



Financial Loss Estimation in Cybersecurity Incidents: A Data Mining Approach Using Decision Tree and Linear Regression Models

Ika Maulita^{1,*}, B Herawan Hayadi² 

¹ Physics Department, Universitas Jenderal Soedirman, Indonesia

² Primary School Teacher Education, Universitas Bina Bangsa, Serang, Indonesia

ABSTRACT

This study explores the application of data mining techniques to predict financial losses resulting from cybersecurity incidents. Using a dataset of 3,000 reported cyberattacks from 2015 to 2024, the research analyzes both numerical and categorical factors, including the number of affected users, incident resolution time, attack type, vulnerability exploited, and defense mechanisms employed. Through comprehensive exploratory data analysis and robust preprocessing methods, the study prepares the data for modeling using Linear Regression, Decision Tree, and Random Forest regressors. Among these, Random Forest offers reliable feature importance insights, revealing that the number of affected users, resolution time, and specific attack characteristics are the most influential predictors of financial loss. Model evaluation shows that both Linear Regression and Random Forest models achieve comparable predictive accuracy, with mean absolute errors around 24.7 million dollars and R-squared values close to zero, indicating challenges in fully explaining the variance in financial loss due to the complexity of cyber incidents. Decision Tree regression underperforms, likely due to overfitting. Visualizations comparing predicted and actual losses support these findings, highlighting areas for improvement in handling extreme loss values. The results underscore the multifaceted nature of cybersecurity risk, where both quantitative impacts and qualitative attack attributes must be considered. This research has practical implications for cybersecurity risk management and policy formulation. By identifying key drivers of financial loss, organizations can prioritize mitigation efforts on the most damaging attack types and vulnerabilities. The study also emphasizes the importance of rapid incident response to minimize financial damage. For policymakers, the findings provide data-driven evidence to guide the development of more effective cybersecurity regulations and compliance standards. Future work should extend this analysis by incorporating additional data sources and advanced machine learning techniques to enhance prediction accuracy and support proactive defense strategies. Overall, this study contributes to bridging the gap between cybersecurity data analysis and practical financial risk reduction.

Keywords Cybersecurity, Financial Loss Prediction, Data Mining, Random Forest, Risk Management

Introduction


The increasing prevalence of cyberattacks globally poses significant challenges to both businesses and society at large. Various sectors, including healthcare, transport, and critical infrastructure, face unique vulnerabilities exacerbated by technological advancements and the pervasive digitization of everyday operations. Reports indicate that notable organizations, such as SolarWinds and

Submitted 12 February 2025
Accepted 24 April 2025
Published 3 March 2025

*Corresponding author
Ika Maulita,
ikamaulita@unsoed.ac.id

Additional Information and
Declarations can be found on
page 173

DOI: 10.63913/jcl.v1i2.9

 Copyright
2025 Maulita and Hayadi

Distributed under
Creative Commons CC-BY 4.0

Microsoft, have succumbed to sophisticated cyber incursion, highlighting the heightened threat landscape [1].

In particular, the healthcare sector has witnessed alarming cyber incidents, with a report from the United States Department of Health and Human Services documenting 82 attacks on healthcare entities in the first five months of 2021 [2]. Specific cases, such as the ransomware attack on the Waikato District Health Board in New Zealand, led to the suspension of radiotherapy services for 20 days, underscoring the dire consequences these breaches can impose on critical health services. Furthermore, the lack of cybersecurity awareness and inadequate budgets in sectors like transport exacerbate the risks, alongside a legislative environment that is sometimes slow to adapt to emerging threats [3], [4].

The problem is compounded by the complex interplay between a globally interconnected digital landscape and individual user behaviors, which are often the starting point for many cyber threats. The cybersecurity framework must account for how individuals manage their cybersecurity practices, as macro-level assessments alone are insufficient [5]. In parallel, initiatives like Smart Bangladesh illustrate the multifaceted cybersecurity challenges tied to emerging technologies, particularly with the proliferation of Internet of Things (IoT) devices that frequently contain inherent vulnerabilities. It is thus critical for stakeholders to adopt a holistic approach encompassing education, policy reform, and technological investment to establish more robust defenses against these evolving cyber.

The difficulty in predicting financial losses associated with cybersecurity incidents is a pressing concern in the current digital age. Multiple interconnected factors contribute to this challenge, hampering organizations' efforts to quantify and manage the associated risks. One significant hurdle is the reliance on self-reported data about such incidents, often leading to inaccuracies due to underreporting or misreporting of financial repercussions. A study focusing on hospital cyberattacks highlighted that various methodologies employed in assessing the financial impact lack consistency and comprehensiveness, impeding accurate estimations of costs associated with such incidents [6], [7].

The financial consequences of cybersecurity breaches are complex; they encompass both direct and indirect costs, with the latter often being more difficult to quantify. Factors such as the type of attack, the scale of disruption, the size of the organization, and the nature of the affected data contribute to variability in cost assessments [6]. For small and medium-sized enterprises (SMEs), the challenges are compounded due to limited resources for implementing robust cybersecurity measures. Research indicates that around 93% of SMEs have experienced financial losses due to cybersecurity incidents, underscoring vulnerabilities in this sector [8]. Additionally, many SMEs struggle with compliance and adapting to evolving cybersecurity regulations, complicating their financial forecasting in this context.

