A Cyber Security Survey &

Using machine learning techniques to classify malicious code

Project Plan and Overview

University of Manchester

Word Count: 2289

Kamran Khan 7574393

12/05/2017

Contents

List of Figures

1 Introduction

The first web site came online in 1991 [1], since then web sites, have evolved to become an important communication channel for all, including governments, businesses and people. Governments use web applications to inform its citizens and businesses have adopted the use of web and mobile applications to advertise and sell their products or services. Web applications have grown due to the unlimited number of use cases. As a society we have become dependent on the web for its tools and applications to carry out our day to day activities. It was estimated in 2016, 81% of the developed world is now using the Internet every day [2]. Web applications face a higher security threats than ever before, cisco had reported in 2016, that 49% of organization's had to manage a public security breach [3]. Web applications can be prone to vulnerabilities, even the more well established and hardened web applications still fall vulnerable to the most basic attacks. In 2012, LinkedIn an employment orientated social networking website was hacked. With more than 100m records of data being breached the probable cause was later identified to be a simple SQL injection attack [4]. Many web applications house sensitive information, the potential gain of accessing this information is enough motivation for attackers. The damages caused by a security breach can be very significant, both economically and socially. Perhaps these vulnerabilities are due to time constraints, financial constraints, lack of security awareness and understanding. We also have to consider whether these web applications, their code and data have become too big for us to effectively and efficiently secure them manually.

The purpose of this project is to create a cyber security survey, to investigate, identify and review yesterdays, todays and tomorrows security threats and their solutions. Investigate the security threats that affect web applications, identify the cause of these threats and how they work, review the implications such threats can have and finally determine the possible solutions for these threats. Along side this survey, a system will be developed, where a machine is trained so that it is capable of accurately predicting/classifying code as either malicious or benign, through the use of artificial intelligence specifically machine learning techniques. The specifics of this proposal has been provided in the methodology section of this report.

The main report, will be aimed towards those who are beginners in the field of web application security and those who may already be knowledgeable in this field. The rest of the report will include the following; In section 2 a brief summary of the background. Section 3 describes the methodology of the project and how the project goals will be achieved. Section 4, contains the plan of the project outlining the aims and objectives. Section 5 will take into consideration some of the ethical and professional aspects of the project. Finally, section 6, describes how the project will be tested and evaluated.

2 Background

Web application security deals with the security of websites, web applications and web services. The Open Web Application Project attempts to document the critical security risks to web applications in their OWASP Top 10 report [5]. The aim of the document is to bring awareness to some of the most critical security risks to web applications. Some of the threats listed in their latest 2017 top 10 version include, injection attacks and cross-site scripting attacks. These two types of attack can be categorized as malicious code. In 2015/2016 an annual report was produced by CERT-UK now apart of the National Cyber Security Centre (NCSC), analyzing the state of cyber security incidents. It showed that malicious code accounted for 30% of security incidents which were reported making it the largest incident category [6].

Artificial intelligence specifically machine learning techniques can be useful in various fields of work. In the age of big data, information is being generated very rapidly, people are looking towards automated systems to help them analyze and manage this big data, using AI or machine learning to manage code could be useful. What is machine learning? "Machine Learning is to be intelligent, a system that is in a changing environment should have the ability to learn. If the system can learn and adapt to such changes, the system designer need not forsee and provide solutions for all possible situation" [7].

There are two main machine learning techniques, supervised learning and unsupervised learning. The idea of trying to predict if some code is malicious or benign is simply a supervised learning problem or classification problem. Supervised learning in its simplest case, is learning a class (in this instance a piece of code) from it's positive (malicious) and negative (benign) examples. During the learning process certain attributes or features of the class are extracted to help later identify the same example. Then, when given a new example that it has not seen before, can it make a prediction whether this new example is malicious or benign.

Precedence has already been set in using machine learning in a security environment. Research was conducted to create a machine which could classify emails, as either "phishing/spam" or "not spam". They collected samples of around 6950 non-spam emails and 860 spam emails. They then attempted to train a machine to identify if a new email was "spam" or "not spam" based on the samples they had collected. The results of the experiments showed that the machine the authors had developed had an overall accuracy of 99.5% with a false positive rate of 1% [8]. The author of the paper proved that it is indeed possible to use machine learning to detect security threats.

