Multivariate Time series Clustering and LSSVM on Bitcoin Heist ransomware Address dataset to forecast attacks

M SRIVIDYA
MASTERS IN DATA SCIENCE

Research Proposal

MAY 2023

ABSTRACT

Ransomware attacks have been a significant threat in the cybersecurity landscape, causing financial losses and data breaches worldwide. The anonymous mechanism of bitcoins aid to ransomware activities by ensuring no traceability to the cybercriminal. Understanding and predicting the dynamics ransom bitcoin transfers in time series for short term similar time segments can greatly help in developing pre-emptive security measures. This research explores the application of the multivariate time series clustering and Least Squares Support Vector Machines (LSSVM) on the Bitcoin Heist ransomware address dataset, with the bitcoin exchange rate dataset to forecast future attacks. This study starts with analyzing the Bitcoin Heist ransomware address dataset, which contains information on Bitcoin addresses, year that the transaction was performed on, day of the year the transaction took place, length designed to quantify mixing rounds on Bitcoin, loop that counts the split, move and merge of the bitcoins, count containing the information on the transaction, weight calculating the merge, neighbors, income, and label and the exchange rate dataset, that contains timeseries data of the exchange rate if 1 bitcoin to USD. This study offers insights on the ransomware threats and how they can be forecasted based on the clustering similar time segments. Mainly aiming at understanding trends and patterns of the ransomware attacks by the ransom families and also assesses if there are any effects of exchange rate value on the ransomware attacks in a given time segment. It tries to evaluates the forecasting model using evaluation metrices. The research also tries to answer the question if short term time series forecast is possible.

Table of Contents

List of figures

LIST OF ABBREVIATIONS

ABBREVIATION	DEFENITION
BTC-e	Cryptocurrency exchange company
GB	Gigabyte
LSSVM	Least Squares Support Vector Machines
ML	Machine Learning
ODT	Optimizable Decision Trees
PCA	Principal Component Analysis
RAM	Random Access Memory
SNN	Shallow Neural Networks
TB	Terabyte
TDA	Topological Data Analysis
USD	US Dollar
w.r.t	With respect to

1. Background

Ransomware is a class of malicious software “Malware” performed by cyber criminals, by gaining access to personal system and files, holding them hostage and then demanding the victims to pay a ransom to recover the files, most frequently as bitcoins. This cyberattack causes severe significant damages and disruptions to different sectors such as education, health, business, research, local government and information technology. The two main types of ransomware attacks commonly seen are the Crypto ransomware and Locker ransomware. (Khan et al., 2020) The Crypto ransomware attacks victim’s machine specifically targeting the files that typically contain valuable data and encrypts them and demands ransom in exchange for the decryption keys. In the locker method the systems and files are taken hostage demanding ransom in exchange for the return of user’s files. (Khan et al., 2020) Ransomware families – These are the ransomware samples which are recognized to depict specific similar behaviors that machine learning algorithms segregate as a part of the same own family. (Institute of Electrical and Electronics Engineers, n.d.) By using machine learning algorithms, the types of ransom family can be predicted comparing and analyzing the characteristics of bitcoin transactions. (Xu, 2021) This allows us to get the dataset needed for this paper. The Bitcoin Heist ransomware address dataset consists of the attributes of the transaction address and the Ransomware family the address belongs to. The methods of ransomware attacks continue to evolve with machine learning strategies implemented to avoid detection.

Bitcoin one of the best known cryptocurrency which makes transactions that can only be recognized by network nodes through cryptography and that are recorded in a public distributed ledger called a blockchain. Bitcoin is a distributed, peer-to-peer system. There is no "central" server or control point. (Kolesnikova et al., 2021) The records stored in the block chain are bitcoin transactions that are performed by anonymous candidates called the miners who execute protocols to maintain and extend the public distributed ledger. Bitcoin is a digital banking system without any connection to the physical central banking system with any particular country’s origin. (Jadhav, 2020) The use of cryptocurrencies allows pseudo-anonymous transactions to take place that is comparatively easier for ransomware developers to demand ransom. Bitcoin transactions are made incognito and involvement in the network where this is no identity proof required. The payments are requested using a public bitcoin address through an anonymity network using Tor. (Dingledine, n.d.) These has been ransom payment transactions linked to 35 ransomware families from 2013 to mid-2017 and the amount of ransom payments had at the least value of 12,768,536 USD (22,967.54 bitcoins).

