MACHINE LEARNING APPROACH FOR IDENTIFYING RELATIVE POVERTY OF URBAN HOUSEHOLDS: A CASE STUDY OF CHINA

Susan (Sixue) JIA[1], Pengling YU[2]

This paper proposes a new perspective for identifying relative poverty of urban households based on the MPI. We construct a multidimensional indicators system based on 4 dimensions: demographic characteristics, education, basic livelihood security, and living conditions, taking into account the situation of urban poverty in China. A Random Forest model and a Logistic Regression model were constructed using the China Family Panel Studies data. Research results show the overall accuracy of the RF model prediction is at 80.00%, which proves that the constructed multidimensional indicators system and identification prediction model have a good performance.

Keywords: Urban relative poverty, Machine learning, Indicator construction, Random Forest model, Logistic Regression model

1. Introduction

Poverty, listed by the United Nations (UN) as the top three themes of social development issues, has plagued the development of the world today and has been an important issue of global concern. One of the major obstacles facing China's Targeted Poverty Alleviation policies is the accuracy of identification, which leads to the inability to effectively intervene against poverty. Meanwhile, poverty identification is also a key aspect and a major challenge for governments, organizations and institutions around the world in the process of poverty eradication. Accurate poverty identification not only affects the fairness and efficiency of resource allocation for poverty alleviation, but also has a bearing on the subsequent poverty monitoring and governance [1].

China has a large population and a high incidence of poverty and has made great contributions to the cause of poverty reduction in recent years [2], so studying poverty in China has a wide audience and great significance. Absolute poverty and regional overall poverty in China are gradually eliminated, but urban poverty and relative poverty formed by layoffs and unemployment has turned into a prominent phenomenon and serious problem with the development of economic globalization

[1] School of Finance and Business, Shanghai Normal University, Shanghai, China, e-mail: qiuxue1220@gmail.com
[2] School of Finance and Business, Shanghai Normal University, Shanghai, China

and social transformation. Although it still has a large gap with rural poverty in terms of aggregate indicators, the growth of urban poverty incidence has accelerated significantly, and the relative poverty incidence is at least comparable to those of many rich countries in 2013 [3], a large part of the countries in China is in relative poverty [4]. Most scholars tend to associating poverty with the countryside, which causes the ignorance on sustainable development of the "urban poverty" in academics, thus affecting the sustainable development of the whole society.

Meanwhile, absolute poverty is a phase, but relative poverty is long-term and permanent [5], and countries that are currently committed to eliminating absolute poverty are very likely to enter the situation of managing relative poverty in future. Therefore the research on this situation has certain general applicability, it is necessary to take China as an example to study the identification problem of relative urban poor.

The pain point of poverty identification is that in the current practice of poverty alleviation, there are obvious identification biases, serious omission errors and inclusion errors caused by the single identification criteria, interference of subjective factors, rent-seeking behaviors of agents, and insufficient technology. Therefore, we construct a poverty identification indicators system for urban households in China to explore the identification of relative poverty of urban households by using machine learning technology. The indicators system follows part of the Multidimensional Poverty Index (MPI), and we select other indicators, taking into account the national conditions and situation, and the characteristics of urban households of China. The results of the Random Forest (RF) model and Logistic Regression (LR) model prove that it is scientific and reasonable.

2. Literature review

2.1 Definition and measurement of poverty criteria

Urban poverty mainly can be divided into absolute poverty and relative poverty. Absolute poverty is the inability to sustain consumption of goods and services necessary for basic survival. Relative poverty refers to a certain proportion of people whose income can reach or maintain the basic survival needs under a certain level of socio-economic development, but have far fewer resources than the average individual or household has at their disposal and are still at a lower standard of living in comparison.

The differences between absolute and relative poverty are illustrated in figure 1 [6], where the world population may be divided into four nonoverlapping groups: (I) under absolute poverty, (II) under both absolute and relative poverty, (III) under relative poverty, and (IV) no poverty. The constitution of the poor is shown schematically in figure 2 (World Bank, 2017). Absolutely poor in the

developing world locates in the area I, with high-income countries are assumed to be no absolute poverty. Relative poverty locates in (II) in developing countries and locates in (III) in high-income countries.

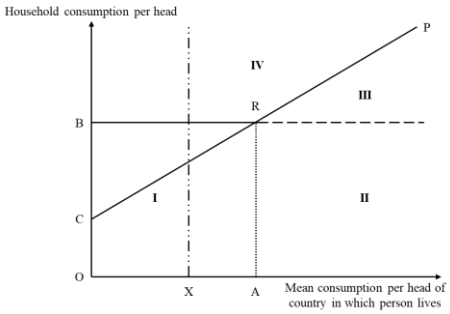

Fig. 1. Absolute and relative poverty

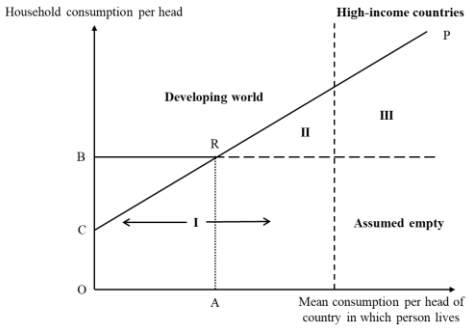

Fig. 2. Global poverty and the developing world

The connotation of poverty is rich and complex, it is dynamic and historical; relative and multidimensional. The phenomenon of urban poverty in contemporary Western societies is more of a relative poverty ("poverty in affluence"). For most developing countries, the main objective of their poverty alleviation policies is to reduce the level of absolute and extreme poverty by meeting the basic needs of the general population, after which "poverty" becomes a "relative" concept. In post-2020 China, where the phase of absolute poverty is history, relative poverty and multidimensional poverty deserves more exploration, the war against poverty will never end, and even if the basic problem of absolute poverty is solved, the problem of relative poverty will still be tough for both developing and developed countries.

