Research Article

Research on Precision Marketing Based on Consumer Portrait from the Perspective of Machine Learning

Mengmeng Zhang

1The Department of Logistics and E-Commerce, Henan University of Animal Husbandry and Economy, Zhengzhou, 450044 Henan, China
2College of Business Administration, University of the Cordilleras, Baguio, 2600 Baguio, Philippines

Correspondence should be addressed to Mengmeng Zhang; zhangmengmeng@hnuahe.edu.cn

Received 9 March 2022; Accepted 28 March 2022; Published 12 May 2022

Academic Editor: Zhiguo Qu

Copyright © 2022 Mengmeng Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the advent of the big data era and the development of artificial intelligence, the traditional marketing model has undergone dramatic changes, and precision marketing has become the focus of current research. Consumer portrait is a labeled consumer model abstracted from the basic attributes, social attributes, behavior characteristics, and psychological characteristics of consumers, which can reflect the real needs of consumers well. Precision marketing based on consumer portraits is a fine division of consumers on the premise of in-depth understanding of consumer-related data. Based on the public dataset of a bank in a certain region, this paper extracts a series of data features such as basic attributes and social attributes of consumers to draw consumer portraits. BP neural network, SVM, and random forest algorithm in machine learning algorithm are used to model and predict whether consumers should take out personal loans, and then, the prediction effect of these algorithms is compared to judge which algorithm has better prediction effect. The results show that the random forest algorithm has the highest prediction accuracy, reaching an average of 94% in 10 calculations, which can help banks to select potential target customers to some extent.

1. Introduction

Nature published a special issue on big data in 2008, which put forward the concept of big data for the first time. After more than 10 years of development, big data has been paid more and more attention by enterprises [1]. How to make more effective use of these data for analysis and evaluation has become the problem of the enterprise based on the big data background after the Internet enters the era of big data. Big data has become as indispensable to Internet companies as water, electricity, and air are to people’s lives. From infrastructure construction to application level, it includes data platform construction and management, data warehouse development, statistical analysis of upper application, report generation and visualization, user portrait modeling, personalized recommendation and precision marketing, and other application directions. Consumer portrait can help big data go out of the data warehouse and carry out personalized recommendation, precision marketing, personalized service, and other diversified services for users, which is an important direction of the application of big data. Enterprises need to establish consumer portraits of their own enterprises if they want to make use of big data to serve fine operation and precise marketing.

In recent years, it is an important strategic plan for China’s economy to develop to a higher level. Stimulating domestic demand is becoming more and more important; one of the important means is to vigorously develop personal credit business [2]. Due to the endless emergence of Internet financial platforms and small loan companies, traditional state-owned banks have encountered many obstacles in promoting the development of individual loan business, and their market shares are not dominant. This paper takes the marketing of a bank’s personal loan business as the research object, establishes the consumer portrait with the help of big data technology, establishes the prediction model of personal credit business management based on machine learning algorithm, and helps the bank to find out
potential target customers to achieve accurate marketing, which is of great significance to the future development of the bank.

2. Consumer Portrait and Precision Marketing in the Big Data Environment

As an effective tool to screen target users, connect user demands with design direction, consumer portrait has been widely used in various fields. Consumer portraits often connect users’ attributes, behaviors, and expected data with the most simple and life-close discourse in practical applications.

2.1. Consumer Portrait. Consumer portrait, the labeling of consumer information, refers to describing a user’s characteristics through direct or indirect data [3]. By collecting the data of social attributes, consumption habits, preference characteristics, and other dimensions of consumers, the characteristic attributes of consumers are described. Through the analysis and statistics of these features, the potential value information can be mined to abstract the overall picture of consumer information. Consumer portraits can be viewed as the foundation of enterprise application of large data, the basis for targeted advertising and personalized recommendation, and the foundation for data-driven operations. Therefore, it is very important to extract valuable information from massive data.

Consumers’ needs can be better understood through consumer portraits. Consumer portrait can effectively help enterprises understand customer groups and provide decision support for enterprises to achieve precision marketing in the business field [4]. In the construction of consumer portrait, the “scene five forces” can be combined to analyze the behavioral needs of different user groups [5], and the accuracy can also be improved by combining the modular calculation method on the basis of k-means clustering algorithm [6]. Consumer portraits are widely used in the recommendation system, especially the recommendation of e-commerce platforms. With the continuous progress of modern science and technology and the Internet, enterprises can accurately locate the personalized needs of consumers through consumer data, so as to achieve the precise marketing of enterprises [7, 8]. In addition, consumer portraits can also be used to recommend scenic spots [9]. Consumer portrait can also be applied in the field of scientific and technological information to sort out users’ needs [10]. The dynamic consumer portrait model can be combined with the recommendation algorithm, which greatly improves the accuracy of the recommendation algorithm and makes the recommendation results better meet the personalized needs of consumers [11]. By using Euclidean distance, cosine distance, and other methods to calculate the similarity between user data and book content, the book consumer portrait is constructed, which provides great convenience for the development of book recommendation system [12].

