Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2024, 12, 479-494
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Exploration of the Impact Mechanism of Government Credibility Based on Variable Screening Method

Jiajun Wu, Yuxiang Ma, Helin Zou, Chun Zhang, Ran Yan

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
Email: 2021311014@email.cufe.edu.cn

How to cite this paper: Wu, J.J., Ma, Y.X., Zou, H.L., Zhang, C. and Yan, R. (2024) Exploration of the Impact Mechanism of Government Credibility Based on Variable Screening Method. *Journal of Data Analysis and Information Processing*, **12**, 479-494.
https://doi.org/10.4236/jdaip.2024.123025

Received: August 13, 2024
Accepted: August 24, 2024
Published: August 27, 2024

Abstract

Government credibility is an important asset of contemporary national governance, an important criterion for evaluating government legitimacy, and a key factor in measuring the effectiveness of government governance. In recent years, researchers' research on government credibility has mostly focused on exploring theories and mechanisms, with little empirical research on this topic. This article intends to apply variable selection models in the field of social statistics to the issue of government credibility, in order to achieve empirical research on government credibility and explore its core influencing factors from a statistical perspective. Specifically, this article intends to use four regression-analysis-based methods and three random-forest-based methods to study the influencing factors of government credibility in various provinces in China, and compare the performance of these seven variable selection methods in different dimensions. The research results show that there are certain differences in simplicity, accuracy, and variable importance ranking among different variable selection methods, which present different importance in the study of government credibility issues. This study provides a methodological reference for variable selection models in the field of social science research, and also offers a multidimensional comparative perspective for analyzing the influencing factors of government credibility.

Keywords

Government Credibility, Variable Selection Models, Social Statistics, Regression Based Approach, Method Based on Random Forest

1. Introduction

Government credibility is an important asset in contemporary national govern-

ance, an important criterion for evaluating government legitimacy, and a key factor in measuring the effectiveness of government governance [1]-[3]. The research on government credibility abroad is deeply influenced by the social contract theory, which states that the community under the social contract should represent the will of the entire community, rather than certain individuals [4]. The level of trust (credibility) of the general public in some leaders of the community determines the stability of the community. Bernard Barber explored the importance of government credibility by studying credit responsibility and technological capability. He believes that only when the political system is thoroughly trusted can the government effectively use its power and better achieve its goals.

In recent years, government credibility has been a hot topic in academic research. Gong Huilian *et al.* [2] demonstrated through the analysis of CGSS2015 survey data that the sense of access to public services and perception of social equity had a positive impact on government credibility. Wu Xiaofeng [5] summarized the research on government credibility and proposed that the definition of government credibility can be divided into two perspectives: one perspective held that credibility is the evaluation and recognition of government behavior by the public, reflecting the level of public trust in the government. For example, Mao Shoulong *et al.* [6] believed that government credibility is manifested as the public's confidence in the government. Another perspective viewed the government as the subject and the public as the object, and regarded government credibility as the ability and degree to gain public trust, believing that it is an authoritative resource possessed by the government. It can be seen that in-depth exploration of the influencing factors of government credibility has important theoretical value for systematically analyzing the process of government governance.

Up to now, there is relatively little literature exploring government credibility from the perspective of residents, and often lacks in-depth and comprehensive analysis. Most of them focus on theoretical and mechanism exploration, and empirical research is usually relatively simple, such as using only relevant analysis [7].

This article intends to use seven different variable selection methods in regression analysis and machine learning (including four regression based methods: stepwise backward selection using p-values and AIC, Lasso variable selection, elastic network variable selection, and three random forest based methods: variable selection using random forest, regularized random forest, Boruta, and gradient boosting feature selection) to conduct a comparative study on the empirical problem of government credibility, in order to achieve the best variable selection effect. This study aims to apply the comparative analysis of variable selection models in the field of social statistics to government credibility, achieve empirical research on government credibility and explore its core influencing factors from a statistical perspective.

2. Related Theoretical Foundations

In this section, a brief introduction is given to the 7 variable selection methods

and related knowledge that will be used in the following text. It mainly includes four regression-based methods and three random forest-based methods.

2.1. Regression Based Approach

1) Stepwise backward selection method using p-value

The stepwise backward selection method using p-values is a classic and widely used variable selection method. This method first incorporates all variables into the regression model, and then simplifies the model by gradually eliminating insignificant variables. Specifically, variables are tested one by one, and if their p-value is higher than the preset significance level, the variable will be removed. The significance level set in the model used in this article is $p < 0.05$ to screen for the most meaningful variables.