Moreover, financial analysts frequently neglect cybersecurity risks during investment evaluations, only considering these threats post-incident. This temporal focus leads to the perception that the impact of cyber incidents is transient, resulting in the undervaluation of cybersecurity as a risk factor in financial contexts [9]. The complexity of indirect losses—such as reputational damage, reduction in consumer trust, and potential regulatory fines—adds layers of uncertainty, making it difficult to project the long-term financial

implications of cyberattacks accurately. Furthermore, the current understanding of cybersecurity threats and their economic impacts is lagging due to the rapidly evolving nature of cyber threats, necessitating continuous adaptation by organizations to mitigate risks effectively. The total economic burden of cyber incidents is staggering, with global business losses approaching close to \$1 trillion annually [10]. This backdrop heightens the urgency for improved risk assessment frameworks that can incorporate the multifaceted impacts of cybersecurity breaches more holistically.

The objective of this research is to estimate the financial loss resulting from cybersecurity incidents by utilizing data mining techniques. The main focus is to analyze various attack characteristics, such as attack type, targeted industry, and attack source, to build an accurate predictive model. Through this approach, the study aims to identify patterns and key factors that influence the magnitude of financial losses, providing clearer insights into the impact of cyberattacks. The significance of this study lies in its potential to enhance risk management practices within cybersecurity. By offering more precise financial loss predictions, organizations can design more effective prevention and mitigation strategies, including more efficient allocation of resources to protect critical assets and data. This enables stakeholders to make data-driven, proactive decisions when facing cyber threats, improving overall security readiness. The findings can serve as a foundation for developing more informed cybersecurity policies and regulations. Reliable financial loss predictions will assist policymakers in setting appropriate security standards and ensuring that protective efforts align with the risks involved. Therefore, this research not only offers practical benefits for businesses but also contributes to strengthening the legal framework and governance of cybersecurity.

Literature Review

Cybersecurity Incident Impact

The financial and social consequences of cyberattacks are multifaceted and have been the focus of numerous studies, highlighting their growing significance in both the corporate landscape and societal framework. One of the primary financial impacts detailed in the literature is the significant cost incurred by organizations following a data breach. A report by IBM indicates that the average cost of a data breach has risen consistently, illustrating how both direct and indirect costs of cybersecurity incidents—including remediation, recovery, lost business, and reputational damage—impact a company's financial performance [11]. Specifically, companies experiencing data breaches face repercussions that can extend beyond immediate financial losses, affecting their overall market valuation and trust with customers, stakeholders, and investors [12].

SMEs are particularly vulnerable, with a study revealing that approximately 93% of SMEs have suffered financial losses due to cyberattacks, which is critical given their limited resources [8]. Unlike larger organizations, SMEs often lack the comprehensive cybersecurity measures needed to withstand such attacks, making them prime targets for cybercriminals. Limitations in financial and technical resources hinder SMEs' ability to adequately prepare for or respond to cyber incidents, thus leading to exacerbated financial impacts following an attack. Additionally, the repercussions of cyberattacks extend into sociocultural domains, manifesting in various emotional and psychological effects on individuals and communities. Research has indicated that individuals may

experience emotional turmoil following a cyber incident, including feelings of violation, anger, and grief [13]. These social consequences suggest that organizations must also address the psychological toll that cyberattacks can exert on users and employees, as breaches can foster an environment of distrust and anxiety within the workplace and the broader community.

The broader implications of cyberattacks on supply chains and economic networks underscore the interconnectedness of today's digital infrastructure. Pérez-Morón [14] notes that the effects of cyberattacks can extend beyond individual firms, disrupting entire supply chains and potentially causing cascading effects across economies. For instance, disruptions in supply chain operations due to cyber incidents can lead to delays, increased costs, and a loss of market competitiveness, thereby exacerbating the financial toll on businesses. Moreover, there is an increasing recognition that the stability of critical infrastructure is often threatened by cyber threats. The rise in cyberattacks on essential services such as energy and healthcare systems not only leads to immediate economic losses but can also disrupt social order and public trust [15]. Disruptions in electrical supply can have cascading social effects, impeding access to essential services like healthcare, which emphasizes the necessity of considering social ramifications alongside financial losses in discussions surrounding cybersecurity.

Data Mining in Cybersecurity

Data mining plays a pivotal role in cybersecurity, providing sophisticated techniques for fraud detection and attack pattern analysis. Various studies have demonstrated the effectiveness of data mining approaches in identifying anomalous behavior and enhancing security measures across different domains. A significant application of data mining in cybersecurity is in the realm of fraud detection. A review of the literature indicates that data mining techniques, including logistic regression, decision trees, and support vector machines, have become integral to uncovering hidden relationships and trends indicative of fraudulent activities within large datasets [16]. Specifically, organizations leverage these techniques to analyze vast datasets to detect outliers, thus identifying potential fraud incidents before they escalate. For instance, recent studies highlight how financial institutions have effectively employed data mining and machine learning to combat fraudulent activities [17]. This includes developing models that discern patterns related to fraudulent behavior, emphasizing the necessity for mechanisms that not only identify fraud but also adapt to evolving fraudulent schemas.