For the measurement of relative poverty line in cities, the internationally prevalent standard is the international poverty line standard (also known as the income proportional method), which takes 50%-60% of the average monthly income of the middle social household in a country or region as the poverty line for that country or region. Gottlieb and Fruman [7] evaluated different relative poverty lines and the results showed that using 60% of the median is the best, which is adopted by the European Commission in practice. Therefore, the subsequent study in this paper will draw on the international mainstream approach.

2.2 Poverty identification

Poverty identification is a prerequisite to all poverty alleviation work. However, in practice, social assistance is difficult to balance efficiency and equity and can suffer from identification bias, there were still sizable targeting errors which allocated the limited resources to the registered non-poor households but excluded non-registered poor households from receiving needed government assistance. Gao et al. [8] found that the mis-targeting rate of the urban grassroots poverty alleviation program "low-benefit" in China has increased, the program failed to lift the target population out of poverty, and the severe poverty rate,

poverty gap and poverty severity still exist. Moreover, the study by Zhu and Li [9] not only provides consistent evidence, but also finds that the simple division of poor groups by income leads to serious disparities in the identification of low income households.

It has been proven that most developing countries, represented by China, have serious biases in their poverty alleviation practices, and the reason for this is, first of all, the obvious interference of subjective factors in identifying poverty based on a single dimensional indicator - income, which can be analyzed from two aspects: criteria definition and resource allocation process. Firstly, in terms of identification criteria, in most developing countries, the criteria for identifying poor households are based on income, but income cannot be accurately measured, and the poor receive unstable income from unstable agricultural production or informal sector employment, making it difficult to collect accurate income and wealth data [10]. Then, in terms of the transmission process of poverty alleviation resources, resources are distributed at each level, subject to the difficulty of defining income criteria to accurately identify the poor, resulting in poverty measured by income criteria has poor targeting performance, and multidimensional poverty-based targeting performance is better than income-based poverty targeting performance. Therefore, this paper will introduce the idea of multidimensional poverty in the subsequent study to explore the solution to the problem. The problem of poverty identification bias is not only attributed to the single identification criteria and subjective factors interfering in the practical implementation. Due to the absence of a monitoring system, the power-seeking behavior of agents is also an important reason for the low accuracy of poverty identification, which is directly manifested in the widespread elite capture phenomenon of elite misuse of designated poverty alleviation resources [11,12]. Therefore, this paper will focus on the technical aspects of the factors in a problem-solving orientation in the subsequent study to explore solutions to the problem of identification bias through the innovation of identification technology.

2.3 Relative poverty indicator system and identification techniques

The above-mentioned problem of poverty identification bias caused by a single poverty criterion may be solved by a MPI system. Since Sen [13] proposed that the identification of poverty should include education, housing, health and other aspects closely related to people's livelihoods, the academic research on poverty has gradually shifted from a single perspective to a multidimensional perspective that includes income, education, health, nutrition, resources, environment and location. After the idea of "capability poverty" was proposed, how to capture or measure this multidimensional capability poverty has become the focus of scholars' attention. Foster [14], Alkire and Foster [15], Alkire et al. [16] give a general approach to poverty identification from a multidimensional perspective: the AF approach. Alkire constructed the MPI criteria, from 3

dimensions of health, education, and living standards, including: nutrition, child mortality, adult education years, child enrollment rate, cooking fuel, toilets, drinking water, electricity, indoor floors, and assets. Compared with traditional methods, the identification of poverty is no longer just measuring income or expenditure, and the multidimensional approach can measure the multilevel and diverse needs of poor areas.

Many scholars made further research in multidimensional methods in the measurement of poverty. In terms of research methods, most of the more mature international measures are based on the MPI method. In constructing the measurement system, scholars in various countries will give a MPI system that is relevant to the time and place in the context of the situation of global poverty or the situation of poverty in a particular country [17-19]. In terms of data sources, most use micro-data to identify and analyze multidimensional poverty groups. Therefore, drawing on the existing research results, this paper chooses to use the China Family Panel Studies (CFPS 2018) as the dataset, to construct a multidimensional poverty identification index system for Chinese urban households based on MPI.

