WALMART DATA REPORT

APRIL 2020

Prepared by Randi King, Anabel Crawford, Michael Phantja, Riley Pinion

OPEN SOURCE INTELLIGENCE BACKGROUND

VALUE

Open Source Intelligence (OSINT) plays a critical role in developing Cyber Threat Intelligence. With the rise of the internet, this also led to a rise in a wealth of information. In addition to OSINT being free, another benefit is that organizations can now use this information and data to go on the offensive and find out what threat actors are up to. For a global company such as Walmart with millions of consumers, the value of OSINT cannot be understated. For this milestone, the team utilized various data sources to collect data and create models to help Walmart identify, detect, and protect against the numerous threats and threat actors that could wreak havoc on their organization.

ACQUISITION

OSINT is generated by a myriad of sources throughout the internet. In fact, much of this data can be generated from the threat actors themselves. The team utilized various traditional sources of OSINT that we learned about in class to collect data for this milestone ranging from social media to cyber intelligence feeds.

Twitter was selected as our first data source. Often, threat actors will actually post their motives or targets on social media sites. Social media can even be a vital step in a Cyber Kill Chain, helping organizations collect certain information through reconnaissance. Social media can aid organizations or governments in finding and interacting with threat actor communities.

RETAIL INDUSTRY

DATA SOURCES

Organizations throughout the industry, like Amazon, Costco, Home Depot, Best Buy and Target are going to facing similar threats to Walmart. All of these organizations are in possession of enormous amounts of consumer PII, as well as having the responsibility for hundreds of thousands of network devices. Therefore, it is likely that many of Walmart's competitors will utilize the same data sources. According to Gartner, "Larger organizations have, for the most part, already invested in various flavors of threat intelligence." A company can decide to purchase services from multiple vendors. It is important to note that often, vendors consume similar information. In fact, many vendors will even share intelligence content. For example, our team used Maltego to gather and integrate data. Maltego has a built-in market place with over 30 data providers ranging from CrowdStrike, Ciphertrace, ShadowDragon, dataprovider.com, and DomainTools, among others. Many of these individual vendors have relationships with Walmart's competitors. For example, CrowdStrike utilizes AWS.

SOCIAL MEDIA

Social media is a vital data source across the industry. For example, Walmart has a presence across social media, including Facebook and Twitter. In fact, Walmart has been operating its Twitter account since 2008, and it currently has 1 million followers. Walmart's Facebook page boasts a much larger following with almost 33 million followers. As these numbers demonstrate, these sites are used to update and interact with customers across the world. Social media can also be used to learn more about their customers. In today's digital marketing age, it is very common, and even expected, for organizations across the industry to have these social media platforms. For example, Walmart's largest competitor, Amazon, also has around 29 million followers on their Facebook page.

OSINT SOURCES WALMART

The next step was to decide what specific OSINT sources we would collect.

1 / MALTEGO

OSINT software to discover connections

2 / WEB SCRAPING FACEBOOK

Web scraping comments on Walmart's Facebook posts

3 / WEB SCRAPING TWEETS

Web scraped tweets related to Walmart on Twitter, Sentiment analysis, Word Cloud

4 / KAGGLE

World's largest data science community

6/ ALTERNATIVE SOURCES

Alternative approaches considered

MALTEGO

MALTEGO IS AN INTERACTIVE, VISUAL DATA MINING AND LINK ANALYSIS TOOL USED TO CONDUCT ONLINE INVESTIGATIONS THROUGH A LIBRARY OF PLUGINS CALLED "TRANSFORMS."

WHAT VALUE DOES IT PROVIDE FOR THE RETAIL INDUSTRY/WALMART?

Any company, especially Walmart, that is serious about cyber OSINT should be familiar with Maltego. Maltego is a freely available easy to use link analysis and visualization product from Paterva. Maltego can provide key insight into websites, can map a networks infrastructure, or research social network graph information. Penetration testers, social engineers, and threat intel researchers use Maltego as an indispensable tool.

WHO GENERATES THE DATA?

Maltego has a unique feature which is the Transform Hub. The hub is a data marketplace built into the desktop client. From the Transform hub, a user can connect data from a variety of public sources(OSINT) as well as connect to their own data. The hub connects to data automatically, and with just one click you can have code installed and run. There are around 30 sources in the transform hub as of March 2020.

WHY DID WE SELECT THIS DATA SOURCE?

We selected this data source for its usability. Maltego automates the process which is useful for a company as large as Walmart. We also chose this source for its link analysis, for its aid in finding relationships between pieces of information from various sources on the internet. Maltego can also be easily adapted to our own unique requirements.

