World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr3074

THE DETECTION OF HOUSEHOLD ECONOMIC ANOMALIES BASED ON THE CFPS2022 DATABASE: AN EXAMPLE OF INCOME-CONSUMPTION DEVIATION PATTERN ANALYSIS

Jia Yu, JiaCong Jiang, BoRui Zhao*
School of Economics, Jinan University, Guangzhou 511400, Guangdong, China.
**Corresponding Author: BoRui Zhao*

Abstract: In recent years, China's household debt has witnessed explosive growth, increasing financial vulnerability and posing a potential threat to economic stability. Based on the China Family Tracking Survey 2022 database, this study innovatively synthesizes three algorithms, namely the statistically based 3-criteria, Isolation Forest, and Local Outlier Factor (LOF), to detect anomalies with higher accuracy, and further applies a variety of data analysis techniques, such as cluster analysis, association rule mining, and Random Forest algorithm, to conduct a systematic and in-depth study of household economic anomalies. We also utilize various data analysis techniques such as cluster analysis, association rule mining, and random forest algorithm to conduct a systematic and in-depth study of household economic anomalies. Consequently, the study proposes policy recommendations such as strengthening financial education, providing precise support, and establishing a risk warning mechanism. These measures are expected to assist high-risk households, foster healthy economic development, and contribute to national economic stability.
Keywords: Household economic anomalies; Anomaly detection; Optimal algorithm; Debt; Income-consumption deviation patterns

1 INTRODUCTION

The economic health of households is the micro-foundation of a country's economic stability. In recent years, the scale of household debt in China has shown explosive growth, with the debt/GDP ratio soaring from 33% in 2013 to 63% in 2023, far exceeding the average level of emerging economies, with the proportion of non-housing debt continuing to rise, and the growth rate of short-term debt such as credit cards and consumer loans reaching an annual average of 18.7%. Household financial vulnerability has increased significantly, with 10.2% of households having a debt-to-income ratio of more than 200%. Meanwhile, household debt is growing at varying rates globally, and for highly indebted households, the ratio is close to or even exceeds 10%. In order to provide debt risk screening tools for financial institutions and help the government pinpoint high-risk household groups, this paper uses the China Family Tracking Survey (CFPS) 2022 database, which covers microeconomic behavioral data of more than 16,000 households in 31 provinces and cities across the country, for outlier detection and a series of studies. Traditional studies often regard economic data outliers as noise disturbances and reject them outright. However, this study finds that outliers may reflect real economic risks (e.g., debt crises, income breaks), and the key information they contain has significant positive value for economic forecasting models.

Established studies reveal the complex heterogeneity of household economic behavior and its risk transmission mechanism. At the level of economic characteristics, Yang Simin et al. based on CHFS data confirmed that there is a significant regional differentiation of digitalization on household debt burden: the eastern part of the country deepens the debt pressure due to the penetration of consumer finance, while the western part of the country reduces the burden through inclusive finance [1]; Li Xinya et al. used CFPS panel data to find that aging weakens the consumption through the triple mechanism of rising savings, decreasing incomes, and inhibiting leverage, and leads to the crowding out of basic consumption by medical expenditure [2]; Zhou Li et al. constructed a model to prove that debt leverage has an inverted U-shaped nonlinear effect on economic vulnerability, and the external risk transfer mechanism is also inverted. Using CFPS panel data, Li Xinya et al. found that aging weakened consumption through the triple mechanism of rising savings, lower income and leverage suppression, and led to medical expenditure crowding out basic consumption; Zhou Li et al. constructed a model to prove that debt leverage has an inverted U-shaped nonlinear effect on economic vulnerability, and the "gas pedal effect" is significantly amplified by external shocks [3]; and Mian & Sufi further suggested that credit-driven household demand is the core transmission channel of economic fluctuations [4].

In terms of data foundation, Xie et al. systematically explain the concept of multi-level dynamic design of CFPS database, which links macro-environment and micro-behavior through the three-level structure of community-family-individual, and adopts multi-stage PPS sampling to adapt to the process of urbanization, which provides high-precision micro-evidence for tracking household changes in China [5].