(Paquet-Clouston et al., 2018) When these financial transactions are tracked, it was found that ransomware criminal had cashed them out through BTC-e, a now-defunct Bitcoin exchange. (Huang et al., 2018) Studies have been undertaken to looking at ransomware activities from an industrial point of view. (Wang et al., 2022) But the paying of ransom does not necessarily guarantee the return of the files. This research proposal aims to apply multivariate time series clustering and Least Squares Support Vector Machines (LSSVM) on the Bitcoin Heist ransomware address dataset and bitcoin exchange rateto USD dataset to forecast future attacks.

2. Related Research

With growing threat of ransomware demands paid out in cryptocurrencies, research works to understand the patterns and characteristics of ransomware attacks to develop proactive strategies for prevention and mitigation is the primary goal. Several machine learning algorithms and techniques have been incorporated to tackle ransomware based on bitcoin transactions. The topological data analysis (TDA) methodology proposed a well-organized and tractable data analytics framework that can detect new malicious addresses automatically in a ransomware family and furthermore proposed techniques that can display high efficacy to detect the appearance of new ransomware families instantly when there had not been any ransomware records of transactions recorded for the new type previously. (Akcora et al., 2019)

In 2021, the common patterns associated with fraudulent activities are identified in graphs of Bitcoin transactions and the method is applied to find other ransomware actors which is done by local clustering and supervised graph machine learning. (Dalal et al., n.d.) The techniques random forest classifier was applied to make likely classifications into recognizedransomware families by mark out the transactions into their exact ransomware labels, to eventually cover this classification method to identifying the unidentified types of ransomwares. (Al Harrack, 2021) High performance classification models that employ two supervised ML methods shallow neural networks (SNN) and optimizable decision trees (ODT) that help learning the distinguishing patterns in Bitcoin payment transactions, were used to recognize ransomware payments for diverse bitcoin networks. (Al-Haija and Alsulami, 2021)

Our proposal mainly aims to understand the trends and patterns of the dataset and understand the possibility of short-term forecasting of possible ransomware threats based on different variants of similar time periods.

3. Research Questions

- Is there any particular threat forecast possible based on existing data on timeseries front?
- Do variants such as the exchange rate of bitcoin w.r.t USD on that particular time period impact/increase the possibility of threat of ransomware?
- Are there any trends and patterns with respect to time on the ransomware attack?
- Are short-term time series threats in forecasting possible for ransomware?

4. Aim and Objectives

The main aim of this research is to propose the understanding of application of multivariate timeseries clustering on subsets of data on similar time periods on bitcoin ransomware dataset and forecast for future threats. We aim to perform analysis based on timeseries on how often a ransomware attack had happened in the past and the various factors it was dependent on. We also perform the clustering based on the variants for similar time periods and predict if short term forecasting based on time for such threat is possible. We identify the malicious transactions in the dataset, apply PCA with respect to the various columns of the dataset and apply multivariate timeseries clustering technique to create time segment subsets and apply LSSVM to forecast for the future.

The research objectives are framed based on the aim of this study which are as follows:

- To analyze the Bitcoin Heist ransomware address dataset and bitcoin exchange rate dataset to identify temporal patterns and trends.
- Apply multivariate time series clustering techniques to group similar ransomware attacks based on various features.
- To develop forecasting model for predicting future ransomware attacks based on classification of similar timeseries subsets
- To evaluate the performance of forecast on the percentage of likeliness of ransomware attack happening based on past data

5. Significance of the Study

The significance of the study is to understand the possible trends and patterns of reported malicious attacks which holds significant value in the field of cybersecurity and ransomware prevention. We try to investigate any time series trends and patterns that maybe present in the independent variants i.e., columns of the dataset including the variant of the exchange rate of bitcoin to USD. This study mainly focuses on the significance of the days the reported malicious activities are planned to be performed on. By applying multivariate time series clustering and

LSSVM, the goal is to elevate the understanding of ransomware attack patterns and provide a proactive approach to predict and prevent future attacks.

