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{11}$$

(2) Specificity in binary classification denotes the model's capacity to accurately forecast negative instances. It gauges the model's ability in correctly identifying true negative cases as negatives [54].

$$Specificity = TN/(TN + FP) \tag{12}$$

(3) Sensitivity, often referred to as Recall, represents the ratio of correctly identified positive instances by the model among all actual positive samples [54].

$$Sensitivity = TP/(TP + FN) \tag{13}$$

(4) The G-means metric, derived from the square root of the product of Precision and Recall, serves as a comprehensive indicator to assess both model precision and recall. This measure proves particularly valuable in scenarios involving imbalanced datasets. Elevated values of the geometric mean signify enhanced overall model performance [55].

$$G - means = \sqrt{Accuracy \times Specificity} \tag{14}$$

(5) AUC depends on the area under the ROC curve. The ROC curve was mapped with Sensitivity as the ordinate and 1-Specificity as the abscissa. Simply put, the larger the AUC value, the better the model performs [56].

4.2. Prediction Performance

The models underwent rigorous evaluation through the five-fold cross-validation method to ensure the validity of the experiment. Ten benchmark models commonly utilized in credit scoring were employed: linear discriminant analysis (LDA), logistic regression (LOG), Gaussian Naive Bayes (NB), K-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), stochastic gradient descent (SGD), random forest (RF), XGBoost (XGB), and LightGBM (LGB). The selection of these ten benchmark models was predicated on their demonstration of diverse learning methods and technologies: LDA is situated within statistical classification methodologies [57]. LOG represents a prevalent linear classification algorithm [58]. NB stands as an uncomplicated yet effective probabilistic model hinged on the Bayesian theorem and assumptions of feature conditional independence [59]. KNN operates as an instance-based learning mechanism that classifies by gauging distances between distinct samples [60]. DT leverages tree structures to make decisions, functioning as a non-parametric supervised learning tool [61]. SVM serves as a binary classification model, achieving classification by identifying the maximum margin hyperplane [62]. SGD embodies an optimization method [63]. RF is classified under ensemble learning mechanisms [64]. XGB signifies an improved rendition of gradient boosting decision trees [65]. LGB embodies an ensemble learning approach grounded in the gradient boosting technique [51]. These models encompass both conventional statistical learning techniques and sophisticated machine learning models, spanning from basic linear associations to intricate nonlinear representations and encompassing individual models as well as ensemble models. The nine machine learning models cover a wide array of prevalent machine learning methodologies, it contributes to facilitating a thorough assessment of the performance and applicability of manifold machine learning models within the realm of online consumer default risk prediction scenarios. Simultaneously, it exemplifies the multifaceted applicability of the CDRP framework. Moreover, the selection of these models augments the assurance of comprehensive evaluation and reliability, fosters deeper insights, and provides invaluable benchmarks for future research initiatives.

The analysis was carried out using Python 3.6.5. Table 2 shows the predictive performance of each model, with the best results highlighted in bold.

Table 2. The prediction performance of different machine learning classifier.

Model	Accuracy	Specificity	Sensitivity	G-Means	AUC
LGB	0.71	0.71	0.68	**0.70**	**0.70**
LDA	0.66	0.66	0.70	0.68	0.68
Logit	0.67	0.66	0.70	0.68	0.68
NB	0.63	0.63	0.73	0.68	0.68
KNN	0.74	0.76	0.43	0.57	0.60
DT	0.88	0.94	0.13	0.35	0.53
SVM	0.70	0.71	0.63	0.66	0.67
SGB	0.19	0.14	**0.95**	0.36	0.55
RF	**0.93**	**0.99**	0.01	0.08	0.50
XGB	0.78	0.79	0.55	0.66	0.67

The results indicate a high Accuracy of 0.93 for the RF classifier. However, the Sensitivity and AUC performances are significantly poor, at 0.01 and 0.08, respectively, suggesting a substantial impact from imbalanced data categories. The LGB model exhibited Accuracy, Specificity, Sensitivity, G-means, and AUC values of 0.71, 0.71, 0.68, 0.70, and 0.70, respectively. The LGB model stands out as the most robust and consistent among all considered models. Consequently, the LGB model was selected for further analysis of variable contribution.

4.3. Feature Contribution

4.3.1. Variable Global Importance in Prediction

Variable global importance refers to the contribution of each feature to the overall model performance. In other words, it quantifies the impact of each feature on the model's predictive outcomes, aiding in the understanding and assessment of the roles different features play within the model. Variable global importance helps identify the most critical features influencing model performance, thereby facilitating model optimization or system design.

