Overview

- Introduction
- Background
- Objective
- Methodology
  - Data Collection
  - Data Visualization
  - Machine Learning
  - Model Deployment
- Extension
- Conclusion

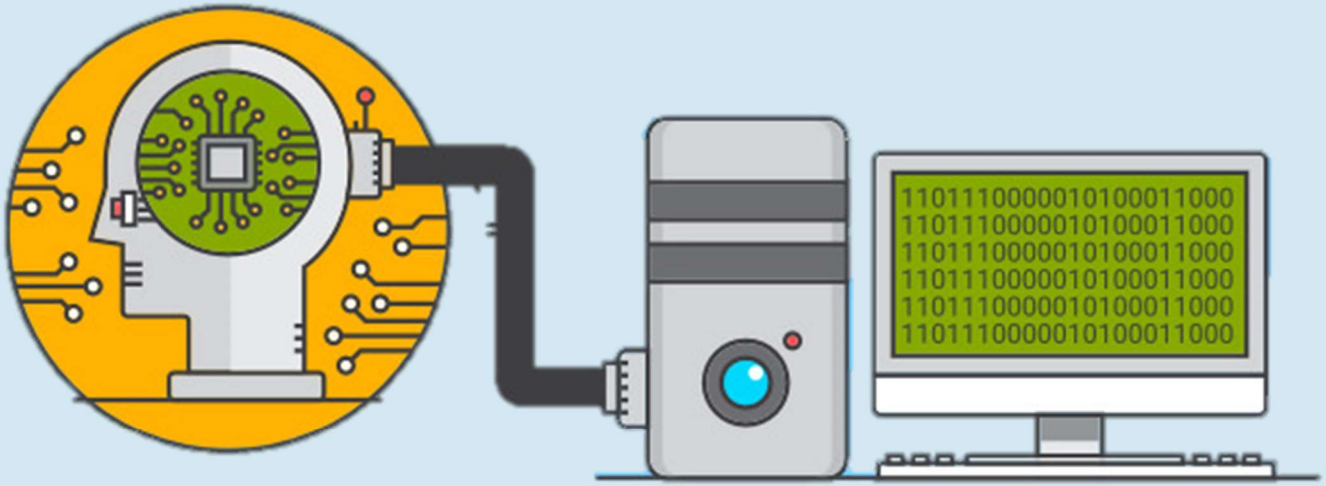
Introduction

Introduction

Phishing – uses e-mails or malicious websites to solicit personal information from an individual or company by posing as a trustworthy organization or entity

Introduction

Therefore, computer-based and data-driven solutions should be implemented as it would enable a computer to detect malicious websites and protect users from interacting with them.



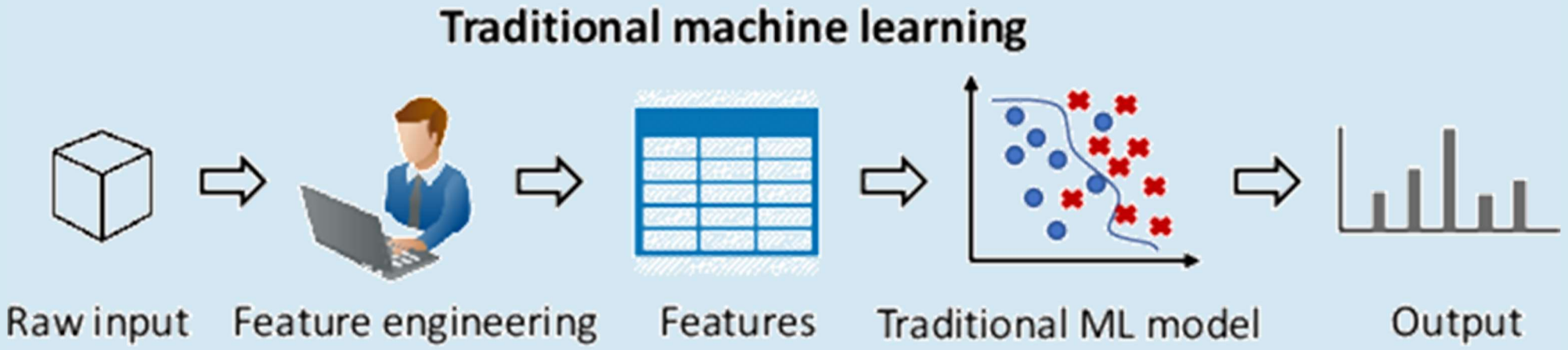
Background

Uniform Resource Locators (URLs) – identify illegitimate websites

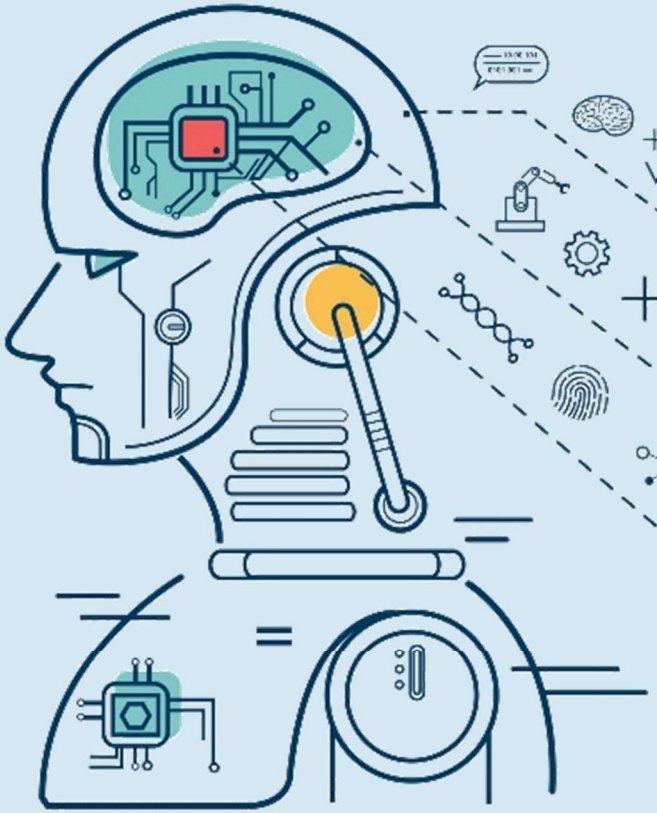
Thus, a possible approach is to use a **blacklist** of malicious URLs developed by anti-virus groups.

Background

To do this, machine-learning based approaches can be used. Machine-learning is a system that can categorize new phishing sites through a model developed using training sets of known attacks.



Objective



Train classification model using 2 existing datasets

Test the trained models

Incorporate the models into a Chrome Extension

Methodology – Data Collection

Dataset #1:
Dataset of Malicious and Benign Webpages
Number of rows: 1,561,934
Number of columns: 11
Source: Malcrawler, Google Safe Browsing API, Mendeley Data

Dataset #2:
Malicious & Benign URLs
Number of rows: 450,176
Number of columns: 2
Source: Phistank

Key Terms

Dataset #1

Variable	Details	Variable	Details
url	URL of webpage	tld	Top level domain of webpage
ip_add	IP Address of webpage	who_is	Whether the WHO IS domain information is complete
geo_loc	Geographic location of the hosted webpage	https	Whether the site uses HTTP or HTTPS
url_len	Length of URL	content	Raw webpage content including JavaScript code
js_len	Length of JavaScript code on website	js_obf_len	Length of obfuscated JavaScript code
label	Benign or malicious		

