Heliyon 9 (2023) e19525

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

What influence farmers' relative poverty in China: A global analysis based on statistical and interpretable machine learning methods

Wei Huang, Yinke Liu *, Peiqi Hu, Shiyu Ding, Shuhui Gao, Ming Zhang

School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China

ARTICLE INFO

Keywords:
Factor identification
Machine learning
Random forest
Relative poverty
XGBoost

ABSTRACT

Poverty eradication has always been a major challenge to global development and governance, which received widespread attention from each country. With the completion poverty alleviation task in 2020, relative poverty governance becomes an important issue to be solved in China urgently. Because of a large population, poor infrastructures, insufficient resources, and long-term uneven development raising the living standard of farmers in rural areas is critical to China's success in realizing moderate prosperity. Therefore, identifying the poor farmers, exploring the influence factors to relative poverty, and clarifying its effect mechanism in rural areas are significant for the subsequent poverty governance. Most of the previous studies adopted the method of apriori assuming the factor system and verifying the hypothesis. We innovatively constructed a relative poverty index system consistent with China's actual conditions, selecting all the possible variables that could affect relative poverty based on the existing literature, including individual characteristics, psychological endowment, and geographical environment, and rebuilt an experimental database. Then, through data processing and data analysis, the main factors influencing the relative poverty of farmers were systematically sorted out based on the machine learning method. Finally, 25 chosen influencing factors were discussed in detail. Research findings show that: 1) Machine learning algorithm is proved it could be well applied in relative poverty fields, especially XGBoost, which achieves 81.9% accuracy and the score of ROC_AUC reaches 0.819. 2) This study sheds light on many new research directions in applying machine learning for relative poverty research, besides, the paper offers an integral framework and beneficial reference for target identification using machine learning algorithms. 3) In addition, by utilizing the interpretable tools, the "black-box" of ML become transparent through PDP and SHAP explanation, it also reveals that machine learning models can readily handle the non-linear association relationship.

1. Introduction

Poverty eradication is a major challenge needed global development and governance focus. There are still 648 million people worldwide living in extreme poverty condition in 2022 [1], without adequate food, safe drinking water, and stable house to live. Even worse, research shows that poverty has been linked to several negative effects including increasing in mortality and illness [2,3],

* Corresponding author.
E-mail addresses: huangwei@ncwu.edu.cn (W. Huang), lyk02241559@163.com (Y. Liu).

https://doi.org/10.1016/j.heliyon.2023.e19525
Received 26 May 2023; Received in revised form 24 August 2023; Accepted 25 August 2023
Available online 30 August 2023

decreasing the sense of well-being [4], and happening of criminal incidents [3]. Eradicating poverty and improving people's livelihoods has become an urgent task deserving of various countries' attention. China is the world's most populous developing country and once had the largest rural poverty-stricken population [5,6], which hinders the economic process and social development seriously. Facing the pattern, the Chinese government has implemented a series of poverty reduction measures to solve poverty including relief-type poverty relief, structural reform promoted poverty relief, development-oriented poverty-relief drive, tackling key problems in poverty relief, consolidation-oriented comprehensive poverty alleviation and targeted poverty alleviation [7], which play an overwhelming role in poverty reductions. Especially, with the completion of China's poverty alleviation task in 2020, absolute poverty has been eliminated, which marks a new historical stage for China's poverty reduction. However, the elimination of absolute poverty does not mean the disappearance of the poverty phenomenon in China, whether in developed or developing countries, relative poverty will persist for a long time [8,9]. According to the statistics, China has a relatively poor population of million farmers in rural China [10]. Besides, a total of 509.92 million farmers are living in rural areas[1] accounting for 36.1% of China total population, who feed on a fifth of the world's population with less than one-tenth of the world's arable land [6]. Due to harsh natural conditions, poor infrastructure and public service, uneven healthcare and education resource contribute to a large number of farmers living under the relative poverty condition, limiting the realization of a moderately prosperous society in all respects. How to alleviate the relative poverty of farmers in China's rural areas has become the top priority of the country, organization, and region.

Together with the wave of the fifth information revolution, emerging information technology provides a novelty way to solve the problems faced by the social sciences fields. As the product of the new era of information technology, demonstrating excellent accuracy and having the ability to handle huge amounts of data, as well as being competent to deal with non-linearity and capture the flexible relationship between independent variables and dependent variables, ML algorithms have deserved enough attention [11]. Additionally, the availability of interpretable tools such as partial dependent plots (PDP) and Shapley additive explanations (SHAP) also help scholars evaluate and explain the effects of variables efficiently. Inspired by the strength, ML is widely welcomed in the field of enterprise innovation performance [12–14]; supply chain demand forecasting [15–17], water resources management [18,19], and risk assessment of coal mining [20,21]. However, ML algorithms are merely mentioned in relative poverty fields, much less to apply the method. So, the overarching goal of this study is introducing ML into relatively poverty areas aiming at exploring whether the method is suitable for the fields; Beyond that, we also attempt to classify the relatively poor in rural China using ML models and sort out the importance of relative poverty. What's more, further analysis is that with the aid of PDP and SHAP, the formation mechanism of relative poverty as well as relative poverty reduction path are well explored, the impact of the indicators also gets presented adequately. The above analysis could also provide good reference for other countries encountering the relative poverty to alleviate relative poverty and achieve common prosperity as soon as possible.

2. Literature review

2.1. *The influencing factors on relative poverty*

In the post-poverty era, the emphasis of China's poverty eradication efforts has shifted from eliminating absolute poverty to improving relative poverty [22,23]. Most of the previous literature focusing on absolute poverty relied on a single indicator of income (e.g., the absolute poverty line set by the state each year) to determine whether a farm household is in poverty [24,25]. In contrast, after entering the period of relative poverty governance, low income may only be a symptom of relative poverty; the root causes of poverty for farmers are the lack of viability, subjective aspirations, and the culture of poverty in poor areas, etc., and using income alone to measure relative poverty is too one-sided [26–30]. Compared to absolute poverty, relative poverty criteria are more diverse and multidimensional, and poverty types will be more multidimensional and complex [31–33], requiring the measurement of multiple influences such as income, health, ability, and psychology [30,34,35]. Angulo focused on five dimensions to analyze poverty, including family education, conditions of children and youth, employment, health and access to public facilities, and housing conditions [36]. Through regression analysis, Li and Shen [37] and Qin and Rong [38] found that public spending on education and direct transfers had a significant impact on reducing relative rural poverty. China is a family-based society, and in addition to individual characteristics, family elements also affect poverty [39], such as per capita household income, proportion of agricultural income, and household consumption, and the external environment, including the individual's family environment, can also have an impact on poverty, such as regional financial support policies [40], and household capital perspectives such as house type [41] and the value of durable goods [42]. In addition to the above dimensions, scholars have also explored poverty status from the perspective of psychological conditions, such as confidence and stress [43]; geographic perspective, such as rain [44]and sunlight [45].

2.2. *The research fields and methods of relative poverty*

In accordance with the content of relative poverty studies and relying on the relative poverty influencing factors, scholars have also used diverse research approaches to study relative poverty-related areas, the attention mainly focuses on relationship evaluation, influencing mechanisms, and relative poverty reduction strategies. In the perspective of relationship evaluation, scholars adopt research methods such as stratified analyses and case-control studies; Hirokazy conducted a 10-year follow-up study and assessed

[1] 2021 National Bureau of Statistics Statistical Yearbook.

mortality according to relative poverty and relationships [46]; Esmael conducted the study which demonstrated a strong association between trachomatous trichiasis and relative poverty [47]. In terms of influencing mechanism, system dynamics, and econometric model has been widely used. In Xin's paper, the system dynamics model is used to analyze interactive mechanisms of eco-environment, geo-disasters, and immigrant poverty in order to understand the impacts of poverty strategies [48]; Based on micro-level data, Stephanie analyzed separately the rate of pretax/transfer poverty and the reduction in poverty using seven types of econometric models [49]. In the field of relative poverty reduction strategies, panel-VAR modeling, Xin points both growth and income distribution are found to aggravate relative poverty by reviewing official reduction poverty strategy [50]; Zulkarnain thinks the government ought to empower the community and encourage mutual support and attentiveness among its members to combat the relative poverty [51]. In particular, the relative poverty measure method also has received significant attention in recent years. A-F, FGT, and MPI multidimensional measure methods have been widely used to evaluate the relative poverty condition by scholars, which has made great achievements [34,52,53]. However, due to the limitation of data scale and research method, these methods couldn't dig out the deep function mechanism of relative poverty, resulting in there aren't fully effective strategies to alleviate relative poverty.

2.3. The advantage of the ML algorithm

The rapid growth of the information technology field has necessitated the need for more novel and accurate methods of analyzing relative poverty, as well as investigating the principle behind the appearance. The ML algorithm with excellent performance suit the needs of scholars naturally. Although traditional models have simple operation and fast computing speed, compared to traditional identification models, ML models enjoy several obvious benefits needed to be classified which have been shown as followings. The first deserved focus point is the aspect of data inspection, Susan Athey points out that traditional statistics and econometrics have put much emphasis on the endogeneity of variables and the configuration of data such as panel data [54], whereas ML has paid less attention on them, which means ML have the ability to work with the data without specific requirement [55]; Besides, ML could handle high-dimensional problems and data-sets as well. For example, Tom Howley investigated the use of the principal component analysis (PCA) to reduce high-dimension and improve the accuracy of predictive performance [56]; Dongdong Sun proposes a multimodal deep neural network by integrating multi-dimension data, the models' results present the proposed methods achieves a better performance [57]. Meanwhile, the ability of ML to achieve high accuracy is another advantage that should get more attention. Yunxin Xie evaluates five typical machine learning methods using data from gas filed, the matrix shows that GTB and RF reach 81.1% and 80.9% accuracy separately [58]; In encrypted traffic areas, Yohei Okada proposed EFM to explore whatever the traffic is encrypted or not, the results indicate EFM using SVM provides overall accuracy 97.2% [59]. More importantly, the appearance of interpretable tools such as partial dependent plots (PDP) and Shapley additive explanations(SHAP) provides exploring tools to analyze the "Black-box" of ML. Traditional regression or classification is possible to obtain the significant factors only if there exists a linear relationship between input and out variables [60], however, ML has the capacity to get the influencing factors when handling complex relationships whether linear or complex relationships [61]. Regarding for above strength. So, we attempt to bring the ML into relative poverty. We also list the whole detailed information of ML techniques in relative poverty identification application which has been shown as Table 1.

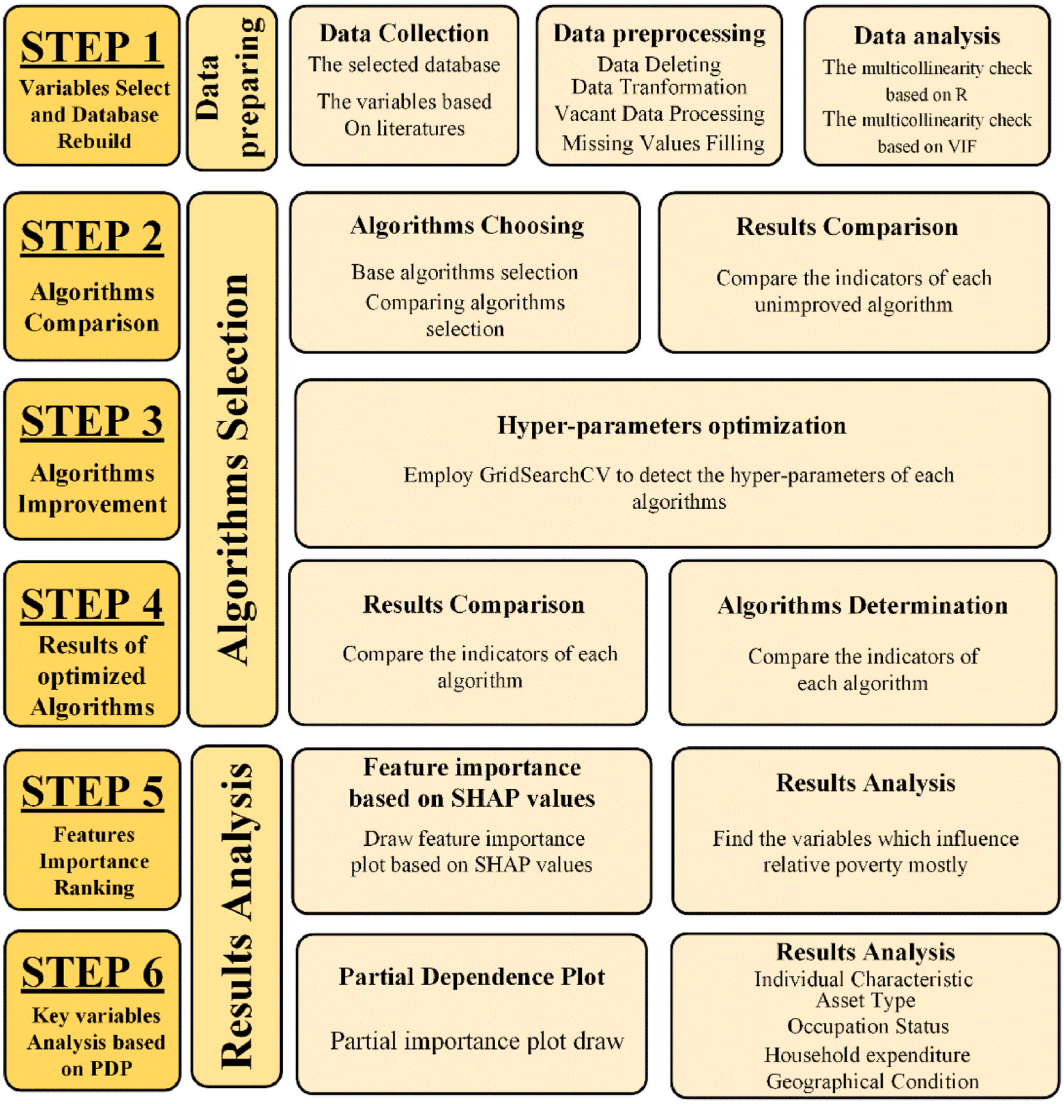
Many researches have made great progress in the field of relative poverty research, which extends today's relative poverty research greatly and brings constructive references to this paper. By summarizing the existing literature, we found that 1): The dimensions of relative poverty influencing indicators have shifted from one-dimensional to multidimensional, besides, the content of indicators has gradually shifted from income to health, education, security, psychological and geographical factors. However, relative poverty is the result of both subjective and objective conditions. In the process of constructing the indicator system, although the above aspects have been taken into consideration, previous authors have failed to measure relative poverty using both subjective and objective dimensions; 2): At present, scholars mainly focus on the relationship identification, influence mechanism, poverty reduction strategies and measurement methods of relative poverty. However, accurate identification of the relatively poor is the premise of carrying out the above research, but there are few scholars involved in this field at the present stage, which means there is a large research gap needed to be filled. 3): Traditional research methods such as stratified analyses, econometric models, case-control study panel-VAR modeling, and satellite image et al. have been widely used in the field of relative poverty and made great progress. Due to the limitations of

