Visualizing and Predicting the type of Malware in Security Dataset with Machine Learning using Python libraries and Tableau

Rahul Bajaj[1], Srishti Kohli[2]
[1]*School of Computing Science and Engineering, VIT Chennai, India*
[2]*Indira Gandhi Technical University for Women, Delhi, India*

***Abstract*- In this paper, the different varieties of malware have been introduced and explained. Using Tableau we have implemented a few visualisation techniques on a malware dataset to visualise the prime characteristics and attributes of the various kinds of malware. Furthermore, we have used python libraries for classification of malware from the data provided in the malware dataset, using machine learning algorithms (primarily Decision Trees and its variations) in jupyter notebook, to obtain a high level of accuracy.**

Index Terms- security, malware, decision tree, random-forest, logistic regression

I. INTRODUCTION

Malware, or malicious software, is any program or file that is harmful to a computer user. Malware includes computer viruses, worms, Trojan horses and spyware. These malicious programs can perform a variety of functions, including stealing, encrypting or deleting sensitive data, altering or hijacking core computing functions and monitoring users' computer activity without their permission.

Machine learning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed.

II. TYPES OF MALWARE

A. Adware

A whole new class of threats was recorded as early as in 1987 aiming to gather marketing information and also to display advertisements which generate revenue. These threats have become more advanced over the time as they are being developed by professionals with their targets more and more widespread.

● Delivery Medium

Adware programs can be installed in different manners. Most of the times users have no idea from where the programs got installed.

● Social Engineering banner ads

Social Engineering banner ads serves as the best way to lure user to install an adware program. Automatic Refresh

● P2P Installation

A relatively new delivery medium is infecting the peer network with adware programs. The files are renamed to intimidating file names are often bundled with pirated media.

● Exploits

Some adware programs make use of the vulnerabilities in Internet Explorer to install programs without any user consent.

B. SDBot

SDBot malware propagate through exploited unpatched vulnerabilities and network shares. They have a number of backdoor capabilities and also information theft routines. They also exhibit a number of backdoor capabilities and some information theft routines. Most of the bots that use Internet Relay Chat (IRC) protocol communication.

C. Muldrop

Muldrop is a Linux Trojan developed in late 2017 as a means to target Rasberry Pi devices with the objective to mine cyptocurrency. When the SSH port 22 is left open the Trojan can infect the device and change its password. Muldrop installs the libraries

like ZMap and SSHPass to mine cryptocurrency and it shuts down the processes.

D. Ransomware

Most widespread ransomwares make an intensive use of file encryption as an extortion mean. Basically, they encrypt various files on victim's hard drives before asking for a ransom to get the files decrypted.

E. Trojan Horse

It is a novel network attack program at present. It is remote control-based software controlling the other computer based on a specific program. With its implantation function or the characteristic of accessory with carrying virus, this virus can enter into user's computer to steal personal information and password, tamper data, or destroy files.

Trojan horses are classified as below:

- Remote access Trojan
- Data sending Trojan
- Destructive Trojan
- Security software disabler Trojan
- Denial-of-Service attack Trojan

F. Backdoor

Backdoors applications that open computers to remote access—play a crucial role in targeted attacks. Often initially used in the second (point of entry) or third (command-and-control [C&C]) stage of the targeted attack process, backdoors enable threat actors to gain command and control of their target network.

- Backdoor Techniques
- Port Binding
- Connect-Back Technique

G. RBot

Rbot is a Trojan program which allows an attacker remote access to the compromised system. Rbot is a component of "rundat.exe" downloaded by another malicious program Blazebot from FTP server:

W32/Rbot-ZW is a network worm and IRC backdoor Trojan for the Windows platform. W32/Rbot-ZW spreads using a variety of techniques including exploiting weak passwords on computers and SQL servers, exploiting operating system vulnerabilities (including DCOM-RPC, LSASS, WebDAV and UPNP) and using backdoors opened by other worms or Trojans.