6. Scope of the Study

The scope of the study is to find trends and patterns with respect to time periods in the malicious activities performed by cyber criminals. To understand if forecasting can be performed to detect future ransomware threats based on similar short-term time segments. The outcome of this research can assist individuals, organizations, and cybersecurity professionals in effectively combating ransomware threats.

7. Research Methodology

Methodology deployed involves key processes such as data collection, data preprocessing, multivariate time series clustering, forecasting model deployment, evaluation and fact finding.

7.1. Data collection:

In this step, we proceed to import Bitcoin Heist ransomware address dataset which includes information about ransomware attacks, such as the timestamp, transaction details, ransom amount, etc. and Bitcoin USD exchange rate dataset which will be merged together into one single dataset, which will provide data on the exchange rate of bitcoin with USD on everyday basis.

7.2. Data preprocessing:

We perform cleanse and pre-process the dataset by handling missing values, outliers, performing label encoding, performing feature engineering based on domain knowledge and normalizing the data if necessary. Perform feature selection to identify relevant variables for analysis.

7.3. Data analysis:

This step involves analyzing the dataset to understand the temporal patterns, trends, and characteristics of ransomware attacks. This includes exploring the distributions of variables, defining time intervals, identifying outliers, creating a time series by grouping the data based on the chosen time interval, aggregating the variable values and plotting the time series data to visualize the patterns and trends over time.

7.4. Model training and evaluation:

7.4.1. Multivariate Time Series Clustering:

Applying PCA to perform dimensionality reduction and applying the appropriate multivariate time series clustering algorithm to group similar

ransomware attacks. Evaluating the features such as attack frequency, ransom amounts, and attack vectors such as length, weight, count, neighbor, looped, and exchange rate to form meaningful clusters.

7.4.2. Forecasting Model Development:

Implementing LSSVM as a forecasting model to predict possible future ransomware attacks based on the clustered data. Performing tuning of the model parameters using appropriate techniques and validating its performance.

7.5. Evaluation:

Evaluate the forecasting model's performance using suitable evaluation metrics, such as accuracy, precision, recall, and F1-score.

7.6. Expected Outcomes:

The anticipated outcomes of this research project are as follows:

- Identification of temporal patterns and trends in the short- term time series data segments.
- Grouping of similar ransomware attacks in clusters for similar time period segments on various features.
- Development of a forecasting model using LSSVM to predict future ransomware attacks.
- Evaluation of the proposed approach's performance and understanding if there is possibility of short-term time series forecast

8. Requirements Resources

Hardware requirement: Windows 10 - intel core i5, 8GB RAM, 1TB disk Software requirement: Anaconda Navigator Python version 3.10.11 or later / Google Colab Python 3

9. Research Plan

The estimated timeline for the research project is as follows:

- Literature review: 4 months
- Topic selection: 2 months
- Research proposal preparation: 2 months
- Research code development: 2 months
- Report/ thesis writing: 2 months
- Code evaluation: 1 month

