Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 1, 2025

Research on Global Cybercrime Prediction Based on Regression Analysis and Machine Learning

Guanyi Hao*

School of Criminal Investigation, People's Public Security University of China, Beijing 102623, China
* **Corresponding author**: Guanyi Hao (Email: 15092749283@163.com)

Abstract: This article establishes a comprehensive model for evaluating cybersecurity policies, revealing the global distribution pattern of cybercrime through data-driven analysis, and exploring which national policies have performed well in preventing, combating, and prosecuting cybercrime. By analyzing the global Cybersecurity Exposure Index (CEI) and the heat map of cybercrime distribution, it reveals that some countries in Europe, Asia, and the Americas face higher cybercrime risks due to advanced information technology and frequent online activities, while some countries in Africa and Oceania have relatively fewer cybercrimes due to lower digitalization levels. In addition, through the gray correlation analysis (GRA) model, the study found that political stability, economic development and cultural factors are significantly associated with cybercrime. This article develops a network security policy evaluation model. The decision tree regression model is used to evaluate the impact of different network security policies. The results indicate that organizational measures have the highest characteristic importance (66.60%), followed by technical measures (18.00%). Although legal measures are important, their impact is relatively small (4.00%). This indicates that effective organizational strategies and technological implementation are crucial for reducing cybercrime.

Keywords: Network security governance, decision tree regression model, multiple linear regression model, policy effectiveness evaluation.

1. Introduction

With the rapid development of information technology around the world, cybercrime is gradually becoming an increasingly serious global problem. Not only do cybercrimes cross national borders, complicating investigations and prosecutions, but many victims often choose to conceal their attacks, further increasing the invisibility of the problem. To address this challenge, countries have developed and implemented cybersecurity policies, but there are significant differences in the effectiveness of the policies. The International Telecommunication Union (ITU) has played an important role in promoting global cybersecurity cooperation and standards development. In the future, global cooperation and continuously improved cybersecurity policies will be the key to effectively responding to cybercrime. Qian et al[1] conducted a comprehensive survey on 3D object detection for autonomous driving, highlighting the importance of advanced detection techniques in ensuring safety and efficiency. Their work underscores the role of technology in addressing complex real-world problems, which parallels our focus on using technological measures to combat cybercrime. The methodologies discussed in their survey, such as pattern recognition and machine learning, are directly applicable to our analysis of cybersecurity policies and their effectiveness. Mafla et al[2]. explored real-time lexicon-free scene text retrieval, emphasizing the importance of efficient data processing and retrieval systems. Their research on pattern recognition and text detection aligns with our need to analyze large datasets related to cybercrime and cybersecurity policies. The techniques they employed, such as stroke width transform (Epshtein et al.)[3] and word spotting in natural scenes (Wang & Belongie)[4], provide a foundation for our data preprocessing and feature extraction processes, particularly in handling unstructured data from various sources. Xu and Ding[5] discussed the development trends and global collaboration in cybercrime governance, emphasizing the need for international cooperation to address the transnational nature of cybercrime. Their work supports our analysis of how different countries' cybersecurity policies, particularly those involving international cooperation, impact the effectiveness of cybercrime prevention. Bai[6] further elaborated on the theoretical support for establishing a global cybercrime convention under the United Nations framework, which aligns with our exploration of legal and organizational measures in cybersecurity policies.

These studies collectively highlight the importance of technological, legal, and organizational measures in addressing cybercrime. They provide a theoretical and methodological foundation for our research, particularly in understanding the global distribution of cybercrime, the effectiveness of national cybersecurity policies, and the role of demographic factors in shaping cybercrime trends. By integrating insights from these works, we aim to develop a comprehensive model that not only identifies effective cybersecurity policies but also provides actionable recommendations for policymakers.

2. The Analysis of Cybercrime Model

2.1. 2020 Global Cybersecurity Exposure Index Heatmap

The Cybersecurity Exposure Index (CEI) is a metric used to assess the level of exposure to cybercrime across countries/regions, with a range from 0 to 1, where a higher index indicates a greater degree of cyber exposure in the region.

From a geographical perspective, certain European countries, such as Russia, exhibit a relatively high CEI due to their advanced information technology and frequent network

activities. In Asia, China, India, and some countries in the Middle East face significant cybersecurity challenges owing to their large internet user bases and vibrant online economic activities. In the Americas, the United States sees an increasing trend in cybersecurity exposure risk due to its rich network activities, while Brazil experiences a similar uptick attributable to its rising internet penetration and rapid digitalization process. Compared to other regions, most African countries have a relatively low level of digitalization, resulting in a correspondingly lower CEI. In Oceania, Australia has a relatively high CEI due to its high level of digitalization and network activity; whereas other island nations, with their relatively lower levels of digitalization, maintain a lower CEI, as shown in Figure 1.