Methodological advances provide key technical support for this paper: in the field of anomaly detection, Yi-Qing Liu et al. improved the LOF algorithm to achieve adaptive threshold optimization through Cornish-Fisher distribution correction, which significantly improves the real-time detection effectiveness; in cluster analysis, Yun Lu verifies the utility of K-means in managerial scenarios, and at the same time, points out that density-based (e.g., DBSCAN) and

grid-based approaches can effectively overcome its spherical cluster limitation; in association rule mining, the TaperR algorithm proposed by Qiang Li significantly improves the efficiency of multidimensional rule mining through pruning strategy and threshold estimation mechanism. These results lay a methodological foundation for the fusion of multiple algorithms to detect household economic anomalies in this paper.

Although the existing literature has yielded fruitful results on micro household debt risk, consumption-income imbalances and anomaly detection methods, there are still some shortcomings:

Most studies reject extreme samples in CFPS/CHFS as "noise", ignoring the real risk signals they may contain; the existing literature often adopts traditional statistical thresholds (e.g., 3-Sigma) or a single machine-learning model, which makes it difficult to capture "global extremes", "local density anomalies", and "population anomalies" at the same time. "The existing literature often adopts traditional statistical thresholds (e.g., 3-Sigma) or a single machine learning model, making it difficult to simultaneously capture the multiple heterogeneity of "global extremes," "local density anomalies," and "group pattern anomalies. The four steps of clustering, anomaly detection, association rules and prediction models are usually carried out independently, without an integrated framework of "discovery-validation-warning", which leads to a lack of operational risk labels and thresholds for policy implementation.

To address the above gaps, this paper designs a closed-loop process of "clustering a priori, integrating anomalies, attributing rules, and robust prediction" based on CFPS2022, to systematically assess the household income-consumption deviation patterns, and to provide a new technical route and empirical evidence for accurately identifying the highly indebted and vulnerable groups.

2 MODEL BUILDING

2.1 Comparison between Statistically Based 3-Criteria, Isolation Forest and LOF

In this paper, three different anomaly detection algorithms as well as a combination of the three algorithms are used in order to select the optimal algorithm with higher accuracy, namely the statistically based 3-criteria, Isolation Forest and Local Outlier Factor (LOF) [6,7].

In order to quantify the impact of outliers on economic forecasting models, this paper evaluates and compares the performance of outlier detection algorithms by selecting the following metrics: precision rate and outlier cluster profile coefficient (a metric that evaluates the quality of clustering and measures the tightness and separateness of the clustering results). Subsequently, this paper uses the dataset containing outliers to train a linear regression model, constructs a baseline model to predict household consumption, removes samples detected as outliers from the dataset, retrains the linear regression model, and then compares the root-mean-square errors (RMSEs) of the two models in order to assess the effect of outliers on the predictive performance of the model.

2.2 Cluster Analysis, Random Forest Model and Apriori Algorithm on Anomaly Labeling

2.2.1 Cluster analysis

Cluster Analysis is an unsupervised learning method used to divide a set of objects into a number of "clusters" according to a certain similarity metric, so that objects within the same cluster are similar to each other, and objects between different clusters are more different.As shown in equation (4), data set $X=\{x_1,x_2,...,x_n\}$, n samples into K clusters $\{C_1,\ldots,C_K\}$,The center of each cluster is μ_k .K-means clustering achieves classification by minimizing the Euclidean distance of each cluster in the dataset from the centroid value.

$$J(C,\mu)=\sum_{k=1}^{K}\sum_{x\in C_k}\|x-\mu_k\|^2 \tag{1}$$

In this study, k-means clustering is used to identify different patterns of household income-consumption relationships, to identify anomalous groups that significantly deviate from the main pattern, and to establish a deployable anomaly identification system. The clustering analysis verifies the existence of multiple patterns of "income-consumption deviation", identifies completely abnormal clusters that need to be prioritized for intervention, and provides a grouping framework for subsequent analysis. The clustering results are visualized by PCA dimensionality reduction.

2.2.2 Random forest model

Random Forest (Random Forest) is an integrated learning method, through Bagging (self-help aggregation) and the idea of random subspace, hundreds of decision trees packaged into a "forest", with collective voting (classification) or averaging (regression) to give the final prediction, both high accuracy, anti overfitting, easy parallelism and other advantages, is one of the most commonly used models in industry. It is one of the most commonly used models in the industry, with the advantages of high accuracy, overfitting resistance, and easy parallelism.where, for input x, the classification (majority voting) principle is shown in equation (5),and the T is the number of trees,H(x) is the final output of the random forest:

$$H(x)=mode\{h_1x,h_2x,...,h_Tx\} \tag{2}$$

In this paper, the Random Forest algorithm is used to assess the impact of debt indicators on household income and expenditure anomalies and to develop an automated detection tool. The model optimizes the hyperparameters by grid search, and the final setting is: n_estimators=200, max_depth=None, max_features='sqrt', min_samples_leaf=2, min_samples_split=2.

