Cyber security attack prediction Using Machine Learning

Shathursan R
dept. of Electrical and Information Engineering
University of Ruhuna
Galle, Sri Lanka
shathursan_r_e22@engug.ruh.ac.lk

Himosh R
dept. of Electrical and Information Engineering
University of Ruhuna
Galle, Sri Lanka
himosh_r_e22@engug.ruh.ac.lk

***Abstract*—The objective of this project is to apply machine learning techniques to predict cyber security attacks. We have used the Kaggle website to gather data that is publicly available. The dataset was analysed and modelled using two algorithms, namely logistic regression and decision tree. The logistic regression technique achieved an impressive 90.16% accuracy, while the decision tree algorithm achieved 88.33% accuracy. This study contributes to the improvement of predictive modelling in cyber security by proving the effectiveness of machine learning algorithms in identifying possible risks and improving the overall security standing.**

I. Introduction

The research, named *"Cybersecurity Threat Prediction Model"*, makes use of machine learning to predict possible risks associated with the internet. In a time where digital environments are constantly changing, identifying threats accurately is essential to reducing security risks. By using an extensive dataset that contains temporal and geolocation data, the objective is to construct a prediction model that detects trends and abnormalities, hence augmenting cybersecurity safeguards. By overcoming the knowledge gap between proactive and reactive cybersecurity approaches, this research hopes to strengthen defenses against constantly changing cyber threats.

II. Methodology

A. Data Description

The dataset used in this research is sourced from Kaggle and is based on the number of cyber security attacks across various geographical places in India over the last three years (2020-2023). The dataset, which is available at Keggle, includes 25 features. The provided link ensures transparency and access for additional validation and testing. It is crucial to highlight that the data is true, accurate, and representative of cyber security attacks occurring in India during the provided time period. This extensive dataset provides a base for our analysis and exploration of machine learning models for predicting and detecting cyber security threats.

Identify applicable funding agency here. If none, delete this.

	Protocol	Malware Indicators	Attack Type	IDS/IPS Alerts	weekday_name	Device browser	Anomaly Scores Normalized
count	40000	40000	40000	40000	40000	40000	40000
unique	3	2	3	2	7	2	2
top	ICMP	IoC Detected	DDoS	No Alerts	Tuesday	Mozilla	high
freq	13429	20000	13428	20050	5813	31951	36007

Fig. 1. Data Description

B. Data Processing

Our dataset contains 40,000 records, with 25 attributes initially. To streamline our study, we intentionally reduced the features to a narrower collection of 8 significant attributes. The goal is to understand the interrelationships between these features.

A critical part of this is determining the correlation within the dataset. This includes determining how each feature interacts with the others, allowing us to analyze their impact on the overall prediction model. Furthermore, we carefully analyzed particular features, such as the Timestamp, which previously displayed extensive temporal information. To simplify and improve our knowledge, we separated the Timestamp into discrete components, specifically month and weekday, allowing for a more precise analysis of their interactions. By performing these steps, we hope to find the underlying patterns and relationships in the data, allowing for a more informed and effective development of our prediction model.

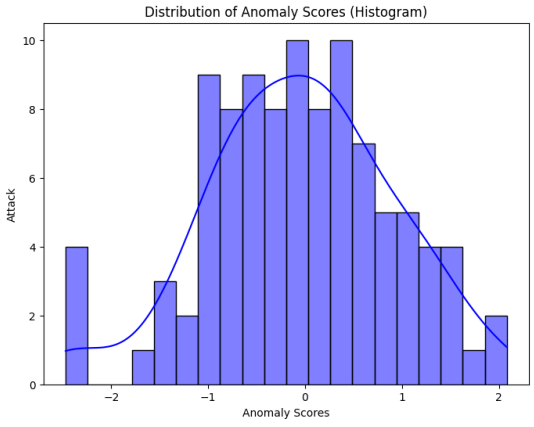

Fig. 2. Distribution of Anomaly Scoresn

C. Classification

To ensure data consistency and accuracy, we performed a complete pre-processing step before beginning analysis. This involved handling missing values, normalizing anomaly scores using MinMaxScaler, and addressing any outliers. Categorical variables were appropriately encoded, and the dataset was split into training and testing sets for model evaluation.

After shuffling the dataset, it is divided into 80% training and 20% testing sets. Afterwards, two classification machine learning algorithms, such as Logistic Regression and Decision Tree, were applied to develop a cyber security threat prediction model.