Figure 1. 2020 Global Cybersecurity Exposure Index Heatmap

2.2. Global Cybercrime Distribution Heatmap

From a regional perspective, Europe emerges as a hotspot for cybercrime, with notably high concentrations in certain countries. This could be attributed to the region's advanced technological infrastructure and the high frequency of network activities, which may provide more opportunities for cybercriminals to exploit vulnerabilities. Specifically, countries like Russia, as indicated in previous discussions, stand out due to their high levels of information technology development and network engagement.

In Asia, the situation is equally concerning, with China and India facing significant cybersecurity challenges. The vast internet user bases in these countries, combined with the dynamic and rapidly growing online economic activities, create a fertile ground for cybercriminal activities. Moreover, parts of the Middle East also exhibit high levels of cybercrime, potentially linked to similar factors as those in Asia.

The Americas also show a considerable presence of cybercrime, particularly in the United States and Brazil. The United States, with its rich network activities, and Brazil, with its increasing internet penetration and rapid digitalization, both contribute to the region's heightened cybersecurity exposure risk.

In contrast, Africa appears to have relatively lower levels of cybercrime, which could be attributed to the continent's generally lower level of digitalization. This suggests that, while digitalization brings numerous benefits, it also introduces new cybersecurity challenges that may not yet be fully manifested in less digitized regions.

Lastly, in Oceania, Australia stands out with a relatively high incidence of cybercrime, likely due to its high level of digitalization and network activity. Other island nations in the region, however, exhibit lower levels of cybercrime, potentially reflecting their lower levels of digitalization compared to Australia.

Overall, this heatmap provides valuable insights into the global distribution of cybercrime, highlighting regions and countries that may require additional focus and resources to enhance their cybersecurity defenses, as shown in Figure 2.

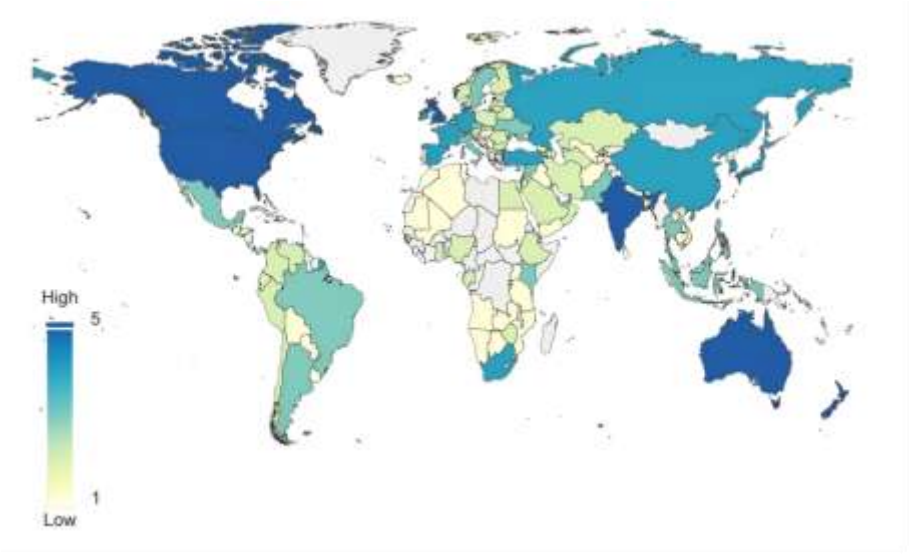


Figure 2. Global Cybercrime Distribution Heatmap

2.3. Grey Relational Analysis (GRA) model

2.3.1. Calculate the Grey Relational Coefficients

Compute the grey relational coefficients between the cybercrime sequence and each of the political, economic, and cultural factor sequences. The grey relational coefficient measures the degree of similarity or correlation between two sequences. It is calculated using a specific formula that takes into account the difference between corresponding elements of the sequences and a resolution coefficient that controls the sensitivity of the analysis.

$$\overline{D}_i(t) = \frac{\min\limits_i \min\limits_t \left|X_0(t) - X_i(t)\right| + \beta \max\limits_i \max\limits_t \left|X_0(t) - X_i(t)\right|}{\left|X_0(t) - X_i(t)\right| + \beta \max\limits_i \max\limits_t \left|X_0(t) - X_i(t)\right|} \tag{1}$$

2.3.2. Determine the Grey Relational Degrees

Aggregate the grey relational coefficients for each factor sequence to obtain the grey relational degree. This represents the overall correlation between cybercrime and the respective factor.

$$\gamma_i = \frac{1}{n}\sum_{i=1}^{n} \overline{D}_i(t) \tag{2}$$

2.3.3. Interpret the Results

Finally, interpret the grey relational degrees to understand the relative importance of each factor in relation to cybercrime. Factors with higher grey relational degrees indicate a stronger correlation with cybercrime, suggesting that changes in these factors may have a more significant impact on cybercrime levels.