As for the problem of poverty identification bias brought by the deficiencies of the system, mechanism itself and identification technology in poverty alleviation practice, with the development and the continuous innovation of research methods, statistical analysis and machine learning technologies have been rapidly introduced into the study of poverty [20-22], providing new poverty identification solutions, including regional targeting, type targeting and individual targeting, which have improved the identification accuracy to a certain extent. Regarding regional targeting, some scholars have combined geographic information system (GIS) and remote sensing technology with machine learning techniques to comprehensively assess and analyze the extent and distribution characteristics of regional poverty [23-25]. Regional targeting brings new research perspectives and ideas, but due to the research techniques and data, the fineness of identification can generally only reach the county level and cannot be accurate to households, which has certain limitations. Regarding type targeting, allocating poverty alleviation or relief resources to focus on or even only to members of a specific type of society not only has serious inclusionary and exclusionary errors, but also has equity problems. Regarding individual targeting, technical challenges require identification techniques with accurate information and are free from the limitations of income and expenditure measurement techniques, and some scholars have already applied identification models to the process of poor household identification. Most of the poor household identification models in China still use logistic regression to build models, such as Wu et al. [26], but such regression models based on assumptions have a higher possibility of error, which can lead to less accurate identification. There are some studies suggests that RF models are superior to other models. Such as Sohnesen and Stender [27] found it is more accurate, Liang and Wang [28] found

that RF has the lowest error rate, Zhao et al. [29] found it has a good generalization ability. What's more, the importance of predictors that explain the variation of poverty can be measured [30].

Combined with the above analysis, achieving individual targeting should be the optimal goal of poverty identification, which is also the title of the targeted poverty alleviation policy. And machine learning models have more excellent properties and have shown better results compared with traditional statistical methods. Therefore, this paper will use machine learning models and aims to achieve effective individual targeting. In the subsequent research of this paper, the above factors will be fully considered to explore an effective identification scheme.

3. Construction of relative poverty identification index system

3.1 Data sources

We adopt the data of the fifth national survey of CFPS (2018) in our research, and the data of Chinese household subgroups provided by the urban part of this database are selected. The CFPS, implemented by Peking University, is the most up-to-date micro-survey data information currently available in China, allowing timely grasp and analysis of residents' living conditions, with questionnaires covering household economic, pediatric, and adult sections, covering both urban and rural household and demographic information. Besides, this paper also uses relevant data from the China's Bureau of Statistics.

3.2 Data pre-processing

The questionnaire of CFPS survey contains individual questionnaire, household questionnaire and community questionnaire, and the research on urban household poverty in this paper needs to collect data from both household and individual questionnaires. The CFPS 2018 database contains a total of 14,241 household data, and based on the urban and rural classification indicators of the National Bureau of Statistics, we filtered and removed the data classified as rural and missing, and only kept the data classified as urban. For the missing data, this paper adds the data according to the plural of urban household data in the questionnaire. In addition, for the data that cannot be filled and the outlier data (such as the outliers of age, education, etc.), this paper adopts the deletion process, and the data of the whole household is deleted, and the data of 6,548 households are finally obtained.

Since the methods used here belong to supervised machine learning, in the process of modeling, the data set needs to be divided, with part of the data used for training the model, part of the data used for prediction, and the other part used for validation. A random division of the collated sample data is chosen in the ratio of 5:3:2. Here, based on the research results of Gottlieb and Fruman (2011) and the

actual practice in the EU, we choose to define relative poverty line as 60% of the median income of urban households in China.

Table 1

Data subsets

Data Set	Relatively poor	Non-poor	Total
Training Set	938	2336	3274
Test Set	553	1411	1964
Verification Set	379	931	1310

As seen from Table 1, the data between non-poor and relatively poor households are imbalanced data. For imbalanced data sets, how to classify the samples from the minority class correctly is of utmost importance [31]. Most algorithms are sensitive to the nature of the datasets, as well as different calibrations which can lead to large differences in performance, accuracy or false positives [32], so the data should be processed for data balancing when training the model. In this paper, the Borderline-SMOTE method [33] is used to expand the non-poor data before training the model, so that the data can reach equilibrium before training the model. In addition, the categorical variables such as whether owning a car are treated as dummy variables, and the absence of cars is used as a control, i.e., households with cars are set to 1.

3.3 Construction of the indicators system

Compared with urban poverty, rural multidimensional poverty indicator system is more complete and mature, so it has important reference significance for the construction of urban poverty identification index system. The most typical MPI system selects 10 indicators from three dimensions to measure rural poverty: health (child mortality, nutritional status), education (educational attainment, child enrollment), and living standards (electricity, drinking water, indoor space area, fuel for daily use, consumer durables, sanitation). Therefore, this paper selects indicators to construct urban household poverty identification index system based on MPI indicators, taking into account the national conditions and situation, and the characteristics of urban households of China.

(1) Demographic characteristics

Demographic characteristics mainly reflect household size and the composition of household members, and this paper adds demographic characteristics dimensions to the indicator system, mainly including two indicators of household size and age structure. The household age structure refers to the proportion of each age group, and the age division nodes are 18 and 65, mainly because 18 is the legal age of adulthood in China and 65 is the upper limit of working age in China.

(2) Education

A large amount of literature on poverty research show that the education dimension is an indispensable dimension in the study of poverty, and children's school attendance and educational attainment are the two education dimension in MPI. In our study, the child enrollment indicator is dropped because of the introduction of nine-year compulsory education in China, and the family members are divided into junior high school and below, high school, undergraduate, and master's degree and above according to the current education level.