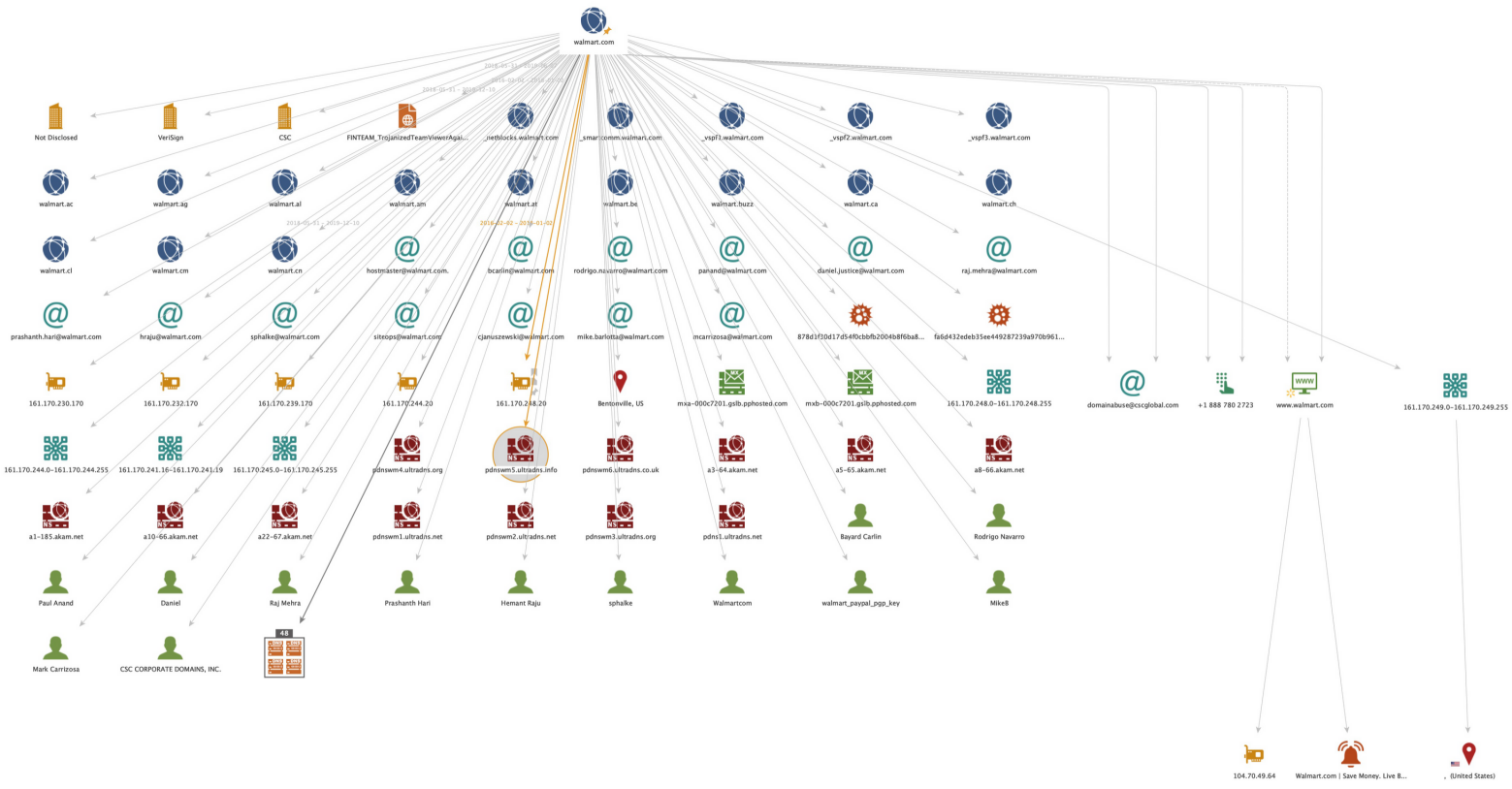
MALTEGO

HOW WAS THE DATA COLLECTED? WHAT APPROACHES WERE USED TO COLLECT IT?

For our collection, we began by downloading the free Maltego Community edition. From there, we created a new graph. To gather the data, the first action was to create a domain entity. We changed the name to “walmart.com.” Since this was the free community edition, transforms were limited. We selected “All transforms” to gather information from Maltego, threatminer, and Paterva. After the transforms, we were able to see a map of all the entities. These approaches were taken due to the constraints of the free version. We attempted to gather as much information as possible with the resources of Matlgo, Threatminer, and Paterva.

THE PROCESS OF CREATING THIS VISUALIZATION

Maltego has the benefit of automatically creating visualization. However, the visualization outputted was far too complex to display all entities.



TOP 10 ENTITIES

Total number of entities	131 (84 nodes)
Total number of links	134 (85 edges)

Ranked by Incoming Links

Rank	Type	Value	Incoming links
1	Email Address	domainabuse@cscglobal.com	2
2	Website	www.walmart.com	2
3	Phone Number	+1 888 780 2723	2
4	DNS Name	blog.walmart.com	2
5	DNS Name	gateway.walmart.com	2
6	DNS Name	time.walmart.com	1
7	DNS Name	ww.walmart.com	1
8	DNS Name	wireless.walmart.com	1
9	Email Address	prashanth.hari@walmart.com	1
10	Email Address	raj.mehra@walmart.com	1

Ranked by Outgoing Links

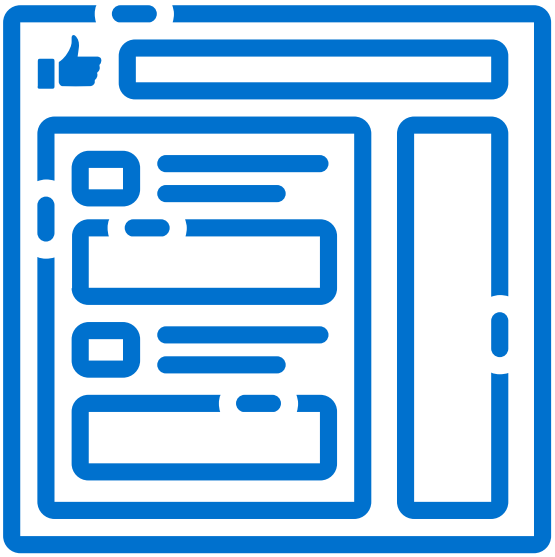
Rank	Type	Value	Outgoing links
1	Domain	walmart.com	131
2	Website	www.walmart.com	2
3	Netblock	161.170.249.0-161.170.249.255	1
4	Email Address	domainabuse@cscglobal.com	0
5	Phone Number	+1 888 780 2723	0
6	DNS Name	blog.walmart.com	0
7	DNS Name	gateway.walmart.com	0
8	DNS Name	time.walmart.com	0
9	DNS Name	ww.walmart.com	0
10	DNS Name	wireless.walmart.com	0