Dataset #2

Variable	Details
url	URL of webpage
label	Benign or malicious

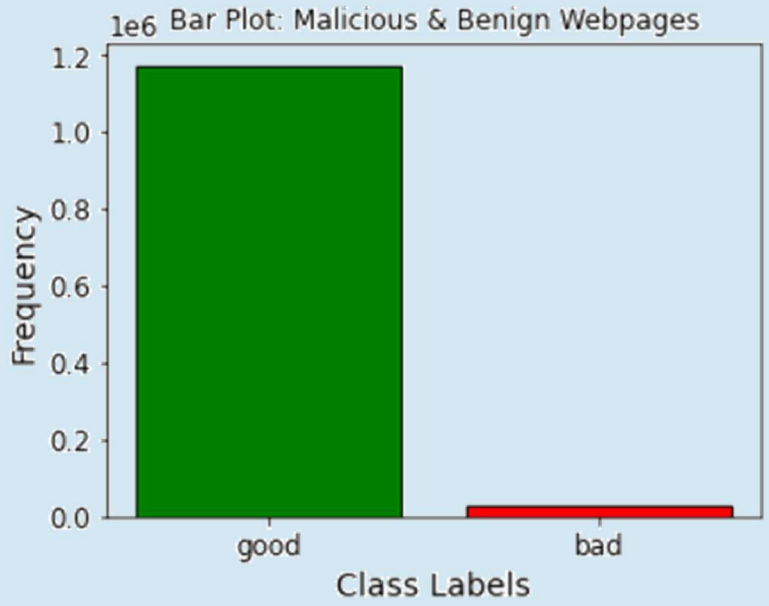
Data Visualisation

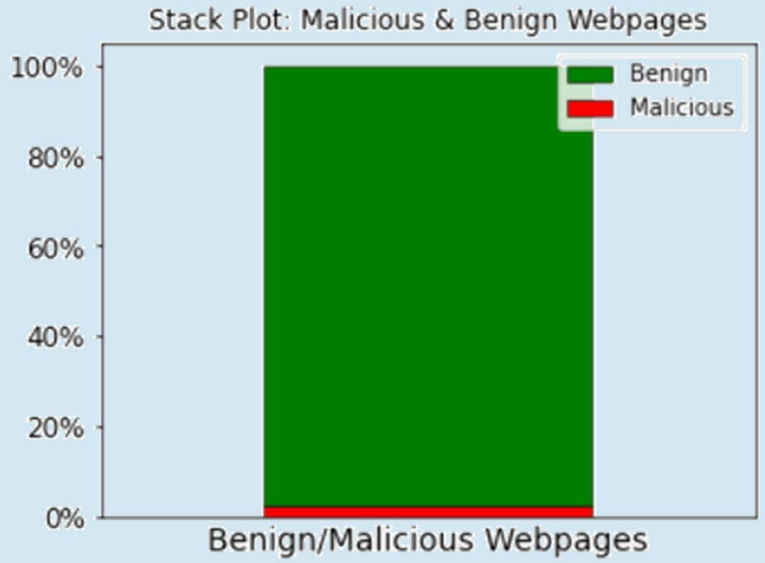
	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

Analysis of Class Label

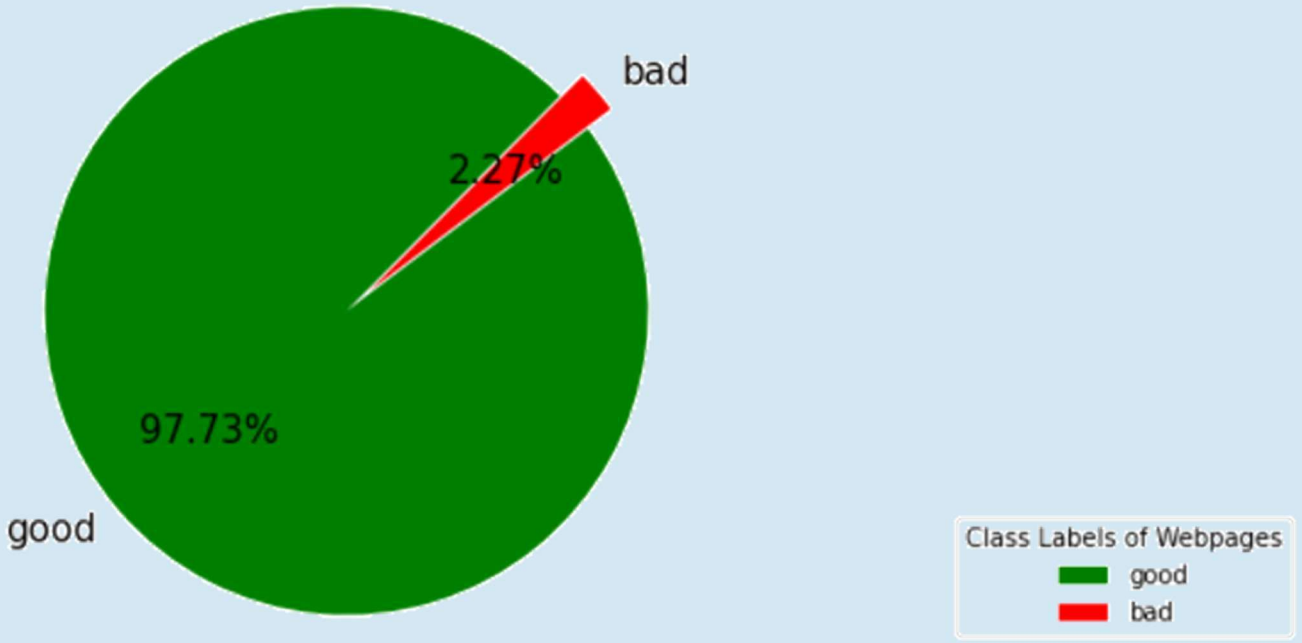
```
label
bad       27253
good    1172747
dtype: int64
```





Analysis of Class Label

```
Total of Samples: 1200000
Malicious: 27253 (2.27% of total)
Benign: 1172747 (97.73% of total)
```



Data Visualisation

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

Analysis of 'url' Attribute

- ?
- \\
- .
- ;
- /
- \

url	url_vect
http://members.tripod.com/russiastation/	members tripod russiastation
http://www.ddj.com/cpp/184403822	ddj cpp 184403822
http://www.naef-usa.com/	naef-usa
http://www.ff-b2b.de/	ff-b2b
http://us.imdb.com/title/tt0176269/	us imdb title tt0176269
...	...
http://csrc.nist.gov/rbac/	csrc nist rbac
http://www.unm.edu/~hist/	unm ~hist
http://www.syfyportal.com/news423380.html	syfyportal news423380 html
http://www.wardkenpo.ie	wardkenpo
http://homepages.gotadsl.co.uk/~jgm/ekmm/	homepages gotadsl ~jgm ekmm