(3) Basic living security

The indicators on the life dimension in MPI include drinking water, electricity and fuel for daily life. In this paper, drinking water, electricity and fuel are used as indicators of the basic living security dimension, among which, the indicator of drinking water is reflected by water cost, the indicator of electricity consumption is reflected by electricity cost, and the indicator of fuel is reflected by the sum of natural gas cost or gas cost or other fuel cost.

(4) Living conditions

The MPI indicators in the living dimension mention three indicators which are environmental health, indoor space area and consumer durables. Since the data on the degree of tidiness of home (fhz3) and the area of current housing (fq801) are seriously missing in the CFPS 2018 database, they are not included in this paper when constructing the indicator system. Regarding the indicator of consumer durables, this paper directly selects the indicator of the total value of consumer durables (fs6v) in the CFPS 2018 database, which can directly reflect a household's ownership of consumer durables.

To make up for the shortage of living condition indicators, this research added three indicators, namely Engel coefficient, health care expenditure, and car ownership, which are related to urban household poverty.

Based on the above analysis, combined with the CFPS 2018 database, this paper takes four indicators, namely the total value of household durables (fs6v), Engel's coefficient obtained from the proportion of total food expenditure (food) in total household expenditure (expense), household health care expenditure (med) and car ownership (fp5070), as indicators of the living standard dimension of urban household poverty identification. In summary, the indicator system constructed in this paper is shown in Table 2.

Table 2

Indicator system of relative poverty identification of urban households in China

Dimensions	Secondary index	Index code	Index description
Demographic characteristics	Family size	size	Number of family members
	Age structure	nl	Proportion of family members by age
Education	Education level	wh	Proportion of family members by educational level
	Drinking water	sf	Water charges

Basic living security	Electricity Fuel	df rlf	Electricity Fuel cost
Living conditions	Consumer durables	nyp	The total value of products with unit price above 1000 yuan and natural service life above 2 years
	Engel coefficient	en	Total food expenditure as a proportion of total household consumption expenditure
	Health care expenditure	med	Household health care expenditure for the year
	Car ownership	car	Whether the family owns a car

3.4 Model variables

3.4.1 Setting of explanatory variables

The explanatory variable (Y) of this paper is whether urban households are relatively poor or not, and 60% of the average median income of urban households in China is taken as the standard, and the average income of each household is compared with it, and those below the relative poverty standard are regarded as poor households, Y=1; otherwise, they are non-poor households, Y=0. Finally, we get 1,595 poor households families and 4,456 non-poor households families.

3.4.2 Setting of explanatory variables

According to the poverty identification index system of Chinese urban households based on CFPS (2018) data constructed in Table 2, the explanatory variables used in this paper are divided into four dimensions of demographic characteristics, education, basic livelihood security, and living conditions, with a total of 14 variables, among which there are 13 numerical variables and 1 sub-type variable. The definitions of each variable are shown in Table 3 below.

Table 3

Setting of explanatory variables

Dimension	Index Name	Variable Name	Variable Description	Variable Type
Demographic characteristics	Family size	X_1	Number of family members	Value
	Age structure	X_2	Population under 18 years old/number of family members	Value
		X_3	Population aged 18-65/number of family members	Value
		X_4	Population over 65 years old/number of family members	Value
Education	Education level	X_5	Number of people in junior high school and below/number of family members	Value
		X_6	Number of high school population/number of family members	Value
		X_7	Population of junior college and above/population of family members	Value
	Drinking water	X_8	Water charges	Value

Basic living security	Electricity	X_9	Electricity	Value
	Fuel	X_{10}	Fuel cost	Value
	Consumer durables	X_{11}	The total value of products with unit price above 1000 yuan and natural service life above 2 years	Value
Living conditions	Engel coefficient	X_{12}	The proportion of total food expenditure to total household expenditure	Value
	Health care expenditure	X_{13}	Household health care expenditure for the year	Value
	Car ownership	X_{14}	Whether the family owns a car	Category

3.5 Feature selection

Feature selection is a method of selecting features from all features of the original dataset that have a greater impact on the study of that dataset and reducing the dimensionality of the data, which optimizes the system for specific metrics, can improve the generalization ability of the learning algorithm, and is a particularly critical part of the data preprocessing process. We choose to use the RF algorithm to calculate the feature importance of each explanatory variable, and use 50% of the mean value of the feature importance as the selection threshold to eliminate the relatively unimportant variables, so as to obtain a new feature set.

The 14 explanatory variables used in this paper constitute the original set of features, and the feature selection results in the retention of all explanatory variables. Table 4 shows the basic descriptive statistical analysis of the variables used in models.

