Research on the regularity of network security crimes based on mathematical models

Luze Yang *

Northwestern Polytechnical University, Xi'an, China

* Corresponding Author Email: ylz@mail.nwpu.edu.cn

Abstract. In order to address the current lack of universal research on network security issues in literature, and the absence of tailored research for different countries and regions, this paper conducts research using clustering analysis models, ARIMA time series prediction models, decision tree prediction models, and PLSR models. This article first collects data from authoritative platforms such as VCDB database to ensure the accuracy of input data, and then visualizes the data through heat maps to observe initial patterns. To verify the hypothesis, this paper adopts the K-means clustering model and uses the elbow rule to determine four clusters as the "elbow points" of the dataset. This article chooses GCI as the indicator for evaluating national policies. Considering the uniqueness of policy implementation, the data was normalized over time and the ARIMA model was used to predict the occurrence rate, while the decision tree model was used to determine the importance of each dimension. The results indicate that organizational measurement dimensions account for 88.5% of importance. Given that the data spans multiple countries and the sample data is limited, this article chooses to use the PLSR model together to alleviate this issue. Calculations show that GDP is the most closely related demographic feature to cybercrime, with an average value of 1.4202. Other demographic features show varying degrees of correlation.

Keywords: Cybercrime, Decision Tree, Policy analysis, Partial Least Squares Regression (PLSR) model.

1. Introduction

In existing literature, due to differences in the types, methods, and characteristics of victims of cybersecurity crimes in different countries and regions, research ignores these regional differences, and the research perspective is relatively single. Moreover, cybersecurity crimes involve multiple fields such as law, information technology, and psychology, and existing research often lacks interdisciplinary integration perspectives, making it difficult to fully reveal the patterns of crime. At present, the research on legal regulation and policy response to cybersecurity crimes is not deep enough, and many studies remain at the surface level, failing to propose effective legal countermeasures and targeted policy improvement plans for different countries and regions in different directions [1].

International organizations such as the United Nations, Interpol, and the World Economic Forum promote cross-border research on cybersecurity crimes and facilitate the formation of international standards and best practices. Government agencies fund and conduct research on cybersecurity crimes to better understand crime trends and develop corresponding defense strategies. National security agencies pay attention to the potential impact of cybersecurity crimes on national security and invest resources in monitoring and analyzing them.

Firstly, in order to determine the global distribution pattern of cybercrime, the article collected cybercrime data from authoritative institutions such as VCDB database, World Bank, and United Nations. Using this data, the article employed cluster analysis to explore the characteristics of cybercrime events in different countries. Secondly, the article quantified the effectiveness of cybercrime by first using the ARIMA model for prediction, removing the influence of other factors on cybercrime, and then using decision tree prediction to obtain feature importance in different dimensions. Later, the article collected data on GDP, education level index, Internet access and government efficiency of countries in the past decade. After processing, a PLSR regression model was established to analyze the correlation between cybercrime incidents and demographic data. Finally,

considering the potential risk of overfitting and instability in the optimized model, sensitivity analysis was conducted to test the stability, effectiveness, and reliability of the model.

The data collection standards and sharing mechanisms for cybersecurity incidents vary among countries and regions, making it difficult to unify and analyze global cybersecurity crime data. Different countries mostly adopt universal rules for the prevention of cybercrime, lacking a tendency to adapt to local conditions. With the rapid development of network technology and the updating of criminal methods, researchers often find it difficult to keep up with these changes, resulting in lagging prevention and crackdown strategies. The research on the activity patterns, economic systems, and criminal network structures of cybercrime is not deep enough.

Firstly, the article collected data from authoritative institutions such as VCDB, the World Bank, and the United Nations, and used these reliable data to conduct cluster analysis on most countries around the world. By utilizing the similarities within different clusters, the article can better unify the standards for dealing with cybercrime security issues in various countries. Secondly, to better analyze the differences in impact caused by different policy priorities, the article uses the five dimensions of the GCI index to quantify different policies into data, in order to better compare the impact weights of each dimension. Combining our previous research, the article can make targeted policy improvements for different regions with different characteristics and needs. Finally, considering that the problem of cybercrime is influenced by multiple factors, the article focused on predicting and verifying the relationship between demographic and socio-economic characteristics. This not only effectively solves the problem of cybercrime, but also fundamentally reduces the impact of existing problems on the existing economy.

Empirical results demonstrate that our framework analyzes the regularity of cybercrime and quantifies the impact of policies in various countries by developing and optimizing appropriate models, GDP The relationship between demographic characteristics and cybercrime. This allows us to clearly analyze which policies have a significant impact on reducing cybercrime. These findings can help different government agencies develop more targeted and reasonable prevention policies, thereby promoting the establishment and improvement of national cybersecurity mechanisms.

Our contributions can be summarized as follows:

We first collected cybercrime data from the VCDB database and the World Bank, and using this data, we created a global map of cybercrime heat distribution.

We used K-means clustering analysis to explore the characteristics of cybercrime events in different countries and determined suitable clusters through the elbow rule.

Using ARIMA model to predict existing data, with the aim of removing the influence of other factors on cybercrime.

We use GCI parameters for decision tree prediction, and evaluate the importance of each parameter's impact on the characteristics of cybercrime.

We established a PLSR(Partial Least Squares Regression) regression model to analyze the correlation between cybercrime incidents and demographic data.

2. RELATED WORK

2.1. K-means Clustering

K-means clustering analysis is a machine learning algorithm used to divide a set of data points into K clusters, where the data points within each cluster are as similar as possible, while the data points in different clusters are as different as possible. The K-means algorithm is an iterative algorithm that aims to minimize the squared error within a cluster, which is the sum of squared distances between data points and their cluster centers [2]. This algorithm is easy to understand and implement, and can handle large datasets. The article first selects K data points as the initial cluster centers. Next, for each data point, the article calculates its distance from the center of each cluster and assign it to the cluster represented by the nearest cluster center. Afterwards, we recalculate the center point of each cluster. The cluster center is the mean of all data points within the cluster, and the above steps are repeated

until the stopping condition is met. The article used an iterative algorithm, but this algorithm converges quickly and is also fast for large datasets.

