Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

A Quantitative Study on the Impact of National Demographic Characteristics on Cybercrime Distribution Based on Spearman and XGBoost Models

Qijia Xu [*, #], Zhengyang Sun [#], Xidong Yang [#]

School of Insurance and Economics, University of International Business and Economics, Beijing, China, 100029

* Corresponding Author Email: xuqijia123@163.com

#These authors contributed equally.

Abstract. This study focuses on quantitatively analyzing the relationship between national demographic characteristics and the distribution of cybercrime, aiming to identify the most influential macro-factors. The research first employed Spearman's rank correlation analysis to assess the linear correlations between twelve national indicators—encompassing economic, social, educational, and internet usage metrics—and the cybercrime factor. The results revealed strong linear correlations for GDP per capita, GDP, aging rate, and three internet-related data points. Subsequently, to incorporate both linear and non-linear relationships, this study utilized the XGBoost machine learning model for in-depth analysis.Feature importance analysis from the XGBoost model indicated that GDP per capita is the most influential national characteristic affecting cybercrime distribution, with a correlation coefficient of 0.24633. The number of secure internet servers and GDP were also identified as high-impact indicators. The findings suggest that while regions with high GDP per capita host active online financial activities and advanced technology—potentially offering more criminal opportunities—overall, economically developed areas may exhibit lower rates of cybercrime incidence. In contrast, the unemployment rate demonstrated the lowest impact, with a coefficient of only 0.01276. The model achieved a mean squared error (MSE) of 0.339 and a coefficient of determination (R^2) of 0.744, demonstrating strong explanatory power and predictive accuracy regarding the factors influencing cybercrime distribution.

Keywords: Cybercrime distribution; National characteristics; XGBoost model.

1. Introduction

Modern technology has vastly interconnected world, significantly boosting global productivity while simultaneously increasing vulnerability to cybercrime. Cybercrimes, such as hacking and data theft, have caused enormous economic losses worldwide, amounting to a staggering $9.5 trillion in 2024 alone, representing a growing threat. Furthermore, the transnational nature of cybercrime complicates jurisdiction, and some institutions choose not to report incidents to avoid information leakage, further increasing governance challenges[1-2]. In response to these challenges, many countries have formulated and issued cybersecurity policies. However, most research focuses on technical prevention, lacking quantitative analysis of policy effectiveness and macro-level influencing factors[3-4].

Against this backdrop, this study aims to address the following core questions: How is global cybercrime distributed? How effective are the cybersecurity policies issued by various countries, and how are they related to crime distribution? And how do national demographic characteristics influence the distribution of cybercrime. To address these questions, this study adopts a multi-model integrated research plan: Firstly, the K-means clustering and Entropy Weight Method-TOPSIS model are used to comprehensively evaluate and define a "Cybercrime Factor," thereby mapping the global distribution of cybercrime[5]. Secondly, the Random Forest model and Grey Prediction Model are employed to quantify and predict the effectiveness trends of national cybersecurity policies (based on the five dimensions of GCI). Finally, to fully incorporate both linear and non-linear effects, this study

combines Spearman's rank correlation analysis and the XGBoost machine learning model to delve into the relationship between national demographic characteristics and cybercrime distribution[6-7].

2. Spearman's Rank Correlation Analysis and Machine Learning XGBoost

This research aims to complete data cleaning by pre-processing the data and handling missing values. After obtaining the cleaned data, this research conducts a correlation analysis between national demographic statistics and the distribution of cybercrimes. Following the Spearman's coefficient test, this research identifies both linear and non-linear relationships. Since XGBoost can simultaneously handle both linear and non-linear relationships, this research performs in-depth machine learning using the XGBoost model based on the Spearman model[8-9].The Overall Idea is shown in figure 1.