2) Using AIC's stepwise backward selection method

Similar to the stepwise backward selection method using p-values, the stepwise backward selection method using AIC (Akaike Information Criterion) [8] is also a method of gradually eliminating insignificant variables. The difference is that its stopping rule is based on the value of AIC, not the p-value. AIC considers the complexity of the model, and as the number of selected variables increases, AIC will penalize the model. Therefore, this method tends to choose simpler but more explanatory models.

3) Lasso variable selection method

Lasso variable selection method is a regularization method based on linear regression, which selects variables by applying L1 penalty to regression coefficients. L1 punishment will reduce some regression coefficients to zero, thereby excluding the corresponding variables from the model. This process not only simplifies the model, but also improves its predictive performance. This article will use cross validation to evaluate the performance of the model.

4) Variable selection method for elastic network

Elastic networks combine L1 and L2 penalties and are an extension of Lasso variable selection method [9] [10]. The introduction of L2 penalty makes the elastic network have a grouping effect, that is, variables with high correlation are either retained or excluded from the model at the same time. Compared to Lasso, elastic networks perform better in handling high-dimensional data. This article will select the model with the minimum MSE to determine the selected variables.

2.2. Method Based on Random Forest

Random forest [11] is a composite model composed of hundreds to thousands of decision trees, whose prediction results are based on the average output of all trees. Each decision tree is built by recursively segmenting a random subset of data, and the selection and segmentation points of these subsets are based on the goal of maximizing the difference or information gain between data subsets.

1) Variable screening using random forest (VSURF)

The VSURF [12] method utilizes the variable selection mechanism built into random forests to select the simplest model that is lower than the minimum error plus its standard deviation. This method generates two subsets of variables, one for interpretation, containing all variables highly correlated with the results, and the other more concise, containing only the core variables required for prediction. In practical applications, the variables in the explanatory subset selected by the VSURF method are considered the most important variables.

2) Regularized Random Forest (RRF)

RRF is a technique based on random forests, characterized by imposing penalties on new variables when constructing each tree if their split information gain is not as good as the previous split. This method aims to select the minimum subset of variables for prediction. This article will use ten-fold cross validation to determine the optimal parameter configuration, thereby achieving regularization of the model [13]. The selected variables will then be used for performance evaluation on the validation set.

3) Boruta method

The main goal of the Boruta method [14] is to identify the variables that are most important for predicting results. Firstly, randomly shuffle all variables in the dataset to generate shadow variables, which correspond one-to-one with the original variables. Secondly, train a random forest model using data that includes both raw and shadow variables. Then, calculate the importance score of each variable and compare the score of each original variable with the highest score of all shadow variables. If the importance score of a primitive variable is significantly higher than the highest score of the shadow variable, it is considered important and the variable is selected. If it is significantly lower than the highest score of the shadow variable, it is considered unimportant and rejected. Remove rejected variables, regenerate shadow variables, and repeat the above process until all variables are clearly classified as important or unimportant.

3. Data Processing

3.1. Data Source and Introduction

The Chinese Family Panel Studies (CFPS) is a nationwide, comprehensive, long-term longitudinal survey project initiated and managed by the Chinese Social Science Survey Center (ISSS) at Peking University. The survey is conducted every two years, targeting individuals, families, and communities, and gradually covering most parts of the country, providing high-quality data on Chinese society, economy, population, education, and health. This article uses the most recent publicly available 2020 data from CFPS and selects its personal database for research.

Calculate the mean of qn6011, qn1101, and qn10025 to obtain the dependent variable of government credibility (Gov); The remaining 20 explanatory variables are to be selected, as shown in **Table 1**.

Table 1. Variable table.

