

**Project Report**

**Strengthening Business Security: Predictive Analytics of Cybersecurity Trends Using CISSM Data**

**Harish Sakaray**

**1. Business Understanding**

**Business Problem Definition**

The project addresses the growing concern of the cybersecurity incidents which threaten the organizations across the different types of sectors. As the number of cyber-attacks is rapidly increasing, the organizations are struggling to anticipate and mitigate the risks effectively. The primary objective of this project is to build a machine learning model using the CISSM data to forecast trends in the types of cybersecurity incidents (e.g., data breaches, service disruptions) based on historical data, so that it can help the organizations to forecast the emerging cyber threats and enhance their security strategies.

**Significance of the Problem**

Cybersecurity incidents will have the devastating effects on the organizations that includes the financial loss, reputational damage, legal repercussions etc., Studies have shown that the cost of cybercrimes is projected to reach trillions of dollars annually. This emphasizes the need for the effective measures in enhancing the cybersecurity posture. By leveraging the historical data of the global cyber events, this project aims to provide the organizations with the insights into the trends and patterns in the cyber incidents. This enables them to allocate the resources in an efficient way, implement the proactive measures and reduce the vulnerability to the cyber threats.

**Specific Objectives and Goals of the Analysis**

Our Project aims to cater the following objectives and goals.

**Trend Analysis**: Identifying and analyzing the historical trends on different types of events in cybersecurity incidents such as data breaches, service disruptions etc., and their characteristics such as actor types, motives etc.,

**Predictive Modelling**: Developing a machine learning model that helps in forecasting the future cyber incidents based on the historical data. This helps the organizations to anticipate the potential cyber threats.

**Resource Allocation**: This project aims to assist the organizations in making the informed decisions in allocating resources for cybersecurity by understanding which sectors are having he most risk.

**Enhancing Security Strategies**: This project will help in providing the actionable insights to the organizations so that they can adjust their cybersecurity strategies, improving their resilience against the emerging cyber threats.

**2. Data Understanding and Preparation**

**Dataset Specification**

In the initial phase, Our team is going to use the Cyber Events Database provided by Center for International Security Studies at Maryland, University of Maryland, an open source dataset which has a collection of data related to various cyber incidents. The details of the dataset is as follows:

**Dataset Name**: Cyber Events Database

**Authors**: Charlie Harry, Nancy Gallagher, and Lauren Samuelsen

**Institution**: Center for International Security Studies at Maryland, University of Maryland

**Reference**: Harry, C., & Gallagher, N. (2018). Classifying Cyber Events. Journal of

Information Warfare, 17(3), 17-31.

**Access**: The dataset is opensource and publicly accessible through their website.

**Data Format**: The dataset in in Excel (.xlsx) format.

**Size**: The dataset has a size of 2.78 MB approximately.

**Dimensions**: The dataset has a total of 16 columns and 13,842 rows.

**Attributes and Description**:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| slug | A unique identifier for each cyber event. |
| event\_date | The date when the cyber event occurred. |
| year | The year the incident took place. |
| month | The month the incident occurred. |
| actor | The name of the responsible individual or organization. |
| actor\_type | Classification of the actor into categories like criminal, nation-state, hacktivist etc., |
| organization | The entity impacted by the cyber event. |
| industry\_code | A two-digit classification code indicating the target organization’s industry. |
| industry | Describes the industry sector corresponding to the NAICS code. |
| motive | The reason for the attack (e.g., financial gain, espionage, sabotage). |
| event\_type | Classifies the event as disruptive, exploitative, or a combination. |
| event\_subtype | Specifies the method of attack (e.g., Denial of Service, Data Manipulation). |
| description | A description of what occurred during the incident. |
| source\_url | The link to the original public information about the event. |
| country | The ISO 3-letter code for the country where the target organization is located. |

**Data Preparation:**

Our team has cleaned the existing data set and preprocess it before building the model.

**Importing the data**: Our team imported the cyber events dataset into the environment (R or Python).

**Handling missing values**: We checked for any missing values in the dataset including null values using functions such as isnull() in pandas. We handled the missing values by removing the rows that has critical missing values. We imputed the missing values in less critical fields with place holders or statistical measures such as mean or median depending upon the field.

