January 4, 2020

MICROSOFT x UW MALWARE PREDICTION PROJECT

PROJECT PLAN

PRESENTED BY:
Saruul Khasar
Alex Honeycutt
Wayne Zhang
Andy Cahill

TABLE OF CONTENTS

I. PROBLEM STATEMENT

Malware families, and malware variants inside the families, cause a major problem for computer security. Computer security is important not only because it maintains the computer's overall health, allows programs to run quicker and smoother but also it keeps the owner's information protected. Malware damage can range from loss of files to loss of security - even identity theft. Therefore, in order to prevent such damages from happening, computer companies and antivirus companies work on taking actions before it happens.

The purpose of this UW and Microsoft collaboration project is to group variants of malware samples on the basis of their characteristics, build a model to predict the likelihood of a machine that could be infected, and communicate the results using interactive dashboards. The outcomes from this project will allow the project sponsors or users to see different perspectives in analyzing malware data and hopefully utilize some findings in the database system, prediction models, and interactive visualization in their everyday operation. The executors of this project will gain industry level experience in building database systems from raw data, building machine learning models, and creating web-based interactive visualizations.

II. PROJECT CHARTER

Outline:

1. General Information
2. Project Scope
3. Resource Requirements
4. Timeline
5. Project Team Roles and Responsibilities

1. General Information

Project title: Microsoft malware prediction
Project sponsor: Kenny Kim
Project advisor: Greg Hay (gthay@uw.edu)
Project team lead: Saruul Khasar (saruul@uw.edu)

2. Project Scope

Situation/Problem/Opportunity:
The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.

Project goal:
The purpose of this UW and Microsoft collaboration project is to group variants of malware samples on the basis of their characteristics and build a model to predict the likelihood of a machine that could be infected.

Objectives/Deliverables (if known):

i) ERD in 3NF to build database system for malware datasets (SQL server)
ii) Classification model for predicting infected machines (Python)
iii) Web based interactive dashboard (Python Dash)

3. Resource requirements

i) People

- Executive sponsor
- Project advisor
- Project team lead
- Independent task leads in the project

ii) Time

The initial estimate for the project duration is approximately 10 weeks.

iii) Budget

Hardware	$0	(provided by the team members)
Software	$0	(provided by the University of Washington)
Monetary budget	$90	

iv) Timeline (preliminary)

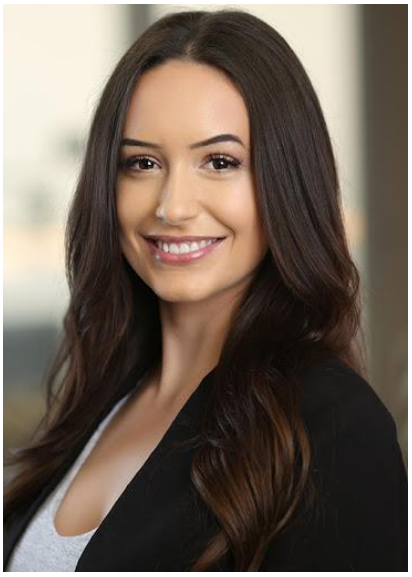
- Project documents approved: **February 14, 2020**
- Team meetings: **Tuesday and Thursday 1:30 - 3:15 pm**
- Preparing data & creating ERD: **February 14, 2020**
- Populating data: **February 21, 2020**
- Exploring data: **March 13, 2020**
- ML modelling: **March 27, 2020**
- Delivering results: **April 10, 2020**