Variable Type	Variable Symbol	Measuring Method
Explained Variable	*Gov*	The average sum of "government integrity", "evaluation of local county and city governments", and "trust in local officials"
Explanatory Variable	*JS* (*Job Satisfaction*)	Quantify and assign values based on five dimensions: "very dissatisfied", "not very satisfied", "average", "quite satisfied", and "very satisfied"
	SH (*State of Health*)	Quantify and assign values based on five dimensions: "very healthy", "relatively healthy", "average", and "unhealthy"
	EC (*Environmental Conservation*)	Scored by respondents from 0 to 10, where higher scores indicate more serious environmental issues
	II (*Income Inequality*)	Scored by respondents from 0 to 10, where higher scores indicate greater income inequality
	Emp (*Employment*)	Scored by respondents from 0 to 10, where higher scores indicate more serious employment issues
	Edu (*Education*)	Scored by respondents from 0 to 10, where higher scores indicate greater education issues
	Hea (*Healthcare*)	Scored by respondents from 0 to 10, where higher scores indicate more serious healthcare issues
	Hou (*Housing*)	Scored by respondents from 0 to 10, where higher scores indicate more serious housing issues
	SS (*Social Security*)	Scored by respondents from 0 to 10, where higher scores indicate more serious social security issues
	LS (*Life Satisfaction*)	Scored by respondents from 1 to 5, where 1 indicates very dissatisfied and 5 indicates very satisfied
	LIL (*Local Income Level*)	Scored by respondents from 1 to 5, where 1 indicates very low and 5 indicates very high local income level
	LSS (*Local Social Status*)	Scored by respondents from 1 to 5, where 1 indicates very low and 5 indicates very high local social status
	TA (*Trust in American*)	Scored by respondents from 0 to 10, where higher scores indicate more trust in Americans
	Age	Age of the respondent
	YE (*Years of Education*)	Determined by the length of education received by the respondent
	Gen (*Gender*)	Gender of the respondent
	CMS (*Current Marital Status*)	Current marital status of the respondent
	WWH (*Weekly Working Hours*)	Weekly working hours of the respondent, measured in hours
	DIU (*Daily Internet Usage*)	Daily internet usage of the respondent, measured in minutes per day
	AWI (*Annual Work Income*)	Annual work income of the respondent, measured in yuan per year

3.2. Data Cleaning

Due to the fact that the selected data is questionnaire data, there are many difficulties in data processing such as jumping. This article processes the initial data as follows:

1) Select data from the population aged over 18 and entering the workforce;

2) Delete data that contains missing values, is not applicable, or is missing;

3) Process the values of "don't know" and "refuse answer", replace "don't know" with the mean, and replace "refuse answer" with 0.

After the above processing, a total of 2957 data points were obtained, and their descriptive statistical results are shown in Table 2. Among them, the dependent variable is government credibility (Gov), with an average value of 5.274 and a standard deviation of 1.515. This indicates that the public's trust in the government is at a moderate level, and the evaluation of trust is somewhat scattered in the sample. Some members of the public have a high level of trust in the government, while others may have a lower level of trust in the government.

Table 2. Descriptive statistics.

Variable	sample size	average value	standard deviation	minimum value	maximum value
Gov	2957	5.274	1.515	0.667	10
JS	2957	6.46	1.953	0	10
SH	2957	5.322	2.28	0	10
EC	2957	7.026	2.231	0	10
II	2957	7.234	2.007	0	10
Emp	2957	6.523	2.014	0	10
Edu	2957	6.655	2.308	0	10
Hea	2957	6.664	2.305	0	10
Hou	2957	6.778	2.311	0	10
SS	2957	6.105	2.301	0	10
LS	2957	3.846	0.825	1	5
LIL	2957	2.859	0.794	1	5
LSS	2957	2.815	0.839	0	5
TA	2957	2.741	2.427	0	10
Age	2957	33.656	9.037	19	74
YE	2957	14.287	2.645	0	22
Gen	2957	0.576	0.494	0	1
CMS	2957	0.701	0.458	0	1
WWH	2957	47.504	14.326	0.1	150
DIU	2957	437.957	318.339	2	2880
AWI	2957	68026.235	60246.921	0	1000000

Among the explanatory variables to be selected, the mean of job satisfaction (JS) is 6.46, with a standard deviation of 1.953. A higher mean indicates that the majority of respondents are satisfied with their work. A larger standard deviation indicates significant differences in job satisfaction among the samples. The

mean of health status (SH) is 5.322, with a standard deviation of 2.28, indicating that the respondents' evaluation of their own health status is above average, but there are significant differences in health status among different individuals.

The mean values of environmental protection (EC) and wealth gap (II) are 7.026 and 7.234, respectively, with standard deviations of 2.231 and 2.007, respectively. A higher mean indicates that respondents generally believe that environmental issues and wealth inequality are more serious, and there are significant differences in their views on these issues. The mean values of employment (Emp), education (Edu), healthcare (Hea), housing (Hou), and social security (SS) are 6.523, 6.655, 6.664, 6.778, and 6.105, respectively, indicating that these socio-economic issues are considered relatively serious and have significant differences in severity among individuals.

The mean of self-life satisfaction (LS) is 3.846, with a standard deviation of 0.825, indicating that the majority of respondents are at a moderate to relatively satisfied level with their lives. The mean values of local income level (LIL) and local social status (LSS) are 2.859 and 2.815, respectively, indicating that the respondents have a low evaluation of local income level and social status, and the differences are significant. The mean value of trust in Americans (TA) is 2.741, and the standard deviation is 2.427, which shows that the respondents' trust in Americans is low, and there are large differences in trust among different individuals.

