

Cyber Security Analysis & Using machine learning techniques to classify malicious code

Project Plan and Overview
University of Manchester

Kay Khan 7574393

01/05/2017

Contents

List of Figures	ii
List of Tables	iii
1 Introduction	1
2 Background	2
3 Methodology	3
3.1 Project Workflow	3
3.2 Tools	4
4 Plan	5
4.1 Aim & Objectives	5
4.2 Milestones	6
4.3 Gantt Chart	7
5 Ethics and Professional Considerations	8
5.1 Research	8
5.2 Data	8
5.3 Software	8
6 Evaluation	9
6.1 Success Criteria	9
6.2 Quantitative Method - Performance Metrics	9
6.3 Cross Validation Framework	9
A Appendix	10
Bibliography	11

List of Figures

3.1 Project Workflow 3

3.2 Popular languages for machine learning 4

4.1 Gantt Chart 7

List of Tables

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. 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 [3]. 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 investigate, identify and review yesterdays, todays and tomorrows security threats and their solutions. Investigate the cyber 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 such threats. In addition to the above, the purpose of this project is to also develop a system or rather teach a machine, so that it is capable of accurately classifying code as either malicious or benign, through the use of artificial intelligence specifically machine learning techniques.

The main report, it's analysis and findings will be aimed towards those who are beginners in the field of web application security and cyber security in general 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 and why this is an important topic. Section 3 describes the methodology of the project, 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 in carrying out this project. Finally, section 6, how will the project will be tested and evaluated.

2 Background

TODO:

- why, motivation - Article on popular web security threats - Linked In Case Study?
- Precedence has already been set for using machine learning in a security aspect. (Article on detecting spam emails)

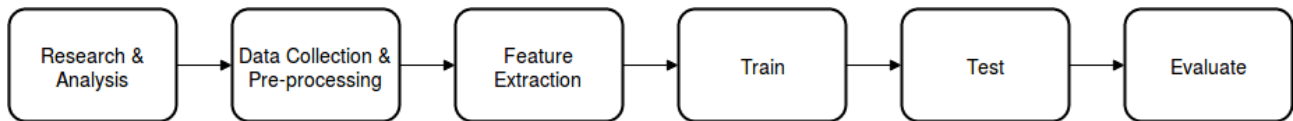
3 Methodology

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

3.1 Project Workflow

To achieve the goals of this project, there are 6 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 & Analysis** - Research and analyze web security threats, their cause, the impact they have and their solution.
2. **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.
3. **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 a greater probability of being malicious.
4. **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.
5. **Test** - Test the machine with a new set of data (code) which it hasn't seen before i.e data which it wasn't trained on. To evaluate the performance of the machine.
6. **Evaluate** - Evaluate the performance of the machine, calculating the accuracy, recall, precision and f1 results and evaluation of the methodologies used and objectives met.

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.

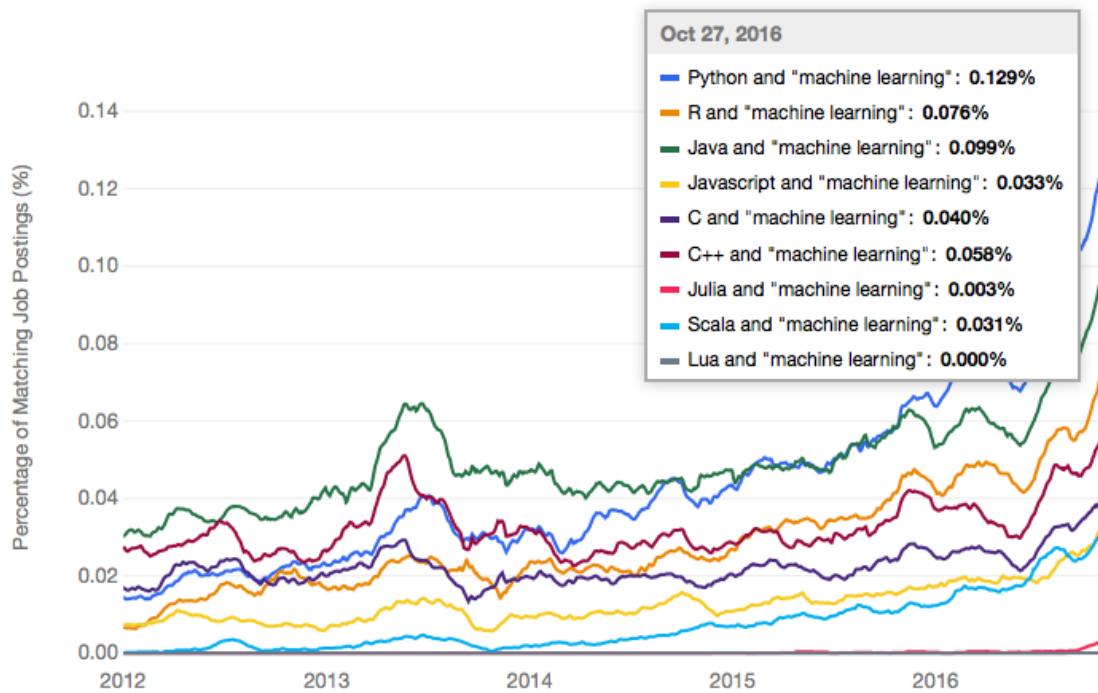


Figure 3.2: Popular languages for machine learning [4]

The language R is another popular statistical analysis programming language for machine learning as it is very friendly for beginners. However, python's 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 project plan and overview of around 2500 words, secondaly 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 the project. The objectives have been outlined below outlined below.