To assess the significance of factors in predicting overdue payments, we computed the Shapley values of 23 variables. Figure 2 illustrates the absolute values of the variable Shapley in a visual representation. The x-axis indicates the contribution of each variable, while the y-axis displays the 23 variables ordered by their contribution levels. The figure provides a clearer visualization of the significance of each variable in predicting overdue payments. The top ten crucial variables are as follows: "Repayment periods", "Loan amount", "Debt to income type", "Revolving line utilization", "Gender", "Age", "Lending number", "Number of days online takeaway search", "Working industry", and "Revolving line". Notably, the "Repayment periods" exhibit a significant lead in predicting defaults, succeeded by "Loan amount" and "Debt to income type", with the importance of other variables gradually diminishing.

Figure 3 provides a direct representation of the Shapley values for each variable, aiding in understanding the positive or negative impact of features on predicted outcomes. The x-axis illustrates the contribution of each variable, while the y-axis displays all variables arranged in descending order based on their Shapley values. When compared to Figure 2, Figure 3 distinctly reveals that the Shapley values for "Number of Days in Online Search" and "Lending Organization Number" are negative, indicating an adverse effect on the predicted default outcome. In contrast, the remaining variables exhibit a positive influence on the predicted default outcome.

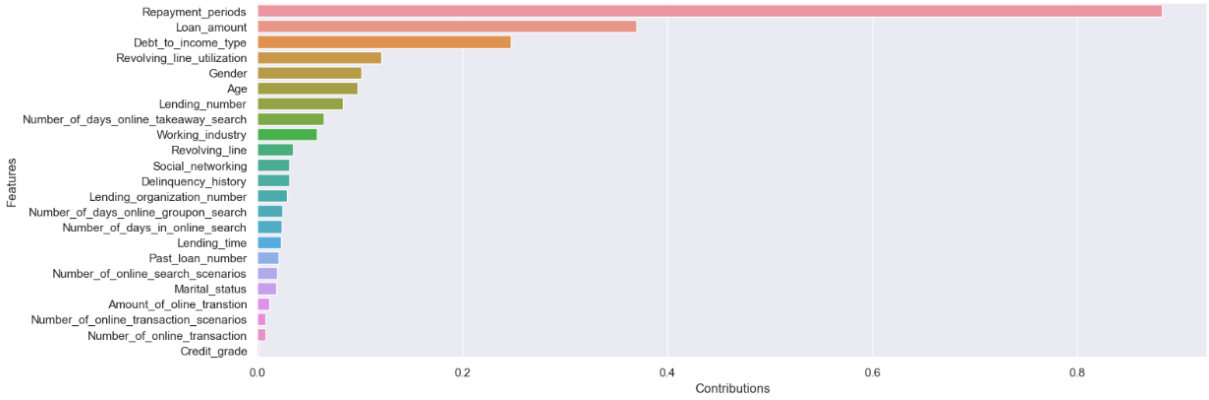

Figure 2. Absolute value of Shapley for the variable. The x-axis denotes the absolute value of each variable's contribution, while the y-axis represents the 23 variables ranked based on their respective contribution values.