The backdoor component of W32/Rbot-ZW can be instructed by a remote user to perform the following functions:

- start an FTP server
- start a Proxy server
- start a web server
- take part in distributed denial of service (DDoS) attacks
- log keypresses
- capture screen/webcam images
- packet sniffing
- port scanning
- download/execute arbitrary files
- start a remote shell (RLOGIN)
- The worm copies itself to a file named scmss.exe in the Windows system folder and creates registry entries

H. Spam

Spam is unsolicited email, normally with an advertising content sent out as a mass mailing. Spammers try to obtain as many valid email addresses as possible, i.e. actually used by users. They use different techniques for this, some of which are highly sophisticated:

- Mail lists
- Purchasing user databases from individuals or companies
- Use of robots (automatic programs) that scour the Internet looking for addresses in web pages, newsgroups, weblogs, etc
- DHA (Directory Harvest Attack) techniques

Techniques used:

- Division of message subject line using bogus line breaks
- Use of null characters (Quoted-Printable type encoding)
- Interchanging letters in the words used. The message is still legible to the recipient, but the filters do not recognize the words used
- Inverting text using the Unicode right-to-left override, expressed as HTML entities
- Encapsulating a <map> tag with an HREF tag, so that a legitimate URL appears instead of a malicious one
- Use of ASCII characters to "design" the message content

I. Normal

● Keylogger

A keylogger (keystroke logging) is a type of surveillance software that once installed on a system, has the capability to record every keystroke made on that system.

● RootKits

A Rootkit is defined as malicious computer software hidden deep inside a PC and remains undetectable. Although Rootkits on their own may not be harmful, they hide worms, bot & malware. Types Of Rootkits:

● Bootkit

● Firmware Rootkits

There are different types of Rootkit virus such as Bootkits, Firmware Rootkits, Kernel-Level Rootkits & Application Rootkits.

III. PROPOSED SYSTEM

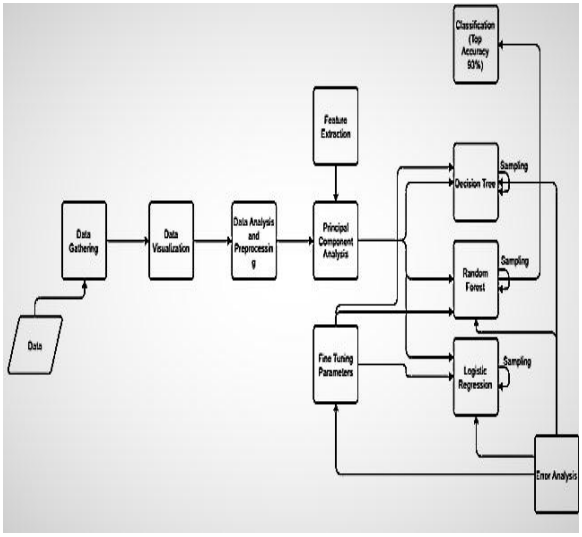

Fig.1. Block Diagram of the Proposed System

The different modules of this project are:

1. Data Gathering
2. Data Visualization
3. Data Analysis and preprocessing
4. Feature Extraction/Principal Component Analysis
5. Algorithms used
   a. Decision Tree
   b. Random Forest
   c. Logistic Regression
7. Error Analysis
8. Fine Tuning Parameters
9. Classification

1. DATA GATHERING

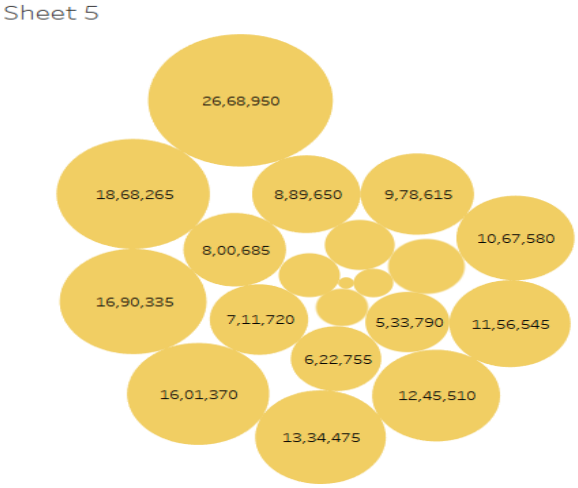

Fig.2. Representation of data using "Packed Bubbles" technique visualizing the Median and bin values of "lat RVA" against the Median "Image Version"

The malware dataset is built using open source software. It is free software and you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or any later version.