Ranked by Total Links

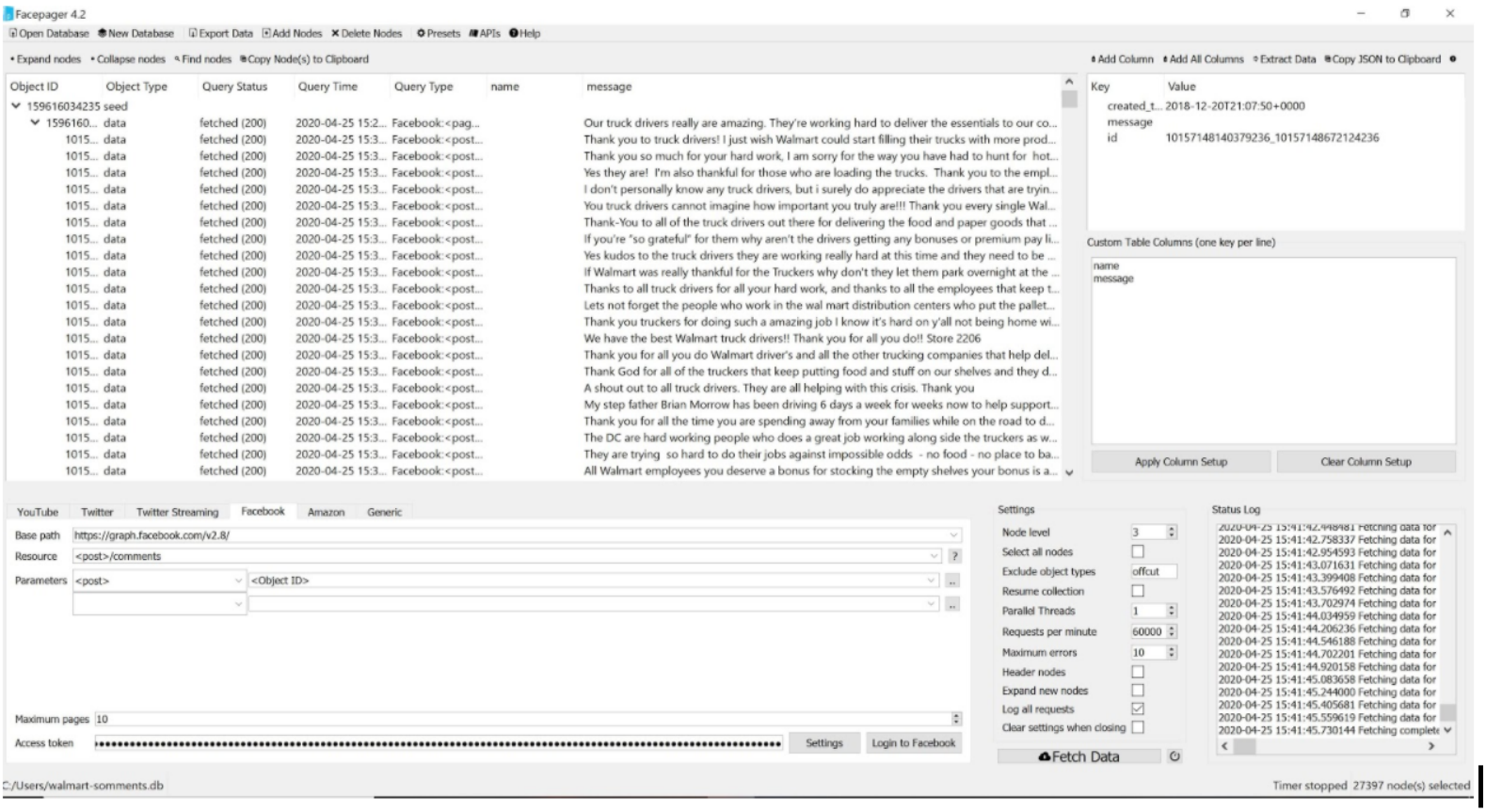
Rank	Type	Value	Total links
1	Domain	walmart.com	131
2	Website	www.walmart.com	4
3	Netblock	161.170.249.0-161.170.249.255	2
4	Email Address	domainabuse@cscglobal.com	2
5	Phone Number	+1 888 780 2723	2
6	DNS Name	blog.walmart.com	2
7	DNS Name	gateway.walmart.com	2
8	DNS Name	time.walmart.com	1
9	DNS Name	ww.walmart.com	1
10	DNS Name	wireless.walmart.com	1

WEB SCRAPING FACEBOOK

Traditional social media sources can add significant value to CTI due to their widespread usage, and many hacker groups utilize such outlets to hint at their next targets or share their exploits. This is why we scraped data regarding comments on Walmart Facebook posts. Walmart's Facebook page has over 32 million followers and over 34 million likes. The company creates a new Facebook post almost every day and has amassed over 8,000 timeline photos. These posts get an abundance of comments. We tried but were unable to scrape data from other areas of Facebook successfully. We collected comment data by using Facepager 4.2. On April 25th, we scraped data on the last 10 pages of Wamart posts and 15 pages of comments data on each of those posts. Overall, we collected a little over 64,000 observations. This data is located in a csv file titled "walmart-post-comments.csv".



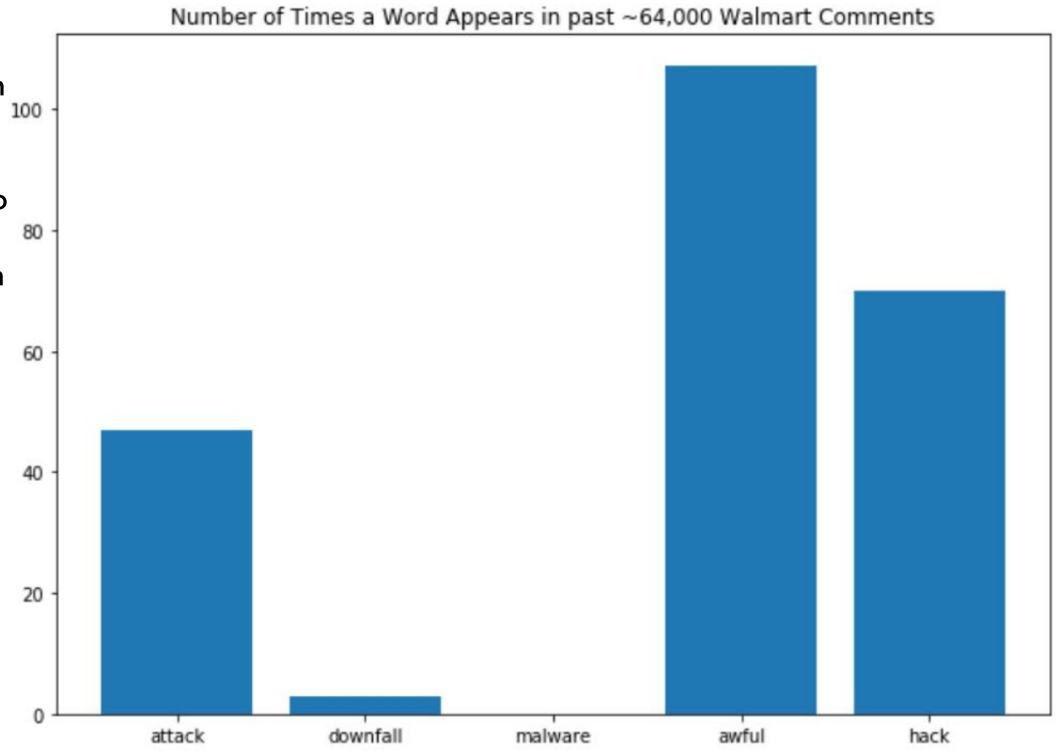
Below is an image showing the Facepager 4.2 screen after we collected our data.