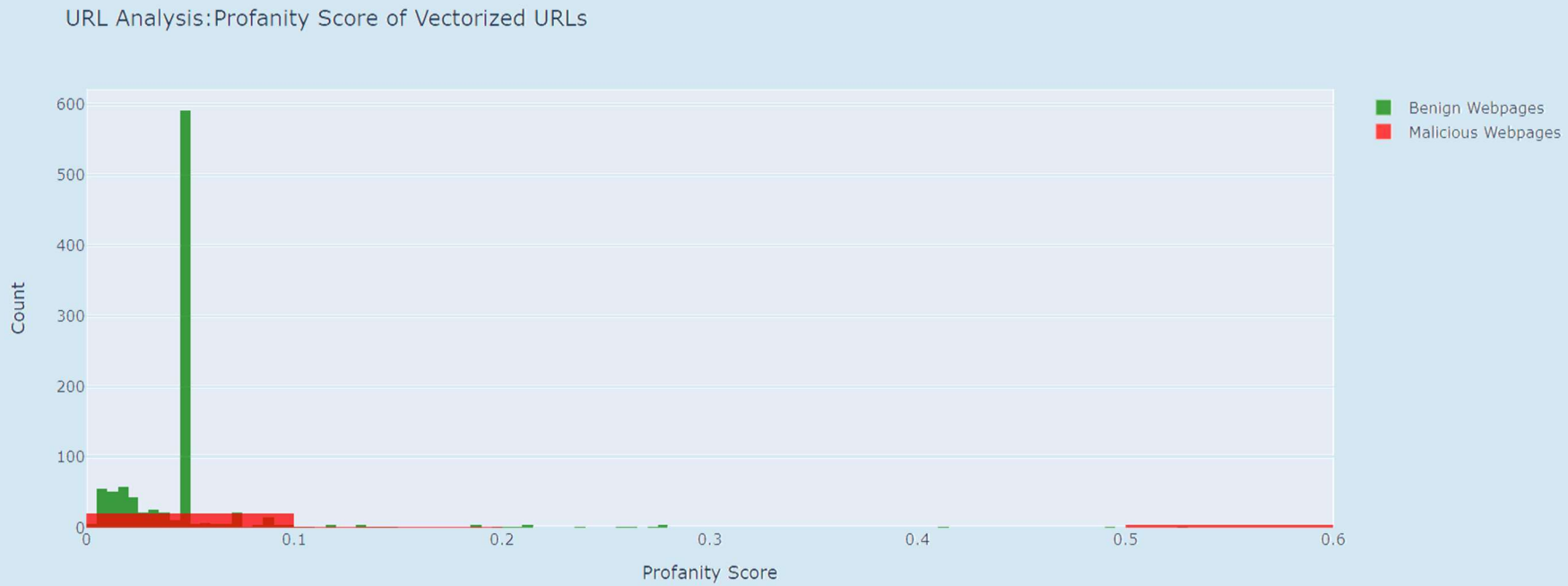
Analysis of 'url' Attribute

Profanity check

- returns a 1 if it is offensive
- else 0

Analysis of 'url' Attribute

Average Malicious Webpage Profanity Score: 0.09000000
Average Benign Webpage Profanity Score: 0.04468303

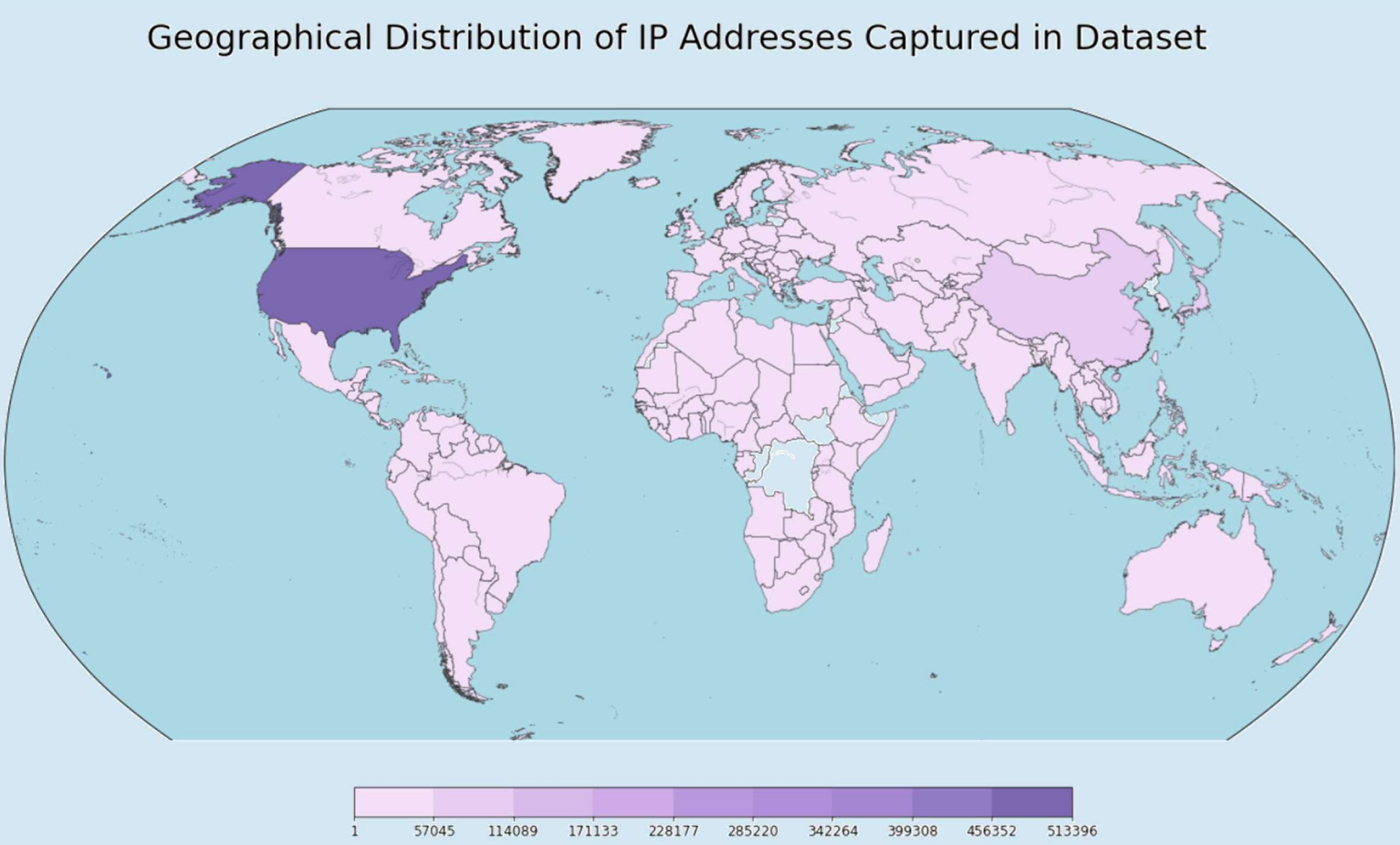


Data Visualisation

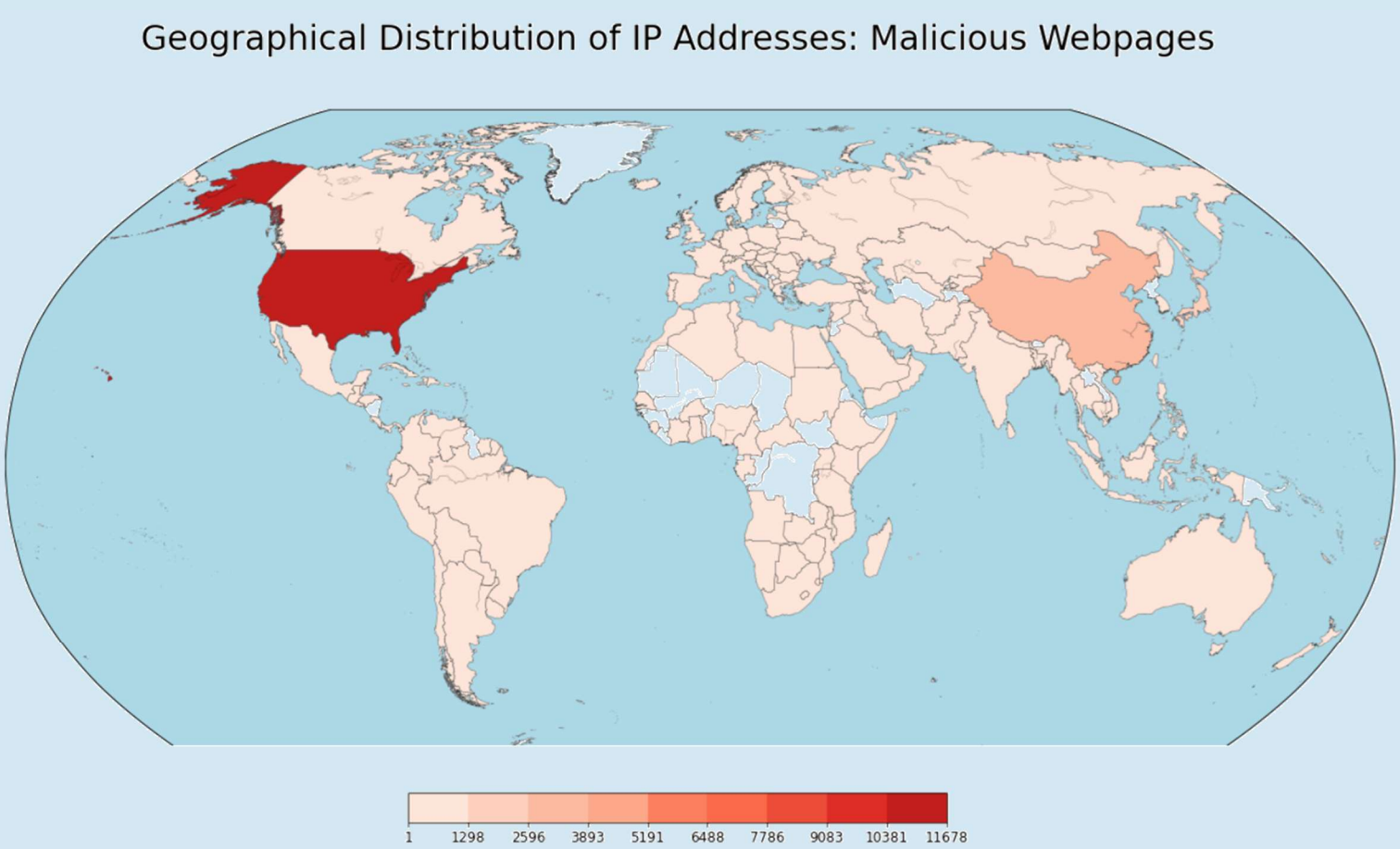
	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

