PREDICTING A BITCOIN HEIST RANSOMWARE ATTACK WITH DATA SCIENCE

Dr. K. Priya
Assistant Professor, Department of Computer Applications and Technology
SRM Arts and Science College, Kattankulathur, Tamil Nadu, India
kpriyaa@gmail.com
https://doi.org/10.34293/9789361639715.shanlax.ch.021

Abstract

Ransomware attacks have emerged as a significant factor contributing to malware incursion in the past few years. A suggested method for anticipating ransomware attacks involves using machine learning and artificial intelligence algorithms to extract information attributes [1]. Identifying variables and comprehending data are the initial steps in creating a successful model, and using the data science approach helps create a model that predicts results more accurately. The pre-processed data is subjected to a various machine learning algorithms, and the best performing algorithm is identified by comparing its accuracy. A machine learning model can forecast how a ransomware attack will turn out.

Keywords: *Bitcoin Heist, Ransomware, Pre-processing, Classification, Cryptocurrency, Voting classifier*

Introduction

Digital currencies intended to function outside of the conventional banking system are called cryptocurrencies, like Bitcoin. As the first and most widely recognized cryptocurrency, Bitcoin is leading the charge as a revolutionary force against the traditional and long-standing financial payment systems that have remained unchanged for many years [10]. The buying and selling of digital currency through cryptocurrency transactions is typically handled by a crypto-exchange platform. These transactions attract the attention of cybercriminals because they typically involve significant amounts of cryptocurrency and are typically anonymized via the blockchain. Like any other system, cryptocurrency exchanges and platforms are vulnerable to cyberattacks. Even though all Bitcoin transactions are forever recorded and publicly visible, existing techniques for detecting ransomware merely rely on a few heuristics and/or time-consuming data collection processes. None of the earlier approaches automatically detected transactions connected to ransomware and illegal Bitcoin addresses by using advanced data analytics techniques. The proposed methodology, which can be used to automate ransomware detection, provides notably higher recall and precision for ransomware transaction identification than existing heuristic-based methods, using the known ransomware data sets. Any quantity of Bitcoin that is sent or received is recorded, together with the recipient's and sender's Bitcoin addresses, in a publicly accessible ledger. This refers to records or ledgers as block chain. Every Bitcoin transaction is encrypted using public key cryptography, and a duplicate of every transaction is stored in every system that handles Bitcoin transactions.

Literature Review

In [2], Tracing cryptocurrencies payments due to malicious activity and criminal transactions is a complicated process. Therefore, the need to identify these transactions and

label them is crucial to categorize them as legitimate digital currency trade and exchange or malicious activity operations. Machine learning techniques are utilized to train the machine to recognize specific transactions and trace them back to malicious transactions or benign ones.

Sabira Karim et al. proposed Bitcoin might be a suburbanized form of payment system wherever the general public ledger is correctly supported in a very distributed manner [3]. The unknown anonymous members referred to as miners, capital punishment a protocol that maintains and extends a distributed public ledger that records Bitcoin transactions is termed a block chain. Block chain is enforced as a series of blocks. Bitcoin is that the known crypto-currency business. The transactions of Bitcoin area unit utterly digital and unknown to a good extent. Ransomware may be a form of malware that infects a victim's information and resources, and demands ransom to unleash them.

Among all those ransomware attacks [4] could be more impacting owing to attack methodology where victim systems become unusable until a ransom is paid, typically have attacker-defined timelines to respond, and can cause more monetary loss. Ransomware attacks, one of the malware attacks affect all types of security issues availability which causes monetary losses, and sensitive information loss. Crypto ransomware, locker ransomware, and hybrid ransomware are standard types of ransomware. In crypto-ransomware attacks, data files are encrypted and the decryption key is provided only after paying the ransom. In locker ransomware attacks, the resources are blocked and are released only after paying the ransom. In hybrid ransomware attacks, both concepts of crypto ransomware and locker ransomware are used.

In [5], the authors proposed Ransomware is a type of malware that infects a victim's data and resources, and demands ransom to release them. In two main types, ransomware can lock access to resources or encrypt their content. In addition to computer systems, ransomware can also infect IoT and mobile devices. Ransomware can be delivered via email attachments or web based vulnerabilities. More recently, ransomware have been delivered via mass exploits. Once resources are locked or encrypted, the ransomware displays a message that asks a certain amount of Bitcoins to be sent to a Bitcoin address. This amount may depend on the number and size of the encrypted resources. After payment, a decryption tool is delivered to the victim.