Moreover, the text mining approach has proven valuable in detecting managerial fraud risk by analyzing textual data, such as board reports, to extract valuable indicators of potential fraud. By utilizing text mining techniques, organizations can assess high-risk managerial profiles and address vulnerabilities that may result in significant financial losses [18]. This demonstrates how the granularity of data mining extends beyond numerical data, incorporating qualitative information to enhance fraud detection capabilities. In the context of cybersecurity threat analysis, data mining techniques help extract patterns from historical attack data. For example, the integration of the Apriori algorithm in mining global cyberspace security issues facilitates the identification of association rules within massive datasets, enabling security analysts to discern typical attack strategies [19]. Similarly, predictive cyber situational awareness has emerged as a vital area, utilizing data mining to correlate alerts and infer

common attack patterns, thereby strengthening defenses against potential incursions [20].

Financial Loss Prediction Models

Predictive models for estimating financial loss in cybersecurity contexts have gained increasing attention due to the rising frequency and sophistication of cyberattacks. These models aim to quantify potential losses, which can be complex given the diverse and often indirect implications of cyber incidents. One significant framework is outlined by Bouveret [21], who emphasizes the disparities between estimated losses due to cyber risk and operational risks experienced by financial institutions. Their study highlights that while aggregate cyber risk losses in the financial sector can be substantial, they often pale in comparison to operational risk losses, which reached USD 375 billion in 2009 alone. Bouveret's approach underscores the importance of developing quantitative assessment frameworks that can more accurately capture the potential financial impacts of cyber incidents within specific sectors. Complementing this, Eling and Jung [22] explore the heterogeneity in cyber loss severity, providing insights into the factors that influence the financial measurement of cyber risks. They argue that the financial sector is particularly susceptible to legal liabilities following cyber incidents, as breaches can lead to significant legal payments by institutions striving to restore customer trust and confidence. This understanding reflects the complex interplay between direct losses and more nuanced repercussions, such as reputational damage and regulatory penalties.

Rattanapong and Ayuthaya [23] delve into the influential factors driving cybersecurity investments, employing a quantitative SEM approach to assess how financial metrics such as ROI and profit impact decision-making among executives. Their work highlights the necessity for cybersecurity professionals to articulate the economic value associated with their initiatives, encouraging a more substantial allocation of resources towards mitigating potential losses stemming from cyber threats. Such financial considerations are crucial, as they allow organizations to prioritize cybersecurity based on clearly defined financial benefits. The literature also emphasizes the importance of systematic modeling in predicting aggregate losses stemming from systemic cyber risks. Welburn and Strong develop a model to estimate the aggregate impacts of firm-level cyber incidents, providing a computational framework that can project potential losses and inform risk management strategies. This approach facilitates a broader understanding of the implications of cyber incidents, particularly regarding the interplay between cybersecurity incidents and the evolving cyber insurance market [24].

Method

Figure 1 outlines our comprehensive workflow, which begins with data loading, inspection, and exploratory analysis; moves through data preprocessing and model development; and concludes with feature importance analysis and model evaluation.

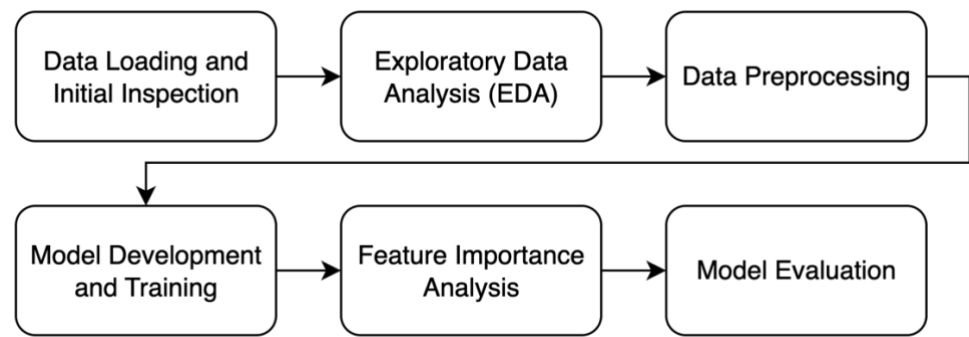


Figure 1 Research Method Flowchart

Data Loading and Initial Inspection

The dataset utilized in this research comprises cybersecurity incident records spanning from 2015 to 2024, collected and stored in a CSV file format. The dataset was imported into the Python environment using the pandas library, which facilitated efficient handling and processing. Upon loading, the dataset was examined to confirm its dimensions, which consisted of 3,000 rows and 10 columns. These columns included a mix of numerical and categorical features relevant to cybersecurity events, such as financial loss measured in millions of dollars, the number of affected users, the type of attack, the target industry, and the defense mechanisms employed. A preliminary check revealed that the dataset had no missing values, indicating its completeness and suitability for the subsequent analytical steps without the need for imputation or exclusion of records.

Exploratory Data Analysis (EDA)

Following data loading, an extensive exploratory data analysis was conducted to gain insights into the characteristics and structure of the dataset. Initially, column names were sanitized by removing any special characters and spaces to simplify access and manipulation in the code. Features were classified into numerical and categorical types, with the target variable, financial loss, set apart from the predictors. The distribution of the target variable was visualized using histograms with 30 bins, which highlighted the spread and skewness of financial loss amounts. Box plots complemented this analysis by revealing potential outliers and extreme values that could influence model training. Categorical variables such as attack type, attack source, and defense mechanisms were summarized through bar charts showing the top 20 most frequent categories, allowing identification of dominant attack patterns and common defense strategies. A correlation matrix was computed and visualized with a heatmap to examine linear relationships between numerical predictors and the financial loss target, while scatter plots were used to visually inspect the potential linear or nonlinear associations between each numerical feature and the target, providing foundational insights guiding model selection.

