Hindawi
Security and Communication Networks
Volume 2022, Article ID 1496170, 10 pages
https://doi.org/10.1155/2022/1496170

Research Article

Statistics and Analysis of Targeted Poverty Alleviation Information Integrated with Big Data Mining Algorithm

Meizhen Gao,[1] **Li Li**,[1] **and Yetong Gao**[2]

[1]*Jiaozuo Normal College, Jiaozuo 454000, Henan, China*
[2]*Xi'an University of Technology, Xi'an 710048, Shanxi, China*

Correspondence should be addressed to Meizhen Gao; gmz1970@jzsz.edu.cn

Received 18 February 2022; Revised 8 March 2022; Accepted 18 March 2022; Published 23 April 2022

Academic Editor: Chin-Ling Chen

To gain a more comprehensive and systematic understanding of the impact of government assistance to poor households on poverty reduction targets, a targeted poverty alleviation information statistics and analysis integrated with big data mining algorithm is proposed. Combined with the big data knowledge of the new era, according to the machine learning (ML) pipeline module in spark, a big data computing framework, combined with known data mining algorithms, massive sample data are used to replace random stratified sampling data for modeling and analysis, and random forest model, logistic model, and newly proposed waterfall model are constructed for poor households. Finally, through the comparative evaluation of several poor household identification models, the results show that when 100 real data test the accuracy of the three poor household models, the random forest model and logistic model are slightly reduced, which are 82% and 72%, respectively, but the waterfall model is basically unchanged, which is 83%, and the three models have little change. The new waterfall design proposed in this article has the advantage of a high percentage of sample reuse and can effectively prevent overfitting, and there is no need for massive data. It is a stable and reliable new model. The combination of targeted poverty reduction algorithms and big information technology and mining data can get the most common causes more accurate and convincing results. The right rib trunk and rib are often separated from the common cause because of the population.

1. Introduction

In the middle of the 20th century, getting rid of poverty has become the primary problem of governments in the international community and one of the problems that must be solved in the national economic development strategy. The main reason is that the global economic level is depressed. All countries regard economic development as the primary task of national development and the core content of international organizations to solve the problem of poverty alleviation [1]. Then, in the 1980s, the international community began to gradually realize that the reason for poverty is not the lack of economic income, but a social problem involving social resources, national policies, medical education, and other aspects, even related to race. China began to move towards the road of not relying solely on economic growth to solve the problem of poverty alleviation, in which sustainable development is one of the main strategies to solve the problem of poverty [2]. From the perspective of the state for the majority of poor households who need to be supported, the previous extensive poverty alleviation model has long been inapplicable to the poverty alleviation population under the current conditions. Afterwards, General Secretary Xi proposed targeted poverty alleviation. Targeted poverty alleviation is no longer the extensive regional poverty alleviation model but targeted at the poor population, that is, the smallest unit in poverty alleviation. The purpose of targeted poverty alleviation is to offset the economic that is one of the necessary measures to be taken to reduce the effectiveness of poverty alleviation due to growth [3, 4].

Poverty has always been one of the major problems faced. In order to ensure that China can build a well-off society as soon as possible, we must solve the problem of

poverty from the root [5]. The first step to solve the problem of poverty is to find out the real poor households and eliminate poverty from the source through targeted assistance to the real poor households. As an important part of big data technology, mining algorithm is mainly responsible for mining the potential logical relationship between data. The application core of data mining algorithm is to design the data mining model, use the data mining model to calculate the data, and then mine the potential associations. We need to collect all the data required by the data mining model as the basic data to provide data support for data mining. Then, the type of data is determined as the pattern division standard of data mining model, and the framework of data mining model is determined. Finally, the data mining model is established by summarizing the data characteristic parameters. Although detailed identification studies are not yet fully developed in China, in recent years, with the rapid development of Internet technology and big data technology, it may be faster and more efficient to use large data technology and data mining algorithms to accurately reduce poverty and more accurately identify real poor households. Compared to traditional machine learning algorithms, due to technological limitations and independent storage, they are used only for small amounts of data and are based on data sampling. The advent of big data technology can help us run machine learning activities, modeling on large amounts of data [6, 7]. Figure 1 shows the in-depth integration of big data and targeted poverty reduction.

2. Literature Review

Since the proposal of targeted poverty alleviation has only been put forward for four or five years, and the time in practical application is shorter, there is less research on targeted poverty alleviation. Targeted poverty reduction has been achieved in a short period of time, but with the strong support of the central government and the joint efforts of local governments, targeted poverty reduction has achieved remarkable results in recent years. Today, many Chinese scholars are making creative proposals to reduce poverty and make a detailed description and analysis in combination with the difficulties encountered in practical work. From the perspective of family structure, Zhao S. and others used the poverty measurement method to study the Multidimensional Poverty Situation of different types of family poverty. Finally, they came to the conclusion of what kind of family is more likely to fall into poverty [8]. Awajan and others take advantage of big data knowledge to help targeted poverty alleviation. They compare big data poverty alleviation with traditional poverty alleviation methods and conclude that big data targeted poverty alleviation has more advantages than traditional poverty alleviation methods, which is more conducive to identifying and helping poor households [9]. By Gaye and others traditional poverty alleviation technologies and models have been challenged. In the past, large-scale poverty alleviation models have been less regional-oriented, making it difficult to identify the poorest households. At the same time, many new problems will appear if the data of poor households are not updated for a long time; there is an urgent need to change the traditional model of poverty reduction. The new model of poverty reduction should be the comprehensive use of big data technology, to reduce target poverty and increase the efficiency of targeted poverty reduction [10]. Cui and others believe that there are still some limitations in simply taking the economic income of poor households as the poverty standard of poor households. Therefore, based on the third-party assessment and research task of targeted poverty alleviation, it is found that 13.5% of the surveyed farmers believe that some poor households who really need government assistance are missed by the government in the process of filing and card [11]. Vachkova and others analyzed and defined what targeted poverty alleviation is and expounded their views on why targeted poverty alleviation should be carried out. This paper gives a complete description of the poverty alleviation model since the reform and opening up, analyzes that the previous extensive poverty alleviation model is no longer applicable to the current environment, summarizes the key and difficult points of targeted poverty alleviation, and expounds the new working methods [12]. Visuwasam and others discussed and studied the mechanism of targeted poverty alleviation through big data technology and concluded that big data technology can be widely used in targeted poverty alleviation [13]. Granat and others proposed the research on the rural targeted poverty alleviation mechanism in Guizhou Province under the background of big data. Accurate identification is the most important link in targeted poverty alleviation. How to find out the poor households is the top priority [14]. Srivani and others studied the research on accurate identification in targeted poverty alleviation. The most important thing in exploring accurate identification is to explore the research on accurate identification through Multidimensional Poverty Measurement [15].

