SMU Data Science Review

Volume 2 | Number 2 Article 9

2019

Machine Learning to Predict the Likelihood of a Personal Computer to Be Infected with Malware

Maryam Shahini
Southern Methodist University, mshahini@smu.edu

Ramin Farhanian
Southern Methodist University, rfarhanian@smu.edu

Marcus Ellis
Northeastern University, mellis@bu.edu

Machine Learning to Predict the Likelihood of a Personal Computer to Be Infected with Malware

Maryam Shahini, Ramin Farhanian, Marcus Ellis

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{mshahini, rfarhanian}@smu.edu, mellis@bu.edu

Abstract. In this paper, we present a new model to predict the probability that a personal computer will become infected with malware. The dataset is selected from a Kaggle competition supported by Microsoft. The data includes computer configuration, owner information, installed software, and configuration information. In our research, several classification models are utilized to assign a probability of a machine being infected with malware. The LightGBM classifier is the optimum machine learning model by performing faster with higher efficiency and lower memory usage in this research. The LightGBM algorithm obtained a cross-validation ROC-AUC score of 74%. Leading factors and feature importance are also identified by LightGBM technique. Our research revealed that variables related to location, firmware version, operating system, and anti-virus software are the most important variables that have the highest weight in predicting malware detection.

1 Introduction

Malware is a malicious software that is designed to damage or disable computers. Once a computer is infected by malware, criminals can hurt consumers and enterprises, cause damage, and steal private information without the consent of the user. It can significantly impact a computers performance in many different ways such as disrupting network connection and operations, installing additional software and switching computer settings.

In todays' technology-driven world, we do not have the tools to predict the probability of a malware infection before it happens. To solve this problem, It is imperative to identify the factors that increase the risk of malware infection and take precautions necessary to prevent infections.

With more than one billion enterprises and consumer customers, Microsoft takes the malware infection problem seriously and is deeply invested in improving the security of its platform. This company is challenging the data science community to develop techniques to predict malware infection. As with their previous Malware Challenge (2015)[1], Microsoft has provided a malware dataset

[1] Microsoft Malware Classification Challenge. [online] Available at: https://www.kaggle.com/c/malware-classification

to encourage implementation of an effective technique to predict malware occurrences.

The dataset is provided by a Kaggle competition supported by Microsoft containing different machines' properties. The machine infections were generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender.

We present a machine learning model in an efficient way to avoid malware security issues for Microsoft clients before they happen. The first step in the process is to perform exploratory data analysis to understand the patterns and remove the unnecessary variables from our analysis[1]. Exploratory data analysis is done by eliminating variables with a high percentage of missing observations and suspected features. As a result, the dataset was left with 57 variables to use in the model[15].

Identifying the right machine learning algorithm to perform an analysis on Microsoft data is the next step. Many attempts are made with various algorithms; however, LightGBM has been determined to be the best for this application. The main advantages are faster training speed, high efficiency, scalability, higher accuracy, and lower memory usage. The ROC-AUC score of 74% is the best result that we have achieved .

LightGBM has also helped us identify specific characteristics and properties of the machine that have the higher weight in malware infection prediction. Some of the most important features are the "CityIdentifier"[2] and "Census_FirmwareVersionIdentifier"(The version id of the firmware[3]).

In the following section, a background of cybercrime is presented. Then we introduce a discussion on cybersecurity in section 3. Section 4 discusses more details regarding our collected data, steps in data preparation and initial insight. A description of different evaluation metrics is provided in section 5. Section 6 is dedicated to our optimum model and results. Ethical issues are addressed in section 7. Finally, the conclusions of our work and the main points of evidence are summarized in section 8.

2 Cybercrime

The proliferation of digital technology, and the convergence of computing and communication devices, has transformed the way in which we socialize and do business. While these aspects of digital technology are very positive, there has also been a dark side to these developments. Virtually every advance has been accompanied by a corresponding niche to be exploited for criminal purposes[4].

The magic of digital cameras and sharing photos on the Internet is exploited by child pornographers. Electronic banking and online sales have provided fertile

[2] the individual ID for the city that the machine is located in

[3] Firmware is a layer of software between hardware and the operating system with the main purpose to initialize and abstract enough hardware so that the operating systems and their drivers can further configure the hardware to its full functionality[19].

ground for fraud. Electronic communication such as email and text messaging may be used to stalk and harass. Our increasing dependence on computers and digital networks makes technology a tempting target for gaining information or as a means of causing disruption and damage.

The idea of a separate category of computer crime arose about the same time that computers became more mainstream in the society. As early as the 1960s there were reports of computer manipulation, sabotage, computer espionage and the illegal use of computer systems. The 1970s saw the first serious treatments of "computer crime". In subsequent decades, the increasing networking of computers and the proliferation of personal computers transformed computer crime and saw the introduction of specific computer crime laws.

The evolution of such legislation followed successive waves, reflecting concerns surrounding the misuse of computers. Initial concerns which related to unauthorized access to private information expanded into concern that computers could also be used for economic crimes. As computers became more and more centralized, the concern was to protect against unauthorized access to computer data. Increasing connectivity not only magnified these concerns. It gave rise to new problems, such as remote attacks on computers and networks, and gave new life to old offences such as infringement of copyright, distribution of child pornography, and global fraudulent schemes.