Table 1
The summary of ML algorithms in relative poverty identification application.

Identification Requirement	The theoretical ability to meet the requirement
Ability to deal with multi-source and high-dimension data	ML have capable of handling high dimensions (>1000) [61] and multi-source data(e.g. qualitative and quantitative data) [62]
Ability to achieve high accuracy and precision based on the selected datasets	ML is widely applied in the field of economics [63], engineering [64], and species identification [65]. All of them get a satisfactory experiment accuracy score.
Ability to emphasis less attention on the data requirement so that simplify the experiment process	ML requires preparing data through feature engineering including dimensionality reduction and data denoising [66], exclude on the endogeneity of variables, and the configuration of data
Ability to extract key variables information and capture influencing mechanisms by learning from results	PDP, SHAP, and ICE have become useful in interpreting the relationship [67].
Ability to handle complex relationships between input and output variables rather than a single relationship	ML algorithms (e.g., ANN, RF, SVM) are capable of learning from results about the complex relationship [68,69]
Ability to adapt to the changing environment and could further be applied to other fields using similar research form	ML is feasible to adapt to changing environments automatically is a major strength of ML [70,71]

sample size and research methods, mechanisms of relative poverty couldn't be analyzed deeply as well as the causes of relative poverty couldn't investigate fully, therefore, new approaches and tools are needed to fill the research gap.

Based on the above analysis, we mainly put forward innovative points from the research content perspective, index system construction, and research methods 1) Different from the prior studies on relative poverty which emphasize capturing the macro-aspect in policy formulation, industrial development, risk challenges, and evolutionary features, our manuscript focuses on the individuals that constitute relative poverty population to explore the mechanism from micro-aspect, attempting to seek a practical pathway that is suitable for China's special condition so that to alleviate the phenomena of relative poverty in China. 2) In the aspect of the indicator system construction, the already existing measuring method such as A-F, FAG, and MPI put a great emphasis on the perspective of livelihood capital, social capital, and family condition [34,51,52], ignoring that the relative poverty is the combination of multifactorial indictors. Based on that, on the one hand, this manuscript covers the more comprehensive dimensions reflecting relative poverty than former literature by combining subjective and objective aspects from individual characteristics, psychological endowment, and geographical environment, which provides a stable foundation to identify relative poverty. On the other hand, this manuscript selects the indicators constructing the above three dimensions as much as possible to provide a wide horizon to reflect the real condition of rural China. For example, the individual characteristic dimension selects education, land use, and so on. The psychological endowment dimension chooses confidence, pressure, etc. as the final indicators. 3) The traditional statistical identification methods such as the linear regression model owes the limitation of couldn't handing the multi-source and high-dimension as well as having no ability to achieve high accuracy and precision based on the selected datasets. By bringing the machine learning algorithms which overcome the above defects to the field of relative poverty, this paper achieves the accurate identification of relative poverty, solving the problem of



STEP 1
Variables Select and Database Rebuild

Data preparing

Data Collection
The selected database
The variables based On literatures

Data preprocessing
Data Deleting
Data Tranformation
Vacant Data Processing
Missing Values Filling

Data analysis
The multicollinearity check based on R
The multicollinearity check based on VIF

STEP 2
Algorithms Comparison

Algorithms Selection

Algorithms Choosing
Base algorithms selection
Comparing algorithms selection

Results Comparison
Compare the indicators of each unimproved algorithm

STEP 3
Algorithms Improvement

Hyper-parameters optimization
Employ GridSearchCV to detect the hyper-parameters of each algorithms

STEP 4
Results of optimized Algorithms

Results Comparison
Compare the indicators of each algorithm

Algorithms Determination
Compare the indicators of each algorithm

STEP 5
Features Importance Ranking

Results Analysis

Feature importance based on SHAP values
Draw feature importance plot based on SHAP values

Results Analysis
Find the variables which influence relative poverty mostly

STEP 6
Key variables Analysis based on PDP

Partial Dependence Plot
Partial importance plot draw

Results Analysis
Individual Characteristic
Asset Type
Occupation Status
Household expenditure
Geographical Condition

Fig. 1. Experimental research framework.

"who should we study in the field of relative poverty". At the same time, inspired by the interpretable tools named feature importance and SHAP, the inherent mechanism and functional relationship of relative poverty are revealed wholly, which answers the question "how should we alleviate relative poverty". The application of the machine learning model method in this manuscript provides an example for realizing accurate identification of relatively poor people, and its research conclusions offer strong theoretical support and practical significance for alleviating relative poverty and improving people's well-being. Following the research steps, this manuscript also provides the realization process of mechanism exploration for researchers in other fields.

3. Material and methodology

3.1. Experiment framework

The research experimental process is divided into 6 steps, 1) *Variables select and database rebuild*. 2) *Algorithms comparison*. 3) *Algorithms improvement*. 4) *Results of optimized Algorithms*. 5) *Features Importance Ranking*. 6) *Key variables Analysis based on PDP*. The detailed experimental framework is shown in Fig. 1.

Variables select and database rebuild. This section consists of three main steps. i) Data Collection. Data collection mainly includes which database we should choose and which variables we ought to select. These databases and variables must stand with the Chinese actual situation. To better reflect the conditional of relative poverty, this paper selects three variables' dimensions including individual characteristics, psychological endowments, and geographic environment variables. ii) Data preprocessing. Data preprocessing is the process of filtering, detecting, and correcting inappropriate records from the raw data to make the data fit the models and met the requirement of the actual situation. A four-stage procedure is used to implement the preprocessing operation of the data, which mainly includes: missing values deletion; categorical values transformation, and outlier records removing. iii) Data analysis. We discard the variables from the database as they suffer from strong multicollinearity.

Algorithms Comparison. This section consists of two steps. i) Algorithms choosing. We choose logistic regression as the traditional algorithm. Random forest, decision tree, BernoulliNB, KNeighbors, and GaussianNB are chosen as traditional machine learning algorithms. And XGBoost, Lightgbm, and Catboost are selected as ensemble learning. At the same time, the results of unimprovement algorithms are comprised as well.

Algorithms I*mprovement*. The method of grid search is employed to improve the performance of the algorithms.

Results of Optimized Algorithms. i) Results Comparison. Compare the indicators of each algorithm. ii) Algorithms Determination. Based on the results, select the most suitable algorithm as the base algorithm.

Feature importance ranking. i) Features importance based on SHAP values. Draw the feature's importance plot based on SHAP values. ii) Results analysis. Find the variables which influence relative poverty mostly.

Key Variables Analysis based on SHAP. This section consists of two main steps. i) Partial Dependence Plot. The partial dependence function of the model describes the expected effect of a feature after marginalizing the effects of all other features. ii) Results Analysis. Based on the analysis results, the dimension of the selected variables is analyzed.

3.2. Methods

3.2.1. Algorithms introduction

(1) Random forest (RF) is a non-parametric machine learning method for classification and regression analysis(Breiman 2001) [72]. It was first proposed by Breiman in 2001 and has significant advantages in processing high-dimensional data. There is a lot of research indicating that the random forest algorithm reduced prediction errors and outperformed other algorithms when dealing with high-dimensional data.

(2) Decision tree is a machine-learning algorithm that makes decisions based on a tree-like structure [73]. Common decision tree algorithms include C4.5, ID3, SLIQ, CRAT, etc. [74–76]. The decision tree algorithm is suitable for discrete data and is easier to interpret than methods such as neural networks. The decision tree algorithm has the advantages of low computational complexity, convenience, and efficiency in solving classification problems [73]. However, the decision tree algorithm is difficult in dealing with missing data, and in addition, it may over-divide the sample space thus leading to over-fitting problems [76].

(3) BernoulliNB is one of the scikit-learn plain Bayesian class libraries for processing binary discrete-valued or very sparse multivariate discrete-valued sample features and is commonly used to process text classification data. The BernoulliNB algorithm is mainly applied to Boolean features or binary values. BernoulliNB assumes a binary Bernoulli distribution for the prior probability of the characteristics, the mathematical expression is seen as equation (1).

$$P\left(X_j = x_{jl} | Y = C_k\right) = P(j|Y = C_k)_{xjl} + (1 - P(j|Y = C_k))\left(1 - x_{jl}\right) \tag{1}$$

$P(X_j = x_{jl} | Y = C_k)$ is the l-th conditional probability of the j-th dimensional feature of the k-th category.There are only two values for l, indicating two parameters. x_{jl} can only take the value 0 or 1.

(4) The K-neighbor classifier classifies data points according to the class of their nearest neighbors in the feature space. Due to the large amount of computation involved, the K-neighbor classifier requires a large amount of computational memory to be

allocated. Because the modeling process needs time, this also increases the difficulty of hyper-parameter adjustment. Nevertheless, it has achieved excellent accuracy in many applications and is easy to understand and illustrate [77].

(5) In the field of poverty, there have been many studies applying regression algorithms to poverty problems. Zhao and Herencsar combined mobile big data and logistic regression algorithms to identify poor households [78]. Peng et al. used quantile regression models to examine the differential effects of poverty determinants in the poverty spectrum [79]. Logistic regression is used for binary classification problems, such as positive or negative, yes or no, etc., which is a powerful tool for traditional statistical analysis [80]. Logistic Regression has been performed with a number of predictive analysis techniques and depends on the total likelihood [81,82]. This model uses a very complex cost identification function, which can be interpreted as a "Sigmoid function". It is also referred to as a "Logistic function" rather than a linear function. The mathematical expression is seen as equation (2).

$$S(z) = P\left(\frac{1}{1+e^{-z}}\right) \tag{2}$$

S(z) = output between 0 and 1 (probability estimate); z = input to the function (e.g. y = mx + b); e = base of natural log.

(6) GaussianNB is one of the scikit-learn plain Bayesian class libraries, which is a priori a Gaussian distribution of plain Bayes. GaussianNB is suitable if the distribution of the sample features is mostly continuous values.GaussianNB assumes that the prior probabilities of the features are normally distributed, and the mathematical expression is seen as equation (3).

$$P\left(X_j = x_j | Y = C_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{\left(x_j-\mu_k\right)^2}{2\sigma_k^2}\right) \tag{3}$$

In this function, C_k is the k-th class category of Y; μ_k and σ_k^2 are the values to be estimated from the training set; The parameters μ_k and σ_k^2 are estimated using maximum likelihood [83].

(7) XGBoost is a tree-integrated model that uses the cumulative sum of the predicted values of the samples in each tree as the predicted values of the samples in the XGBoost system [84]. XGBoost is designed to properly utilize resources and overcome the limitations of previous gradient boosting by being a highly scalable, flexible, and versatile tool. XGBoost uses a new regularization technique to control overfitting, which is the main difference between XGBoost and other gradient boosts [85]. The regularization technique is accomplished by adding a new term to the loss function. The mathematical expression is seen as equation (4).

$$L(f) = \sum_{i=1}^{n} L(\widehat{y}_i, y_i) + \sum_{m-1}^{M} \Omega(\delta_m) \tag{4}$$

with $\Omega(\delta) = \alpha|\delta| + 0.5\beta\|w\|^2$; where $|\delta|$ is the number of branches, w is the value of each leaf and Ω is the regularization function. XGBoost uses a new gain function, the mathematical expression is seen as equations (5)–(7).

$$P\left(X_j = x_j | Y = C_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{\left(x_j-\mu_k\right)^2}{2\sigma_k^2}\right) \tag{5}$$

$$G_j = \sum_{i\in I_j} g_i H_j = \sum_{i\in I_j} h_i \tag{6}$$

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\beta} + \frac{G_R^2}{H_R+\beta} - \frac{\left(G_R+G_L\right)^2}{H_R+H_L+\beta}\right] - \alpha \tag{7}$$

where $g_i = \partial_{\hat{y}_i} L(\widehat{y}_i, y_i)$ and $h_i = \partial_{\hat{y}_i}^2 L(\widehat{y}_i, y_i)$. G is the score of the right child, H is the score of the left child and *Gain* is the score in the case no new child [86].