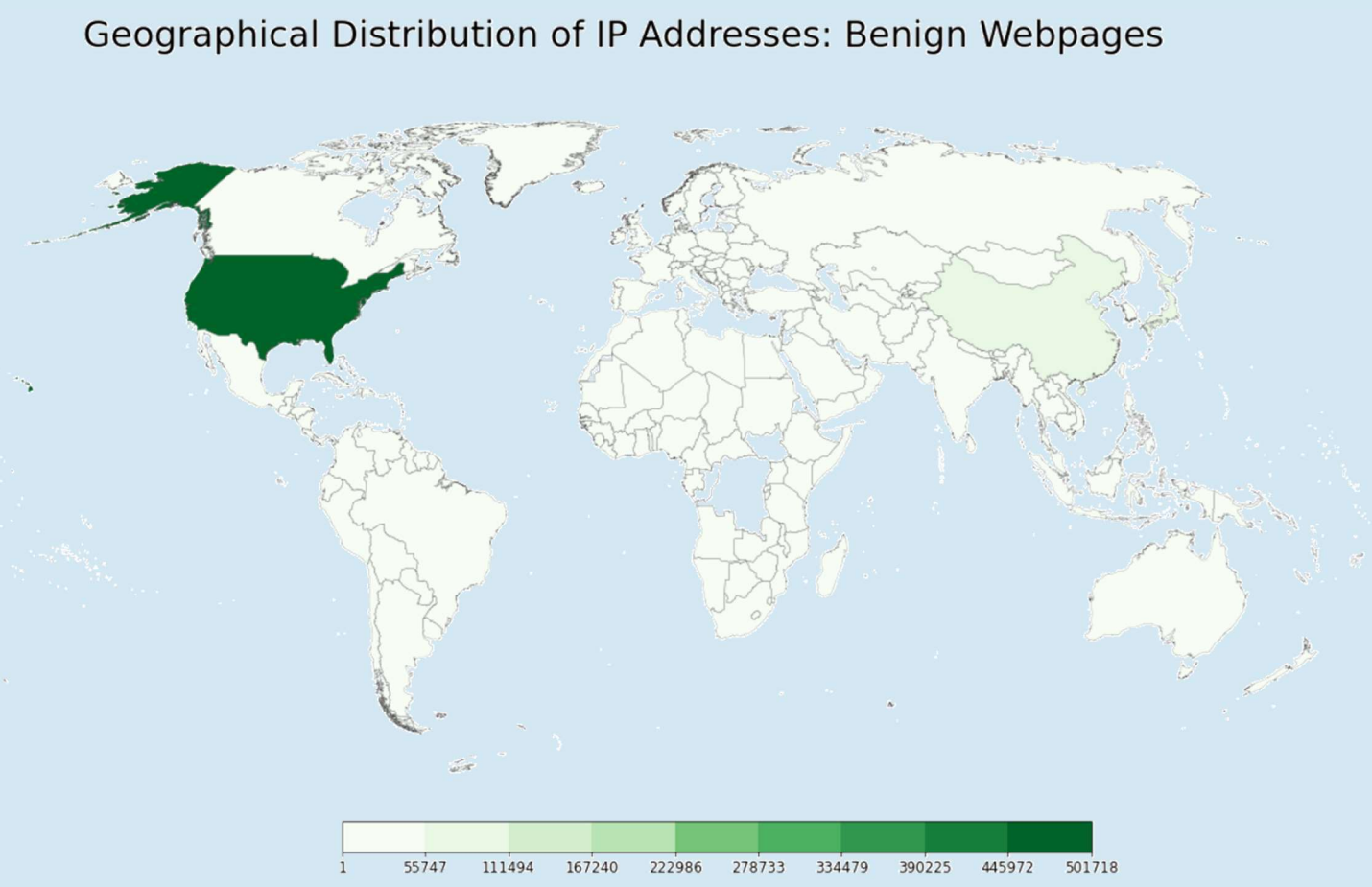
Analysis of 'ip_add' & 'geo_loc' Attributes