Table 4

Results of descriptive statistics of variables

Variable	Observed value	Minimum	Maximum	Average	Standard deviation
X_1	6548	1	17	3.50	1.84
X_2	6548	0	100	15.87	18.66
X_3	6548	0	100	68.61	30.09
X_4	6548	0	100	15.52	29.35
X_5	6548	0	100	64.69	34.88
X_6	6548	0	100	19.14	27.03
X_7	6548	0	100	16.17	31.27
X_8	6438	0	1000	39.73	49.36
X_9	6454	0	5000	137.77	176.60
X_{10}	6476	0	4000	86.35	126.32

X_{11}	6373	0	500	6.56	14.76
X_{12}	5835	0	89	32.32	16.93
X_{13}	6490	0	389000	6699.69	17272.29
X_{14}	6548	0	1	0.38	0.49

4. Model building and testing

Classification and clustering are important Data Mining methods that could group objects into groups. Clustering is a common unsupervised learning algorithm, which can group samples in a given dataset according to the nature of the feature, for example, density-based aggregation clustering [34] or DBScan with similarity join [35] are flexible clustering algorithm which could produce accurate results. However, there is no label in this process. For the specific problem to be solved in this paper, the difficulty of using clustering lies in how to interpret the labels of the output clusters. And the desired household classification is known in advance. The RF model and LR model can divide the sample individuals into four categories through the confusion matrix, which has good explanatory and practical significance. Therefore, in the following research, we choose to use classification methods instead of clustering methods.

4.1 Establishment and verification of RF model

RF algorithm, proposed by Breiman [36], is a supervised learning method that combines multiple classification trees or regression trees and classifies or regresses them. Fernández-Delgado et al. [37] found the classification performance of RF algorithm is the best after conducting a large number of experiments to compare the performance of classification algorithms. The algorithm does not need to adjust too many parameters, can quickly deal with large sample data, rarely occurs overfitting, has fast regression (classification) speed, has good anti-noise ability and can evaluate the importance of characteristic variables in the model and other excellent characteristics, and is thus widely used in prediction and classification in many fields such as biology, finance, economics, etc.

The establishment of RF model is to use the Random Forest Classifier algorithm package to establish a RF model. After importing the algorithm package, the model should be trained and adjusted. The training set should be used for model training, and the test set should be used for parameter adjustment to obtain the optimal parameters, and then the obtained optimal parameters predict and verify the accuracy of the model on the verification set. The algorithms for the parameters is Grid Search. The optimal parameters of the random forest model are as follows: the estimators is 51. The type of estimators used is CART and the type of splitting heuristic used is Gini Index. After training the model, the parameter adjustment result is that the depth of the optimally selected tree is 9, and when the sample is

less than 59, the sample is no longer pruned. We use the optimal parameters to predict the verification set to obtain the confusion matrix as shown in Table 5, and the resulting ROC curve is shown in Figure 3.

Table 5

Confusion matrix of Random Forest verification set

True value / Predicted value	Non-relative poverty	Relative poverty	Total
Non-relative poverty	794	137	931
Relative poverty	125	254	379
Total	919	391	1310

The time to build of RF model is 0.3151 seconds, demonstrating the excellent running speed of RF model. The confusion matrix shows that the RF model has a good overall accuracy of (794+254)/1310=80%. It can also be seen that among the 931 non-relatively poor households 794 households were predicted to be non-relatively poor and among the 379 relatively poor households, a total of 254 households were predicted to be relatively poor. The recall of the model was 85.28%, indicating that 85.28% of the non-poor households delineated by the relative poverty line were accurately predicted as non-poor. The model specificity of 67.02% indicates that among the relatively poor households delineated by 60% of the median household income, 67.02% of the 379 households in the validation set were accurately predicted to be relatively poor. As shown in Figure 2, the AUC value obtained from the RF model is 85.23%, suggesting that the overall performance of the model is accurate and has scientific validity.

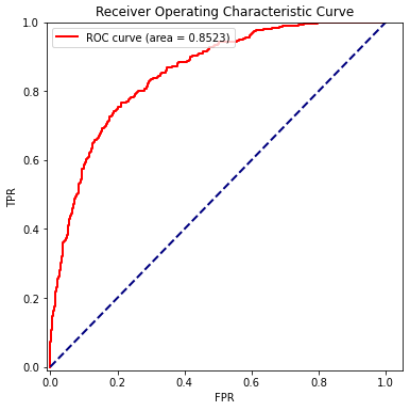

Fig. 3. ROC curve of Random Forest model

4.2 Establishment and Testing of the LR model

LR models are commonly used in the specific analysis of poverty-causing factors. For LR model building the most important is to build the model using the Logistic Regression. After that, the model is trained on the training set to get the

best parameters, and the obtained cases of each parameter are compared with the RF model, in order to verify the accuracy of the LR model by using the same data set for prediction when the model is predicted. Thus, the prediction of the LR model using the validation set was performed and the confusion matrix of the validation set and the overall ROC curve and AUC values of the model were obtained.

The regression basics are as follows: Table 6 shows the parameters of each variable obtained and Table 7 shows the confusion matrix obtained based on the test set, and Figure 2 shows the ROC curves obtained based on the model case.

Table 6

Parameters of each variable

Variable	Parameter	Variable	Parameter
X_1	0.7936	X_8	-0.3598
X_2	0.2252	X_9	-0.1518
X_3	-0.0985	X_{10}	0.0532
X_4	-0.0483	X_{11}	-0.8377
X_5	0.3911	X_{12}	-0.1619
X_6	-0.0854	X_{13}	-0.1254
X_7	-0.3768	X_{14}	-0.6917

Based on the parameters of each variable, it can be seen that among the main influences on relative poverty of urban households, (household size), (consumer durables) and (whether or not the household owns a car) have a significant effect on whether the household is in relative poverty.

