IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 3, June 2019
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Preventing Crypto Ransomware Using Machine Learning

[1] Jitti Annie Abraham; [2] Susan M George

[1] CSE Department, MBCCET
APJ Abdul Kalam Technological University, Kuttikkanam, Peermade, Idukki, Kerala, India.

[2] CSE Department, MBCCET
APJ Abdul Kalam Technological University, Kuttikkanam, Peermade, Idukki, Kerala, India.

Abstract- Ransomware is a kind of malware that forestalls or confines clients from getting to their framework, either by locking the framework's screen or by locking the clients' records except if a payoff is paid. Due to the changing conduct of ransomware, conventional type and detection techniques do not correctly stumble on new variants of ransomware. Our data set includes some of the most up-to-date ransomware samples available, providing an assessment of the category accuracy of device studying algorithms on the present day evolving repute of ransomware. Two primary parts of this work are identification of the behavioral attributes which can be used for choicest class accuracy and type of ransomware the using machine learning classification algorithms. After classifying the ransomware editions, a prevention mechanism is also completed to the cryptographic ransomware variants.

Keywords- *Classification, Machine Learning Ransomware, Ransomware prevention*

1. Introduction

Ransomware attacks are becoming a serious cyber threat to organizations and individuals around the world. Ransomware is a sort of malicious software program from cryptovirology that threatens to put up the sufferer's records or continuously block get admission to it unless a ransom is paid. While a few easy ransomwares may additionally lock the device in a manner, which is not always tough for an informed individual to opposite, extra advanced malware makes use of a technique referred to as cryptoviral extortion, in which it encrypts the sufferer's files, making them inaccessible, and needs a ransom payment to decrypt them.

Regularly, recent types of malware are not recognized from their ancestors because of the impediments of order frameworks depending just on static investigation. Accordingly, methods like static based, signature-based and design coordinating methods for malware investigation are ending up less viable to identify and order new variations of ransomware and give knowledge data about the risk, objectives and practices of ransomware. Efforts have been made to increase behavior-based totally type techniques. Classification of malware samples primarily based on their behavior calls for implementation of algorithms that are successful to supply models and research via the type procedure. The potential of system getting to know to analyze with facts at some stage in the system of type, makes them appealing and powerful for malware classification [1].

Classification of malware samples based on their behavior requires implementation of algorithms that are capable to produce models and learn through the classification process. The ability of machine learning to learn with data during the process of classification, makes them attractive and effective for malware classification. Using machine learning classification algorithms, ransomware samples can be identified with different behaviors from other samples that are part of the same family. The reason behind this study is to identify new modified variants of ransomware based on their behavior using machine learning algorithms. Two main parts of study are identification of the behavioral attributes which can be used for optimal classification accuracy and classification of ransomware. After classifying the ransomware variants, a prevention mechanism is also done to the cryptographic ransomware families.

The rest of this paper is organized as follows. A brief literature review of existing works on classification and prevention mechanism of ransomware variants is given in Section 2. The proposed method is explained in Section 3.

The experimental results are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

M. I. Jordan, T. M. Mitchell, described that Artificial Intelligence is anywhere. Possibility is that the use of it in a single way or the opposite and also you don't even know approximately it. One of the famous programs of AI is Machine Learning, wherein computer systems, software, and devices perform through cognition which could be very just like human mind. Machine mastering a subject of synthetic intelligence that makes use of statistical strategies to give computer systems the capability to "research" from data, without being explicitly programmed [2]. Some of the trending applications of device mastering includes: virtual personal assistants, predictions at the same time as commuting, social media offerings, electronic mail spam and malware filtering, seek engine end result refining and product tips. Machine mastering might be going to be a standout among the most transformative innovations of the 21st century.

Sandhya Ndhage, Charanjeet Kaur Raina [3], Machine learning is a multi displinary field in artificial intelligence, likelihood insights data hypothesis, reasoning, human science, and neurobiology. Machine learning tackles this present reality issues by building a model that is great and valuable estimation to the information. The study on machine learning taking in has developed from the endeavours of investigating regardless of whether computer could figure out how to imitate the human mind, furthermore, a field of measurements to a wide control that has created central factual computational speculations of learning forms. The fundamental objective and commitment of this audit paper is to display the diagram of machine learning and gives machine-learning procedures. Additionally, paper surveys the benefits and limitations of different machine learning algorithm in diverse methodologies.