2.1 The Challenges of Cybercrime

Rapid technological development continues to present new challenges. The increasing uptake of broadband allows many home users to leave their computers connected to the Internet, thus making them more vulnerable to external attack. Peer-to-peer technology may not only be used to transfer illegal content, but also to orchestrate massive attacks. The convergence of telecommunications and computing has transformed mobile phones into miniature networked computers, with attendant potential for criminality.

According to Jonothon Clough, the three necessary factors to commit a crime are motivated offenders, opportunity, and the absence of capable guardians. While there are many opportunities for the offenders in the digital world, we summarize the key features of digital technology that facilitates the crime in table 1.

Challenge	Description
Scale	the Internet allows users to communicate with each other. Internet users provide an unprecedented pool of potential offenders and victims. This acts as a "force multiplier", allowing offending to be committed on a very large scale. The ability to automate several processes further amplifies this effect.
Accessibility	The technology has become ubiquitous and increasingly easy to use, ensuring its availability to both offenders and victims.
Anonymity	Anonymity is an obvious advantage for an offender, and digital technology facilitates this in different ways. Offenders may deliberately conceal their identity online by the use of proxy servers, spoofed email or IP addresses.
Portability	The ability to store enormous amounts of data in a small space, and to replicate it with no appreciable diminution of quality. Storage and processing power which would once have occupied rooms, now fits into a pocket. Images or sounds may be simply transmitted and at negligible cost to potentially millions of recipients. The convergence of computing and communication technologies has made this process a seamless one.
Global reach	Criminal law is traditionally regarded as local, being restricted to the territorial jurisdiction in which the offence occurred. Modern computer networks have challenged that paradigm. As individuals may now communicate overseas as easily as next door, offenders may be present, and cause harm, wherever there is Internet access. There is no need for offenders and victims to be in the same jurisdiction. Not only does this provide a world of opportunity for offenders, it presents enormous challenges to law enforcement.
Absence of capable guardians	An important factor which affects offending behavior is the risk of detection and prosecution. In this respect, the Internet presents law enforcement with a range of challenges. The volatile nature of electronic data requires sophisticated forensic techniques to ensure its retrieval, preservation and validity for use in a court. Apart from the sheer volume of users, the networked nature of modern communications makes surveillance extremely difficult.

Table 1: The Challenges of Cybercrime.

3 Cybersecurity

Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks. These cyberattacks[4] are usually aimed at accessing, changing, or destroying sensitive information; extorting money from users; or interrupting normal business processes. The best defense against the attacks is layers of protection that spreads across computers, networks, and software programs[5].

3.1 Types of Cyber Threats

Different types of cyber threats[6] are listed in table 2.

Cyber Threats	Description
Social Engineering	The process of psychologically manipulating internet users to perform certain actions or giving away information.
APT*	An unauthorized user infiltrates a network undetected and stays in the network for a long time.
Malware	Software that is specifically designed to gain access or damage a computer without the owner consent.

Table 2: Cyber Threats

* Advanced Persistent Threats

3.2 Malware

Malware is a large and growing part of the cybercrime industry. McAfee Labs Global Threat Intelligence identified more than 41 million new malware samples[18]. Total malware samples have grown 34% over the past four quarters to more than 774 million samples. Cybercriminals distribute malware by either infecting a popular website, attaching to emails or source codes of trusted application. Based on what the attacker hopes to achieve, the malware might collect information for commercial purposes, ask for ransom, mine cryptocurrencies, or remotely use the infected computers to launch DDos attacks[7] or Spam emails. Some types of malware[8] are summarized in table 3.

[4] Cyberattack is a deliberate exploitation of computer systems, technology-dependent enterprises and networks. [online] Available at: `https://bit.ly/2tpsTLG`

[5] What is cybersecurity? [online] Available at: `https://tinyurl.com/yxl3hovp`

[6] Types of cyber threats [online] Available at: `https://tinyurl.com/yxl3hovp`

[7] What is a DDoS Attack? [online] Available at: `https://bit.ly/2Ku5jGw`

[8] Understanding malware & other threats [online] Available at: `https://bit.ly/2G2CzCk`

Type	Description
Coin miners	With the rise of digital currencies *, criminals reconfigure malware to infiltrate an organization and secretly mine for coins.
Exploits	They take advantage of security holes in a software. They exploit these holes to bypass security safeguards to infect a device.
Macro Malware	Macros are a powerful way to automate common tasks in some software programs. However, macro malware uses this functionality to infect your device on personal computers.
Phishing	Phishing attacks attempt to steal sensitive information through emails, websites, text messages, or other forms of electronic communication that often seem to be official communication from legitimate companies or individuals.
Ransomware	A type of malware that encrypts files and folders, preventing access to important files. Ransomware usually attempts to extort money from victims by asking for money, in form of cryptocurrencies, in exchange for the decryption key.
Tech support scams	Scammers use scare tactics to trick users into paying for unnecessary technical support services.
Trojans	Trojans are a common type of malware which cannot spread on their own. They either have to be downloaded manually or another malware needs to download and install them.
Worms	A type of malware that can copy itself and spread through a network by exploiting security vulnerabilities. It can be spread through text messages, email attachments, file-sharing programs, social networking sites, network shares, removable drives and software vulnerabilities.

