Engineering Solutions to Mechanics, Marine Structures and Infrastructures

Eng. Solut. Mech. Mar. Struct. Infrastruct.
2024.1(3).8
ISSN：3006-2837
https://doi.org/10.58531/esmmsi/1/3/8

Article

Risk Prediction and Influencing Factors Analysis of Poverty Monitoring Households —— based on Liangshan Yi Autonomous Prefecture, Sichuan Province

Siting Hong[1], Chuantao Li[1], Jielang Huang[2]

[1] College of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China
[2] College of Economics, Guangdong Ocean University, Zhanjiang 524088, China

Academic Editor: Dapeng Zhang < zhangdapeng@gdou.edu.cn>

Received: 1 August 2024; Revised: 25 August 2024; Accepted: 30 August 2024; Published: 2 September 2024

Abstract: This study utilizes Lasso Regression and MAHAKIL over-sampling to address multicollinearity and label imbalance issues in poverty-stricken household data from Liangshan Yi Autonomous Prefecture. Then a Bayesian-optimized LightGBM based on the TPE procedure was used for classification prediction and post-processing of the model using order-preserving regression. The results indicates that the accuracy of Xide increased by 6.62% and 1.35%, and the F1 score increased by 6.07% and 5.22%, respectively. Final, This study analyzes the impact of various factors on the risk of returning to poverty through decision plots, interaction maps, and waterfall charts focusing on per capita income, subsistence allowances, and working hours. This study not only demonstrates the important role of data mining in accurately identifying and preventing the recurrence of poverty, but also provides a scientific basis for the government and social organizations to help them formulate more effective poverty alleviation policies and development strategies.

Keywords: Lasso Regression; Bayesian Optimization; Shapley Additive Explanations; Isotonic Regression; MAHAKIL

1. Introduction

On November 23,2020, The State Council of the People's Republic of China officially announced that all 832 state-level poverty counties have been lifted out of poverty, marking

Citation: Hong Si., Li C., Huang J. Risk Prediction and Influencing Factors Analysis of Poverty Monitoring Households —— based on Liangshan Yi Autonomous Prefecture, Sichuan Province. Eng. Solut. Mech. Mar. Struct. Infrastruct., 2024, 1(3), doi: 10.58531/esmmsi/1/3/8

the successful completion of the national poverty alleviation target. However, many families are still at risk of returning to poverty, so effective measures and policies need to be taken to avoid the problem of poverty to ensure that the achievements of poverty alleviation are consolidated and sustained. Article 59 of the Law on Promoting Rural Revitalization of the People's Republic of China stipulates that people's governments at all levels should establish and improve dynamic monitoring and early warning and assistance mechanisms for people who are prone to return to poverty and cause poverty, so as to consolidate and expand the achievements of poverty alleviation and effectively link them with rural revitalization. The Opinions of the CPC Central Committee and The State Council on the Key Work of Comprehensively Promoting Rural Revitalization in 2023 pointed out that consolidating and expanding the achievements of poverty alleviation is the bottom line task of comprehensively promoting rural revitalization, and the monitoring and assistance mechanism should be constantly improved to resolutely prevent the phenomenon of returning to poverty.

In order to prevent the return to poverty, it is necessary to explore the influencing factors of the return to poverty and apply the right medicine. The traditional view mostly holds the "income poverty theory", believing that income is the only standard to measure poverty. Economist Aatiassen put forward the "lack of ability" theory, that poverty stems from the lack of feasible ability to obtain and enjoy a normal life. According to the reality of China, domestic scholars expanded the influencing factors of the return to poverty. Zheng Ruiqiang, Cao Guoqing, Yao Xiaoping and Fan Hesheng respectively studied the influence of policy environment, natural environment and the subject of on the return to poverty.

However, Zou Wei and Fang Yingfeng believe that compared with the single income dimension, the prevention of poverty from the perspective of multidimensional poverty faces a stronger impact in both spatial and temporal dimension. Restricted by development conditions and other conditions, different regions face different poverty factors, and there may be multiple poverty factors in the same region [1].

In addition to the return to poverty factors in the spatial and temporal dimensions, Wang Sangui et al also conducted in-depth research from internal risk factors such as potential risk objects and external risk factors such as economic environment. Yu-jie luo and Zhang Jigang scholars under the perspective of sustainable living area Chinese problem to explore, the livelihood capital is divided into natural, financial, material, human and social capital, through the livelihood capital to identify Chinese poverty risk measure, to judge the Chinese risk and implement targeted preventive measures to [2].

Previous studies have laid the foundation for this paper, and also provided valuable experience for people to understand the factors of poverty return and the corresponding solutions. But most scholars from the macro level, Chinese standing in the perspective of enterprise, government thinking and put forward solutions, however, poverty in different regions of different, different factors between different degree of interaction, so not one-sided through the analysis of a single area, single characteristics has a universal conclusion.

To solve the above problems, this paper combined with big data and the background of the era of artificial intelligence, established a prediction model with strong generalization ability, and in two poverty county, for example in the macro perspective of each characteristics of positive and negative contribution, in the microscopic perspective quantify the influence of each sample process and analysis of any two characteristics between the interaction of the dependent variables.

2. Methodology

2.1 Lasso Regression

As the number of features in the initial dataset is up to 74, direct training is affected by multicollinearity. To address the above problems, lasso regression was used to feature screen the initial dataset. Lasso regression is an unbiased estimation for handling high-dimensional complex collinearity data, and the basic idea is to add the L1 norm penalty term [3] to fitting a generalized linear regression. Compared with the L2 norm penalty term of ridge regression, Lasso Regression can better compress the coefficients of unimportant variables to 0, so it is widely used in feature reduction and resistance to overfitting. Where the constructed penalty function is as follows:

$$\min_{w}\frac{1}{2n}||y\text{-}Xw||_2^2+\lambda||w||_1=\min_{w}f(x)\text{=}g(x)\text{+}h(x) \tag{1}$$

Among them, y is the poverty alleviation risk matrix, X is the matrix of factors influencing poverty alleviation risk, w is the regression coefficient matrix of the factors affecting poverty alleviation risk, λ is the regularization coefficient. Since the first term of $\min_{w} f(x)\text{=}g(x)\text{+}h(x)$ is a differentiable convex function, while the second term is a non-differentiable convex function, and the full-Lipschitz condition, the Lasso Regression can be solved using the proximal gradient descent method. Where the L-Lipschitz condition is defined as a normal number L for any x_1 and x_2 have:

$$||\nabla f(x_1)\text{-}\nabla f(x_2)||_2^2\leq L||x_1\text{-}x_2||_2^2 \tag{2}$$

In this paper, the second order Taylor expansion at x_k and expressed by the Lipschitz constant L:

$$\hat{f}(x)\approx f(x_k)+\nabla f(x_k)^T(x\text{-}x_k)+\frac{L}{2}||x\text{-}x_k||_2^2+\lambda||x||_1 \tag{3}$$

$$=\frac{L}{2}\left\|x\text{-}(x_k\text{-}\frac{1}{L}\nabla f(x_k))\right\|_2^2\text{+const} \tag{4}$$

Where $\frac{1}{L}$ is the iteration step, allowing $x=x_k-\frac{1}{L}\nabla f(x_k)$ to achieve the maximum minimization, so the iteration:

$$x_{k+1}=\underset{x}{\operatorname{argmin}}\frac{L}{2}\left\|x-(x_k-\frac{1}{L}\nabla f(x_k))\right\|_2^2+\lambda\|x\|_1 \tag{5}$$

The equation (5) is reorganized as follows:

$$x_{k+1}=\frac{L}{2}\sum_{i=1}^{n}(x^{(i)}-(x_k^{(i)}-\frac{1}{L}\nabla f(x_k)))^2+\lambda\sum_{i=1}^{n}|x_i| \tag{6}$$

