Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security

Harisha Dheeravath (202595477)
Introduction

The rapid digitization of supply chains, while heralding unprecedented operational efficiencies, has concurrently escalated their vulnerability to sophisticated cyber threats. This study addresses the critical need to fortify these digital supply chains against evolving threats like malware, ransomware, and Advanced Persistent Threats (APTs). Our focus pivots on enhancing cyber supply chain security through the innovative application of predictive analytics and machine learning. We aim to transition from traditional reactive cybersecurity methods to a proactive, anticipatory stance. This involves a thorough analysis of data patterns and anomalies within supply chain activities, leading to the development of predictive models that can foresee and mitigate potential cyber risks. Additionally, we undertake a comprehensive evaluation of various machine learning classifiers, enriching our understanding of their effectiveness in different threat scenarios. Our objectives—Predictivity, Analysis, Evaluation, and Proactivity—collectively strive to minimize attacker opportunities, ensuring the robustness and continuity of supply chain operations in an increasingly digitalized world.
Problem statement

Digital advancements in supply chains have led to greater efficiency but also increased exposure to cyber threats such as malware, ransomware, and phishing attacks. Advanced Persistent Threats (APTs) pose a significant challenge by stealthily aiming to disrupt or steal data over time. Traditional cybersecurity methods, which are reactive and limited to known threats, are often inadequate against these sophisticated attacks, particularly in detecting zero-day vulnerabilities. This underscores the need for a shift towards predictive analytics and machine learning to proactively identify and counteract potential cyber threats, enhancing the resilience and security of supply chains while ensuring operational integrity.
Objectives

Predictivity
Enhance supply chain security using predictive analytics.

Evaluation
Compare machine learning classifiers for cyber threat assessment.

Analysis
Analyze data for trends, anomalies in supply chain activities.

Proactivity
Shift from reactive to proactive cyber security strategy.
| Sl. No. | Title                                                                 | Author                        | Year | Context                        | Research Design                                           | Major Theme                                      |
|--------|------------------------------------------------------------------------|-------------------------------|------|--------------------------------|----------------------------------------------------------|--------------------------------------------------|
| 1      | Security and Resilience in Sustainable Smart Cities through Cyber Threat Intelligence | K. Nova                       | 2022 | Smart city cyber resilience   | Predictive analytics                                      | Enhancing cyber resilience                      |
| 2      | Trends and Challenges Regarding Cyber Risk Mitigation by CISOs         | M. Zwilling                   | 2022 | Cyber risks for CISOs         | Systematic literature and expert opinion review          | Cyber risk mitigation strategies                 |
| 3      | Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security | A. Yeboah-Ofori et al.       | 2021 | Cyber supply chain security  | Predictive analytics and machine learning techniques     | Proactive cyber threat identification            |
| 4      | Cyber security threat modeling for supply chain organizational environments | A. Yeboah-Ofori and S. Islam | 2019 | Supply chain security        | Threat modeling                                          | Key strategies for supply chain security         |
| Sl. No. | Title                                                                 | Author                               | Year | Context                      | Research Design     | Major Theme                     |
|--------|-----------------------------------------------------------------------|---------------------------------------|------|------------------------------|---------------------|---------------------------------|
| 5      | Modelling Activity of a Malicious User in Computer Networks           | A. D. Lazarov and P. Petrova          | 2022 | Cyber defense strategies     | Behavioral modeling | Malicious user behavior analysis |
| 6      | Anomaly Detection Using XGBoost Ensemble of Deep Neural Network Models | S. T. Ikram et al.                   | 2021 | Network anomaly detection   | Ensemble machine learning | Effectiveness of XGBoost and DNNs |
| 7      | A Comparative Analysis of Cyber-Threat Intelligence Sources, Formats and Languages | A. Ramsdale, S. Shiaeles, and N. Kolokotronis | 2020 | Cyber-threat intelligence analysis | Comparative analysis | Evaluation of intelligence sources |
| 8      | Threat intelligence sharing between cybersecurity vendors              | A. Zrahia                            | 2018 | Cybersecurity collaboration | Network analysis    | Benefits of intelligence sharing |
Research Gap