2.2. ARIMA Model

The ARIMA model (Autoregressive Integrated Moving Average Model) is a statistical model for predicting time series data. It consists of three parts: autoregression (AR), differencing (I), and moving average (MA), and is suitable for predicting and analyzing trends, seasonality, and random fluctuations in time series data. It can determine the stationarity of a time series by drawing a time series graph, autocorrelation plot (ACF), and partial autocorrelation plot (PACF) to check the stationarity of the series., If the time series is non-stationary, it is stabilized by differencing [3]. The model uses maximum likelihood estimation to estimate the parameters of the ARIMA model, and tests the fitting effect of the model through residual analysis. Finally, the fitted ARIMA model is used to predict future values.

2.3. Decision Tree Model

The decision tree model is a commonly used machine learning algorithm that mimics the human decision-making process to predict data. A decision tree is a tree structure where each internal node represents a test on an attribute, each branch represents an output of the test, and each leaf node represents a class label. Its structure is intuitive, easy to explain and understand, and does not require any assumptions about the data, making it suitable for various types of data [4]. When constructing a decision tree, the first step is to select the optimal feature to segment the data set. Then, the above process is repeated for each subset until the stopping condition is met. The decision tree will automatically select important features and ignore irrelevant features during the construction process, making it suitable for processing large-scale data.

2.4. Partial Least Squares Regression (PLSR) model

PLSR is a multivariate statistical data analysis method mainly used to handle regression analysis with multicollinearity problems between variables. The core of PLSR is to extract latent components that can explain the covariance structure between the independent and dependent variables to the greatest extent possible. The extraction of components is achieved by optimizing the correlation between the independent and dependent variables, and unlike traditional linear regression, PLSR uses a bilinear model, which considers the linear combination of independent and dependent variables simultaneously. This means that PLSR considers not only the structure of the independent variable but also the structure of the dependent variable when extracting components. By extracting components, PLSR actually achieves dimensionality reduction of data. This helps to solve the problem of multicollinearity between independent variables and also improves the predictive ability of the model. When there is a high correlation between independent variables, traditional regression methods (such as ordinary least squares) may fail, while PLSR can effectively handle this situation.

3. METHOD

3.1. Preliminaries

The article denote the index of development as Z which is specified in J, the article denote coordinated index of development as CI which is specified in J, the article denote sustainability index of development as DI which is specified in J, the article denote economic index of development as A which is specified in J, the article denote social index of development as B which is specified in K, the article denote environment index of development as C which is specified in K, the article denote Impact index value as F which is specified in km^2,the article denote reality index value as S which is specified in m and the article denote influence coefficient as β which is specified in m^2.

3.2. Assumption

The cybersecurity policies implemented by countries (e.g., the legal, technical, and organizational measures reflected in the GCI score) directly impact the frequency and types of cybercrimes. In other words, countries with stronger policies will have lower cybercrime rates, while countries with weaker policies will experience higher rates of cybercrime. Equations.

All types of cybercrime incidents (such as data breaches, ransomware, fraud, etc.) have comparable impacts on the economy and society of each country, and can be uniformly measured and compared based on criteria like financial losses and number of victims.

The collected cybercrime data (such as that from VCDB) is complete, and all countries report cybercrime incidents according to the same standards. While some countries may have issues with concealment or low reporting rates, most will provide data in an open or reliable manner.

3.3. K-means Clustering

Using the elbow method previously, the article determined that the optimal number of clusters (K) is 4. Therefore, we selected 4 initial cluster centers. Calculate the distance between each data point and the centroid, and assign it to the closest centroid. the article uses the Euclidean distance to calculate the distance from each sample point to the center:

$$d(x, c_i) = \sqrt{\sum_{j=1}^{d} (x_j - c_{ij})^2} \tag{1}$$

x represents the data point, c_i represents the i-th cluster center, d represents the dimension of the data object, x_j and c_{ij} refer to the values of the j-th dimension of the data point and the cluster center, respectively.

The objective function of K-Means is:

$$J = \min \sum_{i=1}^{m} \sum_{k=1}^{K} \omega_{ik} \left\| x_i - c_k \right\|^2 \tag{2}$$

ω_{ik} is the indicator function. If a data point is within the clusterm, ω_{ik}=1, or ω_{ik}=0.

The goal is to minimize in order to achieve the optimal value, so we need to differentiate the two variables ω_{ik} and c_k. In this step, we first differentiate with respect to ω_{ik}:

$$\frac{dJ}{dc_k} = 2 * \sum_{k=1}^{K} \omega_{ik} \left\| x_k - c_k \right\|^2 \tag{3}$$

After the calculation, we obtain:

$$c_k = \frac{\sum_i^m \omega_{ik} x_i}{\sum_i^m \omega_{ik}} \tag{4}$$

This formula indicates that the new centroid c_k of the cluster is the weighted average of all data points x_i within the cluster, where the weights are the distances. This averaging process ensures that during the update, the centroid moves towards the center of the data points within the cluster. The process is repeated until the centroids no longer change or the maximum number of iterations is reached. the article performed cluster analysis by comparing the number of cybercrime incidents with crime success rates, reporting rates, and prosecution rates. Since the data for reporting rates and prosecution rates exhibit significant variations, the article used bar charts to better highlight their differences and characteristics.