When each dimension of the parameter x in formula (6) is optimized, the result is optimal, so solving the Lasso Regression is the following goal:

$$\underset{x}{\operatorname{argmin}}\frac{L}{2}(x-z)^2+\lambda|x| \tag{7}$$

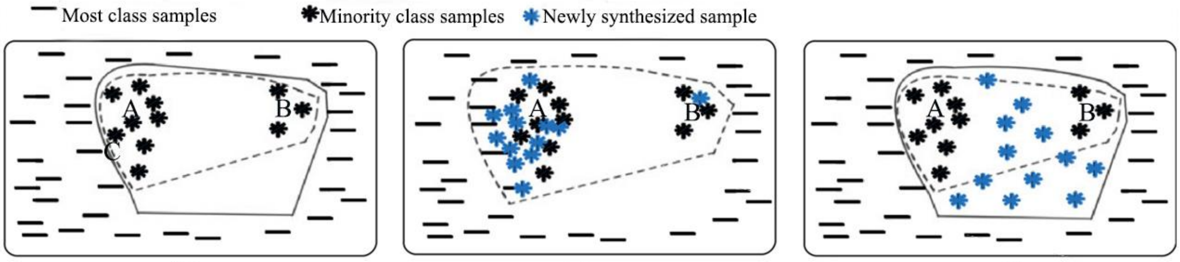
For formula (7), the near-gradient solution is used, and the solution result is:

$$x=\begin{cases} z+\frac{\lambda}{L}, & z<-\frac{\lambda}{L} \\ 0, & |z|\leq\frac{\lambda}{L} \\ z-\frac{\lambda}{L}, & z>\frac{\lambda}{L} \end{cases} \tag{8}$$

From formula (8), the optimal result depends on a suitable regularization coefficient.

2.2 MAHAKIL Oversampling

Since the number of positive labels of poverty risk elimination in the initial training set was 11900, while the number of negative labels of poverty risk not elimination was only 3714, the proportion of both was close to 3.204:1. To overcome the influence of imbalance on model learning, we oversample negative example label samples using the MAHAKIL method. Traditional oversampling methods such as SMOTE and ADASYN use the K nearest neighbor method to select the nearest point and the sample point to produce new samples, but this method has a drawback, that is, the new samples will widen the boundary of the collection of a few samples, making the resulting new samples easily confused with the majority of samples. Among them, the schematic diagram of the traditional oversampling method widening the boundary of a few samples is shown in Figure 1 [4].



(A) (B) (C)

Figure. 1. The boundary map by traditional oversampling. (A) Before sampling. (B) After sampling. (C) Ideal sampling

As shown in Figure 1 (A) and Figure 1 (B), the dashed line is the estimated decision boundary, while the solid line is the real decision boundary. The oversampling method based on the K-neighbor method will produce close to the repeated or wrong sample points and broaden the estimated decision boundary. As shown in Figure 1 (C), the ideal sampling situation should ensure that the estimated and true decision boundaries are basically equal. To overcome the above problems, an oversampling method based on genetic chromosome theory, MAHAKIL, was proposed by Kwabena et al. Specifically, a schematic diagram of the process of synthesized sample points by this method is shown in Figure 2.

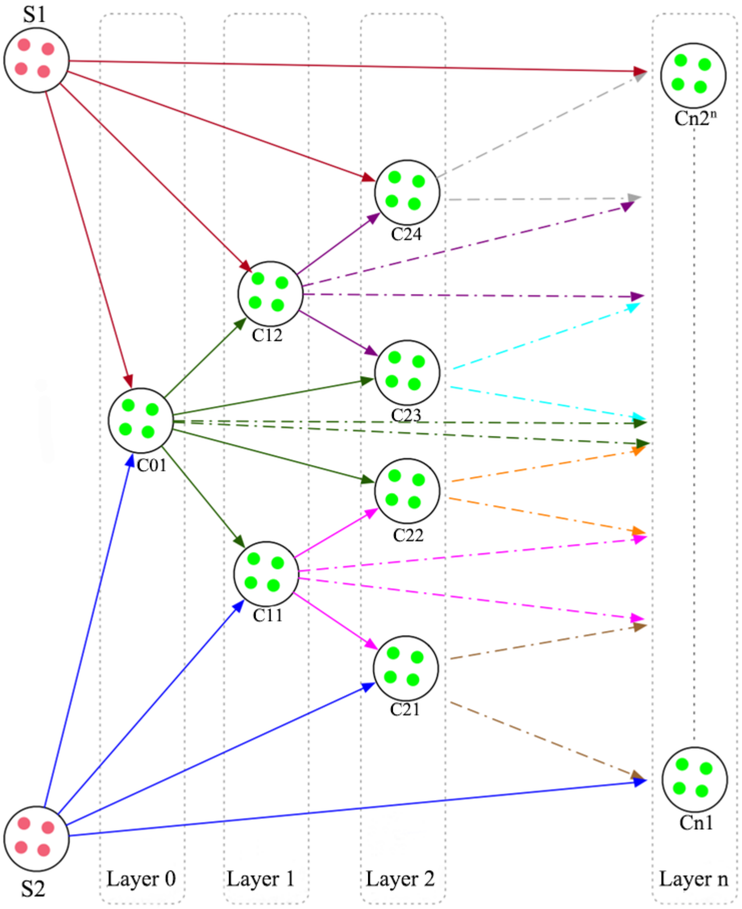


Figure. 2. MAHAKIL Schematic of synthetic sample points.

As shown in Figure 2, the red points represent the minority sample points of poverty alleviation risk, and the green points represent the synthetic sample points of poverty alleviation risk. The goal of this method is to generate a synthetic example with unique and common attributes. The algorithm steps are as follows:

（1） Mark most sample point sets as N_{max}, and a few sample point sets as N_{min}.

（2） The number of sample points to be synthesized is calculated to obtain the length k of the final generated set of sample points P and N_{min}.

（3） The array X_{new} storing the synthetic samples and the variable X_{newchk} recording the number of synthetic samples were initialized to 0.

(4) The Mahalanobis distance $D^2=(x-\mu)^T\Sigma^{-1}(x-\mu)$ is calculated for N_{min}, where μ is the mean and Σ is the covariance matrix.

(5) Save the matrix D^2 and store its corresponding samples in the array $N_{mindist}$ in descending order of Mahalanobis distance.

(6) The index of the sample median is found and is indicated by $N_{min}=\frac{k}{2}$.

(7) N_{min} is divided into two sets $N_{bin1}=\{y_1,y_2,\ldots,y_{mid}\}$ and $N_{bin2}=\{y_{mid+1},y_{mid+2},\ldots,y_k\}$ by the index N_{mid} and array $N_{mindist}$.

(8) For each $y_i\in N_{bin1}$ and $y_i\in N_{bin2}$, a unique label $l_i(i=1,\ldots,mid)$ is assigned in order.

(9) Select two samples y_a and y_b with the same subscript from 1 to mid, then take the average $x=average(y_a,y_b)$, add x to X_{new} and add the value of X_{newchk} by 1.

(10) If $X_{newchk}<T$, it means that the number of synthesized samples is insufficient. Then, the newly synthesized samples in X_{new} are paired with N_{bin1} and N_{bin2}. Respectively, and the step 9 is repeated to obtain $X_{new[j]}(j=1,2)$. If $X_{newchk}<T$ is still true, pair the current $X_{new[j]}$ with its direct parent set, then with the precursor set of the direct parent set, and repeat step 9 in subsequent operations.

(11) If $X_{newchk}\geq T$, all samples in X_{new} are assigned minority class labels and added to dataset N.

2.3 LightGBM Classification and Prediction Model

LightGBM is an integrated learning model that mainly improves on the histogram algorithm and gradient-based unilateral sampling algorithm based on XGBoost, which can effectively solve the problem [5], which takes a long time in the process of determining the optimal split point. Therefore, in this paper, the LightGBM classification prediction model was used to predict the risk of return to poverty.