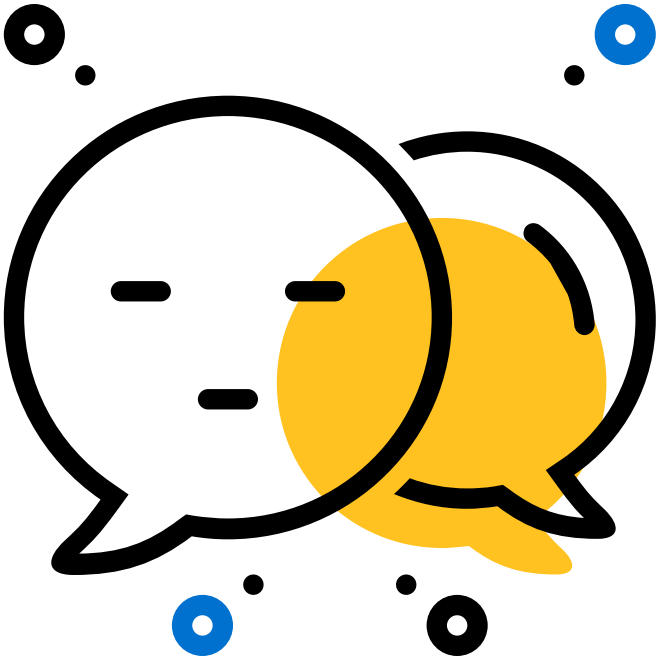


WEB SCRAPING FACEBOOK

FACEPAGER CONTINUED

From these thousands of comments on Walmart Facebook posts, we searched for how many times certain "sketchy" words such as malware, awful, downfall, attack, and hack were used. We created a bar graph to show the number of times a comment used one of these words in it as a visualization. We have attached the python file used to process the data and create visualization as a file named CTI.csv. To run this file, you would need to put the folder we turned in, CTI Milestone 2 (Final), directly into your google drive under the section "My Drive." If not, you will have to adjust the base path to point towards where you are storing the two datasets included in this folder in order to run properly.

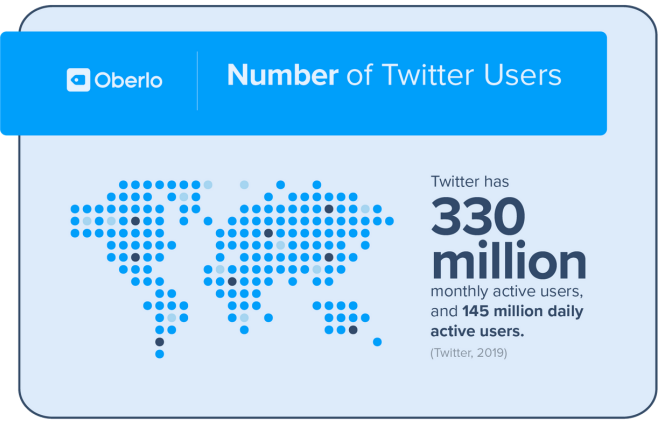




Tracking the number of times a suspicious word is used that might indicate a possible threat would allow the company to further investigate the individual comment and the user that posted the comment. As you can see, some words normally occur more often. The word "awful" can be used in many circumstances. The words "downfall" and "malware" are much more rare, so a spike in occurences in these words may indicate suspicious activity. If suspicious words are plotted often, any stray from normal plots can indicate malicious activity.

WEB SCRAPING TWITTER

TWITTER IS A WIDE PLATFORM THAT ENABLES ALL USERS TO POST THEIR THOUGHTS ABOUT ANYTHING. IT HAS OVER 330 MILLION USERS AND 145 MILLION ACTIVE DAILY USERS. THERE IS A LOT OF INFORMATION AND CONTENT OUT ON TWITTER AT ALMOST 500 MILLION POSTS IN A DAY GENERATED BY THE PUBLIC. THIS MAKES TWITTER A GREAT TRADITIONAL OSIT DATA SOURCE.



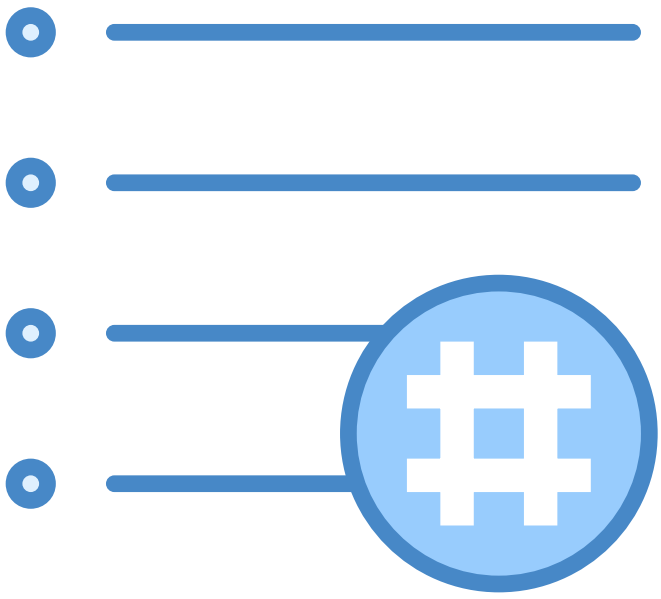
WHAT VALUE DOES IT PROVIDE FOR THE RETAIL INDUSTRY/WALMART?

Regarding Walmart, the largest retail store in America, there is a good amount of information on Twitter on personal opinions towards Walmart, whether positive or negative. These opinions can turn into negative sentiment towards an organization leading to an increase in disgruntled people to potentially cause harm to Walmart. Checking this sentiment and high used words can help Walmart detect future threats and threat periods. This data was primarily selected to identify the public's opinion and threats to Walmart and the whole retail industry as a whole..

WHAT APPROACH?

Common groups that try to analyze Twitter data take the approach of sentiment analysis. These groups search for everything, from Donald Trump to cats and ping the most recent tweets to analyze current sentiment.

https://hackernoon.com/twitter-scraping-text-mining-and-sentiment-analysis-using-python-b95e792a4d64



Following this, the team decided to run a sentiment analysis of our own on Twitter related to Walmart. This type of analysis allows for an organization to be proactive. It can be used to monitor social media sites and categorize certain statements or posts from a user by urgency.