Table 7

Confusion matrix for the LR validation set

True value / Predicted value	Non-relative poverty	Relative poverty	Total
Non-relative poverty	694	237	931
Relative poverty	86	293	379
Total	780	530	1310

The confusion matrix in Table 7 shows that the LR model has a overall accuracy of (694+293)/1310 = 75.34%. It can also be seen that 237 of the 931 relatively non-poor households are predicted to be relatively poor and 694 are predicted to be non-poor. Among the 379 relatively poor households, 293 were predicted to be relatively poor, and 86 were predicted to be non-poor. In addition, the completeness (recall) of the model was 74.54%, indicating that 74.54% of the non-poor households delineated by the relative poverty line were accurately

predicted as non-poor. The model specificity of 77.31% indicates that among the relatively poor households delineated by the line, 77.31% of the 379 households in the validation set were accurately predicted by the model to be relatively poor. As shown in Figure 3, the AUC value obtained based on the LR model is 83.94%, indicating that the overall performance of the LR model is also pretty well and has predictive ability.

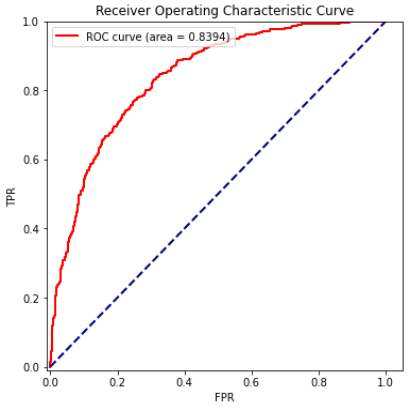

Fig. 4. ROC curve of LR model

4.3 Results

For the identification problem of relative poverty, the LR model and RF model were established, and the confusion matrix under different models was obtained by using the model to verify the prediction for urban household data, and the evaluation indexes can reflect the classification performance. In the following, the prediction ability of the models is firstly discussed in comparison with each type of judgment indicator.

Table 8

Comparison of model prediction ability

Index	RF	LR
Overall accuracy	80%	75.34%
Recall rate	85.28%	74.54%
Specificity	67.02%	77.31%
AUC value	85.23%	83.94%
F1 value	85.84%	81.12%

In the specific problem of this paper, the overall accuracy indicates the probability of identifying the true relatively poor and non-relatively poor households accurately; the recall indicates the probability of identifying the true non-relatively poor households accurately; the specificity represents the probability of identifying the true relatively poor households accurately; the AUC value indicates that given a randomly selected true relatively poor household and a true non-relatively poor household, the probability value of the classifier outputting a

real relatively poor household as a relatively poor household is more likely than that of the classifier outputting a real non-relatively poor household as a relatively poor household, and this indicator can well reflect the performance of the model; the F1 value reflects the classification effect of the RF classifier for relative poverty identification.

According to Table 8, the RF model outperforms the LR model in terms of overall accuracy, recall, AUC value, and F1 value, but the LR model outperforms the RF model in terms of specificity. Since the recall rate of the RF is better than that of the LR, the RF model can be considered for non-relative poverty prediction. As the specificity of the LR model is higher than that of the RF model, the LR model can be considered for screening relatively poor samples. In terms of the AUC values, the RF model is slightly better than the LR model. In general, the RF model is better than the LR model.

In order to better measure the predictive ability and importance of each indicator variable, the IV (Information Value) value of each variable calculated by the RF model in feature engineering is invoked, as shown in Figure 4. The role of the IV value is to measure the overall predictive ability of a variable, through which it is possible to compare the information value contribution of each variable, i.e., the contribution of information of a variable to determining whether a household is relatively poor or non-relatively poor, the greater the contribution, the greater the IV value. From the figure, it can be seen that indicators such as education level, household size, whether or not to own a car, and household age structure are specifically significant and important information value contributions for predicting household relative poverty.

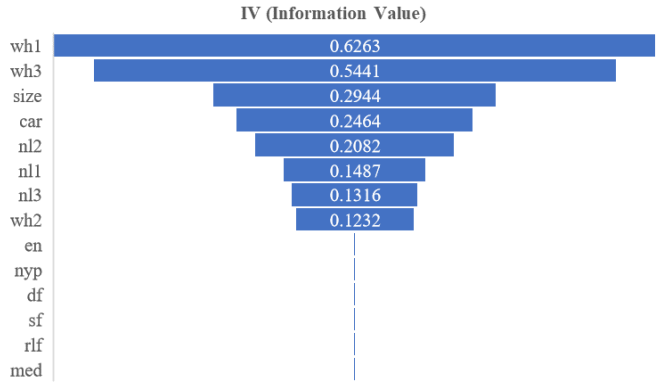

Fig. 5. IV value of each index variable

MPI contains three dimensions of health, education, and living standard, which have been widely used. However, we added the dimension of demographic characteristics (household size and household age structure) in the construction of the relative poverty identification indicator system. From the IV values of the indicators in the prediction model, the demographic characteristics indicators

contribute significant information value in predicting relative poverty, which is consistent with the assumptions made in the construction of the indicator system. Demographic characteristics are closely related to poverty, and this factor should be fully taken into account in future poverty identification studies. In addition, this paper does not include the income factor in the poverty identification index system, consistent with the MPI, to avoid the identification bias caused by the difficulty of defining income and the problem of a single indicator, but focuses more on other indicators, which to a certain extent also brings a solution to the identification problem of critically poor households (a category of people whose net per capita income fluctuates around the national poverty line).