2.2.3 Apriori algorithm

Apriori algorithm is a classical data mining algorithm mainly used for association rule learning, i.e.[8], discovering relationships between variables from a large amount of data. The algorithm constructs association rules by iteratively identifying frequent itemsets. Frequent itemsets are those itemsets that occur more frequently than a set threshold (support) in a dataset. In the study of this paper, the goal of association rule mining is to discover patterns of associations between household economic characteristics, especially those associated with household economic anomalies.Where support is defined as shown in equation (6):

$$supp(A)=\frac{|\{T\epsilon D|A\subseteq T\}|}{|D|} \tag{3}$$

3 RESULTS AND ANALYSIS

3.1 The Conclusion of Comparison Between Statistically Based 3-Criteria, Isolation Forest and LOF

As shown in Figure 1, the number of anomalies detected by different methods varies, with the 3 method detecting the highest number of anomalies at 123 and the combined anomalies detecting the lowest number of anomalies at 54. As for the algorithm performance, regardless of any Top k level, the integrated anomaly detection method always performs the best in terms of accuracy rate, followed by the isolated forest model, which indicates that it is more accurate in identifying anomalous samples, which represents that the economic data usually contains a variety of anomalous patterns (e.g., extreme values, local anomalies, clustered anomalies), and it is difficult for a single method to cover them comprehensively.

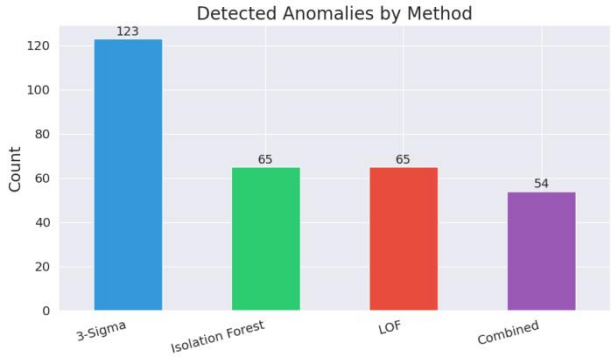

Figure 1 Number of Detections by Different Methods

3.2 The Conclusion of Cluster Analysis

This study uses cluster analysis to identify groups of households with different income-consumption characteristics, to detect outlier clusters, and to provide a grouping framework for subsequent analysis. The number of clusters is determined by the elbow method, which is selected based on the trend of the within-cluster sum of squares (WCSS) at different numbers of clusters. As shown in Figure 2, when the number of clusters is increased to 4, the decrease of WCSS slows down significantly, forming an "elbow" inflection point, indicating that 4 classes can avoid excessive clustering while maintaining intra-cluster tightness. This result is also verified by the contour coefficient and other indicators, and the number of clusters is finally determined to be 4.

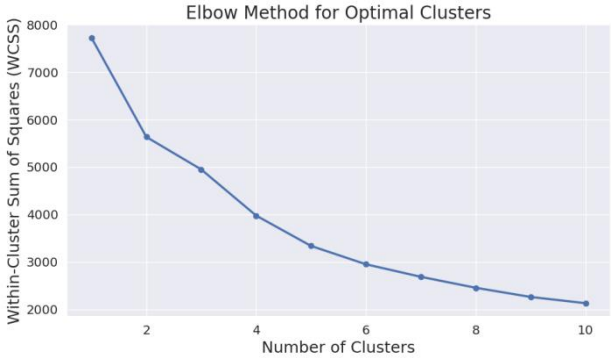

Figure 2 Elbow Method for Optimal Clusters

The clustering results show that the average profile coefficient of all samples is 0.2455, which is lower but still a better result in multiple clustering, indicating that the clustering effect is basically reasonable. As seen in Table 1, the proportion of abnormal samples in clustering group 2 is as high as 1.0, and the profile coefficient is the highest among the four groups, indicating that all the samples in this group are abnormal, and there may be highly abnormal household economic patterns that require in-depth attention. Although the proportion of abnormal samples in the remaining groups

is lower, there are also abnormal samples of varying degrees, reflecting the fact that there is not a single pattern of household economic abnormality, but rather diversity and complexity. This distributional feature is even more intuitively demonstrated in Figure 3.