Table 3: Malware Types

* It is also known as cryptocurrency. It is a virtual currency that uses encryption techniques for security.

3.3 Incidents

Some of the most destructive malware programs of all time[9] are shown in table 4. These are only a few examples of the damages that malware attacks have caused. Viruses, worms, Trojan horses and ransomware are severely damaging, wreaking havoc across business, government, and computers[2].

Masked by internet anonymity, cybercriminals are evolving quickly, constantly unleashing new and improved malware which is a threat to our online safety. Despite years of research and development, it has not been possible to develop security design and implementation techniques that systematically exclude security flaws and prevent all unauthorized actions which might lead to such attacks[17].

In the absence of such foolproof techniques, it is useful to have a set of widely agreed principles that can help us avoid malware issues[11]. Certain practices are suggested to deal with malware hacks[6]. Being diligent on downloading content from email attachments or Internet, updating the installed software programs and antiviruses, and checking the personal accounts and credit reports are different ways to avoid the issue[14]. It is also suggested to back up the files in case the computer is infected [10].

[9] The 8 most Famous Computer Viruses of All Time. [online] Available at: `https://tinyurl.com/y5fo8xkh`

[10] McAfee Security Tips. [online] Available at: `https://bit.ly/2Kvolfv`

Malware Incident	Type	Date
Stuxnet is a computer worm that was originally aimed at Irans nuclear facilities and has since mutated and spread to other industrial and energy-producing facilities. The original Stuxnet malware attack targeted the programmable logic controllers (PLCs) used to automate machine processes.	Worm	2010
CryptoLocker spread through email attachments and encrypted the user's files so that they could not access them. The hacker later exchange the encryption key in return for a sum of money.	Ransomware	2013
ILOVEYOU was a worm that was downloaded by clicking on an attachment. It overwrote system files and personal files and spread itself over and over again.	Worm	2000
MyDoom is the fastest-spreading email-based worm ever. It hit tech companies with a Distributed Denial of Service attack.	Worm	2004
Storm Worm was a Trojan horse that infected computers, sometimes turning them into bots to continue the spread of the virus and to send a huge amount of spam mail.	Trojan	2006
Sasser & Netsky are actually two separate worms, but they are often grouped together because the similarities in the code led experts to believe they were created by the same person. Sasser spread through infected computers by scanning random IP addresses [*] and instructing them to download the virus. Netsky was the more familiar email-based worm. It was more viral, and caused a huge amount of problems.	Worm	2004
Slammer is a very viral virus and did spread itself every few seconds. It impacted on banks, police and airline IT operations.	Worm	2013

Table 4: The Most Destructive Malware of all Time.

[*] IP is a unique string of numbers separated by periods that identifies each computer using the Internet Protocol to communicate over a network.

3.4 The Importance of Data Science in Cybersecurity

Data science was initially used for malware detection and attribution[11]. It is critically important for the future of cybersecurity for three reasons. First, security is all about data. When we seek to detect cyber threats, we are analyzing data in the form of files, logs, network packets, and other artifacts. Traditionally, security professionals did not use data science techniques to make detection based on these data sources. Instead they used file hashes, custom written rules like signatures, and manually defined heuristics. Although these techniques have their merits, they required handcrafted techniques for each type of attack, necessitating too much manual work to keep up with the changing cyber threat landscape. In recent years, data science techniques have become crucial in bolstering our ability to detect threats. They are important to cybersecurity because the number of cyber-attacks on the internet has grown dramatically[12]. Figure 1 illustrates the recent growth of malware.

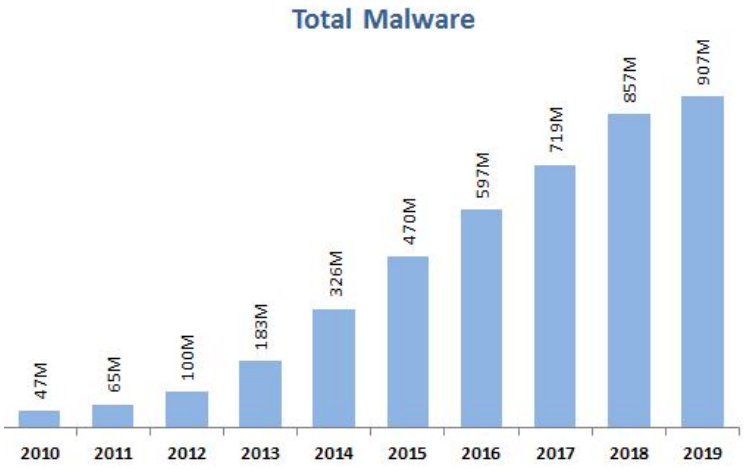

Fig. 1: Total Malware Throughout the Years

3.5 Microsoft and Malware

All platforms have to invest in detecting malware to protect their clients. However, when it comes to the market share, Microsoft has a well established market of personal computers. As shown in figure 2, there are various sizes of malware

[11] Microsoft Malware Classification Challenge.[online] Available at: `https://www.kaggle.com/c/malware-classification/overview`

[12] Malware. [online] Available at: `https://www.av-test.org/en/statistics/malware/`

for different platforms and it exposes the hassle of such malicious software for Microsoft business. The extensive usage of machine learning has enabled the company to secure its platforms. However, every once in a while, one of the practices might be compromised due to negligence, cost or applying the security at the last iteration. Using the current cloud computing power, no key exists which cannot be copied, no lock which cannot be picked, and no fortification which cannot be breached. The only factor is the inconvenience of those who should have access and those who should not[13]. That is why innovation is essential to avoid security issues or at least minimize the cost and damage.