3.4. ARIMA Model

The ARIMA model is a method used for analyzing and modeling various types of time series data. The time series to be predicted is generated by a random process. If the random process generating the series does not change over time, the structure of this random process can be precisely characterized and described. By utilizing the known past observations of the series, future values of the series can be extrapolated. In the ARIMA model, the general form of the model is as follows:

$$Y_t = c + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \cdots + \beta_q \varepsilon_{t-q} \tag{5}$$

3.5. Decision Tree Model

A Decision Tree is a classification and regression method based on the conditions required for various events to occur, forming a tree diagram to maximize the expected outcome. In a decision tree, the root node represents the entire dataset space, and each internal node represents a test of a single variable. The process of constructing a decision tree is as follows: First, identify the initial split. The entire training set serves as the set from which the decision tree is generated. All attribute domains are exhausted, and the quality of each attribute's split is quantified to determine the best possible split. Next, repeat the first step until each leaf node contains records that all belong to the same class, growing into a complete tree.

The tree uses the optimal features to split the data:

$$IG(D, A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} H(D_v)$$

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{6}$$

Here, H(D) represents the information entropy of the dataset D, Value(A)denotes the feature, D_v is the value of the feature A, and P_i is the probability of the i class within the subset v.In decision tree algorithms, the model selects splitting nodes by calculating the Gini index for different features, with the goal of making each child node as 'pure' as possible. The reduction in Gini impurity caused by the split of each feature in the dataset determines the importance of that feature. In other words, the lower the Gini index, the higher the purity of the data after the split, and the greater the contribution of the feature.

3.6. Partial Least Squares Regression (PLSR) model

Partial Least Squares Regression (PLSR) is a regression modeling method that handles multiple dependent variables and multiple independent variables, extending the least squares method. It is used to address the interdependencies between two sets of highly correlated variables and to examine the predictive relationship of independent variables with dependent variables [5]. PLSR combines the characteristics of principal component analysis, canonical correlation analysis, and linear regression analysis during the modeling process. As a result, it provides a more robust and reasonable regression model in the analysis.

For the regression problem with P dependent variables $y_1 \cdots y_p$ and m independent variables, the first component u_1 is extracted from the set of independent variables (u_1 as $x_1 \cdots x_m$ linear combination that maximizes the extraction of variance information from the original set of independent variables). At the same time, the first component v_1 is also extracted from the set of dependent variables, with the condition that it maximizes the correlation with u_1 and v_1. A regression model is then established between the dependent $y_1 \cdots y_p$ and independent u_1 variables. This process is repeated until a sufficient number of components have been extracted.

To use PLSR more accurately, x and y need to be standardized first:

$$A=\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}, B=\begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix} \tag{7}$$

Next, we extract the first pair of components from both the X and Y variable sets u_1 and v_1, ensuring that their correlation is maximized. The first component from both sets is assumed to be $[\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{m}]^\mathrm{T}$ a linear combination of the independent variables, denoted as $\mathrm{Y}=[\mathrm{y}_1,\ldots,\mathrm{y}_\mathrm{p}]^\mathrm{T}$ a linear combination of v_1:

$$\begin{cases} u_1 = \rho_1^T X \\ v_1 = \gamma_1^T Y \end{cases} \tag{8}$$

For the purposes of regression analysis, the following requirements must be met:

(1) u_1 and v_1 should extract as much variance information as possible from their respective variable groups;

(2) The correlation between u_1 and v_1 should be maximized.

Therefore, we first calculate ρ_1 and γ_1:We begin by maximizing the covariance to ensure the maximum correlation between u_1 and v_1, which can be computed using the dot product of the vectors $\hat{u}_1$ and $\hat{v}_1$:

$$\max\langle \hat{u}_1, \hat{v}_1 \rangle = \rho_1^T AB\gamma_1 \tag{9}$$

$$\mathrm{s.t.} = \begin{cases} \rho_1^\mathrm{T}\rho_1 = 1 \\ \gamma_1^\mathrm{T}\gamma_1 = 1 \end{cases} \tag{10}$$

Using the Lagrange multiplier method, the problem is transformed into unit vectors ρ_1 and γ_1 such that γ_1 is maximized. To solve the problem, the article only need to compute the eigenvalues and eigenvectors of $\mathrm{M} = \mathrm{A}^\mathrm{T}\mathrm{B}\mathrm{B}^\mathrm{T}\mathrm{A}$, with the largest eigenvalue of M being θ_1^2. The corresponding eigenvector is the desired solution for ρ_1, which in turn allows us to obtain γ_1, $\gamma_1 = \frac{1}{\theta_1}\mathrm{B}^\mathrm{T}\mathrm{A}\rho_1$. From the standardized observed data matrices X and Y for the two sets of variables, the scores for the first pair of components, denoted as $\hat{u}_1$ and $\hat{v}_1$, $\hat{u}_1=\mathrm{A}\rho_1$, $\hat{v}_1=\mathrm{B}\gamma_1$can be calculated. Then, we establish the regression of $\mathrm{y}_1,\ldots,\mathrm{y}_\mathrm{p}$ on u_1 and $\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{m}$ on u_1, assuming the regression model:

$$\begin{cases} A = \hat{u}_1\delta_1^T + A_1 \\ B = \hat{u}_1\gamma_1^T + B_1 \end{cases} \tag{11}$$

Here, $\sigma_1^T = [\sigma_1,\cdots,\sigma_m]$ and $\tau_1^T = [\tau_1,\cdots,\tau_m]$ are the parameter vectors in the many-to-one regression model, and A_1 and B_1 are the residual matrices. The least squares estimates of the regression coefficients σ_1 and γ_1 are:

$$\begin{cases} \sigma_1 = \dfrac{A^T\hat{u}_1}{\|\hat{u}_1\|^2} \\ \gamma = \dfrac{B^T\hat{u}_1}{\|\hat{u}_1\|^2} \end{cases} \tag{12}$$

Replace A and B with the residual matrices A_1 and B_1, and repeat the above steps until the absolute values of the elements in the residual matrices approximate zero. Each iteration yields a new σ_1 and γ_1.