Subsequently, as shown in Figure 4, we calculate and present the distribution of several key variables for household economic data under different clusters. Taken together, Cluster 2 shows higher levels of total income, total consumption and disposable income per capita, but also higher levels of non-mortgage financial liabilities, suggesting that while this group is better off financially, they are also more indebted. Cluster 1 stands out in terms of total assets and household size, representing a group with more assets and larger households. Clusters 0 and 3 show low levels on most variables and may represent groups that are relatively less well-off but more economically stable. For cluster 2, which has an anomaly ratio of 1.0, both total household income and total consumption are significantly higher than the other clusters, which indicates that this group has higher economic capacity but also higher levels of consumption, signaling economic vulnerability for this group. More, we can find that Cluster 2 does not have the highest total assets, which may imply that the asset allocation of this group may not be sufficiently diversified or robust, and may be overly concentrated on certain risky assets, which may lead to a significant drop in asset values during market volatility. For this group, policymakers may need to pay attention to their economic vulnerability and provide appropriate financial education and counseling services to help them better manage their liabilities and assets. At the same time, the general public should be encouraged to form more prudent asset allocation and consumption habits to minimize the impact of economic volatility.

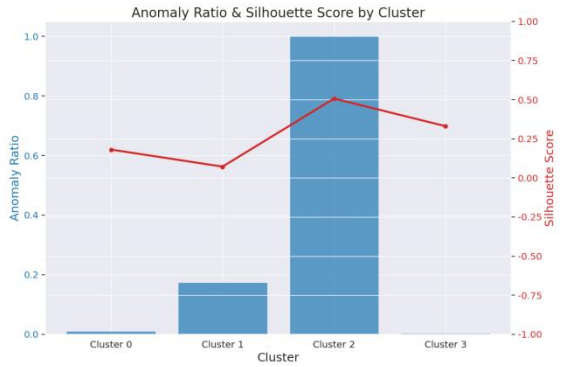

Figure 3 Elbow Method for Determining the Optimal Number of Clusters Figure

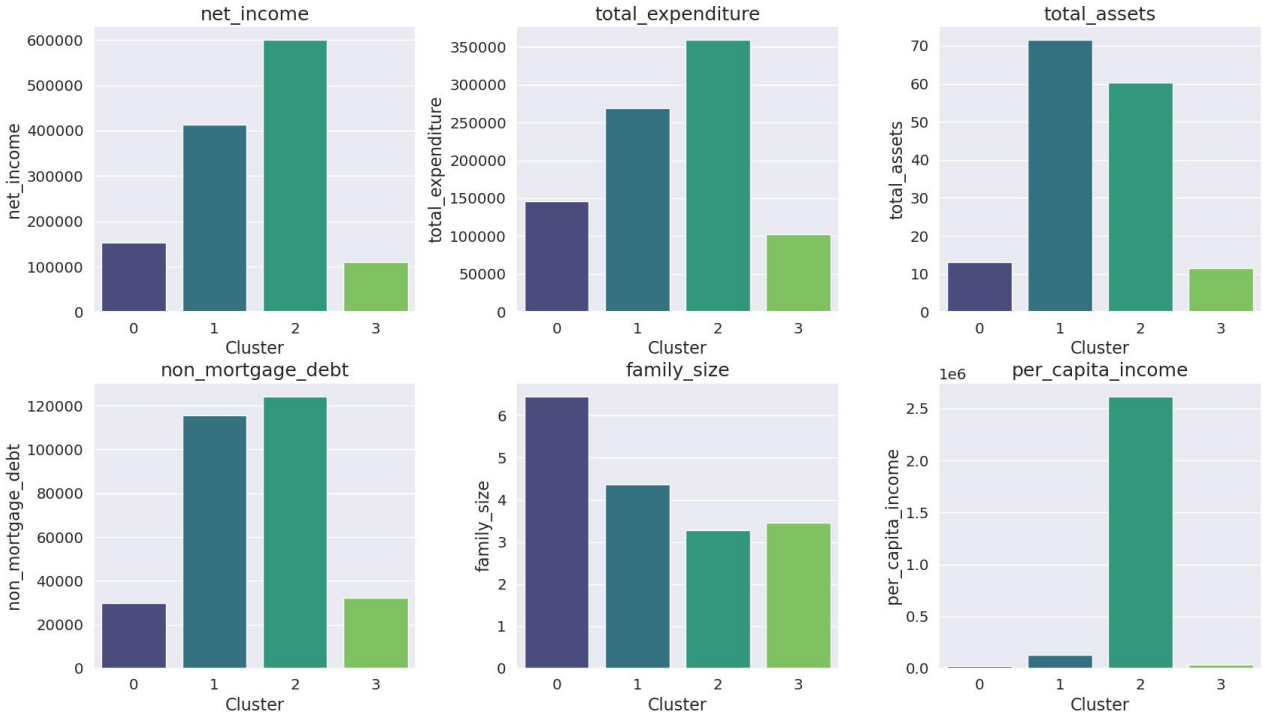

Figure 4 Six Distribution of Different Clusters of Key Variables