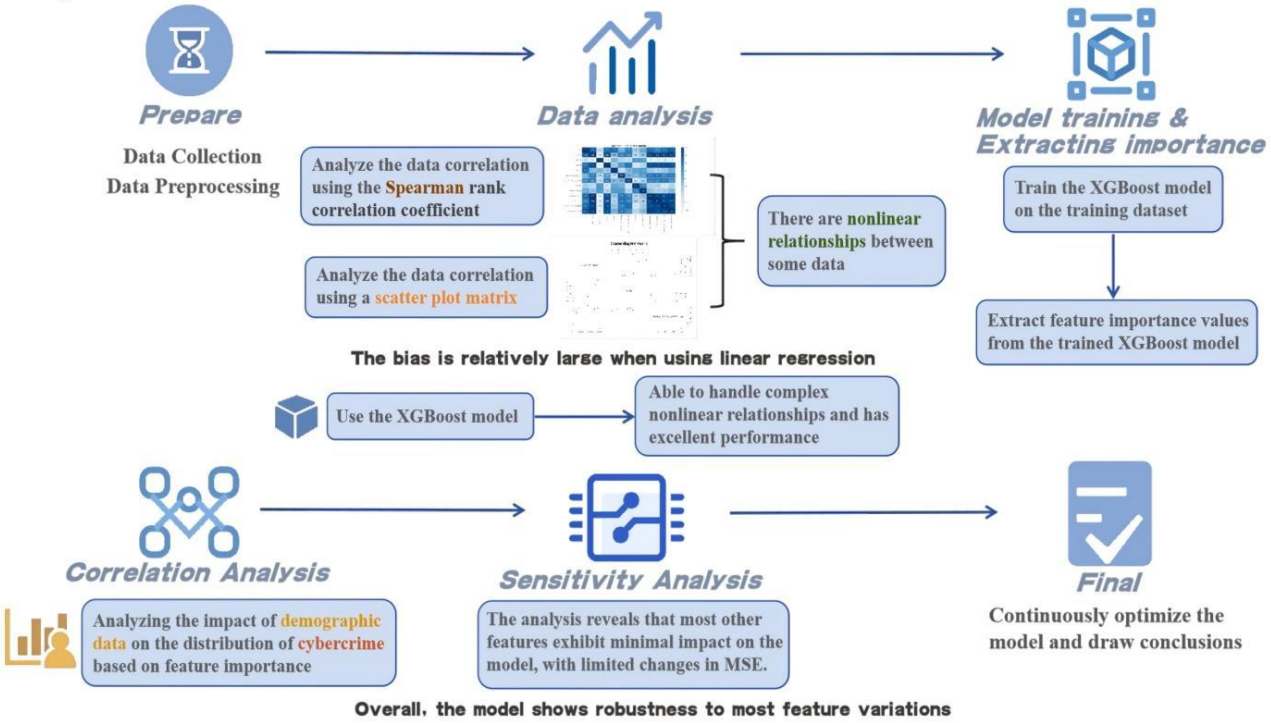

Figure 1. The Overall Idea

2.1. Concept of the Model

2.1.1. Establishment of Index - National Demographic Data

Table 1. Index - National Demographic Data

Category	Specific Indicators
Economic Index	GDP, GDP per Capita, GDP Growth Rate
Social Index	Crime Index, Unemployment Rate
Internet Index	Secure Internet Servers (per million people), Internet Usage Rate, Fixed Broadband Subscriptions (per 100 people)

Index - National Demographic Data is shown in table 1.

2.1.2. Cybercrime Factor Index

In the first large-scale model this research established, within the TOPSIS model section, this research defined the distance to the ideal solution. The cybercrime factor index is the reciprocal of this score[10].

2.2. Data Collection, Cleaning and Sorting

2.2.1. Data Sources

The data is sourced from official institutions such as the World Bank and platforms like Kaggle.

2.2.2. Cleaning and Sorting

(1) Missing Value Processing: After inspection, there are no missing values in the obtained data.

(2) Normalization and Z-Score Standardization:

Normalization (Min-Max Normalization):

Scale the data to the range [0, 1].

Formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Where x_{min} and x_{max} are the minimum and maximum values of the indicator, respectively.

Application Scenario: Suitable when the data range is large and all feature values need to be standardized to the same scale.

Z-Score Standardization:

Transform the data into a distribution with a mean of 0 and a standard deviation of 1.

Formula:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Where x is the original value, μ is the mean of the indicator, and σ is the standard deviation of the indicator.

Application Scenario: Suitable when the data contains positive and negative values or needs to conform to a Gaussian distribution.

2.3. Spearman's Rank Correlation Analysis

2.3.1. Introduction to the Model

The steps to calculate the Spearman rank correlation coefficient (ρ) are as follows:

Rank Transformation:

Convert the observed values of each variable into ranks. The smallest value is assigned rank 1, the second smallest rank 2, and so on. For multiple identical values (ties), their average rank is used.