Ziya Alper Gen, Gabriele Lenzini, Peter Y.A. Ryan, discussed that ransomware is a category of malware whose aim is to extort money. At the point while brought on a framework, a ransomware encodes files or squares functionalities and whilst the interest is achieved it requests a get better. In this paper the survey present day barrier techniques for ransomware, talking about their stable and powerless focuses [4]. Here describe current techniques to mitigate ransomware and speak their boundaries. The current ransomware mitigation systems are built upon the evaluation of gathered samples that is they with the exception of the inefficient and ineffective exercise to returned-up and repair files. The future threats of ransomware include rootkit-primarily based ransomware, obfuscation, white-field cryptography, socio-technical assaults.

D.Nieuwenhuizen [5] performs a prediction that ransomware is a type of malicious software program (malware) that once finished on a pc machine, hinders the user from using the laptop or its facts, annoying an amount of cash (ransom) for the recuperation of the computer. Currently, ransomware attacks preclude laptop operation in three approaches: through blockading gaining access to the computer, this form of ransomware is referred to as locker ransomware; through making person facts unusable with the aid of employing encryption algorithms, known as crypto ransomware and an aggregate of locker/crypto ransomware where a person is blocked from the usage of their pc even as their records is being encrypted. This paper gives inspiration to the utilization of machine-learned conduct for ransomware identification. Ransomware assaults impede PC task in three different ways: by blocking getting to the PC (storage ransomware), by making client information unusable by methods for utilizing encryption calculations (crypto ransomware) and mix of storage/crypto ransomware. The procedures portrayed in this paper are utilized in RansomFlare which is a ransomware counteractive action operator that uses dynamic (social) examinations and AI strategies. Here demonstrates that signature based recognition methods have demonstrated an insufficient resistance. Additionally, the static-based recognition is compelling against known ransomware.

R.Vijaya Kumar Reddy, Dr. U. Ravi Babu, A classification is a technique of predicting comparable facts from the fee of an express target or express elegance variable [6]. It is a useful method for any sort of statistical data. These algorithms are used for diverse functions like photo category, Predictive modelling, facts mining method and so on. The primary reason of supervised learning is to construct an easy and unambiguous version of the allocation of sophistication labels in terms of predictor capabilities. The classifiers are then used to categories elegance labels of the checking out times where the values of the predictor features are known, to the price of the magnificence label that is unknown. In this paper here illustrate numerous class techniques used in supervised gadget getting to know.

Smruti Saxena, Hemant Kumar Soni [7] Ransomware is now grow to be a horrific tool to earn cash, theft records, hack the gadget or to stop the normal functioning of the gadget. Ransomware is a malware that breaches the security of the machine by means of the use of malicious codes. It encrypts the data and available data earlier than noticing it.

Traditional vaccination gadget does not remedy the infected device without acquiring data on ransomware. Since the statistics is encrypted subsequently cannot be recovered without encryption key. Users can keep away from the infections of ransomware via updating vaccination device every so often. However, this approach has confined efficacy. This method cannot trace changed ransomware with new pattern. This paper explores the various ransomware attack. In this paper here converse the analysis of ransomware and the advised movement in opposition to ransomware assault. This paper also discusses ransomware removal and prevention methodology.

Daniel Gonzalez, Thaier Hayajneh, [8] described that crypto-ransomware is a difficult danger that ciphers a user's files at the same time as hiding the decryption key till a ransom is paid by way of the sufferer. This form of malware is a lucrative enterprise for cybercriminals, producing tens of millions of bucks yearly. The spread of ransomware is growing as traditional detection-based totally safety, along with antivirus and anti-malware, has verified useless at stopping attacks. Additionally, this form of malware is incorporating advanced encryption algorithms and expanding the range of report sorts it goals. This paper discusses ransomware strategies of contamination, technology in the back of it and what may be accomplished to assist save you turning into the subsequent victim. The paper investigates the maximum commonplace sorts of crypto-ransomware, numerous payload methods of infection, regular behavior of crypto ransomware, its techniques, how an attack is primarily completed, what files are maximum typically targeted on a victim's computer, and suggestions for prevention and safeguards are listed as well.