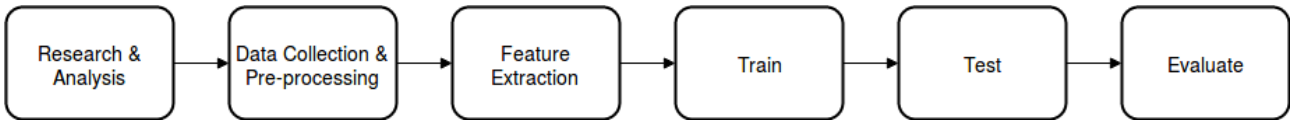
3 Methodology

This part of the report coveres how the project will be achieved and what work will be done. The topics include the project workflow and the tools and techniques that will be used.

3.1 Project Workflow

To achieve the goals of this project, there are 6/7 major steps or activities which will be carried out. Figure 3.1 below outlines the workflow for this project.

Figure 3.1: Project Workflow



1. **Research** - Research the current state of web application security and the existing defences aswell as machine learning techniques.

2. **Analysis** - Produce a survey/state-of-the-art for web application/cyber security threats by analyzing their cause, the impact they have and their solution. Focusing on the top security threats identified by the research.

3. **Data Collection & Pre-Processing** - Collect malicious and benign code. Clean the data removing unnecessary items, errors or inconsistencies. Format the data so it is suitable for the next stage.

4. **Feature Extraction** - Extract features from the code collected using the TF-IDF feature extraction method. The features make up important parts of the code, i.e parts that have a greater probability of being malicious.

5. **Train** - Use a classification algorithm such as NaiveBayes to teach a machine to label a piece of code as either malicious or benign, based on the data collected and the features extracted.

6. **Test** - Test the machine with a new set of data (code) which it hasn't seen before i.e data which it was not trained on. To evaluate the performance of the machine.

7. **Evaluate** - Evaluate the performance of the machine, calculating the accuracy, recall, precision and f1 results and an evaluation of the methodologies used and objectives met.

From the background research, SQL Injection and Cross-Site Scripting (XSS) are two of the more prevalent attacks. The projects system development will focus on these two types of malicious code.

3.2 Tools

There are many different programming languages which can be used for machine learning. In 2016 a survey was conducted, the survey tried to find out which programming languages were popular with machine learning. As shown in the figure 3.2 below, over the past 2 years all of the languages present have grown in popularity, python is regarded as a popular choice for machine learning jobs.

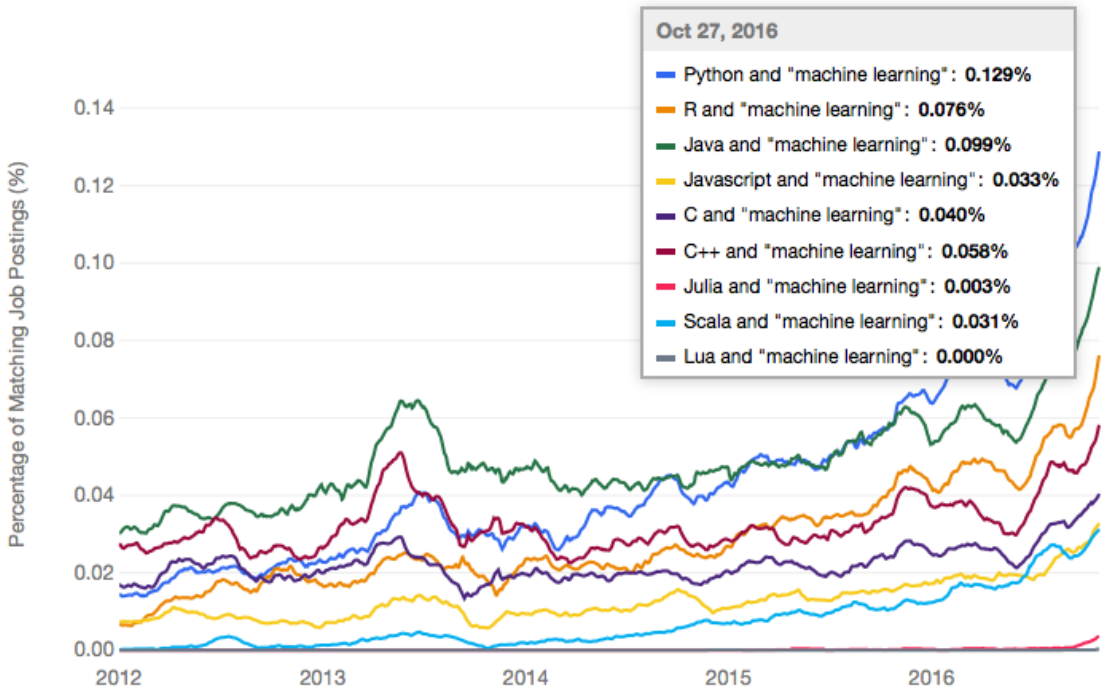


Figure 3.2: Popular languages for machine learning [9]

The language R is another popular statistical analysis programming language for machine learning as it is very friendly for beginners. However, pythons popularity in machine learning is increasing very rapidly, it has many advanced libraries and frameworks such as scikit-learn which is mature enough to be used in a production environment. For these reasons and more i have chosen Python as the programming language to develop in.