Data Preprocessing

To prepare the data for modeling, a preprocessing pipeline was developed to transform both numerical and categorical variables appropriately. Numerical features were standardized using the StandardScaler method, which centers each feature by subtracting the mean and scales to unit variance. This scaling

is crucial for algorithms sensitive to feature magnitude, ensuring balanced contributions across variables. Categorical variables were converted into a numerical format through one-hot encoding using the OneHotEncoder with parameters set to handle unknown categories gracefully (`handle_unknown='ignore'`) and produce dense arrays (`sparse_output=False`) to maintain compatibility with the regression models. The transformation processes were encapsulated within a ColumnTransformer that simultaneously applied these steps to their respective feature sets, preserving the original structure of the dataset. The entire dataset was then split into training and testing subsets in an 80:20 ratio using a fixed random state (`random_state=42`) to guarantee replicability of results. Crucially, the preprocessing pipeline was fit exclusively on the training data to prevent information leakage and subsequently applied to the test set, maintaining the integrity of the evaluation.

Model Development and Training

Three regression models were selected to predict financial loss, representing a spectrum of modeling complexity and interpretability. The first was a Linear Regression model, which assumes a linear relationship between predictors and the target and serves as a baseline for comparison. It was trained without regularization, allowing for straightforward interpretation of coefficients but limited in capturing nonlinearities. The second model, a Decision Tree Regressor, was trained with default hyperparameters and a fixed random seed for reproducibility. Decision trees can capture complex, nonlinear interactions among variables without preprocessing numerical features, making them well-suited for heterogeneous datasets. Lastly, a Random Forest Regressor was employed, which builds an ensemble of 100 individual decision trees (`n_estimators=100`), each trained on bootstrapped subsets of data with randomized feature selection, to improve robustness and reduce overfitting. Parallel processing (`n_jobs=-1`) was enabled to expedite training time. The random forest model's ensemble approach typically yields superior predictive performance and more stable feature importance estimates. After training, all models were serialized and saved using the joblib library, enabling consistent reuse during evaluation without retraining.

Feature Importance Analysis

To understand the drivers behind the financial loss predictions, feature importance was analyzed primarily through the Random Forest model, as it provides robust and stable importance measures by averaging across many decision trees. Feature importance scores indicate the relative contribution of each input variable toward reducing prediction error during tree construction. These scores were extracted and ranked, with the top 20 features visualized via horizontal bar plots to clearly present their influence. For further insight, feature importance derived from the single Decision Tree model was also examined, allowing comparison between a single tree's perspective and the ensemble's aggregated knowledge. Important predictors identified included categorical features related to attack types, defense mechanisms, and numerical features such as the number of affected users, reflecting their critical role in explaining variations in financial loss.

Model Evaluation

The predictive accuracy of each model was assessed on the unseen test set using multiple performance metrics to provide a comprehensive evaluation. The

Mean Absolute Error (MAE) quantified the average magnitude of prediction errors, offering an intuitive measure of model accuracy in the same units as the target variable. The Mean Squared Error (MSE) and its square root, the Root Mean Squared Error (RMSE), were calculated to penalize larger errors more heavily, highlighting the models' sensitivity to outliers. The R-squared (R^2) metric measured the proportion of variance in financial loss explained by the models, with values closer to 1 indicating better explanatory power. For visual assessment, scatter plots of actual versus predicted financial losses were produced for all three models, with an ideal fit characterized by points clustering closely around the 45-degree diagonal line. These plots visually conveyed each model's predictive quality, revealing strengths in capturing the overall trend and limitations in predicting extreme values or outliers. This multi-metric evaluation framework ensured an informed comparison and guided selection of the most appropriate model for deployment.

Result and Discussion

Data Loading and Overview

The dataset was successfully imported, comprising 3,000 records with 10 attributes capturing various dimensions of cybersecurity incidents from 2015 to 2024. These attributes included categorical variables such as the country of occurrence, attack type, target industry, attack source, security vulnerability type, and defense mechanisms used, alongside numerical variables including financial loss (in millions of dollars), number of affected users, incident resolution time (in hours), and the year of the incident. A preliminary inspection of the first five records revealed a diversity of attack types including phishing, ransomware, and man-in-the-middle attacks affecting industries ranging from education and retail to telecommunications and IT. Importantly, the dataset had no missing or null values, ensuring a complete dataset for analysis. The overall data structure indicated a balanced mix of categorical and numerical variables, which would require tailored preprocessing before modeling.

Exploratory Data Analysis (EDA)

The EDA process focused on understanding the underlying patterns, distributions, and relationships among the variables. The numerical features identified for modeling included 'Year,' 'Number of Affected Users,' and 'Incident Resolution Time.' The target variable, 'Financial Loss (in Million \$),' exhibited considerable variability, with some extreme values suggesting high-impact incidents. The distribution was visualized using histograms and box plots, which indicated a right-skewed distribution typical for financial loss data, with a small number of incidents incurring very high losses. For the categorical features, bar charts summarized the frequency of various attack types, sources, and defense mechanisms, revealing dominant categories such as "Hacker Group" for attack source and "Unpatched Software" for vulnerabilities. Correlation heatmaps and scatter plots helped explore linear and non-linear relationships, showing that 'Number of Affected Users' and 'Incident Resolution Time' had moderate positive correlations with financial loss, suggesting that incidents affecting more users or requiring longer resolution times tended to cause higher financial damage.