3. Cybersecurity Policy Assessment Model

3.1. The Establishment of Model

Decision tree regression is a regression analysis method that predicts continuous numerical outputs based on the structure of a decision tree. It divides the dataset into different regions through a series of decision rules and then makes predictions within each region. Each decision node divides the data into different branches based on the value of a particular feature until it reaches the leaf node. In a regression problem, the value of the leaf node is usually the mean or median of the target variable for that region. The core idea of decision tree regression is as follows:

Dividing the data set: a decision tree selects a feature (e.g. X1, X2,..., Xn) and a cut-off point (e.g. t) at each node to divide the data into different subsets. The goal is to divide the data into the most meaningful subsets based on a certain feature that makes the target value (e.g. y) in each subset more similar.

Predicted value: in a regression problem, the predicted value of a leaf node is usually the average of the objective values of all samples in that node.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

N is the number of samples in the leaf node and y_i is the objective value of the i th sample.

Selecting the best features: In order to select the best features for data partitioning, decision tree regression usually uses the Mean Squared Error (MSE) as a measure. For each feature X_j and each possible cut point t, calculate the mean square error before and after the segmentation, and select the segmentation method that minimizes the mean square error. Specifically, the MSE is calculated as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{4}$$

$\hat{y}$ is the predicted value, y_i is the real value of the sample.

Based on the Global Cybersecurity Index (GCI) and the number of global cybercrime cases, we form a decision tree regression model covering the five dimensions of the GCI, i.e., legal, technological, organizational, capacity-building and cooperation, reflecting the direction and implementation of each country's cybersecurity policy, and the average number of cybercrime crimes based on IP behaviors, in order to find out the most effective and critical dimensions, thus providing reference and guidance for the formulation of effective cybersecurity policies.

3.2. The Solution of Model

Based on the calculation of the feature importance of the decision tree regression model, we get the correlation between each dimension of the Global Cybersecurity Index (GCI) and the number of cybercrimes with different degrees of influence, and the results of the calculation are shown in the figure 3 below:

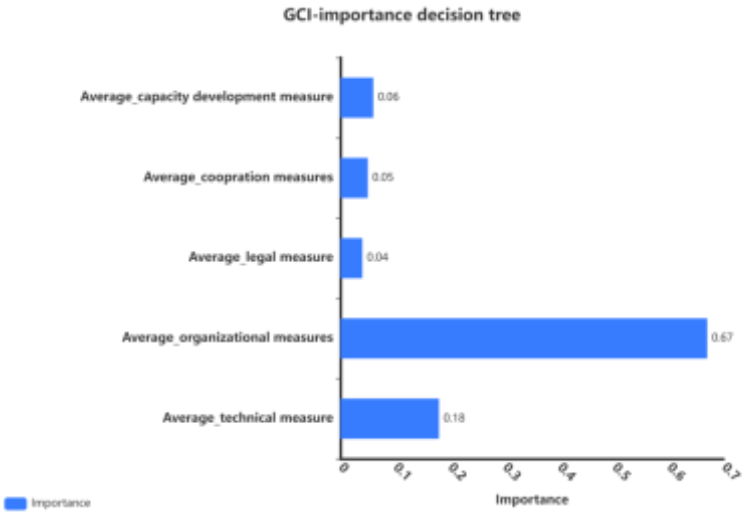


Figure 3. GCI-importance decision tree figure

Based on this data, we can see that organizational measures with 66.60% feature importance tops the list, indicating that organizational level measures have more significant impact on avg_per_year (average number of cybersecurity incidents per year). In contrast, technical measure ranked second with 18.00% importance, indicating that technical-level metrics also had a more significant impact on avg_per_year (average number of cybersecurity incidents per year). This suggests that effective organizational cooperation and the application of technology can play a key role in reducing cybersecurity crimes.

The characteristic importance of legal measure is 4.00%, which is in the lower position, implying that legal-level measures play a limited role in reducing cybersecurity crimes. Cooperation measures and capacity development measures are also relatively less important in reducing cybercrime, indicating that cooperation and capacity development measures contribute to the reduction of cybercrime, but their influence is less significant than that of organizational cooperation and technological applications, or perhaps the three indicators mentioned above are not significant enough in this dataset.

