MALICIOUS URL STRUCTURES & PREDICTIVE MODEL DEVELOPMENT FOR MALICIOUS URL IDENTIFICATION

Analysis Report

October 2023

Claire Lawrence
LinkedIn: https://www.linkedin.com/in/claire-lawrence-senior-analyst/

Key Findings

- The Random Forest Classifier model performed the best in terms of prediction (93% accuracy).

- Phishing and benign URLs were harder for the model to classify; this may be due to errors within the original data set or the URL types may need further close inspection to understand why this may be.

- Malware URLs showed more signs of being obfuscated than the other malicious types (i.e. characters are 'escaped' using % to hide their true details).

- Phishing URLs frequently used battle.net and us.battle – these are type-squatted versions of the genuine site us.battle.net, in order to appear authentic.

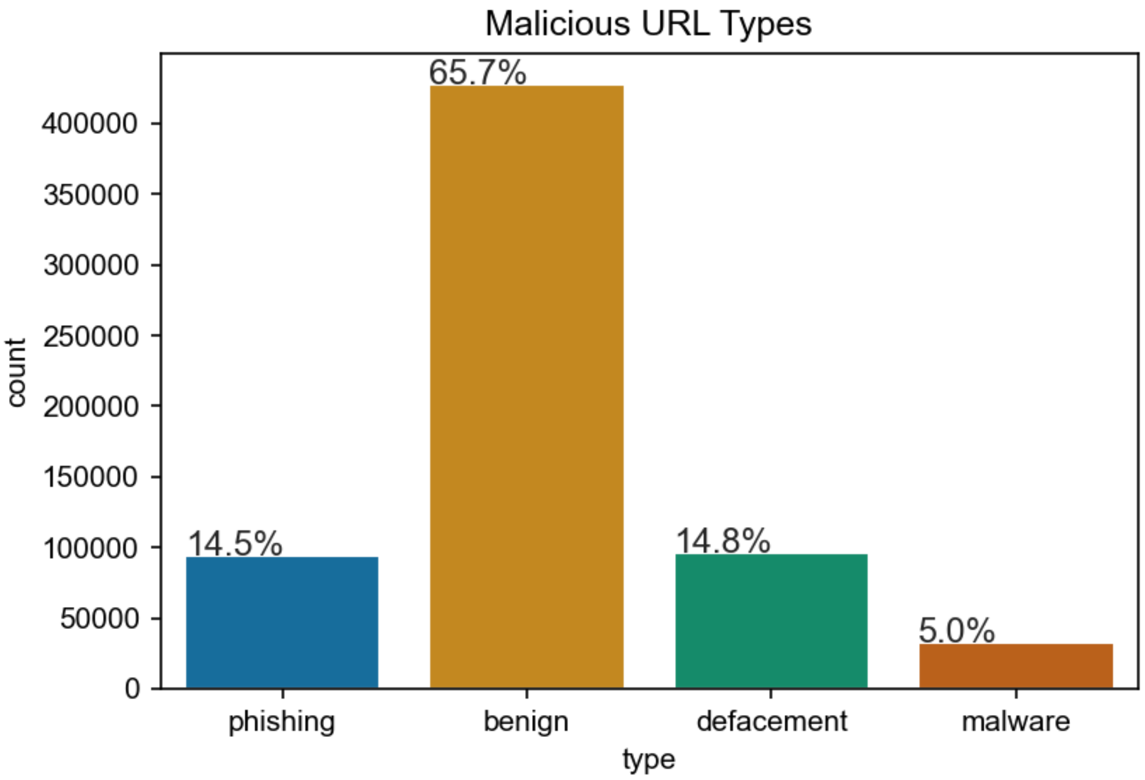
- Phishing URLs frequently try to appear as though they are from educational institutions (particularly from the United States) in order to appear more authentic.

Introduction

Project Purpose

A data set comprising of 6,51,191 URLs has been analysed[1]. The original data contained the URL and its respective classification (benign, defacement, malware or phishing). The split between these categories within the data is shown in the chart below. The aim was to analyse the components of the URLs to get a better understanding of the features of each malicious type. This information was then used to build machine learning models to predict which category an URL belongs to, consequently allowing the automatic identification of malicious URLs and their category.

Three models have been applied in order to determine which most accurately classifies malicious URLs. The models applied were a Decision Tree, a Multinomial Logistic Regression and a Random Forest Classifier.



Limitations

The data set has been taken from an open source third party. Hence, the complete accuracy of the data cannot be guaranteed. The data set also was produced approximately two years ago. Consequently, results should be seen as representative of the process rather than a firm confirmation of fact as regards the nature of malicious URLs.