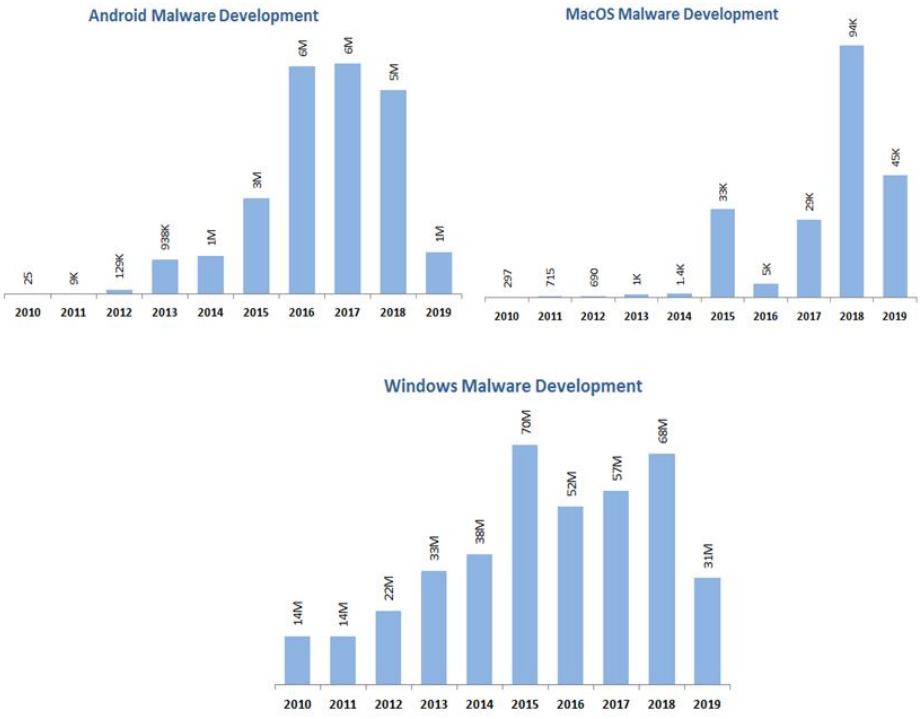

Fig. 2: Size of Malware in Different Platforms by Year

4 Microsoft Data Collection and Preparation

The dataset[14] does represent all Microsoft user machines. The size of the train and test data are about 9 million and 8 million records, respectively. There are 83

[13] Cryptography and Network Security by William Stalling, 7th Edition

[14] Microsoft Malware Data Description.[online] Available at: `https://www.kaggle.com/c/microsoft-malware-prediction/data`

features in total, with 52 being categorical[5]. Each row in this dataset represents a machine that is uniquely identified by a Machine Identifier. "HasDetections" column is the ground truth and indicates if malware was detected on the machine. Using the information and labels in train dataset, the value for "HasDetections" for each machine in test dataset will be predicted.

As shown in figure 3, the number of machines that are infected by malware is equal to the number of machines that are not detected by malware. Therefore, the dataset is very well balanced and that will lead to less challenges in dealing with approaches in statistical machine learning.

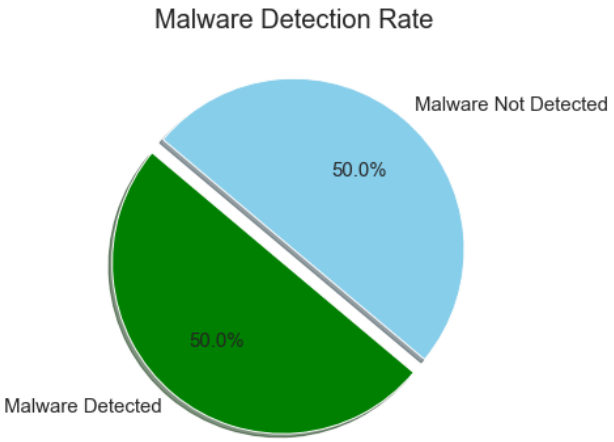

Fig. 3: Distribution of Malware Detection Rate in Microsoft Dataset

In fact, just by looking at the columns, some features are apparent candidates to explore and help to determine if the machine would be infected. Columns such as City[15] and Country Identifier[16], "AVProductsInstalled" [17] and "OSVer"[18].

As shown in Fig. 4, 52 percent of the infected machines have "x64" processors[19]. There is no indication that either 64-bit or 32-bit processor is more secure than the other, but this reveals the preferred hardware of Microsoft clients[20].