In the histogram algorithm, for the previous round of training, the gradient of each sample can be represented as the degree of error in the prediction of the sample. Thus, LightGBM retains all instances with large gradients and employs a strategy of random sampling in a certain proportion for instances with small gradients. For the gradient calculation of each sample, O is the training dataset on a fixed node in the decision tree, and the variance gain of the split feature j of the d node at the point is expressed as:

$$V_{j|O}=\frac{1}{n_O}\left(\frac{\left(\Sigma_{\{x_i\in O:x_{ij}\leq d\}}g_i\right)^2}{n^j_{l|O}(j)}\right)+\left(\frac{\left(\Sigma_{\{x_i\in O:x_{ij}\leq d\}}g_i\right)^2}{n^j_{r|O}(j)}\right) \tag{9}$$

In the gradient-based one-sided sampling algorithm, a is the proportion of larger gradient samples, $b\in(0,1-a)$ is the proportion of randomly selected smaller gradient samples, pre-determined values of a and b. The sample with the top $a\times100\%$ data was taken as dataset A, then randomly selected $b\times100\%$ data in the left $(1-a)\times100\%$ data as dataset B. When calculating the information gain, ensure that discarding some smaller gradient samples does not affect the model training, so the retained smaller gradient samples are multiplied by the coefficient $\frac{1-a}{b}$. After the last iteration, the sample data is sorted in gradient descending order, and the final calculated gain is:

$$\tilde{v}_j(d)=\frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i+\frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{\eta_l^j(d)}\right)+\left(\frac{(\sum_{x_i\in A_r} g_i+\frac{1-a}{b}\sum_{x_i\in B_r} g_i)^2}{\eta_r^j(d)}\right) \tag{10}$$

2.4 a Bayesian Optimization Based on the TPE Process

Bayesian optimization is a method to optimize the objective function by establishing a probabilistic model between the hyperparameters and the performance. By constantly observing the evaluation results of the objective function, it constantly updates the posterior probability distribution of the objective function, so as to choose the next hyperparameters to be evaluated more reasonably. The proxy models commonly used for Bayesian optimization are mainly Gaussian process regression and TPE processes.[6] In the problem of poverty return risk prediction, the effect of model optimization will have an impact on the prediction results. Therefore, this paper uses Bayesian optimization based on TPE process for model hyperparameter optimization . The main steps are as follows:

(1) Each hyperparameter is constrained to follow a known distribution, from which it is subsequently sampled. This paper selects a Gaussian distribution with the mean of 0.

(2) Random sampling of the hyperparameters X (two-dimensional matrix, the number of rows is n, the number of columns is the number of the hyperparameters m to be searched by the model), and the true loss of the model y when applying these hyperparameters is calculated.

(3) The TPE process is set to fit the agent function $X\rightarrow y$, and $p(x|y)$ is divided as follows: when y is less than y^*, $p(x|y)=l(x)$; when y is greater than y^*, $p(x|y)=g(x)$. The following equation can be obtained from the above division:

$$p(x)=\int_R p(x|y)\,p(y)dy=\int_{-\infty}^{y^*} p(x|y)p(y)dy+\int_{y^*}^{+\infty} p(x|y)p(y)dy \tag{11}$$

For the selection y^*, a hyperparameter γ can be set to represent the quantile y, that is $p(y<y^*)=y$, the value is set to 0.25. Therefore, formula (11) can be simplified into the following form:

$$\gamma l(x)+(1-\gamma)g(x) \tag{12}$$

(4) The EI function is selected as the acquisition function to evaluate the acquisition point. The EI function can represent the average elevation level of y relative to the threshold y^* when given x, and the formula is as follows:

$$EI_{y^*}(x)=\int_{-\infty}^{+\infty} \max(y^*-y,0)P_M(y|x)dy \tag{13}$$

(5) Take the $P_M(y_i|x_i)$ of $X_{candidate}$, for each x_i (1 D vector, length m), select the best x^* from the acquisition point, and bring it into the model to calculate the real loss y^*.substitute formulas (10), (11) and (12) into formulas (13):

$$EI_{y^*}(x)=\int_{-\infty}^{y^*}(y^*-y)\frac{l(x)p(y)}{\gamma l(x)+(1-\gamma)g(x)}dy=\frac{\int_{-\infty}^{y^*}(y^*-y)p(y)dy}{\gamma+(1-\gamma)\frac{g(x)}{l(x)}} \tag{14}$$

Specifically, the EI function value is proportional to the reciprocal of its denominator. When γ is determined, the denominator size depends on the ratio $\frac{l(x)}{g(x)}$ of the probabilities at both ends of x, so the best acquisition point is the acquisition point that maximizes the ratio.

(6) Add (x^*,y^*) to the set (X,y) into a new (X,y), and repeat steps 3 through step 5 until the maximum number of iterations is reached.

2.5 Isotonic Regression

In the conventional classification prediction, the probability of the model output is prone to deviate from the probability of the true probability of the category of the sample. In the prediction of the risk of return to the poverty, the deviation between the model output and the actual situation will have an impact on the implementation of the decision. To solve this problem, this paper uses order-preserving regression to post-process the model.

Sequence-preserving regression is a technique for fitting free-form lines to a series of observations, such that the fitting lines are monotonic at any position and as close to the observations as possible [7]. In this paper, PAV algorithm is used to solve the order-preserving regression model. This method has only one constraint, that is, the space of function is monotonically increasing function. The main steps are as follows:

(1) It is assumed that each observed variable y_i satisfies a normal distribution $N(u_i,\sigma^2)$ and satisfies $u_i \geq u_j$ when $x_i \geq x_j$.

(2) Taking step 1 as the monotonicity constraint, the points violating the monotonicity constraint form a monotonicity sequence with its adjacent points, and the points in their range satisfy an identical distribution. That is, when $u_{i-1}<u_i$, the points within the interval $[x_{i-1},x_i]$ satisfy the normal distribution $N(\frac{u_i+u_{i-1}}{2},\sigma^2)$.

(3) If the resulting new distribution also violates the monotonicity constraint in the next point comparison, which is $u_{i+1}>\frac{u_i+u_{i-1}}{2}$, so the points within the interval $[x_{i-1},x_{i+1}]$ satisfy the normal distribution $N(\frac{u_i+u_{i-1}+u_{i+1}}{3},\sigma^2)$.

(4) Steps 2 and 3 are repeated until the distribution does not change when the monotonicity constraint is not violated.

2.6 Shapley Additive Explanations

The shapley additive explanations model is a local accuracy, missing ability, and consistent additive feature attribution method based on game theory and local interpretation, which can produce feature properties for each instance. In the task of analyzing the factors affecting the risk of returning to poverty, the shapley additive explanations model can identify the causal relationship between influential characteristics and whether the risk of returning to poverty is eliminated, but also reveal the interactive influence of multiple influencing factors [8]. Therefore, the shapley additive explanations model model was used to analyze the influencing factors of the return to poverty risk prediction model.

The shapley additive explanations model creates a simplified input by mapping by $x=h_x(z)$ from x to z. Based on z, the original model $f(x)$ can be approximated as a linear function of a binary variable as follows:

$$f(x)=g(z)=\varphi_0+\sum_{i=1}^{M}\varphi_i z_i \tag{15}$$

Where $z=\{0,1\}^M$, M is the number of input features, $\varphi_0=f(h_x(0))$ and φ_i is the feature attribute value, the calculation formula is as follows:

$$\varphi_i=\sum_{S\in F\{i\}}\frac{|S|!(M-|S|!-1)!}{M}[f_x(S\cup\{i\})-f_x(S)] \tag{16}$$

$$f_x(S)=f\left(h_x^{-1}(z)\right)=E[f(x)|x_s] \tag{17}$$

In formula (17), F is the non-zero set of input in z, S is the subset of F excluding the i feature, and φ_i is the unified measure of additive feature properties, namely the shapley additive explanations value.

Due to the extremely high complexity of accurately calculating $E[f(x)|x_s]$, many approximate solution methods such as Tree-SHAP have been developed. Tree-SHAP calculates the original complexity of $E[f(x)|x_s]$ as $O(TL2^M)$, given the number of a tree T and the maximum number of leaves in any tree L; given the maximum depth D of any tree, it calculates the original complexity of $E[f(x)|x_s]$ as $O(TLD^2)$, which reduces the time complexity from the higher order exponential level to the quadratic level. And the tree structure chosen in this paper is the LightGBM after the Bayesian optimization.