This article mainly uses big data technology and data mining algorithms to accurately identify poor households, establishes a variety of different data mining models under the big data computing framework spark, uses the models to identify and classify poor households, and finally compares and evaluates the results of the models. Then it classifies 100 real data with three poverty identification models, checks the identification accuracy of the three models on the real poverty data, and analyzes the modeling time of the three models.

3. Research Methods

3.1. Big Data Analysis Technology. Big data analysis technology: Big data analysis is based primarily on machine learning, cloud computing technology, and several data mining algorithms, but also on some new information analysis technologies. At the same time, it also depends on some new data analysis technologies, mainly including graph-based mining algorithm, group-based mining algorithm, and data network-based mining algorithm. At the same time, it adopts the fusion technology of object-based data connection. In recent years, big data analysis technologies for various fields have emerged one after another,

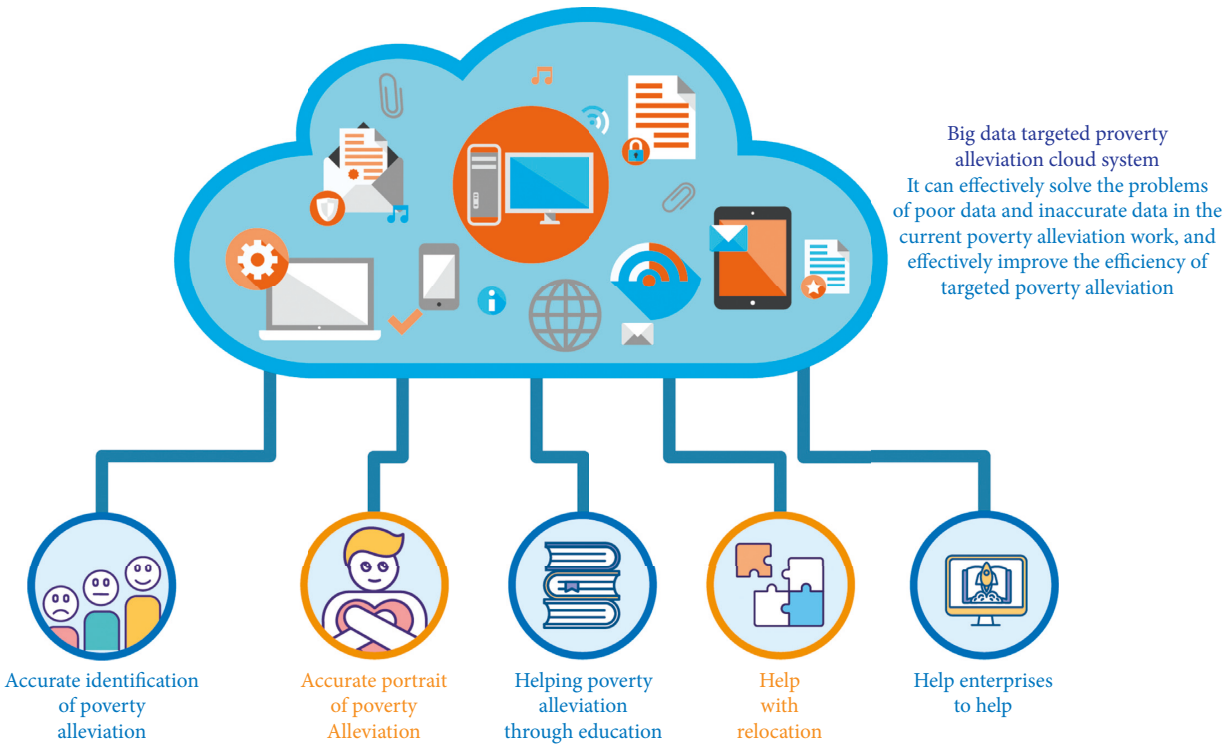

Figure 1: Deep integration of big data and targeted poverty alleviation.

mainly using network behavior analysis, semantic analysis, emotion analysis, syntax analysis, and user interest mining [16].