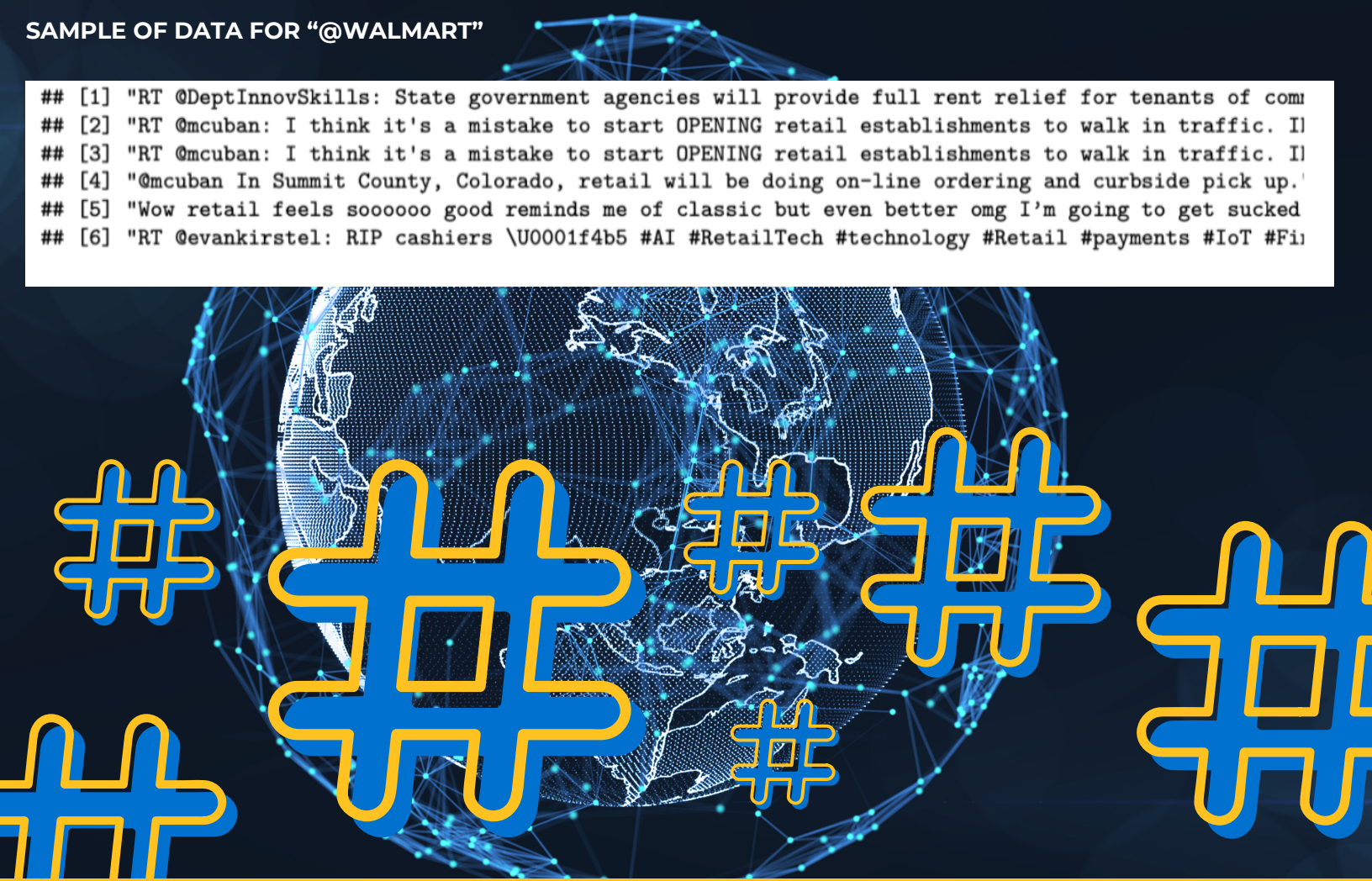
WEB SCRAPING TWITTER

COLLECTION STRATEGIES

This data was collected through twitter's API in R. For our text mining and data scraping, the API key that we held only allowed for the return of 1500 tweets at a time. The keywords that we used to get the data were "@walmart", "walmart + threat", and "retail". We tried approaching other word groups such as "walmart + malware", "walmart + cyber", and "walmart + breach". These attempts yielded a very small number of data points(<10). We ultimately settled on the 3 key words and phrases because of data content, reliability and relatability. We were able to collect the full 1500 for "@walmart" and "retail" and 1225 posts for "walmart + threat". This was perfect in forming a sentiment analysis on walmart since these were things that related to walmart. The data pinged us back a list of related tweets to our key words in long string format. There was no other metadata. We then were able to run our sentiment analysis using the R packages, sentimentr and syuzhet. Lastly, we created word clouds for each analysis to identify words. In the creation of this, we removed stop words, punctuation and created a matrix of the top 20 most used words. The code and the process is attached to the appendix.

SAMPLE OF DATA FOR "@WALMART"

```
## [1] "RT @DeptInnovSkills: State government agencies will provide full rent relief for tenants of comm
## [2] "RT @mcuban: I think it's a mistake to start OPENING retail establishments to walk in traffic. I
## [3] "RT @mcuban: I think it's a mistake to start OPENING retail establishments to walk in traffic. I
## [4] "@mcuban In Summit County, Colorado, retail will be doing on-line ordering and curbside pick up.
## [5] "Wow retail feels soooooo good reminds me of classic but even better omg I'm going to get sucked
## [6] "RT @evankirstel: RIP cashiers \U0001f4b5 #AI #RetailTech #technology #Retail #payments #IoT #Fi
```