(8) CatBoost focuses on categorical columns using the swap technique, one_hot_max_size (OHMS), and objective-based statistics. CatBoost solves the problem of exponential growth of feature combinations by using greedy methods at each new split of the current tree. For features with more than OHMS (one input parameter) per category, CatBoost uses the following steps.

1. Dividing the records into subsets randomly,
2. Converting the labels to integer numbers, and
3. Transforming the categorical features to numerical, the mathematical expression is seen as equation (8).

$$aveTarget = \frac{countInClass + prior}{totalCount + 1} \tag{8}$$

where count In Class is the number of ones in the target for a given categorical feature, total Countis the number of previous objects,

Table 2
Confusion matrix description.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Pr\ ecision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$f1score = 2 * \frac{P * R}{P + R} \tag{12}$$

		Predicted value	
———	–	1	0
True Value	1	TP	FN
	0	FP	TN

and prior is specified by the starting parameters [87–89].

(9) LightGBM supports efficient parallel training of the framework for implementing the GBDT algorithm. LightGBM is proposed to solve the time-consuming problem in large high-dimensional data sample environments [90]. The main difference between LightGBM and XGBoost is that the decision tree in LightGBM grows by leaves, rather than checking all previous leaves for each new leaf. LightGBM has better accuracy, faster training speed, the ability to process data at scale, and support for GPU learning.

3.2.2. *Evaluation methods*

In this study, the accuracy of the model is calculated using a confusion matrix, and the accuracy, precision, recall, F1 score, and ROC-AUC score after 10-fold cross-validation averaging are calculated to evaluate the accuracy of the model. The confusion matrix is one of the most important references to evaluate the accuracy of the classification model, which is represented using a matrix with N rows and N columns, where each column of the matrix represents the predicted values while each row represents the actual values. The schematic of the confusion matrix is shown in Table 2. Each index is calculated as shown in the following equations (9)–(12).

3.2.3. *Interpretable methods*

3.2.3.1. Part Independence plot. The part dependence plots are an interpretable method to analyze the effect of a particular variable on other variables when controls of other variables are unchanged. The core advantage of PDP is that it could show us the relationship between the dependent variable and target variables. The mathematical expression of PDP function is shown as equation (13).

$$\widehat{f}_{xs}(x_s) = E_{xC}[\widehat{f}[(x_s, x_C)]]dP(x_C) = \int \widehat{f}(x_s, x_C)dP(x_C) \tag{13}$$

Among them, x_s presents the characteristic variables, x_C are the other variables expect x_s, $\widehat{f}_{xs}(x_s)$ is the correspondence label estimate under x_s values. And $(x_c) = \int p(x_{all})dx_s$, $P(x_c) = \int p(x_{all})dx_s$ [91]. Through PDP, the whole interpretable explanation could be analyzed.

3.2.3.2. SHAP values. SHAP is a unified tool to measure feature importance. SHAP uses the additive feature attribution method to interpret models. The mathematical expression of SHAP is shown as equation (14).

$$f(x) = g(x') = \varphi_0 + \sum_{i=1}^{M} \varphi_i x_i' \tag{14}$$

Assuming the input variables of the models is x = $(x_1, x_2, x_3 \ldots x_n)$. n present the total amount of variables. The explanation models $g(x')$ with simplified input x' for an original model $f(x)$ is expressed as equation (14).

M presents the number of input variables. φ_0 presents the constant value when the whole input is missing. Input x' and x connect through a mapping function, which is shown in Fig. 2.

4. Variables selection and database Rebuilding

4.1. *Data Collection*

For the analysis throughout the paper, we used 3 waves of databases from China. Aiming at providing current and future researchers with whole and objective data on Chinese society, CFPS (China Family Panel Studies) database is a national, comprehensive,

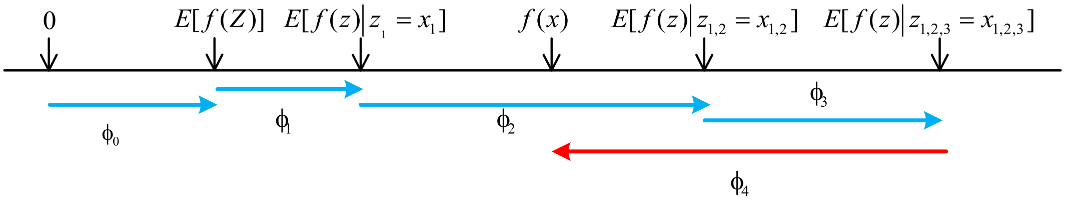


Fig. 2. The achieving function of SHAP values [92].

longitudinal social survey launched by Peking University in 2010 to collect individual-, family-, and community-level data [93]. (The official website is at isss.pku.edu.cn/cfps/). The samples of CFPS cover the sample 25 provinces in China, which provides a general-purpose, national representative as well as longitudinal horizon of China. At the same time, the 2019 National Bureau of Statistics Statistical Yearbook (the official website is China Statistical Yearbook 2018 (stats.gov.cn)) and the Bulletin of the first national census for water (the official website is https://kns.cnki.net/) are adopted.

In terms of independent variables, the current relative poverty definition could be divided into two categories: the "proportional method" and the "index method". The "index method" is mainly applied to developed countries [94]. However, in the context of China's economic development, it is too early to use the "index method" to measure relative poverty, so this paper set 50% of the national per capita disposable income in 2018 as the standard line to measure relative poverty with reference to previous research results [95]. In this study, the relative poverty status can be expressed as a 0–1 dummy variable. When the income of the individual is less than 50% of the national disposable income (28,228 yuan, in 2018), the person is considered as in the relative poverty status and the value of a dependent variable is 1. Otherwise, the individual is considered to be in the normal state and the value is 0.

In addition to the dependent variables, we also find all the possible variables affecting relative poverty indicators from the previous literature in multiple dimensions. Basic individual characteristics are the most direct variables affecting individual poverty [96,97]. At the individual level, we can explore whether there is a relationship between an individual's own conditions and a relative poverty situation. Similarly, individual psychological endowments can also have an impact on relative poverty [98]. Individual psychological endowments are manifested intrinsically as personal psychological qualities and extrinsically as behavioral tendencies that one is willing to make to escape poverty. Due to the differences in the natural environment and socioeconomic conditions, the contiguous poor individuals show significant spatial differences in poverty. Therefore, it is necessary to accurately reveal the spatial details of poverty. Lots of studies have shown that an individual's geographic location can significantly affect poverty status [98]. Based on these aspects, this paper divides the indicators into three major dimensions, individual characteristics and psychological endowments, and physical geographic environment. Details are as follows.

a. Individual characteristics variables reflect the characteristics of the individual's household conditions, work status, and consumption expenditures, presenting the individual's own household situation.
b. Individual psychological endowment variables include indicators of confidence, stress resistance, personal well-being, and satisfaction with current living conditions. Our aim is to investigate whether these indicators can influence the relative poverty of individuals directly.
c. Geographic environment variables mainly contain characteristics such as average temperature, average precipitation, annual sunshine hours, and average altitude in which it is located, which reflects the influence of geographical conditions on relative poverty.

Following the steps, the individual characteristics and psychological endowment variables were obtained from the 2018 CFPS database. A total of 37 variables were selected from the CFPS database. Individual psychological endowments were also obtained from the 2018 CFPS, which totally include 4 variables. Among the geographic environment variables, such as temperature, rainfall, etc., from the 2019 China statistical yearbook, a total of 9 variables were selected, due to the number of reservoirs in each province in the statistical yearbook is missing, and the final number of reservoirs in each province was obtained from the bulletin of the first national census for water. In the end, a total of 50 variables and 17,722 data were obtained, involving both subjective and objective dimensions, which have been listed in Appendi×1.

4.2. Data preprocessing

We adopt a four-stage approach to prepare the data to improve predictive accuracy and computational efficiency. The first stage is database processing. The CFPS database covers individuals at all stages of age. However, the junior and senior groups usually have inherent deficits in income, which influence the experimental precision greatly. Therefore, groups younger than 14 and older than 65 years were removed from the dataset. The second stage is variable transformation. Selected data were transformed according to the data type. Numerical variables are input directly and categorical variables are converted into dummy variables to satisfy the requirement of the experiment. The third step is vacant data processing. After transforming, there are still existing missing values and empty values. Referring to the previous data processing process [99], we delete the records that have a single vacant record, own the field named “not applicable” and exist values more than 45% missing. As for outlier values that present negative values such as total_house_property, we delete them form the dataset directly. The fourth step is missing values filling. For categorical variables with missing values such as gender, marriage, health condition, Internet use, confidence, pressure, etc., the mode value of the record is used to fill [100]. For numerical variables with missing values, such as durable goods, money, etc., the overall average value is used for padding [101]. The data preprocessing framework is shown in Fig. 3.

4.3. *Data analysis*

Multicollinearity is one of the most vexing and intractable problems in all of the regression analysis [102], which drives us to the wrong conclusion of lower importance of one variable over the other [103]as well as decreasing the computation power of the models [104]. In this paper, different from previous studies that ignore the multicollinearity among selected variables [105], we delete those variables with a VIF(variance inflation factor) of more than 10 [106], and the highly correlated variables(those with $r > 0.7$) are also removed. Finally, a total of 25 variables remain as shown in Table 3. At the same time, the heat map of selected variables is shown in Fig. 4.

5. Algorithms selection

5.1. *Algorithms Comparison*

Firstly, to verify whether ML algorithms can be well applied in the field of relative poverty, refer to previous research [85, 107–109], as well as better distinguish the difference between machine learning and traditional models, we mainly choose the “Logistic Regression” algorithms as traditional models, other eight types of algorithms are considered as the comparison models. Among them, five traditional machine learning algorithms including Random Forests, Decision Tree, BernoulliNB, KNeighbors, and GaussianNB, and three ensemble models such as XGBoost, Catboost, and Lightgbm are selected to carry out the experiment. The whole experiment environment is based on Python version 3.7.4 and Sklearn version 0.24.2, Before employing the models, the data are divided into 7:3, which means 70% of the original data are classified as training samples, and the remaining 30% are used as testing set samples to evaluate algorithms. Through the experiments, the output of the models is shown in Table 4. We can see that the RF algorithm has the best performance, which reaches 0.81 scores in accuracy. Besides, other scores of RF such as ROC_AUC, precision, recall, and F1 also make the wonderful manifestation. We also notice “Logistic Regression” reach an accuracy of 0.55. However, compared with the machine learning models, it obviously gets the wore score. It main indicates two aspects. One the one hand, the results reveal that machine-leaning algorithms surely suit the field of relative poverty, and have a better performance than traditional models; On the other hand, the models still have room for improvement, which needed to dig out the potentiality of algorithms. So, following the experiments' logistics, in the next section, we try to improve the accuracy by adjusting the parameters of the algorithms.

5.2. *Algorithms improvement*

The performance of classification algorithms on a particular dataset is highly dependent on the algorithm optimization process. The methods to improve the accuracy of classification algorithms commonly used feature selection [110] and optimizing the parameters of machine learning algorithms [111]. Feature selection methods are based on the relevance or correlation between features and the class labels as its fundamental principle [110]. For a given dataset $D = (X,C)$, the features are list as $X = \{X_1,X_2,\ldots,X_n\}$, the label is seen as C. The purpose of feature selection is to find the subset $\widehat{D}$ ($\widehat{D} \subseteq D$), maximizing the accuracy of models. Another improving accuracy of the models' method is to find the most suitable parameters, which also call as “hyper-parameters”. Approaches like “grid search” and “genetic algorithms” are frequently employed to find suitable parameter values [112]. Grid search is seen as an exploration method to test all the combinations of hyper-parameters. Genetic algorithms regard biological evolution theory and national selection theory as the theory foundation to finding the short path. There also exist related research shows the difference between them main embody that “grid search” cost more time than “genetic algorithms”, however, compared with the latter, the former could gain more accuracy [112]. In this paper, feature selection has been adopted in data preprocessing. So, in this section, we mainly utilize “grid search” to optimize the hyper-parameters so that find optimal results. Specific experimental steps are shown as followings: In this section, we mainly employed “GridSearchCV” models which be contained in Sklearn to detect the hyper-parameters. Although " GridSearchCV " is quite simple, the priority we need to do is confirm the range of parameters. The value ranges of each model's hyperparameters and their final value results are shown in Table 5 below.