Ransomware Background

A type of harmful software known as ransomware encrypts data and makes it impossible for users to access it. It is used by cybercriminals to demand money from people or organizations whose data they have compromised, and they keep the data hostage until the ransom is paid. The data may be irreversibly corrupted or leak to the public if the perpetrators do not pay the ransom within the allotted time. Ransomware is one of the most significant problems that companies deal with. Ransomware has just become a standard and changing danger in the last ten years. This is mainly because developing a technically complex crypto virus is difficult [12].

Price Discrimination

It restricts criminals' earnings to levels equivalent to those of a monopolist with consistent prices, despite the fact that this logical conclusion is confirmed by multiple examples of ransomware outbreaks found in the wild. Price discrimination, which involves setting different ransoms for various victims, is undoubtedly a method for them to enhance their profits [13]. In the following sections, we will explore how this can be accomplished. The method of assigning each individual a price that is specifically tailored to their willingness to pay (WTP) is referred to as first-degree price discrimination. To put it another way, a victim who has a WTP of vi is offered a ransom that is somewhat less than vi. While it may seem far off now, this is the greatest discrimination standard that criminals could eventually reach. More realistically, criminals can take advantage of price discrimination in the second or third degree.

Datasets Description

In [11], the author observes Bitcoin along with other cryptocurrencies to exhibit explosive behaviour. Only the ransomware families included in the Bitcoin Heist dataset will be included in the classification. From January 2009 to December 2018, daily transactions are listed in the original dataset. Since ransomware payments are typically much greater, networks with edges that featured amounts less than 0.3 Bitcoins have been eliminated. As a result, 1,048,576 transactions were left. Every transaction, usually for a single transaction, had an address that resembles a physical address or an email address and is required for a Bitcoin payment [6]. Loops keep track of the number of transactions that split the funds and take various routes throughout the network before coming together at a single address. In the end, Bitcoins can be sold and exchanged for other currencies at their ultimate location. When transactions contain more input nodes than output nodes, Weigh seeks to measure this merging behaviour. The Bitcoin network provides the dataset used to train machine learning algorithms on ransomware payments.

Methodology

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset [14]. If the data volume is large enough to be representative of the population, you may not need the validation techniques. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters [16]. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list.

XG Boost Classifier

Its speed and efficiency are unmatched, consistently surpassing all other algorithms designed for supervised learning tasks. The library's ability to parallelize allows the core algorithm to operate on clusters of GPUs or even across a network of computers [9]. This capability makes it possible to tackle machine learning tasks by training on hundreds of millions of examples with exceptional performance.

Random Forest Classifier

Random Forest is a widely used algorithm in machine learning that falls under the category of supervised learning techniques. It is applicable to both classification and regression tasks in ML. The algorithm is grounded in the principle of ensemble learning, which involves combining multiple classifiers to tackle complex problems and enhance model performance. Random Forest operates as a classifier comprising numerous decision trees applied to various subsets of the dataset, and it averages their results to boost predictive accuracy [8]. Rather than depending on a single decision tree, Random Forest aggregates predictions from each tree and determines the final output based on the majority vote. Increasing the number of trees in the forest enhances accuracy and mitigates the risk of overfitting.

Logistic Regression

Logistic regression is a robust supervised machine learning algorithm employed for binary classification tasks. It can be best understood as a form of linear regression tailored for classification issues. Essentially, logistic regression utilizes a logistic function, as defined below, to model a binary outcome variable [7]. The main distinction between linear regression and logistic regression is that the latter's range is confined between 0 and 1. Furthermore, unlike linear regression, logistic regression does not necessitate a linear relationship between the input and output variables.

Voting Classifier

A homogeneous and heterogeneous kind of ensemble learning, the voting classifier allows basis classifiers to be either the same or diverse in type. As was already mentioned, this kind of ensemble functions similarly to bagging (e.g. Random Forest). In figure 1, voting classifier's architecture consists of a number "n" of machine learning models, each of which has two possible prediction values: hard and soft. The guess that receives "the most votes" wins in hard mode.