1) Logistic Regression: Logistic regression is one of the linear regression algorithms transformed using a sigmoid function to make it a classification algorithm for predicting probabilities of each class classifier capabilities[1].

2) Decision Tree: In addition to Logistic Regression, we implemented a Decision Tree classification model. Decision Trees are known for capturing complex in data. We conducted training, validation, and testing phases for the Decision Tree model, optimizing parameters for enhanced predictive performance. Evaluation metrics were employed to assess its accuracy and effectiveness in predicting cybersecurity threats[6].

III. Results

The outcomes of the training and testing phases of the Decision Tree and Logistic Regression models offer important new information about how well our cybersecurity threat prediction system is working.

With an astounding 89.98% accuracy on the training dataset and a strong 90.16% accuracy on the testing dataset, the Logistic Regression model showed stable accuracy levels. This implies that the model reliably predicts cybersecurity threats and that it generalises well to new, unseen data. A well-fitting model that successfully and non-overfittingly captures the underlying patterns in the data is shown by a little accuracy loss from training to testing.

Conversely, the Decision Tree model demonstrated remarkably high accuracy—99.80%—during the training period. Its testing accuracy did, however, show a minor decline to 88.33%.

The Decision Tree model may be more prone to overfitting the training data, capturing noise or outliers that do not translate well to new instances, as suggested by the difference between training and testing accuracies. To boost generalisation, more research and hyperparameter tweaking could be necessary to maximise the Decision Tree model.

In conclusion, the Logistic Regression model consistently exhibits excellent accuracy in both training and testing datasets, demonstrating its dependability in cybersecurity threat prediction.

Meanwhile, the Decision Tree model, while achieving exceptional accuracy in training, requires careful consideration and refinement to enhance its performance on unseen data. These results provide a foundation for future iterations and improvements in our cybersecurity prediction model, potentially incorporating ensemble methods or additional feature engineering to address any limitations identified.

IV. Discussion

The findings of our cybersecurity threat prediction experiment shed light on the efficacy of Logistic Regression and Decision Tree models in predicting possible security breaches. The Logistic Regression model demonstrated excellent consistency between the training and testing stages, with robust accuracy of 89.98% and 90.16%, respectively. This is consistent with the model's capacity to generalize effectively to fresh data, demonstrating its suitability for real-world cybersecurity applications. The small decrease in accuracy from training to testing indicates a well-balanced model capable of capturing the subtleties of cyber threats without succumbing to overfitting.

In comparison, the Decision Tree model achieved an impressive training accuracy of 99.80%, demonstrating a high ability to learn subtle patterns within the training dataset. However, the decline in testing accuracy to 88.33% indicates potential overfitting, in which the model is overly fitted to the complexities of the training data, limiting its generalizability to previously encountered occurrences. This study emphasizes the significance of continued tweaking and modification to improve the Decision Tree model's adaptability to a variety of cyber threat scenarios.

TABLE I
Performance Metrics of Classifiers

Classifier	**Training Accuracy**	**Testing Accuracy**
Logistic Regression	89.98%	90.16%
Decision Tree	99.80%	88.33%

V. Conclusion

In conclusion, this project provides a thorough examination of machine learning models for cybersecurity threat prediction. The Logistic Regression model emerges as a consistent and dependable performer, with good predictive ability. The Decision Tree model, while good at capturing complicated relationships in training data, requires rigorous tuning to assure its efficacy in real-world circumstances.

Moving forward, the insights acquired from this study will serve as a foundation for further refining our prediction model. Future revisions may include ensemble approaches, new feature engineering, or fine-tuning hyperparameters to improve the Decision Tree model's generalizability. This initiative adds significant knowledge to current efforts to strengthen cybersecurity defenses by providing a practical strategy to identifying and mitigating possible attacks in a changing digital ecosystem.

References

1) Brownlee, J. (2019) 'Logistic Regression Tutorial for Machine Learning', Machine Learning Mastery.

2) Prabhakaran, S. (2021) 'Logistic Regression – A Complete Tutorial With Examples in R', Machine Learning Plus.
3) Simplilearn (2023) 'An Introduction to Logistic Regression in Python',
4) TAE (n.d.) 'Logistic Regression', Tutorial And Example.
5) DataCamp (n.d.) 'Logistic Regression in R Tutorial',
6) HackerEarth (n.d.) 'Decision Tree Tutorials & Notes',