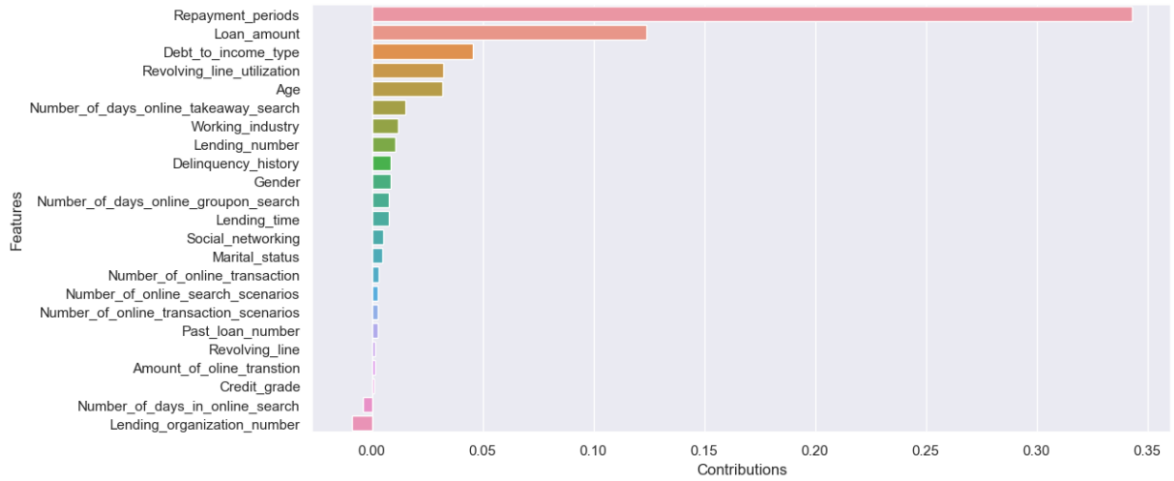

Figure 3. Shapley values for variables. The x-axis represents the contribution of each variable, and the y-axis shows the 23 variables arranged according to their respective contribution.

4.3.2. Variable Local Distribution in Prediction

While the bar chart provides a succinct overview of feature importance rankings, it falls short in providing detailed information. Consequently, amalgamating feature importance and effects through the summary diagram facilitates a more comprehensive understanding of the relationship between the feature values and predictive outcomes. Variable local distribution denotes the dispersion of feature values within a specified sample, offering insights into the specific values of each feature in the sample. This description sheds light on the internal feature values within the given sample. Variable local distribution enables managers to discern the significance and impact of individual features on a particular sample, thereby enhancing their understanding of how the model generates precise predictions.

The summary diagram depicted in Figure 4 aids in comprehending the distribution of Shapley values associated with each feature. Each point in the graph is the Shapley value of a feature and an instance, the position on the y-axis is determined by the feature, the position on the x-axis is determined by the Shapley value, and the color from blue to red represents the feature value from low to high. The overlap points of the sample shake up and down the y-axis, from which we can understand the distribution of Shapley values for each feature. Again, all features are ranked according to their importance.

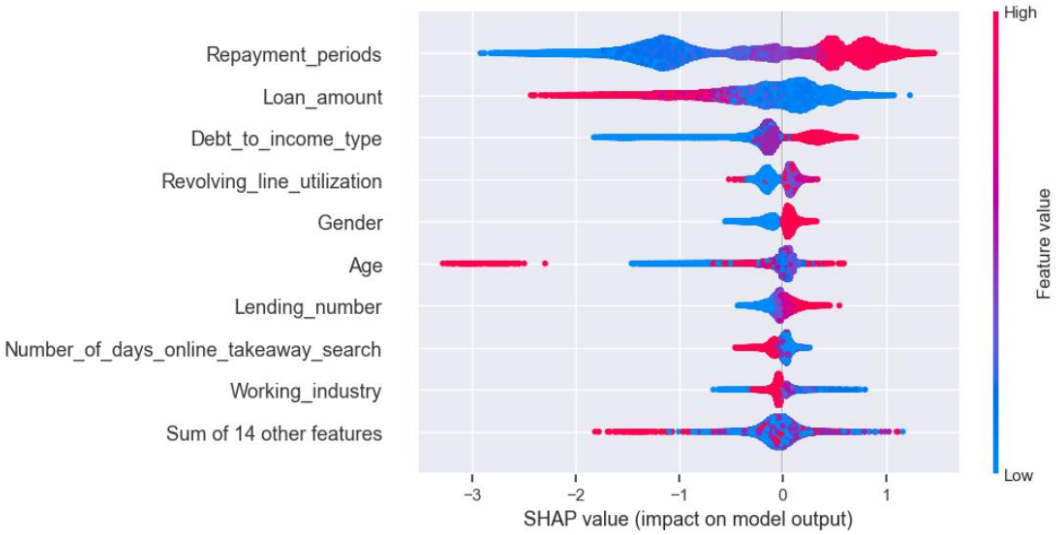

Figure 4. Distribution of local importance. Each point in the graph is the Shapley value of a feature and an instance. The position on the y-axis is determined by the feature, the position on the x-axis is determined by the Shapley value, and the color gradient from blue to red signifies the feature value ranging from low to high.