[15] Identifier for the city the machine is located in.
[16] Identifier for the country the machine is located in.
[17] Number of Anti-Virus installed products.
[18] The version of the operating system
[19] The 64 bit computers can run both 32bit programs and 64 bit programs. 32 bit computers cannot run 64 bit programs, because the bit sizes are fundamentally different. Latest Laptops with pre-installed Windows are usually x64.
[20] Difference Between 32 bit & 64 bit Operating Systems.[online] Available at: http://net-informations.com/q/mis/x86.html

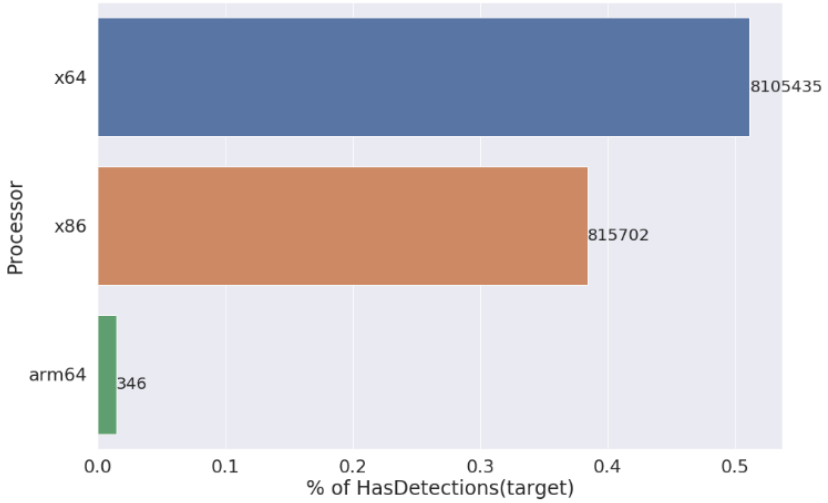

Fig. 4: The malware detection rate on x86, x64, and arm64 processors.

Figure 5 illustrates the total number of machines in each of the 222 country identifiers in the dataset[21]. The data reveals that the distribution of malware infection in different countries is not the same. One of the reasons for the inconsistency across the countries and cities is user education and supporting users to remove infections from their computers. For instance, Japan is one of the countries with the lowest malware infection rate. One of the reasons is that Cyber Clean Center, a cooperative project between Internet Service Providers, major security vendors, and government agencies, have invested on user education and improving the security[22].

[21] Country names are anonymized in the dataset.

[22] Japan Lessons from Some of the Least malware Infected Countries in the World.[online] Available at: `https://tinyurl.com/y5neykwq`

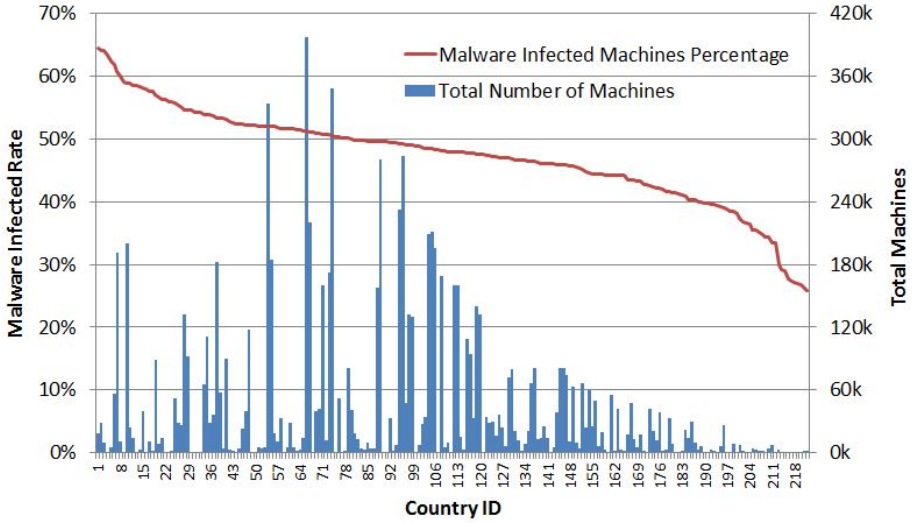

Fig. 5: Total number of machines and malware infection rate for each country

4.1 Checking Columns with Missing Data

As shown in Table 5, columns "PuaMode", "Census_ProcessorClass", "Default-BrowsersIdentifier", "Census_IsFlightingInternal", and "Census_InternalBatteryType" have over 70% null data. Therefore, they are removed from the dataset as they do not provide enough value for target prediction.

Feature	Unique Values	Missing %	Type
PuaMode	2	99.97	category
Census_ProcessorClass	3	99.59	category
DefaultBrowsersIdentifier	1730	95.14	float16
Census_IsFlightingInternal	2	83.04	float16
Census_InternalBatteryType	78	71.05	category

Table 5: Top Features with Missing Data.

4.2 Checking Multicollinearity

The dataset contains some redundant and highly correlated features. Some of them are almost identical. In this section, such columns are identified and removed from train and then later from test data sets. Correlation analysis measures the statistical relationship between two different features. The results will show how the change in one parameter will impact the other. Multicollinearity detection must be executed before building the model and arriving at any conclusion about variable relationships. In fact, correlated features do not improve the model and removing them from the model make the learning algorithm faster and decrease the bias and overfitting. Though correlation analysis is a great help in understanding the association between features in a dataset, it can't explain, or measure, the cause. To achieve the correlations for categorical columns, the first step is to perform encoding process and building a correlation matrix. Table 6 displays the top 10 features with highest correlations.