The mean age is 33.656 years, with a standard deviation of 9.037, indicating that the majority of respondents in the sample are in the young to middle-aged stage, but the age distribution is relatively wide. The mean length of education (YE) is 14.287 years, with a standard deviation of 2.645 years, reflecting that the majority of respondents have a longer education period and some differences in educational levels. The average weekly working hours (WWH) is 47.504 hours, with a standard deviation of 14.326, indicating that respondents generally have longer working hours and significant differences in working hours. The mean duration of internet use (DIU) is 437.957 minutes, with a standard deviation of 318.339, indicating that the surveyed individuals spend a considerable amount of time online each day, and there are significant differences between individuals. The average annual income (AWI) of the surveyed individuals is 68026.235 yuan, with a standard deviation of 60246.921, indicating that their overall annual income is relatively high but the differences are significant.

3.3. Data Preprocessing

Firstly, by using the scale function in R, the data is standardized to eliminate the influence caused by dimensional differences between different variables. Secondly, the dataset is randomly divided into a training set and a testing set in a 7:3 ratio. Using the create Data Artifact function of the Caret package (Hyndman *et al.*, 2018), 70% of the data is used for training and 30% for testing. This partitioning method ensures that the model can fully learn the features of the data during training, while also verifying the model's predictive ability during testing. After this processing, the random forest can be trained and tested, and

the mean square error (MSE) can be calculated to evaluate the predictive performance of the model.

4. Empirical Analysis of Government Credibility Issues Based on Different Variable Selection Methods

4.1. Analysis of Government Credibility Issues Based on Regression Method

4.1.1. Solution by the Backward Selection Method Based on P-Value and AIC

By establishing a multiple linear regression model, two judgment criteria were used, p-value and AIC, with $p < 0.05$ as the boundary or AIC no longer increasing as the boundary, gradually removing variables that did not meet the requirements. Finally, 10 variables that met the requirements were selected from 20 variables using p-value backward selection, and 13 variables that met the requirements were selected from 20 variables using AIC backward selection. The coefficient results are shown in Table 3.

Table 3. Coefficients of regression based variable selection method.

Variable	p-Value Backward Selection Result	AIC Backward Selection Result	LASOO	Elastic Net
JS	0.150	0.150	0.102	0.106
SH	0.062	0.055	0.001	0.006
EC	−0.080	−0.075	−0.038	−0.042
II	−0.127	−0.124	−0.105	−0.106
Edu	−0.163	−.0136	−0.111	−0.113
SS	−0.202	−0.185	−0.143	−0.146
LS	0.076	0.082	0.031	0.034
LSS	0.084	0.087	0.046	0.049
TA	0.112	0.109	0.047	0.053
YE	0.074	0.077	0.008	0.014
Hea	0	−0.051	−0.043	−0.045
DIU	0	−0.037	0	0
CMS	0	−0.080	0	0

4.1.2. Solution by Lasso

By solving the minimum value of the objective function to obtain the coefficients of the LASOO regression model, the objective function is as follows:

$$\beta = \arg\min\left\{\sum_{i=1}^{m}\left(Gov_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j\right)^2\right\} + \lambda\sum_{j=1}^{d}\left|\beta_j\right| \quad (1)$$

M is the sample size of the data, d is the number of variables to be selected, which is 20 in this article, and Gov_i is the i-th value of the target quantity Gov; x_{ij} is the j-th variable corresponding to the i-th influencing factor; β_j is the re-

gression coefficient for each variable.

After establishing the model, the dynamic process of selecting variables is obtained, as shown in Figure 1. It can be seen that as the hyperparameters λ increase, each parameter is also compressed smaller. When the variable parameter is compressed to 0, it means that the variable is not important and is removed from the model.