4. Project Team Roles and Responsibilities

Team member	Roles	Responsibilities
Saruul Khasar	Team lead, Machine Learning Modelling Lead	• Develop a strategy by which team members can use to reach the project goal • Assign tasks to team members and give team members an ownership of tasks • Determine completion timeline and monitor progress to ensure project is on track • Lead Machine learning modelling and Python programming tasks • Develop web-based interactive dashboard for data analysis and data modelling • Contribute to the design of the relational database ERD • Contribute to the SQL queries for data analysis
Alex Honeycutt	Database Design Lead	• Design the relational database that will help create our model • Ensure the database is implemented correctly and the data is stored correctly • Ensure fields meet the required data types • Contribute to the data analysis on SQL • Contribute to the data visualization on interactive dashboard
Wayne Zhang	Data Visualization Lead, SQL Query Lead	• Develop web-based interactive dashboard for data analysis and data modelling • Develop SQL scripts for data analysis • Contribute to the design of the relational database ERD • Contribute to the data analysis on Python • Contribute to the design of the relational database ERD
Andy Cahill	SQL Query Lead	• Contribute to the design of the relational database ERD • Build the ERD in Lucidchart from the design and data types in train.csv • Import the ERD from Lucidchart, creating the tables and relationships • Develop SQL queries to help determine which combination of columns predicts the presence of malware • Contribute to building data visualizations from our findings

III. PROJECT CHARTER | USER STORIES

Name: Allison Tech
Market Segment: IT in Business
Gender: Female
Age: 32 years old
Education: Bachelor's degree (Systems Management)
Occupation/income: Systems Administrator
Family situation: Married
Location: Austin, TX
Hobbies: Gym-workouts, reading, hackathons, volunteering with kids



Motivation and goal:

While watching the news, Allison sees another segment that shows malware attacks are on the rise. Having experience with security and networks, she wants to make her systems safer and protect internal documents as well as customer privacy. She wants to know why most of these machines are infected in order to find the best way to protect her company.

Challenges:

1. Find a common link between machines that can help her anticipate whether a machine will be infected or not
2. Send warnings out to customers and employees advising them to not have a specific set of settings, or at least have an anti-virus active

How this project helped them achieve their goal:

Allison will be provided with a model which she can use to determine which machines are more prone to malware. With this, she can proactively anticipate attacks and create better safeguards for them. Safeguards could include warning customers and employees about potentially compromising configurations or informing suppliers about what parts are more prone to being infected. Instead of waiting for the machine to become infected and reacting to it, we want her to be able to run it through our model and know whether that machine has a higher likelihood of being infected or not.

Name: Nora Lum

Market Segment: IT in Business
Gender: Female
Age: 30 years old
Education: Bachelor's degree (Business Administration)
Occupation/income: Data Analyst
Family situation: Married
Location: Sunnyvale, CA
Hobbies: Hiking, swimming

Motivation and goal:

Nora is a data analyst in a Tech firm and she is focused on building prediction models. She is currently working on a project to predict the probability of machine's likelihood of getting infected. She needs a different perspective on modelling this large amount of data. She also wants to know what methods are used in this area and how were their success rate was. She is interested in sponsoring this project.

Challenges:

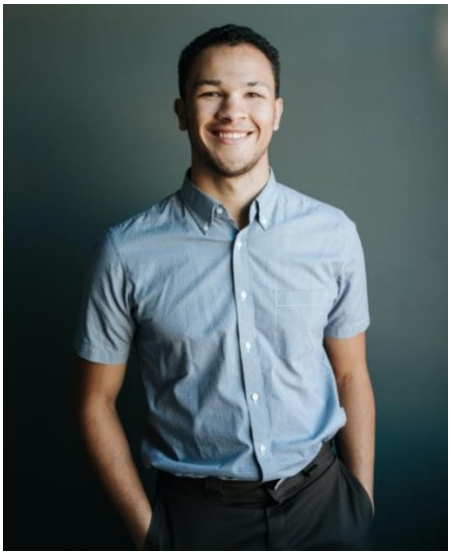
1. She is using decision tree model to predict the machine's likelihood of getting infected, but she doesn't know if this is a good method for this data.
2. She is struggling to run this model on Python because of large datasets and she is interested in other methods to make the data more runnable.

How this project helped them achieve their goal:

Nora will get the overview of materials on the internet regarding this problem. This project will provide her the literature review on this problem, so she can learn from other people success and failure stories. She doesn't know if she should invest more on the decision tree model and this project will tell her if that method is appropriate or not. And what other modelling methods are available.