5. Conclusions

Absolute poverty will be history one day, yet relative poverty is becoming increasingly serious, and there is less attention and corresponding research from all sectors at present. In the current practice of poverty alleviation, there are obvious identification biases. Achieving individual targeting should be the optimal goal of poverty identification. Therefore, we propose a new perspective to explore the identification of relative poverty of urban households by using machine learning technology, which provides a solution to the problem of difficult and time-consuming poverty identification and serious identification bias in practical implementation. This paper firstly adds the dimension of demographic characteristics and constructs a multidimensional poverty identification index system, considering the national conditions of developing countries and the characteristics of urban households. Based on the objective of targeting the relative poor population in urban areas, machine learning models with RF and LR are built using CFPS 2018 data. These two models verify the importance of the demographic characteristics dimension (household size and household age structure) for poverty identification, and the evaluation indexes show that the RF model has a better performance of identification. In the process of model parameter optimization, using a 5:3:2 ratio to randomly divide the sample data into training set, test set and verification set, and the final optimized model is proved to have a better prediction performance. There are still some limitations in this study which may provide directions for future research, such as the construction of the MPI may be limited by the data, and the index system still needs to be improved. In the future shift from absolute poverty management to relative poverty management, information resources and machine learning technology should be fully utilized, and a networked platform for poverty population prediction and discrimination could be built.

Acknowledgements

This research was funded by National Natural Science Foundation of China [Grant 72072118].

R E F E R E N C E S

[1]. *C. Lu,* "Who is Poor in China? A Comparison of Alternative Approaches in Rural Yunnan", in The Journal of Peasant Studies**, vol. 37**, no. 2, 2010, pp. 407-428.
[2]. *Y. Zhou, Y. Guo, Y. Liu, etc.*, "Targeted Poverty Alleviation and Land Policy Innovation: Some Practice and Policy Implications from China", in Land Use Policy, **vol. 74**, 2018, pp. 53-65.
[3]. *B. Gustafsson and D. Sai,* "Growing into Relative Income Poverty: Urban China, 1988–2013", in Social Indicators Research, **vol. 147**, no.1, 2020, pp. 73-94.
[4]. *L. Xu, X. Deng, Q. Jiang*, etc., "Identification and Alleviation Pathways of Multidimensional Poverty and Relative Poverty in Counties of China", in Journal of Geographical Sciences, **vol**. 31, no. 12, 2021, pp. 1715-1736.
[5]. *Y. Liu, J. Liu and Y. Zhou,* "Spatio-Temporal Patterns of Rural Poverty in China and Targeted Poverty Alleviation Strategies", in Journal of Rural Studies, **vol. 52**, 2017, pp. 66-75.
[6]. *World Bank,* Monitoring Global Poverty: Report of the Commission on Global Poverty, World Bank, 2017.
[7]. *D. Gottlieb and A. Fruman,* "A Quality-index of Poverty Measures", ECINEQ working, 2011.
[8]. *Q. Gao, S. Yang and S. Li,* "Welfare, Targeting, and Anti-poverty Effectiveness: The Case of Urban China", in The Quarterly Review of Economics and Finance, **vol. 56**, 2015, pp. 30-42.
[9]. *M. Zhu and S. Li,* "The Key to Precise Poverty Alleviation Rests in the Precise Indentification of Impoverished Populations-An Analysis of the Targeting Effectiveness of China's Rural Dibao Program", in Social Sciences in China, **vol. 40**, no. 2, 2019, pp. 60-76.
[10]. *S. R. Tabor,* "Assisting the Poor with Cash: Design and Implementation of Social Transfer Programs", in World Bank Social Protection Discussion Paper, **vol. 223**, 2002, pp. 79-97.
[11]. *M. Li and R. Walker,* "Targeting Social Assistance: Dibao and Institutional Alienation in Rural China", in Social Policy & Administration, **vol. 52**, no. 3, 2018, pp. 771-789.
[12]. *Y. Yang, J. Chen and M. Jin,* "Who are the Asset-poor in China: A Comprehensive Description and Policy Implications", in Journal of Social Policy, **vol. 48**, no. 4, 2019, pp. 1-23.
[13]. *A. Sen,* "Poverty: An Ordinal Approach to Measurement", in Econometrica: Journal of the Econometric Society, **vol. 44**, no. 2, 1976, pp. 219-231.
[14]. *A. J. Foster,* "Counting and Multidimensional Poverty Measurement", in Journal of Public Economics, **vol. 95**, no. 7, 2007, pp. 476-487.
[15]. *S. Alkire and J. Foster,* "Understandings and Misunderstandings of Multidimensional Poverty Measurement", in Journal of Economic Inequality, **vol. 9**, no. 2, 2011, pp. 289-314.
[16]. *S. Alkire, J. M. Roche and A. Vaz,* "Changes over Time in Multidimensional Poverty: Methodology and Results for 34 Countries", in World Development, **vol. 94**, 2017, pp. 232-249.
[17]. *P. A. Ervin, L. Gayoso de Ervin, J. R. Molinas Vega, etc.*, "Multidimensional Poverty in Paraguay: Trends from 2000 to 2015", in Social Indicators Research, **vol. 140**, no. 3, 2018, pp. 1035-1076.
[18]. *K. Decancq, M. Fleurbaey and F. Maniquet,* "Multidimensional Poverty Measurement with Individual Preferences", in The Journal of Economic Inequality, **vol. 17**, no. 1, 2019, pp. 29-49.
[19]. *S. Alkire and Y. Fang,* "Dynamics of Multidimensional Poverty and Uni-dimensional Income Poverty: An Evidence of Stability Analysis from China", in Social Indicators Research, **vol. 142**, no. 1, 2019, pp. 25-64.