**Type conversion**: We ensured that each column in the dataset has correct data type. We formatted the column to correct data type such as converting event\_date to datetime format, year and month columns to integers etc.,

**Outlier detection**: We identified any existing outliers in the numerical fields by using the methods such as Z-scores and IQR method.

**Standardization and Normalization**: We standardized all the existing categorical variables such as actor\_type, industries etc., in order to ensure the consistency among the values. We also normalized the numerical attributes for efficient analysis using techniques such as min-max scaling etc.,

**Key Variables and Attributes we focused on**:

**event\_date**: Helps in understanding distribution of events across time

**actor** and **actor\_type**: Useful in analyzing the types of actors involved in various incidents.

**industry**: Helps in understanding which industries are more targeted.

**event\_type**: Important variable in categorizing the incidents to identify the trends and patterns.

**motive**: Helps in understanding the objectives behind the attacks.

**Feature Selection/Engineering:**

Feature engineering was a key step in optimizing the model’s performance by creating the new and meaningful features from the raw data. Temporal features such as the month, quarter, day\_of\_week, and is\_weekend were designed in order to capture the time-based patterns in the cybersecurity incidents which revealed trends like increased attacks during specific periods. Interaction features, such as actor\_x\_motive and industry\_x\_event\_type, were created in order to understand the combined effects of the actors’ motives and the targeted industries, which enables deeper insights into the attack patterns. Aggregated features, such as the sector\_event\_ratio and actor\_event\_count has provided a summarized view of the incidents across multiple records, by offering a macro-level perspective on the most impacted sectors.

To capture the recurring temporal trends, cyclical features such as month\_sin and month\_cos were used in order to represent cyclic patterns such as seasonality, avoiding artificial discontinuities in the data. Categorical features were transformed into numerical representations using encoding techniques. Frequency encoding replaced categories with their occurrence rates for features like actor\_type and motive, while label encoding assigned unique integers to event types. For features without natural order, such as industry or country, one-hot encoding created binary columns to ensure all categories were treated equally by the model.

**Plots and Visualizations:**

A graph with a green and orange rectangle

Description automatically generated

This bar chart displays the number of cyber threat events attributed to different actor types (e.g., Hacktivist, State-Sponsored). It helps to understand which types of actors are most frequently involved in cyberattacks.

A graph with multiple colored squares

Description automatically generated with medium confidence

This bar chart illustrates the distribution of different types of cyber threat events (e.g., Disruptive, Exploitive). It provides insights into the prevalence of various attack strategies.

A graph with blue lines

Description automatically generated

This visualization presents a histogram showing the frequency of cyber threat events over the years in the dataset. It provides insights into the overall trend of cyberattacks, highlighting periods with higher or lower activity levels.

A graph showing a line

Description automatically generated

This time series plot depicts the trend of cyber threat events over time, aggregated on a monthly basis. It helps to visualize any seasonal patterns or significant changes in cyberattack activity over the observed period.

A diagram of a heatmap

Description automatically generated

This heatmap visualizes the correlations between numerical features in the dataset, such as year, industry code, and month. It helps to identify potential relationships between these variables, which can inform further analysis or feature engineering.

A green square with red lines

Description automatically generated

This box plot illustrates the distribution of industry codes affected by cyberattacks. It shows the median, quartiles, and outliers, allowing for the identification of industries that are more prone to cyber threats.

A graph showing different colored bars

Description automatically generated with medium confidence

This bar chart shows the distribution of cyber threat events across different industries. It helps to understand which industries are most frequently targeted by cyberattacks.

A diagram of a distribution of event types

Description automatically generated

This pie chart presents the same information as the previous bar chart but in a circular format, emphasizing the relative proportions of each event type.

**3. Methodologies Applied**

In our project, we have used the XGBoost (Extreme Gradient Boosting) machine learning model. We have divided the dataset into 80% for training and 20% for testing so that the model could be trained on a very large set of data and the testing can be done on the unseen data. The implementation of this classifier for training the model involved utilizing XGBoost, an efficient and accurate gradient boosting algorithm widely used for pattern recognition in various domains, including cybersecurity incidents. This was measured through accuracy and a classification report.

The trained model was then used to predict cybersecurity trends over the next 5 years. The forecasts revealed the types and frequency of events across industries, and are visualized with a heat map to show trends by industry and year. XGBoost stands out as the top choice because of and for its exceptional performance and versatility. It consistently outperforms other algorithms in various machine learning competitions and in real-world applications, making it a reliable alternative to prediction.