Repeat the above steps to obtain:

$$\begin{cases} A = \hat{u}_1\delta_1^T + \cdots + \hat{u}_1\delta_r^T + A_r \\ B = \hat{u}_1\tau_1^T + \cdots + \hat{u}_1\tau_r^T + B_r \end{cases} \tag{13}$$

$$y_j = c_{j1}x_1 + \cdots + c_{jm}x_m, j = 1,2,\cdots,p \tag{14}$$

It gives the Partial Least PLSR equation for the P dependent variables

4. EXPERIMENTS

4.1. Data Analysis

Remove missing values: Check if there are missing values in the dataset. the article supplemented the missing value processing method with Lagrange interpolation to ensure the reliability and continuity of the data.

Correction of Outliers: Identify outliers in the data and use clustering analysis to handle them.

Data source integration: Data collected from multiple different sources (such as VCDB database, global network security index GCI, Internet penetration, social-economic characteristics, etc.) need to be integrated. Ensure that country, region, and other identifiers in various data sources are aligned during integration to avoid data misalignment caused by inconsistent names or units.

Correlation analysis: By calculating the correlation coefficient matrix and using a heatmap to analyze the correlation between various variables, identify key factors that affect the distribution of cybercrime and the effectiveness of cybersecurity policies.

Check data consistency: Ensure national data consistency for data from different countries. For example, for cybersecurity indices or crime data, confirm whether the scoring and statistical standards of each country are consistent, and whether there are differences in reporting methods among different countries.

4.2. Experimental Result

Firstly, the article selects a series of K values. For each K value, run the K-means algorithm and calculate the sum of squared errors (SSE). SSE reflects the closeness of clustering, with larger K values typically resulting in smaller SSE values because the cluster centers are closer to the sample data points. Plot the K value and its corresponding WSS on a graph, with the X axis representing the K value and the Y axis representing Subfigures and Tables. The result is shown in the Figure 1:

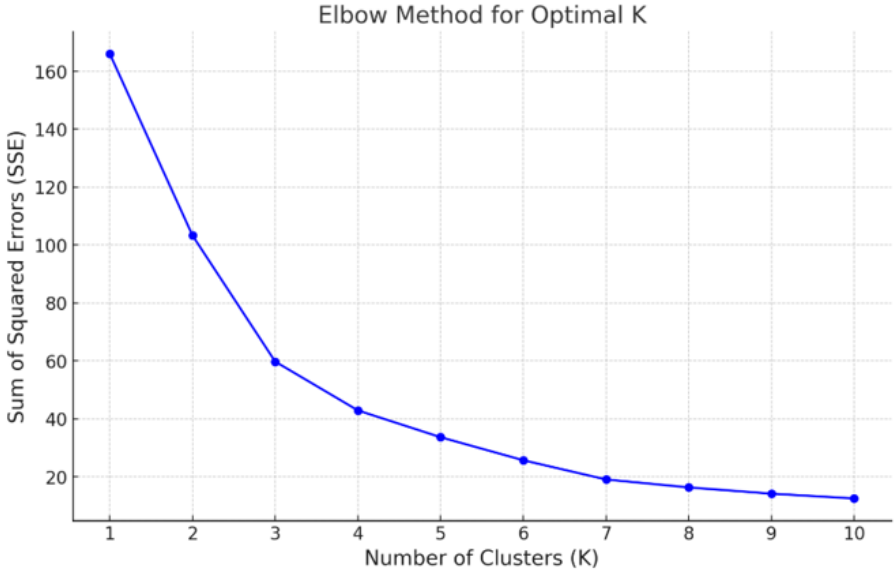

Figure 1. The relationship between SSE and K

The curve indicates that as the K value increases, WSS gradually decreases, but as the K value increases, the magnitude of error reduction gradually decreases. At the "elbow" position, there will be a significant change in the descent speed of the curve, which is the optimal choice for the K value.

the article conducted the above clustering analysis on the number of cybercrime incidents with crime success rate, reporting rate, and prosecution rate. Due to the significant differences in reporting rate and prosecution rate data, we used a bar chart to better illustrate their differential characteristics We analyze the performance characteristics of these countries in cybercrime prevention and control based on each cluster. Assuming that the characteristics of each category are as follows:

Category 1: Low success rate, low reporting rate, low prosecution rates the prevention and control of cybercrime in these countries are relatively weak, with a high success rate and low reporting and prosecution rates. There may be significant legal or technical loopholes, or the government's attention to cybercrime may be relatively low. Such countries may need to strengthen technological protection, raise public awareness of cybersecurity, and improve reporting and legal enforcement mechanisms. Example countries: India, Argentina, Nepal, Bangladesh, etc. Category 2: High success rate, low reporting rate, low prosecution rates These countries have high success rates in cybercrime, but there are significant issues with reporting and prosecution. Although cybercrime occurs frequently and successfully, these countries may lack effective mechanisms for tracking and enforcing laws against cybercrime, or may be unwilling to publicly report cybercrime for reasons such as corporate interests. Policy makers need to enhance transparency in cybercrime and strengthen reporting and prosecution capabilities for cybercrime. Example countries: China, Brazil, Argentina. Category 3: High success rate, high reporting rate, low prosecution rates the success rate and reporting rate of cybercrime in these countries are relatively high, but the prosecution rate is relatively low. These countries may have established strong cybercrime reporting systems, but there are certain loopholes in prosecution, possibly due to complex legal procedures or limited law enforcement resources. Suggest strengthening the legal system and law enforcement capabilities to ensure smooth prosecution of cybercrime cases. Example countries: United States, France, Germany, South Korea. Category 4: High success rate, high reporting rate, high prosecution rates These countries have shown excellent performance in cybercrime prevention and control, with high success rates, reporting rates, and prosecution rates. These countries usually have strong technological capabilities, well-established legal systems, and policies that highly value cybersecurity. These countries can serve as role models for other countries, sharing experiences and best practices to help them improve their cybercrime prevention and control capabilities. Example countries: United Kingdom, Switzerland, Israel, Netherlands, Norway. The result is shown in the Figure 2:

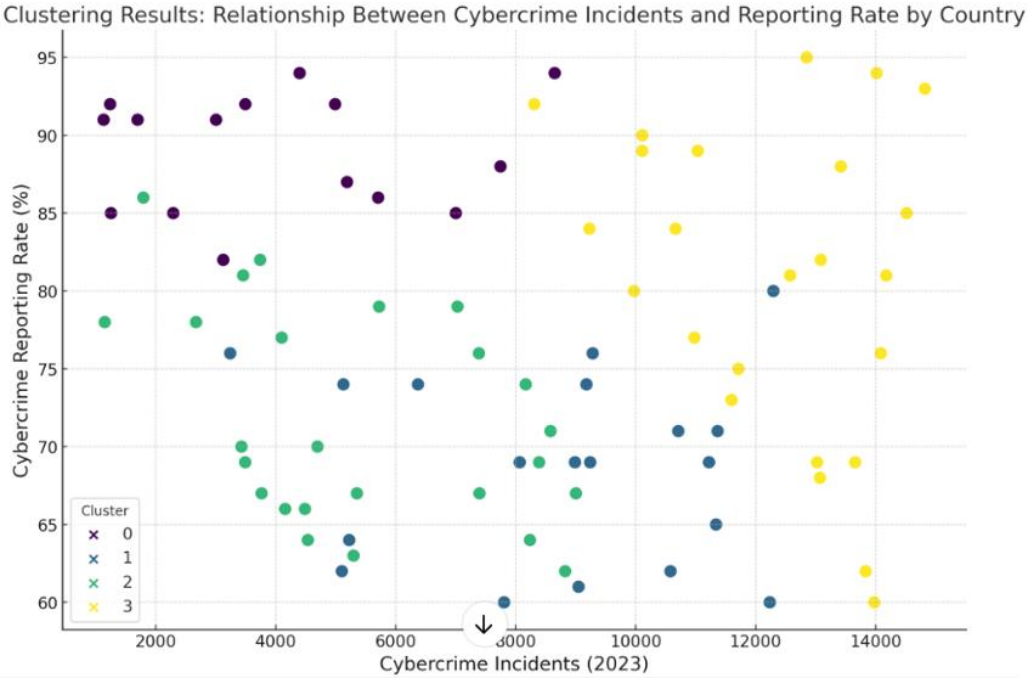

Figure 2. Cluster scatter plot of the number of success rate

By using the elbow rule and cluster analysis to divide countries into four categories, we can draw the following conclusions There are significant differences in the prevention and control of cybercrime: different countries have significant differences in their response to cybercrime, especially in terms of reporting and success rates. Countries with high success rates may not necessarily have high reporting or prosecution rates, which may reflect differences in legal enforcement or technological means among countries. Next, we focus on analyzing the policies of various categories of countries. Due to the significant differences in data dimensions between different data sources, the article standardized the data to ensure that each feature can be compared at a unified scale. The result is shown in the table 1:

Table 1. Data on the different aspects of the GCI for certain countries

Economy ISO3	Capacity development measure	Coopration measures	Legal measure	Organizational measures	Technical measure
CN	18.46	17.7	20	18.34	17.14
CA	18.98	20	18.9	20	15.3
GB	20	20	20	20	20
AU	19.44	18.85	20	20	17.95
DE	19.92	15.99	20	20	17.93

In order to measure the effectiveness of policies in curbing cybercrime, the article quantify the effect of cybercrime by the percentage change between predicted and actual values, and quantify policies using the five parameters in GCI to better find their relationships. In practical scenarios, the effectiveness of policies in curbing cybercrime is influenced not only by the GCI parameters but also by various other factors. Therefore, before conducting feature importance analysis using the decision tree model, the article first performs ARIMA time series forecasting [6]. We assume that the changes in other factors over time follow a regular pattern. Since the enactment of policies is instantaneous, the article can effectively isolate the impact of other factors by leveraging this characteristic and focus solely on evaluating the effect of policies on curbing cybercrime. Following this, the article tests the significance of the model parameters, the validity of the model itself, and check whether the residuals follow a white noise process. If the model passes these tests, it is considered correctly specified; otherwise, the model form needs to be re-determined and re-evaluated until a valid model is obtained. Finally, the article uses the established ARIMA model to make predictions for each country and compare the predicted values with the actual values. The result is shown in the Figure 3:

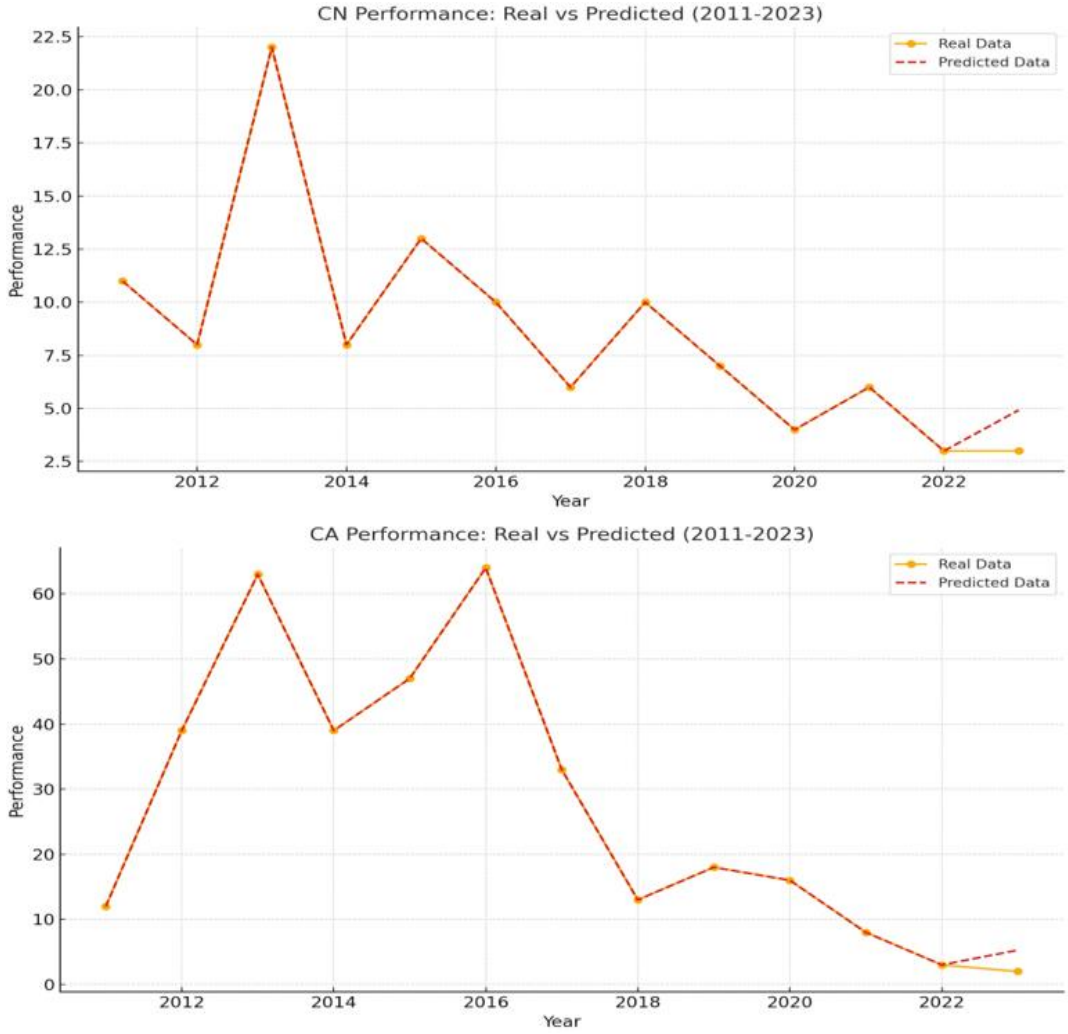

Figure 3. Line chart of predicted values versus actual values for each country

Afterwards, the article will use the percentage change between the predicted data and the actual value as a quantitative parameter to measure the effectiveness of cybercrime containment, and combine it with the five-dimensional parameters of GCI for decision tree prediction. Finally, the article obtains the weight ratio of the five-dimensional parameters of GCI on the impact of cybercrime in this environment, that is, the feature importance. Using decision trees, we successfully obtained the feature importance of each parameter in GCI. To ensure the rigor of the results, we conducted a series of tests, The result is shown in the table 2:

Table 2. Model Evaluation Results Table

	MSE	RMSE	MAE	MAPE	R^2
training set	0	0	0	0	1
test set	0.013	0.116	0.112	732.702	0.488

MSE (Mean Squared Error): The expected value of the square of the difference between the predicted and actual values. A smaller value indicates higher model accuracy.

RMSE (Root Mean Squared Error): The square root of MSE. A smaller value indicates higher model accuracy.

MAE (Mean Absolute Error): The average of the absolute errors, reflecting the actual prediction error. A smaller value indicates higher model accuracy.

MAPE (Mean Absolute Percentage Error): A variant of MAE, expressed as a percentage. A smaller value indicates higher model accuracy.

R^2 (Coefficient of Determination): Compares the predicted values to those obtained using only the mean. The closer the result is to 1, the higher the model's accuracy [7].

Since the known data contains both positive and negative values, and the overall absolute values are relatively small, we applied an exponential transformation to smooth the entire dataset. After this processing, we conducted the tests again, and the results are as follows in the table 3:

Table 3. Model Evaluation Results Table

	MSE	RMSE	MAE	MAPE	R^2
training set	0	0	0	0	1
test set	0.013	0.116	0.114	10.866	0.521

As observed, after processing the data, the MAPE significantly decreased in the subsequent tests, indicating a substantial reduction in errors. The gap between the predicted values and the actual values narrowed considerably, which aligns with our expectations. Based on the data processing outlined above, we performed feature importance analysis, and the results are as follows in the Figure 4:

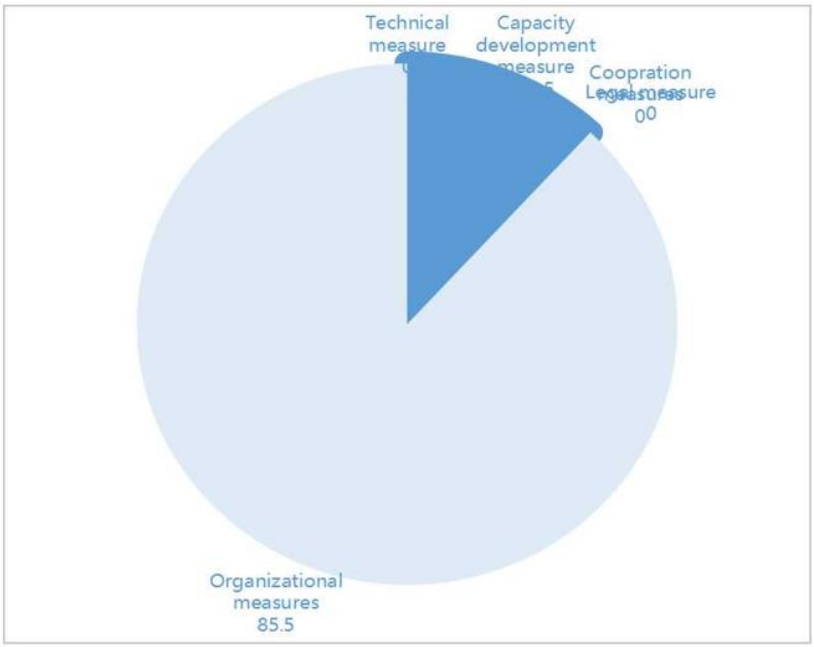

Figure 4. Feature Importance Bar Chart for GCI Parameters

After the above analysis, the article has identified some reliable demographic data that are somewhat related to the occurrence of cybercrime. We need to use the PLSR model to quantitatively calculate the correlations of these data.