Calculate Rank Differences:

For each pair of observed values, compute the difference between their ranks in the two variables:

$$d_i = \text{rank}(X_i) - \text{rank}(Y_i) \tag{3}$$

Calculate the Coefficient:

Use the following formula to compute the Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{4}$$

Where n is the number of observation pairs, and d_i is the rank difference of each pair of observed values.

Interpretation of Results:

$\rho = +1$: Indicates a perfect positive monotonic relationship (when one variable increases, the other always increases).

$\rho = -1$: Indicates a perfect negative monotonic relationship (when one variable increases, the other always decreases).

$\rho = 0$: Indicates no monotonic relationship between the variables.

2.3.2. Summary of the Model

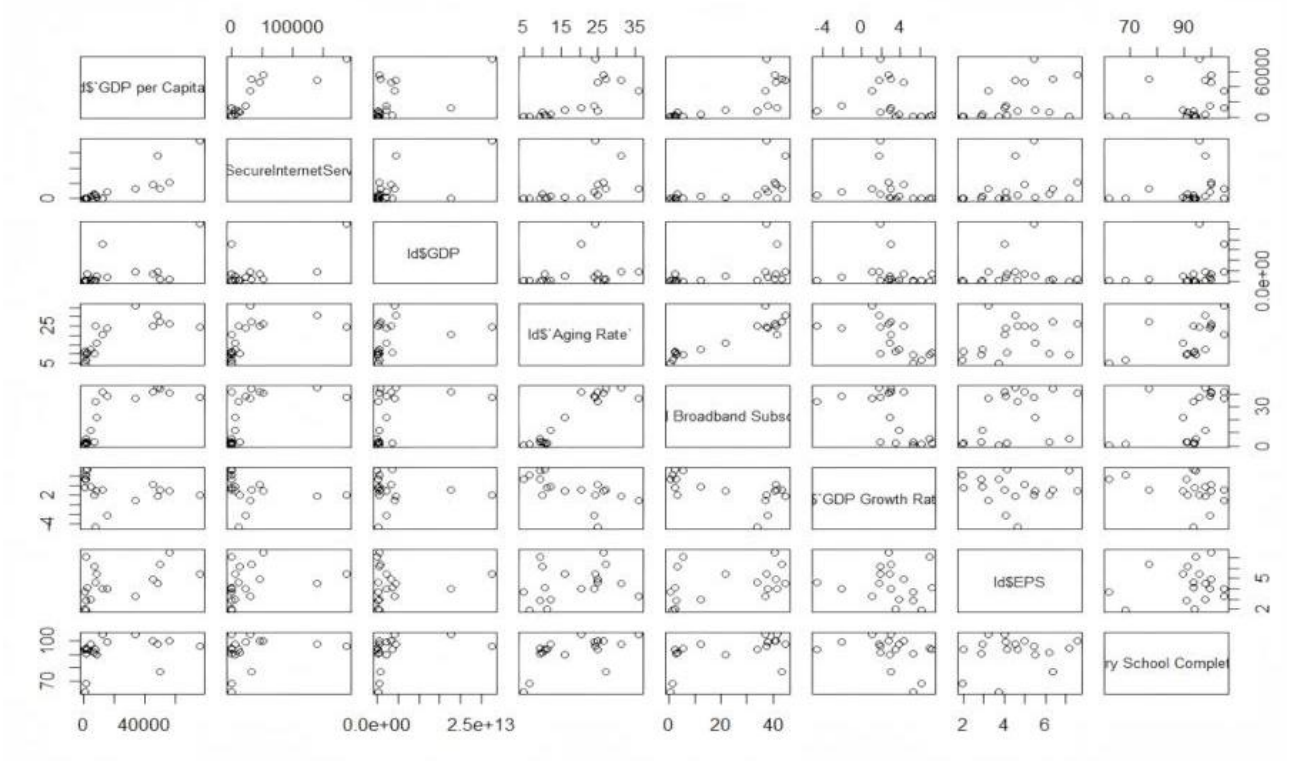

Figure 2. Scatter Diagram Matrix

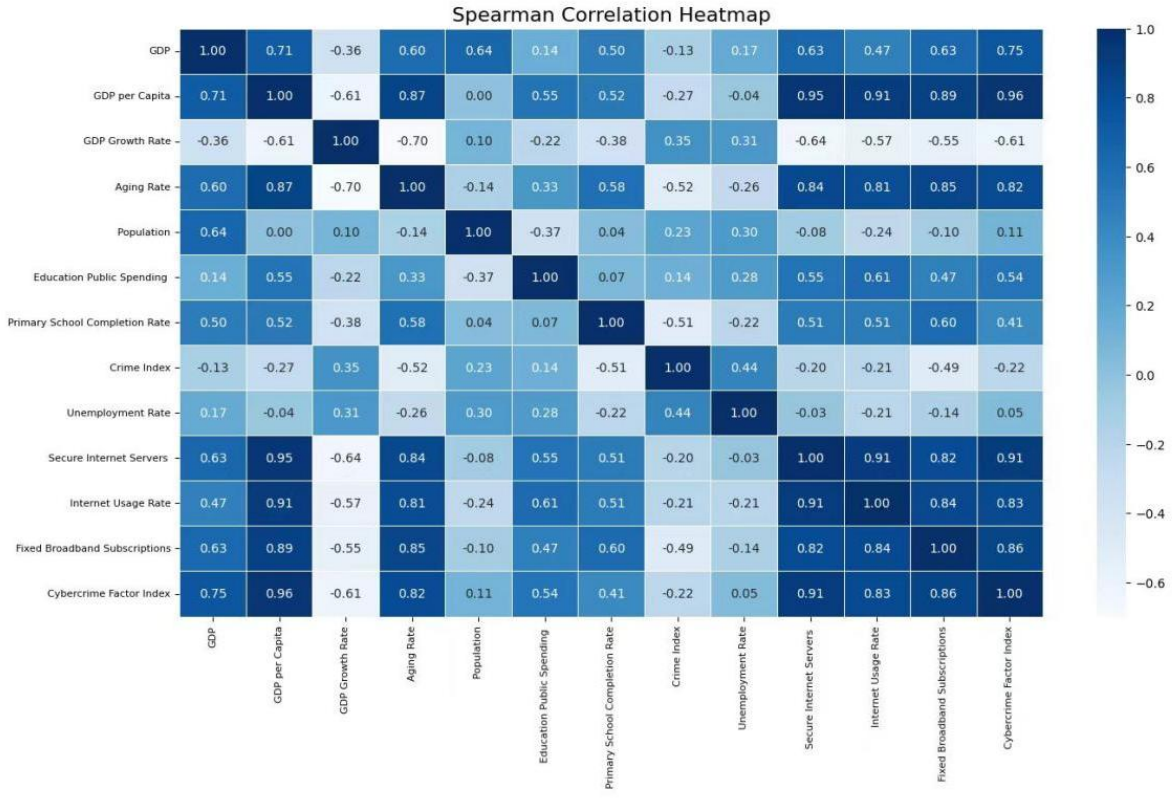

Figure 3. Spearman Correlation Analysis

Scatter Diagram Matrix and Spearman Correlation Analysis are shown in figure 2 and figure3, respcectively. Strong Linear Correlation: GDP per Capita, GDP, Aging Rate, and three Internet-related indicators (Secure Internet Servers, Internet Usage Rate, Fixed Broadband Subscriptions).

Weak Linear Correlation: GDP Growth Rate, Growth Rate of Public Education Expenditure as a Percentage of GDP, Primary School Graduation Rate.

Almost No Linear Correlation: Population, Crime Index, Unemployment Rate.

2.3.3. Conclusion

Reasons for Strong Linear Correlation: Countries with high GDP and GDP per Capita invest heavily in technology and infrastructure, leading to widespread Internet usage. Frequent online activities increase opportunities for cybercrimes—especially in aging populations, where individuals often have weaker online security awareness. Additionally, complex and active network ecosystems further amplify cybercrime risks.

Reasons for Weak Linear Correlation:

GDP growth reflects economic growth rate but not directly cybercrime rates, as the latter depend on supporting factors like security measures and legal frameworks. Improvements in basic education, while socially significant, do not directly prevent cybercrimes (which require specialized technical knowledge).

Reasons for Almost No Linear Correlation:

Population size has no clear link to cybercrimes, as demographics and Internet usage patterns vary across countries. Traditional crime rates differ fundamentally from cybercrimes, and unemployed individuals often lack the technical skills or motives to commit cybercrimes.

2.4. Machine Learning - XGBoost

2.4.1. Why This Research Chooses XGBoost

(1) Handling Linear and Non-Linear Relationships: The Spearman's coefficient test confirms both linear and non-linear relationships between indicators and cybercrime distribution. XGBoost is selected for its ability to process both relationship types simultaneously.