4. Demographic-Cybercrime Impact Regression Model

4.1. The Establishment of Model

4.1.1. Heat Matrix Diagrams

A heat matrix plot is a visualization tool that shows the correlation between multiple variables through color. It centers on the Pearson correlation coefficient, which is used to measure the linear relationship between two variables. The Pearson correlation coefficient is calculated by the formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (5)$$

X_i and Y_i are the observed values of the two variables, and $\bar{X}$ and $\bar{Y}$ are their means, and the value of E ranges from -1 to +1. Values of E close to +1 indicate a strong positive correlation, close to -1 a strong negative correlation, and close to 0 no significant linear relationship.

In the heatmap, the A-axis and B-axis represent the different variables, and each cell in the matrix shows the correlation coefficient between the corresponding variables. The color shades in the graph are used to reflect the strength of the correlation: warm colors indicate positive correlation, cold colors indicate negative correlation, and neutral colors indicate weak correlation or irrelevance. The heat matrix plot allows the strength and direction of relationships between variables to be visually identified, thus helping to uncover underlying patterns and structures in the data, shown in the figure 4 below.

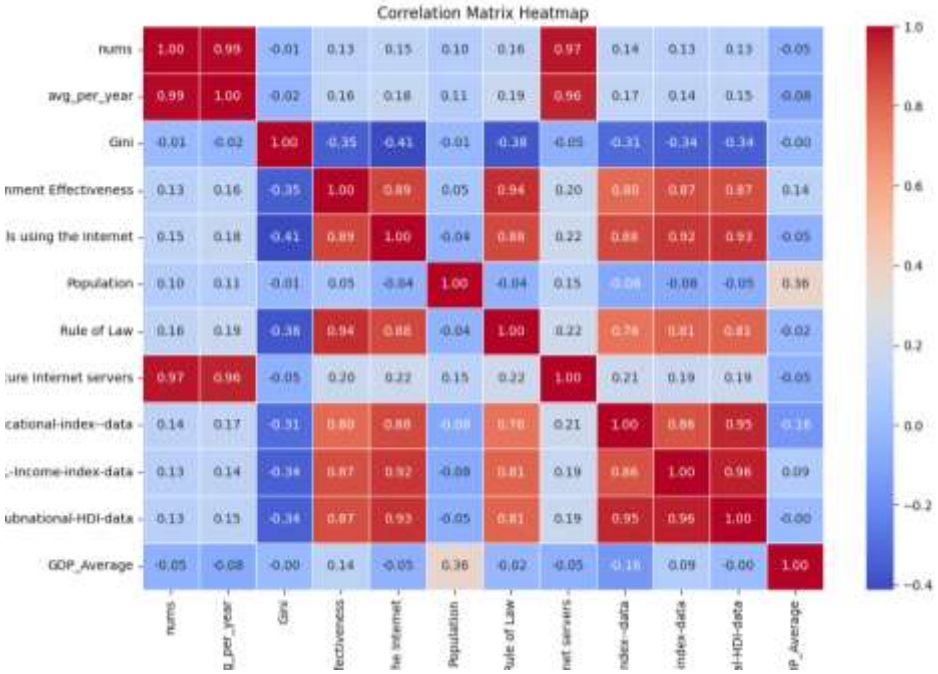


Figure 4. Correlation Matrix Heatmap

This matrix distribution graph shows the correlation between several variables. Each small cell represents the correlation coefficient of the two variables, helping us to visualize the correlation between them. Correlation coefficients can take values from -1 to 1. The closer the correlation coefficient is to 1 or -1, the stronger the linear relationship between the two variables; closer to 0 means there is little or no linear relationship. The color in the heat map indicates the magnitude of the correlation coefficient, usually red indicates a positive correlation, blue indicates a negative correlation, and the darker the color, the stronger the correlation. Through these plots, we can observe the correlation between variables, such as the strength of significant linear relationships or correlations between certain variables.

For example, the correlation between "Future Internet servers" and other variables (e.g., "avg_per_year") may show some correlation, indicating that there is some correlation between the use of the Internet and the occurrence of cybercrime. Some correlation between Internet usage and the occurrence of cybercrime.

This type of graph is suitable for use in the pre-processing stage of the data to help identify the strength of the correlation between the variables, which in turn lays the groundwork for the next step of in-depth analysis.

4.1.2. Regression analysis

Multiple linear regression is a method in regression analysis that is used to study the linear relationship between multiple independent variables (predictor variables) and a dependent variable (response variable). The basic idea is to describe how the dependent variable is affected by multiple independent variables by building a linear model.