3. Data Processing

3.1 Index Selection

3.1.1 County Selection

Due to the difference in the characteristics of the data set, this paper selected the tracking data of the households aged 18 to 78 in 15 counties in Liangshan Yi Autonomous Prefecture, Sichuan Province, respectively. The sample size distribution of each county is shown in Table 1.

Table 1. Sample size distribution by county in the original total data set.

The name of the county	Sample size	Percentage of volume
Xichang City	56	0.27%
Huidong County	1162	5.54%
Huili City	151	0.72%
Ningnan County	581	2.77%
Ganluo County	1812	8.64%
Yuexi County	2463	11.75%
Xide County	1465	6.99%
Zhaojue County	1272	6.07%
Jinyang County	1876	8.95%
Bushi County	939	4.48%
Meigu County	3860	18.41%
Leipo County	1686	8.04%
Yanyuan County	1534	7.32%
Muli Tibetan Autonomous County	399	1.90%
Pugh County	1706	8.14%

In this paper, Xide and Meigu counties were used as the test set, and the data from the remaining counties were combined as the training set.

3.1.2 Feature Selection

The influencing factors that induce the return to poverty can be roughly divided into intrinsic inducing factors and extrinsic inducing factors, and the subordinate factors are shown in Table 2.

Table 2. Table of internal and external induction factors.

Intrinsic predisposing factors	Extrinsic precipitating factors
Individual-level factors	Natural factors
Relevant factors at the family level	Pro-poor policy factors
Factors that cause poverty due to illness	Factors of the quality of grassroots cadres

In order to facilitate the accurate further study of return to poverty factors, this paper takes 108 characteristics, including income, as key factors, based on Table 2 and existing studies, affecting whether to return to poverty, and conducts an in-depth study based on this.

3.2 Data Cleaning

3.2.1 Missing Value Processing

In this paper, missing values were detected and visualized by the invalid matrix, and found a large number of missing values in some features, as shown in Figure 3.

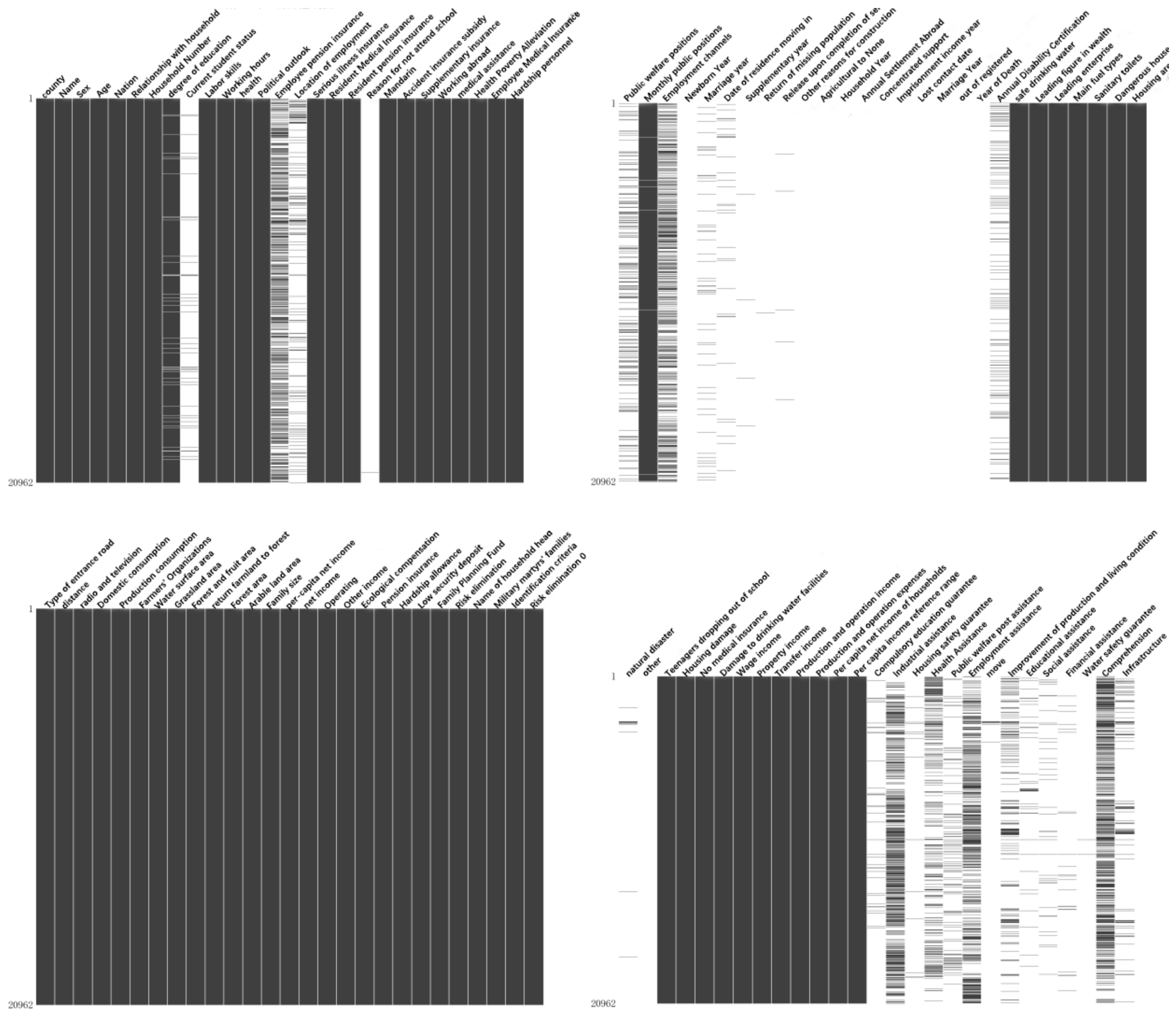


Figure. 3. Visualization of feature missing values cases.

In order to improve the robustness of the model, the missing values in each column were counted, and the proportion of missing features reached 80%. For the deleted features and the proportion of missing values are shown in the appendix.

After excluding the above features with more than 80% of missing values, there are still a few missing values in the remaining features. After analysis, it is known that the meaning of missing values in binary variables is the same as "no", so this missing value is filled as "no" in this paper.

3.2.2 Outlier Processing

Some data in the data set are not standardized, which are regarded as outliers and removed as shown in Table 3.

Table 3. Outlier processing table.

The name of the feature	Culling reasons	Reject sample size
Public welfare positions	There are non-standard data such as "09" and "07".	4

The name of the feature	Culling reasons	Reject sample size
The return of the lost population	The risk of returning to poverty for the sample of missing persons in the sample is 10.	19
Labor skills	There is non-standard data such as "empty".	1

After outlier processing, the total data set has 20938 data and 108 features.

3.3 Discrete Type Data Coding

In order to input the data into the model for training, this paper encodes the discrete-type data. For discrete data for one element with multiple values, and labels for discrete data with only one value for an element. Among them, the correspondence of single-thermal coding features is shown in Table 4, and the characteristics of label coding are shown in Table 5.

Table 4. Characteristic single-heat encoding table.