**Key advantages of XGBoost:**

**High prediction accuracy:** The XGBoost model excels at capturing the complex patterns in the data, resulting in the highly accurate predictions.

**Regularization:** This includes built-in regularization mechanisms to prevent overfitting, ensure that the model generalizes well to unseen data and avoid overly complex models that memorize training data but fail to perform well on new data.

**Feature Importance**: XGBoost provides insight into the importance of different factors in the model, supports feature selection and understands predictive factors.

**Efficiency and scalability:** Designed for speed and scalability, making it suitable for handling big data and real-time applications.

**Implementation**:

The XGBoost classifier is imported into our code by importing xgboost from XGBClassifier. The classifier is instantiated using specific parameters developed for our project such as the parameter objective='multi:softmax' which configures the classifier for multiclass classification, where the objective is to predict one of several possible cases.

num\_class=len(df['event\_type'].unique()): This sets the class number based on the unique events in your dataset, ensuring that the model is aware of all possible outcomes

random\_state=42: Setting the random state enables repeatability of results, and allows for consistent results as you execute the code repeatedly.

The model is trained using the fit method, providing it with training data (X\_train, y\_train). During training, XGBoost reconstructs decision trees, reducing error and refining its predictions. Data are tested using a predictive approach, and the trained model is applied to unobserved data to evaluate its performance.

Example: model = XGBClassifier(value = 'quantity: softmax', num\_class = len(df['article\_type'].unique()), random\_rank = 42);

**Model-Building Processes**:

**Data Splitting**: To check the ability of the model to generalize to new data, the data set is split into training test sets using train\_test\_split. This ensures that the model not only memorizes the training data but can accurately predict outcomes in unseen examples.

**Feature Scaling**: Numerical features are scaled using StandardScaler to ensure they have comparable ranges. This prevents features with larger values from dominating the model and ensures that all features contribute equally during training.

**Validation Methods**: Validation techniques, such as accuracy and classification reports are used to evaluate the model’s performance.

**4. Results and Findings**

**Model’s performance:**

**A screenshot of a computer code

Description automatically generated**

After analyzing the results, we have found that our model has an accuracy of 71.7% which means that our model has correctly predicted the event type for approximately 71.7% of the instances from the test dataset. The precision value tells us that how often the model is correct when the model predicts the event of a particular event type. In our case, the precision for the disruptive event is 0.75 which means that when the model predicts the event as disruptive, it is correct at 75% of the time. Similarly, the precision for exploitive events is also 0.75 and for mixed events it is slightly higher at 0.83.

The recall score measures how often the model correctly predicts a particular type of event out of all actual instances of that type of event. For disruptive events the recall score is 0.63 which means that the model had correctly identified the 64=3% of actual disruptive events in the dataset. The recall for the exploitive events is higher at 0.82 which indicates a better performance in identifying those events. The F1 scores for the disruptive, exploitive and mixed events are 0.68, 0.77 and 0.62 respectively. These findings has provided with a comprehensive overview of the model’s performance.

**Predicted Event Types Over the Next 5 Years by Industry:**

A graph of a number of blue and red squares

Description automatically generated with medium confidence

Using the model we have built, we have predicted the event types over the next five years for all the industries. This heatmap visualizes the predicted distribution of cybersecurity event types across various industries over the next five years, from 2025 to 2029. The industries are listed on the vertical axis, while the years are represented on the horizontal axis. The color intensity within each cell indicates the predicted prevalence of a specific event type in a given industry and year. Red signifies disruptive events, blue represents exploitive events, and white indicates mixed events.

It is observed that the model predicts a predominance of "Exploitive" events (blue) across most industries in the coming years. This suggests that attacks aimed at exploiting vulnerabilities and gaining unauthorized access to systems will likely remain prevalent. "Disruptive" events (red), which aim to disrupt operations and services, are also predicted, particularly in industries like "Mining, Quarrying, and Oil and Gas Extraction" and "Utilities." This highlights the potential for targeted attacks on critical infrastructure. Interestingly, "Mixed" events (white) are less prevalent in the predictions, suggesting that attacks with a combination of disruptive and exploitive motives might be less common. This heatmap provides valuable foresight into the potential cybersecurity landscape, enabling businesses to proactively allocate resources and strengthen their defenses against the most likely threats in their respective sectors.