Name: Ryan Lee

Market Segment: Consulting
Gender: Male
Age: 22 years old
Education: Master's degree (Business Administration)
Occupation/income: Associative Researcher
Family situation: Single
Location: Seattle, WA
Hobbies: Meeting new people



Motivation and goal:

Ryan is a recent graduate hired by a consulting firm. He is currently working on his customer's request to research the effectiveness of anti-virus products and the potential demand in the market. Therefore, he needs to evaluate the spread of risks among Microsoft and Apple laptops, define what their likelihood of getting infected when there is an antivirus product. And how widespread the use of antivirus product is.

Challenges:

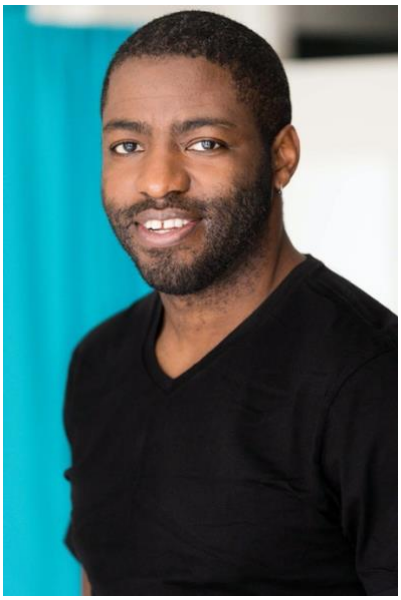
1. Needs to determine the likelihood of a machine to get infected when there is no antivirus product versus when there is an antivirus product.
2. Needs to determine how widespread the use of antivirus product but don't know how to analyze this large dataset.

How this project helped them achieve their goal:

Ryan will be provided an interactive data visualization where he can see the analysis to get answers to his questions. He will be able to see some descriptive statistics on the usages of antivirus product on Microsoft machines; however, he will provide an information about Apple machines.

Name: Steven Maxwell

Market Segment: Digital/Software
Gender: Male
Age: 29 years old
Education: MS in Human Resource Management
Occupation: HR Manager
Location: Seattle, WA
Hobbies: Skiing, hitting the gym, cooking, yoga



Motivation and Goal:

His laptop recently started to get alerts from Windows Defender of potential threats. Given the important corporate files on his laptop, he wished to get rid of the danger, but he couldn't achieve it without a proper analysis on his system, since he did not know much about how to manage the security of his laptop with confidence.

Challenge:

1. Find the potential causes of threat alarm and eliminate them;
2. Re-adjust security settings of the machine to inject highest possible security measures;

How this project helped him achieve his goals:

This project provides Steven a fundamental and comprehensive analysis of machine (device) parameters and settings that allowed easier malware targeting. He used the dashboard we provided as the outcome of the analysis to run a prediction of his machine and, got a full report of factors making his laptop prone to attack. He changed the settings accordingly and could concentrate on his work without worrying about compromising corporate data.

IV. PROJECT PLAN | GANTT CHART

		Lead	Week 1		Week 2		Week 3		Week 4		Week 5		Week 6		Week 7		Week 8		Week 9		Week 10	
			M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
I. Preparing data & creating ERD	1. Obtain .csv file on hay08	Everyone																				
	2. Create ERD	Alex																				
	3. Fine-tuning the ERD	Alex																				
	4. Import data with Python	Saruul																				
	5. Creating data dictionary	Saruul & Alex																				
II. Populating data	1. Import data with SQL	Wayne																				
	2. Create new tables on SQL	Andy																				
	3. Test population code	Wayne																				
	4. Populating data on ERD	Wayne																				
III. Exploring data	1. Find patterns in data on SQL	Andy																				
	2. Find patterns in data on Python	Saruul																				
	3. Create preliminary visualizations	Wayne																				
IV. ML modelling	1. Conducting literature review	Saruul																				
	2. Creating literature review summary	Saruul																				
	3. Choosing ML models	Saruul																				
	4. Train & predict data	Saruul																				
V. Delivering the results	1. Visualization on patterns	Wayne																				
	2. Visualization on prediction	Wayne & Saruul																				
	3. Preparing PPT	Everyone																				
VI. Documenting the procedure	1. Writing meeting minutes	Take turns																				
	2. Project documentations	Saruul																				

	- SQL database related tasks
	- ML model related tasks
	- Overall project related tasks