3. Proposed System

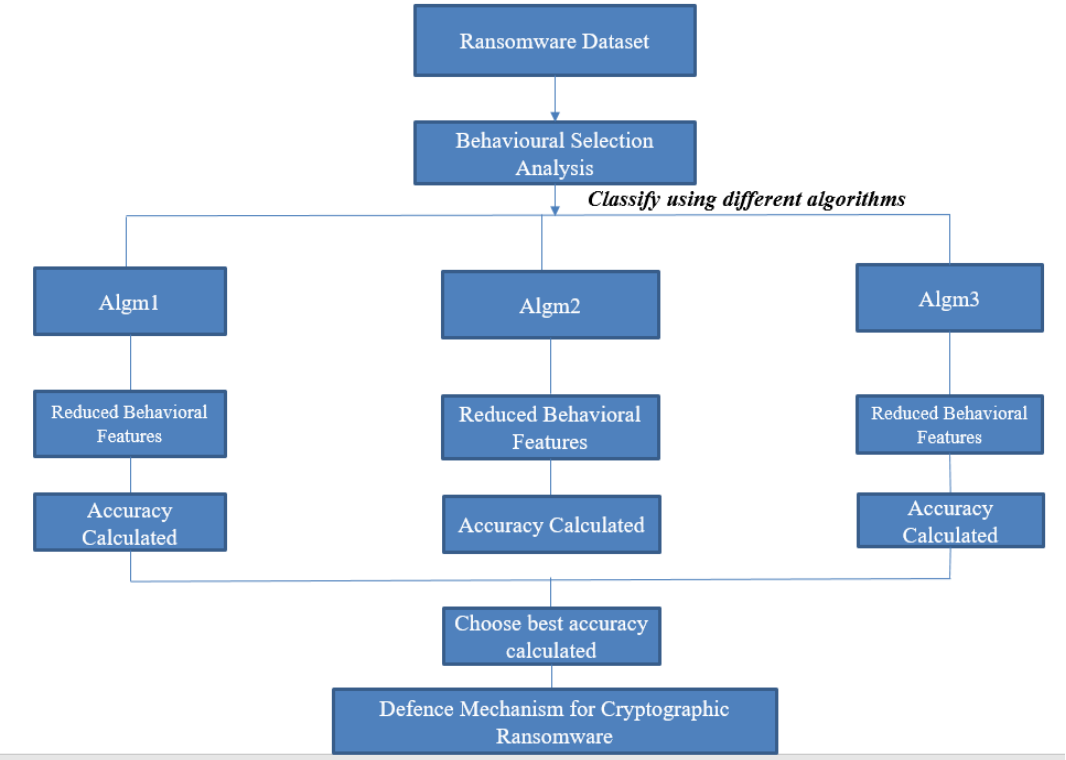

Fig 1 Proposed Architecture

The proposed architecture is shown in Fig 1. The study consists of three main phases: data collection, extraction of behavioral attributes and selection of behavioral attributes for optimal classification accuracy. In the data collection phase, we collect behavioral reports from VirusTotal for every ransomware sample. In the next step, behavioral attributes are extracted from the behavioral reports. For ultimate classification accuracy, we perform behavioral attributes selection analysis to identify behavioral attributes which should be used for classification in the next phase. Using the selected behavioral attributes, we evaluate classification accuracy of machine learning algorithms.

The main goal of the behavioral attributes extraction phase is to obtain a dataset which best represents the behavior of a ransomware sample without missing any relevant information. Therefore, spent a considerable time and effort extracting the behavioral attributes from all the behavioral reports. Identified behavioral attributes appear at least in one of the behavioral reports. For each of the behavioral attributes, based on the type of information contained in the behavioral reports, we determine the attribute type to be used to assign a value to the attribute.

After calculating the accuracy of classification algorithms, we conclude by best accuracy algorithm. Then describes a prevention mechanism for cryptographic ransomware families using machine learning techniques. Here BitLocker Drive Encryption method is used as a prevention mechanism.