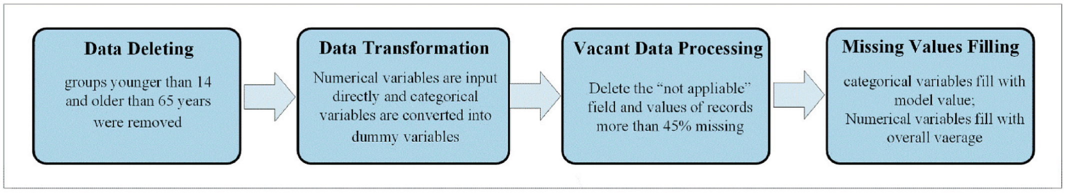

Data Deleting
groups younger than 14 and older than 65 years were removed
Data Transformation
Numerical variables are input directly and categorical variables are converted into dummy variables
Vacant Data Processing
Delete the “not appliable” field and values of records more than 45% missing
Missing Values Filling
categorical variables fill with model value;
Numerical variables fill with overall vaerage

Fig. 3. The data preprocessing framework.

Table 3
25 selected variable names and explanations.

Name	Description
information_chance	Whether respondents were able to use modern information technology to obtain more information
Social_Risk	Whether the respondents have commercial insurance
Gender	Gender of the respondents
present_house_property	The property value of the respondent's current home
Water_conservancy_facilities	The number of DAMS per person in the respondent's province
medicel_cost	The respondent's household expenditure on medicine
police_chance	Whether the respondents received any national policy funding
Production_Risk	Weekly working hours of respondents
book	The number of books in the respondent's home
Education	The number of years the respondent has been in education
major_thing	Whether there was a major incident in the household of the respondent in the previous year
social_relationship	The amount of money the respondent's household spent on favors in the previous year
Ecomicial_Risk	Weekly working hours of respondents
family_machine_value	Home mechanical value of respondents
migrant_worker	Whether respondents work outside the home
education_cost	The respondent's household expenditure on education
duriable_good	The value of durable goods in respondents' households
loan	The per capita loan amount of the respondents' households
food_cost	Respondent's household food expenditure
fuel	Respondent's living and cooking fuel
money	Total household cash and savings of respondents
total_house_property	The property value of the respondent's total house
Ecomicial_Chance	Whether respondents received funds from family and friends when they were in difficulty
hire_other_land	Whether the respondent's family rents other people's land
work_type	Whether the respondent was self-employed or employed

5.3. *Algorithms Determination*

The final selected 25 variables and the optimal hyper-parameters of each algorithm were input into the selected model, and the results were shown in Table 6. Meanwhile, according to the five dimensions of evaluation indicators, the radar diagram of each model is shown in Fig. 5 (a)-(i). The results show that, compared to the unadjusted model, the performance of the model gets improved by hyperparameter optimization. For example, RF improved the accuracy from 0.814 to 0.818. However, considering the data volume of the inputting data, the improvement effect is still significant. We also found that relevant indicators of some algorithms, such as Lightgbm and Logistic Regression, decreased after the above operation. The possible reason for the above situation is that grid search is applicable to the group of parameters with the highest score in cross-validation, but not necessarily to the overall data. At the same time, we found XGBoost has an excellent performance by adjusting the hyper-parameters, the accuracy changes from 0.806 to 0.819, which performs best in all algorithms, also indicating machine learning algorithms can achieve better accuracy than traditional statistical models.

6. Results and discussions

6.1. *Features Importance Ranking*

Features importance and visualization are important and widely used interpretable tools in machine learning [113], which help us with a better understanding of the total effect of every feature on the final results. In section 5, the XGBoost model's algorithms have proved to have the best performance in all selected algorithms. So, in this section, the XGBoost is selected as the baseline model to illustrate the key features and visualized the results by utilizing SHAP. The results have shown in Fig. 6. In Fig. 6, the left plot through the mean (|SHAP value|) presents the importance ranking of the top 20 variables, and the right plot calculated by the features impacting the model output presents the results as well. In the right plot, the y-axis indicates the input variables in order of importance from top to bottom. Every dot is painted by the value of the input variables, from low(blue) to high(red). The density as Fig. 6 presents the distribution of the point in the database [114].

Aiming to explain the results better, according to the results of the experiment, 20 top selected variables impacting relative poverty are subdivided: the dimension of individual characteristics are divided into four aspects including *asset variables*, *occupational condition variables, house expenditure variables, and individual characteristic variables*. Geographical environment variables remain constant. Of the top 20 most influencing relative poverty, seven are the asset variables, two are occupational conditional variables, three are house expenditures variables, six are individual characteristic variables and one is geographical environment variables. Based on the above classification, further discussion is conducted as follows.

6.1.1. *The asset variables*

The asset variables mainly include "money", "duriable_good", "family_machine_value", "fuel", "total_house_property", "present_house_property" and "book". Among them, "money" is the variable that influences relative poverty mostly. "The born of the

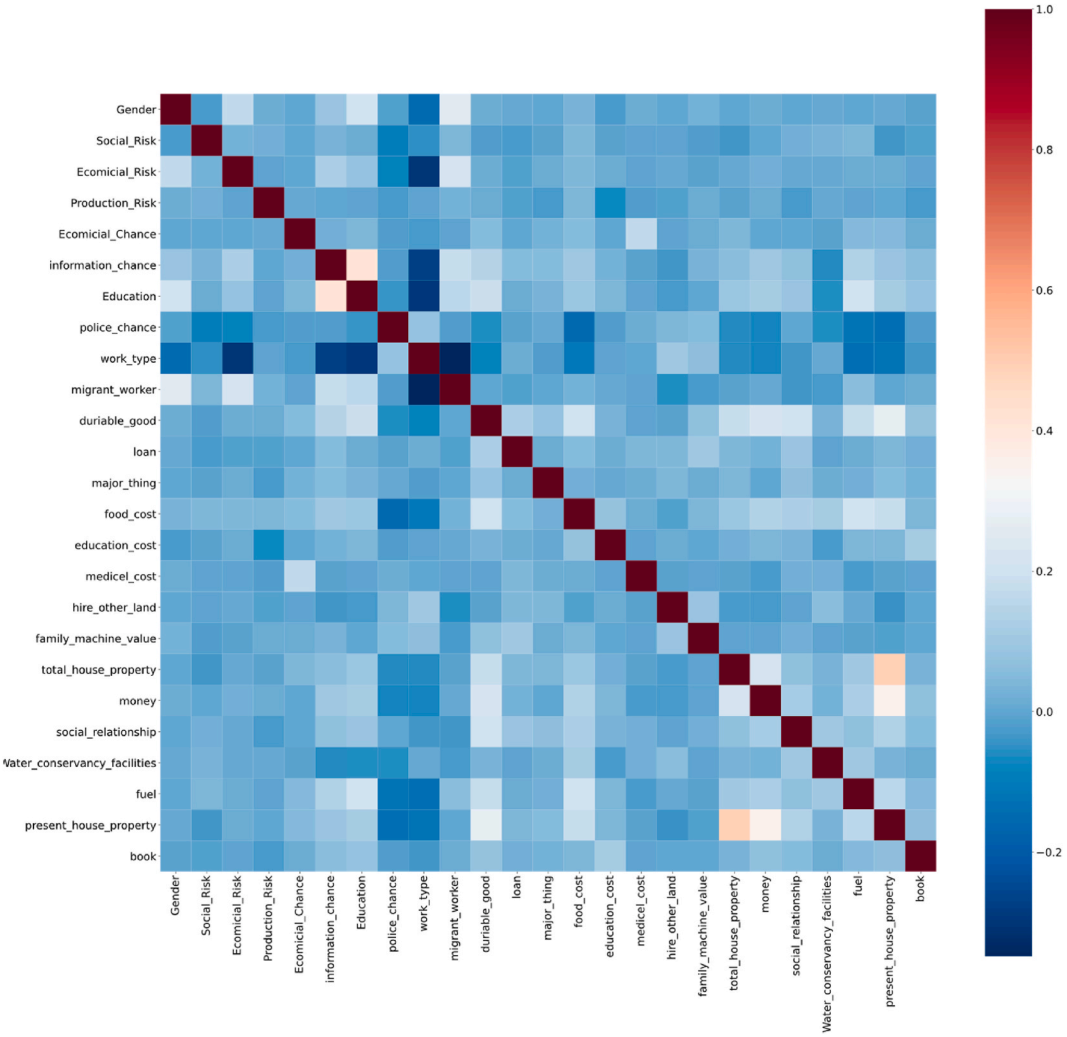


Fig. 4. The heat map of selected variables.

Table 4
The comparison between unimproved traditional algorithms and machine learning algorithms.

	Model	Accuracy	ROC_AUC	Precision	Recall	F1
Traditional Model	Logistic Regression	0.554	0.559	0.649	0.267	0.379
Traditional Machine Learning Model	Random Forest	0.814	0.813	0.819	0.815	0.816
	Decision Tree	0.767	0.767	0.771	0.771	0.771
	BernoulliNB	0.698	0.698	0.697	0.719	0.708
	KNeighbors	0.675	0.674	0.664	0.732	0.696
	GaussianNB	0.600	0.595	0.567	0.908	0.698
Ensemble Learning Models	XGBoost	0.806	0.806	0.802	0.821	0.811
	Catboost	0.763	0.762	0.759	0.788	0.772
	Lightgbm	0.803	0.803	0.802	0.813	0.808

laboring poor due to they have no assets" [115]. Money plays a key role in promoting family harmony and stability, disease prevention, career planning, and self-efficacy [116]. Therefore, improving the asset allocation of rural residents, paying attention to the asset accumulation of rural households, and implementing differentiated asset support policies for relatively poor families with a lack of assets are the solutions to alleviate the problem of relative poverty.

The value of the total property, the value of durable goods, and the value of the present property accounted for the second, third, and fourth place of the variable types of assets. The value of durable goods refers to the total value of durable goods including color

Table 5
The selection of hyper-hyperparameters.

Algorithms Types	Hyper-parameters Values Search Range	Best Parameters
Random Forest	n_estimators: linspace(start = 50,stop = 3000,num = 60)	2800
	max_features: [‘auto’, ‘sqrt’]	auto
	max_depth: linspace(start = 10,stop = 500,num = 50)	430
	min_samples_split [2,5,10]:	2
	min_samples_leaf [1,2,4,8]:	1
Decision Tree	criterion: [‘Gini’, ‘entropy’]	Gini
	min_samples_leaf: linspace(start = 1,stop = 50,num = 10)	46
	min_impurity_decrease: linespace(start = 0,stop = 0.5,num = 20)	0.0
	max_depth [1–10]:	6
BernoulliNB	Default	Default
KNeighbors	n_neighbors [1–11]:	11
	‘Metric’: [‘euclidean’,’manhattan’,’cheebyshev’,’minkowski’]	minkowski
	weights: [‘uniform’,’distance’]	distance
	p [1–6]:	4
Logistic Regression	‘C’: [0.25,0.5,0.75,1]	0.5
	solver: [’newton-cg’, ’lbfgs’, ’liblinear’, ’sag’, ’saga’]	newton-cg
	penalty: [L1,L2]	L2
GaussianNB	Default	Default
XGBoost	n_estimators: linespace(start = 50,stop = 3000,num = 60)	250
	max_features [1,3,5,7,9]:	1
	max_depth [3–10]:	10
	min_samples_leaf [1–10]:	9
	min_samples_split [1–10]:	9
Catboost	iterations: linspace(start = 50, stop = 3000,num = 10)	600
	max_depth [1–10]:	9
	subsample [1–10]:	1
Lightgbm	learning_rate: [0.01,0.02,0.05,0.1,0.15]	0.1
	feature_fraction: [0.6, 0.7, 0.8, 0.9, 0.95]	0.6
	max_depth [15,20,25,30,35]:	30
	bagging_fraction: [0.6, 0.7, 0.8, 0.9, 0.95]	0.6
	lambda_l1: [0, 0.1, 0.4, 0.5, 0.6]	0
	lambda_l2: [0, 10, 15, 35, 40]	0

Table 6
The comparison between improved traditional algorithms and machine learning algorithms.

	Model	Accuracy	ROC_AUC	Precision	Recall	F1
Traditional Model	Logistic Regression	0.748	0.747	0.741	0.776	0.758
Traditional Machine Learning Model	Random Forest	0.818	0.818	0.816	0.829	0.822
	Decision Tree	0.726	0.725	0.714	0.77	0.74
	BernoulliNB	0.698	0.700	0.700	0.719	0.708
	KNeighbors	0.746	0.745	0.735	0.782	0.759
	GaussianNB	0.600	0.595	0.567	0.908	0.698
Ensemble Learning Models	XGBoost	0.819	0.819	0.822	0.821	0.822
	Catboost	0.815	0.814	0.806	0.836	0.821
	Lightgbm	0.802	0.802	0.801	0.812	0.806

televisions, cars, and computers in the home. Gregg points out that gains in car ownership and the telephone, are likely to be important in terms of connecting families to social networks and to employment [98]. The property value has a similar action mechanism to the value of durable goods, that is, the more the total real estate value of the family, the higher the opportunities to access social resources and the degree of family economic stability, so they increase the channels to obtain income, thus reducing the possibility of falling into relative poverty. At the same time, the plot indicates the value of the number of books and cooking fuel stand a low situation in importance, indicating that the current people’s living standards are significantly improved, social contradictions caused by infrastructure are becoming closer and closer, and the gap between the rich and the poor is gradually narrowing.