Analysis of 'ip_add' & 'geo_loc' Attributes



Analysis of 'ip_add' & 'geo_loc' Attributes

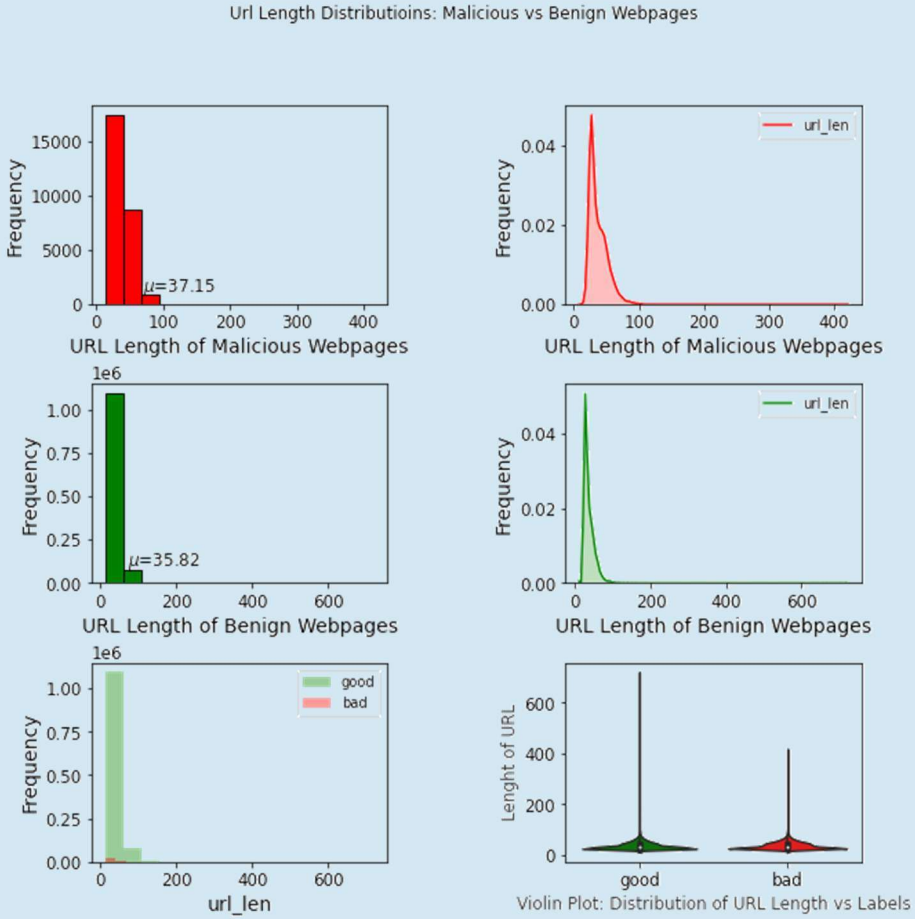


Data Visualisation

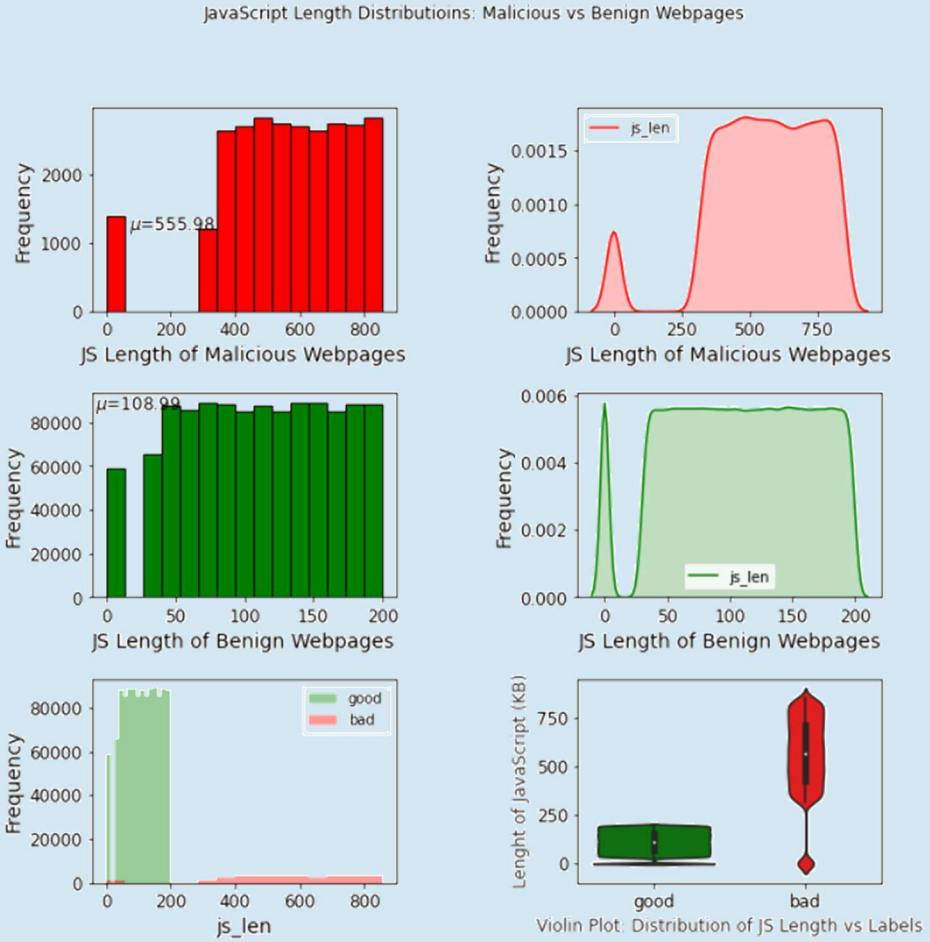
	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

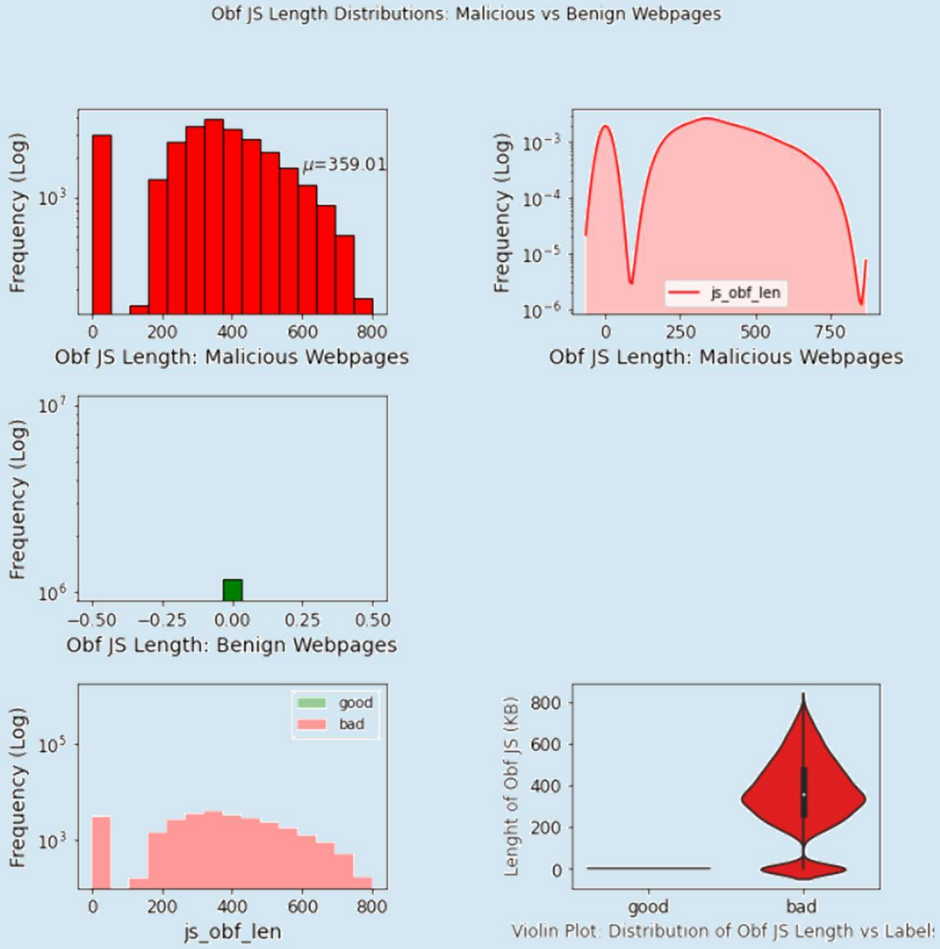
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



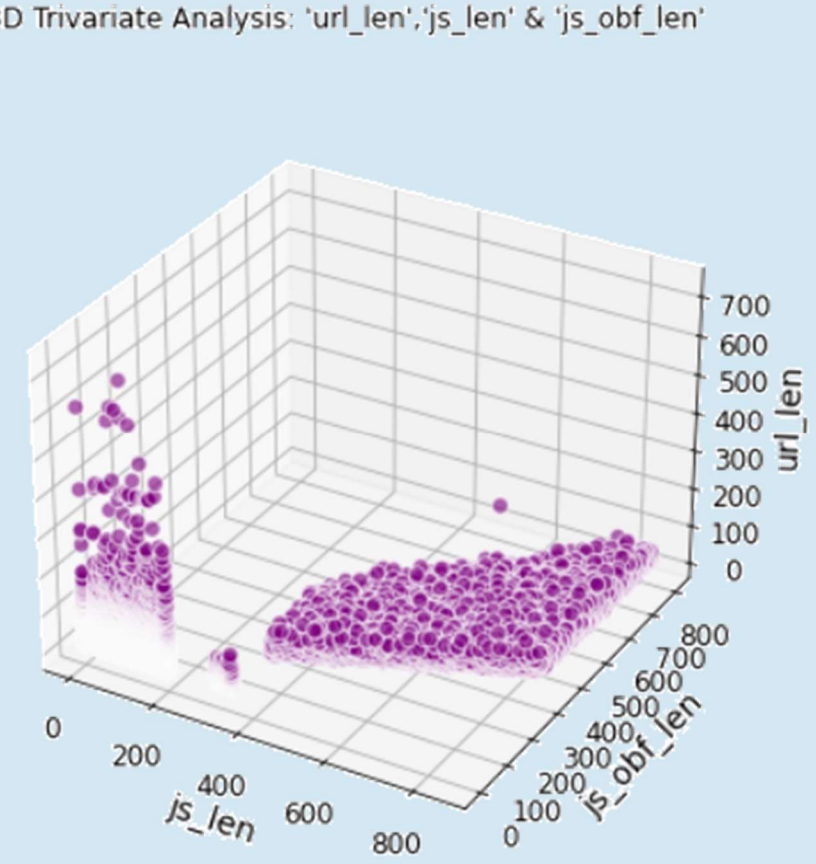
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'

	url_len	js_len	js_obf_len	url_vect
count	1.200000e+06	1.200000e+06	1.200000e+06	1.200000e+06
mean	3.585337e+01	1.191463e+02	8.153424e+00	1.118906e-01
std	1.441089e+01	9.046649e+01	6.001398e+01	3.581809e-02
min	1.200000e+01	0.000000e+00	0.000000e+00	3.000000e-03
25%	2.600000e+01	6.650000e+01	0.000000e+00	1.030000e-01
50%	3.200000e+01	1.120000e+02	0.000000e+00	1.210000e-01
75%	4.200000e+01	1.580000e+02	0.000000e+00	1.210000e-01
max	7.210000e+02	8.541000e+02	8.028540e+02	1.000000e+00