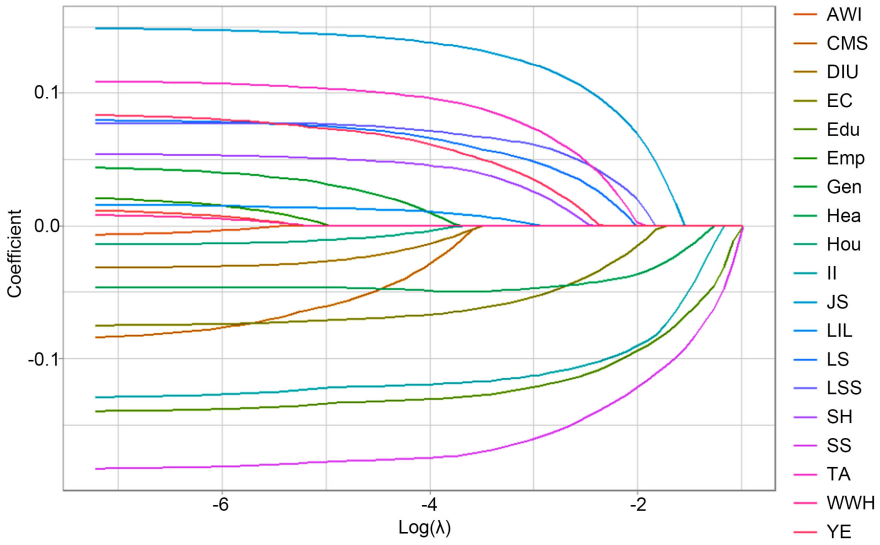

Figure 1. Regression coefficient path diagram.

To establish the LASOO model, it is also necessary to determine the values of hyperparameters λ. This article conducts a ten-fold cross validation, chooses the largest λ with MSE within one standard deviation (which is also the turning point from low to high mean square error growth rate in cross validation). The cross-validation graph is shown in Figure 2. The final selected λ value is 0.0811.

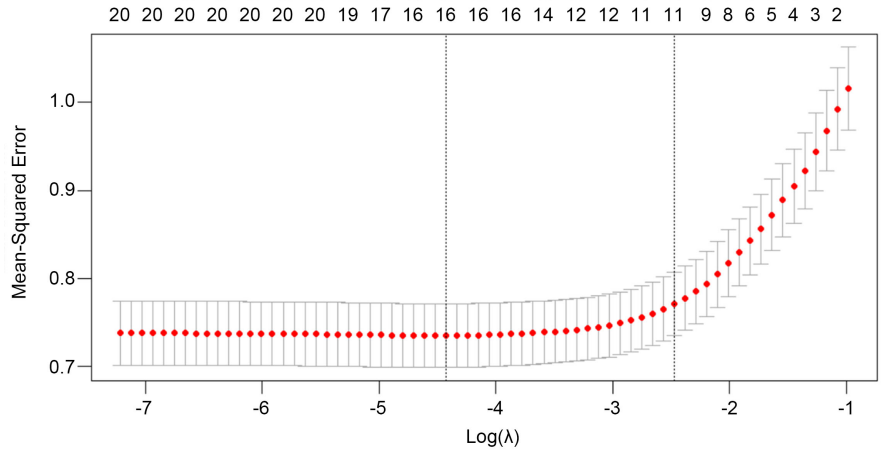

Figure 2. Cross validation diagram.

Through the LASSO model, 11 variables were ultimately selected from the 20 candidate variables. The specific results are shown in Table 3.

4.1.3. Solution by Elastic Network

Due to the characteristic of compressing variable coefficients to 0, LASSO regression performs poorly when dealing with variables with high dispersion. For this, ridge return can be used as a supplement. Elastic network combines the advantages of LASOO regression and ridge regression, the elastic network regression objective function established in this article is as follows:

$$\beta = \arg\min\left\{\sum_{i=1}^{m}\left(Gov_i - \sum_{j=1}^{d} x_{ij}\beta_j\right)^2\right\} + \lambda\sum_{j=1}^{d}\left(\alpha\left|\beta_j\right| + (1-\alpha)\beta_j^2\right) \quad (2)$$

This objective function is a convex linear combination of ridge regression and LASSO regression objective functions. When $\alpha = 0$, the elastic network becomes ridge regression, and when $\alpha = 1$, it is LASOO regression. The α value determines the degree to which the elastic network prefers LASOO regression or ridge regression. In order to obtain the optimal parameter value, this paper traverses the weight coefficient α with each change of 0.005 between (0, 1). In this process, the maximum MSE within one standard deviation in the ten-fold cross-validation is taken as the penalty coefficient λ, and finally, α that can minimize MSE of the overall model is selected as the model parameter, as shown in Figure 3. Finally, $\alpha = 0.4$ is selected, the elastic network is biased towards ridge regression, and $\lambda = 0.185$ at this time.