By observing the variance and other data of the model, we continuously adjust the reliability of the data and create tables to record the results when the error is minimized. This calculation is then used to derive the feature importance and the results are as follows in the table 4, table 5 and table 6:

Table 4. Summary table of independent variable VIP

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Government effectiveness	1.339	0.93	1.063	1.043	1.022
Population	0.659	0.975	0.915	0.901	0.878
Computer service rate	0.005	0.424	0.816	1.286	1.188
Urban population rate	1.343	0.94	0.859	0.781	0.906
GDL-Educational-index-data	1.369	0.952	0.915	0.862	0.802
Individuals using the Internet	1.293	0.907	0.815	0.736	0.684
Real GDP per capita	0.468	0.741	0.833	0.812	1.153
Total labur force	0.523	1.001	0.901	0.855	0.812
GDP	0.964	1.686	1.619	1.47	1.362

Table 5. Component Matrix Table

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Government effectiveness	0.446	0.227	0.568	-1.041	1.267
Population	-0.22	0.277	0.295	-0.653	0.855
Computer service rate	-0.002	0.195	0.605	0.04	0.033
Urban population rate	0.448	0.303	-0.033	0.117	-0.583
GDL-Educational-index-data	0.456	0.231	0.324	-0.606	0.652
Individuals using the Internet	0.431	0.18	0.009	0.031	-0.097
Real GDP per capita	0.156	-0.222	-0.448	0.809	0
Total labur force	-0.174	0.336	0.17	-0.422	0.562
GDP	0.321	0.894	-0.151	0.334	-0.202

Table 6. Factor Loading Coefficients Table

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Government effectiveness	0.458	-0.178	0.09	-0.042	0.3
Population	-0.44	0.333	-0.148	-0.049	0.124
Computer service rate	-0.11	0.023	1.575	0.877	0.058
Urban population rate	0.417	0.097	-0.033	-0.366	-0.646
GDL-Educational-index-data	0.468	-0.101	0.003	-0.112	0.168
Individuals using the Internet	0.464	-0.043	0.004	-0.023	-0.107
Real GDP per capita	0.326	-0.188	-0.074	0.942	0.675
Total labur force	-0.414	0.412	-0.2	-0.049	0.099
GDP	-0.075	0.856	-0.255	0.244	0.201

Finally, perform cross-validation. For each iteration, omit the i observation and use the remaining n-1 data points to fit the regression equation after extracting h components using the least squares regression method. Then, substitute the omitted j observation from the independent variable group into the fitted regression equation to obtain the predicted value of $y_j(j = 1,2,\ldots,p)$ at the i-observation point. Repeat the above validation for $i = 1,2,\ldots,n$. The predicted sum of squared errors for the j dependent variable $y_j(j = 1,2,\ldots,p)$ when h components are extracted is:

$$PRESS_j(h) = \sum_{i=1}^{n}(b_{ij} - b_{ij}(h))^2, j = 1,2,\ldots,p \tag{15}$$

The sum of squared prediction errors for $Y=[y_1,\ldots,y_p]^T$ is:

$$PRESS(h) = \sum_{j=1}^{p} PRESS_j(h) \tag{16}$$

Additionally, using all the sample points, fit a regression equation with h components. In this case, let the predicted value of the i-sample point be $\hat{b}_{ij}(h)$, and the sum of squared errors for y_i can be defined as:

$$SS(h) = \sum_{j=1}^{p} SSj(h) \tag{17}$$

When $PRESS(h)$ reaches its minimum value, the corresponding h is the number of components sought. Typically, there is always $PRESS(h) > SS(h)$, and $SS(h) < SS(h-1)$, therefore, when extracting components, we always hope that $\frac{PRESS(h)}{SS(h-1)}$ is as small as possible. The article set a threshold of 0.05, with the decision rule being: $\frac{PRESS(h)}{SS(h-1)} \leq (1-0.05)^2$, indicating that the new component helps improve the regression performance [8]. Therefore, we can define cross-validation effectiveness as:

$$Q_h^2 < 1 - 0.095^2 \tag{18}$$

Before completing each calculation step, the article calculates cross-validation effectiveness. When the model reaches the desired accuracy, the component extraction process stops. Through the above validation method, we can confidently conclude that the model is important for both prediction and feature importance analysis. To further validate that the model's error is within an acceptable range, the article compares the model's predicted values with the actual values to demonstrate this. The result is shown in the Figure 5:

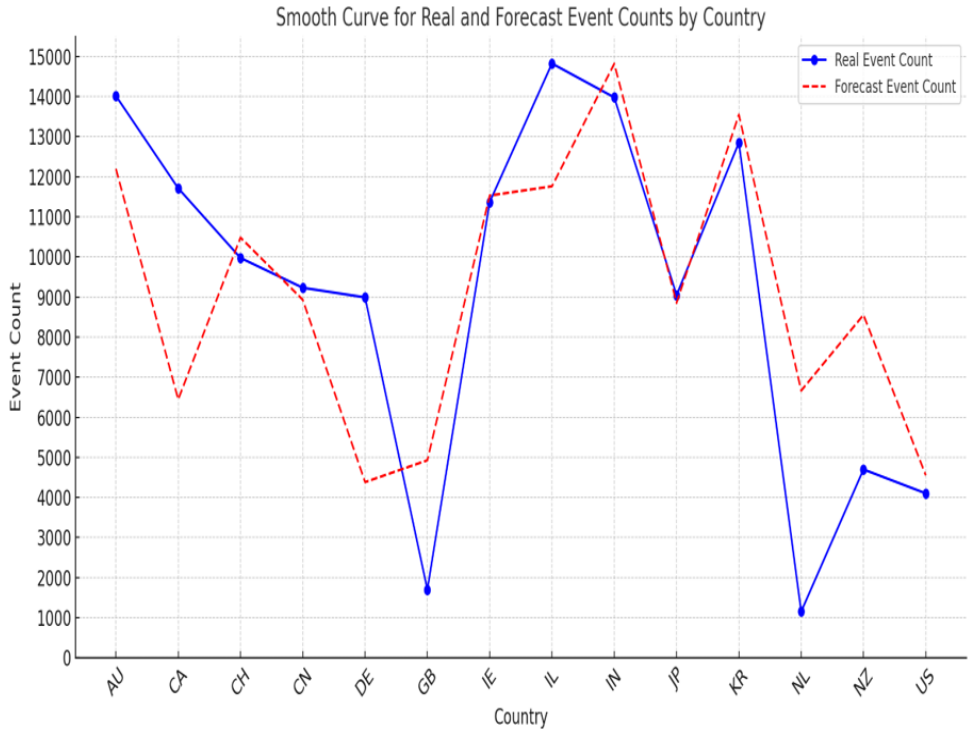

Figure 5. PLSR model predicted values vs actual values line chart

The fitted trend closely matches the actual values, indicating that the model has a certain degree of 7reliability. The output results are as follows in the table 7:

Table 7. Model Coefficients Results Table

	Event Count
constant	9117.286
Government effectiveness	-7319.342
Population	-4377.131
Computer service rate	-1774.078
Urban population rate	784.648
GDL-Educational-index-data	-4381.035
Individuals using the Internet	-391.327
Real GDP per capita	3137.316
Total labur force	-3029.849
GDP	-166.177
R^2	0.558

4.3. Analysis Result

Through clustering analysis of cybercrime numbers and their corresponding success rates, prosecution rates, and reporting rates in several countries worldwide, and visualizing the results in a heatmap, the article classified these countries into four categories. We found that countries with higher cybercrime numbers generally have less developed or incomplete mechanisms for crime prevention and legal frameworks. In contrast, countries with lower cybercrime numbers tend to have more well-established mechanisms. This conclusion is also supported by the GCI scores. We further concluded

that there is no clear geographic pattern among countries with high cybercrime rates; these countries are distributed globally. By establishing ARIMA prediction models and decision tree models to analyze the relationship between the GCI scores and cybercrime numbers (extracted from the VCDB) of five representative countries, we discovered that the "organizational measures" feature holds the highest importance score. This indicates that organizational measures have a significant impact on cybercrime [9]. Countries with higher organizational measures scores have likely already established effective policies to curb cybercrime. In contrast, countries with lower scores in this area may still need to implement more reasonable laws and policies focused on combating cybercrime organizations [10]. The article used the PLSR model to analyze the correlation between demographic features and cybercrime across 14 countries. Through modeling and analysis, we calculated the VIP (Variable Importance in Projection) values for each feature. For example, GDP has a VIP value of 1.686 for Factor 2, and GDP generally has a high VIP value across all five factors. This suggests a strong correlation between GDP and cybercrime. Similarly, features like GDP, population size, and GDP per capita show strong correlations with cybercrime, while features such as the proportion of computer services in the economy show a very low correlation.

5. Conclusion

This article quantifies the GCI scores of various countries by developing appropriate models GDP The relationship between demographic characteristics and cybercrime. This allows us to clearly analyze which policies have a significant impact on reducing cybercrime. These findings can help government agencies develop more targeted and reasonable prevention policies, thereby promoting the establishment and improvement of national cybersecurity mechanisms. However, due to the imbalance of categories in the dataset, the model may tend to predict larger categories of data, resulting in biased results. The segmentation process in decision tree models reduces data capacity, and this segmentation method inevitably introduces bias. Moreover, the situation of cybercrime may change due to factors such as policy changes, technological advancements, and social transformations. Therefore, regularly updating models, maintaining data freshness, and adjusting policy effectiveness analysis in real-time will help improve policy responsiveness. Attempt to combine machine learning methods such as random forest, XGBoost, etc. to improve the shortcomings of ARIMA models in handling nonlinear and multivariate data. For example, combining ARIMA with neural networks and LSTM to enhance the model's ability to handle nonlinear and long-term dependent data. Before modeling, strengthen the handling of outliers and missing data, especially by using more advanced time series preprocessing techniques such as time series decomposition.

References

[1] Sen P, Roy M, Pal P. Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization [J]. Energy, 2016, 116: 1031 - 1038.

[2] Cybercrime in social media: Theory and solutions [M]. CRC Press, 2023.

[3] Akram M S, Mir M J, Rehman A. Dimension of cyber-warfare in Pakistan’s context [J]. Journal of Positive School Psychology, 2023, 7 (6): 82 - 94.

[4] Veena K, Meena K, Tee Karaman Y, et al. C SVM classification and KNN techniques for cybercrime detection [J]. Wireless Communications and Mobile Computing, 2022, 2022 (1): 3640017.

[5] Alagappan A, Venkata chary S K, Andrews L J B. Augmenting Zero Trust Network Architecture to enhance security in virtual power plants [J]. Energy Reports, 2022, 8: 1309 - 1320.

[6] Wang H. Big data security management countermeasures in the prevention and control of computer network crime [J]. Journal of Global Information Management (JGIM), 2021, 30 (7): 1 - 16.

[7] Yokotani K, Takano M. Predicting cyber offenders and victims and their offense and damage time from routine chat times and online social network activities [J]. Computers in Human Behavior, 2022, 128: 107099.

[8] Lubis M, Handayani D O D. The relationship of personal data protection towards internet addiction: Cybercrimes, pornography and reduced physical activity [J]. Procedia Computer Science, 2022, 197: 151 - 161.

[9] Timofeyev Y, Dremova O. Insurers' responses to cybercrime: evidence from Russia [J]. International Journal of Law, Crime and Justice, 2022, 68: 100520.

[10] Tsolekile S. The plausibility of developing a digitised township economy: A study of townships in Cape Town [J]. 2021.