Malware is a malicious software, kind of code designed with unintended purpose for compromising the privacy and security of the user. The various kind of malware are virus, trojan, adware, backdoor, muldrop, Sdbot, spam, Rbot and ransomware. The malware datasets are generated using virtual instances running in a virtual environment. The malware samples are collected from various repositories and they are executed in the virtual environment to understand the impact in the virtual machines. The parameters that are collected are debug size, latRVA, export size, image version, resource size, virtual size and number of sections. The malware dataset is in the excel format.

2. DATA VISUALIZATION

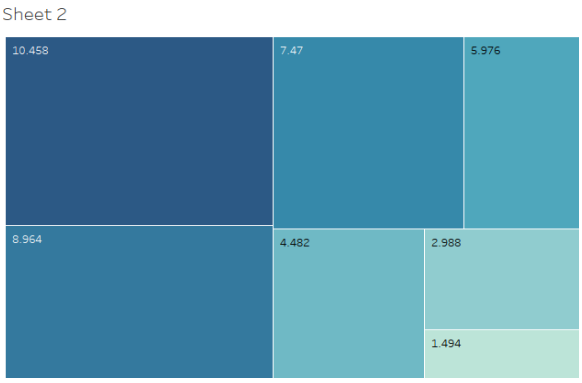

Fig.3. Representation of data in the form of Heatmap

The above diagram is the representation of our dataset in the form of a Heatmap, visualising the relationship between "Number of Sections" and the "Resource Size".

We have used the Average value of the "Resource Size" and the Median, Average and Bin values of "The Number of Sections".

Moreover, the color of the blocks emphasises the resultant values, i.e., higher the value, deeper the color of the block.

Sheet 6

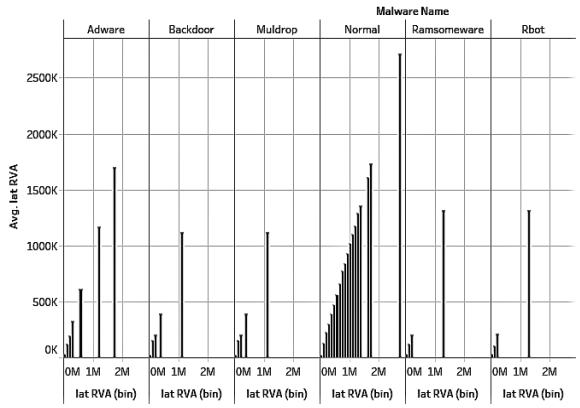

Fig.5. Representation of data in the form of Histogram

This is the representation of "malware name" and "lat RVA" in the form of histogram.

To make the visualization more detailed we have plotted the average and bin values of "lat RVA" against each other for every category of malware.

For every variation in "lat RVA" values, the image version is also displayed.

Sheet 8

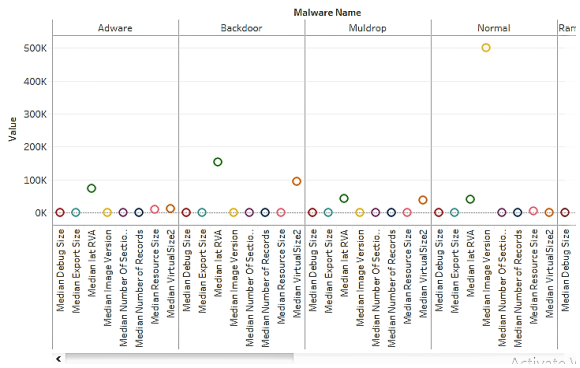

Fig.4. Representation of data in the form of "side-by-side circles"

This diagram visualizes all the measure values from the dataset for every malware using the "side-by-side circles" visualization technique.

In our data set we have eight attributes which describe the properties of the malware. Using this visualization technique we have shown the SUM value of all the attributes for each malware type.