Feature 1	**Feature 2**	**Coeff**
OSUILocaleIdentifier	OSInstallLanguageIdentifier	0.99
OSBuildNumber	OsBuild	0.94
InternalResolutionHorizontal	InternalResolutionVertical	0.90
IsSxsPassiveMode	RtpStateBitfield	0.89
ProcessorModelID	ProcessorManufacturerID	0.85

Table 6: Top 5 Pair of Features with Highest Correlation

4.3 Categorical Feature Encoding

The existence of more than 50 categorical features creates a major challenge in this research. It is very common to see categorical features in a data set. However, machine learning algorithm can only read numerical values. It is essential to encode categorical features into numerical values. LabelEncoder and One-HotEncoder have been used as an encoding categorical feature process in this research. Filling missing values is essential before encoding, therefore, missing values have been imputed by the most frequent values. Then, One-Hot-Encoding is performed to create new features encoded in binary formats, to use in prediction model.

After removing the columns that are either unique, unnecessary, highly correlated, and have high percentage of missing data, the final data set specified 57 variables.

5 Model Evaluation Metric

Using the right evaluation metric for classification system is crucial to choose the best model with the highest precision[7]. Choosing the wrong metric may lead to a model that seems to be performing well with the training data, but it will not be as accurate with the test data.

After considering all evaluation metrics for classification systems, we ended up using ROC Curve. Area under ROC Curve[23] is a performance metric for binary classification problems.

In statistics, a receiver operating characteristic curve, ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. In fact, a ROC curve can be used to select a threshold for a classifier which maximizes the true positives, while minimizing the false positives. We usually use ROC when the detection of both classes are as important.

The AUC represents a models ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random. Most classifiers have AUCs that fall somewhere between these two values. Therefore, the overall model performances can be compared by considering the AUC. A brief description of each one[24] is listed in table 7. In addition to ROC-AUC metric, we have used other classification metrics in our models as well.

Metric	Description	Equation
Precision	Proportion of positive cases that were correctly identified	(TP) / (TP + FP)
Recall	Proportion of actual positive cases which are correctly identified	(TP) / (TP + FN)
Accuracy	Proportion of the total number of predictions that were correct over all kinds predictions made	(TP + TN)/(TP + FP + FN + TN)

Table 7: Other Types of Metrics

[23] AUC for short

[24] Evaluation Metrics in Python.[online] Available at: `https://tinyurl.com/y2rvvkvb`

6 Analysis

6.1 LightGBM Model Selection

The LightGBM[25] is one of the most successful machine learning frameworks that uses tree based learning algorithms. It is a gradient boosting framework and is designed to be efficient and powerful in solving problems[13]. Some of the most important advantages of this framework[26] is listed in table 8.

Index	Advantages
1	Lower memory usage
2	Faster training speed
3	Higher efficiency
4	Support of parallel and GPU learning
5	Capable of handling large-scale data
6	Better accuracy in results

Table 8: LightGBM Framework Advantages

LightGBM grows tree vertically, leaf-wise while most decision tree learning algorithms grows trees horizontally, level-wise. As shown in Figure 6, LightGBM choose the leaf with max delta loss to grow. When growing the fixed leaf, Leaf-wise algorithm can achieve the lower loss and reduce more loss than a level-wise algorithm[27]. Figure 6 represents how leaf-wise and level-wise tree growth work. Leaf-wise tree growth is the representation of how LightGBM works[9].

6.2 LightGBM Execution and Results

A 5-fold cross validation with a train test split of 80/20 was used when creating the model[28]. The folds preserved the percentage of samples for each class using

[25] LightGBMs documentation.[online] Available at: `https://lightgbm.readthedocs.io/en/latest/`

[26] "LightGBM: A Highly Efficient Gradient Boosting Decision Tree".[online] Available at: `https://tinyurl.com/y2ectjlo`

[27] CatBoost vs. Light GBM vs. XGBoost .[online] Available at: `https://tinyurl.com/y2unc3d6`

[28] The GitHub repository. [online] Available at: `https://bit.ly/2NaiY7r`

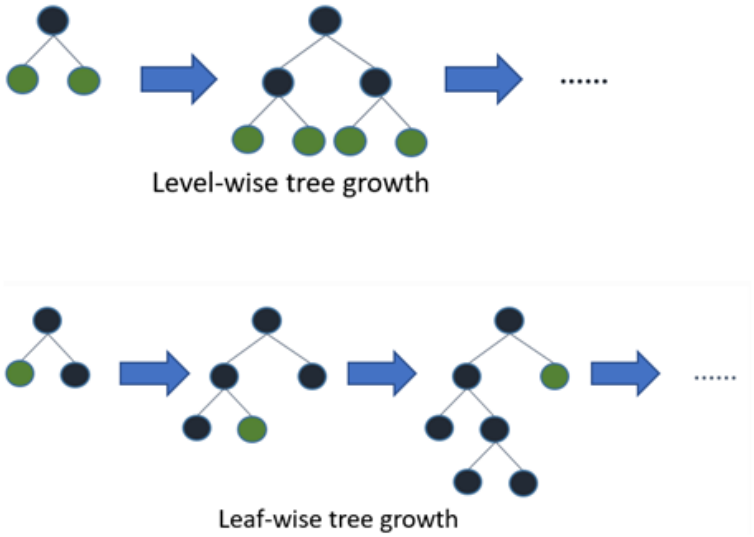

Fig. 6: The Difference Between Level-Wise and Leaf-Wise Tree Growth

stratified kfold. Simply, it helps to ensure that target relative class frequencies is approximately preserved in each train and validation fold.