6.1.2. The occupational conditional variables

The occupational conditional variables mainly include “work_type” and “migrant_worker”. The type of work held by respondents was the biggest variable affecting relative poverty. According to whether the respondents work in government institutions, state-owned enterprises, and holding enterprises, they can be classified as working in the formal sector and informal sector. Considering that the respondents were from China rural areas in 2018, this phenomenon can be explained as follows: On the one hand, from the perspective of national development strategy, the government’s investment in agricultural and rural areas in recent years plays a crucial role in promoting agricultural production efficiency, expanding agricultural scale operation and improving the agricultural

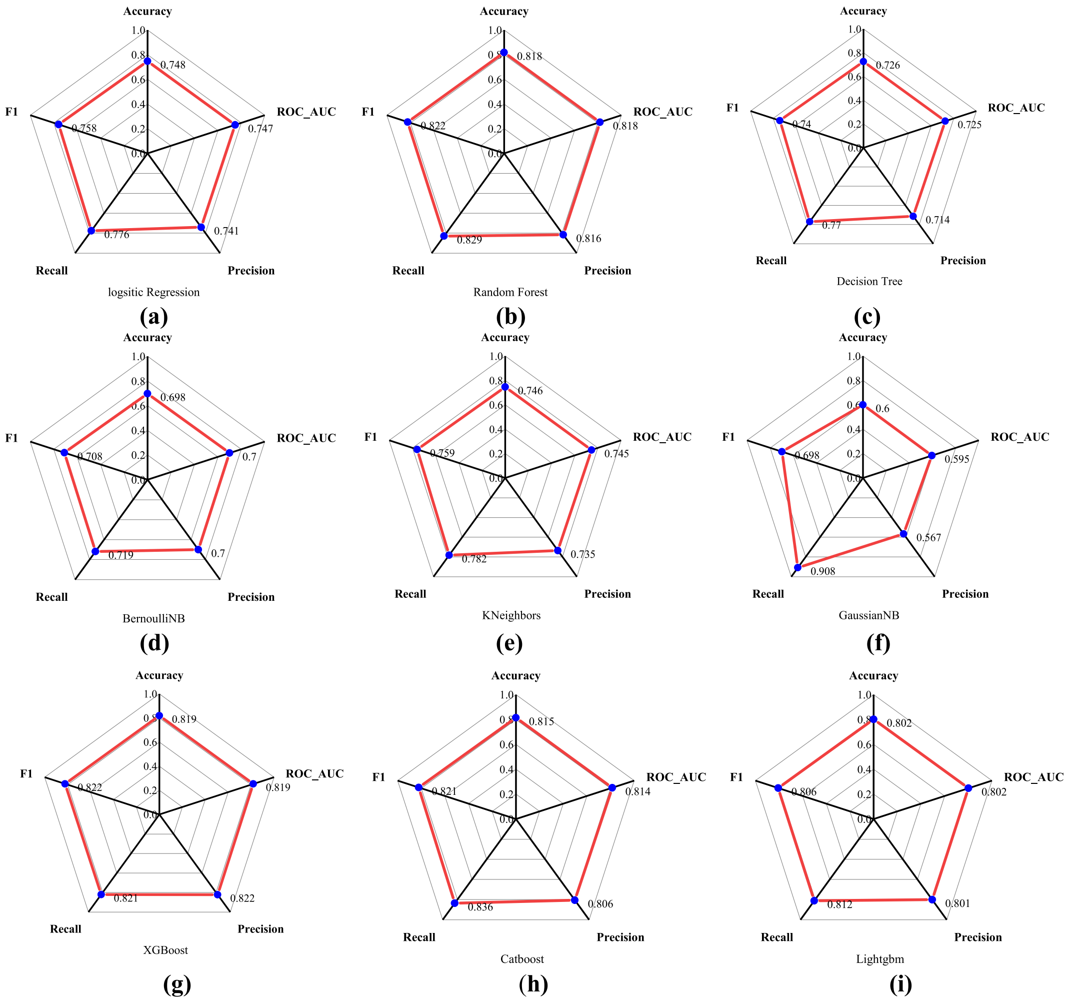


Fig. 5. The radar diagram of each algorithm. The five dimensions of evaluation indicators(Accuracy, F1, Recall, Precision and ROC_AUC) are calculated by logistic Regression(a), Random Forest(b), Decision Tree(c), BernoulliNB(d), KNeighbors(e), GaussianNB(f), XGBoost(g), Catboost(h) and Lightgbm(i).

socialized service system. In particular, the implementation of the rural revitalization strategy in 2018 has taken safeguarding the fundamental interests of farmers as its foothold, greatly expanded the channels for increasing farmers' income, and made up for the shortcomings of agriculture and rural areas, which is of great significance for increasing farmers' income and increasing production. On the other hand, we thought that the rural government agencies are not the most efficient economic units, therefore, to get rid of relative poverty, choosing private enterprises with more prominent economic benefits and independent self-employment models has become the choice of rural residents. At the same time, in terms of market economic development, compared with the low wages of rural grass-roots government departments, the production and operation of private enterprises are more dynamic. Statistics show that since China economic reform and opening up, the private sector uses less than 40% of social assets, but contributes more than 50% of tax revenue. These results can well explain the effect of work variable types on relative poverty in rural areas and also provide a reference value for government policymaking in rural areas.

Whether the respondent is a migrant worker also stands a large weight among all variables in terms of importance. In terms of self-development ability, out-migrating for work can promote the development of the rural labor force and improve their labor skills [117]. In stabilizing the effect of poverty alleviation, out-migrating for work can significantly reduce the vulnerability of families to absolute poverty [118]. However, different from absolute poverty, according to the analysis of data results, the impact of migrant work on relative poverty is small and the effect is not obvious. It indicates that the country's agricultural policies have gradually taken effect in recent years.

To sum up, among the variables of occupational conditional variables, the variable of job type has the largest impact, while whether to work outside the home and whether to have part-time production, have a small impact on relative poverty. In this regard, this paper suggests that the quality of employment of rural residents should be effectively guaranteed, social security services of grassroots government departments should be strengthened, and the ability of private enterprises to integrate social resources in the practice of

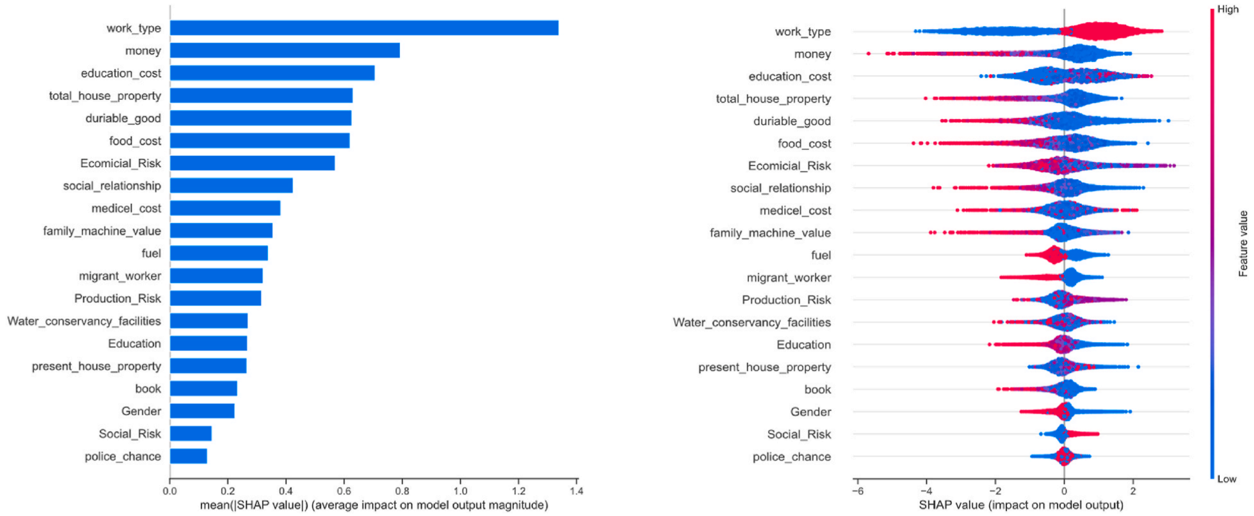


Fig. 6. The importance ranking based on SHAP.

rural revitalization should be enhanced, so as to actively guide them to undertake social responsibilities.

6.1.3. *The house expenditures variables*

The house expenditures variables mainly include “education_cost”, “food_cost”, “medicel_cost” and “social_relationship”. According to 《Lohnarbeit und kapita》written by Marx, consumption expenditure in a family can be divided into subsistence consumption, developmental consumption, and enjoyment consumption [100]. Consumption follows a logical process from subsistence consumption to developmental consumption and then to enjoyment consumption, or it can be considered as a process from material consumption to service consumption [100]. Some scholars say that the poor are limited to the main types of consumption (such as food), while the upper class invests more in technology consumption (such as tourism) and information consumption (such as education, etc.) [101]. Thus, compared with the conservative views of the relatively poor, the non-poor are more willing to invest in education, maintain social networks, and increase capital, as a result, consumer spending shows a high level. The experimental results of this paper show that food expenditure is a key indicator affecting relative poverty, and it is also confirmed that in the subsequent construction of the relative poverty indicator system, the impact of food expenditure on relative poverty should be considered.

At the same time, A large number of scholars have explained the impact of medical expenses and education expenditures on absolute poverty [103,104]. According to the experimental results of this paper, the above three variables still have a great impact on relative poverty. However, different from medical security expenditure and education investment expenditure, social relationship expenditure has a negative correlation with relative poverty, the larger the amount of expenditure, the smaller the probability of falling into relative poverty. Studies have shown that social expenditure promotes communication and integration among social members [105], and strengthens the emotional bond and attachment with relatives and friends [106]. Through social expenditure, "gift-givers" can expand and enhance their social contacts, and make full use of the relationship network of the gift-giving to seek corresponding economic benefits and benefits for themselves. Therefore, social expenditure can affect the relative poverty of the respondents.

6.1.4. *The individual characteristic variables*

The individual characteristic variables mainly include “ecomicial_risk”, “production_risk”, “gender”, “social_risk”, “education” and “police_chance”. The importance ranking of variables shows that the above variables have a low impact on relative poverty. We thought that relative poverty is more reflected in social stratification, while in rural areas, due to the influence of the environment, the backward education level, the limitation of traditional cultural concepts, and the loss of rural elite talents, the differences of individualization within rural towns are not obvious, and the phenomenon of homogeneity among rural residents is more prominent. Therefore, the different results of individual characteristic variables in the representation of relative poverty in rural areas are not obvious. At the same time, we noticed, such as “gender”, “police_chance”, and “education”, listed at the bottom of the relative importance of poverty, showing that the more open view of rural society, the more complete the social security policy and the higher status of rural women in the family, all of these make more obvious female freedom to make decisions.

6.1.5. *The geographical environment variables*

The geographical environment variable mainly includes “water_conservancy_facilities”. Different from absolute poverty, geographical environment has a low effect on relative poverty, total effects of the above variables on relative poverty are limited, some of them even have been deleted. It might have a relation to the implementation of targeted poverty alleviation and regional imbalances elimination strategies in China in recent years. To overcome the constraints of geographical conditions, especially in the mountainous areas and far away areas, relocation policies are set up in rural areas, such as "Plan to connect every village with the last mile of targeted poverty alleviation" and "rural water supply Project" devoting to strengthen drinking water safety. It also verifies all these efforts made by the Chinese government are effective and greatly eliminated the influence of geography on relative poverty in China. Limited production conditions caused by geographical constraints have been greatly improved.

6.1.6. Robustness inspection

In order to ensure the stability of experiment results, we also employ the function of importance based on “weight”. The weight in XGBoost is the number of times to split the selected data across all three. The plot has been shown in Fig. 7. It can be observed from Fig. 7 that “work_type” is the most important feature in all selected features equally. Although the ordering of features may have a little different from SHAP, the major influencing features such as “money”, “fuel”, “total_house_property” also stand in an important position. This also proves the reliability of SHAP from the other side.