3. DATA ANALYSIS

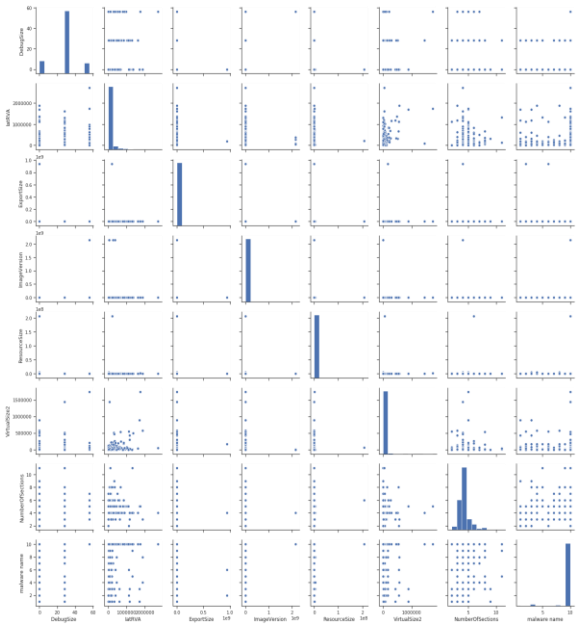

Fig.6. Feature Pair Plot

Anomalies.xls- The dataset contains the attributes such as debug size, latRVA, export size, image version, resource size, virtual size and number of sections. The records have class labels based on the attack/normal behavior such trojan, adware, backdoor, muldrop, Sdbot, spam, Rbot and ransomware and normal. The total number of records in the dataset is 5724.

The parameters that are used for data collection are as follows:

- Debug size: Files contain an optional debug directory that indicate what form of information present and where it is. The field that is considered is size.
- latRVA: Relative virtual address in an image file that address of an item after it is loaded into memory., with the base address of the image subtracted from it. The RVA of an item almost always differs from its position within the file on disk.

- Export size: Export directory table information size.
- Image version: The version of the image that is used for processing the operating system.
- Resource size: Resource directory table has the format of offset, size and field. The field that is used is resource size.
- Virtual size: Virtual size occupied by the particular sample.
- Number of sections: The basic unit of code or data within a portable executable file. Virtual size occupied by the particular sample.

Sheet 3

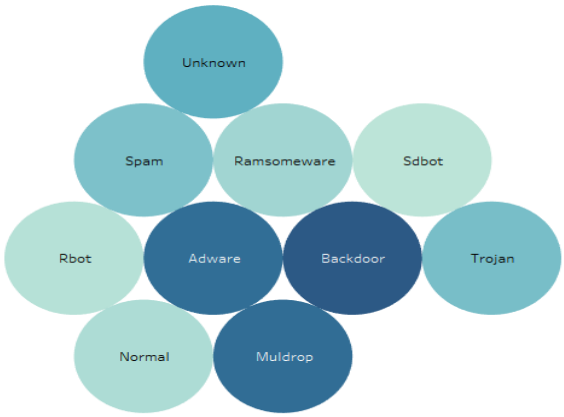

Fig.7. Representation of data using "Packed Bubbles" technique

The above diagram is a representation of the average value of the attribute "Virtual Size" against the malware type for a record, using "Packed Bubbles" visualization technique.

For one record, the virtual size of every category of malware is displayed.

The color of each bubble furthers simplicity of visualization by describing the magnitude of virtual size for each malware, i.e., higher the value of average virtual size, deeper the color of the bubble.

4. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. If there are n observations with p variables, then the number of distinct principal components is min(n-1,p). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

5. ALGORITHM

A. DECISION TREE

Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

B. RANDOM FOREST

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

C. LOGISTIC REGRESSION

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

6. ERROR ANALYSIS

7.

For each bootstrap sample, there is one third of data which was not used in the creation of the tree, i.e., it was out of the sample. This data is referred to as out of bag data. In order to get an unbiased measure of

the accuracy of the model over test data, out of bag error is used. The out of bag data is passed for each tree is passed through that tree and the outputs are aggregated to give out of bag error. This percentage error is quite effective in estimating the error in the testing set and does not require further cross validation.

7. FINE TUNING PARAMETERS

Cost functions try to find most homogeneous branches, or branches having groups with similar responses.

Regression : sum(y—prediction)²

Classification : G = sum(pk * (1—pk))

A Gini score gives an idea of how good a split is by how mixed the response classes are in the groups created by the split.

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). We use standard deviation to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous its standard deviation is zero.

- Standard Deviation (S) is for tree building (branching).
- Coefficient of Deviation (CV) is used to decide when to stop branching. We can use Count (n) as well.
- Average (Avg) is the value in the leaf nodes.