Figure 4 indicates that a longer "Repayment period" correlates with an increased predicted default risk. The larger the "Loan amount", the lower risk of default is predicted. This effect explains the behavior of the model, but it does not necessarily imply causality in reality. In a study conducted in 2012 on the factors of default in farmers' loan repayment, Dadson concluded through the Probit model that a longer repayment period is more likely to reduce the probability of default [66]. This is contrary to the conclusion obtained in our online consumer credit scenario and also means that, in online consumer credit services, a longer repayment period may give customers the illusion of delayed repayment, making them less sensitive to repayment responsibilities, resulting in more frequent loan repayment delays, thus increasing the potential risk of default. It also reminds managers how to set repayment deadlines more scientifically.

4.4. Risk Personality Portrait

Global and local analyses offer insights into the average contribution of features, yet from the perspective of personalized profiling, understanding individual users becomes paramount. In this context, both managers and users must grasp the model's interpretation of each user's anticipated behavior. The final stage of this framework involves constructing a waterfall graph utilizing Shapley values to create personalized default risk portraits for individual users, aligning with the user portrait theory. This tailored approach not only enhances risk assessment accuracy but also empowers both managers to comprehend and respond effectively to personalized risk profiles, thereby fostering a more proactive and targeted risk management strategy.

Figure 5 displays the Shapley waterfall diagram of the first six sample data. $E[f(x)]$ represents the mean value of the model's predictions for the sample dataset, i.e., the predictions start from a uniform baseline $E[f(x)]$. The x-axis is the value of log-odds pairs of prediction results, so the prediction results can be converted into a continuous range of values, making the interpretation results more general and comparable. The y-axis shows the value for each feature of the sample. The Shapley value for each feature is an arrow, with positive values pushing the prediction and negative values reducing the prediction. For example, for the first sample corresponding to Figure 5a, compared with the base value, the user's "Repayment Periods" was 93 days, which greatly reduced the probability of being predicted to default; on the other hand, the "Loan Amount" was 1980, which increased the probability of being predicted to default to some extent. By looking at the

Shapley waterfall diagram of the first six sample predictions, it can be found that not every user fits the conclusions obtained in the feature contribution analysis. For example, for the second sample corresponding to Figure 5b, characteristic "Repayment Periods" are not the most influential factor, but "Loan Amount" has a greater impact. This further shows that only the macro analysis of features is not enough to understand the heterogeneity of samples, and it also reflects the importance of personalized user portraits.