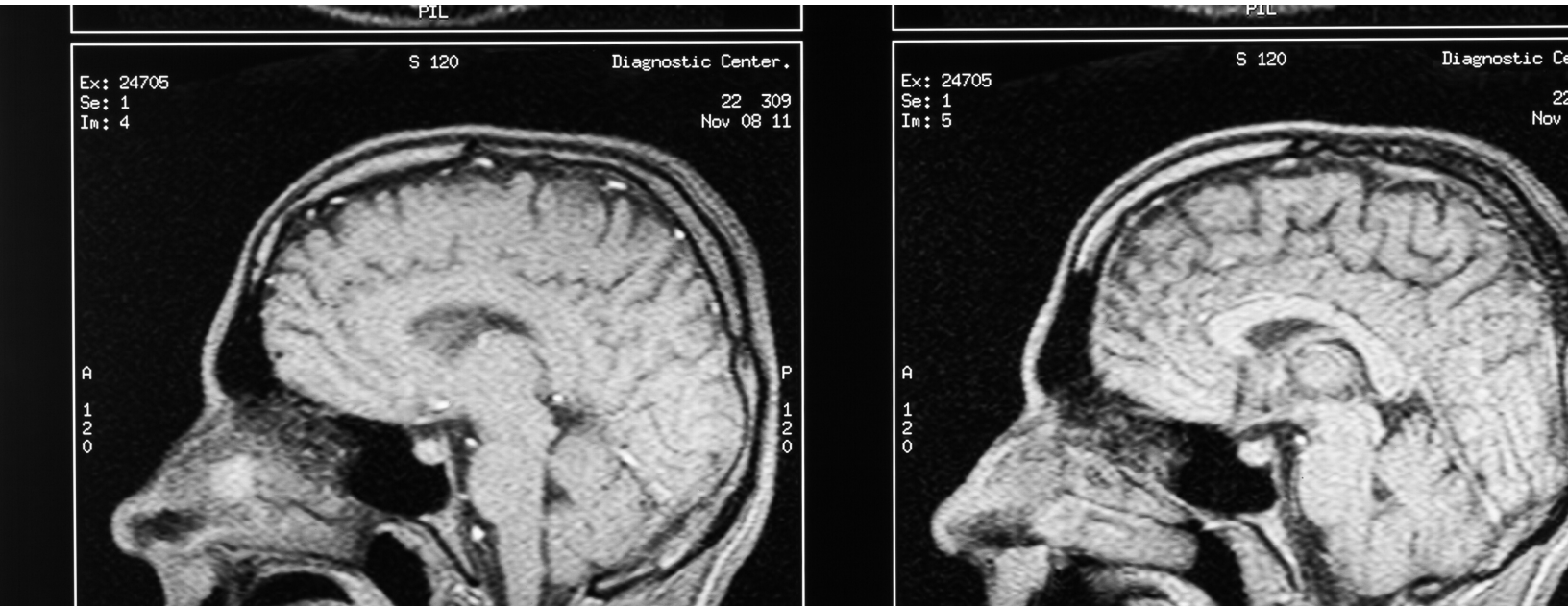


SENTIMENT ANALYSIS

SENTIMENT ANALYSIS USES A RANGE OF MODELS AND ALGORITHMS TO ANALYZE DATA. THE NAÏVE BAYES ALGORITHM, LINEAR REGRESSION, SUPPORT VECTOR MACHINES, AND DEEP LEARNING ARE ALSO COMBINED IN THIS PROCESS.

With our sentiment analysis, we were able to conclude on a number of different topics. For the "@walmart", the analysis gave us our highest mean of 0.03699895. This was slightly positive above relative neutral. The general public at twitter wasn't outwardly negative towards them, nor they were outwardly positive. This puts Walmart in a situation that they can be flexible in their cyber model. For the "walmart + threat" model, the analysis gave us a mean of -0.1958167. This was slightly below negative since the worst is -10. This concludes that Walmart should be slightly worried about cyber threats but these threats are primarily associated with the coming of the Corna Virus. This is still something that Walmart should look to secure.

Lastly, our last sentiment analysis of "retail" yielded an almost neutral sentiment of -0.01023695. We wanted to approach this analysis towards the whole industry as a whole since Walmart is still a part of the retail industry. Considering Walmart had a higher mean than "retail", we were able to conclude with this that Walmart was better off in public opinion than other retailers and they were not to be cyber attacks first go to for retailers. With this sentiment analysis, it gives Walmart the edge to be proactive and current. They are able to monitor social media and detect possible problems and threats while it's still out there.

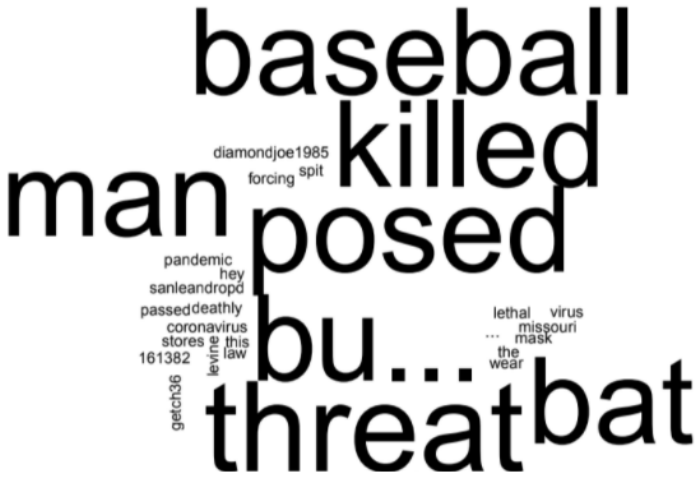


WORD CLOUDS

After our sentiment analysis, we were able to create a few word clouds. During the creation of this, we removed stop words, punctuation and created a matrix of the top 20 most used words. This brings value in that we can have our ear against the wall and listen to current things that can potentially be a threat. In retail, the public is talking about “business”, “mistakes", “manufacturing” and “traffic”. It could mean various things but this releases common ideas that are going around in the walmart and retail community. When combined with current events, the retail industry wants to be opened up by the public, potentially causing a lot of issues with the public and the retail industry. When looking at “Walmart + threat” some of the most counted words were “man”, “kill”,”bat”, relating to the topic of these things. Lastly when looking at “@walmart ", the public is talking about the “politicaltwtwar”, “workplace”, “safe”, concerning people about the current environment in the world. This identifies key issues in place for Walmart and Retail in general and where the most cyber protection would be required. For example, since retail is at a low sentiment, and manufacturing and mentioned a good amount of time, there is good reason to secure the cyber assets related to manufacturing.

"RETAIL"

“WALMART + THREAT”



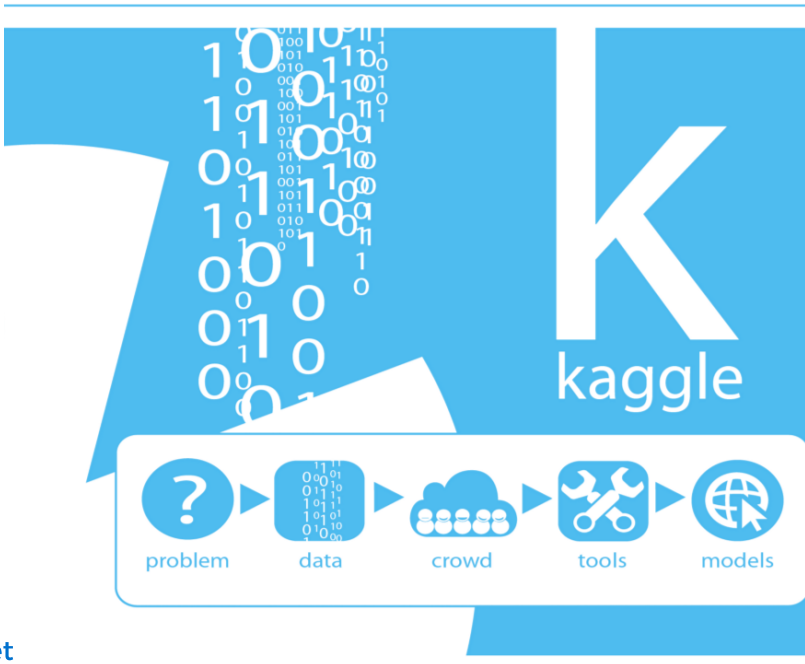
“@WALMART”

Other steps we were considering were going to track each word count and see if anything malicious pops up but we were not able to receive specific comments about cyber related threats since the data set was limited to 1500. Another thing to consider was, the analysis we ran on Facebook was similar and we didn’t want to release the same analysis. Hence, Running a sentiment analysis and word cloud of top 20 to generate a more holistic point of view of the public towards Walmart.