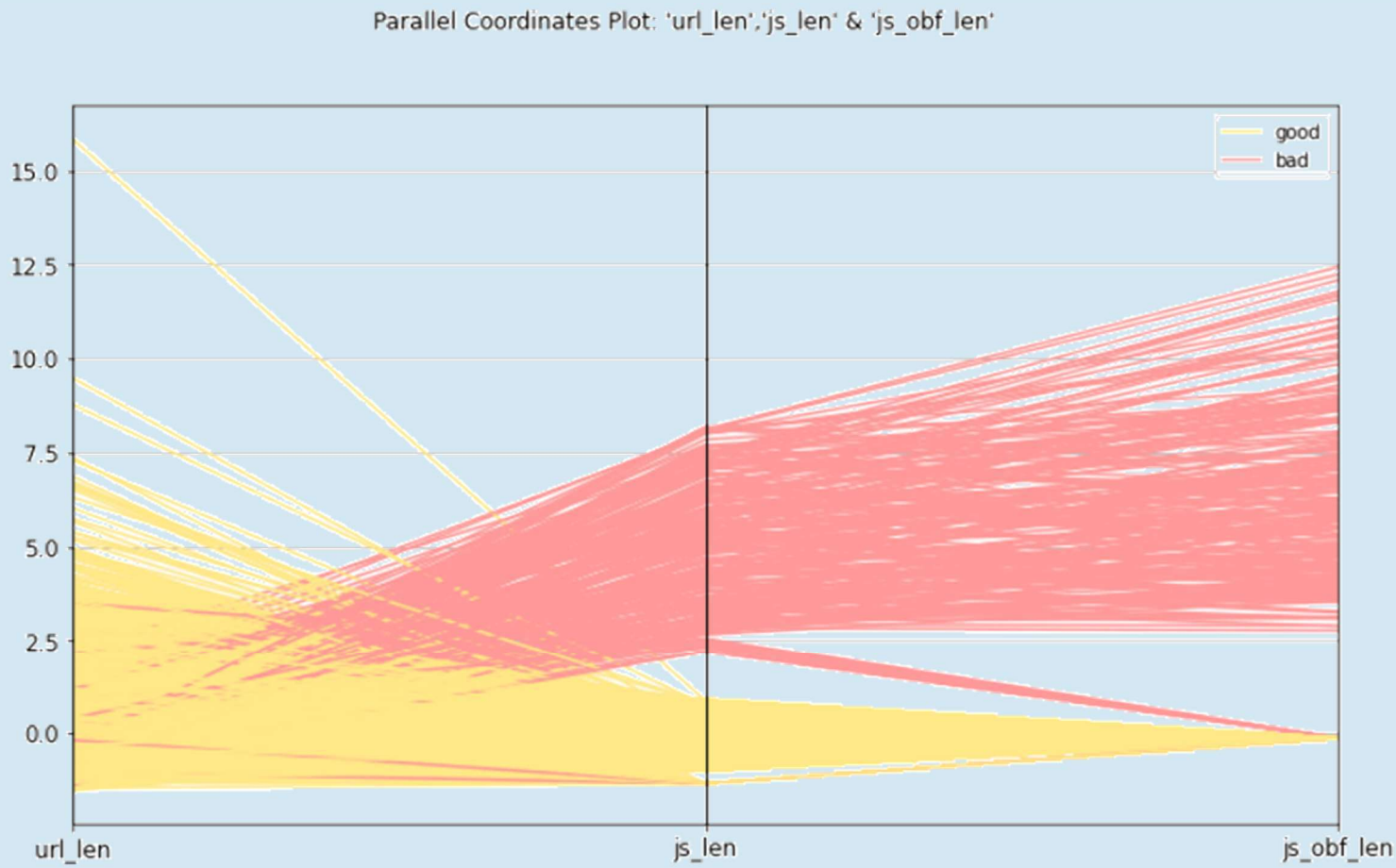
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'

	Benign Webpages Statistics			Malicious Webpages Statistics		
	url_len	js_len	js_obf_len	url_len	js_len	js_obf_len
count	1172747.00	1172747.00	1172747.0	27253.00	27253.00	27253.00
mean	35.82	108.99	0.0	37.15	555.98	359.01
std	14.42	53.97	0.0	14.02	199.36	180.63
min	12.00	0.00	0.0	13.00	0.00	0.00
25%	26.00	65.50	0.0	27.00	431.10	261.86
50%	32.00	110.00	0.0	33.00	569.70	361.35
75%	42.00	155.00	0.0	45.00	714.60	478.86
max	721.00	199.50	0.0	416.00	854.10	802.85

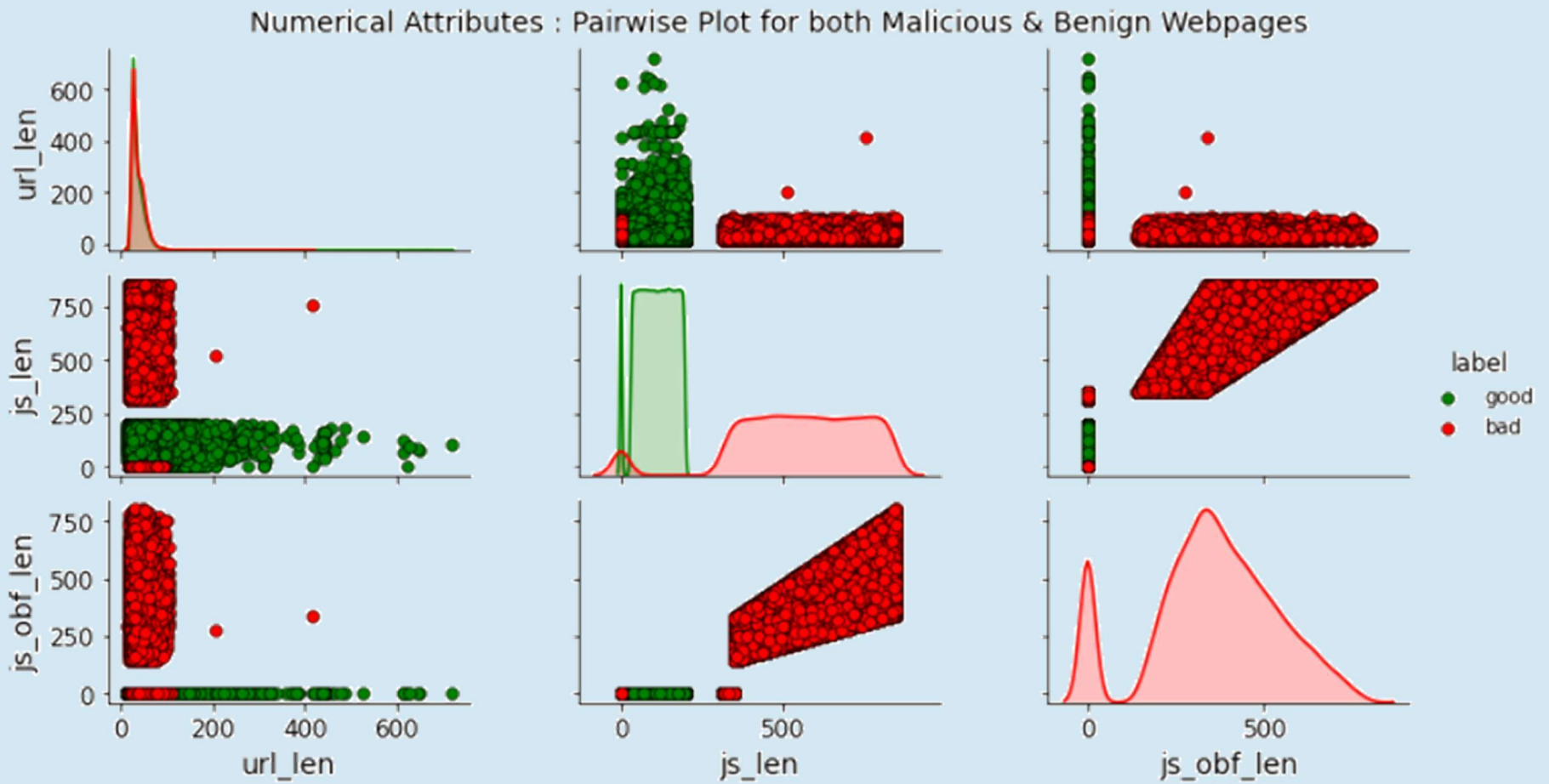
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