(2) Feature Selection Capability: XGBoost automatically assigns weights based on feature importance and ignores unimportant features, eliminating the need for manual removal of irrelevant variables.

(3) Interpretability: Feature importance analysis in XGBoost helps clarify the impact of each variable on the target variable (cybercrime distribution).

(4) Data Preparation: After initial data cleaning, the dataset is split into a training set and a test set. To avoid model overfitting, 80% of the data is used as the training set.

(5) XGBoost Model Construction: The construction process includes the following key steps:

Define the target variable (cybercrime factor index) and feature variables (national demographic indicators).

Adjust hyperparameters (e.g., tree depth, learning rate, subsample ratio) to optimize model performance.

Use grid search or cross-validation to select the optimal hyperparameters.

Core Components of the XGBoost Model:

Loss Function: XGBoost uses a loss function to measure the error between predicted and true values:

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k) \tag{5}$$

Where:$l(y_i, \hat{y}_i)$ is the error loss function (e.g., Mean Squared Error (MSE) or Logarithmic Loss (Logloss)).

$\Omega(f_k)$ is a regularization term for model complexity, which prevents overfitting.

Parameter Update via Taylor Expansion: XGBoost performs a second-order Taylor expansion on the objective function and uses first-order and second-order derivatives to update model parameters:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \text{ (First-order derivative)} \tag{6}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \text{ (Second-order derivative)} \tag{7}$$

(6) Model Training and Prediction:

Train the model using the training set.

Evaluate model performance on the test set using metrics including MSE (Mean Squared Error) and R^2 (Coefficient of Determination).

(7) Feature Importance Analysis: Extract feature importance scores from the trained XGBoost model to identify which variables have the greatest impact on cybercrime distribution.Feature Importance is shown in figure 4.