KAGGLE

KAGGLE IS A LARGE DATA SCIENCE COMMUNITY THAT HAS NUMEROUS LARGE DATASETS THAT CAN BE USED FOR DATA SCIENCE RESEARCH AND TO CREATE MODELS. KAGGLE IS A WEBSITE THAT IS OPEN TO ANYONE TO VIEW AND CONTAINS THOUSANDS OF LARGE DATASETS.

KAGGLE PHISHING URLS



While we used Kaggle information to create our model, most modern companies have their own more reliable methods of collecting URL information.There has been heavy efforts by top tech companies such as Google and Facebook to build a platform that works all together for one cause of preventing users from visiting malicious URLs. Traditional methods such as blacklisting may not be as effective as machine learning for preventing users from visiting malicious URLs. Phishing attacks are one of the most common attacks and being able to detect phishing URLs and block them could greatly decrease the chances of a phishing attack.

The link to the Kaggle dataset used in our study is:

https://www.kaggle.com/murataltay3504/phishing
https://www.kaggle.com/akashkr/phishing-website-dataset

ANALYTIC APPROACH - CLASSIFICATION:

This dataset contains 32 attributes and 11,055 observations. This dataset is used for classification. A classification of 1 indicates a malicious URL while a classification of o indicates a benign URL. This dataset allows us to use 31 attributes to create a model to correctly label a URL as malicious or benign. We trained the data using 66% of observations in the training set and tested our model on the remaining 33% to see how accurate our results were. We converted the categorical attributes to use 0 for false and 1 for true. We had to use the get_dummies method on attributes that had more than two options. We also dropped any NaN values. We used a machine learning logistic regression model as well as a random forest classifier model in order to predict whether the URL was malicious. We have attached the python code for the machine learning models in a file named CTI2.csv. To run this file, you would need to put the folder we turned in, CTI Milestone 2 (Final), directly into your google drive under the section “My Drive.” If not, you will have to adjust the base path to point towards where you are storing the two datasets included in this folder in order to run properly.

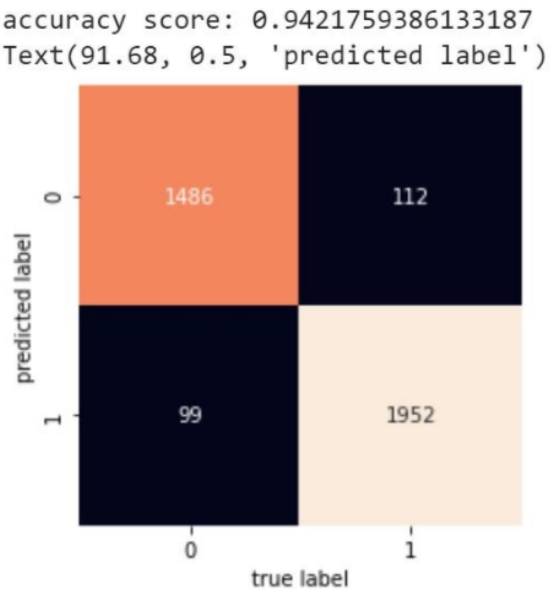
Classification of URLs as malicious or bening is important to cybersecurity as phishing, spam, and drive-by-downloads are extremely common and cost businesses billions of dollars every year. One way to reduce the costs of these threats is to apply filtering methods to emails so that less phishing emails are able to get through to the user. By developing an accurate algorithm to filter out these malicious URLs, companies can reduce the number of times employees visit a malicious URL. Classification is an appropriate approach when determining what label would apply to s set of inputs. We used a random forest approach by using python and packages from sci kit learn. Random forest models are simple to implement but have a strong performance. Logistic regression was also used because it is easy to implement and trains data very efficiently. While we got this information from Kaggle, many other companies can collect similar information through internal intelligence, commercial data feeds, or from intelligence collected by other companies within the industry.

KAGGLE

THE LOGISTIC REGRESSION MODEL VISUALIZATIONS:

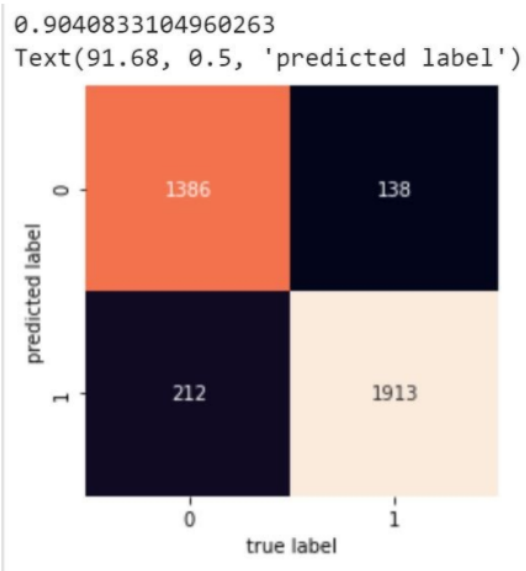
Below are some summary statistics of the logistic regression model.

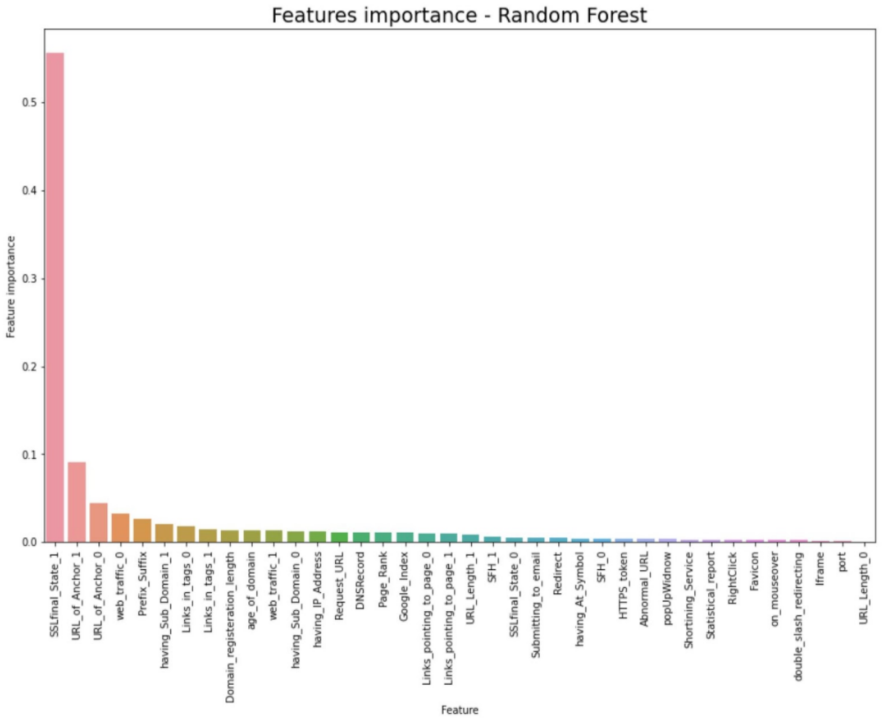
```
The error rate is 0.057824061386668128
The precision score is 0.9457364341085271
The recall score is 0.9517308629936616
The F1 score is 0.9487241798298907
```