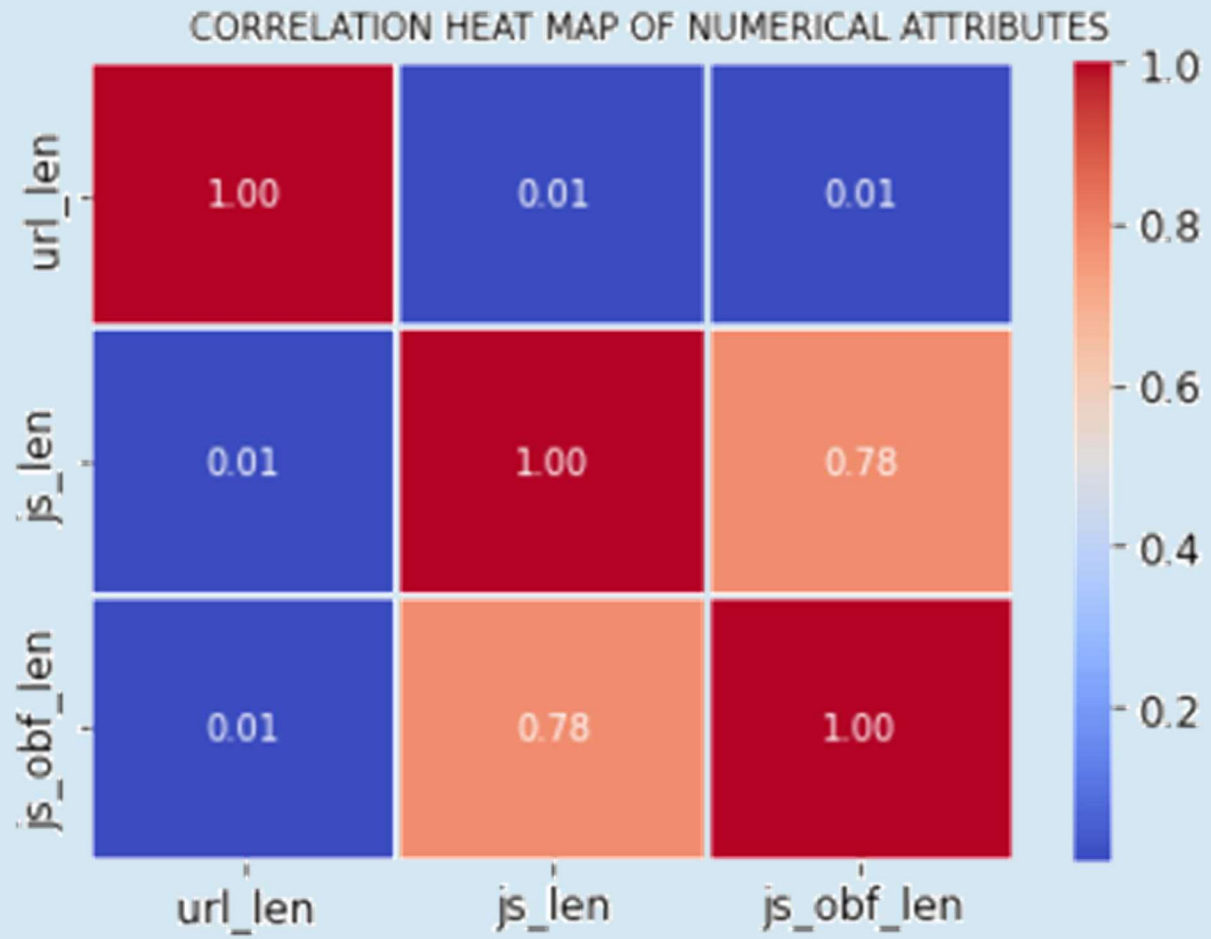
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



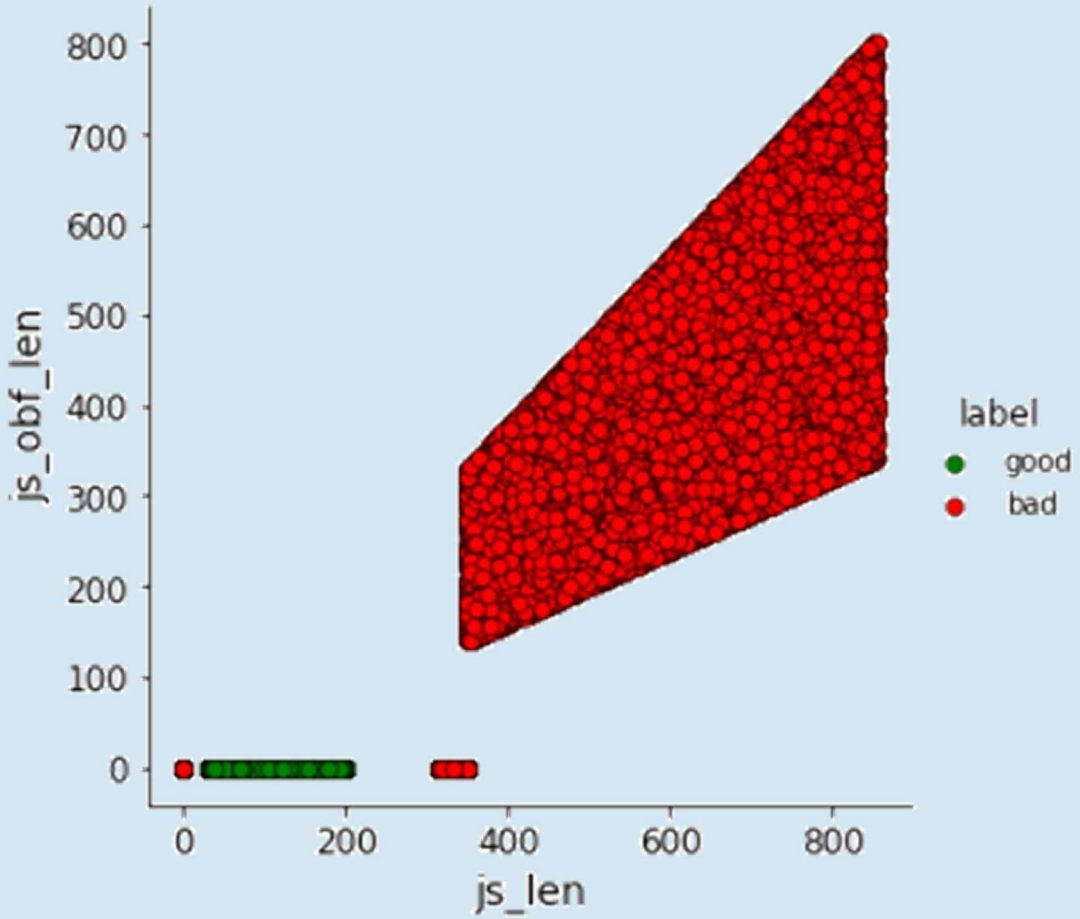
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