The name of the feature	Correspondence between monothermal encoding features
Compulsory education guarantees	Persuasion to return, living allowance for boarders, and door-to-door teaching
Industry support	Aquaculture industry, processing industry, forestry and fruit industry, consumption assistance, planting industry
Housing security	Renovation of dilapidated houses
Health support	Medical assistance, participation in basic medical insurance for urban and rural residents, individual payment subsidy, serious illness insurance
Public welfare post assistance	Cleaners, rangers, grass rangers
Employment assistance	Cash-for-work, labor export, subsidies for migrant workers, skills training, and employment of business entities
Production and living conditions have improved	Production and living conditions have improved
Educational assistance	Production and living conditions have improved
Social assistance	Social donations
Financial assistance	Microfinance
Comprehensive coverage	Temporary assistance, subsistence allowance, disability allowance, special poverty support, poverty prevention insurance
Infrastructure development	Infrastructure development

Table 5. Characteristic tag encoding table.

variable	A description of the coding situation
Degree	Blank value:0

variable	**A description of the coding situation**
	Illiterate or semi-literate:1
	elementary school:2
	Bachelor's degree or above:3
	junior high school:4
	high school:5
	The second year of higher vocational college:6
	College:7
	First year of undergraduate studies:8
	The third year of undergraduate:9
	The third year of higher vocational college:10
	Second year of undergraduate:11
	The third year of regular high school:12
	Fourth year of technician college:13
	Second year of technician college:14
	The second year of regular high school:15
	The first year of higher vocational college:16
	Fourth year of undergraduate:17
	First year of Technician College:18
	The third year of secondary vocational education:19
	Ninth grade:20
	Second year of secondary vocational education:21
	The first year of regular high school:22,
	Third year of technician college:23
	The first year of secondary vocational education:24
	Master's degree or above:25
Labor skills	Blank value:0
	General labor:1
	Weak or semi-labor force:2
	No labor:3
	Loss of labour:4
	Skilled Labor:5
Health status	Blank value:0
	Healthy:1
	Suffering from a serious illness:2
	Handicapped:3
	Long-term chronic illness, disability:4
	Suffering from a serious illness, disability:5
	Disability, long-term chronic illness:6

variable	A description of the coding situation
	Long-term chronic illness:7
	Disabled, suffering from a serious illness:8
Employment channels	Blank value:0
	Self-directed migrant work:1
	Poverty alleviation public welfare posts:2
	Organized migrant work:3
	Poverty alleviation workshop:4
Main fuel types	Blank value:0
	Firewood:1
	Clean energy:2
	Other:3
	Dry animal manure:4
	Coal:5
Type of entrance route	Blank value:0
	Hardened road:1
	Dirt roads:2
	Gravel road:3

4. Results And Discussion

4.1 Lasso Regression Feature Screening

Since the number of features in the initial data set is as high as 74, if it is trained directly, the model will not only be affected by multicollinearity, but also bear the burden of training time. To address the above problems, lasso regression was used to feature screen the initial dataset.

In this paper, we first draw the ridge plot of each regression coefficient about the regularization coefficient, and choose the appropriate regularization coefficient search interval by observing the convergence of the regression coefficient. Among them, the ridge map drawn in this paper is shown in Figure 4.

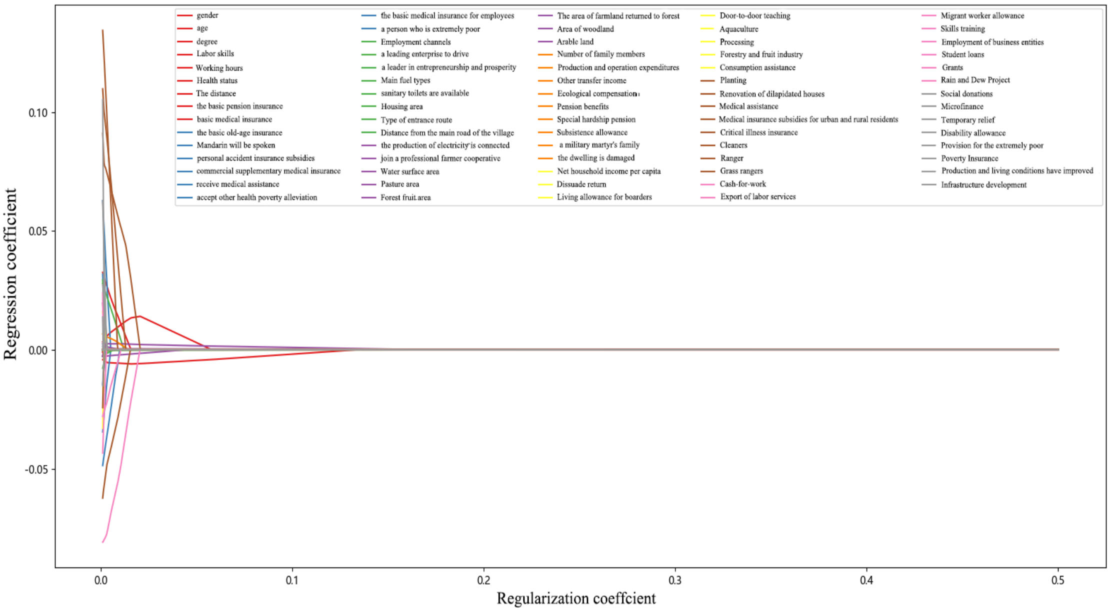


Figure. 4. Regression coefficient ridge plot.

According to the convergence of Figure 4, we search in the range of 0.001 to 0.5 by cross validation, and the best regularization coefficient is 0.0057428714357178594. Finally, the number of selected features was 28, and the average variance expansion factor decreased from 11.54318645 to 3.535605151, which significantly reduced the effect of multicollinearity on the model. Specifically, the feature names and the variance inflation factors after feature screening using lasso regression are shown in Table 6.

Table 6. Feature names and variance expansion factor after feature screening.

The name of the feature	Variance inflation factor
age	11.47438206
Labor skills	8.20400106
Working hours	1.992588038
Health status	2.791092343
The distance between the place of work and the hometown	1.402549839
Whether Mandarin will be spoken	2.316154285
Main fuel types	5.524835284
Housing area	12.32007085
Distance from the main road of the village	1.500268393
Pasture area	1.418060483
Forest fruit area	1.109111046
The area of farmland returned to forest	1.233426835
Area of woodland	1.44517691
Arable land	2.046919796

The name of the feature	Variance inflation factor
Number of family members	7.684832927
Production and operation expenditures	1.575012376
Other transfer income	1.298602592
Ecological compensation	1.186866115
Pension benefits	2.006317097
Special hardship pension	1.015630256
Subsistence allowance	1.963926689
Net household income per capita	18.60370631
Planting	1.507090407
Medical assistance	1.35807478
Participate in the individual payment subsidy of basic medical insurance for urban and rural residents	1.665854759
Cleaners	1.095686065
Export of labor services	1.672394505
Skills training	1.584312132

4.2 Analysis of the Model Prediction Effect

In conclusion, the final model established in this paper is: Lasso Regression + MAHAKIL + LightGBM prediction + sequence regression, while the benchmark model is LightGBM prediction. Before each model prediction, Bayesian optimization based on the training set is used by a ten-fold cross-validation. Among them, the best hyperparameter combination of each model is shown in Table 7.

Table 7. Best hyperparameter combination for each model.

model	The best combination of hyperparameters
benchmark model	'n_estimators': 481 'max_depth': 35 'learning_rate': 0.1182827604856063 'subsample': 0.581614058601332 'min_child_weight': 1 'gamma': 0.923517035793613 'colsample_bytree': 0.6129998201452649 'reg_alpha': 0.3623935688054989 'reg_lambda': 0.4860325677688167

model	The best combination of hyperparameters
Lasso Regression	'n_estimators': 768 'max_depth': 59 'learning_rate': 0.02696576158228669 'subsample': 0.3048122257429559 'min_child_weight': 1 'gamma': 0.00014084157110738627 'colsample_bytree': 0.363624289678255 'reg_alpha': 0.5612039044735743 'reg_lambda': 0.35629932751865506
Lasso Regression+MAHAKIL	'n_estimators': 1000 'max_depth': 75 'learning_rate': 0.04667427362300975 'subsample': 0.33791404773553085 'min_child_weight': 1 'gamma': 0.6199397820063137 'colsample_bytree': 0.2438288140136796 'reg_alpha': 0.05775296096018548 'reg_lambda': 0.7253815663352604
Lasso Regression+MAHAKIL+Order-preserving regression	

Comparing the prediction effect of the final model, ablation experiments were conducted on two test sets in Xide County and Meigu County. The accuracy and F1 scores of each model are shown in Table 8.