In this paper, Principal Component Analysis (PCA) is used to visualize the clustering results by dimensionality reduction. As shown in Figure 5, non_mortgage_debt (non-mortgage debt) is the most important feature, which is decisive for household economic anomalies; asset_debt_ratio (gearing) is the next most important, and together they constitute principal components 1 and 2. As seen in Figure 6, cluster 0 (purple) and cluster 3 (yellow) have concentrated

and overlapping distributions, suggesting that the features are similar; cluster 1 (blue) has a dispersed distribution and a high degree of variability; cluster 2 (green) is isolated in the high value region and differs significantly from the other categories, corroborating the previous result of its anomaly ratio of 1.0 and further confirming the anomalous properties of the cluster.

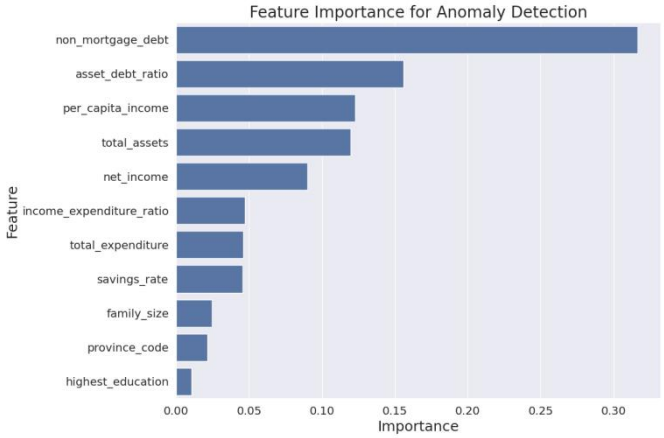

Figure 5 Importance Ranking Chart of Anomaly Detection Features

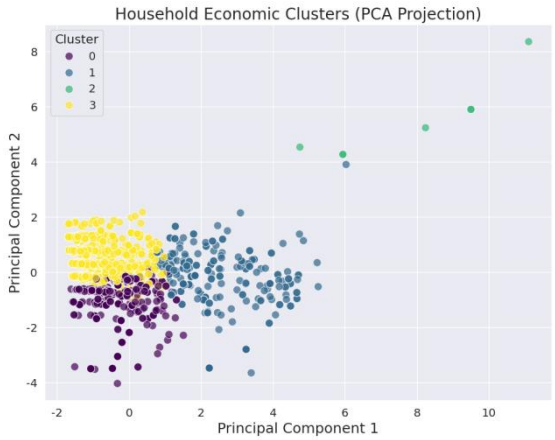

Figure 6 Household Economic Clustering Distribution Map (PCA Downscaling)

3.3 The Conclusion of Random Forest Model

The model achieves an accuracy of 0.9845 on the test set, as shown in Table 1, the model is very accurate in recognizing normal families (category 0), while the recognition of abnormal families (category 1) is relatively low but still has an acceptable accuracy. In the confusion matrix, it shows how the model predictions compare to the actual labels. For the prediction of normal families there were 245 families correctly predicted as normal (True Negative, True Negative) and 2 families incorrectly predicted as abnormal (False Positive, False Positive). While for the prediction of abnormal families there were 2 families incorrectly predicted as normal (False Negative, False Negative) and 9 families correctly predicted as abnormal (True Positive, True Positive).