The results reveal that LightGBM algorithm is the best approach based on ROC/AUC evaluation metric with a score of 0.74. This means that the area between the ROC curve and the axis is 0.74. It indicates that, the model performed relatively well in distinguishing the given classes, in terms of the malware probability prediction.

6.3 Identifying Leading Factors and Feature Importance

LightGBM feature importance analysis is the best performing classifier in this research to determine the variables that play the biggest role in malware prediction on machines[10]. As shown in table 9, the features are accompanied by variable description along with the weight that it has on the LightGBM model. In this paper, we proceed to find the level of importance of each feature and select relevant features to guarantee best model performances using recursive feature elimination.

Feature selection is a process where we automatically select those features that contribute most to the outcome we are trying to predict. Having irrelevant features in the data can decrease the accuracy of the models. Some of the benefits of feature selection include reduce overfitting, improving accuracy and reducing computation time.

Feature	Description	Weight
CityIdentifier	Identifier of the city the machine is located in	42595
FirmwareVersionIdentifier	Identifier of the firmware version	39648
SystemVolumeTotalCapacity	The size of the partition that the System volume is installed on in MB	37285
OEMModelIdentifier	Original equipment manufacturer model identifier	36104
ProcessorModelIdentifier	Processor model identifier	34616
CountryIdentifier	Identifier of the country the machine is located in	31726
OSBuildRevision	Operating System Build revision	28113
AVProductStatesIdentifier	Identifier of the specific configuration of a user's antivirus software	27793
PrimaryDiagonalDisplaySize	Retrieves the physical diagonal length in inches of the primary display	23463
OEMNameIdentifier	Original equipment manufacturer name identifier	23152
GeoNameIdentifier	Identifier of the geographic region the machine is located in	21705
Wdft_RegionIdentifier	–	21148
LocaleEnglishNameIdentifier	English name of Locale identifier of the computer owner	19614
IeVerIdentifier	Internet Explorer version identifier	19051
OSInstallLanguageIdentifier	Identifier of the main language installed on the operating system	18274

Table 9: Feature Importance of the Top 15 Variables used to Predict Malware Infection