6.2. Key Variables Analysis based on part dependence plot

The importance ranking plot denotes the most significant variable in predicting the final output result. However, it doesn’t show whether this feature affects output positively or negatively. Different from it, Partial dependence plots (PDPs) can display the expected target response as a function of the input features of interest and reveal whether the relationship between the independent variables and dependent variables is linear, monotonic, curvilinear, or more complex [119]. In this study, a partial dependence plot(blue line) between variables is shown in Fig. 8. We also apply ICE as a local interpretation technique to observe the influence of each variable on the outcome of each observation, the PDP line defines the average of the line of the ICE plot(grey shadow). From the plot,

it shows that most of the variables have obvious nonlinearity correlation. For example, there is an obvious nonlinearity relationship between the selected variables and whether they are in relative poverty, such as “book”, “education_cost” and “medical_cost”. We also noticed that only a few selected variables were linearly correlated with the dependent variable, such as “social_relationship” and “food_cost”. In summary, the information extracted from Fig. 8 can summarize as following:1) Most relations between existing independent variables and dependent variables are non-linear, inspiring us to use machine learning especially ensemble algorithms to handle the non-linearity; 2) most of the variables present positive relationships such as “duriable_good”, “food_cost” and “education”;

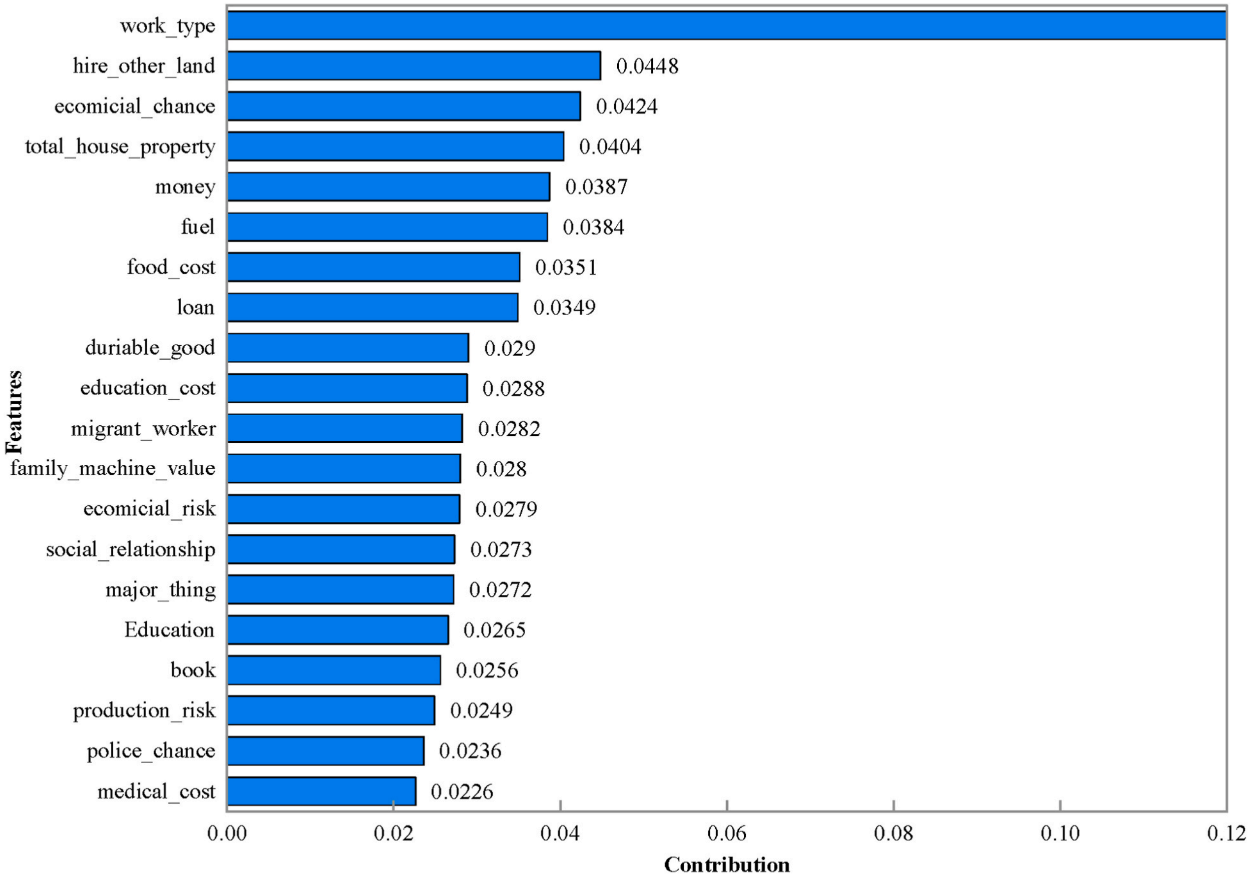


Fig. 7. The importance ranking based on weight.

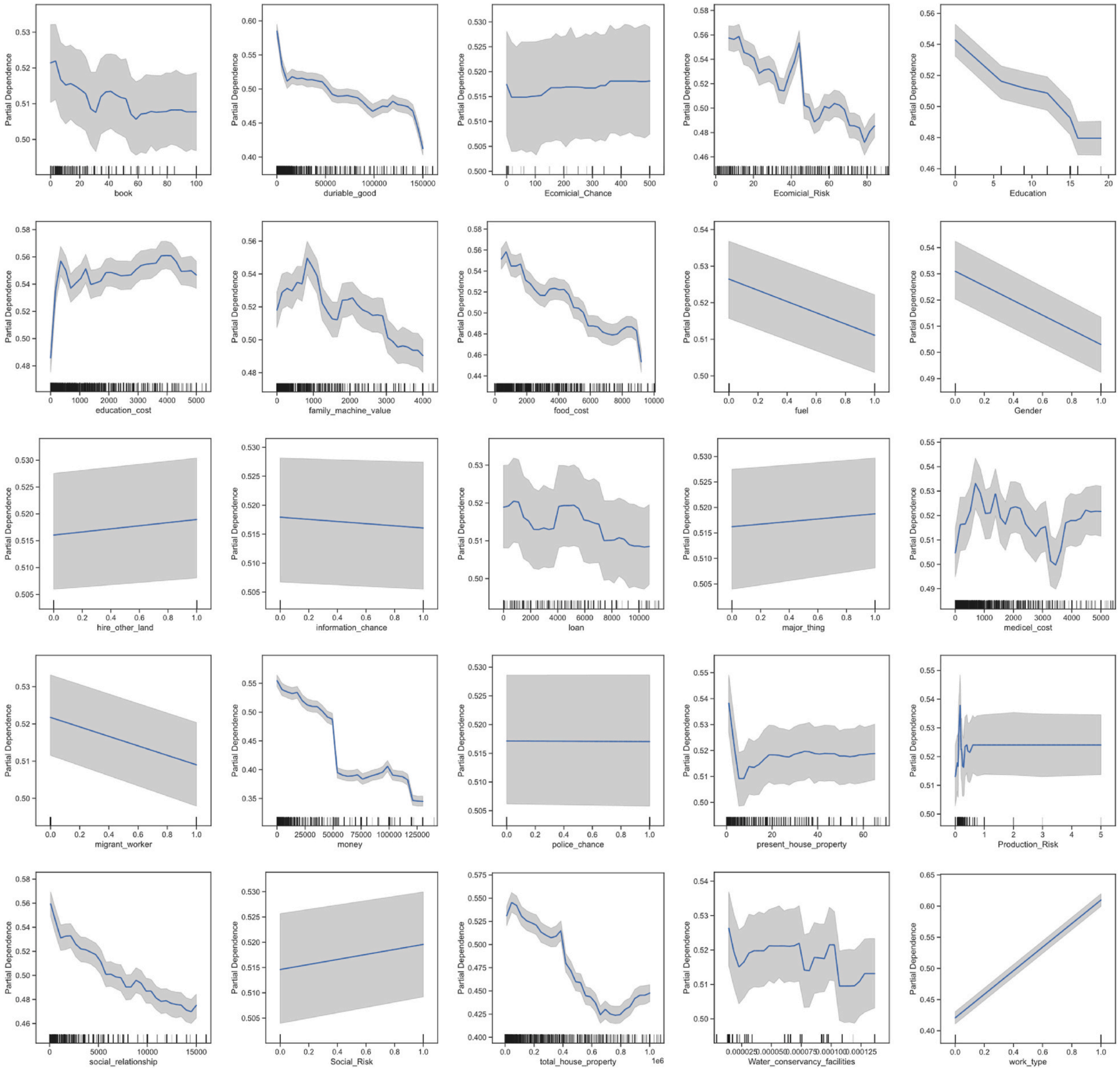


Fig. 8. Partial dependence plots of variables.

some features stand the negative relationship with the relative poverty such as “education_cost” and “work_type”. There also presents the complex relationship between independent variables and dependent variables such as “medical_cost” and “present_house_property”.

7. Conclusion, limitation, and future research

The phenomenon of relative poverty is rooted in the development of a country’s economy. Accurate identification of the relatively poor is a prerequisite for alleviating the problem of relative poverty. Constrained by the data dimension and the data volume, the previous research more focus on the relationship evaluation, influencing mechanism, and relative poverty reduction strategies and et al. by using traditional research methods, such as stratified analyses, econometric models, case-control study panel-VAR modeling, which not only couldn’t recognize the relatively poor accurately but also couldn’t dig out the relative poverty formation mechanism deeply. This paper is one of the first attempts to bring the machine learning algorithm into the identification of the relatively poor in the relative poverty field, achieving accurate identification of the relatively poor. Meanwhile, the article completely describes the machine learning operation process, by using the interpretable model tools, the main influence factors are found and the function mechanism is explored fully. The conclusions of the article are listed as followings: 1) Machine learning algorithm especially ensemble learning is proved it could be well applied in relative poverty fields, especially XGBoost algorithm, which achieves 81.9% accuracy and the score of ROC_AUC reaches 0.819. In other words, compared with the performance of other ML and traditional models in the same datasets, results show XGBoost is the most suitable model for the relatively poor identification in China's rural areas and also provides a novelty research approach in relative poverty fields. 2) This study sheds light on many new research directions in applying machine learning for relative poverty research, besides, the paper offers an integral framework and beneficial reference for target identification using a machine learning algorithm, presenting the whole experiment process of machine learning including variables select and database rebuild, data preprocessing, identification algorithms proposed and algorithms interpretability. 3) In addition, by utilizing the interpretable tools, the "black box" of ML become transparent through PDP and SHAP explanation, it also reveals that machine learning models can readily handle the non-linear association relationship. Based on the above experiment, the key influencing factors are found that "work_type" could account for the condition of relative poverty mostly, the "money" and "education_cost" stand second and third place respectively. The global explain ability components of the PDP and SHAP indicate relative poverty differs from absolute poverty, the function mechanism as well as influencing factors show a different situation.

Notwithstanding the contributions of the research, it is not without limitations. Firstly, considering the accessibility of data, this paper only chooses "year" as the time scale to build the dataset. Second, strict to analyzing technique, the experimental process of this paper focuses on the influence of a single variable on the relative poverty state, which fails to consider the cumulative effect between variables, as well as has not been able to analyze the intermediate and moderating effects between variables. Third, the whole relative poverty condition of rural China is analyzed, however, limited by geographical conditions and economic development, there existed significant differences in the field of relative poverty among different regions. There also have spatial heterogeneity of relative poverty in different time and space. Fourth, a major classification judgment criterion in this paper is whether the target individual is in a relative poverty state, which can be described binary classification model.

Based on the above description, future research should explore the effect of time on relative poverty, especially induced by time variation, the change in relative poverty state should get more emphasized; Besides, further research should adopt diversified research methods such as combining the machine learning and traditional empirical tools, demonstrate the interaction of multi variables together; Further studies also should subdivide the classification criterion of relative poverty at great length, like general relative poverty, moderate relative poverty, and severe relative poverty. Furthermore, satellite data has gotten widely available attention either land use in land classification [120] or sea surface temperature prediction [121], using satellite datasets is the direction of relative poverty research, and should deserve more scholars' focus.

Author contribution statement

Wei Huang: Conceived and designed the experiments.

Yinke Liu: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Peiqi Hu; Shuhui Gao: Performed the experiments.

Shiyu Ding: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Ming Zhang: Contributed reagents, materials, analysis tools or data.

Data availability statement

Data included in article/supplementary material/referenced in article.

Additional information

No additional information is available for this paper.

Declaration of competing 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.

Acknowledgement

This study was funded by the National Social Science Foundation of China (Grant No. 19BGL224); Major Science and Technology Projects in Henan Province, China (Grant No.201110210300); Postgraduate Innovation Project of North China University of Water Resources and Electric Power: Assessment measurement, evolution law and path selection of rural livelihood resilience in China under the background of rural revitalization (Grant No, NCWUYC-2023081).