- Existing studies lack real-time analysis of cyber threat data for predictive supply chain security measures.
- Comparative analysis of machine learning classifiers in supply chains remains superficial, without in-depth performance benchmarks.
- Research on integrating behavioral modeling with anomaly detection for comprehensive cyber defense is limited.
- Few studies address the efficiency of threat intelligence sharing across diverse cybersecurity vendors.
- There is a scarcity of research on agile response strategies incorporating big data analytics for incident response.
Proposed Methodology

01 Data Acquisition

02 EDA

03 Statistical Approach

04 Preprocessing

05 ML Algorithm

06 Model Evaluation

07 DL Algorithm

08 Prediction
Current Situation & problems

Current situation
Cyber threats are outstripping traditional defenses in supply chains, with sophisticated attacks causing disruptions and breaches. This necessitates a shift towards predictive security measures for better protection and resilience.

1. Vulnerability
Exposed to increasing cyber attacks.

2. Obsolescence
Outdated security systems easily breached.

3. Fragmentation
Disconnected security measures and weak links.
Result and Discussion

This dataset, exhibiting zero missing values across all features, greatly streamlines the preprocessing stage. Its completeness eliminates the need for imputation, ensuring data integrity and reliability. This optimal state of the dataset facilitates straightforward, accurate analysis and predictive modeling, enhancing the robustness of the derived results.

Pie chart for Target column

The pie chart shows an 87.1% to 12.9% imbalance in the `HasDetections` dataset, with most entries indicating "No Detection." This skewness may bias models towards the majority class.
Result and Discussion

Histograms of the dataset display varied distributions: right-skewed in "IsBeta" and "RtpStateBitfield," low variance in "IsSxsPassiveMode," and multimodality in "AVProductStatesIdentifier." "HasTpm" and "CountryIdentifier" show bimodal patterns, suggesting distinct groups. Uniform distributions and potential outliers in "OrganizationIdentifier" are observed, while "HasDetections" is sharply binary. These varying tendencies require specific preprocessing for effective analysis and modeling.
The correlation matrix reveals `RtpStateBitfield` and `HasDetections` have a strong correlation (0.828848), indicating a predictive link. High correlations like `HasTpm` and `CountryIdentifier` (0.937366) suggest multicollinearity concerns. Moderate correlations with `HasDetections` point to key predictive features, while low correlations signal the need for selective feature inclusion in models.
SMOTE addresses class imbalance in datasets by generating synthetic minority class samples, interpolating between selected instances and neighbors. This balances class distribution, aiding algorithm learning. However, excessive synthetic sample generation risks overfitting, necessitating careful application.
## Model Summary

| Metric                          | Value          | Metric                          | Value          |
|---------------------------------|----------------|---------------------------------|----------------|
| Accuracy                        | 0.999950606406506 | Accuracy                        | 0.9999810214268091 |
| Precision                       | 0.9999208234362629 | Precision                       | 0.999978317634859 |
| Recall                          | 0.99980280280963465 | Recall                          | 0.9999214454045562 |
| F1 Score                        | 0.9998614492409398 | F1 Score                        | 0.999947628983739 |
| Specificity                     | 0.9999825946199971 | Specificity                     | 0.999942053843569 |
| Negative Predictive Value       | 0.999954876880152 | Negative Predictive Value       | 0.9999826163545337 |
| False Positive Rate             | 1.74053800022958916e-05 | False Positive Rate             | 5.79461564314439e-06 |
| False Negative Rate             | 0.0001979179036535645 | False Negative Rate             | 7.85549544383346e-05 |

### Decision tree with and without K-best

| Metric                          | Value          | Metric                          | Value          |
|---------------------------------|----------------|---------------------------------|----------------|
| Accuracy                        | 0.9998430477277591 | Accuracy                        | 0.999966787496916 |
| Precision                       | 0.9998811740008714 | Precision                       | 0.9999476261554979 |
| Recall                          | 0.9992479119661165 | Recall                          | 0.999869756742603 |
| F1 Score                        | 0.99956442684617  | F1 Score                        | 0.9999083493721932 |
| Specificity                     | 0.9999738919299955 | Specificity                     | 0.999884107687137 |
| Negative Predictive Value       | 0.9998346719107576 | Negative Predictive Value       | 0.9999710274254391 |
| False Positive Rate             | 2.610807000443837e-05 | False Positive Rate             | 1.158923128628878e-05 |
| False Negative Rate             | 0.0007520880338835451 | False Negative Rate             | 0.00013092432573972245 |