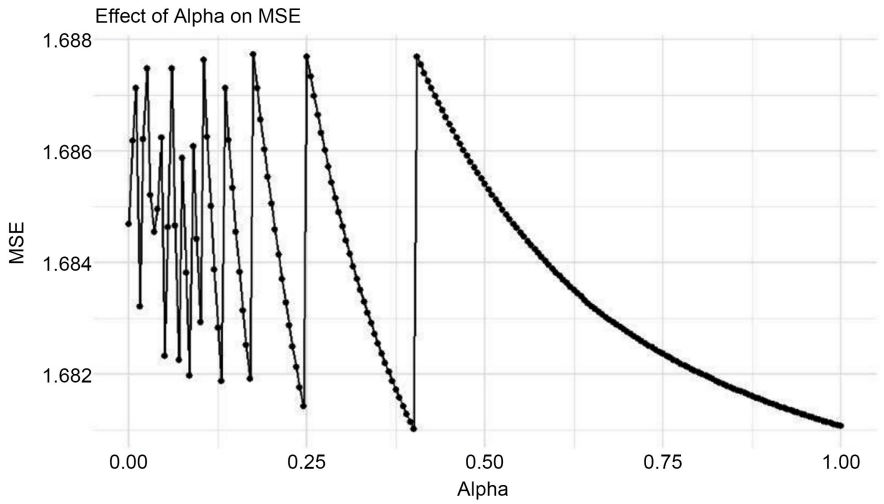

Figure 3. Traverse graph.

The four variable selection methods all selected job satisfaction (JS), health status (SH), environmental protection (EC), gap between rich and poor (II), education (Edu), social security (SS), satisfaction with their own lives (LS), local social status (LSS), trust to Americans (TA) and years of education (YE). Lasso and Elastic Network regularization methods also selected medical (Hea) variables, while AIC based variable selection methods selected more variables, including daily internet usage duration (DIU) and current marital status (CMS), demonstrating the importance of these variables in the model.

4.2. Analysis of Government Credibility Issues Based on Random Forests

For general regression-based methods, after normalization, the absolute value of the coefficient corresponding to the independent variable can represent its degree of influence on the dependent variable, so the absolute value of the coefficient is the degree of importance. For random forests, there is no concept of coefficients, but the importance of variables can be determined based on the following two indicators:

- %IncMSE (Percentage Increase in Mean Squared Error): This indicator measures the percentage increase in model error when variables are removed. The higher the value, the greater the impact of the variable on the predictive performance of the model, therefore the more important the variable is.
- IncNodePurity (Increase in Node Purity): This indicator measures the degree to which a variable improves node purity when used to segment nodes. The higher the value, the greater the impact of the variable on the decision-making process of the model, therefore the more important the variable is.

The larger the two indicators, the better, but the ranking of the same variable on these two indicators may be different. We choose to consider both rankings comprehensively and take the average of the variables in both rankings as the overall ranking. In terms of specific implementation, this article selects a random forest of 500 trees as a reference.

4.2.1. VSURF Method

The VSURF method selects variables by measuring their importance in a random forest model, and evaluates the predictive performance of the model by comparing the predicted results with the actual observed values and calculating the mean square error. The optimal mtry selected by the VSURF method is 4. Finally, 14 variables were selected from 20 variables, and the specific results are shown in Table 4.

It can be seen that in the random forest based on the VSURF method, social security, wealth gap, education issues, and healthcare issues are the four variables that have the greatest impact on government credibility. These four issues represent the four major areas of income, education, healthcare, and elderly care, which are also the most essential four areas in a person's life. Compared to others, local income level, social status, age, and health status are not as important.

4.2.2. RRF Method

Use the trainControl function to set the cross-validation method (method = "cv") and specify the number of folds for cross validation (number = 2). Call the train function to train the RRF model, and use cross validation during the training process to evaluate the performance of each set of parameters, in order to select the best combination of parameters. Use the trained RRF model to predict the test dataset (testData_rrf), calculate the mean square error between the predicted results and the actual observed values, and evaluate the predictive performance of the model. In order to accelerate the training process, this article

also attempted to use parallel computing methods through the doParallel package, which effectively reduced the running time and achieved good results. The RRF method effectively improves the model's generalization ability and prediction accuracy by utilizing the ensemble learning ability of random forests and combining it with regularization techniques. By cross validation and parameter tuning, the optimal model parameters were selected to further optimize the performance of the model. The optimal regularization coefficient coefReg selected for the RRF method is 0.9, the optimal importance coefficient coefImp is 0.5, and the optimal mtry is 3. Finally, 13 variables were selected from 20 variables, and the specific results are shown in Table 5.

Table 4. Analysis results of VSURF method.

Variable	%IncMSE	IncNodePurity
SS	0.13155353	203.05088
Edu	0.09905857	171.90519
II	0.09404796	179.92668
Hea	0.06888147	150.73287
JS	0.05215137	168.08401
Hou	0.04293124	129.23171
EC	0.03629889	132.59746
Emp	0.03799304	105.17284
LS	0.02477383	85.79587
TA	0.02728953	121.08674
LSS	0.01949882	87.94910
Age	0.01516251	181.22060
SH	0.01191190	123.44444
LIL	0.01246424	85.42013

Table 5. Description of RRF method results.