2.2. Precision Marketing. The concept of “precision marketing” was clearly defined by American marketing expert Kotler at the Global Speech Forum, who indicated that companies should develop more accurate, measurable, and high return on investment marketing communications [13]. Later, Farris and other scholars discussed how to quantitfy and evaluate marketing indicators in their book Marketing Metrics [14]. Erudeng, a domestic scholar, believes that precision marketing is “standard” and “certainty”; “standard” refers to continuous replication and promotion, while “certainty” represents a complete grasp of the current market [15]. Some scholars believe that precision marketing is a way to achieve measurable low-cost expansion of enterprises by establishing a personalized customer communication service system [16]. In addition, specific methods of precision marketing were put forward in Research on Precision Marketing Methods published in 2008, which were divided into three methods: database-based, Internet-based, and relying on others [17]. With the development of science and technology, Caijie elaborated on the need to establish a precision marketing system with the network and information technology as the core and began to use the concept of big data [18]. Mingshuo et al. believe that traditional CRM tools are difficult to meet the needs of business development. In order to achieve customer expansion and quality improvement, banks need to innovate ideas, such as introducing popular machine intelligence, big data, and other advanced technologies to reshape business processes and break traditional inefficient organizational structures, so as to improve customer experience [19].

3. Construction of Consumer Portrait Based on Personal Credit Management

Modeling consumer portraits is actually “labeling” consumers. From the perspective of labeling methods for consumers, there are generally three types. (1) Statistical labels: it is the most basic and common. For a user, their gender, age, city, constellation, and so on can be derived from user registration data and consumption data, which forms the basis of the consumer portrait. (2) Rule class labels: it is based on consumer behavior and established rules. For example, the definition of “consumer activity” on a platform is the “number of transactions in the past 30 days > 2.” Rules for rule class labels are determined by the operator and data personnel through consultation. (3) Machine learning mining class labels: it is generated by machine learning mining to predict certain attributes or behaviors of consumers. For example, a user’s preference for a commodity is determined by his or her consumption habits.

In project engineering practice, labels of general statistics and rules occupy a large proportion in development. Machine learning mining labels are mostly used to predict scenarios, such as judging users’ preference for purchasing goods and users’ churn intention. Generally, the development cycle of machine learning mining labels is long and the cost is high. This paper takes machine learning mining labels as the goal to play its role in precision marketing.

3.1. Data Selection. With the continuous development of many emerging Internet finance, such as Ant Cash Now, Ant Credit Pay, and Jingdong Finance, the performance of traditional commercial banks has declined seriously. Many
Internet financial products are trusted by consumers and their market share keeps expanding. Traditional banks have to formulate relevant countermeasures to ensure their normal operation and steady development. Among them, the application and practice of big data and other Internet technologies have great advantages, and traditional banks can improve their marketing accuracy to attract more high-quality consumers.

The dataset selected in this paper is the public dataset of a bank in a certain region, including 10142 customers and their related attribute indicators, including gender, region, education level, age, average monthly income, credit qualification, average monthly consumption, certificate type, marital status, occupational status, and personal loan. In this paper, the potential consumer mining model of personal loans will be established based on the dataset. Firstly, the attribute information related to personal credit loans will be extracted according to the consumer attribute information provided by the dataset to depict the consumer portrait. Consumer portrait can reasonably organize and store the massive data which is complex, diverse, widely distributed, and heterogeneous [20]. Based on the established consumer portrait, modeling and prediction are carried out through machine learning algorithm to mine potential target consumers and achieve accurate marketing of bank personal credit loans.

3.2. Data Analysis. Firstly, the original data are cleaned by eliminating invalid attributes and deleting missing values. Through statistical analysis of the data, it can be obtained as follows.

As shown in Figure 1, the proportion of male consumers is 54%, while that of female consumers is 46%. This shows that the gender distribution of consumers is balanced and there is no significant difference. Additionally, after statistics, the average age of customers is about 41, the highest is 85, and the lowest is 18.