[1] Source: Kaggle.com (https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset), credit to Manu Siddhartha.

Analytical Approach

The data has been analysed using Python and relevant libraries for analysis, visualisation and model building (including Pandas, Scikit-Learn, Seaborn, Matplotlib and urllib.parse).

The analysis followed the below process:

1. Data importation and sense-checking;
2. Data wrangling to clean and prepare the data for analysis;
3. Exploratory data analysis of each URL category (primarily the malicious URLs);
4. Preparation of data for model building;
5. Building of the three model types and evaluating their performance.

The library urllib.parse was used to break the URL down into its constituent parts. The parts of most interest are:

- Scheme (e.g. http, https);
- Netloc (i.e. the root domain of the URL, such as google.com);
- Directories (components such as links to other pages or files):
- Queries (parameters at the end of the URL to define specific content or actions).

Patterns, Trends and Insights

Directories

It was noted that malware URLs showed more signs of obfuscation than the other two malicious types (i.e. they contained more % symbols, often used to escape characters and hide the URL's real content). Obfuscated URLs also tend to be longer in length. However, the % can also appear frequently in benign URLs and so this aspect may only have relevance in comparison to the other malicious types and may not be a strong factor by which URLs can be categorised. However, this is assuming the data set is entirely reliable.

The below table shows the top 10 directories (or directory element combinations) seen in each malicious URL type, along with how many URLs contained them:

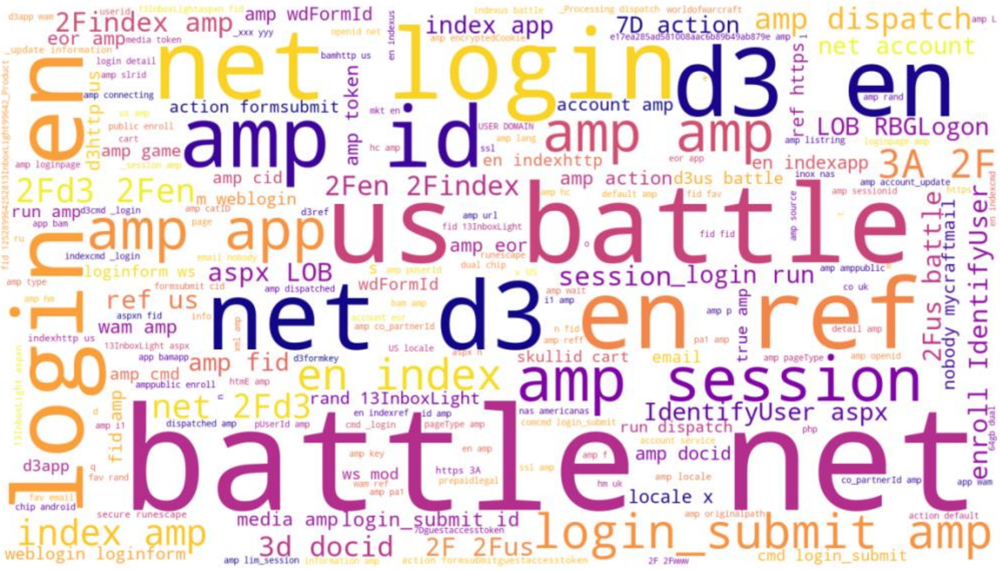
	PHISHING		MALWARE		DEFACEMENT	
TOP 10	DIRECTORIES	COUNT	DIRECTORIES	COUNT	DIRECTORIES	COUNT
1	*None*	2525	Mozi.m	4100	index.php	39141
2	js	372	.i	553	index.html	3145
3	images	293	index.php	385	sejeal.jpg	1283
4	login, en, login.html	255	app, member, SportOption.php	288	component, mailto, index.html	1078
5	www.webring.com, hub	209	download	285	x.txt	663
6	chase, home.php	81	cl	238	portal, index.php	384
7	js, index.htm	73	css, detail, mysite, siteconfig, pro_control.css	180	index.php, component, mailto, index.html	369
8	www.tek-tips.com, threadminder.cfm	72	uc	180	cms, index.php	293
9	Pages, ResponsePage.aspx	66	wiki, lib, exe, css.php	146	component, virtuemart, index.html	285
10	globetrotter-games.com, index.htm	66	mips	112	site, index.php	271

- Clear differences are noticeable across the different URL types, aligning with their individual purposes.
- Phishing URLs most commonly showed up with no identified directories. Existing research into phishing URL formats indicates that they often use login, account and activate, among others, given that they seek to get a victim to enter personal information into the website.
- Malware URLs seem to frequently use 'css', an element in an URL often used to link to a resource. This likely causes a victim to download a malicious malware file. Mozi.m stands out as very commonly present.
- Index.php and index.html appear frequently in defacement URLs. Again, this indicates the replacement of the victim webpage with that of the attacker.