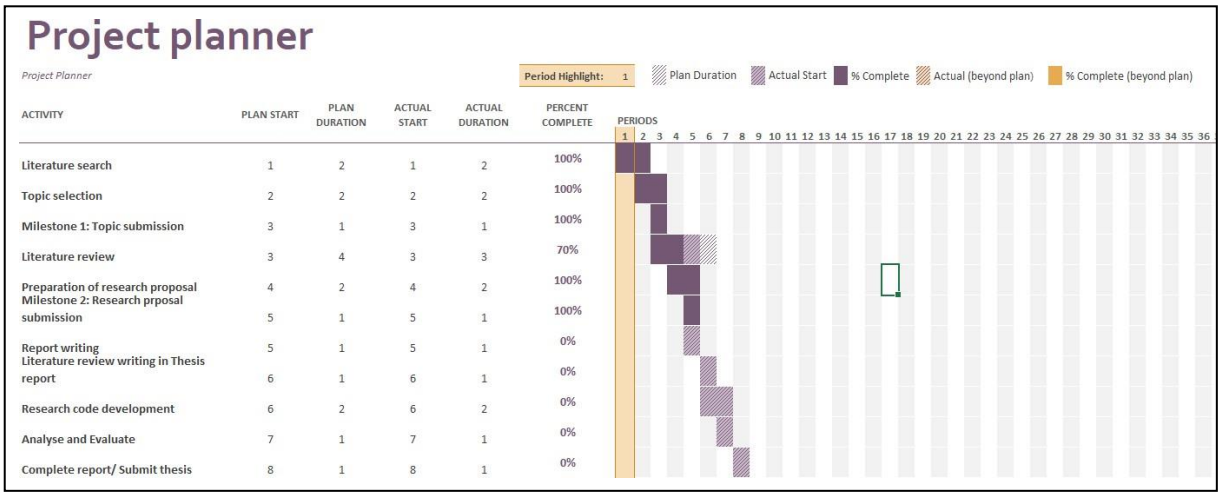

Fig 1.1: Gantt Chart of the Research project planner

References

Akcora, C.G., Li, Y., Gel, Y.R. and Kantarcioglu, M., (2019) BitcoinHeist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain. [online] Available at: http://arxiv.org/abs/1906.07852.

Al-Haija, Q.A. and Alsulami, A.A., (2021) High performance classification model to identify ransomware payments for heterogeneous bitcoin networks. *Electronics (Switzerland)*, 1017.

Dalal, S., Wang, Z. and Sabharwal, S., (n.d.) *IDENTIFYING RANSOMWARE ACTORS IN THE BITCOIN NETWORK*.

Dingledine, R., (n.d.) *Tor: The Second-Generation Onion Router*.

Al Harrack, M., (2021) The BitcoinHeist: Classifications of Ransomware Crime Families. *International Journal of Computer Science and Information Technology*, 135, pp.75–81.

Huang, D.Y., Aliapoulios, M.M., Li, V.G., Invernizzi, L., Bursztein, E., McRoberts, K., Levin, J., Levchenko, K., Snoeren, A.C. and McCoy, D., (2018) Tracking Ransomware End-to-end. In: *Proceedings - IEEE Symposium on Security and Privacy*. Institute of Electrical and Electronics Engineers Inc., pp.618–631.

Institute of Electrical and Electronics Engineers, (n.d.) *Proceedings of the 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies ICICICT-2019 : 5th & 6th July 2019*.

Jadhav, K., (2020) *Investigating Machine Learning Approaches for Bitcoin Ransomware Payment Detection Systems*. [online] *International Journal of Innovative Science and Research Technology*, Available at: www.ijisrt.com.

Khan, F., Ncube, C., Ramasamy, L.K., Kadry, S. and Nam, Y., (2020) A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning. *IEEE Access*, 8, pp.119710–119719.

Kolesnikova, K., Mezentseva, O. and Mukatayev, T., (2021) Analysis of Bitcoin Transactions to Detect Illegal Transactions Using Convolutional Neural Networks. In: *SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies*. Institute of Electrical and Electronics Engineers Inc.

Paquet-Clouston, M., Haslhofer, B. and Dupont, B., (2018) Ransomware Payments in the Bitcoin Ecosystem. [online] Available at: http://arxiv.org/abs/1804.04080.

Wang, K., Pang, J., Chen, D., Zhao, Y., Huang, D., Chen, C. and Han, W., (2022) A Large-scale Empirical Analysis of Ransomware Activities in Bitcoin. In: *ACM Transactions on the Web*. Association for Computing Machinery.

Xu, S., (2021) The Application of Machine Learning in Bitcoin Ransomware Family Prediction. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp.21–27.