From Figure 2, it can be seen that 1356 customers made personal loans, accounting for about 13.37%, while the remaining 8786 customers did not make personal loans.

3.3. Variable Index Selection. It can be seen from the above analysis that the number of consumers with personal loans is less than the number of consumers without personal loans. If all these data are imported into the model for calculation, it will lead to unbalanced training samples, which often leads to overfitting of the model, thus affecting the classification effect [21]. Therefore, this paper randomly extracts 400 sets of data from the data of all consumers, in which the number of consumers who process personal loans and those who do not process personal loans are 200, and forms a new dataset.

Six relatively key indicators are selected according to the degree of correlation affecting individual loan business, which are the characteristic variables of the mining model of potential individual loan customers, including gender, education level, age, average monthly income, average monthly consumption, and marital status. These six indicators include basic customer information such as gender and age and customer financial behavior information such as income level and consumption level. Then, some non-data indicators are coded, as shown in Table 1.

All characteristic data are integrated to obtain a new dataset. The six attribute data of gender, education level, age, average monthly income, average monthly consumption, and marital status will be used as input to the model, while whether the customer has handled personal loans will be used as output of the model. Table 2 shows part of the sample data.
have been applied in many discipline in the 21st century, including research methods that Machine learning has gradually developed into a very complete

4. Prediction of Personal Credit Processing Based on Machine Learning

Machine learning has gradually developed into a very complete discipline in the 21st century, including research methods that have been applied in many fields of scientific research [22]. In the field of machine learning, supervised learning, unsupervised learning, and semisupervised learning are three kinds of learning technologies that are widely used and studied. (1) Supervised learning: a function is generated that maps the input to the corresponding label. The network finally compares the error from back to front [23]. In the former process, the signal is input from the input layer, processed by the hidden layer, and eventually sent to the output layer. The difference between the results of the output layer and the label value of the sample data is the error of the network under the current weight and threshold. At this time, it is necessary to adjust the weight and threshold through the error, so as to make the final predictive output of the neural network gradually approach the corresponding label value of the sample.

4.1. BP Neural Network. BP neural network is one of the most widely used supervised learning neural networks. The neural network with tutor learning means that in the learning process of neural network, it needs to provide a set of data samples corresponding to input and output correctly, and each set of input sample data has a fixed label corresponding to it. The network finds the corresponding relationship between them by learning the potential rules. In each learning process, if the value calculated by the network is different from the label value of the group of data, it needs to calculate its error and then change its internal weight and threshold value to minimize the error. The neural network without tutor learning refers to that the adjustment of network weights and thresholds is only affected by network input, and the input sample data is not labeled with the corresponding label, so it cannot be adjusted by calculation error after learning. This kind of neural network basically uses clustering method to classify the input sample data simply.

BP neural network continuously advances the sample training data from front to back, compares the value calculated by the network with the label value of the data, and propagates the error from back to front [23]. In the former process, the signal is input from the input layer, processed by the hidden layer, and eventually sent to the output layer. The difference between the results of the output layer and the label value of the sample data is the error of the network under the current weight and threshold. At this time, it is necessary to adjust the weight and threshold through the error, so as to make the final predictive output of the neural network gradually approach the corresponding label value of the sample.

Table 1: Assignment of nondata indicators.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Gender</th>
<th>Education level</th>
<th>Marital status</th>
<th>Whether to apply for a loan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute values</td>
<td>0—male</td>
<td>0—high school or below</td>
<td>0—married</td>
<td>0—no</td>
</tr>
<tr>
<td></td>
<td>1—female</td>
<td>1—bachelor</td>
<td>1—unmarried</td>
<td>1—yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2—master</td>
<td>2—divorced</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3—doctor</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Sample feature data.

<table>
<thead>
<tr>
<th>The serial number</th>
<th>Gender</th>
<th>Age</th>
<th>Average monthly income</th>
<th>Marital status</th>
<th>Average monthly consumption</th>
<th>Education level</th>
<th>Whether to apply for a loan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
<td>4000</td>
<td>1</td>
<td>1500</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>25</td>
<td>5300</td>
<td>1</td>
<td>2000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>30</td>
<td>3800</td>
<td>1</td>
<td>1000</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>32</td>
<td>6100</td>
<td>0</td>
<td>2800</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>51</td>
<td>12000</td>
<td>0</td>
<td>5000</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>40</td>
<td>9500</td>
<td>0</td>
<td>4600</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>39</td>
<td>4200</td>
<td>0</td>
<td>1300</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

...
According to the known Step 2. the connection weights \( w_{ij} \) and \( w_{jk} \) between neurons in the input layer, hidden layer, and output layer are determined. The threshold of the hidden layer and the threshold of the output layer are initialized, and the network learning rate and neuron excitation function are preliminarily determined.