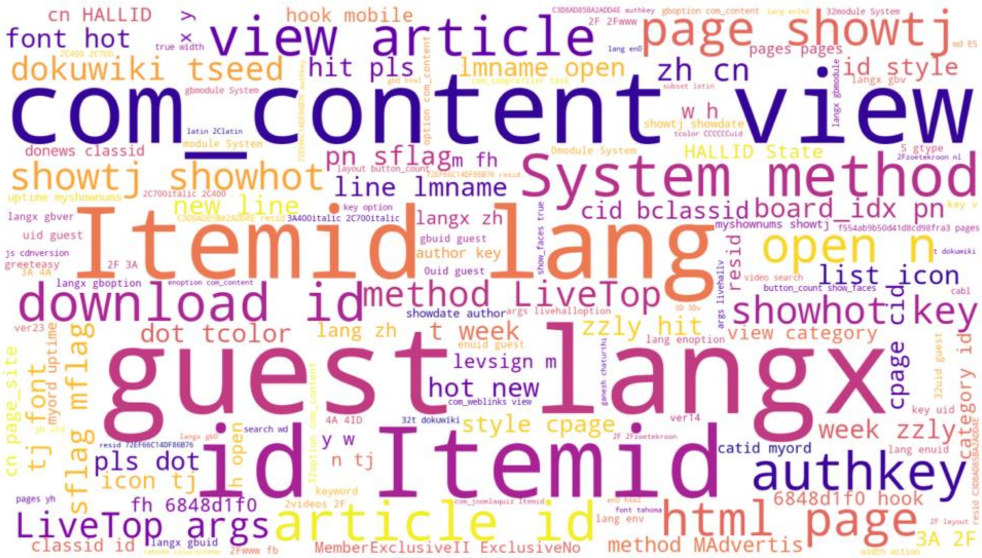
- These outstanding elements in each type's top 10 could be used by a machine learning model to help identify URLs within each type category.

Queries

The most frequently-occurring elements within the query part of phishing URLs are shown below. Again, login is seen frequently, as well as battle.net and us.battle. Of note is that us.battle.net was seen frequently amongst the benign URLs – battle.net and us.battle may therefore be type-squatted to appear like an authentic site:



Malware URLs regularly contained 'guest' and 'lang'. These may refer to a user ID as 'guest' and 'lang' may refer to the setting of the language in use:



‘Component’ features frequently in defacement URL queries, as well as ‘article_id’ and ‘view_article’:

Schemes

	PHISHING		MALWARE		DEFACEMENT	
TOP 10	SCHEME	COUNT	SCHEME	COUNT	SCHEME	COUNT
1	http	17886	http	24546	http	96457
2	https	6966	https	6764		
3	www.mit.edu	5	77.228.191.183	9		
4	ilpubs.stanford.edu	3	escuelanet.com	1		
5	www-vs.informatik.uni-ulm.de	2				
6	www.ripn.net	2				
7	dbpubs.stanford.edu	2				
8	ftp	2				
9	gopher.quux.org	2				
10	www.ee.ryerson.ca	2				

- http was most common across all malicious types. However, the same is noted with benign URLs too, suggesting that this feature may not be overly reliable in determining whether an URL is malicious or not.
- Of interest is that phishing URLs have often used those that appear to be from educational institutions.
- Malware URLs showed some usage of an IP address – Domain Tools shows this to be an IP of Vodafone España.
- Usage by a malware URL of escuelanet.com also indicates both a Spanish and educational institution link, given that ‘escuela’ means school in Spanish.

Domains

TOP 10	PHISHING		MALWARE		DEFACEMENT	
	DOMAIN	COUNT	DOMAIN	COUNT	DOMAIN	COUNT
1	pastehtml.com	944	9779.info	3984	allaroundrental.com	265
2	docs.google.com	275	mitsui-jyuku.mixh.jp	2879	bruynzeelmultipanel.be	222
3	firebasestorage.googleapis.com	127	apbfiber.com	1147	ninopizzaria.com.br	209
4	storage.googleapis.com	116	pastebin.com	987	tandemimmobilier.fr	191
5	naylorantiques.com	114	toulousa.com	501	zibae.ir	188
6	cheaproomsvalencia.com	110	grasslandhotel.com.vn	354	holidayclub-mtb.com	108
7	playarprint.com	83	hotlinegsm.com	349	zjtfjt.com	102
8	distrimarsanitarios.soydg.com	75	3cf.ru	295	niobestudio.com	102
9	forms.office.com	67	chinesevie.com	290	enprofil.nl	102
10	drive-google-com.fanalav.com	66	onedrive.live.com	289	klavierhaus-alber.de	102

- Each malicious type has evidently been using a different set of most-used domains. Of interest is the high usage of pastehtml.com for phishing URLs and 9779.info for malware URLs. Pastebin.com, the fourth most common URL used in malware URLs, is a well-known depository for stolen credentials and frequently used by cyber criminals.