Variable	%IncMSE	IncNodePurity
JS	0.04881507	174.32713
SH	0.01130591	136.33819
EC	0.03308799	142.70942
II	0.08589480	180.53781
Emp	0.04184010	120.57413
Edu	0.10011374	176.47246
Hea	0.07973504	163.10666
Hou	0.04657975	145.74367
ss	0.12456012	210.70754
Ls	0.01749683	91.48309
LIL	0.01250760	92.06090
LSS	0.01298612	95.96894
TA	0.02896957	132.67629

4.2.3. Boruta Method

Firstly, the Boruta algorithm is used to select features from all variables in the training set to predict the target variable GOV and extract the selected features from the Boruta algorithm. Then generate new training and testing datasets that only contain the selected features and target variable GOV. Train the model on a newly created training dataset using the random forest algorithm, and use the trained random forest boruta_model to predict and generate predicted values on the test dataset. Calculate the mean square error (MSE) between the predicted and actual values based on the predicted set to evaluate the performance of the model. The optimal mtry selected by the Boruta method is 6. Finally, 19 variables were selected from 20 variables, and the specific results are shown in Table 6.

Table 6. Description of boruta method results.

Variable	%IncMSE	IncNodePurity
JS	0.0461320005	134.17053
SH	0.0088382713	88.82949
EC	0.0256668314	102.47680
II	0.0798574620	157.60446
Emp	0.0248835927	75.81821
Edu	0.0800532394	151.16008
Hea	0.0599475620	120.51307
Hou	0.0385381305	106.39238
SS	0.1118911063	176.00574
LS	0.0210023587	66.09119
LIL	0.0089915932	59.73633
LSS	0.0143407587	65.91882
TA	0.0169561173	89.60364
Age	0.0083642639	120.46459
YE	0.0089283678	64.97606
WWH	0.0010888679	127.09424
DIU	0.0044517464	124.09338
AWI	0.0008317609	137.37886
CMS	0.0065816756	15.55762

It can be seen that in the random forest based on Boruta, social security, wealth gap, job satisfaction, education issues, and healthcare issues are the five variables that have the greatest impact on government credibility. And Gender was not selected, indicating that there is no significant difference in the views of men and women on government credibility; Alternatively, it can be considered that the differences reflected by the Gender variable are reflected in the first 19 variables. That is, the variable has multicollinearity with other variables.

4.3. Analysis of Practical Significance

4.3.1. Regression Method Variable Analysis

The variable selection and importance ranking between regression methods are slightly different, but the difference is not significant. Only one model with the best performance can be analyzed as a representative of the regression method model. In the previous section, we compared the simplicity and accuracy of the model. The MSE of the model with p-value backward selection is 0.657, and the number of variable selections is 10, which combines simplicity and accuracy, making it the best model for comprehensive selection in regression methods. Hence, we will take this model result as the main body to analyze the model results and their significance.

The importance of the variables selected backward by the p-value is in the order of SS (social security), Edu (education), JS (job satisfaction), II (gap between rich and poor), TA (trust in Americans), LSS (local social status), EC (environmental protection), LS (satisfaction with their own lives), YE (education), SH (health status) according to the absolute value of the coefficient.

Overall, SS, Edu, and JS are the most important. These issues are the most concerning for the people and the most important for the government to implement. Making improvements to these issues can greatly enhance the credibility of the government, and in situations where time and resources are limited, priority should be given to addressing these issues. Due to the limited selection of variables in the backward selection model for P-values, these 10 variables are all important for government credibility and should be given due attention when studying government credibility. Therefore, the following six indicators should also be the direction of government efforts.

4.3.2. Variable Analysis of Random Forest Model

The order of variable selection and variable importance ranking between random forests is also similar, but there are some differences from the results of the regression model. Therefore, the best random forest model is selected to represent the random forest method for analysis. Based on the simplicity and accuracy analysis in the previous text, the results of the RRF model are selected as the main analysis model for the results of the random forest model.

The number of variables selected by RRF is 13, and the order of importance is: SS (social security), Edu (education), Hea (medical care), II (wealth gap), JS (job satisfaction), Hou (housing), Emp (employment), EC (environmental protection), LS (satisfaction with their own lives), TA (confidence in Americans), SH (health status), LIL (local income level), LSS (local social status). The most important of these are the first four: SS, Edu, Hea, and II. These four questions represent the four major areas of income, education, healthcare, and elderly care, which are also the core four areas of a person's life. It is interesting that the local income level LIL, education level YE, and health status SH corresponding to income, education, and healthcare are ranked in the fourth level of importance. That is to say, these four major areas are more important for people's under-

standing of these fields, rather than their own status at these levels. For example, a person may have a low level of education, but they may not consider it a social issue. The focus is on human perception. So in order to enhance the credibility of the government, it should focus on implementing the above five areas. Of course, this does not mean raising everyone's indicators to the highest level, but rather achieving fair and reasonable solutions to these issues. Taking education as an example, it is not about ensuring that everyone is admitted to university, but about achieving fairness, justice, and rationality, so that everyone believes that the system is correct.