3.1 DARPA Dataset

The dataset utilized for leading the test is "DARPA". DARPA IDS assessment dataset is valuable for testing interruption discovery frameworks in that great execution against it is a fundamental yet not adequate condition to showing the capacities of a propelled IDS. This dataset was built for system security examination purposes. Analysts scrutinized DARPA because of issues related with the counterfeit infusion of assaults and benevolent traffic. DARPA incorporates exercises, for example, send and get mail, peruse sites, send and get documents utilizing FTP, the utilization of telnet to sign into remote PCs and perform work, send and get IRC messages, and screen the switch remotely utilizing SNMP. It contains assaults like DOS, surmise secret key, cradle flood, remote FTP, syn flood, Nmap, and rootkit. Sadly, it doesn't speak to genuine system traffic and contains abnormalities, for example, the nonattendance of false positives, and is obsolete for the successful assessment of IDSs on current systems as far as assault types and system foundation. In addition, it does not have the real assault information records.

3.2 Classification Algorithms

Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category or class to which a new data will fall under. Order is strategy to sort information into an ideal and unmistakable number of classes where we can relegate mark to each class. Utilizations of classification includes discourse acknowledgment, penmanship acknowledgment, biometric distinguishing proof, record arrangement and so forth. The classification algorithms used here for identification and classification of ransomwares based on their behaviors are:

i. Linear Regression
ii. Adaboost
iii. Random Forest
iv. Extra Trees
v. Gradient Boost
vi. Multilayer Perceptron

3.3 Modules

In programming, a module is a piece of a program. Projects are made out of at least one freely created modules that are not consolidated until the program is connected. A solitary module can contain one or a few schedules. The work can be combined into following modules:

a) Data Collection
b) Classification Processing
c) Prevention

The data collection method comprises feature extraction and fitness package. Feature extraction refers to the extraction of linguistic items from the documents to provide a representative sample of their content. Feature extraction begins from an underlying arrangement of estimated information and constructs determined qualities (highlights) expected to be instructive and non-excess, encouraging the resulting learning and speculation steps, and now and again prompting better human translations. Highlight extraction is identified with dimensionality decrease. At the point when the information to a calculation is too huge to possibly be handled and it is suspected to be excess (for example a similar estimation in the two feet and meters, or the monotony of pictures introduced as pixels), at that point it very well may be changed into a diminished arrangement of highlights (additionally named a component vector). Deciding a subset of the underlying highlights is called include determination. The chose highlights are relied upon to contain the pertinent data from the information, with the goal that the ideal errand can be performed by utilizing this decreased portrayal rather than the total starting information.

The classification processing module does the classification of ransomware variants using various machine learning classification. Here also calculated the accuracy of each algorithm in each model. Totally three models evaluated with different classification algorithm.

In the prevention module, a prevention mechanism for crypto ransomware family is implemented. The encryption technique “BitLocker Driven Encryption” method is used as prevention technique.