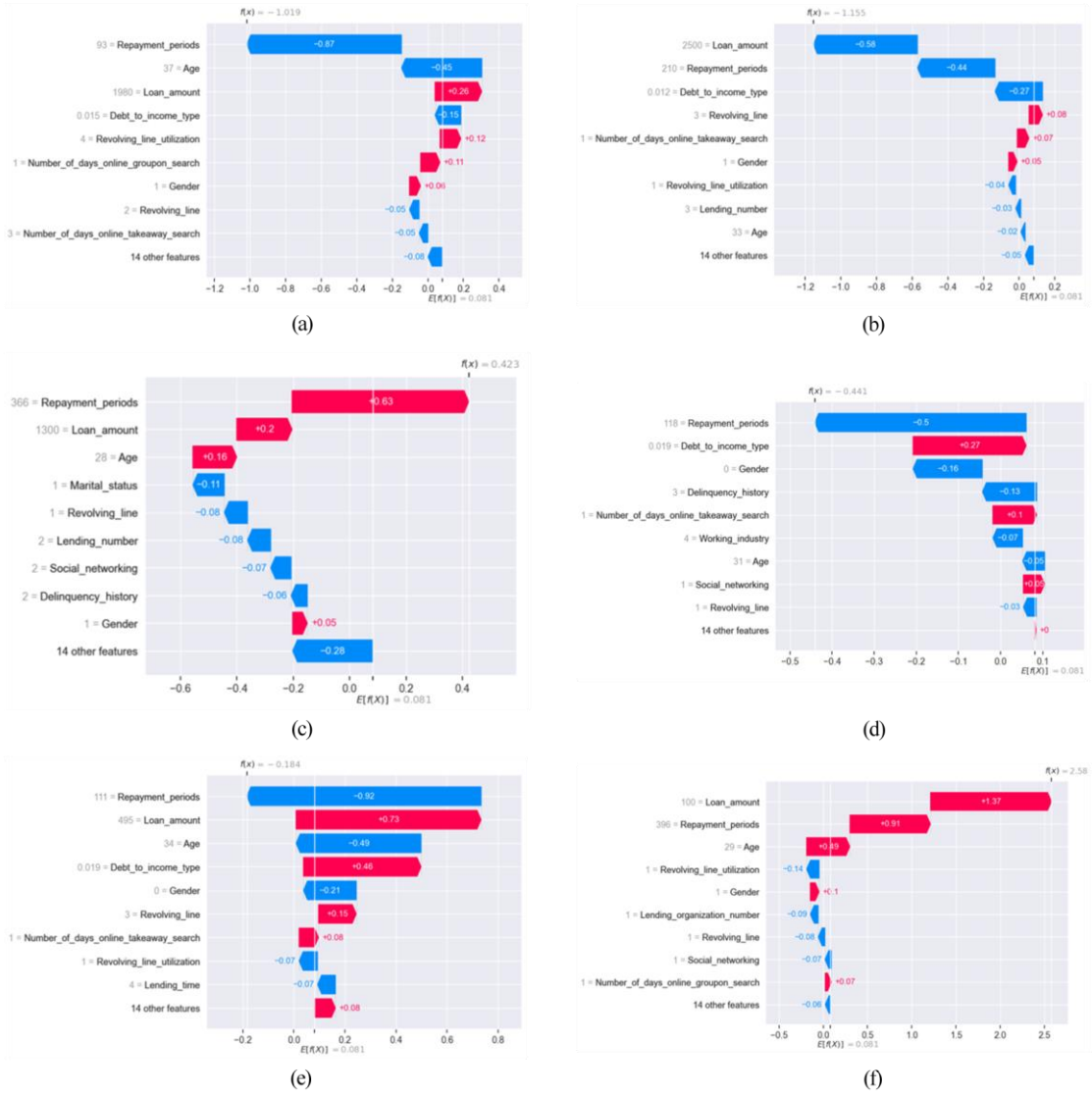

Figure 5. The Waterfall diagram of Shapley values. (**a**–**f**) present the waterfall diagram of the first six sample datasets. Blue signifies a negative impact of the feature on the prediction outcomes, while red indicates a positive effect.

In theory, the concept of default risk personality portrait serves to elucidate the foundational principles of machine learning models for individualized risk prediction in risk management, thereby enhancing model transparency and interpretability. In practice, it provides guidance for model refinement and feature engineering enhancements, aiming to boost the precision and robustness of credit risk management models.

5. Conclusions and Discussion

Given the widespread adoption and advancement of online consumer credit services, this study proposes CDRP, an intelligent management framework for assessing default risk in online consumer loans. The framework enables the integration of diverse data

from various platforms to create personalized user default risk portrait. To illustrate its implementation, consumer credit data from a commercial bank is utilized, incorporating both traditional credit information and Internet platform information. Data preprocessing involves near-miss undersampling to address data imbalance. Following predictions made by the LGB algorithm, interpretable machine learning methods are employed to evaluate the contribution of each indicator during the prediction process, encompassing discussions on global importance and local contributions. Furthermore, leveraging the user portrait theory, managers gain valuable insights through the personalized analysis of individual samples, aiding their decision-making process.

This paper provides a detailed explanation of 23 variables. By effectively comparing these variables on a standardized scale, this study determines their contribution ranking in predicting default and provides managerial guidance. Furthermore, an in-depth analysis reveals the varied performance of each variable in default prediction, indicating their positive or negative impact on consumer loan default. This expands managerial perspectives on controlling default risks. Additionally, the creation of personalized user default risk portrait offers a fresh managerial perspective. In conclusion, the framework presented in this study broke current research limitations and enhances comprehension regarding the management of consumer credit default risks.

