Microsoft Malware Prediction

Bhavna Arora
Dept. of CSE
PES University
Bengaluru, India
bhavnaaro28@gmail.com

Akanksha Tonne
Dept. of CSE
PES University
Bengaluru, India
tonne.akanksh99@gmail.com

Athira AD
Dept. of CSE
PES University
Bengaluru, India
athiraasha1126@gmail.com

***Abstract*— Malware is a catch-all term to refer to any software designed to cause damage to a single computer, server, or computer network. Not only the effects of malware can generate damage to systems, they can also destroy a country. However, in combating malware, the prediction on malware behavior or development is as crucial as the removing of malware itself. This is because the prediction on malware provides information about the rate of development of malicious programs in which it will give the system administrators prior knowledge on the vulnerabilities of their system or network and help them to determine the types of malicious programs that are most likely to taint their system or network**

Keywords— Malware Prediction, Microsoft Malware, Computer and Network security, Potential threats

I. Introduction

Malware is designed to bypass security systems and avoid detection, making it extremely difficult for security teams to ensure that users and the wider business are not adversely impacted. Malware authors implement a variety of methods to achieve this circumvention, including using obscure filenames, modifying file attributes, mimicking legitimate program operations, and hiding processes and network connections. These obfuscation and evasion techniques are helped along by the sheer volume of emerging malware; it is estimated that 350,000 new variants are discovered every day.

Measured in terms of worldwide user numbers, Windows remains the number one operating system. The players in the malware industry are in full agreement, and so Microsoft systems are still cybercriminals main target of attack.

Malware Prediction is inherently a time series problem. There are also many other considerations. Are certain hardware configurations more prone to specific types of malware? Did a series of anti-virus definition updates reduce detection rates because they are stopping the threats using advanced security techniques? Did a Microsoft Windows feature reduce detection of various malware? (i.e. SmartScreen, PUA mode) Are certain web browsers (including Internet Explorer) reducing malware detection rates more than others? These are important considerations.

This project uses the Microsoft Malware Dataset released during the 2015 Malware Challenge, available on Kaggle.

II. Literature Survey

In [2] Detecting internet worms using data mining techniques, Journal of Systemics, Cybernetics and Informatics, proposed the idea is extracting variable length instruction sequences that can identify worms from clean programs using data mining techniques. According to the general statistics obtained from the instruction sequences the problem is formulated to be a binary classification problem and built using decision trees, bagging and random forest. The results had 95.6 % detection rate on novel worms.

In [3] Scalable, Behavior-Based Malware Clustering, explores the idea of identifying and grouping malware samples that exhibit similar behavior. Var

III. Problem Statement

The goal of this competition is to predict a Windows machine's probability of getting infected by various families of malware, based on different properties of that machine. The telemetry data containing these properties and the machine infections was generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender.

Malware detection is inherently a time-series problem, but it is made complicated by the introduction of new machines, machines that come online and offline, machines that receive patches, machines that receive new operating systems, etc. The dataset provided here has been roughly split by time.

IV. Approach

A. Understanding the dataset

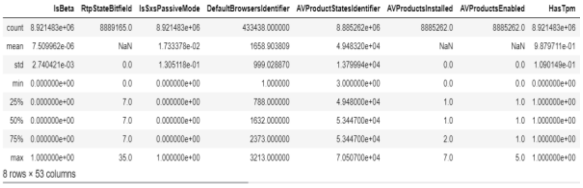
The size of the training and testing data is 9 million and 8 million rows, respectively. There are 81 features in total, with 52 being categorical, 23 of which are encoded numerically to protect the privacy of the information.

On analyzing the dataset we realized a few challenges faced are listed as follows:

- Large Dataset
- Many attributes

- Missing Values
- Categorical features

It is essential to resolve each of the above issues in the preprocessing stage before moving forward with the any visualization and classification.



	IsBeta	RtpStateBitfield	IsSxsPassiveMode	DefaultBrowsersIdentifier	AVProductStatesIdentifier	AVProductsInstalled	AVProductsEnabled	HasTpm
count	8.921483e+06	8889165.0	8.921483e+06	433438.000000	8.885262e+06	8885262.0	8885262.0	8.921483e+06
mean	7.509962e-06	NaN	1.733378e-02	1658.903809	4.948320e+04	NaN	NaN	9.879711e-01
std	2.740421e-03	0.0	1.305118e-01	999.028870	1.379994e+04	0.0	0.0	1.090149e-01
min	0.000000e+00	0.0	0.000000e+00	1.000000	3.000000e+00	0.0	0.0	0.000000e+00
25%	0.000000e+00	7.0	0.000000e+00	788.000000	4.948000e+04	1.0	1.0	1.000000e+00
50%	0.000000e+00	7.0	0.000000e+00	1632.000000	5.344700e+04	1.0	1.0	1.000000e+00
75%	0.000000e+00	7.0	0.000000e+00	2373.000000	5.344700e+04	2.0	1.0	1.000000e+00
max	1.000000e+00	35.0	1.000000e+00	3213.000000	7.050700e+04	7.0	5.0	1.000000e+00

8 rows × 53 columns

Fig. Description of the training dataset

B. Preprocessing

- Large Dataset:

  The dataset is huge with datatypes of the columns occupying significant amount of memory. To combat this problem, the columns' datatypes are converted to a lower datatype reducing the memory occupied with decrease in data load time.

  Eg: a column with datatype int16 is converted to int8

- Missing Values

  If not dealt appropriately, the missing data could lead to wrong inference from the data.

  1. **Missing value ratio**: Columns with more than 50% missing values are dropped from the data frame. There were 7 such columns.
  2. Impute the missing values: Since, most the data is categorical, the rows having missing values cannot be replaced mean or median.

  The solution we used for this problem was to replace the missing categories with the most occurring category in that column i.e. with the mode the column.

- Categorical Features:

  A variable having numeric type does not imply that it is quantitative, the numbers could represent categories/levels also.

  1. **Low variance filter:** From inspection of the dataset (nature and domain of the values an attribute can take) and calculating unique values under each. If a column contains categorical data but 80% of the rows fall in the same category, then that attribute would not play much significance in prediction. 26 columns are dropped from the dataset using this concept.
  2. **Frequency Encoding**: Using this encoding technique we replace the unique category with the frequency at which that unique value occurs throughout training data.

V. References

[1] Microsoft Malware Challenge, https://arxiv.org/abs/1802.10135

[2] Muazzam Siddiqui, Morgan C. Wang, and Joohan Lee. Detecting internet wormsusing data mining techniques.Journal of Systemics, Cybernetics and Informatics,pages 48–53, 2009.

[3] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel,and Engin Kirda. Scalable, behavior-based malware clusterin