In practical application, most of the data are data without any rules. The amount of these data is usually very large, but they cannot be stored through relational database. Most of the data are incomplete and have a large amount of noisy data. However, this part of the data hides a large amount of valuable user data, which has very important potential value and academic knowledge. To obtain these valuable data, we must clean these data through data mining algorithms, filter out useless data, and finally retain valuable potential data. The more common data mining methods include clustering, anomaly and trend discovery, association rules, dependency models, etc. The data objects of data mining mainly come from the current mainstream relational databases, text data, web data sources, heritage databases, etc. After data cleaning, it enters the statistical stage. Statistical algorithms mainly include regression analysis and cluster analysis based on data multidisciplinary regression and autoregressive algorithms. Data mining is also known as knowledge discovery in a database. As the name suggests, data mining is from massive and complex data. The process of using certain algorithmic research means to find effective information with certain potential value and revealing significance hidden in data. As a process of decision support, data mining can automatically and intelligently analyze data by using artificial intelligence technology, statistical principle, machine learning, and pattern recognition methods, realize data visualization through visualization technology, make inductive and summary reasoning operations, and mine the laws contained in the data and provide correct guidance for decision makers in formulating strategies, reducing risks and preventing major mistakes [17]. As a multistep processing process, the knowledge in the database includes the operation stages of data selection, preprocessing, data conversion, data mining, result interpretation, and evaluation (see Figure 2).

3.2. Pretreatment of Multidimensional Poverty Characteristics of Farmers

3.2.1. Overview of Maslow's Hierarchy of Needs Theory. Maslow's demand hierarchy theory is one of the theories of behavioral science. According to the pyramid model, human needs from the bottom of the pyramid are divided into five levels: physiological needs, security needs, love and belonging needs, respect, and self-realization needs (Figure 3).

3.2.2. Selection of Poverty Characteristics Based on Maslow's Demand Level. The measurement indicators selected for the multidimensional poverty characteristics mainly form the following at the current stage of development: the basic needs of farmers and the overall characteristics of poor groups, the correlation between research indicators and poverty, the existing poverty research framework, etc. In the selection process, there is not only a lack of corresponding guiding theory, but also many restrictions on the acquisition of data due to various reasons. Therefore, it is difficult to make corresponding trade-offs and lack of norms and systems for the measurement and selection of multidimensional poverty characteristics.

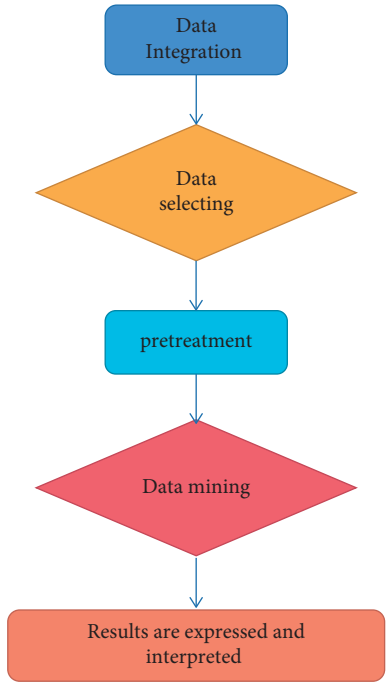

Figure 2: Process diagram of knowledge mining.

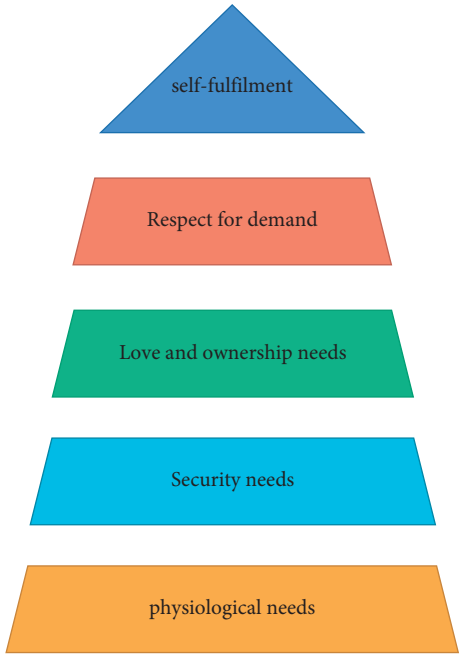

Figure 3: Maslow demand hierarchy pyramid model.

In the targeted poverty alleviation big data platform, various data of farmers are recorded. In addition to farmers' personal information, it also includes the most basic information of various types of farmers, such as food, clothing, housing, and transportation, as well as relatively high-level data on farmers' income, subsidies, medical treatment, and education. As we all know, the impact of various dimensions on the causes of poverty is different. For example, the importance of factors such as food, housing, and transportation must be greater than that of family labor force and education. Only after meeting the conditions of farmers' food, clothing, housing, and transportation can farmers have a higher level of pursuit. Moreover, in the big data platform for targeted poverty alleviation, there are some redundant records of farmers' basic information unrelated to the research objectives, which will affect the results of the experiment. Here, the basic information recorded by farmers in the platform is screened and divided, as shown in Figure 4.

3.2.3. Unsupervised Discretization of Basic Household Information. For the convenience of data mining, for the continuous data contained in the basic information of farmers, the unsupervised discretization of static attributes needs to be carried out before using association rule analysis.

Uncontrolled sorting of farm information means adjusting the interval values of the relevant digital attributes stored in the database and using separate intervals to represent all continuous data, which drastically reduces the number of attribute values and simplifies the initial data. The mining results can be improved with more buttons. The easier-to-operate, more logical representation, equally wide interval method is used for unsupervised discretization of farmer data and differentiation between two continuous data: household income per farmer age and heavy network. For dividing the age structure (Table 1).

Since the national poverty recognition standard for farmers' income is that the per capita annual net income is below 2300, the discretization of per capita net income of farmers' families is formulated according to the national poverty line standard (see Table 2).