In summary, variables such as SS, Edu, JS, and II are all located in relatively important positions in the two major categories of models. Especially SS and Edu, they almost rank in the top two among the seven models. Most of all, social security is the biggest influenced factor. It shows that social security and education are the issues that our people are most concerned about, and also the issues that people think the government should shoulder the responsibility most. The government needs to focus on implementing social security to enhance the trust of local residents in the government and ensure the well-being of the people.

5. Summary

As an important field in the field of social sciences, it is a meaningful research topic to determine the statistical significance of the influencing factors of government credibility. This article conducts empirical analysis and research on various factors that affect government credibility through social statistics. The models for variable analysis include seven types: stepwise backward selection method (based on p-value, AIC), LASSO, elastic network, random forest, regularized random forest, and Boruta. The research results of this article indicate that there are certain differences in simplicity, accuracy, and variable importance ranking among different variable selection methods, which screen different influencing variables for different dimensions of government credibility. Through the research in this article, it is expected to provide some reference for the relevant theories on government credibility issues.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hu, X.M. (2021) Logical Connotation and Implementation Path of Enhancing Government Credibility. *People's Forum*, No. 34, 73-75.

[2] Gong, H.L. and Li, W.Y. (2023) Research on the Impact of Perception of Public Service Access and Social Equity on Government Trust: Empirical Analysis-Based on CGSS. *Administrative Science Forum*, **10**, 58-64.

[3] Lv, W.-X. and Wang, Y.-G. (2010) The Influential Mechanisms of Public-Perceived Administrative Service Quality on the Reputation of Governments. *Journal of Renmin*

University of China, No. 4, 117-126.

[4] Greene, W.H. (2002) Econometric Analysis. Prentice Hall.

[5] Wu, X.F. (2008) Review of Research on Government Credibility in Recent Years. *China Administrative Management*, No. 5, 63-67.

[6] Mao, S.L. and Tan, Y.D. (2012) The Four Realms of Public Trust in Government: The Path to the Growth of Government Credibility. *People's Forum*, No. 18, 24-25.

[7] Lai, X.Y. (2023) How Do Citizens' Perceptions of Livelihood Issues Affect Their Trust in the Government?—Empirical Analysis Based on CFPS (2020). *Modern Management*, **13**, 1770-1781. https://doi.org/10.12677/mm.2023.1312223

[8] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. https://doi.org/10.1109/tac.1974.1100705

[9] Fan, H., Wu, Y., Liu, X., *et al.* (2024) Construction and Validation of a Column Chart Prediction Model for Active Pulmonary Tuberculosis Based on LASSO Regression. *Chinese Clinical Research*, **37**, 424-429.

[10] Liu, P., Hu, J.J. and Xie, L.L. (2024) Ranking and Comparison of Seismic Motion Parameters Based on Elastic Network Regression. *Journal of Harbin Institute of Technology*, **56**, 54-62.

[11] Zhao, H., Lu, Y.J., Gao, J., *et al.* (2020) Algorithm and Application of Guided Regularized Random Forest SMOTEBoost. *Statistics and Decision*, **36**, 9-14.

[12] Ye, N., Morgenroth, J. and Xu, C. (2023) Improving Neural Network Classification of Native Forest in New Zealand with Phenological Features. *International Journal of Remote Sensing*, **44**, 6147-6166. https://doi.org/10.1080/01431161.2023.2264496

[13] Thakur, D. and Biswas, S. (2024) Permutation Importance Based Modified Guided Regularized Random Forest in Human Activity Recognition with Smartphone. *Engineering Applications of Artificial Intelligence*, **129**, Article ID: 107681. https://doi.org/10.1016/j.engappai.2023.107681

[14] Karbasi, M., Ali, M., Bateni, S.M., Jun, C., Jamei, M., Farooque, A.A., *et al.* (2024) Multi-Step Ahead Forecasting of Electrical Conductivity in Rivers by Using a Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) Model Enhanced by Boruta-XGBoost Feature Selection Algorithm. *Scientific Reports*, **14**, Article No. 15051. https://doi.org/10.1038/s41598-024-65837-0