Table 8. Comparison of the prediction effects of each model on the test set.

model	Xide County Accuracy	Xide County F1 score	Meigu County Accuracy	Meigu County F1 score
LightGBM forecast	0.8784	0.6529	0.8651	0.7571
Lasoo Regression+LightGBM forecast	0.8913	0.6723	0.8750	0.7767
Lasso Regression+MAHAKIL+LightGBM forecast	0.8715	0.6488	0.8692	0.7801
Lasso Regression+MAHAKIL+LightGBM forecast+Order-preserving regression	0.9446	0.7136	0.8786	0.8093

As can be seen from Table 8, the final model increased by 6.62% and 6.07%; and F1 by 1.35% and 5.22% compared with the other three models. In conclusion, the model established in this paper has good generalization ability on the test set, providing a set of schemes for the risk prediction of residents in different counties that can be applied in a smaller number of features and wider scenarios.

4.3 Analysis of the Factors Influencing The Risk of Return to Poverty

4.3.1 Macro-Analysis of the Influencing Factors

Decision graph is a powerful visualization tool for macroscopic analysis of shapley additive explanations model, which shows how each sample is affected by individual features and then changes from the benchmark value to the [9] of the final predicted value. In the classification problem, the shapley additive explanations model benchmark value is the expected value of the predicted probability, and the benchmark value calculated in this paper is 0.8647, and the decision maps drawn in Xide counties and Meigu counties are shown in Figure 5.

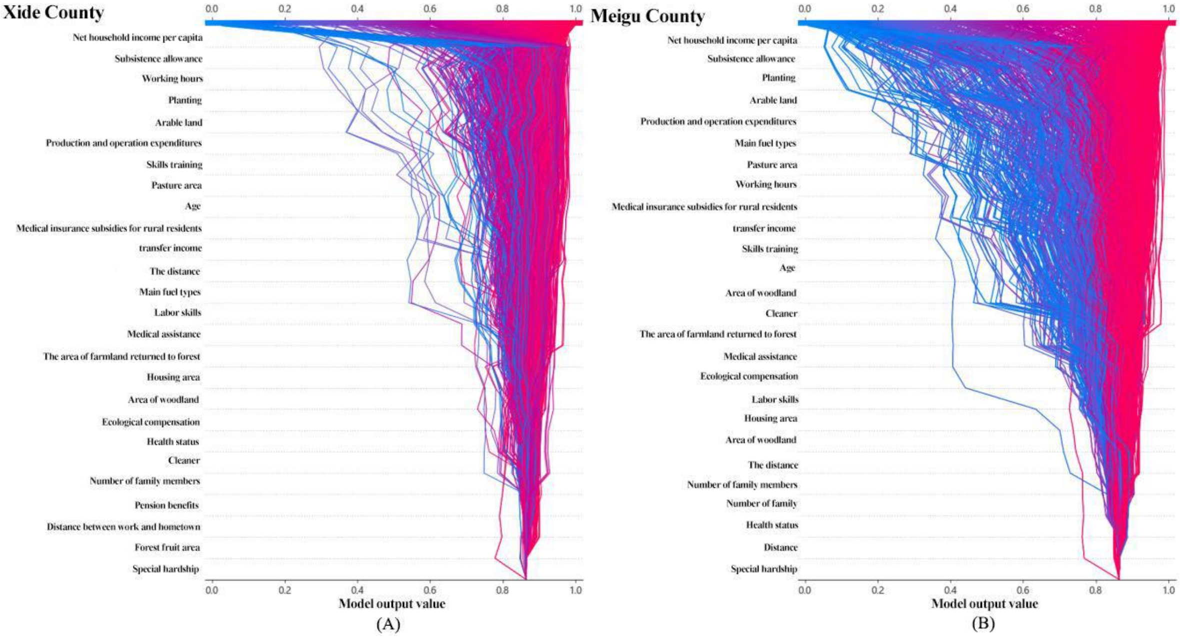


Figure. 5. Visualization of decision maps of Xide counties and Meigu counties. (A) Hide County decision map. (B) Meigu County decision map

As can be seen from Figure 5, compared with Meigu County, the poverty-stricken households in Xide County are more affected by the per capita net income and subsistence allowance, and the influence is significant, while the poverty-stricken households in Meigu County are more smoothly affected by various factors compared with Xide County. For Xide County, 13 factors, such as family per capita net income, are prone to the risk of return to poverty, while 5 factors, such as the number of family population, have no significant impact on the risk of return to poverty; for Meigu County, the decision chart has a clear dividing line, and 10 factors, such as ecological compensation, have no significant impact on the risk of return to poverty.

4.3.2 Interaction Analysis of the Influencing Factors

Interaction dependence diagram is a shapley additive explanations model analysis of any two features interaction between the dependent variables of visualization tools, from figure 5, family per capita net income is the two counties affect the risk of Chinese factors, so this paper choose low gold, working time the two biggest factors, respectively draw Xide County and Meigu county about family per capita net income interaction diagram, to explore the family per capita net income and the three factors affecting the interaction between [10]. For each contributing factor, the color of the scatter depends on the household per capita net income, thus reflecting the effect of the household per capita net income on the shapley additive explanations values of the current characteristics. The interactive dependence maps of Xide County and Meigu County drawn on subsistence allowance and working time are shown in Figure 6 and Figure 7, respectively.

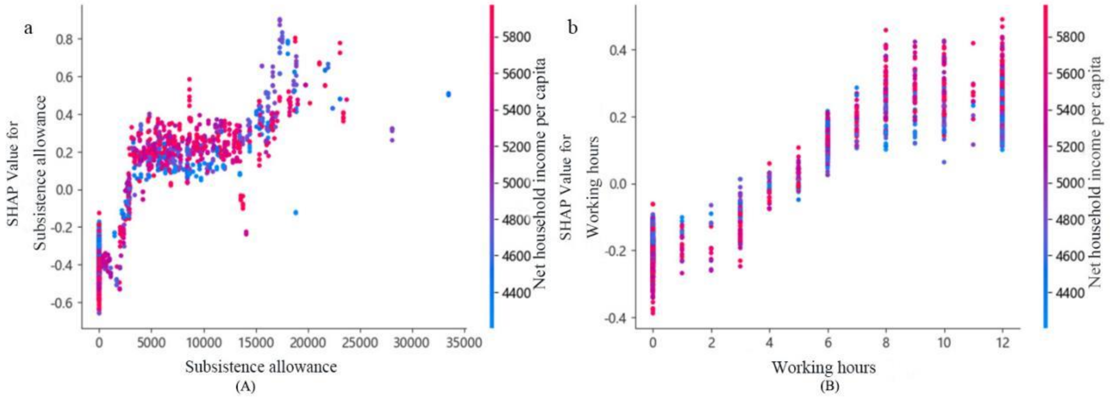


Figure. 6. Heide County interaction dependency graph visualization. (A) Interactive dependence chart of subsistence allowance fund. (B) Interactive dependence diagram of working time.

As can be seen from Figure 6 (A), when the minimum allowance of about 4000 yuan is the threshold of the positive and negative influence of this factor, the shapley additive explanations value increases significantly in the range of 0 to 5000 yuan, the shapley additive explanations value increases significantly in the range of 5000 to 150,000 yuan, and the change increases significantly in the range of 15,000 to 20,000 yuan. The higher per capita net income of families is concentrated when the subsistence allowance is 0 to 15,000 yuan, which indicates that many families out of poverty still rely on the subsistence allowance for the per capita net income.

As can be seen from Figure 6 (b), when the working time is about 5 months, the threshold is the positive and negative influence of this factor. The shapley additive explanations value of different months of working time is large, and the per capita net income of the family is evenly distributed, etc. The length of the working time will not significantly cause the income gap.

To sum up, the risk of returning to poverty in Xide County is mainly caused by the employment skills, labor ability, consumption level and transportation problems of the poverty-stricken households. The traditional forms of migrant work cannot play a significant positive impact on consolidating the achievements of poverty alleviation.