1. Research

- (a) Research security threats, their cause, impact and solution.
- (b) Research machine learning techniques.
- (c) Research testing and evaluation techniques.

2. Analysis

- (a) Identify security threats
- (b) Investigate the cause of the security threats
- (c) Review the impact of the security threats
- (d) Outline the solution to the security threats

3. Development

- (a) Data Collection & Pre-processing
- (b) Extract important features from the data
- (c) Train a machine (model) using the features

4. Testing

- (a) Cross-validate the model
- (b) Test model with new set of data

5. Evaluation

- (a) Evaluate the system & methodologies used.
- (b) Evaluate project against initial aims and objectives
- (c) Discuss areas for further work

4.2 Milestones

1. (12/05/2016) - Submission of the project plan and overview.
2. (26/06/2017) - Draft of the security threat analysis.
3. (17/07/2016) - Completed or working version of the system/machine.
4. (07/08/2016) - 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

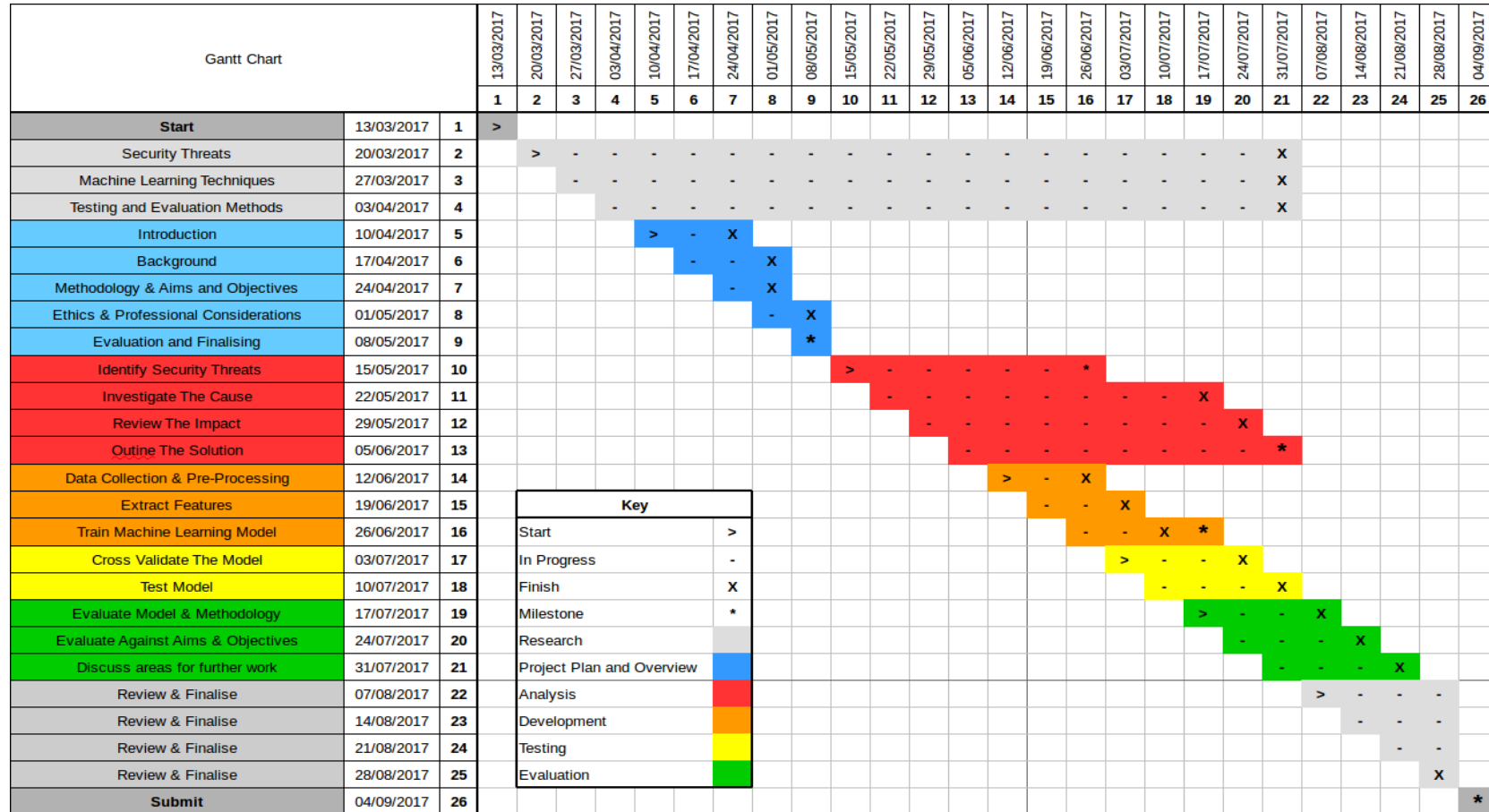


Figure 4.1: Gantt Chart

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 [5].

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 [6]. 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 objective 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

We can use numerical metrics 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.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

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

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

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

$$F = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

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.

A Appendix

Bibliography

- [1] Cern, 1991.
- [2] Ict facts and figures 2016, 2016.
- [3] 2012 linkedin security breach, 2016.
- [4] JeanFrancoisPuget. What languages is best for machine learning and data science, 2016.
- [5] Needed. Needed. *Needed*, Needed(Needed):Needed, Needed.
- [6] Ethical review system.