The evaluation scores are discretized according to the hundred mark system, with excellent, good, medium, and poor, as shown in Table 3.

3.3. Identification Model of Random Forest Poor Households

3.3.1. Random Forest Algorithm. Random forest algorithm is a new and highly flexible data mining algorithm, an algorithm for integrating multiple decision trees. Therefore, this integrated learning method is often better than the prediction made by any single classification. The basic idea of random forest mainly lies in "random" and "forest." Each time the decision tree is established, some characteristic variables are randomly selected to build the tree through the selected characteristic variables, and then the tree building steps are repeated until the number of decision trees needs to be established. In this way, many independent trees will be established. These trees are the "forest" in the random forest, and the final classification result is determined by the results obtained by the established decision tree through voting [18, 19].

In the application scenario of random forest, it is generally used for classification prediction. The principle of solving the classification problem is usually as follows: generally, when using random forest algorithm for classification, because the random forest is constructed by many decision trees, in order to obtain the final result of the

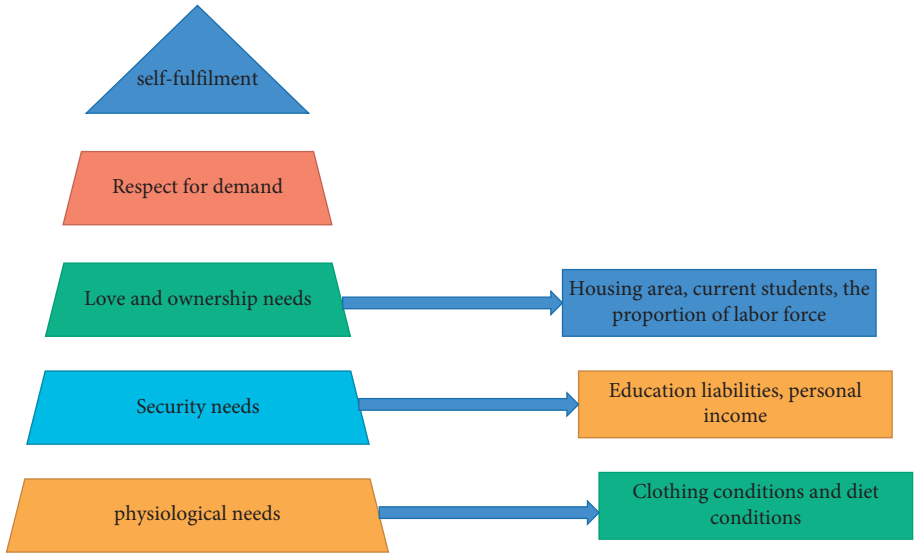

Figure 4: Screening and division of farmers' information.

Table 1: Age discretization structure.

Number	Age	Discretization
1	30 and below	Youth
2	30–49	Mid-life
3	50–59	Quinquagenarian
4	60 and above	Old age

Table 2: Income discretization structure.

Number	Per capita family net income is discrete
1	Under 2300
2	2300–4999
3	5000 and above

Table 3: Evaluation of separation and dispersion structure.

Number	Assess the score	Evaluation scores were discretized
1	More than 89	Ample
2	80–89	Good
3	60–79	Center
4	Less than 60	Difference

random forest classification, it is necessary to vote and summarize the classification results for each decision tree erected. Each decision tree is based on a relatively independent random sample with the same distribution. The creator of random forest combines bagging with random extraction method. This operation not only effectively reduces the correlation between each decision tree, but also improves the performance of random forest tree.

3.3.2. Model Construction. The information in this article is taken from detailed file and card information on poverty reduction in the province, but in order to make the analysis results true and effective, the data also includes noncard filing data. According to the card filing data of poor households and nonpoor households, a total of 34 million pieces of data are generated in proportion, including the data of poor households and nonpoor households. Different from the traditional statistical methods, this paper adopts large data massive sample data set to replace the traditional hierarchical random sampling to construct the data set. Based on both practical and technical considerations, this paper mainly refers to the international multidimensional poverty index in the selection of characteristic variables, which reflects which poverty attributes are easy to make people fall into poverty.

Spark, the big data computing framework used in this paper, is used to model and analyze poor households. First, spark needs to be started and the packages needed to build the model need to be imported. Next, import the original sample data set and preprocess all the data, including feature transformation, which helps to speed up the iterative calculation and model modeling. Finally, the processed data and the random forest model are jointly established to establish pipeline. In the process of random forest modeling, the parameters of the model can be set by setter or paramMap. In this paper, paramMap is selected to set the parameters, and the obtained model parameters are the best model parameters obtained by the final cross validation. Then, by importing the training set data for model training, the random forest poor household identification model, pipeline model, can be obtained, and the pipeline model is used to classify and predict the test set data [20]. The model prediction results are derived together with the real results of the test set and the probability value that the model prediction is a poor household for analysis and comparison. The original sample data used in this paper is 34 million family data. Each family selects the ten characteristic variables described above, and the sample data size is 10.36. According to the above steps, the sample size of the test package, which randomly divided the training package and

the test package between 70% and 30%, is about 10.2 million. The results of the random forest model classification on the experimental data are shown in Table 4, and the random forest model evaluation index is shown in Table 5.

The total running time of the random forest poor household identification model under spark is 3 hours and 36 minutes. In the confusion matrix, 1 represents poor households and 0 represents nonpoor households. The overall accuracy of the final result is 89.48% compared to the actual results of the test package, the precision is 91.65%, the recall is 86.68%, the FPR is 0.077, the specificity is 93.23%, and the AUC value is 0.9718. The sum of the number of poor households and nonpoor households identified by the model accounts for 89.48% of the total sample size, 86.68% of all real poor households are identified, and 91.65% of all identified poor households are real poor households. The ROC curve of the model is shown in Figure 5.