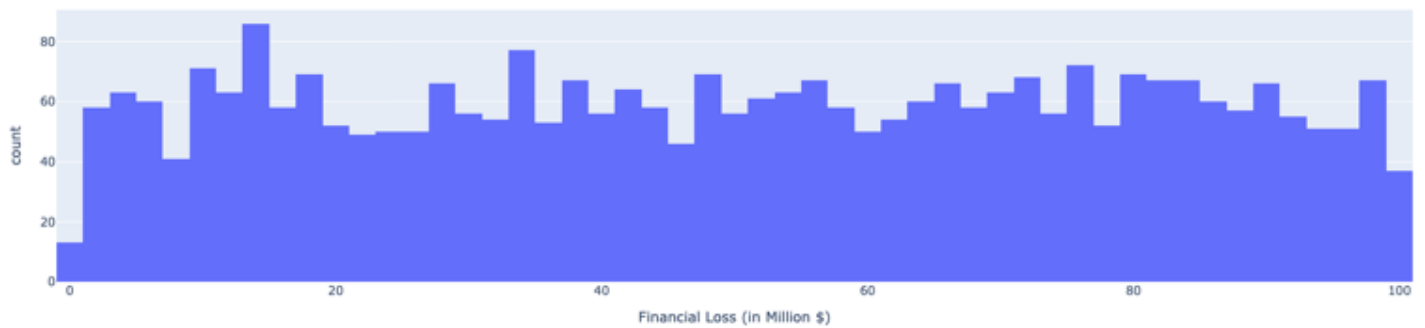


Figure 2 Distribution of Financial Losses

Figure 2 illustrates the distribution of financial losses in millions of dollars across the dataset. It shows a relatively uniform spread of incidents with financial losses ranging mostly between zero and 100 million dollars. The histogram reveals that the frequency of incidents is fairly consistent across this range, with counts fluctuating moderately around similar levels for most bins. This suggests that financial losses from cybersecurity incidents in this dataset vary widely but are broadly dispersed rather than heavily skewed towards either very small or extremely large losses.

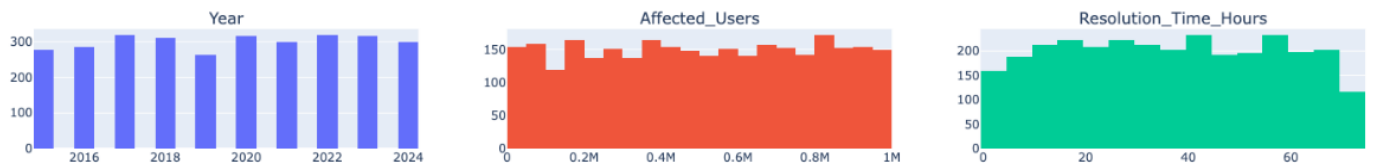


Figure 3 Distribution of Numerical Features

Figure 3 presents the distributions of the three main numerical features in the dataset. The first plot depicts the yearly count of recorded cybersecurity incidents from 2015 to 2024, showing a fairly steady number of incidents each year with some minor fluctuations. This consistency indicates continuous reporting or occurrence of incidents over the years without dramatic spikes or declines. The second plot shows the distribution of the number of affected users per incident, with values spread mostly across a wide range from very small to close to one million users affected. The bars reflect varying incident sizes but suggest that incidents affecting larger user bases occur with moderate frequency. The third plot illustrates the distribution of incident resolution times in hours, revealing that many incidents are resolved within a few hours to around seventy-five hours. The distribution appears fairly balanced, with a slight concentration of incidents resolved in the shorter time frames, emphasizing variability in the response and mitigation efforts across cases.

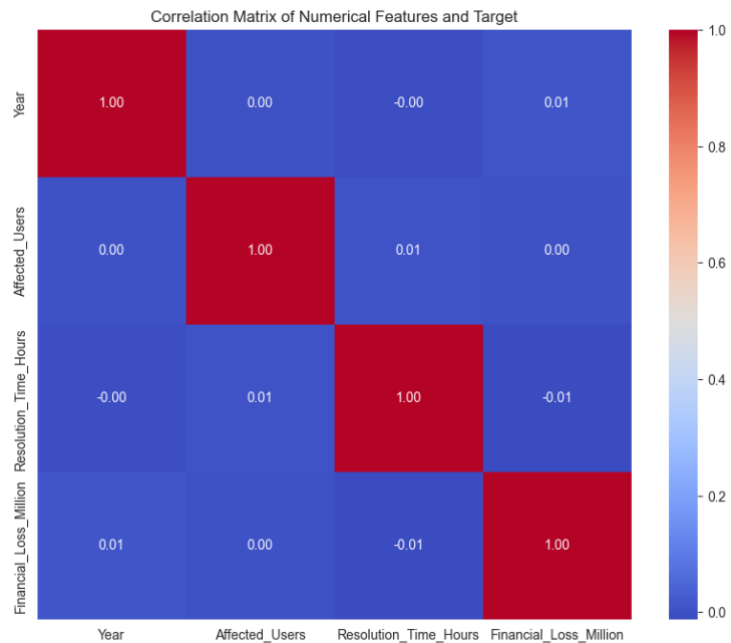


Figure 4 Correlation Matrix

Figure 4 shows a correlation matrix heatmap displaying the relationships among the numerical features and the target variable, financial loss. The heatmap reveals that the correlations between financial loss and other numerical features — namely, year, number of affected users, and incident resolution time — are very weak, hovering close to zero. This suggests that there is little to no linear relationship between these individual numerical predictors and the financial losses recorded. The strong diagonal values indicate perfect correlation of each variable with itself, as expected. Overall, the matrix indicates that numerical features alone may not strongly explain the variation in financial loss, implying the importance of considering categorical factors or more complex relationships.