	DebugSize	latRVA	ExportSize	ImageVersion	ResourceSize	VirtualSize2	NumberofSelctions	Malwarename
Count	540.0	5,4e+02	5.4e+02	540.0	5,4e+02	540.0	540.0	540.0
Mean	4.356	1.7e+05	1.04e+07	77844.44	2.8e+05	53056.92	4.72	4.85
std	10.16	3.3e+05	9.8e+07	193301.01	9.5e+05	133575.7	1.69	2.58
min	0.0	5.9e+03	0.0e+00	0.0	0.0e+0	12.0	2.0	1.00
25%	0.0	2.5e+04	0.0e+00	0.0	7.4e+02	2398.0	3.0	3.00
50%	0.0	7.3e+04	0.0e+00	0.0	1.8e+04	4793.0	5.0	4.00
75%	0.0	1.6e+05	0.0e+00	0.0	1.8e+05	32768.0	5.0	8.00
max	28.0	1.8e+06	9.4e+08	600000.0	4.6e+06	888832.0	11.0	9.00

8. CLASSIFICATION

CART(Classification and Regression Trees) uses Gini index as the classification metric. Iterative Dichotomiser 3(ID 3) uses Entropy function as the classification metrics.

Initially we are presented with 7 feature variables and Malware Type is the target variable. The target variable consists of 10 different categories of malware labels.

Data preprocessing and Principal Component Analysis was performed

The problem was approached with different machine learning algorithms for prediction purpose.

```
In [50]: prob

Out[50]: [0.9272409778812573,
          0.9272409778812573,
          0.9289871944121071,
          0.9272409778812573,
          0.9272409778812573,
          0.9278230500582072,
          0.9266589057043073,
          0.9272409778812573,
          0.9284051222351571,
          0.9289871944121071,
          0.9278230500582072,
          0.9266589057043073,
          0.929569266589057,
          0.9284051222351571,
          0.9278230500582072,
          0.9278230500582072,
          0.9278230500582072,
          0.9307334109429569,
          0.9278230500582072,
          0.9266589057043073,
          0.9266589057043073,
          0.9266589057043073,
          0.9254947613504074,
          0.9278230500582072,
          0.9289871944121071]
```

Fig.8. Accuracy Array

As a result IatRVA, number of sections and VirtualSize2 came out to be the principal features to determine the Malware Type.

Logistic Regression	91.1%
Decision Tree	92.1%
Random Forest	93.07%

IV. ACKNOWLEDGMENT

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

REFERENCES

[1] Eric Chien, Symantec Security Response, "Techniques of Adware and Spyware".
[2] New Jersey Cybersecurity and Communications Integration Cell, "Muldrop", link: https://www.cyber.nj.gov/threat-profiles/trojan-variants/muldrop
[3] Dr. Web Antivirus, "Linux.MulDrop.14", link:https://vms.drweb.com/virus/?_is=1&i=15389228
[4] Loucif Kharouni ,Trend Micro Threat Research, "SDBOT IRC Botnet Continues to Make Waves", link: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.626.6378&rep=rep1&type=pdf
[5] Ronny Richardson and Max North, "Ransomware: Evolution, Mitigation and Prevention", link:http://scholarspress.us/journals/IMR/pdf/IMR-1-2017.%20pdf/IMR-v13n1art2.pdf
[6] Alexandre Gazet, "Comparative analysis of various ransomware virii", link:http://download.adamas.ai/dlbase/ebooks/VX_related/Comparative%20analysis%20of%20various%20ransomware%20virii.pdf
[7] ZHU Zhenfang, "Study on Computer Trojan Horse Virus and Its Prevention", link:https://www.ijeas.org/download_data/IJEAS0208024.pdf
[8] Abdelshakour Abuzneid, "Detection of Trojan Horses by the Analysis of System Behavior and Data Packets", link: https://www.researchgate.net/publication/275891938_Detection_of_Trojan_Horses_by_the_Analysis_of_System_Behavior_and_Data_Packets
[9] Dove Chiu, Shih-Hao Weng, and Joseph Chiu, "Backdoor Use in Targeted Attacks", link:https://www.trendmicro.de/cloudcontent/us/pdfs/security-intelligence/white-papers/wp-backdoor-use-in-targeted-attacks.pdf
[10] Vangie Beal, "keylogger (keystroke logging)", Webopedia, link: https://www.webopedia.com/TERM/K/keylogger.html