The framework development and research findings presented in this study significantly contribute to the existing knowledge body within the realm of online consumer risk management. The consumer default risk portrait (CDRP) framework not only effectively harnesses consumer default risk data but also enhances the interpretability of risk assessment models. It is crucial for both managers and technicians to comprehend how the model interprets the anticipated behaviors of individual users. The framework proposed in this research serves to elucidate the rationales behind the model's predictions for individual users through personalized portraits. It aids technicians in enhancing prediction model performance and ensuring model fairness. It also assists managers in devising targeted management strategies and optimizing resource allocation. For instance, our analysis reveals that "Repayment periods", "Loan amount", and "Debt to income type" are the top three variables significantly influencing default risk prediction outcomes. In the section of variable local distribution in prediction, we observe that as the "Repayment period" increases, the predicted risk of default increases. This suggests that in online consumer credit services, longer repayment periods might create a perception of delayed repayment among customers, potentially reducing their sensitivity towards repayment obligations and leading to more frequent delays in loan repayments, consequently elevating the risk of default. This finding alerts managers of the importance of setting repayment deadlines judiciously. Furthermore, we provide personalized portraits for six samples individually, shedding light on the specific feature contributions to each predicted outcome. Through these individualized results, the significance of crafting personalized portraits in the context of default risk management is further underscored.