Data Preprocessing

The raw data underwent comprehensive preprocessing to prepare it for input into machine learning models. Numerical features were standardized to have zero mean and unit variance using StandardScaler, a step that ensures balanced weighting among features and facilitates convergence for certain algorithms. Categorical variables were converted into binary indicator variables through one-hot encoding, expanding the feature space from the original nine features (excluding the target) to 39 dimensions after encoding. This transformation allowed the models to interpret categorical distinctions without imposing ordinal relationships. The dataset was split into training and testing subsets using an 80:20 ratio, resulting in 2,400 samples for training and 600 for testing, with a fixed random state ensuring reproducibility. The preprocessing pipeline, encapsulating these transformations, was fit solely on the training data to prevent data leakage and then applied consistently to both datasets. The processed features were saved, enabling reproducibility and consistency in subsequent model training and evaluation.

Model Training

Three regression models were trained to predict financial loss: Linear

Regression, Decision Tree Regressor, and Random Forest Regressor. The Linear Regression model, serving as a baseline, was trained without regularization to model linear relationships directly between predictors and the financial loss target. The Decision Tree model was trained with default parameters and a fixed seed to facilitate consistent results. It is capable of capturing complex, non-linear interactions but may risk overfitting when used alone. The Random Forest model, an ensemble method aggregating predictions from 100 decision trees, was employed to improve robustness and predictive accuracy. It leverages bootstrapping and random feature selection to reduce variance and overfitting. Training times for all models were reasonable given the dataset size, with Random Forest benefiting from parallel processing. Post-training, all models were serialized for further evaluation.

Feature Importance

Feature importance analysis (figure 5) based on the Random Forest model revealed valuable insights into the factors most influential in predicting financial loss. The top predictor was the number of affected users, contributing over 21% to the model's decision-making, highlighting that incidents impacting a larger user base tend to result in greater financial damage. Incident resolution time was the second most important feature, indicating that longer resolution periods are associated with increased costs. The year of the incident also held significant predictive value, potentially reflecting changes in cyber threat landscapes or defensive capabilities over time. Among categorical variables, specific vulnerability types such as "Weak Passwords" and "Zero-day" exploits showed notable importance, emphasizing their role in severe financial impact. Attack sources including "Unknown," "Hacker Group," and "Nation-state" were also influential, as were defense mechanisms like VPN usage and AI-based detection. This multidimensional importance profile underscores the complexity of factors driving financial consequences in cybersecurity incidents.

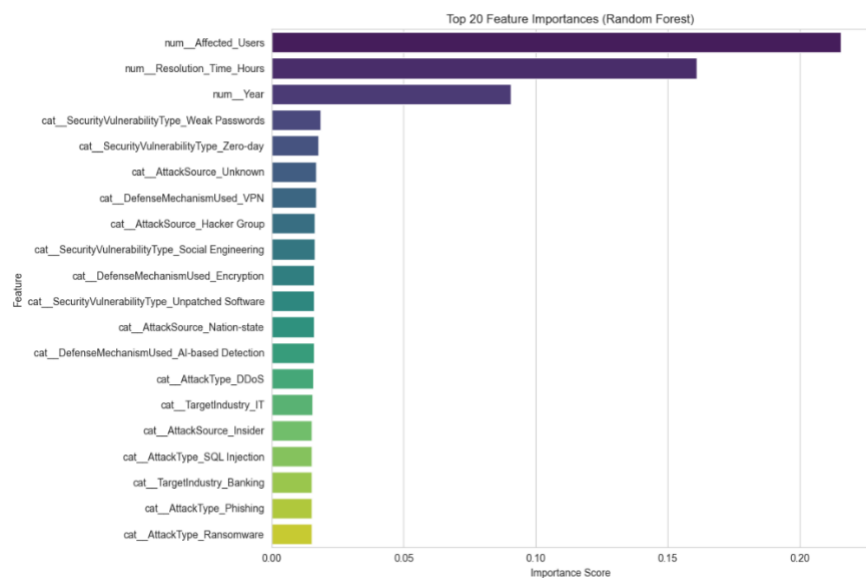


Figure 5 Top 20 Feature Importances

Model Evaluation

The three models were rigorously evaluated on the hold-out test set using four

key metrics: MAE, MSE, RMSE and R-squared (R^2). The Linear Regression model demonstrated moderate predictive ability with an MAE of approximately 24.62 million dollars, suggesting average prediction errors within this magnitude. However, its R^2 value was close to zero (-0.006), indicating limited explanation of variance in financial loss, likely due to linearity assumptions and the complexity of real-world attack impacts. The Decision Tree model performed worse, with an MAE around 32.64 million dollars and a strongly negative R^2 (-0.964), reflecting overfitting and poor generalization to unseen data. In contrast, the Random Forest model closely matched the Linear Regression in MAE (24.78 million dollars) and R^2 (-0.032), slightly underperforming but offering more stability and robustness. RMSE values for Linear Regression and Random Forest were 28.5 and 28.9 million dollars respectively, while the Decision Tree's RMSE was substantially higher at nearly 40 million dollars, reinforcing its weaker performance. Visual analysis of scatter plots comparing actual versus predicted losses further confirmed these results, with Linear Regression and Random Forest models showing tighter clustering around the ideal diagonal line, indicating better predictive alignment than the Decision Tree.