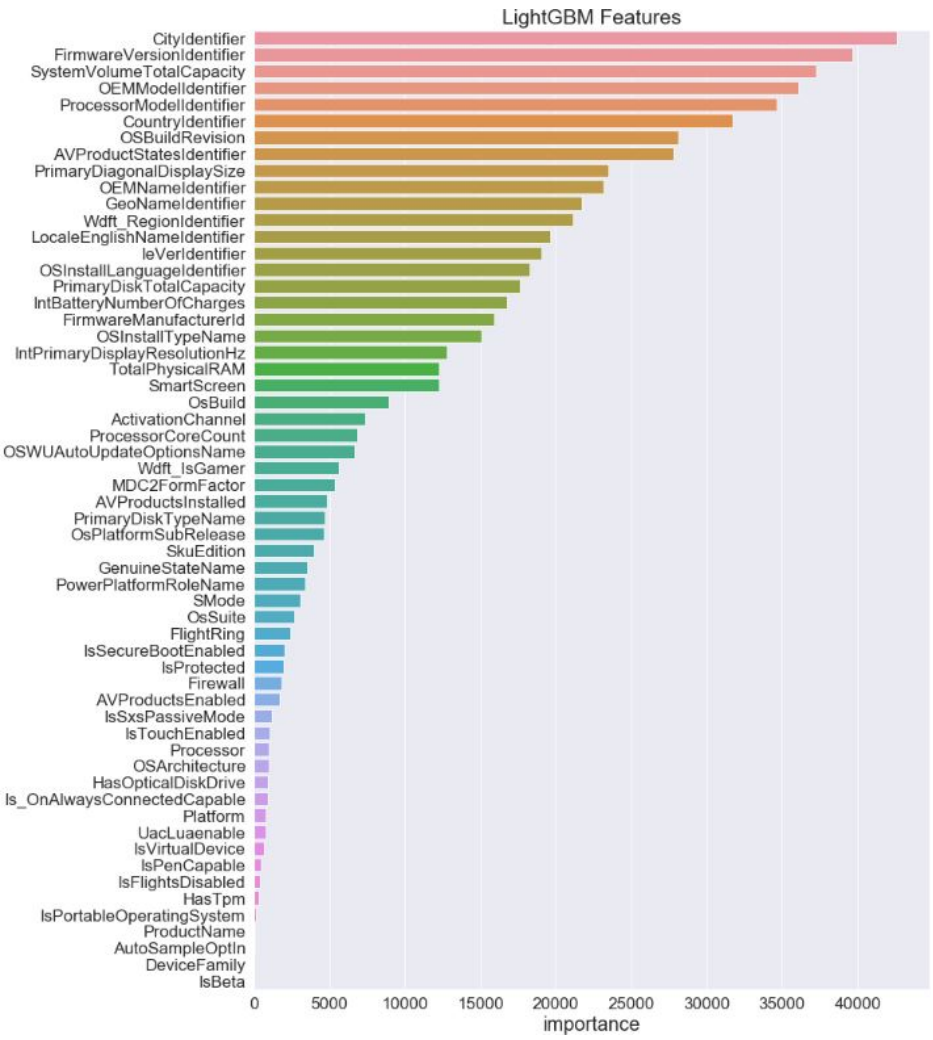

Fig. 7: Feature Importance of Variables Used to Predict Malware Infection

7 Business Ethics

Predicting malware infection needs different kinds of information about the owner of the computer, the installed software and how it is configured. User privacy will be compromised as we try to detect hackers trends in time.

According to the statistics published by Gartner[29], global security spending hits $96 Billion by the end of 2018, which is 8% more than what it was in 2017

[29] Gartner [online] Available at: `https://www.gartner.com/en/about`

and will be 35% more by 2020[30]. This is due to the arrival of new regulations (like GDPR[31]), development of online business strategy, and sophistication of hacker tools. Today, as Gartner claims, more than 53% of companies state data breaches as their number one risk. This trend might lead IT companies to choose a less costly reactive approach than a secure proactive one.

When we consider the overwhelming trend analysis techniques that have been recently introduced to our lives[12], compromising security will cause an unpredictable impact in our societies. Certain reactive strategies to predict the hack should be used as the last resort. Table 10 shows the new equipment maintenance techniques using trend analysis. Security becomes more and more essential as new trend analysis techniques are changing the way elevators, automobiles and so many other devices around us fail.

New Equipment Maintenance Techniques
Detecting failures in early stages and preventing them
Finding "Remaining life of asset"
Schedule predictive maintenance
Maintaining "Right level of inventory" for spare parts
Evaluate "What if" alternate scenarios
Determine right warranty period for the assets at the design time
Predict breakdown
Notify operator at right time
Prevent risk of collateral damage and secondary failure
Prevent high production downtime
Maximizing equipment life

Table 10: New equipment maintenance techniques using trend analysis.

[30] Cybersecurity is becoming more and more expensive.[online] Available at: `https://tinyurl.com/y3tp9pvv`

[31] GDPR definition [online] Available at: `https://tinyurl.com/yxgjkgvj`

8 Conclusions

Although best practices such as updating security patches and installing antivirus software on a computer are useful, they cannot solely protect a computer from malware infection. We conclude that legal and technical cybersecurity institutions and frameworks[16], policy coordination institutions, national training programs[3], and information sharing networks are the most important factors that reduce the malware infection[8] at the national level.

References

[1] Usukhbayar Baldangombo, Nyamjav Jambaljav, and Shi-Jinn Horng. "A static malware detection system using data mining methods". In: *arXiv preprint arXiv:1308.2831* (2013).

[2] Ulrich Bayer et al. "A View on Current Malware Behaviors." In: *LEET*. 2009.

[3] Razvan Beuran et al. "Towards effective cybersecurity education and training". In: (2016).

[4] Jonathan Clough. *Principles of cybercrime*. Cambridge University Press, 2015.

[5] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support". In: *arXiv preprint arXiv:1810.11363* (2018).

[6] Josiah Dykstra. *Essential cybersecurity science: build, test, and evaluate secure systems*. " O'Reilly Media, Inc.", 2015.

[7] Peter A Flach. "The geometry of ROC space: understanding machine learning metrics through ROC isometrics". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 194–201.

[8] Energy Index. "Global". In: *Nature* 522.7556 (2015), S1–27.

[9] Akhil Kadiyala and Ashok Kumar. "Applications of python to evaluate the performance of decision tree-based boosting algorithms". In: *Environmental Progress & Sustainable Energy* 37.2 (2018), pp. 618–623.

[10] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3146–3154.

[11] George Kostopoulos. *Cyberspace and cybersecurity*. Auerbach Publications, 2017.

[12] Jay Lee et al. "Intelligent prognostics tools and e-maintenance". In: *Computers in industry* 57.6 (2006), pp. 476–489.

[13] Qi Meng et al. "A communication-efficient parallel algorithm for decision tree". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1279–1287.

[14] Abbas Moallem. *Human-Computer Interaction and Cybersecurity Handbook*. CRC Press, 2018.

[15] Karthik Raman et al. "Selecting features to classify malware". In: *InfoSec Southwest* 2012 (2012).

[16] Regner Sabillon, Victor Cavaller, and Jeimy Cano. "National cyber security strategies: global trends in cyberspace". In: *International Journal of Computer Science and Software Engineering* 5.5 (2016), p. 67.

[17] Eric C Thompson. *Cybersecurity Incident Response: How to Contain, Eradicate, and Recover from Incidents*. Apress, 2018.

[18] Ronghua Tian et al. "Differentiating malware from cleanware using behavioural analysis". In: *2010 5th international conference on malicious and unwanted software*. IEEE. 2010, pp. 23–30.

[19] Vincent Zimmer et al. *Embedded Firmware Solutions: Development Best Practices for the Internet of Things*. Apress, 2015.