4 Plan

The project began on the 13th of March 2017 and has an expected finish date on 8th of September 2017. There are 3 main deliverables for this project. Firstly, a report containing the project plan and overview of around 2500 words, secondly a dissertation consisting of around 18,000 words and finally a working machine/system.

4.1 Aim & Objectives

The aim of this project is to investigate, identify and review yesterdays, todays and tomorrows security threats and their solution. Additionally, the aim of this project is to train a machine to see if it can classify whether a piece of code is either malicious or benign. The objectives for this project are based on trying to achieve the main aims of this project and they have been outlined below.

1. Research
    (a) To research security vulnerabilities so that i have a comprehensive understanding to be able to analyze their cause, their impact and their solution.
    (b) To investigate machine learning and evaluation techniques so that i can build a machine capable of labeling code and evaluating its performance.
2. Analysis
    (a) To combine and collate security vulnerabilities for analysis so that i can investigate their cause, review their impact and describe their solution for the intended audience.
3. Development
    (a) To collect and pre-process data (code) so that i can extract the important features from them and train a machine to label a new set of data as malicious or benign.
4. Testing
    (a) To test my machine with a new set of data so that i can evaluate the performance of the machine.
5. Evaluation
    (a) To evaluate the performance of my system using standard evaluation metric researched to determine if this method is useful in helping to secure web applications.
    (b) To evaluate the system & and methodologies used, so i can then discuss areas for further work to build upon or improve this technique.

4.2 Milestones

1. (12/05/2017) - Submission of the project plan and overview.

2. (26/06/2017) - Draft of the security threat analysis.

3. (17/07/2017) - Prototype version of system/machine

4. (07/08/2017) - Draft of the final dissertation for review.

4.3 Gantt Chart

The gantt chart as seen in figure 4.1 shows the deliverables and timeline for this project. It is based upon the objectives outlined in section 4.1

Gantt Chart			13/03/2017	20/03/2017	27/03/2017	03/04/2017	10/04/2017	17/04/2017	24/04/2017	01/05/2017	08/05/2017	15/05/2017	22/05/2017	29/05/2017	05/06/2017	12/06/2017	19/06/2017	26/06/2017	03/07/2017	10/07/2017	17/07/2017	24/07/2017	31/07/2017	07/08/2017	14/08/2017	21/08/2017	28/08/2017	04/09/2017
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Start	13/03/2017	1	>																									
Security Threats	20/03/2017	2		>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X					
Machine Learning Techniques	27/03/2017	3			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X					
Testing and Evaluation Methods	03/04/2017	4				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X					
Introduction	10/04/2017	5					>	-	X																			
Background	17/04/2017	6						-	-	X																		
Methodology & Aims and Objectives	24/04/2017	7							-	X																		
Ethics & Professional Considerations	01/05/2017	8								-	X																	
Evaluation and Finalising	08/05/2017	9									*																	
Identify Security Threats	15/05/2017	10										>	-	-	-	-	-	*										
Investigate The Cause	22/05/2017	11											-	-	-	-	-	-	-	-	X							
Review The Impact	29/05/2017	12												-	-	-	-	-	-	-	-	X						
Outine The Solution	05/06/2017	13													-	-	-	-	-	-	-	-	*					
Data Collection & Pre-Processing	12/06/2017	14														>	-	X										
Extract Features	19/06/2017	15															-	-	X									
Train Machine Learning Model	26/06/2017	16																-	-	X	*							
Cross Validate The Model	03/07/2017	17																	>	-	-	X						
Test Model	10/07/2017	18																		-	-	-	X					
Evaluate Model & Methodology	17/07/2017	19																			>	-	-	X				
Evaluate Against Aims & Objectives	24/07/2017	20																				-	-	-	X			
Discuss areas for further work	31/07/2017	21																					-	-	-	X		
Review & Finalise	07/08/2017	22																						>	-	-	-	
Review & Finalise	14/08/2017	23																							-	-	-	
Review & Finalise	21/08/2017	24																								-	-	
Review & Finalise	28/08/2017	25																									X	
Submit	04/09/2017	26																										*

Key	
Start	>
In Progress	-
Finish	X
Milestone	*
Research	
Project Plan and Overview	
Analysis	
Development	
Testing	
Evaluation	

Figure 4.1: Gantt Chart

5 Ethics and Professional Considerations

This section of the report considers the ethical and professional considerations when carrying out the project. A set of considerations and their risks have been outlined.