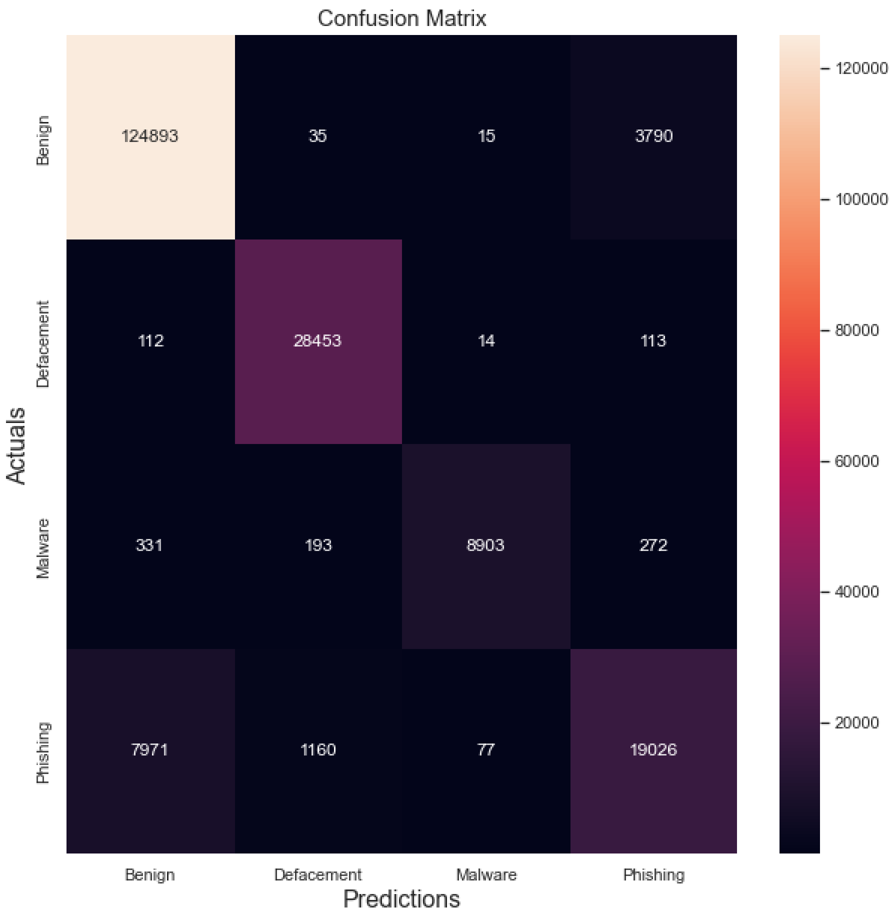
Predictive Model Performance

Based on the analysis of the URLs, the features were used to train three different models. The overall performance of these are shown below:

MODEL	**ACCURACY**
Decision Tree	86%
Multinomial Logistic Regression	79%
Random Forest Classifier	93%

Of the three models implemented, the Random Forest Classifier performed the best. Across all models, the prediction of phishing URLs was least effective across all four types. The Random Forest has a 67% chance of correctly predicting a phishing URL, far better than the other models (albeit with a considerable margin for error).

The matrix below shows predicted versus actual classifications by the Random Forest Classifier. It appears that the model frequently confuses benign and phishing URLs. This may be due to either similarities in the two types that need further exploration, or could be the result of incorrect labelling within the original data set:



Recommended Further Work

- Assess why phishing and benign URLs are more frequently mixed up by the model and do any necessary cleaning and re-running of the models.
- Use an ‘un-shortening’ service on the data to return any shortened URLs to their original format in order to potentially improve the models.
- Create more lexical features based on characters within the URLs, such as @ and &, to further potentially improve model accuracy.
- Run ‘whois’ checks against the URLs to identify number of days since registration to use as an additional modelling feature (i.e. malicious URLs may have been set up for short periods). This would be potentially very time-consuming and data would need to be more timely for this to be effective; this could be performed on an up-to-date data set.
- Extract features from the web page itself; again, this could be done and could improve the models further, but would be a time-consuming endeavour.