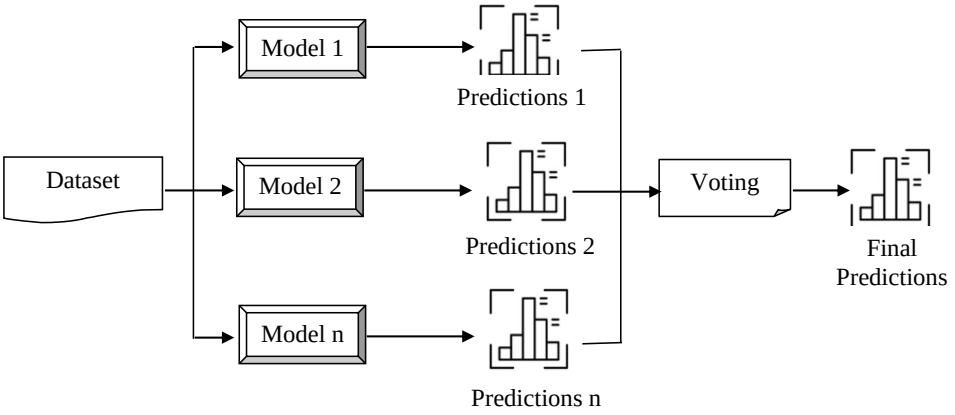

Figure 1: Voting Classifier Architecture

Proposed Method

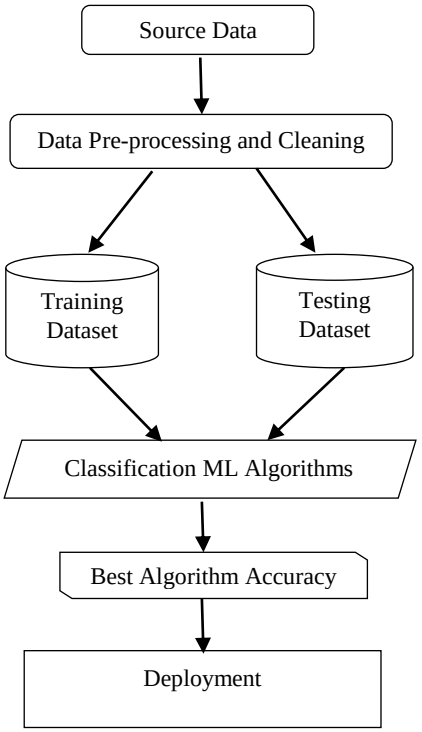

Figure 2: Implementation of Deployment

The proposed strategy aims to create a model that can anticipate the different kinds of ransomware attacks. As seen in Figure 2, finding the target column and identifying dependent and independent variables are the first phases of the process. In order to construct a model, pre-processed data is divided with 70% of the data being used for training and 30% for testing. This allows for the assessment of the method's effectiveness. Pre-processing techniques are then applied to address missing values. The Bitcoin heist ransomware attack

types can be predicted using the classification model. Every model will have unique performance attributes. Voting classifiers and Bitcoin ransomware attacks are implemented. The deployment procedure is complete.

Results and Discussion

Performance metrics to calculate, Accuracy = (TP + TN) / (TP + TN + FP + FN)

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

To calculate the accuracy, test data, training data, length of the test data and train data are the parameters to calculate the accuracy for XG Boost, Logistic Regression, Random Forest Classifier. As the ratio of correctly predicted observations to total observations, accuracy is the most straightforward performance metric to understand. Measuring accuracy requires only symmetric datasets with nearly identical false positive and false negative results.

Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance [15]. It is a good idea to visualize a newly acquired dataset using several approaches so that you can examine the data from many angles. One approach to accomplish this is to display the variance, average accuracy, and other characteristics of the model accuracy distribution using various visualization techniques. Figure 3 shows the accuracy for different algorithms.