[20]. *A. Alsharkawi, M. Al-Fetyani, M. Dawas, etc.,* "Improved Poverty Tracking and Targeting in Jordan Using Feature Selection and Machine Learning", in IEEE Access, **vol. 10**, 2022, pp. 86483-86497.
[21]. *M. Gao, L. Li and Y. Gao,* "Statistics and Analysis of Targeted Poverty Alleviation Information Integrated with Big Data Mining Algorithm", in Security and Communication Networks, 2022.
[22]. *F. Wu, Q. Zheng, F. Tian, etc.,* "Supporting Poverty-stricken College Students in Smart Campus", in Future Generation Computer Systems, **vol. 111**, 2020, pp. 599-616.
[23]. *Y. Wang and L. Qian,* "A PPI-MVM Model for Identifying Poverty-Stricken Villages: A Case Study from Qianjiang District in Chongqing, China", in Social Indicators Research, **vol. 130**, no. 2, 2017, pp. 497-522.
[24]. *G. Li, Z. Cai, X. Liu, etc.,* "A Comparison of Machine Learning Approaches for Identifying High-Poverty Counties: Robust Features of DMSP/OLS Night-Time Light Imagery", in International Journal of Remote Sensing, **vol. 40**, no. 15, 2019, pp. 5716-5736.
[25]. *J. Yin, Y. Qiu and B. Zhang,* "Identification of Poverty Areas by Remote Sensing and Machine Learning: A Case Study in Guizhou, Southwest China", in ISPRS International Journal of Geo-information, **vol. 10**, no. 1, 2020, pp. 11.
[26]. *Y. X. Wu, Z. N. Hu, Y. Y. Wang, etc.,* "Rare Potential Poor Household Identification with a Focus Embedded Logistic Regression", in IEEE Access, **vol. 10**, 2022, pp. 32954-32972.
[27]. *T. P. Sohnesen and N. Stender,* "Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment", in Poverty & Public Policy, **vol. 9**, no. 1, 2017, pp. 118-133.
[28]. *T. Liang and X. Wang,* "A Statistical Analysis Model of Big Data for Precise Poverty Alleviation Based on Multisource Data Fusion", in Scientific Programming, **vol. 2022,** Article ID 5298988, 2022.
[29]. *X. Zhao, B. Yu, Y. Liu, etc.,* "Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh", in Remote Sensing, **vol. 11**, no. 4, 2019, pp. 375.
[30]. *C. Kllestl, E. Blandón Zelaya, R. Peña, etc.,* "Predicting Poverty. Data Mining Approaches to the Health and Demographic Surveillance System in Cuatro Santos, Nicaragua", in International Journal for Equity in Health, **vol. 18**, no. 1, 2019, pp. 1-12.
[31]. *F. Li, X. Zhang, X. Zhang, etc.,* "Cost-Sensitive and Hybrid-Attribute Measure Multi-Decision Tree Over Imbalanced Data Sets", in Information Sciences, **vol. 422**, 2018, pp. 242-256.
[32]. *C. O. Truica and C. A. Leordeanu,* "Classification of An Imbalanced Data Set Using Decision Tree Algorithms", U. P. B. Sci. Bull., Series C, **vol. 79**, no. 4, 2017, pp.69-84.
[33]. *H. Han, W. Y. Wang and B. H. Mao,* "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", in International Conference on Intelligent Computing, 2005, pp. 878-887. Springer, Berlin, Heidelberg.
[34]. *I. M. Rădulescu, A. Boicea, C. O. Truică, etc.,* "DenLAC: Density Levels Aggregation Clustering-A Flexible Clustering Method", In Computational Science-ICCS: 21st International Conference, Krakow, Poland, Proceedings, Part I, 2021, pp. 316-329. Cham: Springer International Publishing.
[35]. *I. M. Rădulescu, C. O. Truică, E. S. Apostol, etc.,* "Performance Evaluation of DBSCAN With Similarity Join Algorithms", in the 34th International-Business-Information-Management-Association (IBIMA) Conference, Madrid, Spain, 2019.
[36]. *L. Breiman,* "Random forests", in Machine Learning, **vol. 45**, no.1, 2001, pp. 5-32
[37]. *M. Fernández-Delgado, E. Cernadas, S. Barro etc.,* "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?", in The Journal of Machine Learning Research, **vol. 15**, no. 1, 2014, pp. 3133-3181.