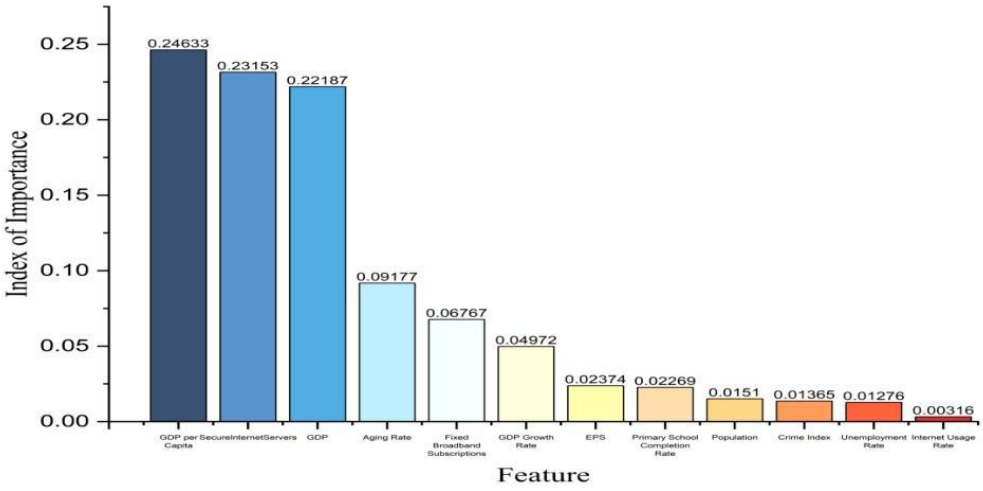

Figure 4. Feature Importance

2.4.2. Model Analysis and Evaluation

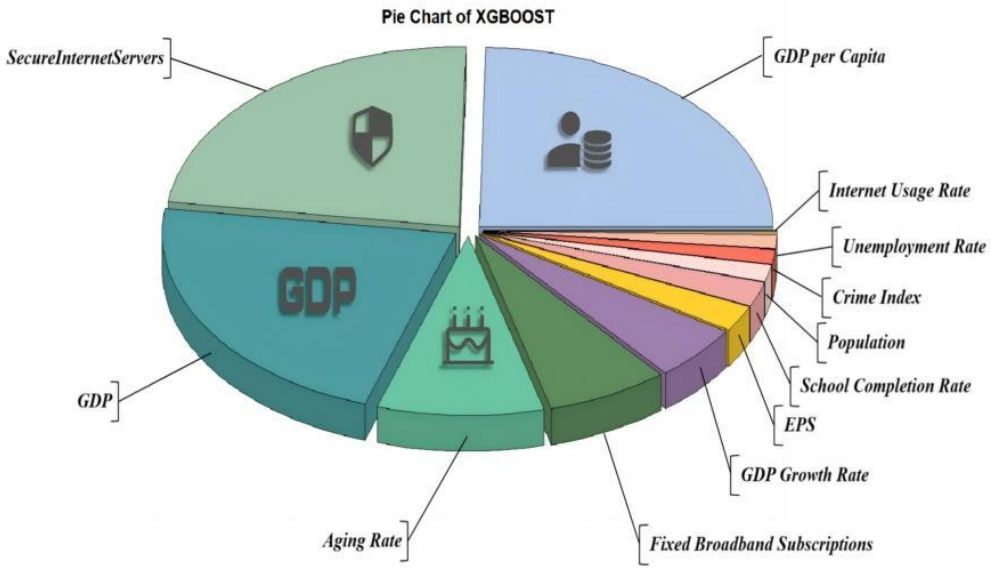

Figure 5. Comparison of Indicators Under the XGBoost Model

Comparison of Indicators Under the XGBoost Model is shown in figure 5.

(1) Interpretation of Results:

High-Impact Indicators:

GDP per Capita, Secure Internet Servers, and GDP have the most significant impacts on the cybercrime factor. High GDP per Capita correlates with active online financial activities (creating greater economic incentives for cybercrimes) and higher reliance on networks. Secure Internet Servers store high-value data, making them prime targets for attacks. High GDP reflects active economic activity and advanced technology, which indirectly increases cybercrime opportunities.

Low-Impact Indicators:

Indicators such as the Aging Rate, Fixed Broadband Subscriptions, and GDP Growth Rate have minimal impacts. The elderly have limited online activity, reducing their influence on cybercrime trends. Fixed Broadband providers often implement security measures that curb risks. GDP Growth Rate has no direct link to cybercrimes. Other indicators (Earnings per Share, Primary School Graduation Rate, Population, Crime Index) also show weak correlations. While unemployment may create motivation for crimes, cybercrimes require specialized skills. Internet Usage Rate does not directly equate to cybercrime probability, as security measures and user awareness also play critical roles.

(2) Model Optimization and Evaluation:

Sensitivity Analysis:

Based on the XGBoost model results, sensitivity analysis is conducted by applying small perturbations (±10% of the standard deviation) to all features to assess their impact on prediction outcomes. Key findings:

GDP per Capita and GDP are the most sensitive features—perturbations to these indicators significantly increase the Mean Squared Error (MSE), indicating the model's strong dependence on them and their critical importance to the target variable.

Most other features have minimal impacts, with negligible changes in MSE, demonstrating the model's stability and robustness to feature fluctuations.

Overall, the model is robust to variations in most features, but high-sensitivity features require special attention. It is recommended to enhance data preprocessing and improve data collection quality for these indicators to ensure stable and reliable predictions. Sensitivity Analysis of Features is shown in figure 6.