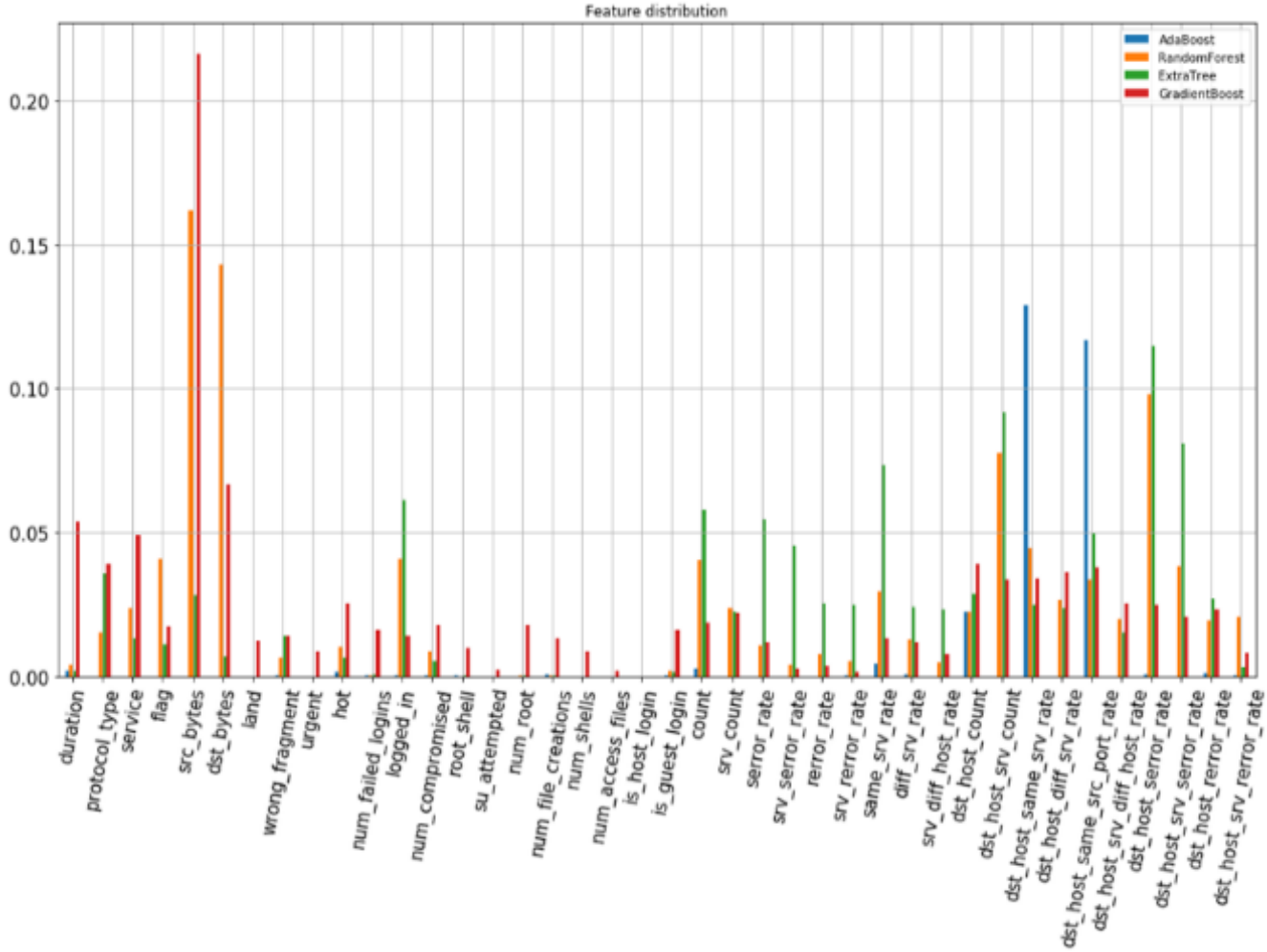

Fig 2 Feature selection of ransomware variants

3.4 BitLocker Driven Encryption

BitLocker is a full volume encryption highlight included with Microsoft Windows renditions beginning with Windows Vista. It is intended to secure information by giving encryption to whole volumes. Of course, it utilizes the AES encryption calculation in figure square anchoring (CBC) or XTS mode with a 128-piece or 256-piece key. CBC isn't utilized over the entire plate; it is connected to every individual division. BitLocker is a PC hard drive encryption and security program discharged by Microsoft Corporation as a local application in its Windows 7 Enterprise and Ultimate releases, Windows Vista Enterprise and Ultimate, and Windows Server 2008, R2 and 2012 working framework variants. It is a drive security and encryption program that shields drive substance and information from any disconnected assault.

BitLocker is basically intended to keep a client's information from being seen, extricated or recovered on the off chance that a drive is stolen. It doesn't secure a framework when it's running in light of the fact that the on the web/operational/live assurance is kept up by the working framework. BitLocker utilizes an AES encryption calculation with a 128-piece key or 256-piece key to scramble plate volumes. It secures the information when a hard drive is stolen and is being utilized on another PC or when somebody has physical access to the drive. To get to the drive in a disconnected mode, BitLocker requires a recuperation key. BitLocker is by and large pointed toward individual clients who may fall prey to PC/PC robbery.