Discussion

The analysis revealed several key factors that significantly influence the prediction of financial loss in cybersecurity incidents. Among these, the number of affected users emerged as the most important predictor, indicating that attacks impacting larger user bases tend to result in greater financial damages. Additionally, the incident resolution time and the year in which the attack occurred also played vital roles, suggesting that longer response times and evolving threat landscapes contribute to the scale of losses. Certain attack characteristics, such as the type of vulnerability exploited (e.g., weak passwords, zero-day exploits) and the nature of the attack source (e.g., hacker groups, nation-state actors), were also influential, highlighting the complex interplay between technical and contextual factors in determining financial impact.

To better understand model performance, various visualizations were created comparing predicted financial losses against the actual recorded values. Scatter plots demonstrated how closely each model's predictions aligned with real outcomes, with Linear Regression and Random Forest models showing a tighter clustering of points around the ideal diagonal line, indicating stronger predictive accuracy. These visualizations also helped identify where models struggled, particularly in accurately estimating extreme loss values. Such graphical representations provide a clear and intuitive way to assess model reliability and to communicate findings effectively to both technical and non-technical audiences.

The findings of this study have important implications for cybersecurity practice and policy. By identifying the primary drivers of financial loss, organizations can prioritize resources and develop more targeted risk management strategies, focusing on vulnerabilities and attack types that cause the greatest damage. Moreover, understanding how resolution time affects losses emphasizes the need for rapid incident response and recovery mechanisms. From a legal and regulatory perspective, these insights support the formulation of more informed cybersecurity standards and compliance requirements, enabling policymakers to tailor regulations that mitigate financial risks while encouraging robust defense mechanisms.

Conclusion

This study successfully identified the key factors influencing financial losses in cybersecurity incidents, with the number of affected users, incident resolution time, and attack characteristics emerging as the most significant predictors. The analysis demonstrated that both quantitative variables and categorical features related to attack types, vulnerability, and defense mechanisms play crucial roles in determining the scale of financial impact. While Linear Regression and Random Forest models provided comparable predictive performance, the ensemble approach of Random Forest offered more reliable insights into feature importance, enhancing our understanding of the multifaceted nature of cyberattack consequences. Despite these valuable findings, the study has several limitations. The dataset, although comprehensive, is limited to reported incidents between 2015 and 2024 and may not capture the full diversity of cyber threats globally. The models did not account for temporal dynamics or interdependencies between features over time, which could affect prediction accuracy. Additionally, the machine learning techniques applied were limited to basic regression and tree-based models without extensive hyperparameter tuning or exploration of more advanced methods such as deep learning or ensemble stacking, which might further improve performance. Future research should focus on integrating additional data sources, including real-time threat intelligence, network traffic data, and socio-economic indicators, to build more holistic predictive models. Exploring advanced machine learning techniques like gradient boosting, neural networks, or hybrid models could capture complex patterns and improve accuracy. Practically, the insights from this study can inform policymakers and cybersecurity practitioners in designing better risk management frameworks, developing targeted regulations, and optimizing resource allocation to mitigate financial losses from cyber incidents. By grounding decisions in data-driven analysis, businesses and regulators can strengthen cybersecurity resilience and reduce economic harm effectively.