Step 3. The predictive output \( O \) can be obtained from the known connection weights \( w_{jk} \), thresholds \( b \) and \( H \) obtained in the previous step.

\[
O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k. 
\]

Step 4. Calculate the prediction error \( e \) by using the label values \( Y \) of the sample and \( O \) calculated in the previous step.

\[
e_k = Y_k - O_k, \quad k = 1, 2, \ldots, m. 
\]

Step 5. Calculate \( w_{ij} \) and \( w_{jk} \) of network connections again through the prediction error \( e \).

\[
w_{ij} = w_{ij} + \eta H_j (1 - H_j) x(i) \sum_{k=1}^{m} w_{jk} e_k, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, l, \\
w_{jk} = w_{jk} + \eta H_j e_k, \quad j = 1, 2, \ldots, l, k = 1, 2, \ldots, m. 
\]

Among them, \( \eta \) is the learning rate.

Step 6. Calculate the threshold \( a, b \) of network node again based on the error \( e \) of network prediction.

\[
a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^{m} w_{jk} e_k, \quad j = 1, 2, \ldots, l, \\
b_k = b_k + e_k, \quad k = 1, 2, \ldots, m. 
\]

Step 7. Evaluate whether the iteration is complete or not, and continue to Step 2 if it is not.

4.2. SVM. Support vector machine (SVM) is very powerful and suitable for classification, which is to find the most appropriate hyperplane as the decision surface by training data samples, as shown in Figure 4. The data is divided by using this surface to maximize the distance from the closest point in space to itself, which is easy to calculate and allows functions to be built in a wider set of functions [24]. It is based on statistical learning theory. Statistical learning theory uses structural risk minimization criteria to ensure that all sample points are classified with minimal errors and that the accuracy of data classification is guaranteed while minimizing the dimensions as possible. This can enhance the model’s ability to adapt to
different samples without being limited by the number of independent parameters in the sample data [25].

Support vector machines were originally proposed to solve linear separable problems. Let \( \{ (x_i, y_i), i = 1, 2, \cdots, l \} \) be the sample data of the total number \( l \), which is mainly composed of two categories. If \( x_i \) is the first type of data, it will be represented as \( y_i = 1 \). If it is the second type of data, it will be represented as \( y_i = -1 \).

If there is a hyperplane,
\[
w x + b = 0. \tag{7}
\]

Training data can be accurately divided into two categories, and the data belonging to the same category can be assigned to the same side of the hyperplane, so the training data is linearly separable:
\[
\begin{align*}
wx_i + b &\geq 1, \quad y_i = 1, \quad i = 1, 2, \cdots, l, \\
wx_i + b &\leq -1, \quad y_i = -1,
\end{align*}
\tag{8}
\]

The distance from a point \( x_i \) in the sample to the hyperplane is
\[
\epsilon_i = y_i (wx_i + b) = |wx_i + b|. \tag{9}
\]

\( w \) and \( b \) in the above formula are normalized to \( w/\|w\| \) and \( b/\|w\| \), and the normalized distance is called the geometric distance.
\[
\delta_i = \frac{wx_i + b}{\|w\|}. \tag{10}
\]

At the same time, the geometric distance of the sample data point closest to the hyperplane is called the distance.
\[
\delta = \min \delta_i, \quad i = 1, 2, \cdots, l. \tag{11}
\]

Among them, the error classification times \( N \) of samples and the distance \( \delta \) between the whole training set and the hyperplane satisfy the following formula:
\[
N \leq \left( \frac{2R}{\delta} \right)^2. \tag{12}
\]

Among formula (12), \( R = \max \|x_i\|, i = 1, 2, \cdots, l \), is the value with the longest vector length in the whole sample data.

The maximum number of error classification is related to the distance between the whole training set and the hyperplane. Therefore, it is necessary to find the most suitable classification surface to maximize the distance between the whole training set and the hyperplane.