Appendix 1

Table 1
The meaning and assignment of the selected 50 variables

Index	Variable Name	Description	source	Database
Individual Characteristics	Age	Age of respondents	Sun [122]	CFPS
	Gender	Gender of respondents, Male = 1, Female = 0	Sun [122]	
	Province	Province where the survey object is located.	Goh [123]	
	Marriage	Marital status of respondents, 1 for married, 0 for unmarried, divorced, or widowed	Sun [122]	
	family_size	Number of people in the household of the survey respondents	Song [124]	
	Social_risk	Whether the respondents have commercial insurance, 1 if there is insurance, 0 if not	Renahy [125]	
	Ecomicial_Risk	Weekly working hours of respondents	Kenworthy [126]	
	work_type	Respondent's Job Type. formal sector = 1, informal sector = 0	Cichello [127]	
	migrant_worker	Whether the respondent is currently working outside the home, yes = 1, no = 0	Peng [128]	
	farm_forest	Whether the respondent's family is engaged in agriculture, forestry, animal husbandry and by-fishing, yes = 1,no = 0	Ghimire [129]	
	major_thing	Whether there was a major incident in the household of the respondent in the previous year, yes = 1,no = 0	Healey [130]	
	loan	The per capita loan amount of the respondents' households	Burgess [131]	
	total_house_property	Respondent's household total expenditure	Bartoşová [132]	
	food_cost	Respondent's household food expenditure	Bartoşová [132]	
	medicel_cost	Respondent's household medical expenditure	Bartoşová [132]	
	consume_cost	Respondent's household consume expenditure	Bartoşová [132]	
	education_cost	Respondent's household education expenditure	Lin [133]	
Individual Characteristics	Production_risk	Respondent's family dependency ratio	Allen [134]	CFPS
	water	Respondents' domestic water consumption	Adetayo [135]	
	fuel	Respondent's living and cooking fuel	Yu [136]	
	Ecomicial_Chance	Whether respondents received funds from family and friends when they were in difficulty, yes = 1,no = 0	Song [124]	
	police_chance	Whether respondents received policy assistance from government when they were in difficulty, yes = 1,no = 0	Song [124]	
	information_chance	Whether respondents were able to use modern information technology to obtain more information, yes = 1,no = 0	Yang [137]	
	book	Respondent's household book collection	Tran [138]	
	family_machine_value	Gross value of respondents' household farm machinery	Yu [136]	
	Health	Respondent's health status	Sun [122]	
	Education	Respondent's highest education. 0 years for illiterate/semi-literate, 6 years for primary school, 9 years for junior high school, 12 years for high school/secondary school/technical school, 15 years for junior college, 16 years for undergraduate, 19 years for master, and 22 years for doctorate	Sun [122]	
	hire_other_land	Whether the respondent's family rents other people's land, yes = 1,no = 0	Tan [139]	
	duriable_good	Total value of durable goods in respondents' households	Maitra [42]	
	house_type	Type of house respondent currently live in	Adetayo [135]	
	Prensent_house_property	Property value of the respondent's current house	Wang [140]	
	Total_house_property	Property value of the respondent's total house	Wang [140]	
	income_state	Respondent's self-rated income status	Yu [136]	
	money	Total household cash and savings of respondents	Arcanjo [141]	
	isei	Respondent's socioeconomic status	Yu [136]	
	social_state	Respondent's self-rated social status	Yu [136]	CFPS
	social_relationship	Respondent's family spending on personal exchanges	Wang [140]	
Psychological endowment	confidence	Respondents' confidence in their future	Kuruvilla [142]	

(continued on next page)

Table 1 (*continued*)

Index	Variable Name	Description	source	Database
	pressure	Whether respondents can cope well with stress yes = 1,no = 0	Kuruvilla [142]	
	likehold	Respondent's personal well-being	Sun [122]	
	satisfaction	Respondents' Satisfaction	Kuruvilla [142]	
Geographical Environment	temperture	Average temperature in the province where the respondents are located	Lin [133]	2019 Statistical Yearbook by Province
	rain	Average precipitation in the province where the respondents are located	Lin [133]	
	sunlight	Number of sunshine hours in the province where the respondents are located	Liu [143]	
	sea_level	Average altitude of the respondents' province	Wang [114]	
	fertilizer_consuptiom	Per capita agricultural fertilizer application in the province where the respondents are located	Shita [144]	
	agricultive_crop	Per capita gross output value of agriculture, forestry, animal husbandry and fishery in the province where the respondents are located	Ghimire [129]	
	machine_power	Total power of agricultural machinery per capita in the province where the respondents are located	Yu [136]	
	crop_production	Total sown area of crops per capita in the province where the respondents are located	Wang [114]	
	Water_conservancy_facilities	Number of reservoirs per capita in the province where the respondents are located (sets)	Madulu [145]	BFNCW

References

[1] W. Bank, Poverty and Shared Prosperity 2022: Correcting Course, The World Bank, 2022.

[2] E. Yelin, J. Yazdany, L. Trupin, Relationship between poverty and mortality in systemic lupus erythematosus, Arthritis Care Res. 70 (7) (2018) 1101–1106.

[3] Y. Sawada, Y. Takasaki, Natural disaster, poverty, and development: an introduction, World Dev. 94 (2017) 2–15.

[4] K. Plax, et al., An essential role for pediatricians: becoming child poverty change agents for a lifetime, Academic Pediatrics 16 (3) (2016) S147–S154.

[5] L. Yansui, C. Zhi, Supply-side structural reform and its strategy for targeted poverty alleviation in China, Bull. Chin. Acad. Sci. 32 (10) (2017) 1066–1073.

[6] Y.S. Liu, Y.Z. Guo, Y. Zhou, Poverty alleviation in rural China: policy changes, future challenges and policy implications, China Agric. Econ. Rev. 10 (2) (2018) 241–259.

[7] Y. Guo, Y. Zhou, Y. Liu, Targeted poverty alleviation and its practices in rural China: a case study of Fuping county, Hebei Province, J. Rural Stud. 93 (2019) 430–440.

[8] A. Sen, Poor, relatively speaking, Oxf. Econ. Pap. 35 (2) (1983) 153–169.

[9] M. Ravallion, S.H. Chen, Weakly relative poverty, Rev. Econ. Stat. 93 (4) (2011) 1251–1261.

[10] H. Lian, et al., Changes and decomposition of rural relative poverty in China:2002 ~ 2018, J Quantitative Tech Eco 38 (2) (2021) 132–146.

[11] E.Y. Barri, et al., Understanding transit ridership in an equity context through a comparison of statistical and machine learning algorithms, J. Transport Geogr. 105 (2022), 103482.

[12] X. Jin, et al., Knowledge source strategy and enterprise innovation performance: dynamic analysis based on machine learning, Technol. Anal. Strat. Manag. 30 (1) (2018) 71–83.

[13] B. Luo, A method for enterprise network innovation performance management based on deep learning and Internet of Things, Math. Probl Eng. (2022) 2022.

[14] T. Eom, C. Woo, D. Chun, Predicting an ICT Business Process Innovation as a Digital Transformation with Machine Learning Techniques, Technology Analysis & Strategic Management, 2022, pp. 1–13.

[15] R. Carbonneau, K. Laframboise, R. Vahidov, Application of machine learning techniques for supply chain demand forecasting, Eur. J. Oper. Res. 184 (3) (2008) 1140–1154.

[16] Z.H. Kilimci, et al., An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain, Complexity 2019 (2019) 16.

[17] J. Feizabadi, Machine learning demand forecasting and supply chain performance, Int. J. Logist. Res. Appl. 25 (2) (2020) 119–142.

[18] E. Rozos, Machine learning, urban water resources management and operating policy, Resources 8 (4) (2019) 173.

[19] A.S. Abowarda, et al., Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale, Rem. Sens. Environ. 255 (2021), 112301.

[20] S. Matloob, Y. Li, K.Z. Khan, Safety measurements and risk assessment of coal mining industry using artificial intelligence and machine learning, Open J. Bus. Manag. 9 (3) (2021) 1198–1209.

[21] D.K. Zhao, et al., Using random forest for the risk assessment of coal-floor water inrush in Panjiayao Coal Mine, northern China, Hydrogeol. J. 26 (7) (2018) 2327–2340.

[22] J. Guo, S. Qu, T. Zhu, Estimating China's Relative and Multidimensional Poverty: Evidence from Micro-level Data of 6145 Rural Households, vol. 26, World Development Perspectives, 2022, 100402.

[23] F. Xia, Z. Zhang, X. Wang, Hometown attachment or urban dependence? The reciprocal effects between multi-dimensional relative poverty of migrant workers and urban-rural land dependence, Habitat Int. 137 (2023), 102850.

[24] B. Gustafsson, D. Sai, Growing into relative income poverty: urban China, 1988–2013, Soc. Indicat. Res. 147 (1) (2020) 73–94.

[25] G. Wan, et al., From equality of deprivation to disparity of prosperity: the poverty–growth–inequality triangle in post-reform China, China World Econ. 26 (2) (2018) 50–67.

[26] A. Sen, Development as Freedom, Oxford Paperbacks, 2001.

[27] X. Meng, R. Gregory, G. Wan, Urban poverty in China and its contributing factors, 1986–2000, Rev. Income Wealth 53 (1) (2007) 167–189.

[28] G. Xibao, Z. Qiang, Long-term multidimensional poverty, inequality and poverty-causing facto rs, Econ. Res. 51 (6) (2016) 143–156.

[29] J. You, A. Kontoleon, S. Wang, Identifying a sustained pathway to multidimensional poverty reduction: evidence from two Chinese provinces, J. Dev. Stud. 55 (1) (2019) 137–158.

[30] Z. Wang, et al., Multidimensional poverty alleviation effect of different rural land consolidation models: a case study of Hubei and Guizhou, China, Land Use Pol. 123 (2022), 106399.
[31] R.M. Vijaya, R. Lahoti, H. Swaminathan, Moving from the household to the individual: multidimensional poverty analysis, World Dev. 59 (2014) 70–81.
[32] A. Berihuete, C.D. Ramos, M.A. Sordo, Welfare, inequality and poverty analysis with rtip: an approach based on stochastic dominance, R J 10 (1) (2018) 328.
[33] Q. Chen, et al., Coupling analysis on ecological environment fragility and poverty in South China Karst, Environ. Res. 201 (2021), 111650.
[34] S. Alkire, J. Foster, Counting and multidimensional poverty measurement, J. Publ. Econ. 95 (7–8) (2011) 476–487.
[35] S. Alkire, J.M. Roche, A. Vaz, Changes over time in multidimensional poverty: methodology and results for 34 countries, World Dev. 94 (2017) 232–249.
[36] R. Angulo, From multidimensional poverty measurement to multisector public policy for poverty reduction: lessons from the Colombian case, OPHI Working Papers (102) (2016).
[37] Y. Li, K. Shen, Fiscal expenditure structure, relative poverty and economic growth, Manag. World 11 (2007) 14–26.
[38] J. Qin, A. Rong, Analysis of the impact of fiscal expenditure structure on rural relative poverty, Econ. Issues 11 (2012) 95–98.
[39] D. Benjamin, L. Brandt, J. Giles, Inequality and Growth in Rural China: Does Higher Inequality Impede Growth?, 2006.
[40] S. Cai, A. Park, S. Wang, Microfinance Can Raise Incomes: Evidence from a Randomized Control Trial in China, HKUST Business School Research Paper, 2020-006, p. 2020.
[41] O.A. Adekoya, Analysis of farm households poverty status in Ogun states, Nigeria, Asian Econ. Financ. Rev. 4 (3) (2014) 325.
[42] S. Maitra, The poor get poorer: tracking relative poverty in India using a durables-based mixture model, J. Dev. Econ. 119 (2016) 110–120.
[43] A. Kuruvilla, K. Jacob, Poverty, social stress & mental health, Indian J. Med. Res. 126 (4) (2007) 273–278.
[44] H. Lin, et al., Measurement and identification of relative poverty level of pastoral areas: an analysis based on spatial layout, Environ. Sci. Pollut. Control Ser. 29 (58) (2022) 87157–87169.
[45] Y. Liu, Y. Xu, A geographic identification of multidimensional poverty in rural China under the framework of sustainable livelihoods analysis, Appl. Geogr. 73 (2016) 62–76.
[46] H. Tanaka, et al., Relationship of relative poverty and social relationship on mortality around retirement: a 10-year follow-up of the Komo-Ise cohort, Environ. Health Prev. Med. 23 (1) (2018) 64.
[47] E. Habtamu, et al., Trachoma and relative poverty: a case-control study, PLoS Negl Trop Dis 9 (11) (2015), e0004228.
[48] X. Cheng, et al., Building a sustainable development model for China's poverty-stricken reservoir regions based on system dynamics, J. Clean. Prod. 176 (2018) 535–554.
[49] S. Moller, et al., Determinants of relative poverty in advanced capitalist democracies, Am. Socio. Rev. (2003) 22–51.
[50] G.H. Wan, X.S. Hu, W.Q. Liu, China's poverty reduction miracle and relative poverty: focusing on the roles of growth and inequality, China Econ. Rev. 68 (2021), 101643.
[51] Z.A. Hatta, I. Ali, Poverty reduction policies in Malaysia: trends, strategies and challenges, Asian Cult. Hist. 5 (2) (2013) 48.
[52] J. Foster, J. Greer, E. Thorbecke, A class of decomposable poverty measures, Econometrica: J. Econom. Soc. (1984) 761–766.
[53] LiuHong, ZhangXiangxiang, Relative poverty:connotative characteristics , multidimensional dilemmas and research prospects, World Agric. (6) (2022).
[54] A. Abadie, M.D. Cattaneo, Econometric methods for program evaluation, Annual Review of Economics 10 (2018) 465–503.
[55] T. Wuest, et al., Machine learning in manufacturing: advantages, challenges, and applications, Production & Manufacturing Research 4 (1) (2016) 23–45.
[56] T. Howley, et al., The effect of principal component analysis on machine learning accuracy with high dimensional spectral data, December 2005, in: Applications and Innovations in Intelligent Systems XIII: Proceedings of AI-2005, the Twenty-Fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, Cambridge, UK, 2006.
[57] D. Sun, M. Wang, A. Li, A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data, IEEE/ACM Trans Comput Biol Bioinform 16 (3) (2018) 841–850.
[58] Y.X. Xie, et al., Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances, J. Petrol. Sci. Eng. 160 (2018) 182–193.
[59] Y. Okada, et al., Comparisons of machine learning algorithms for application identification of encrypted traffic, in: 10th International Conference on Machine Learning and Applications and Workshops. 2011, IEEE, 2011.
[60] S. Mangalathu, S.H. Hwang, J.S. Jeon, Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach, Eng. Struct. 219 (2020), 110927.
[61] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[62] A.K. Mahlein, et al., Quantitative and qualitative phenotyping of disease resistance of crops by hyperspectral sensors: seamless interlocking of phytopathology, sensors, and machine learning is needed!, Curr. Opin. Plant Biol. 50 (2019) 156–162.
[63] V. Galdo, Y. Li, M. Rama, Identifying urban areas by combining human judgment and machine learning: an application to India, J. Urban Econ. 125 (2021), 103229.
[64] E. Ceylan, F.O. Kutlubay, A.B. Bener, Software defect identification using machine learning techniques, in: *32nd* EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO'06), IEEE, 2006.
[65] J. Wäldchen, P. Mäder, N. Cooper, Machine learning for image based species identification, Methods Ecol. Evol. 9 (11) (2018) 2216–2225.
[66] C. Fan, et al., Deep learning-based feature engineering methods for improved building energy prediction, Appl. Energy 240 (2019) 35–45.
[67] H. Yu, et al., Integrating machine learning interpretation methods for investigating nanoparticle uptake during seed priming and its biological effects, Nanoscale 14 (41) (2022) 15305–15315.
[68] M. Raissi, G.E. Karniadakis, Hidden physics models: machine learning of nonlinear partial differential equations, J. Comput. Phys. 357 (2018) 125–141.
[69] Q.R. Fan, et al., Advancing theoretical understanding and practical performance of signal processing for nonlinear optical communications through machine learning, Nat. Commun. 11 (1) (2020) 3694.
[70] S.C. Lu, Machine learning approaches to knowledge synthesis and integration tasks for advanced engineering automation, Comput. Ind. 15 (1–2) (1990) 105–120.
[71] H.A. Simon, Why should machines learn?, in: Machine Learning Elsevier, 1983, pp. 25–37.
[72] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[73] A.J. Myles, et al., An introduction to decision tree modeling, J. Chemometr.: J ChemSoc 18 (6) (2004) 275–285.
[74] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[75] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.
[76] Y.-Y. Song, L. Ying, Decision tree methods: applications for classification and prediction, Shanghai archives of psychiatry 27 (2) (2015) 130.
[77] N. Wang, B.W. Hughes, Efficiency and the Mitigation of Carbon Emissions in Semi-truck Transportation, 2022.
[78] W. Zhao, N. Herencsar, Logistic regression analysis of targeted poverty alleviation with big data in mobile network, Mobile Network. Appl. (2022) 1–12.
[79] C. Peng, et al., Determinants of poverty and their variation across the poverty spectrum: evidence from Hong Kong, a high-income society with a high poverty level, Soc. Indicat. Res. 144 (2019) 219–250.
[80] N.F. Da Silva, E.R. Hruschka, E.R. Hruschka Jr., Tweet sentiment analysis with classifier ensembles, Decis. Support Syst. 66 (2014) 170–179.
[81] W. Ramadhan, S.A. Novianty, S.C. Setianingsih, Sentiment analysis using multinomial logistic regression, in: 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), IEEE, 2017.
[82] S. Symeonidis, D. Effrosynidis, A. Arampatzis, A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis, Expert Syst. Appl. 110 (2018) 298–310.
[83] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[84] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016.