Table 4: Results of classification of poor households by the random forest model.

Confusion matrix		Predicted value	
		0	1
True value	0	4631822	387851
	1	662031	4270040

Table 5: Model evaluation indicators.

Model test index	
Overall model identification accuracy	88.37%
Precision ratio	90.54%
Recall ratio	85.57%
Specificity	91.22%
FPR	7.6%
AUC	96.07%

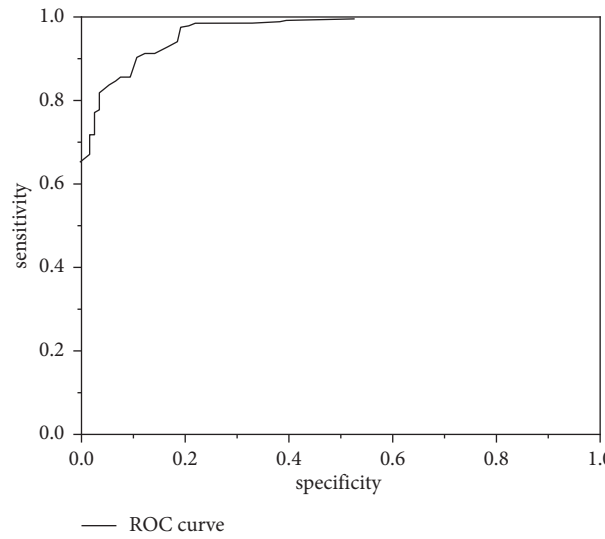

Figure 5: ROC curve of the model.

3.4. Logistics Poverty Identification Model

3.4.1. Introduction to the Logistic Algorithm. Assuming that there is a random variable X, the distribution form of the function of the random variable X is given by equation (1), and the probability density function of the random variable X is given by equation (2).

$$F(s) = P(X \le s) = \frac{1}{1 + e^{-(x-u)/r}}, \tag{1}$$

$$f(s) = F'(s) = \frac{e^{-(x-u)/u}}{\gamma\left(1 + e^{-(x-u)/r}\right)^2}. \tag{2}$$

Then it can be considered that x obeys the logistic distribution; in the above formula $\lambda > 0$ can be called shape parameter and u can be called position parameter. Logistic $F(s)$ distribution function is a sigmoid function with different properties. The shape of the sigmoidal function is a curve similar to the s-shape, and the center of symmetry of the function is $(\mu, 1/2)$, so it satisfies

$$F(-s + \mu) - \frac{1}{2} = -F(s + u) + \frac{1}{2}. \tag{3}$$

Because the sigmoidal function is an S-shaped curve, the function image grows around the point of central symmetry and changes significantly and changes very little at both ends of the curve. In the distribution function according to formula (3), if the value of the shape parameter Y is smaller, the growth range of the function image S-shaped curve around the center point will be larger [21].

3.4.2. Binomial Logistic Regression Model. Although the binomial logistic regression model is called the regression model, it is actually a classification model, which can be classified according to different segmentation points. The binomial logistic regression model can be expressed by conditional distribution $P(Y|X)$ in the form of parametric logistic distribution. In binomial logistic regression, variable y can be taken as 1 or 0, where x is still a random variable. The conditional distribution of the binomial logistic regression model is shown in

$$P(Y = 1|X) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)}, \tag{4}$$

$$P(Y = 0|X) = \frac{1}{1 + \exp(w \cdot x + b)}, \tag{5}$$

where $x \in \wp^n$ represents the input vector, $Y \in \{0, 1\}$ represents the predicted output variable, and $x \in \wp^n$ and $b \in \wp^n$ are the parameters of the two models. W is the model weight vector to be estimated by the algorithm, and B is also the parameter to be estimated by the algorithm, representing the offset of the function. Then, compare the probability of 1 with that of 0, take the maximum value, and divide this classification into the category with larger probability value.

In order to facilitate observation and calculation, the model can be simplified by expanding the weight vector to $w = (w^{w(1)}, w^{w(2)}, \ldots, w^{w(n)}, b)^T$ and the input vector to $w = (w^{w(1)}, w^{w(2)}, \ldots, w^{w(n)}, 1)^T$. The internal product of the vectors of this generation is the same as the first, and a simplified logistic regression model is shown in

$$P(Y = 1|X) = \frac{\exp(w \cdot x)}{1 + \exp(w \cdot x)}, \tag{6}$$

$$P(Y = 0|X) = \frac{1}{1 + \exp(w \cdot x)}. \tag{7}$$

The characteristics of logistic regression model: at this time, it is necessary to introduce a new concept, odds ratio odd. Assuming that the probability of an event is p, the probability of its nonoccurrence is $1 - p$. Odds ratio odd refers to the ratio between occurrence and nonoccurrence probability, which is $p/1 - p$. At this time, the logit function of log odds of the event is shown in the following equation:

$$\text{logit}(p) = \log\left(\frac{p}{1 - p}\right). \tag{8}$$

For logistic regression, equation (9) can be obtained by

$$\log\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = w \cdot x. \tag{9}$$

In the logistic regression model, the odds ratio (odds) of the probability of an event is expressed by the linear function of the input vector x. At the same time, through the definition of logistic regression model in the above two formulas, based on the linear function, the logistic regression model can be transformed into a probability calculation formula (see equation (10)).

$$P(Y = 1|X) = \frac{\exp(w \cdot x)}{1 + \exp(w \cdot x)}. \tag{10}$$

At this point, if you want the probability value of P to be infinitely close to 0, the value of the linear function must be infinitely close to negative infinity. If you want the probability value of P to be infinitely close to 1, the value of linear function should be infinitely close to positive infinity. The final model is called logistic regression model.