If \( \varepsilon = |wx_i + b| = 1 \), the distance between the sample points of the two types is \( 2|wx_i + b|/\|w\| = 2/\|w\| \). So the goal is
\[
\begin{align*}
\min & \quad \frac{\|w\|^2}{2} \\
\text{s.t.} & \quad y_i (wx_i + b) \geq 1, i = 1, 2, \cdots, l
\end{align*}
\tag{13}
\]

The problem can be solved by solving the saddle points of the Lagrange function.
\[
\Phi(w, b, \alpha_i) = \frac{\|w\|^2}{2} - \sum_{i=1}^{l} \alpha_i [y_i (wx_i + b) - 1]. \tag{14}
\]

Among the formula, \( \alpha_i > 1, i = 1, 2, \cdots, l \), is the Lagrange coefficient.

If it is solved directly, the calculation will be too complicated. According to the Lagrange duality theory, it can be transformed into a duality problem.
\[
\begin{align*}
\max & \quad Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i, x_j) \\
\text{s.t.} & \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geq 0
\end{align*}
\tag{15}
\]
Such duality problems are usually solved by quadratic programming. Assuming that $\alpha^* = [\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*]^T$ is the optimal solution obtained by solving, then

$$w^* = \sum_{i=1}^l \alpha_i^* x_i y_i,$$

$$b^* = -\frac{1}{2} w^*(x_r + x_s).$$

$x_r$ and $x_s$ are any pair of support vectors of the two classes.

Finally, $f(x)$ is

$$f(x) = \text{sgn} \left[ \sum_{i=1}^l \alpha_i^* y_i (x x_i) + b^* \right].$$

Generally, the optimal classification hyperplane cannot be found due to some sample data in the sample dataset. In this case, relaxation variables can be introduced to correct the target and constraint conditions:

$$\min \left\{ \frac{||w||^2}{2} + C \sum_{i=1}^l \xi_i \right\},$$

$$\text{s.t.} \quad y_i (w x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l,$$

$$\xi_i > 0$$

In the equation, $C$ is the penalty factor, which is conducive to making a compromise between the classification error of the sample training set and the complexity of the algorithm. The rest is similar, but the constraint condition becomes

$$\left\{ \begin{array}{l} \sum_{i=1}^l \alpha_i y_i = 0, \quad i = 1, 2, \ldots, l. \\ 0 \leq \alpha_i \leq C. \end{array} \right.$$ (19)

The final form of the classification function $f(x)$ is the same as before.

All of the above analysis is based on the assumption that the training sample dataset is linearly separable. However, most of the problems are nonlinear. In this regard, nonlinear mapping $\Phi: \mathbb{R}^d \rightarrow H$ is required to map the training data input at the beginning into the high-dimensional feature space $H$.

By mapping it to a high-dimensional feature space, the point product calculation is also required, but this will greatly increase the amount of data calculation and make the problem very complex. In this case, the point product operation needs to be replaced by a kernel function that satisfies the Mercer condition.

$$K(x_i, x_j) = \Phi(x_i) \Phi(x_j).$$ (20)

This method of using kernel function instead of dot product in high-dimensional characteristic space can reduce computation and complexity. After mapping to the high-dimensional feature space, the following equation can be obtained:

$$\max_{Q(\alpha)} \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{s.t.} \quad \left\{ \begin{array}{l} \sum_{i=1}^l \alpha_i y_i = 0, \quad i = 1, 2, \ldots, l. \\ 0 \leq \alpha_i \leq C. \end{array} \right.$$ (21)
Assuming that $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*)^T$ is the solution of the above equation, then

$$w^* = \sum_{i=1}^{l} \alpha_i^* y_i \Phi(x_i).$$  \hspace{1cm} (22)

Then, the optimal classification function is

$$f(x) = \text{sgn} \left[ \sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^* \right].$$  \hspace{1cm} (23)

Figure 5 shows the structure of SVM. The output of SVM is calculated by the linear combination of intermediate nodes, each of which corresponds to the support vector one by one.

Currently, there are several kernel functions [26]:

1. Linear kernel function

$$K(x, x_i) = xx_i.$$  \hspace{1cm} (24)

2. Kernel function of order polynomial $d$

$$K(x, x_i) = (xx_i + 1)^d.$$  \hspace{1cm} (25)

3. Radial basis kernel function

$$K(x, x_i) = \exp \left( -\frac{\|x - x_i\|^2}{2\sigma^2} \right).$$  \hspace{1cm} (26)

4. Sigmoid kernel functions with arguments $k$ and $\theta$

$$K(x, x_i) = \tanh (k xx_i + \theta).$$  \hspace{1cm} (27)

Among them, the radial basis kernel function is also called RBF kernel function, which is adopted by default by most support vector machines for modeling.