Despite the contributions of this study to online consumer default forecasting, it also has some limitations. On the one hand, as mentioned above, credit data are subject to variability due to the "smarts" of fraudsters. How to deal with dynamically updated data to complete dynamic modeling is also a major challenge facing the management framework. On the other hand, CDRP ignores the economic costs of the management process. From an economic perspective, cost and profitability are also factors that managers pay close attention to. Therefore, it is worth developing a more "cost-effective" regulatory framework to help promote economic growth. Lastly, by explaining the rationale behind each prediction through the CDRP framework, we also hope that we can gain a deeper understanding of the model's decision-making process, further refining the model to improve interpretability while maintaining accuracy.

Author Contributions: Conceptualization, J.L.; Methodology, M.Z.; Software, M.Z.; Validation, M.S.; Resources, B.-C.S.; Data curation, M.Z.; Writing—original draft, M.Z.; Writing—review & editing, J.L.; Supervision, B.-C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the financial support provided by the High-level talent research of Huaqiao University, China (Grant (22SKBS028).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly accessible because they originate from a commercial bank and involve user privacy.

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

References

1. Rona-Tas, A.; Guseva, A. Consumer credit in comparative perspective. *Annu. Rev. Sociol.* **2018**, *44*, 55–75. [CrossRef]
2. Kshetri, N. Big data's role in expanding access to financial services in China. *Int. J. Inf. Manag.* **2016**, *36*, 297–308. [CrossRef]
3. Wang, Z.; Jiang, C.; Ding, Y.; Lyu, X.; Liu, Y. A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electron. Commer. Res. Appl.* **2018**, *27*, 74–82. [CrossRef]
4. Peñaloza, L.; Barnhart, M. Living US capitalism: The normalization of credit/debt. *J. Consum. Res.* **2011**, *38*, 743–762. [CrossRef]
5. Thakor, A.V. The financial crisis of 2007–2009: Why did it happen and what did we learn? *Rev. Corp. Financ. Stud.* **2015**, *4*, 155–205. [CrossRef]
6. Mishkin, F.S. Over the cliff: From the subprime to the global financial crisis. *J. Econ. Perspect.* **2011**, *25*, 49–70. [CrossRef]
7. Xiao, J.J.; Tao, C. Consumer finance/household finance: The definition and scope. *China Financ. Rev. Int.* **2021**, *11*, 1–25. [CrossRef]
8. Bhatore, S.; Mohan, L.; Reddy, Y.R. Machine learning techniques for credit risk evaluation: A systematic literature review. *J. Bank. Financ. Technol.* **2020**, *4*, 111–138. [CrossRef]
9. Gaganis, C.; Papadimitri, P.; Pasiouras, F.; Tasiou, M. Social traits and credit card default: A two-stage prediction framework. *Ann. Oper. Res.* **2023**, *325*, 1231–1253. [CrossRef]
10. He, H.; Wang, Z.; Jain, H.; Jiang, C.; Yang, S. A privacy-preserving decentralized credit scoring method based on multi-party information. *Decis. Support Syst.* **2023**, *166*, 113910. [CrossRef]
11. Mourtas, S.D.; Katsikis, V.N.; Stanimirović, P.S.; Kazakovtsev, L.A. Credit and Loan Approval Classification Using a Bio-Inspired Neural Network. *Biomimetics* **2024**, *9*, 120. [CrossRef] [PubMed]
12. Boustani, N.; Emrouznejad, A.; Gholami, R.; Despic, O.; Ioannou, A. Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks. *Ann. Oper. Res.* **2023**, 1–18. [CrossRef]
13. Tripathi, D.; Edla, D.R.; Kuppili, V.; Bablani, A. Evolutionary extreme learning machine with novel activation function for credit scoring. *Eng. Appl. Artif. Intell.* **2020**, *96*, 103980. [CrossRef]
14. Bücker, M.; Szepannek, G.; Gosiewska, A.; Biecek, P. Transparency, auditability, and explainability of machine learning models in credit scoring. *J. Oper. Res. Soc.* **2022**, *73*, 70–90. [CrossRef]
15. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
16. Ashofteh, A.; Bravo, J.M. A conservative approach for online credit scoring. *Expert Syst. Appl.* **2021**, *176*, 114835. [CrossRef]
17. Zhang, X.; Yu, L. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Syst. Appl.* **2023**, *237*, 121484. [CrossRef]
18. Gao, M.; Zhang, Y.; Gao, Y. Research Progress of User Portrait Technology in Medical Field. In Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences, Beijing, China, 29–31 October 2021; pp. 500–504.
19. Cooper, A. *The Inmates Are Running the Asylum*; Software-Ergonomie '99. Berichte des German Chapter of the ACM, vol 53; Vieweg+Teubner Verlag: Wiesbaden, Germany, 1999. [CrossRef]
20. Yuan, T. User Portrait Based on Artificial Intelligence. In Proceedings of the International Conference on Frontier Computing, Tokyo, Japan, 12–15 July 2022; Springer Nature: Singapore, 2022; pp. 359–366. [CrossRef]
21. Peng, J.; Choo KK, R.; Ashman, H. User profiling in intrusion detection: A review. *J. Netw. Comput. Appl.* **2016**, *72*, 14–27. [CrossRef]
22. Yao, W.; Hou, Q.; Wang, J.; Lin, H.; Li, X.; Wang, X. A personalized recommendation system based on user portrait. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan, China, 12–13 July 2019; pp. 341–347.
23. Su, M.; Cheng, D.; Xu, Y.; Weng, F. An improved BERT method for the evolution of network public opinion of major infectious diseases: Case Study of COVID-19. *Expert Syst. Appl.* **2023**, *233*, 120938. [CrossRef]
24. Niu, W.; Liu, J.; Ishikawa. *User Network Behavior Profiling: Analysis and Content Recommendation Application of User Network Behavior Profiling in Big Data*; Beijing Book Co., Inc.: Linden, NJ, USA, 2016.
25. Xie, X.; Zhang, J.; Luo, Y.; Gu, J.; Li, Y. Enterprise credit risk portrait and evaluation from the perspective of the supply chain. *Int. Trans. Oper. Res.* **2023**, *31*, 2765–2795. [CrossRef]
26. Liu, Y. Computer Method Research on Risk Control Identification System Based on Deep Learning. In Proceedings of the 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 27–28 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 561–565.
27. Zhang, Z.; Han, L.; Chen, M. Multi-label learning with user credit data in China based on MLKNN. In Proceedings of the 4th International Conference on Information Technology and Computer Communications, Guangzhou, China, 23–25 June 2022; pp. 105–111.