The mathematical model for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (6)$$

Y is the dependent variable; $X_1, X_2, \ldots, X_n$ is the independent variable; β_0 is the intercept (the value of dependent variable when all independent variables are zero); $\beta_1, \beta_2, \ldots, \beta_n$ is the regression coefficient, which indicates how much each independent variable affects the dependent variable;ϵ is the error term, which indicates the part of the model that cannot be explained.

The regression coefficients reflect the magnitude and direction of the effect of each independent variable on the dependent variable. Positive values indicate that the independent variable is positively correlated with the dependent variable, while negative values indicate a negative correlation. The larger the absolute value of the regression coefficient, the stronger the effect of the independent variable on the dependent variable.

Intercept β_0: The intercept is the predicted value of the dependent variable when all values of the independent variables are zero.

Error term ϵ: The error term represents the difference between the model's predictions and the actual values, and it is usually assumed that the error terms are independent and follow a normal distribution.

In multiple linear regression, the most commonly used method is Ordinary Least Squares (OLS), whose goal is to minimize the sum of squares of the errors between the predicted values of the model and the actual observations by finding a set of regression coefficients. In other words, minimize the following objective function:

$$\text{Minimize} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \quad (7)$$

Y_i is the actual value, and $\hat{Y}_i$ is the value predicted by the model.Multiple linear regression is based on a number of assumptions, including:Linear relationship: there is a linear relationship between the dependent and independent variables.Independence of Error: The error terms for each observation are independent of each other.Normal distribution of error: the error term follows a normal distribution.Homoskedasticity: the error term has a constant variance.The goodness of a multiple linear regression model can be assessed by several statistical indicators: R^2 : coefficient of determination, which indicates the proportion of variation in the dependent variable explained by the independent variables. The closer the value is to 1, the better the fit of the model. Adjusted R^2: adjusted while considering the number of independent variables, to avoid overfitting. P-value: Used to test whether the regression coefficients are significant or not. If the P-value of an independent variable is

less than the significance level (usually 0.05), it means that the independent variable has a significant effect on the dependent variable.

5. Conclusions

This study relies on multiple machine learning models to deeply analyze the numerous factors influencing the occurrence of cybersecurity incidents, and integrates different regression analysis techniques to successfully uncover the complex relationships between cybersecurity incidents and various influencing factors. Identification of key features and the importance of Grey Relational Analysis: By applying Grey Relational Analysis, the study clarifies the significant roles of key features such as national economic development level, government governance effectiveness, legal enforcement capability, and technological innovation capability in predicting cybersecurity incidents. Among numerous models, the Grey Relational Analysis model demonstrates excellent predictive ability, especially in capturing nonlinear relationships between features. The results of its feature importance analysis provide a powerful basis for understanding which factors have a greater impact on cybersecurity incidents.

Quantitative analysis of influencing factors and the dominant role of technology: With the training results of decision trees and other models, it is found that technological factors are significant influencers of cybersecurity incidents, highlighting the crucial role of technological and organizational measures in reducing the frequency of cybersecurity incidents. Meanwhile, legal enforcement capability and government effectiveness also have certain importance, indicating that strong legal safeguards and government governance capabilities are equally indispensable for preventing cybersecurity threats. In addition, economic factors such as national income level and internet penetration rate have a negative impact on the occurrence of cybersecurity incidents, meaning that high-income countries and countries with higher internet penetration rates are relatively more likely to take effective measures to prevent and respond to cybersecurity incidents.

References

[1] R. Qian, X. Lai, X. Li, 3D object detection for autonomous driving: A survey, Pattern Recognit. (2022) 108796, http://dx.doi.org/10.1016/j.patcog. 2022. 108796.

[2] A. Mafla, R. Tito, S. Dey, L. Gómez, M. Rusiñol, E. Valveny, D. Karatzas, Real-time lexicon-free scene text retrieval, Pattern Recognit. (2021) 107656.

[3] B. Epshtein, E. Ofek, Y. Wexler, Detecting text in natural scenes with stroke width transform, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, http://dx.doi.org/10.1109/cvpr.2010.5540041.

[4] K. Wang, S. Belongie, Word spotting in the wild, in: ECCV 2010, in: Lecture Notes in Computer Science, 2010, pp. 591–604.

[5] XU Zhanpeng, DING Libai. International Cooperation on Cybercrime Governance: Development Trend, Global Collaboration and Chinese Program[J]. Academic Forum, 2024, 47(2): 123-133.

[6] Bai Songyu. Theoretical support and rule conception of establishing a global cybercrime convention under the framework of the United Nations[J]. Journal of Economic Research, 2022(26):158-160.