3.4.3. Estimation of Model Parameters. Finally, the logistic regression model can be obtained, as shown in equations (11) and (12).

Hypothesis:

$$P(Y = 1|X) = \pi(X), \tag{11}$$

$$P(Y = 0|X) = 1 - \pi(X). \tag{12}$$

Probability function (13) is shown in the following equation:

$$\prod_{I=1}^{N} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1 - y_i}. \tag{13}$$

See equation (14) for log likelihood function is as follows:

$$L(w) = [y_i(w \cdot x_i) - \log(1 + \exp(w \cdot x))]. \tag{14}$$

Assuming that w is calculated with the maximum probability of w, then the probability formulas for the true logistic regression model can be obtained by controlled study, as shown in

$$P(Y = 1|X) = \frac{\exp(\widehat{w} \cdot x)}{1 + \exp(\widehat{w} \cdot x)}, \tag{15}$$

$$P(Y = 0|X) = \frac{1}{1 + \exp(\widehat{w} \cdot x)}. \tag{16}$$

3.5. Waterfall Model for Poor Households

3.5.1. Waterfall Design. A waterfall model can be used to make classification assumptions. This model is mainly inspired by the normal distribution graph created by the rapid sand experiment and uses the graph model of the triangular structure (Figure 6).

The middle node of layer 0 accounts for 100% of the total data, while the middle node of layer 2 accounts for 50% of the total data, and the middle node of layer 4 accounts for 37.5% of the total data. By analogy, the proportion of the middle node of layer $2n$ in the total data is $C2n$. This is because by Stirling formulas:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{(1/12n) - (1n(9n)/(9n)^{\pi} - (9n)^{-\pi})}, \tag{17}$$

$$(n!)^2 \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^{2n} e^{(1/6n) - (21n(9n)/(9n)^{\pi} - (9n)^{-\pi})}, \tag{18}$$

$$(2n!) \approx \sqrt{2\pi 2n}\left(\frac{2n}{e}\right)^{2n} e^{1/24n - (1n(18n)/(18n)^{\pi} - (18n)^{-\pi})}. \tag{19}$$

Therefore, when $n \longrightarrow \infty$, the proportion of intermediate nodes in the total data tends to be close to 0; that is, the limit is 0.

3.5.2. Construction of the Waterfall Model. Due to the waterfall, it is more convenient to select the parameters of the model [22, 23]. First, start the big data computing framework spark. Since the waterfall model is a new model; there is no model in spark's machine learning library pipeline. When importing the required package, you only need to import other packages required for building the model. The sample data is also 34 million pieces of data, with a size of 10.36. The results of the waterfall model classification and assumptions in the experimental set of data are shown as follows. The actual class label in the left test set represents the waterfall model, which predicts the class label of the test set data and the predicted value on the right. The evaluation parameters of the waterfall model are shown in Tables 6 and 7.

The overall running time of spark is only 1 hour and 42 minutes. The actual results of the test package were

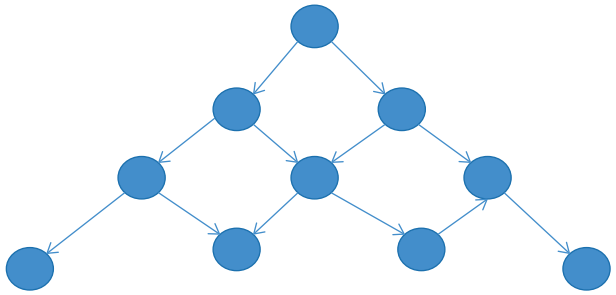

Figure 6: Waterfall model.

Table 6: Results of waterfall model classification of poor households.

Confusion matrix		Predicted value	
		0	1
True value	0	4452260	561631
	1	958013	4077148

Table 7: Model evaluation indicators.

Model test index	
Overall model identification accuracy	83.71%
Precision ratio	86.54%
Recall ratio	80.72%
Specificity	87.70%
FPR	10.1%
AUC	92.34%

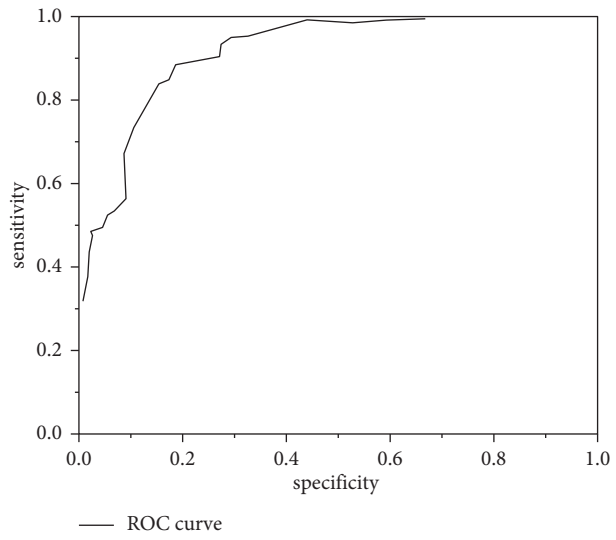

Figure 7: ROC curve of the model.

compared with the results predicted by the model, and the overall identification accuracy of the model was 83.71%, the accuracy was 86.54%, the recall rate was also the recall rate, and the TPR was 80.72%. . The feature is 87,700%, the FPR value is 0.1120, and the AUC value of the area along the ROC curve is 0.9234. The model is shown in the ROC curve in Figure 7.