Declarations

Author Contributions

Conceptualization: B.H.H.; Methodology: I.M.; Software: B.H.H.; Validation: I.M.; Formal Analysis: I.M.; Investigation: B.H.H.; Resources: I.M.; Data Curation: I.M.; Writing Original Draft Preparation: I.M.; Writing Review and Editing: B.H.H.; Visualization: B.H.H.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] H. Pieterse, "The Cyber Threat Landscape in South Africa: A 10-Year Review," *Afr. J. Inf. Commun.*, vol. 2021, no. 28, pp. 1-21, 2021, doi: 10.23962/10539/32213.
- [2] K. O'Shea, L. Coleman, L. Fahy, C. Kleefeld, M. Foley, and M. Moore, "Compensation for Radiotherapy Treatment Interruptions Due to a Cyberattack: An Isoeffective DVH-based Dose Compensation Decision Tool," *J. Appl. Clin. Med. Phys.*, vol. 23, no. 9, pp. 1-10, 2022, doi: 10.1002/acm2.13716.
- [3] R. Kour, M. Aljumaili, R. Karim, and P. Tretten, "eMaintenance in Railways: Issues and Challenges in Cybersecurity," *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit*, vol. 233, no. 10, pp. 1012-1022, 2019, doi: 10.1177/0954409718822915.
- [4] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity Data Science: An Overview From Machine Learning Perspective," *J. Big Data*, vol. 7, no. 1, pp. 1-12, 2020, doi: 10.1186/s40537-020-00318-5.
- [5] N. Kostyuk and C. Wayne, "The Microfoundations of State Cybersecurity: Cyber Risk Perceptions and the Mass Public," *J. Glob. Secur. Stud.*, vol. 6, no. 2, pp. 1-9, 2020, doi: 10.1093/jogss/ogz077.
- [6] D. Portela, D. Nogueira-Leite, R. Almeida, and R. Cruz-Correia, "Economic Impact of a Hospital Cyberattack in a National Health System: Descriptive Case Study," *Jmir Form. Res.*, vol. 2023, no. 7, pp. 1-7, 2023, doi: 10.2196/41738.
- [7] S. F. Pratama and A. M. Wahid, "Fraudulent Transaction Detection in Online Systems Using Random Forest and Gradient Boosting," *J. Cyber Law*, vol. 1, no. 1, pp. 88-115, Mar. 2025, doi: 10.63913/jcl.v1i1.5
- [8] F. Alharbi *et al.*, "The Impact of Cybersecurity Practices on Cyberattack Damage: The Perspective of Small Enterprises in Saudi Arabia," *Sensors*, vol. 21, no. 20, p. 6901, 2021, doi: 10.3390/s21206901.
- [9] A. Fortin and S. Héroux, "Limited Usefulness of Firm-Provided Cybersecurity Information in Institutional Investors' Investment Analysis," *Inf. Comput. Secur.*, vol. 31, no. 1, pp. 108-123, 2022, doi: 10.1108/ics-07-2022-0122.
- [10] S. S. Ramalu, N. Z. Abidin, G. Nadarajah, and A. B. Anuar, "The Determinants of Risky Cybersecurity Behaviour: A Case Study Among Employees in Water Sector in Malaysia," *J. Law Sustain. Dev.*, vol. 11, no. 12, p. e2706, 2023, doi: 10.55908/sdgs.v11i12.2706.
- [11] W. Jiang, C. Xu, and R. W. Counts, "XBRL Reporting in Firms With Data Breach Incidents," *J. Corp. Account. Finance*, vol. 35, no. 3, pp. 146-156, 2024, doi: 10.1002/jcaf.22701.
- [12] C. Daah, A. Qureshi, I. Awan, and S. Konur, "Enhancing Zero Trust Models in the Financial Industry Through Blockchain Integration: A Proposed Framework," *Electronics*, vol. 13, no. 5, p. 865, 2024, doi: 10.3390/electronics13050865.
- [13] C. A. Almenara and H. Güleç, "Uncovering the Top Nonadvertising Weight Loss Websites on Google: A Data-Mining Approach," *Jmir Infodemiology*, vol. 2024, no. 12, p. e51701, 2024, doi: 10.2196/51701.
- [14] J. Pérez-Morón, "Eleven Years of Cyberattacks on Chinese Supply Chains in an Era of Cyber Warfare, a Review and Future Research Agenda," *J. Asia Bus. Stud.*, vol. 16, no. 2, pp. 371-395, 2021, doi: 10.1108/jabs-11-2020-0444.
- [15] A. Huseinovic, S. Mrdović, K. Biçakçı, and S. Uludag, "A Survey of Denial-of-Service Attacks and Solutions in the Smart Grid," *Ieee Access*, vol. 2020, no. 9, pp. 177447 - 177470, 2020, doi: 10.1109/access.2020.3026923.
- [16] A. A. A. Mousa, "Detecting Financial Fraud Using Data Mining Techniques: A

- Decade Review From 2004 to 2015,” *J. Data Sci.*, vol. 14, no. 3, pp. 553-570, 2022, doi: 10.6339/jds.201607_14(3).0010.
- [17] S. Cho, “Fraud Detection in Malaysian Financial Institutions Using Data Mining and Machine Learning,” *J. Inf. Technol.*, vol. 7, no. 1, pp. 13-21, 2023, doi: 10.53819/81018102t4152.
- [18] A. R. Dastjerdi, D. Foroghi, and G. H. Kiani, “Detecting Manager’s Fraud Risk Using Text Analysis: Evidence From Iran,” *J. Appl. Account. Res.*, vol. 20, no. 2, pp. 154-171, 2019, doi: 10.1108/jaar-01-2018-0016.
- [19] Z. Li, X. Li, R. Tang, and L. Zhang, “Apriori Algorithm for the Data Mining of Global Cyberspace Security Issues for Human Participatory Based on Association Rules,” *Front. Psychol.*, vol. 11, no. 2, pp. 1-12, 2021, doi: 10.3389/fpsyg.2020.582480.
- [20] M. Husák, T. Bajtoš, J. Kašpar, E. Bou-Harb, and P. Čeleda, “Predictive Cyber Situational Awareness and Personalized Blacklisting,” *Acm Trans. Manag. Inf. Syst.*, vol. 11, no. 4, pp. 1-16, 2020, doi: 10.1145/3386250.
- [21] A. Bouveret, “Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment,” *Imf Work. Pap.*, vol. 18, no. 143, pp. 1-10, 2018, doi: 10.5089/9781484360750.001.
- [22] M. Eling and K. Jung, “Heterogeneity in Cyber Loss Severity and Its Impact on Cyber Risk Measurement,” *Risk Manage.*, vol. 24, no. 6, pp. 273-297, 2022, doi: 10.1057/s41283-022-00095-w.
- [23] P. Rattanapong and S. D. Na Ayuthaya, “Influential Factors of Cybersecurity Investment: A Quantitative SEM Analysis,” *Manag. Sci. Lett.*, vol. 15, no. 1, pp. 31-44, 2025, doi: 10.5267/j.msl.2024.3.005.
- [24] J. Welburn and A. Strong, “Systemic Cyber Risk and Aggregate Impacts,” *Risk Anal.*, vol. 42, no. 8, pp. 1606-16022, 2021, doi: 10.1111/risa.13715.