A pie chart with different colored circles

Description automatically generated

The pie chart provided a high-level view of the overall distribution of event types for the forecasted period. It showed the following trends:

* Exploitive Attacks are accounted for the largest share, emphasizing the importance of detecting and mitigating breaches early.
* Disruptive Attacks are slightly less frequent, their impact on business continuity makes them a critical focus for incident response teams.
* Mixed Event Types are represented as a smaller proportion but highlighted the need for robust, multifaceted cybersecurity strategies.

These findings underscore the necessity for tailored approaches to managing cybersecurity risks. Organizations cannot afford a one-size-fits-all approach; instead, they need to adapt based on their specific industry, threat profile, and resource constraints.

**5. Limitations and Future Work**

**Data Bias**: One of the most significant challenges was the inherent bias in the CISSM dataset. Since the data relies on reported cybersecurity incidents, certain event types, industries, or geographic regions may be underrepresented. To mitigate this bias, future iterations could incorporate additional datasets from diverse sources, providing a more balanced view of global cybersecurity incidents.

**Feature Selection**: The selection of features in our model had a direct impact on its performance. Striking the right balance was critical to avoid overfitting and improve performance.

**Model selection and optimization**: Choosing the most appropriate machine learning model for predicting cybersecurity trends was challenging and required experimentation with different algorithms

**Evolving Threat Landscape**: Cyber threats evolve rapidly, often outpacing existing defenses and models trained on historical data. For instance, Emerging technologies, such as quantum computing or AI-driven malware, might introduce entirely new types of threats not captured by our model. Our model’s reliance on historical patterns means it might not fully account for these developments. To address this, continuous retraining of the model using updated data is essential. Integrating threat intelligence feeds could also help capture real-time trends.

**6. Conclusion**

This project delves into the critical area of cybersecurity trend prediction, aiming to bolster business security against the ever-evolving threat landscape. Leveraging the CISSM dataset, a comprehensive repository of cyber incidents, we embarked on a journey to understand the patterns and nuances of cyberattacks. Our methodology involved rigorous data preparation, followed by the application of XGBoost, a powerful machine-learning algorithm renowned for its predictive accuracy and efficiency. This model, trained on historical cyber incident data, was then used to forecast potential threats over the next five years, providing a glimpse into the future cybersecurity landscape.

Our XGBoost model, with a commendable accuracy rate of 71.7%, successfully predicted the event type for a significant portion of the test dataset. This predictive capability is a valuable tool for organizations, enabling them to anticipate the nature of potential cyber threats. Furthermore, our model's ability to forecast the types and frequency of cybersecurity incidents across different industries for the next five years is a significant achievement. This foresight, visually represented using a heatmap, empowers businesses to strategically allocate resources and proactively enhance their security strategies.

Despite the inherent limitations, such as potential data bias and the dynamic nature of cyber threats, our project offers invaluable insights for businesses to understand and anticipate future cybersecurity challenges. By predicting trends and understanding potential vulnerabilities, organizations can proactively strengthen their cybersecurity posture and mitigate potential risks, ensuring a safer digital future.

**7. References**

[**Harry, C., & Gallagher, N. (2018). Classifying Cyber Events. Journal of Information Warfare, 17(3), 17-31.**](https://cissm.umd.edu/research-impact/publications/classifying-cyber-events-proposed-taxonomy)

Ahmad, A., Maynard, S. B., & Shanks, G. (2019). A systematic review of data breach prediction models. Journal of Cybersecurity and Privacy, 1(3), 321-348. https://doi.org/10.3390/jcp1030020

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

Verizon. (2023a). 2023 Data Breach Investigations Report. Help Net Security. https://www.helpnetsecurity.com/2023/06/06/verizon-data-breach-investigations-report-2023-dbir/

Verizon. (2023b). Verizon 2023 Data Breach Investigations Report. Noetic Cyber. https://noeticcyber.com/verizon-2023-data-breach-investigations-report/

Verizon. (2023c). Verizon 2023 Data Breach Investigations Report: Frequency and cost of social engineering attacks. Nasdaq. https://www.nasdaq.com/press-release/verizon-2023-data-breach-investigations-report%3A-frequency-and-cost-of-social

ISO/IEC 27001. (2013). Information technology – Security techniques – Information security management systems – Requirements. International Organization for Standardization.