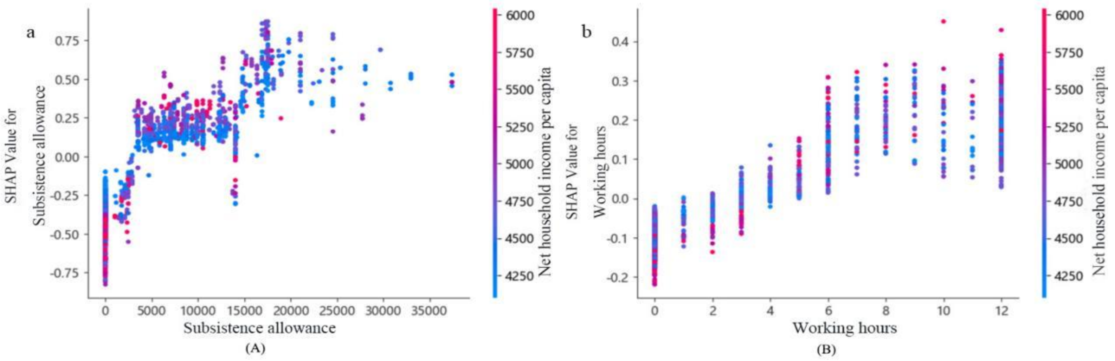


Figure. 7. Meigu County interaction dependency graph visualization. (A) Interactive dependence chart of subsistence allowance fund. (B) Interactive dependence diagram of working time.

As can be seen from Figure 7 (A), when the subsistence allowance is about 5000 yuan, the threshold of the positive and negative influence of the factor, the elimination of the risk is within the range of 0 to 5000 yuan, and the longitudinal distribution is relatively stable within 5000 yuan to 15,000 yuan, and less within the range of 15,000 to 35,000 yuan. The distribution of SHAP values of subsistence allowances in Meigu County is discrete, and most of the subsistence allowances of poverty-stricken households with large per capita net income are less than 5,000 yuan, which indicates that the poverty-stricken households in Meigu County have a low dependence on family income support.

According to Figure 7 (b), the threshold for the positive and negative impact of this factor is when the working time is around 4 months. Within the range of 0 to 4 months, the working time of poverty-stricken households with higher per capita net income has a greater negative impact on eliminating the risk of returning to poverty; Within the range of 4 to 12 months, the working time of poverty-stricken households with higher per capita net income has a greater positive impact on eliminating the risk of returning to poverty, and the longer the working time, the higher the average per capita net income level of households.

To sum up, the poverty-stricken households in Meigu County have a high level of labor skills and working ability, and migrant work is the main driving force for the households to consolidate the achievements of poverty alleviation and promote economic growth.

4.3.3 Microscopic Analysis of the Influencing Factors

Waterfall diagram is a visualization tool of shapley additive explanations model that can observe the changes in dependent variables for each sample. It is designed to show the interpretation of individual predictions. Waterfall at the bottom of the start from the expected value of the model output, each row shows the positive or negative contribution of each feature how the value from the expected model on the data set output to the predicted model output, red for positive contribution, and blue for negative contribution, in order to make the prediction probability quantification, this paper to shapley additive explanations value and output probability processing as follows:

$$x_{new}=-\ln\left(\frac{1}{x}-1\right) \tag{18}$$

In formula (18), x_{new} is the processed data in the waterfall diagram, and x is the raw data predicted by the LightGBM function.

In this paper, one sample is randomly selected to draw the waterfall map from Xide county and Meigu County respectively. The specific data of the selected samples are shown in Table 9 and Table 10, and the visualization results are shown in Figure 8 and Figure 9 respectively.

Table 9. Xide County waterfall map analysis data.

Feature Name	numerical value	Feature Name	numerical value
age	30	Number of family members	6
Labor skills	1	Production and operation expenditures	0
Working hours	6	Other transfer income	0
health	1	Ecological compensation	0
The distance between the place of work and the hometown	1369.009	Pension benefits	0
Main fuel types	1	Special hardship pension	0
Housing area	40	Subsistence allowance	18858
Distance from the main road of the village	2.3	Planting	0
Pasture area	0	Medical assistance	0
Forest fruit area	0	Participate in the individual payment subsidy of basic medical insurance for urban and rural residents	0
The area of farmland returned to forest	0	Cleaners	0
Area of woodland	0	Skills training	1
Arable land	0	Net household income per capita	4086.67

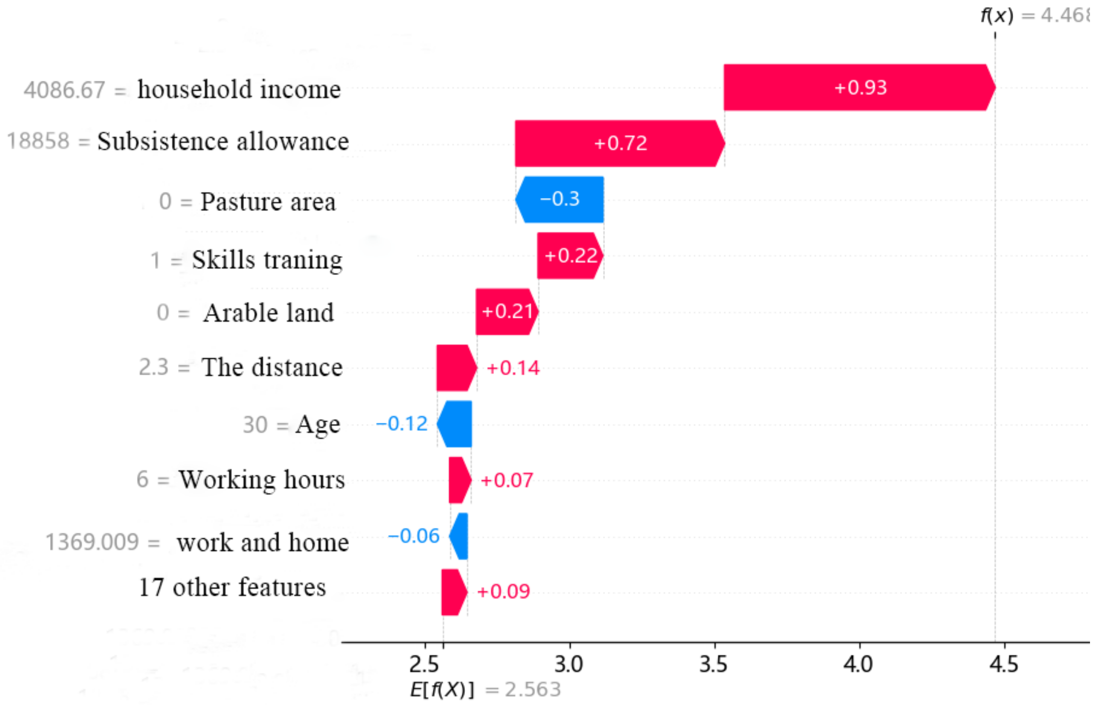


Figure. 8. Visualization of the Xide County waterfall maps.

Combined with table 9 and figure 8, Xide county of the poverty households its workers location from home 1369.009 km and the age of 30 caused 0.18 negative effects, the reason may be the poverty households have six people, and no grassland area and arable land area this means that the family mainly depends on the members of the annual six months migrant workers income to make a living, this single source of income in the face of family spending is insufficient. Fortunately, the government has provided strong assistance to the subsistence allowance and employment for the poor households, making the households out of the risk of returning to poverty. However, for long-term consideration, the government should further develop local employment assistance, promote the development of local industries, and improve the production and economic capacity of such poverty-stricken households.

Table 10. Meigu County waterfall map analysis data.