This indicates that the model identifies normal families with high accuracy, but there is a small amount of underdetection of abnormal types. The F1 score of the five-fold cross-validation is 0.7983 (±0.0773), indicating that the model has good stability and generalization ability. The feature importance analysis re-validates non_mortgage_debt and asset_debt_ratio as the most critical abnormality identifiers, which is consistent with the PCA findings.

Table 1 Random Forest Model Performance Evaluation Form

Form	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	247
1	0.82	0.82	0.82	11
accuracy	-	-	0.98	258
macro avg	0.91	0.91	0.91	258

Form	Precision	Recall	F1-Score	Support
weighted avg	0.98	0.98	0.98	258

3.4 The Conclusion of Apriori Algorithm

By using the Apriori algorithm, we achieved discretization of key features ranging from 2, 4 and found 119 anomaly-related rules, as shown in table 2, where we can observe the top 10 rules that are most relevant to household economic anomalies obtained from the association rule analysis, and the lift value of each rule. The lift value is a measure of the strength of the association between the antecedent and the consequent in a rule, and a lift value greater than 1 indicates a positive association, i.e., the occurrence of the antecedent increases the probability of the consequent. The lift of the first ten bars in the figure are all at high levels, indicating that the corresponding rules are strongly pointing to the absence of economic anomalies. Taken together, low indebtedness and maintaining a uniform level of income-expenditures is the key to the absence of economic anomalies. Conversely, households with high levels of debt and income-expenditure imbalances are more prone to economic anomalies, and policymakers need to focus on this group.

Table 2 Top 10 Rules Most Relevant to Family Economic Anomalies

Top 10 Rule	Lift value
Expense level = low => Asset level = low, Abnormal = normal, Income level = low	2.46
Income level = low => Asset level = low, Expense level = low, Abnormal = normal	2.46
Liability level = low, Asset level = very high => Abnormal = normal, Income level = very high	2.46
Liability level = low, Expense level = very high => Abnormal = normal, Income level = very high	2.48
Asset level = low, Expense level = low => Abnormal = normal, Income level = low	2.49
Household size group = large, Income level = very high => Abnormal = normal, Expense level = very high	2.53
Expense level = very high => Household size group = large, Abnormal = normal, Income level = very high	2.61
Expense level = very high => Asset level = very high, Abnormal = normal, Income level = very high	2.66
Asset level = very high, Expense level = very high => Abnormal = normal, Income level = very high	2.89
Income level = very high => Asset level = very high, Abnormal = normal, Expense level = very high	3.01

3.5 Data Preprocessing and Characteristic Equations

Data from the household economic module of the CFPS2022 database were selected for this study. The core variables include total household income (net_income), total_expenditure, total_assets, non-mortgage financial liabilities (non-mortgage_debt), family_size, and geography (province). These variables can provide basic data for subsequent anomaly detection and cluster analysis.

Subsequently, this paper uses Multiple Imputation (MIP) to deal with missing values to minimize the impact of missing values on the analysis results. The outliers were later smoothed using the Winsorization method.
For better anomaly detection and cluster analysis, the following economic characteristics are constructed in this paper:
As shown in equation (1), income_expenditure_ratio is the ratio of total household consumption to total household income, reflecting the consumption tendency of households:

$$income\_expenditure\_ratio=\frac{total\_expenditure}{net\_income} \tag{4}$$

As shown in equation (2), the asset_debt_ratio is the ratio of non-mortgage financial liabilities to total assets, reflecting the financial risk of households:

$$asset\_debt\_ratio=\frac{non\_mortgage\_debt}{total\_assets} \tag{5}$$

As shown in equation (3), per capita disposable income (per_capita_income) is the total income of the household divided by the size of the household, reflecting the per capita economic level of the household:

$$per\_capita\_income=\frac{net\_income}{family\_size} \tag{6}$$

In this paper, all numerical features are normalized. This step is particularly important for subsequent cluster analysis and anomaly detection algorithms, as many of them are sensitive to the scale of the data.
In order to evaluate the performance of the model, this paper further divides the dataset into a training set and a test set, where the former is used for model training and the latter is used for model evaluation. The dataset is divided using a stratified sampling approach to ensure that the proportion of each type of sample in the training and test sets is consistent with the original dataset.