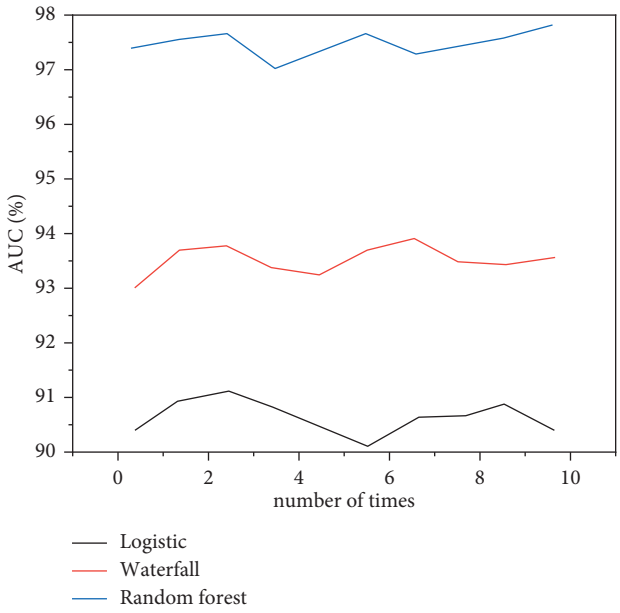

Figure 8: 10 AUC values of three models.

4. Results and Discussion

4.1. Multiple Selection Test Set Evaluation Model. In order to compare and evaluate the three models for identifying poor households, a new set of data that was not included in the model training and experiment was selected from the three sets of pilot models. The data with a data size of 30% of the total sample size are randomly selected from the new test set for times as the test set; that is, the data size of each randomly selected test set is about 10.2 million poverty-stricken households. Three models were tested to compare the average AUC of the three models and the stability of the three models [24]. In 10 experiments, the AUC of the random forest model was consistently higher than that of the logistics model and the waterfall model. The average AUC was 0.934 (Figure 8).

The changes of AUC values of the three models are small, which proves that the stability of the three models is very good. However, the operating time is much shorter than the other two models, with a random forest model running time of 3 hours and 36 minutes, a logistics model running time of 2 hours and 20 minutes, and a waterfall running time of only 1 hour and 42 minutes. It is more than half an hour of running even a small random forest model.

4.2. Testing the Model with Real Data. This is because 34 million pieces of data are generated proportionally from the filing and card filing data of poor households when building the model [25]. In order to test the accuracy of the constructed model on the filing and card filing data of fully real poor households, 100 fully real poor households' data were randomly selected, including 50 real poor households and 50 nonpoor households, and 100 real data were used to test the

Table 8: Results of the classification of poor households by the random forest model.

Confusion matrix		Predicted value	
		0	1
True value	0	41	7
	1	8	40

Table 9: Classification results of poor households by the logistic model.

Confusion matrix		Predicted value	
		0	1
True value	0	36	12
	1	13	35

Table 10: Results of waterfall model classification of poor households.

Confusion matrix		Predicted value	
		0	1
True value	0	41	7
	1	7	41

three models, respectively. The established random forest poverty identification model is used to test 100 real data. For the test results, see Table 8.

According to the classification results of 100 real data by the random forest poor households identification model constructed by spark, the overall identification accuracy of the model is 83%, and 41 of the 50 poor households are identified. 42 out of 50 nonpoor households were identified. The recognition accuracy of random forest poor households recognition model for 100 real data is slightly reduced.

The constructed logistic poor household identification model is used to test 100 real data, and the test results are in Table 9.

Through the classification results of 100 real data by the logistic poor households identification model built by spark, a big data computing framework, it can be seen that the overall identification accuracy of the model is 72, and 36 of the 50 poor households have been identified. The accuracy of the logistic poor household identification model in identifying the real data of 100 households also decreased slightly.

The constructed waterfall poor household identification model is used to test 100 real data, and the test results are in Table 10.

According to the classification results of 100 real data by the waterfall poor households identification model constructed by spark, the big data calculation framework, the overall identification accuracy of the model is 84%; 42 of the 50 poor households and 42 of the 50 nonpoor households are identified. The recognition accuracy of waterfall poor household recognition model for 100 real data is basically unchanged. When 100 real data are used to test the accuracy of the three poor household models, the random forest model and logistic model are slightly reduced, which are 82% and 72%, respectively, but the waterfall model is basically unchanged, which is 83%, and the three models have little change.

5. Conclusion

The essence of targeted poverty alleviation is that the government effectively identifies poor families and members, excavates the causes and extent of poverty, and carries out practical and effective assistance, so as to fundamentally break the barriers of poverty. With the rapid economic development, the national income level is seriously unbalanced. The previous extensive regional poverty alleviation method has long been inapplicable. In this case, targeted poverty alleviation came into being. Through the research, the logistic algorithm, random forest algorithm, and the newly proposed waterfall model in data mining are found. The newly proposed waterfall model has the advantages of high sample reuse rate, can effectively prevent overfitting, and has no demand for massive data. It is a stable and reliable new model. Visualize the research results, design and implement the abnormal poverty assistance model obtained in the above research work, and apply it to the actual poverty alleviation work, so that the staff can get the effectiveness of the government's policy assistance to the village by importing the basic information and assistance data of farmers into the system. It can make timely and effective adjustments to the existing abnormal assistance phenomenon, so as to provide help for the government in the targeted poverty alleviation work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This study was sponsored by Jiaozuo Normal College.