4.3. Introduction to Random Forest Algorithm. To put it simply, the random forest algorithm constructs a forest by randomly picking out a certain number of decision trees and does not allow any communication between each tree. Whenever a new test data appears in a forest, each tree independently determines the category of the test data. By voting, the type that gets the most votes is the final result of this test [27]. In fact, the final classification result for the sample test data is the majority of the independently determined category values in all the trees.

The general flow of the random forest algorithm is as follows.

1. The number of samples of the training set is set as $N$.
2. Firstly, repeated multiple sampling with reset is carried out from all data samples to obtain $T$ training sets $S_1, S_2, \ldots, S_T$ used to generate the decision tree.
3. $T$ training sets generate $T$ decision trees $C_1, C_2, \ldots, C_T$ accordingly. Set the number of input variables as $M$; each node will randomly select $m$ ($m < M$) variables from $M$ input variables and split the node using the best splitting method among $m$ variables (usually, the value of $m$ is fixed when the decision tree is generated).
4. Decision trees can grow as much as possible without pruning.
5. $T$ training sets $S_1, S_2, \ldots, S_T$ were tested by decision tree to obtain their categories $C_1(X), C_2(X), \ldots, C_T(X)$.
6. Vote $T$ decision trees to get the classification of test set $X$.

The number $m$ of feature extraction will deeply affect the correlation of decision tree and the accuracy of classification. The smaller the number $m$ is, the smaller the correlation of decision tree and the accuracy of classification will be. Therefore, the selection of $m$ is very important.

4.4. Comparison between Model Solution and Prediction Results. The 400 groups of data extracted before are divided into two parts: 80% of them are used as training sets and the remaining 20% as test sets to verify the classification effect of the model. Six attribute indicators including gender, education level, age, average monthly income, average monthly consumption, and marital status are taken as the input of the model, while whether to handle personal loan is taken as the output of the model, which constitutes a dichotomous problem.

Write programs through MATLAB [28]; 20% of each machine learning algorithm was randomly selected as the test set. A total of 10 calculations were carried out, and the predicted results were compared with the standard results to calculate the prediction accuracy of the three machine learning algorithms. Table 3 shows the accuracy of the three machine learning algorithms in each calculation in detail.

In order to clearly compare the prediction accuracy of BP neural network, SVM, and random forest algorithm in each run, a broken line graph as shown in Figure 6 is made.

In Figure 6, the random forest algorithm has the highest prediction accuracy among the three machine learning algorithms, with an average prediction accuracy of more than 94%. The prediction accuracy of BP network is the lowest; because BP network adjusts the weights of the network step by step along the direction of local optimization, it is easy to make the weights converge to the minimum point of the interval and make the algorithm fall into the maximum value of the interval, which makes the training effect of the network not good. At present, many scholars have used sparrow search algorithm, particle swarm algorithm, genetic algorithm, and so on to optimize the BP neural network to improve the accuracy [29–31]. The prediction accuracy of support vector machine is higher than that of BP neural network. On the one hand,
support vector machine can get better results with fewer sample training set data and has a better processing method for high-dimensional data. On the other hand, it can avoid the problem of local minimum compared with BP neural network due to structure selection. The prediction accuracy of random forest algorithm is the highest, because as an integrated learning algorithm, random forest has strong adaptability to data, can handle continuous and discrete variables, and is not easy to overfit.

5. Conclusion

The precision marketing based on consumer portrait from the perspective of machine learning is studied by this paper in detail. By extracting the attribute characteristics of consumers and constructing consumer portraits, a model label for mining potential loan customers is created. It is a predictive label, which can predict the future needs of consumers, further target consumers, and improve the possibility of marketing success. At the same time, the machine learning algorithms commonly used in classification are compared in this study, including BP neural network, SVM, and random forest algorithm. The results show that the prediction accuracy of random forest algorithm is the highest in 10 computation solutions, with an average accuracy of 94%, which can lay a foundation for the subsequent research.

The model established in this paper is effective to some extent in predicting whether consumers will handle personal loans. The characteristic data selected in the model can reflect consumers’ intentions. After obtaining the usable prediction model, it can import the information of consumers who want to make predictions, excluding the sample customers, into the model to predict whether these consumers will take out loans. In this way, it can reach the goal of marketers to find personal loan business customers, solve the time consumption of needle marketing in haystack, and improve the marketing success rate.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References