The figure to the left shows the models summary statistics as well as a visualization of the confusion matrix. The confusion matrix shows the model had 1,486 true negatives, 1,952 true positives, 112 false negatives, and 99 false positives. The accuracy of the model is 0.9421 indicating that the model is quite useful for predicting the classification of a URL.

THE RANDOMFORESTCLASSIFIER MODEL VISUALIZATIONS:





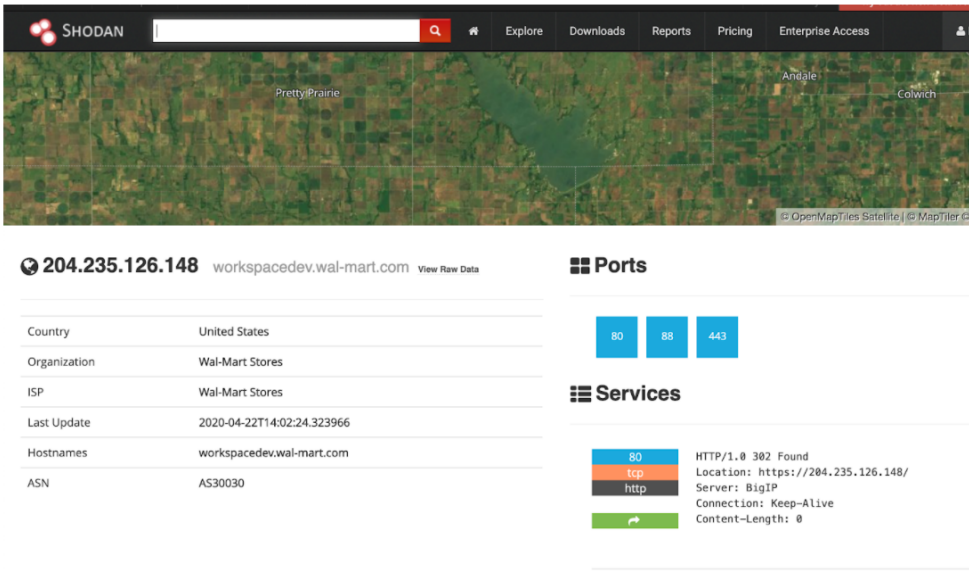
The figures above show the model's confusion matrix as well as a visualization of the feature importance. The confusion matrix shows the model had 1,386 true negatives, 1,913 true positives, 138 false negatives, and 212 false positives. The accuracy of the model is 0.9041, indicating that the random forest model is not as useful for predicting the classification of a URL. The second figure shows a plot of the feature importance in the random forest model for classification. The most important feature by far for predicting if a URL is malicious is SSL_final_State_1.

ALTERNATIVE SOURCES

VARIOUS APPROACHES WERE CONSIDERED THROUGHOUT THIS PROJECT.

SHODAN

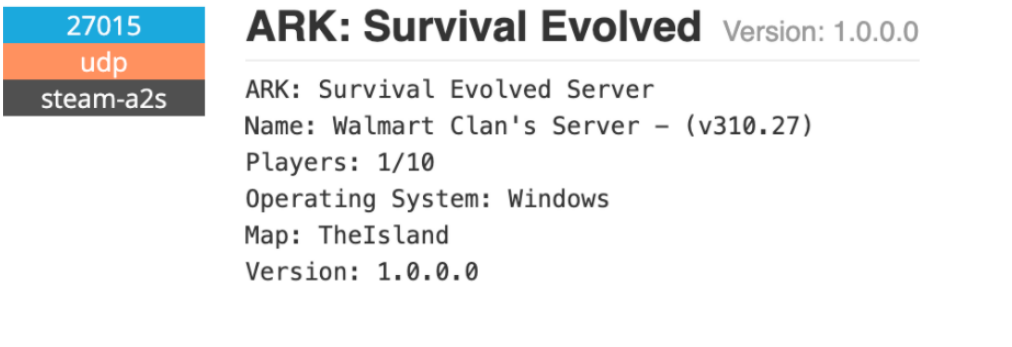
In the beginning, the team attempted to use Shodan to help identify any unprotected devices that would be susceptible to threats. After spending some time on the search engine, the team did find some interesting data. However, it was not enough to perform an in-depth threat analysis, and we wanted to make sure that we were using extensive data sets. One of Shodan's features is its ability to scan for a wide range of ports. Through some preliminary searching, we discovered Ports 80, 88, and 443 on the workspacedev-wal-mart.com host. This information is depicted in the screenshot below. While these ports are susceptible to attacks and should be moved behind a firewall, it is also important to note that Walmart currently has over 500,000 network devices. Walmart must constantly work to make sure that all of these devices are protected. The team also attempted to manually search various IP addresses found through Maltego on Shodan, but it was unsuccessful in bringing back any further data.



LIMESTONE NETWORKS

Limestone Networks is a cloud and server provider. Something interesting that we would also like to note from our time spent on Shodan is that we found a Walmart server from the unprotected Limestone port, 27015. This is important when considering outsourcing certain business operations. Your cyber assets in the hands of a third party vendor are also susceptible to threats. This port can be seen in the screenshot place below:

Pastebin was another traditional OSINT source that the team considered using to collect data. However, we quickly encountered some issues. Recently, Pastebin removed the search bar functionality. This made it much more difficult and time consuming to work with.
The team also considered utilizing RapidMiner for analytics and also for visualizations.



WORKS CITED

https://aws.amazon.com/solutions/case-studies/crowdstrike/

https://monkeylearn.com/sentiment-analysis/

https://www.shodan.io/

https://www.maltego.com/

https://medium.com/@raebaker/a-beginners-guide-to-osint-investigation-with-maltego-6b195f7245cc