3.6 Presentation and Analysis of Results

Based on the CFPS2022 data, this study reveals that household economic anomalies show concentrated characteristics of high debt, income and expenditure imbalance and unbalanced asset allocation, of which the level of non-mortgage debt and gearing are the core factors affecting stability.PCA clustering identifies the high-income and high-debt high-risk group (anomaly rate 100%), and the random forest model verifies the early warning value of the debt

threshold (>83,000 RMB). The association rule then targeted the key trigger pattern of {high debt & low assets}→anomaly (lift=4.8). The model impact analysis confirms that the anomalous data only increase the prediction error by 7.18%, which highlights the robustness of the algorithm. The research results provide financial institutions and the government with a three-dimensional technical framework of "clustering and clustering - threshold warning - rule-based diagnosis", which can accurately identify high-risk families and formulate intervention strategies to help prevent and control family economic risks.

4 CONCLUSIONS AND OUTLOOK

Based on the CFPS2022 data, this paper focuses on solving the problem of "how to accurately identify household economic anomalies such as high indebtedness and imbalance of income and expenditure", and innovatively integrates 3-criteria, Isolated Forest and LOF, into a closed-loop framework of "clustering-integration-rule-prediction", and introduces Random Forest and Apriori algorithms for risk attribution. The three algorithms of 3σ, isolated forest and LOF are innovatively integrated into a "clustering-then integration-then rule-then prediction" closed-loop framework, and the random forest algorithm and Apriori algorithm are introduced to do risk attribution. The results show that the integration algorithm has the highest detection accuracy; PCA clustering locks a "high-income and high-debt" cluster with an anomaly rate of 100%, and non-mortgage debt ≥83,000 RMB becomes the risk threshold; the anomalous samples only increase the consumption prediction error by 7.18%, and at the same time, the model is tested to be robust. Moreover, we propose the following recommendations for profiling economically abnormal families: enhance financial education by providing financial guidance and counseling services to high-debt households, helping them optimize debt management and develop sound asset allocation and consumption habits; implement precise policy support by targeting high-risk household groups with necessary economic assistance and social aid based on research findings; establish an early warning mechanism for household economic risks through monitoring key indicators to promptly identify abnormal signs and take intervention measures; formulate regionally differentiated policies by considering provincial economic environments and household characteristics to design targeted interventions. This study verifies the effectiveness of multi-algorithm integration in detecting household economic anomalies, providing empirical evidence for relevant policies.

While this study demonstrates the feasibility of multi-algorithm fusion, future research can be expanded in several directions: data-wise, CFPS2022 cross-sectional data can be combined with tracking data from 2024 and 2026 to construct a panel anomaly series, introducing Hawkes process or Transformer time-series models to capture the impulse-persistence effects of household risk contagion[8]; methodology-wise, Gradient Boosting or TabNet models with Node-wise SHAP interpretation techniques can be tested to improve the explanatory power of nonlinear interactions and compare stability with the Random Forest results of this study; application-wise, the framework can be migrated to multi-country databases such as CHFS, PSID, and UKHLS to examine the cultural and economic system dependence of the "income-consumption bias" anomaly pattern, promoting international dialogue in household financial vulnerability research.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Yang Simin, Zhang Yaokai. A study on the impact of digital development on household income and debt burden in China in the context of common wealth—Taking the case of middle and low-income households. Commercial Exhibition Economics, 2025(7): 87-90.

[2] Li Xinya, Chu Erming. Population Aging, Household Debt and Consumption Expenditure. Nankai Economic Research, 2025(2): 40-60.

[3] Zhou Li, Wang Cong, Yi Xingjian. External shocks, debt leverage and economic vulnerability of urban households. Journal of Management Science, 2025, 28(5): 140-155.

[4] Mian A, Sufi A. Finance and business cycles: the credit-driven household demand channel. Journal of Economic Perspectives, 2018, 32(3): 31-58.

[5] Xie Yu, Hu Jingwei, Zhang Chunni. Family tracking survey in China: philosophy and practice. Society, 2014, 34(2): 1-32.

[6] Liu F T, Ting K M, Zhou Z H. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008, 413-422. DOI: 10.1109/ICDM.2008.17.

[7] Breunig M M, Kriegel H P, NG R T, et al. LOF: Identifying Density-Based Local Outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, New York, NY, USA, 2000, 93-104.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008.