References

[1] H. Yao and J. Fu, "Analysis of smooth implementation of industry poverty alleviation considering government supervision," *Mathematical Problems in Engineering*, vol. 2021, Article ID 5554595, 13 pages, 2021.

[2] R. Narayan, S. Gadag, S. P. Cheruku et al., "Chitosan-glucuronic acid conjugate coated mesoporous silica nanoparticles: a smart ph-responsive and receptor-targeted system for colorectal cancer therapy," *Carbohydrate Polymers*, vol. 261, no. 2, pp. 117893–117898, 2021.

[3] J. Wu, Z. Feng, and S. Anwar, "Assessing the pollution convergence across Chinese cities by considering ecological indicators: a continuous distribution dynamics approach," *Ecological Indicators*, vol. 126, no. 2, pp. 107552–107559, 2021.

[4] B. He and L. Yin, "Prediction modelling of cold chain logistics demand based on data mining algorithm," *Mathematical Problems in Engineering*, vol. 2021, Article ID 3421478, 9 pages, 2021.

[5] B. Khan, R. Naseem, M. A. Shah et al., "Software defect prediction for healthcare big data: an empirical evaluation of machine learning techniques," *Journal of Healthcare Engineering*, vol. 2021, Article ID 8899263, 16 pages, 2021.

[6] D. Wang, T. Miwa, and T. Morikawa, "Big trajectory data mining: a survey of methods, applications, and services," *Sensors*, vol. 20, no. 16, pp. 4571–4577, 2020.

[7] D. Xia, F. Ning, and W. He, "Research on parallel adaptive canopy-k-means clustering algorithm for big data mining based on cloud platform," *Journal of Grid Computing*, vol. 18, no. 2, pp. 263–273, 2020.

[8] S. Zhao, M. Hu, Z. Cai, Z. Zhang, T. Zhou, and F. Liu, "Enhancing Chinese character representation with lattice-aligned attention," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

[9] I. Awajan, M. Mohamad, and A. Al-Quran, "Sentiment analysis technique and neutrosophic set theory for mining and ranking big data from online reviews," *IEEE Access*, vol. 9, no. 99, pp. 1–16, 2021.

[10] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of support vector machine algorithm in big data background," *Mathematical Problems in Engineering*, vol. 2021, Article ID 5594899, 9 pages, 2021.

[11] Y. Cui, "Intelligent recommendation system based on mathematical modeling in personalized data mining," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6672036, 11 pages, 2021.

[12] S. N. Vachkova, E. Y. Petryaeva, R. B. Kupriyanov, and R. S. Suleymanov, "School in digital age: how big data help to transform the curriculum," *Information*, vol. 12, no. 1, pp. 33–36, 2021.

[13] L. M. M. Visuwasam and D. P. Raj, "A distributed intelligent mobile application for analyzing travel big data analytics," *Peer-to-Peer Networking and Applications*, vol. 13, no. 6, pp. 2036–2052, 2020.

[14] J. Granat, J. M. Batalla, C. X. Mavromoustakis, and G. Mastorakis, "Big data analytics for event detection in the iot-multicriteria approach," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4418–4430, 2020.

[15] B. Srivani, N. Sandhya, and B. Padmaja Rani, "Literature review and analysis on big data stream classification techniques," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 24, no. 3, pp. 205–215, 2020.

[16] J. Liu, G. Shi, J. Zhou, and Q. Yao, "Prediction of college students' psychological crisis based on data mining," *Mobile Information Systems*, vol. 2021, Article ID 9979770, 7 pages, 2021.

[17] Y.-X. Shi, B.-K. Zhang, Y.-X. Wang, H.-Q. Luo, and X. Li, "Constructing crop portraits based on graph databases is essential to agricultural data mining," *Information*, vol. 12, no. 6, pp. 227–231, 2021.

[18] X. Wei, "A classification method of tourism English talents based on feature mining and information fusion technology," *Mobile Information Systems*, vol. 2021, Article ID 5520079, 9 pages, 2021.

[19] S. Khan and S. Alqahtani, "Big data application and its impact on education," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 17, pp. 36–42, 2020.

[20] S. Zhao, M. Hu, Z. Cai, and F. Liu, "Dynamic modeling cross-modal interactions in two-phase prediction for entity-relation extraction," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

[21] T. Ramalingeswara Rao, S. K. Ghosh, and A. Goswami, "Mining user-user communities for a weighted bipartite network using spark GraphFrames and Flink Gelly," *The Journal of Supercomputing*, vol. 77, no. 6, pp. 5984–6035, 2021.

[22] E. Prasetyo and A. Putra, "Implementasi waterfall model dalam pengembangan sistem informasi eksekutif penduduk," *Journal of Information Systems and Informatics*, vol. 3, no. 1, pp. 213–224, 2021.

[23] N. Rofiq, A. Perdananto, and N. Jaya, "Penerapan model waterfall pada aplikasi bank sampah," *Infotech Journal of Technology Information*, vol. 7, no. 1, pp. 19–26, 2021.

[24] D. J. Kalita, V. P. Singh, and V. Kumar, "Two way threshold based intelligent water drops feature selection algorithm for accurate detection of breast cancer," *Soft Computing*, vol. 26, no. 5, pp. 2277–2305, 2021.

[25] G. A. Loosmore and R. T. Cederwall, "Precipitation scavenging of atmospheric aerosols for emergency response applications: testing an updated model with new real-time data," *Atmospheric Environment*, vol. 38, no. 7, pp. 993–1003, 2004.