Feature Name	**numerical value**	**Feature Name**	**numerical value**
age	40	Number of family members	5
Labor skills	1	Production and operation expenditures	698
Working hours	6	Other transfer income	59.71
health	1	Ecological compensation	0
The distance between the place of work and the hometown	1481.43851	Pension benefits	0

Feature Name	numerical value	Feature Name	numerical value
Main fuel types	1	Special hardship pension	0
Housing area	73	Subsistence allowance	0
Distance from the main road of the village	1.2	Planting	0
Pasture area	0	Medical assistance	0
Forest fruit area	0	Participate in the individual payment subsidy of basic medical insurance for urban and rural residents	0
The area of farmland returned to forest	0	Cleaners	0
Area of woodland	7	Skills training	0
Arable land	11.08	Net household income per capita	4760.8

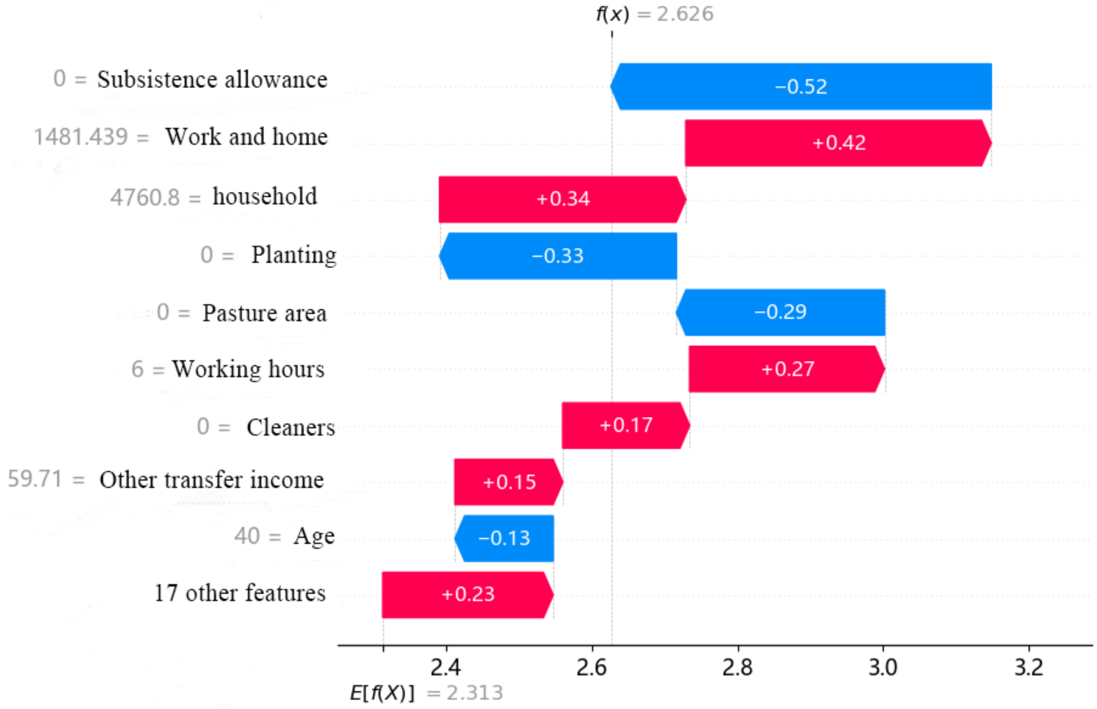


Figure. 9. Waterfall map visualization of Meigu County.

According to Table 10 and Figure 9, it can be seen that the family in Meigu County with a pasture area of 0, no planting, age of 40 and no subsistence allowance had a negative impact of 1.27. This suggests that these factors increase the risk of the family falling back into poverty. Although there is no low gold support, but the poverty still get rid of the Chinese risk, the reason may be that the income source of the poverty-stricken household is not single. Although there is no support of grassland and planting industry, the family has the agricultural support of cultivated land, woodland and other production economic support, which provides agricultural support for the family. In addition, the diversified sources of income enable such poverty-stricken households to leave their hometown to work and guarantee the basic material living standards of the families.

5. Conclusions

5.1 Conclusions Based on the Prediction Model

The accuracy of the model established in this paper in Xide County and Meigu County increased by 6.62% and 1.35%, respectively, and the F1 score increased by 6.07% and 5.22%, respectively. In order to verify the performance of the final model, the model ablation experiment, the results show that the final model relative to the benchmark model and the ablation experiment three model prediction effect is significantly improved, this shows that the model established in this paper has strong generalization ability, can significantly reduce the poverty difference between different regions on the prediction accuracy of the model.The main conclusions of the study may be presented in a short Conclusions section, which may stand alone or form a subsection of a Discussion or Results and Discussion section.

5.2 Conclusions Based on the Shapley Additive Explanations

Through the shapley additive explanations model used in this paper, in Xide County, the risk of returning to poverty is mainly caused by the employment skills, labor ability, consumption level and transportation of the households in the anti-poverty households, and the migrant work is the main driving force for the households to consolidate the achievements of poverty alleviation and promote economic growth.

5.3 The Shortcomings of the Risk Prediction Model

(1) MAHAKIL Oversampling method, which can alleviate the problem of class imbalance, may also lead to overfitting because it increases its number through the replication of a few class samples.

(2) Although using Bayesian optimization based on a TPE procedure enables efficiently finding optimal hyperparameters, the optimal configuration may be missed if the optimization space is improperly defined or the search number is insufficient.

(3) The risk of returning to poverty is a problem that changes over time, and the model needs to be constantly updated and calibrated with new data, otherwise its prediction effect will decrease over time.

5.4 Future Perspectives of the Risk Prediction Model

(1) As more data accumulate, prediction accuracy can be improved by continuously iterating the model. In addition, other more advanced machine learning algorithms or deep learning technologies can be explored to further improve model performance.

(2) Establish a real-time data collection and analysis system to timely detect the potential risk of return to poverty and take preventive measures.

(3) Developing modules that can evaluate the effect of the government's poverty alleviation policies will provide a basis for the formulation of more effective social policies.

Acknowledgments: After two months, we finally completed the writing of topic selection, data collection and model building. First of all, I would like to thank the instructor for his support and help in this process. Under the careful guidance of the teacher for many times, we continue to improve and improve, and finally this paper can be successfully completed. Thanks later to the authors of this citation, whose previous experience and results provide a variety of ideas for this paper. Finally, I would like to thank all the team members. Without their perseverance and joint efforts, this paper would not have been completed so successfully. Thanks again to all the teachers and students who have helped this paper. Without your help, this study would not have been so smooth.

Funding: This research was funded by Guangdong Ocean University Innovation and Entrepreneurship Training Program project, grant number CXXL2024148.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References:

1. Jiang et al. EXAMINING MULTI-LEVEL POVERTY-CAUSING FACTORS OF FARM HOUSEHOLD. ISPRS - International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences. 2019; XLII-4/W20: 49-53. Doi:10.5194/isprs-archives-XLII-4-W20-49-2019
2. Tan et al. Spatial Differentiation and Influencing Factors of Poverty Alleviation Performance Under the Background of Sustainable Development: A Case Study of Contiguous Destitute Areas in Hunan Province, China. Chinese Geographical Science. 2021; 31(6): 1029-1044. Doi:10.1007/S11769-021-1242-4
3. Robert, T. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 2018; 58(1): 267-288. Doi:10.1111/j.2517-6161.1996.tb02080.x
4. Bennin et al. MAHAKIL: Diversity Based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction. IEEE Transactions on Software Engineering. 2018; 44(6): 534-550. Doi: 10.1109/tse.2017.2731766
5. Cheng et al. Prediction of rock mass class ahead of TBM excavation face by ML and DL

algorithms with Bayesian TPE optimization and SHAP feature analysis. Acta Geotechnica. 2023; 18(7): 3825-3848. Doi:10.1007/S11440-022-01779-Z

6. Liu et al. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. Computers & Security. 2021; 106: 102289. Doi:10.1016/j.cose.2021.102289
7. Chakravarti, N. Isotonic Median Regression: A Linear Programming Approach. Mathematics of Operations Research. 1989; 14(2): 303–308. Doi:10.1287/moor.14.2.303
8. Mohd et al. Abnormality Detection and Failure Prediction Using Explainable Bayesian Deep Learning: Methodology and Case Study with Industrial Data. Mathematics. 2022; 10(4): 554-554. Doi: 10.3390/MATH10040554
9. Huan et al. Exploring complex water stress-gross primary production relationships: impact of climatic drivers, main effects and interactive effects. Global change biology. 2022; 28(13): 4110-4123. Doi:10.1111/GCB.16201
10. Dong et al. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. Journal of Environmental Management. 2022; 301: 113941-113941. Doi:10.1016/J.JENVMAN.2021.113941