### Random forest with and without K-best
# Model Summary

### Gradient boosting with and without K-best

| Metric                        | Value           |
|-------------------------------|-----------------|
| Accuracy                      | 0.9999001212813012 |
| Precision                     | 0.9998812116412592  |
| Recall                        | 0.9995645806119622  |
| F1 Score                      | 0.9997228710558613  |
| Specificity                   | 0.999378919299955  |
| Negative Predictive Value     | 0.9999042770743594  |
| False Positive Rate           | 2.61080700443837e-05  |
| False Negative Rate           | 0.0004354193880378419  |

### Logistic Regression with and without K-best

| Metric                        | Value           |
|-------------------------------|-----------------|
| Accuracy                      | 0.9999810214268091  |
| Precision                     | 0.999476302697841  |
| Recall                        | 0.999476302697041  |
| F1 Score                      | 0.999476302697041  |
| Specificity                   | 0.999884107687137  |
| Negative Predictive Value     | 0.999884107687137  |
| False Positive Rate           | 1.158923128628878e-05  |
| False Negative Rate           | 5.2369730295888974e-05  |
## Model Summary

| Metric                      | Value          | Metric                      | Value          |
|-----------------------------|----------------|-----------------------------|----------------|
| Accuracy                    | 0.9997645715916387 | Accuracy                    | 0.9997295553320301 |
| Precision                   | 0.9990505577972941  | Precision                   | 0.9998689143486355 |
| Recall                      | 0.9996437477734236 | Recall                      | 0.9986383870123069 |
| F1 Score                    | 0.999347647593045 | F1 Score                    | 0.9992532718483487 |
| Specificity                 | 0.999791354399645 | Specificity                 | 0.9999710269217843 |
| Negative Predictive Value   | 0.999921665564182 | Negative Predictive Value   | 0.999698762027795 |
| False Positive Rate         | 0.00020886450003550697 | False Positive Rate         | 2.897307821572195e-05 |
| False Negative Rate         | 0.0003562522265764161 | False Negative Rate         | 0.0013616129876931134 |

### Cat boost with and without K-best

![Accuracy comparison of all model](image)

- **Accuracy** comparison of all models for the Cat boost with and without K-best.
Insights from model

- K-best boosts Logistic Regression precision.
- Ensemble methods robust, slightly improve post-selection.
- Feature selection vital for Logistic Regression.
- CatBoost’s performance dips post-feature selection.
- Careful selection essential for optimal performance.
- Feature selection’s impact varies by model.
Conclusion

➢ Comprehensive analysis of classifiers showed ensemble methods like Random Forest maintain accuracy, improving slightly with K-best feature selection.

➢ Logistic Regression significantly benefits from feature selection, highlighting its sensitivity to feature relevance for optimizing model performance.

➢ CatBoost Classifier's slight performance dip post-feature selection suggests potential information loss or initial optimization with the full feature set.

➢ The K-best method’s impact varies by model, demonstrating its effectiveness in enhancing model accuracy through focused feature reduction.

➢ Findings underscore the importance of tailored feature selection and model choice in machine learning for accurate, efficient predictive outcomes.
Future Scope

➢ Advanced feature engineering could yield more insightful variables and interactions, significantly improving model performance and predictive accuracy.

➢ Implementing grid or randomized search for hyperparameter tuning across models promises optimized performance and finer-tuned results.

➢ Employing model stacking techniques, by combining multiple model predictions, could surpass the performance of individual models.

➢ Integrating domain-specific knowledge into feature engineering and model selection may offer more contextually accurate and relevant results.

➢ Exploring deep learning models, given adequate data and resources, might reveal complex patterns missed by simpler models, enhancing predictions.
Thanks!

Do you have any questions?