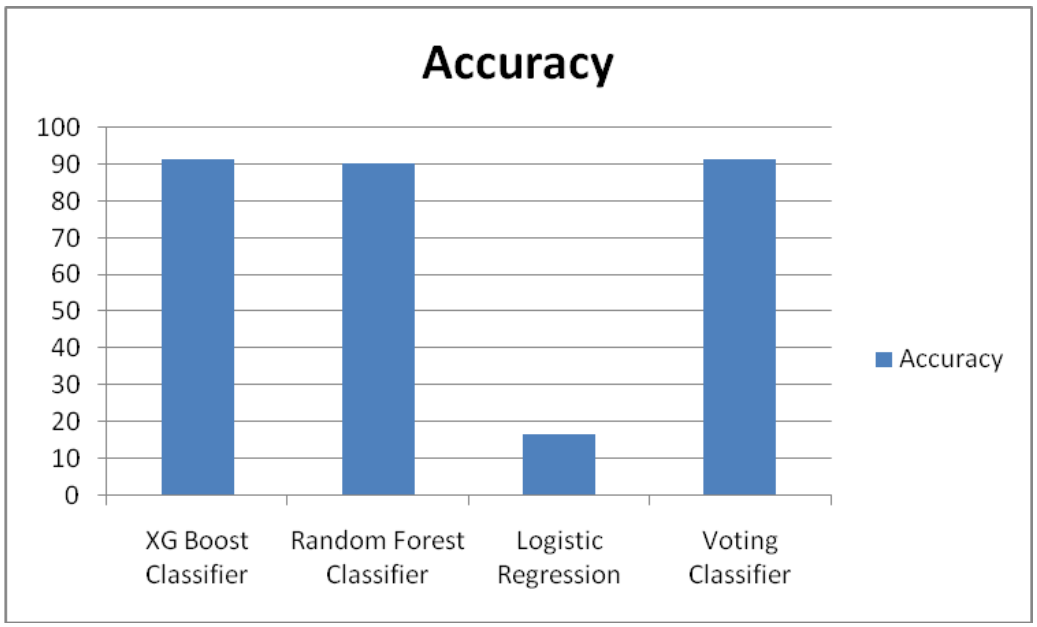

Figure 3: Accuracy for different algorithms

Conclusion

Creating separate dedicated models and finding the accuracy for each of them, we create a single model that trains by these models and predicts output based on their combined majority of voting for each output class. When working on your machine learning challenges, use accuracy as a template and add other, distinct methods for comparison. Models will differ in their performance attributes such as test data and trained data. The initial steps in the analytical process included cleaning and processing the data, searching for missing values, building and assessing models, and conducting exploratory analysis. The most effective accuracy is found on the higher accuracy score algorithm's public test set. Using the created model, a voting classifier can help identify the Bitcoin Heist ransomware attack. The best machine learning algorithm will be deployed.

References

1. J. Hernandez Castro, A. Cartwright, E. Cartwright, "An economic analysis of ransomware and its welfare consequences", 2021.
2. Micheline Al Harrack, "The Bitcoin heist: Classifications of Ransomware Crime Families", 2021.
3. Sabira Karim, Shemitha PA, "A Survey on Detection and Classification of Ransomware Bitcoin Transactions", 2021.
4. Vytarani Mathane, P.V. Lakshmi, "Predictive Analysis of Ransomware Attacks using Context-aware AI in IoT Systems", 2021.
5. Yitao Li, Cuneyt Gurcan Akcora, Yulia R. Gel, Murat Kantarcioglu, "Bitcoin Heist: Topological Data Analysis for Ransomware Detection on the Bitcoin Block chain", 2019.
6. Bitcoin.org, (n.d.) https://Bitcoin.org/en/vocabulary#Bitcoin
7. Meurer, W. J., and Tolles, J. "Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes", 2017.
8. Matthias Schonlau and Rosie Yuyan Zou, "The Random Forest Algorithm for statistical Learning", 2020.
9. Ankush Patil et.al., "XGBoost Algorithm and Its Comparative Analysis", 2022.
10. Peter D DeVries, "An Analysis of Cryptocurrency, Bitcoin, and the Future", 2016.
11. EC Cagli, "Explosive behavior in the prices of Bitcoin and altcoins", 2019.
12. B Khammas, "Ransomeware Detection using Random Forest Technique", 2020.
13. Eugenio J. Miravete, "Price Discrimination Theory", 2005.
14. Kiran Maharana, Surajit Mondal, Bhushankumar Nemade, "A review: Data Pre-processing and data augmentation techniques", 2022.
15. Matthew N.O. Sadiku, Adebowale E. Shadare, Sarhan M. Musa and Cajetan M. Akujuobi, "Data Visualization", 2016.
16. T. Sathya, Keertika N, Shwetha S, Deepti Upodhyay, "Bitcoin Heist Ransomware Attack Prediction Using Data Science Process", 2023.