5.1 Research

Research is an important component of this project, information from various sources will be compiled and collated together. Therefore, the source of the content should be of quality and integrity. The resources obtained should be from reliable and reputable sources in the field of Cyber/Web security. A search engine such as Google scholar should be used to help achieve some of the above. In addition, all research conducted should be independent and impartial.

5.2 Data

An objective of this project is to collect data. To achieve the aim of teaching a machine how to detect malicious and benign code, sample data of malicious (and benign) code would need to be collected. Therefore, this project will comply with the Data Protection Act of 1998 [10].

1. How will the data be stored?
    - The data collected should be stored on an encrypted laptop.
2. For how long will the data be stored for?
    - The data collected should only be stored for the duration of the project.

5.3 Software

The software or machine and the results and findings produced will not be used in a commercial environment and it will not have an effect on decision making in a live environment. The software which is going to be developed in its current state is intended for the sole purpose of research and analysis only.

Finally, after considering the ethical and professional considerations for this project as well as carrying out the online ethics decision tool provided by the University of Manchester [11]. It was concluded that this project does not require ethical approval. However, the ethical and professional consideration outlined above as well as their risks will be considered during the entire duration of the project and as the project progresses further considerations may need to be drawn.

6 Evaluation

6.1 Success Criteria

One method to test and evaluate the project is to test it against the success criteria. At the start of the project the initial aim and objectives have been defined. These are the expectations of the project, at the end of the project, we can test and evaluate whether the objectives had been met and if the initial aim of the project was achieved.

6.2 Quantitative Method - Performance Metrics

Numerical metrics can be used to evaluate the performance of a machine. When trying to accurately classify or label something as malicious or not, we can produce a confusion matrix. The confusion matrix simply tells us the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values of a prediction.

From these values we can produce additional evaluation metrics including but not limited to precision, recall and f-measure. These are popular methods for evaluating the performance of a machine learning system as simply calculating the accuracy is not enough.

- Recall - What % of data the machine simply labelled as malicious.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

- Precision - What % of data the machine labelled as malicious and were actually malicious.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- F-measure - The harmonic mean of Recall and Precision.

$$F = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

6.3 Cross Validation Framework

When testing and evaluating a predictive model such as a one trying to label if code is malicious or benign, it is often good practice to implement the cross validation framework. The idea behind this framework is to test the system using different data than that which it was trained upon. This framework helps to avoid overestimation of the machines performance. K-fold cross validation is one such implementation of this framework which could be used.

References

[1] *Cern*. 1991. URL: http://info.cern.ch/hypertext/WWW/TheProject.html.

[2] *ICT Facts and Figures 2016*. URL: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf.

[3] CISO. *CISO 2017 Annual Cybersecurity Report*. 2017. URL: http://www.cisco.com/c/dam/m/digital/1198689/Cisco%5C_2017%5C_ACR%5C_PDF.pdf.

[4] *2012 LinkedIn Security Breach*. 2016. URL: https://sentinelone.com/blogs/blast-past-2012-linkedin-breach-dumps-100m-additional-records/.

[5] *OWASP Top 10*. URL: https://www.owasp.org/index.php/Category:OWASP%5C_Top%5C_Ten%5C_Project.

[6] NCSC. *CERT-UK Annual Report*. 2016. URL: https://www.ncsc.gov.uk/report/cert-uk-annual-report-201516.

[7] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[8] Ian Fette, Norman Sadeh, and Anthony Tomasic. *Learning to detect phishing emails*. Tech. rep. DTIC Document, 2006.

[9] JeanFrancoisPuget. *What languages is best for machine learning and data science*. 2016. URL: https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Language_Is_Best_For_Machine_Learning_And_Data_Science?lang=en.

[10] Data Protection Act. "London: The Stationery Office". In: (1998). URL: http://www.legislation.gov.uk/ukpga/1998/29/contents.

[11] *Ethical Review System*. URL: http://www.staffnet.manchester.ac.uk/services/rbess/governance/ethics/new-online-system-for-ethics-review-erm/.