[85] E. Al Daoud, Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset, Int J Com Inf Eng 13 (1) (2019) 6–10.

[86] Y.R. Zhang, A. Haghani, A gradient boosting method to improve travel time prediction, Transport. Res. C Emerg. Technol. 58 (2015) 308–324.

[87] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, ArXiv 2018 (2018). /abs/1810.11363.

[88] Q. Meng, et al., A communication-efficient parallel algorithm for decision tree, Adv. Neural Inf. Process. Syst. 29 (2016).

[89] A. Klein, et al., Fast bayesian optimization of machine learning hyperparameters on large datasets, in: Artificial Intelligence and Statistics, PMLR, 2017.

[90] G. Ke, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).

[91] D. Sun, et al., Assessment of landslide susceptibility along mountain highways based on different machine learning algorithms and mapping units by hybrid factors screening and sample optimization, Gondwana Res. (2022) (in press), https://doi.org/10.1016/j.gr.2022.07.013.

[92] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[93] Y. Xie, J. Hu, An introduction to the China family panel studies (CFPS), Chinese sociological review 47 (1) (2014) 3–29.

[94] z.Y. Zhang lin, Dynamic Poverty of Rural Areas in China:Condition Transformation and Durability——Based on Survival Analysis on Micro-data of Health and Nutrition Survey in China, Journal of Huazhong Agricultural University (Social Sciences Edition), 2021.

[95] YechSheng, Zhaorui, Dynamic poverty of rural areas in China:condition transformation and durability——based on survival analysis on micro-data of health and nutrition survey in China, J Huazhong Agri Univ (Social Sciences Edition) (3) (2013) 42–52.

[96] T.Q. Tran, et al., The influence of contextual and household factors on multidimensional poverty in rural Vietnam: a multilevel regression analysis, Int. Rev. Econ. Finance 78 (2022) 390–403.

[97] S. Nosier, R. Beram, M. Mahrous, Household Poverty in Egypt: Poverty Profile, Econometric Modeling and Policy Simulations, 2021.

[98] J. Jalan, M. Ravallion, Geographic poverty traps? A micro model of consumption growth in rural China, J. Appl. Econom. 17 (4) (2002) 329–346.

[99] N. Chaudhuri, et al., On the platform but will they buy? Predicting customers' purchase behavior using deep learning, Decis. Support Syst. 149 (2021), 113622.

[100] C.F. Tsai, M.L. Li, W.C. Lin, A class center based approach for missing value imputation, Knowl. Base Syst. 151 (2018) 124–135.

[101] E. Zinovyeva, W.K. Hardle, S. Lessmann, Antisocial online behavior detection using deep learning, Decis. Support Syst. 138 (2020), 113362.

[102] M. Allen, The problem of multicollinearity. Understanding regression analysis, Plenum Press, 1997, pp. 176–180, https://doi.org/10.1007/b102242.

[103] G. Grekousis, et al., Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: a geographical random forest approach, Health Place 74 (2022), 102744.

[104] J.Y.-L. Chan, et al., Mitigating the multicollinearity problem and its machine learning approach: a review, Mathematics 10 (8) (2022) 1283.

[105] Y. Luo, J. Yan, S. McClure, Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis, Environ. Sci. Pollut. Control Ser. 28 (6) (2021) 6587–6599.

[106] J. Groß, Variance inflation factors, R. News 3 (1) (2003) 13–15.

[107] T.N. Achia, A. Wangombe, N. Khadioli, A logistic regression model to identify key determinants of poverty using demographic and health survey data, 2010.

[108] I. Shhadat, A. Hayajneh, Z.A. Al-Sharif, The use of machine learning techniques to advance the detection and classification of unknown malware, Procedia Comput. Sci. 170 (2020) 917–922.

[109] S.K. Singh, et al., Predicting sustainable arsenic mitigation using machine learning techniques, Ecotoxicol. Environ. Saf. 232 (2022), 113271.

[110] J. Cai, et al., Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79.

[111] A.L.I. Oliveira, et al., GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation, Inf. Software Technol. 52 (11) (2010) 1155–1166.

[112] M. Reif, F. Shafait, A. Dengel, Meta-learning for evolutionary parameter optimization of classifiers, Mach. Learn. 87 (3) (2012) 357–380.

[113] M.M. Ali, et al., Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison, Comput. Biol. Med. 136 (2021), 104672.

[114] S. Mangalathu, S.-H. Hwang, J.-S. Jeon, Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach, Eng. Struct. 219 (2020), 110927.

[115] M. Sherraden, N. Gilbert, Assets and the poor: New American welfare policy, Routledge, 2016.

[116] J. Xianlin, The poor have assets" – a policy proposal to enrich the "working poor, Truth Seeking (1) (2008) 44–46.

[117] P. Gregg, J. Waldfogel, E. Washbrook, Family expenditures post-welfare reform in the UK: are low-income families starting to catch up? Lab. Econ. 13 (6) (2006) 721–746.

[118] G. Ruo-chen, L. Shi, The impacts of the outflow of the rural labor force on its left-behind household's economic condition:an empirical analysis based on the rural household's poverty and vulnerability, J. Beijing Normal Univ. (Nat. Sci.) (4) (2018) 132–140.

[119] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[120] A. Alam, M.S. Bhat, M. Maheen, Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley, Geojournal 85 (2020) 1529–1543.

[121] C.J. Xiao, et al., Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach, Rem. Sens. Environ. 233 (2019), 111358.

[122] H. Sun, et al., Differences and influencing factors of relative poverty of urban and rural residents in China based on the survey of 31 provinces and cities, Int J Environ Res Public Health 19 (15) (2022) 9015.

[123] C.-c. Goh, L. Xubei, Z. Nong, Income growth, inequality and poverty reduction: a case study of eight provinces in China, China Econ. Rev. 20 (3) (2009) 485–496.

[124] J. Song, L. Geng, S. Fahad, Agricultural factor endowment differences and relative poverty nexus: an analysis of macroeconomic and social determinants, Environ. Sci. Pollut. Res. Int. 29 (35) (2022) 52984–52994.

[125] E. Renahy, et al., Connections between unemployment insurance, poverty and health: a systematic review, Eur J Public Health 28 (2) (2018) 269–275.

[126] L. Kenworthy, I. Marx, *In-work poverty in the United States.* Handbook on in-work poverty, 2018, pp. 328–344.

[127] P. Cichello, M. Rogan, A job in the informal sector reduces poverty about as much as a job in the formal sector, Econ. Times 3x3 (2017).

[128] J. Peng, J. Chen, L. Zhang, Gender-differentiated poverty among migrant workers: aggregation and decomposition analysis of the Chinese case for the years 2012–2018, Agriculture 12 (5) (2022) 683.

[129] K. Ghimire, Forest or farm? The politics of poverty and land hunger in Nepal, Manohar Publications, 1998.

[130] J. Healey, Marginality and misfortune: poverty and social welfare in Lancashire, c. 1630-1760, Oxford University, 2008.

[131] R. Burgess, R. Pande, G. Wong, Banking for the poor: evidence from India, J. Eur. Econ. Assoc. 3 (2–3) (2005) 268–278.

[132] J. Bartoşová, V. Bína, Influence of the relative poverty on the structure of household expenditures in the Czech Republic, in: ICABR: VI. International Conference on Applied Business Research Ras Al Khaimah, 2010.

[133] H. Lin, et al., Measurement and identification of relative poverty level of pastoral areas: an analysis based on spatial layout, Environ. Sci. Pollut. Res. Int. (2022) 1–13.

[134] R.H. Allen, The role of family planning in poverty reduction, Obstet. Gynecol. 110 (5) (2007) 999–1002.

[135] A.O. Adetayo, Analysis of farm households poverty status in Ogun states, Nigeria, Asian Econ. Financ. Rev. 4 (3) (2014) 325–340.

[136] C. Yu, et al., Eliminating deprivation and breaking through dependence: A mechanism to help poor households achieve sustainable livelihoods by targeted poverty alleviation strategy, Growth and Change, 2022.

[137] L. Yang, et al., Mobile Internet use and multidimensional poverty: evidence from A household survey in rural China, Soc Indic Res 158 (3) (2021) 1065–1086.

[138] T.D. Tran, S. Luchters, J. Fisher, Early childhood development: impact of national human development, family poverty, parenting practices and access to early childhood education, Child Care Health Dev. 43 (3) (2017) 415–426.

[139] S. Tan, N. Heerink, F. Qu, Land fragmentation and its driving forces in China, Land Use Pol. 23 (3) (2006) 272–285.

[140] H.J. Wang, et al., Poverty and subjective poverty in rural China, Soc. Indicat. Res. 150 (1) (2020) 219–242.

[141] M. Arcanjo, et al., Child poverty and the reform of family cash benefits, J. Soc. Econ. 43 (2013) 11–23.

[142] A. Kuruvilla, K.S. Jacob, Poverty, social stress & mental health, Indian J. Med. Res. 126 (4) (2007) 273–278.

[143] Y.H. Liu, Y. Xu, A geographic identification of multidimensional poverty in rural China under the framework of sustainable livelihoods analysis, Appl. Geogr. 73 (2016) 62–76.

[144] A. Shita, N. Kumar, S. Singh, Technology, poverty and income distribution nexus: the case of fertilizer adoption in Ethiopia, African Development Review-Revue Africaine De Developpement 33 (4) (2021) 742–755.

[145] N.F. Madulu, Environment, poverty and health linkages in the Wami River basin: a search for sustainable water resource management, Phys. Chem. Earth, Parts A/B/C 30 (11–16) (2005) 950–960.