4. Experimental Result and Discussion

This study is carried out to identify and classify ransomware variants using machine learning classification algorithm based on their behavior. For the implementation of the proposed system, the model is created in Python

programming language. For getting more accurate classification algorithm, the different algorithm uses different features in each three models. The following graph shows the accuracy level in each model.

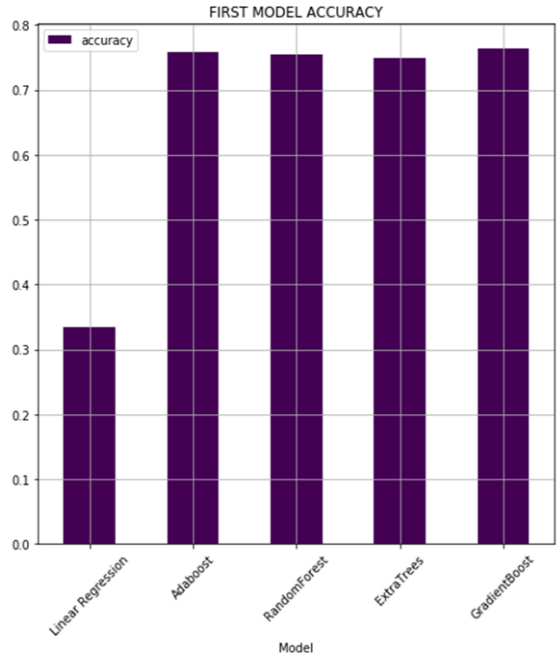

Fig 3 Accuracy of Model 1

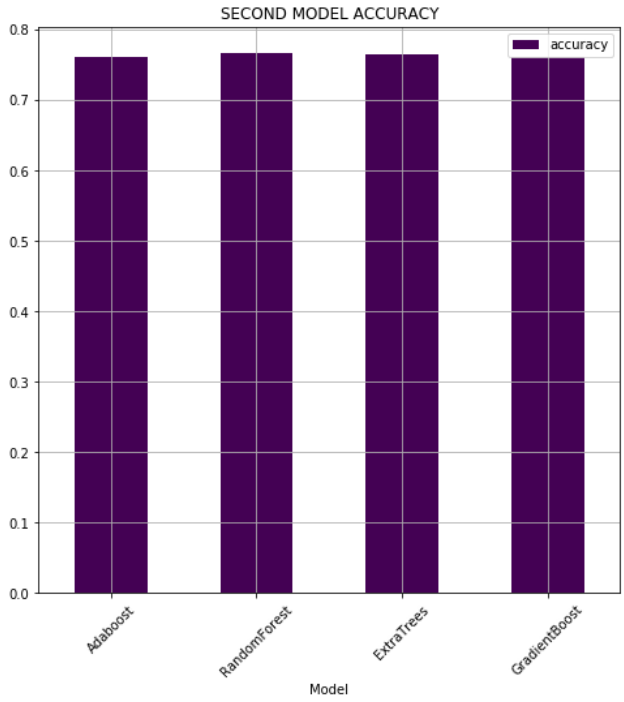

Fig 4 Accuracy of Model 2

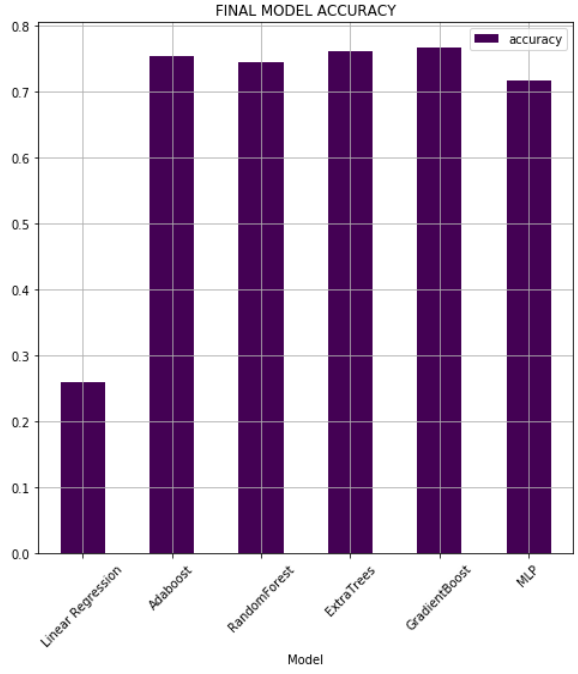

Fig 5 Accuracy of Final Model

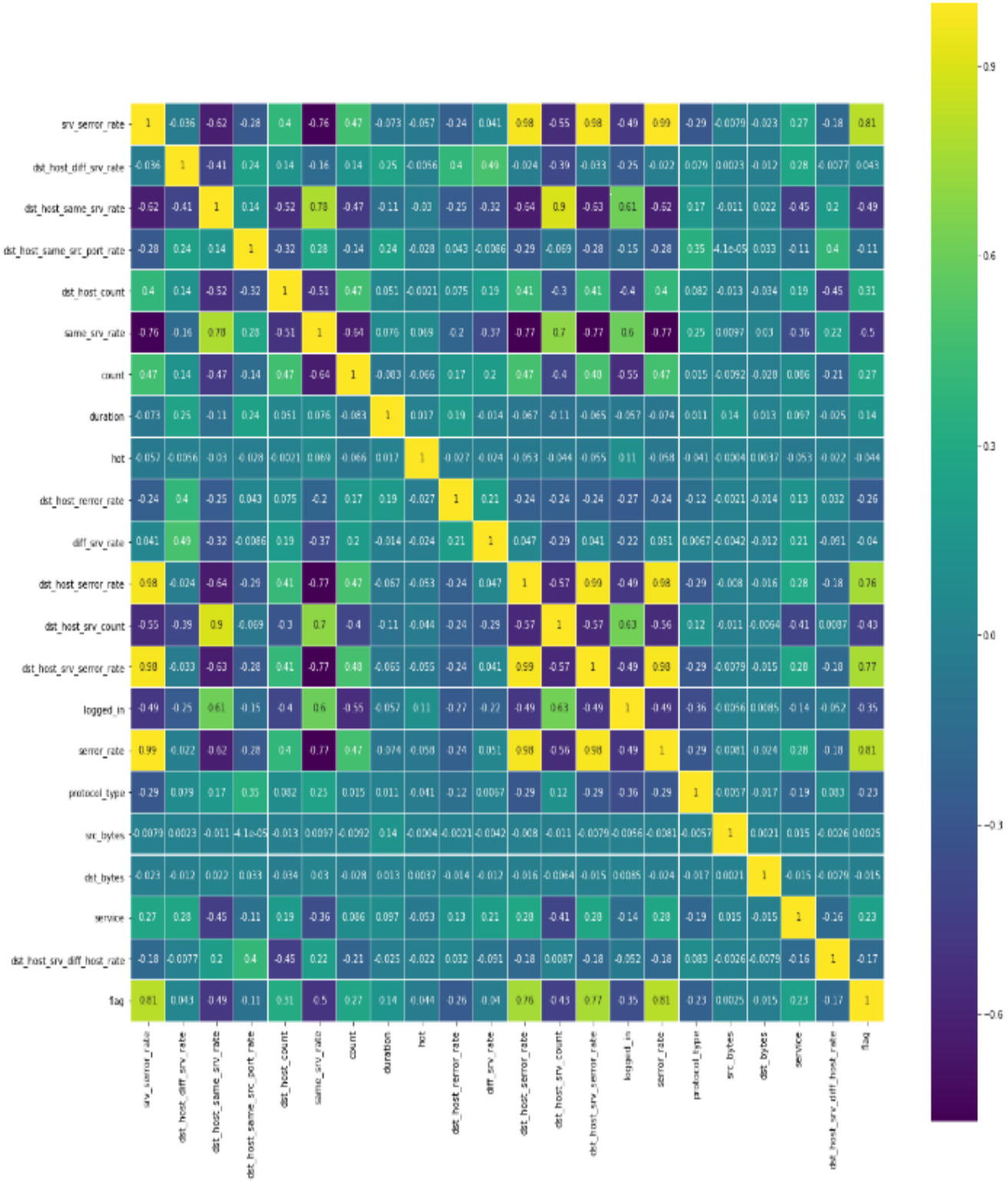

Fig 6 Confusion Matrix

The values of accuracy of final model can be tabulated as follows:

Algorithm	Accuracy
Linear Regression	0.259637
Adaboost	0.753700
Random Forest	0.744700
Extra Trees	0.761200
Gradient Boost	0.766800
Multi-Layer Perceptron	0.716500

Table 1 Accuracy Values of Classification Algorithms

From the above experimental values and figures the result can be summarized as, in case of three models Gradient Boost classification algorithm has greatest accuracy. Thus on further use we can directly choose this algorithm. Also from here, got analyses that the highest attack occurs in the experimental dataset is "probe" attack. It occurs around 11760 times. Probe-response attacks are a new threat for collaborative intrusion detection systems. A probe is an attack which is deliberately crafted so that its target detects and reports it with a recognizable fingerprint in the report. The attacker then uses the collaborative infrastructure to learn the detector's location and defensive capabilities from this report [9].

5. Conclusion

Ransomware variations are expanding step by step. They generally target client savvy and framework shrewd. The principle point of ransomware is to take cash from the person in question. Here studied the implementation of machine learning algorithms for malware classification based on the behavior of malware samples. Using an iterative approach, determined the set of behavioral attributes which can be used for ransomware classification to achieve the optimal classification accuracy. Moreover, here evaluated classification accuracy of five machine learning algorithms. Using machine learning, identified modified variants of ransomware samples, confirming the new trend of malware in evading classification and detection systems by modifying their behavior. The identified ransomware samples from evolving families with a diverse behavior compared to their predecessors. The intention of creating malware variants with various behaviors might be to evade detection systems by presenting a rare behavior on new samples, or to mislead detection and classification systems by using a similar behavior to other ransomware families. Then describes a prevention mechanism named BitLocker Driven Encryption method for crypto ransomware families using machine learning techniques.

Since machine learning is an upcoming trend, in future there may get more accurate algorithm for classifying. Malware detection is an arms race, as defenders provide mitigations, adversaries will modify their techniques. Also can be developed as a web based application, in future because now the study and work is windows based only.

References

[1] Hajredin Daku, Pavol Zavarsky, Yasir Malik, "Behavioral-Based Classification and Identification of Ransomware Variants Using Machine Learning", 2324-9013/18/31.00© IEEE, 2018
[2] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects", Science 349,255 2015.
[3] Sandhya Ndhage, Charanjeet Kaur Raina, "A Review On Machine Learning Techniques", IJRITCC, ISSN: 2321-8169 Volume: 4 Issue: 3 395 – 399, 2016.
[4] Ziya Alper Gen, Gabriele Lenzini, Peter Y.A. Ryan,"The Cipher, the Random and the Ransom: A Survey on Current and Future Ransomware", CECC, November 2017.
[5] D Nieuwenhuizen, "A Behavioural-based Approach to Ransomware Detection" Information Security 2017.
[6] R. Vijaya Kumar Reddy, Dr. U. Ravi Babu, "A Review on Classification Techniques in Machine Learning",ICRTESM March 2018.
[7] Smruti Saxena, Hemant Kumar Soni, "Strategies for Ransomware Removal and Prevention", 978-1-5386-4606-9© IEEE, 2018.
[8] Daniel Gonzalez, Thaier Hayajneh, "Detection and PreventionofCrypto-Ransomware", 978-1-5386-1104-3/17/$31.00 © IEEE, 2017.
[9] Vitaly Shmatikov and Ming-Hsiu Wang, "Security Against Probe-Response Attacks in Collaborative Intrusion Detection", ACM 2007

Author Profile

Jitti Annie Abraham received her B.Tech (CSE) degree from University of Kerala in 2016. She is currently pursuing her Masters in Computer Science & Engineering from APJ Abdul Kalam Technological University. Her research interests areas includes machine learning, artificial intelligence, cyber forensics and cryptography.

Susan M George is working as Assistant Professor in Computer Science and Engineering Department. She has more than 3 years' experience in teaching. Her research interests focus data mining, machine learning and artificial intelligence. She has published several papers on different areas.