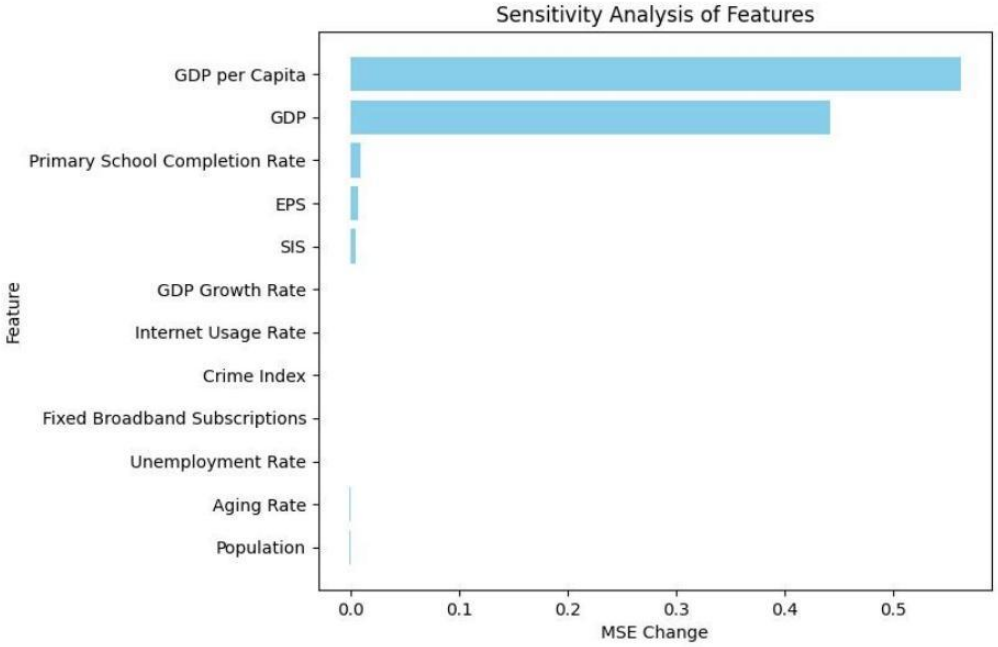

Figure 6. Sensitivity Analysis of Features

Error Analysis of the XGBoost Model: The XGBoost model achieves an MSE of 0.339 (indicating high prediction accuracy) and an R^2 of 0.744 (meaning the model explains 74.4% of the variance in the target variable). The low MSE and high R^2 demonstrate the model's precision and strong explanatory power. Combined with its robust performance under feature perturbations, the model exhibits excellent generalization ability and practical application value.

3. Conclusions

This study systematically evaluated the distribution characteristics of cybercrime, policy effectiveness, and the influence of national demographic factors through a multi-model approach. Comprehensive analysis concludes that regarding global cybercrime distribution, GDP per capita, the number of secure Internet servers, and GDP are the most significant national characteristics affecting cybercrime risk. These economic and technological indicators reflect the value of potential targets and the degree of reliance on the Internet. In terms of policy dimensions, the GCI's "Capacity Building" and "Technical Measures" were proven to be the factors with the greatest impact on national cybersecurity levels. Capacity Building strengthens long-term security defenses through incentive mechanisms and industry development, while Technical Measures provide immediate protection through advanced solutions. The research findings emphasize that when formulating cybersecurity policies, countries should prioritize capacity building and technical investment, while necessarily considering their own demographic characteristics to counter evolving cyber threats. Given the transnational nature of cybercrime, international cooperation, especially technical and capacity support from developed countries to developing and less developed nations, is crucial for collectively enhancing global cybersecurity.

References

[1] Zhang Di. The Logic and Direction of Judicial Governance Regarding the Difficulties in Proving Cybercrimes: An Examination of Judicial Interpretations and Normative Documents from 2004 to 2024[J]. Peking University Law Journal, 2025, 37(3): 605-624.

[2] Cao Jie. Revisiting the Theory of Restricted Accessoriality in the Crime of Assisting Information Network Criminal Activities[J]. Journal of Yibin University, 2025, 25(5): 49-59.

[3] Li Xiaolu. The "Production Order" Clause in the UN Convention on Cybercrime and Its Domestic Legal Response[J]. Evidence Science, 2025, 33(3): 307-319.

[4] Hu Zongjin. A Systematic Examination of the Principal Offender Status of Aiding Behaviors in the Realm of Cybercrime[J]. Jingchu Law Review, 2025(3): 39-49

[5] Xu Yang, Yang Yuan. A Study on the Legalization of Collaborative Governance of Cybercrime[J]. Journal of Shenyang Normal University (Social Science Edition), 2025, 49(3): 35-42.

[6] Dong Yuelin. A Review of Cybercrime Research in China over the Past Five Years[J]. Journal of Shanxi Police College, 2025, 33(2): 104-112.

[7] Zhang Yafei, Li Wenjing. On the Re-identification and Judicial Application of the Types of the Crime of Assisting Information Network Criminal Activities[J]. Journal of Kashgar University, 2025, 46(2): 28-36.

[8] Xie Dengke. Personal Information Protection in Cross-Border Electronic Data Acquisition: From the Perspective of Article 36 of the UN Convention on Cybercrime[J]. Journal of Shanghai University of Political Science and Law (Tribune of Law), 2025, 40(3): 63-79.

[9] Li Xiang, Cao Jiyuan. Development Trends, Generative Mechanisms, and Governance Strategies of Cybercrime in the Web3.0 Era[J]. Journal of Guizhou Police College, 2025, 37(3): 89-99.

[10] Lao Dongyan. The Protected Legal Interest of the Crime of Assisting Information Network Criminal Activities[J]. Legal Forum, 2025, 40(2): 5-16.