28. Zhu, X. Internet financial risk control model based on machine learning algorithm. In Proceedings of the 2022 International Conference on Artificial Intelligence of Things and Crowdsensing (AIoTCs), Nicosia, Cyprus, 26–28 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 380–384.
29. Shapley, L.S. A value for n-person games (1953). In *Classics in Game Theory*; Harold, W.K., Ed.; Princeton University Press: Princeton, NJ, USA, 1997; pp. 69–79. [CrossRef]
30. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4768–4777.
31. Weng, F.; Zhu, J.; Yang, C.; Gao, W.; Zhang, H. Analysis of financial pressure impacts on the health care industry with an explainable machine learning method: China versus the USA. *Expert Syst. Appl.* **2022**, *210*, 118482. [CrossRef]
32. Molnar, C. *Interpretable Machine Learning*; Lulu. com: Morrisville, NC, USA, 2020.
33. Yang, C.; Zhang, H.; Weng, F. Effects of COVID-19 vaccination programs on EU carbon price forecasts: Evidence from explainable machine learning. *Int. Rev. Financ. Anal.* **2023**, *91*, 102953. [CrossRef]
34. Yang, C.; Abedin, M.Z.; Zhang, H.; Weng, F.; Hajek, P. An interpretable system for predicting the impact of COVID-19 government interventions on stock market sectors. *Ann. Oper. Res.* **2023**, 1–28. [CrossRef] [PubMed]
35. Brown, I.; Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.* **2012**, *39*, 3446–3453. [CrossRef]
36. Yi, Z.; Cao, X.; Chen, Z.; Li, S. Artificial Intelligence in Accounting and Finance: Challenges and Opportunities. *IEEE Access* **2023**, *11*, 129100–129123. [CrossRef]
37. Li, W.; Ding, S.; Chen, Y.; Wang, H.; Yang, S. Transfer learning-based default prediction model for consumer credit in China. *J. Supercomput.* **2019**, *75*, 862–884. [CrossRef]
38. Papouskova, M.; Hajek, P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decis. Support Syst.* **2019**, *118*, 33–45. [CrossRef]
39. Costa e Silva, E.; Lopes, I.C.; Correia, A.; Faria, S. A logistic regression model for consumer default risk. *J. Appl. Stat.* **2020**, *47*, 2879–2894. [CrossRef]
40. Hou, J.; Li, Q.; Liu, Y.; Zhang, S. An enhanced cascading model for E-commerce consumer credit default prediction. *J. Organ. End User Comput. (JOEUC)* **2021**, *33*, 1–18. [CrossRef]
41. Wen, H.; Sui, X.; Lu, S. Study on Effect of Consumer Information in Personal Credit Risk Evaluation. *Complexity* **2022**, *2022*, 7340010. [CrossRef]
42. Lappas, P.Z.; Yannacopoulos, A.N. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Appl. Soft Comput.* **2021**, *107*, 107391. [CrossRef]
43. Chang, J.S.; Chang, W.H. Analysis of fraudulent behavior strategies in online auctions for detecting latent fraudsters. *Electron. Commer. Res. Appl.* **2014**, *13*, 79–97. [CrossRef]
44. 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]
45. Xia, Y.; Li, Y.; He, L.; Xu, Y.; Meng, Y. Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electron. Commer. Res. Appl.* **2021**, *49*, 101095. [CrossRef]
46. Zhang, L.; Wang, J.; Liu, Z. What should lenders be more concerned about? Developing a profit-driven loan default prediction model. *Expert Syst. Appl.* **2023**, *213*, 118938. [CrossRef]
47. Zhou, J.; Wang, C.; Ren, F.; Chen, G. Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. *Decis. Support Syst.* **2021**, *149*, 113611. [CrossRef]
48. Mqadi, N.M.; Naicker, N.; Adeliyi, T. Solving misclassification of the credit card imbalance problem using near miss. *Math. Probl. Eng.* **2021**, *2021*, 7194728. [CrossRef]
49. Alsowail, R.A. An insider threat detection model using one-hot encoding and near-miss under-sampling techniques. In Proceedings of the International Joint Conference on Advances in Computational Intelligence: IJCACI 2021, Online, 23–24 October 2021; Springer Nature: Singapore, 2022; pp. 183–196.
50. Gao, X.; Luo, H.; Wang, Q.; Zhao, F.; Ye, L.; Zhang, Y. A human activity recognition algorithm based on stacking denoising autoencoder and lightGBM. *Sensors* **2019**, *19*, 947. [CrossRef]
51. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
52. Weng, F.; Meng, Y.; Lu, F.; Wang, Y.; Wang, W.; Xu, L.; Cheng, D.; Zhu, J. Differentiation of intestinal tuberculosis and Crohn's disease through an explainable machine learning method. *Sci. Rep.* **2022**, *12*, 1714. [CrossRef]
53. Wang, Y.; Jia, Y.; Tian, Y.; Xiao, J. Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring. *Expert Syst. Appl.* **2022**, *200*, 117013. [CrossRef]
54. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2020**, *17*, 168–192. [CrossRef]
55. Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of classification methods on unbalanced data sets. *IEEE Access* **2021**, *9*, 64606–64628. [CrossRef]
56. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
57. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B.; Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear discriminant analysis. *Robust Data Min.* **2013**, 27–33. [CrossRef]

58. LaValley, M.P. Logistic regression. *Circulation* **2008**, *117*, 2395–2399. [CrossRef]
59. Ontivero-Ortega, M.; Lage-Castellanos, A.; Valente, G.; Goebel, R.; Valdes-Sosa, M. Fast Gaussian Naïve Bayes for searchlight classification analysis. *Neuroimage* **2017**, *163*, 471–479. [CrossRef]
60. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
61. Song, Y.Y.; Ying, L.U. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
62. Suthaharan, S.; Suthaharan, S. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Springer: Boston, MA, USA, 2016; pp. 207–235.
63. Ketkar, N.; Ketkar, N. Stochastic gradient descent. In *Deep Learning with Python: A Hands-On Introduction*; Apress: Berkeley, CA, USA, 2017; pp. 113–132.
64. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
65. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
66. Awunyo-Vitor, D. Determinants of loan repayment default among farmers in Ghana. *J. Dev. Agric. Econ.* **2012**, *4*, 339–345.