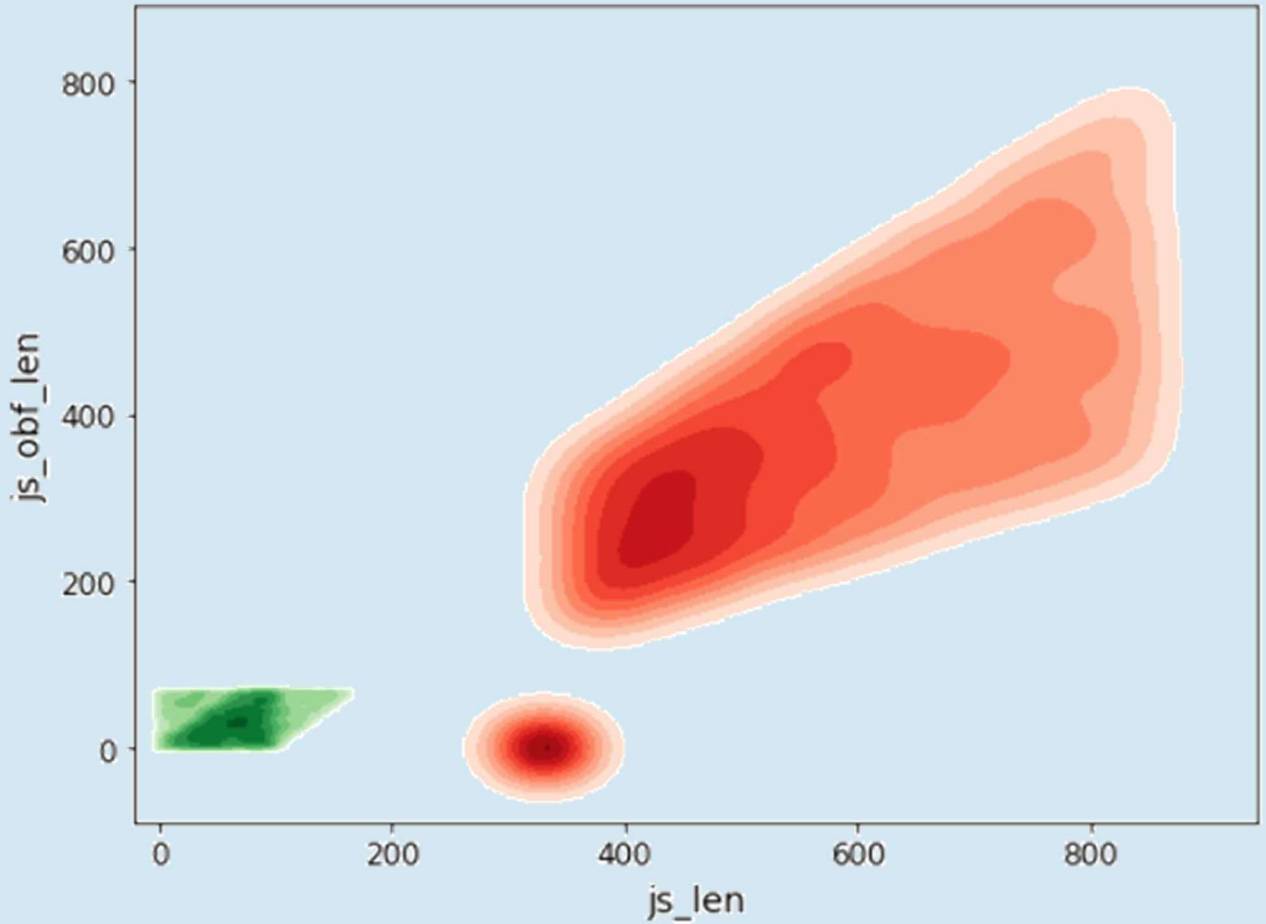
Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'



Analysis of Numerical Attributes: 'url_len', 'js_len' and 'js_obf_len'

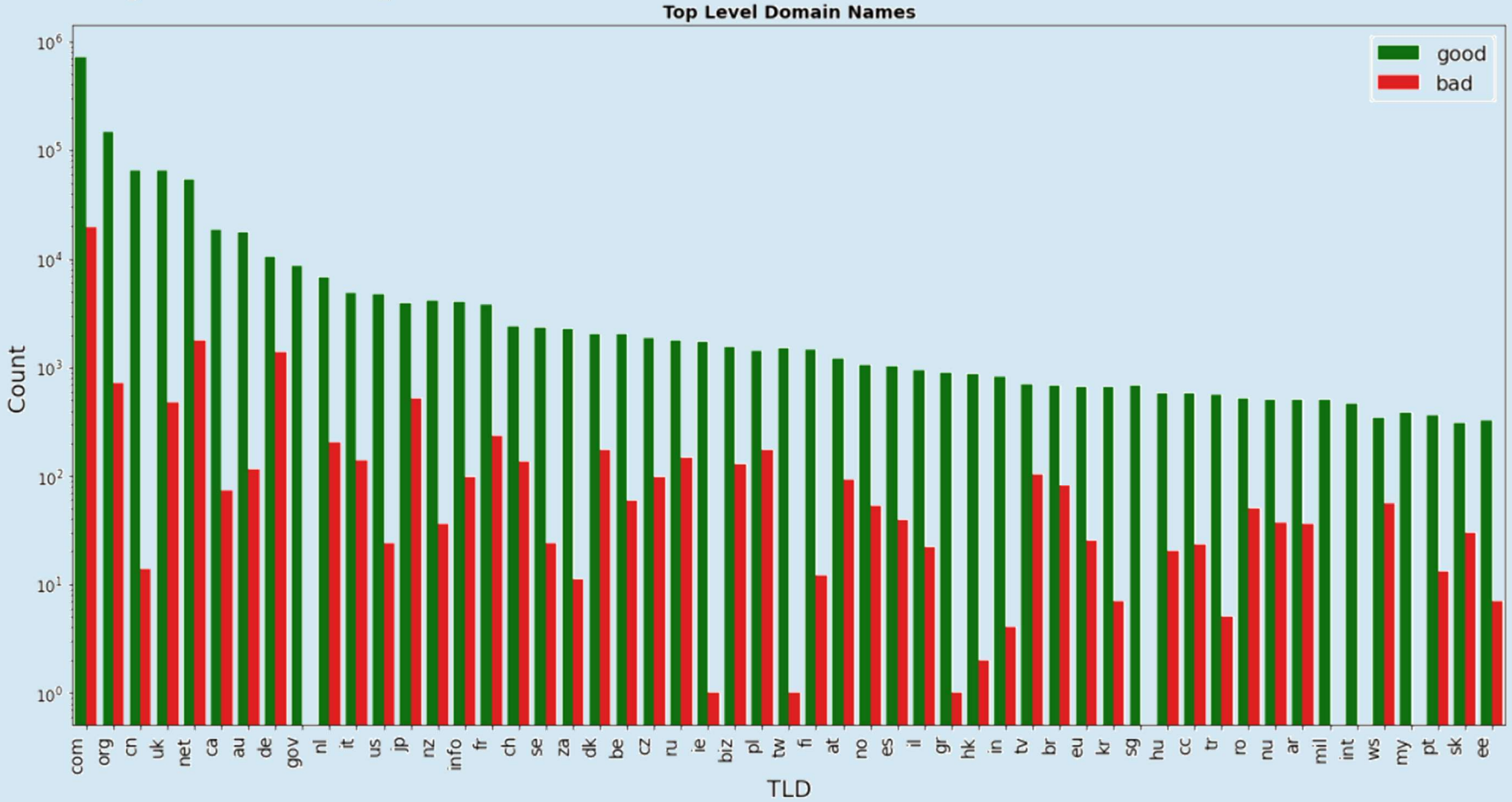


Data Visualisation

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

Analysis of Top Level Domain: 'tld' Attribute



Data Visualisation

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

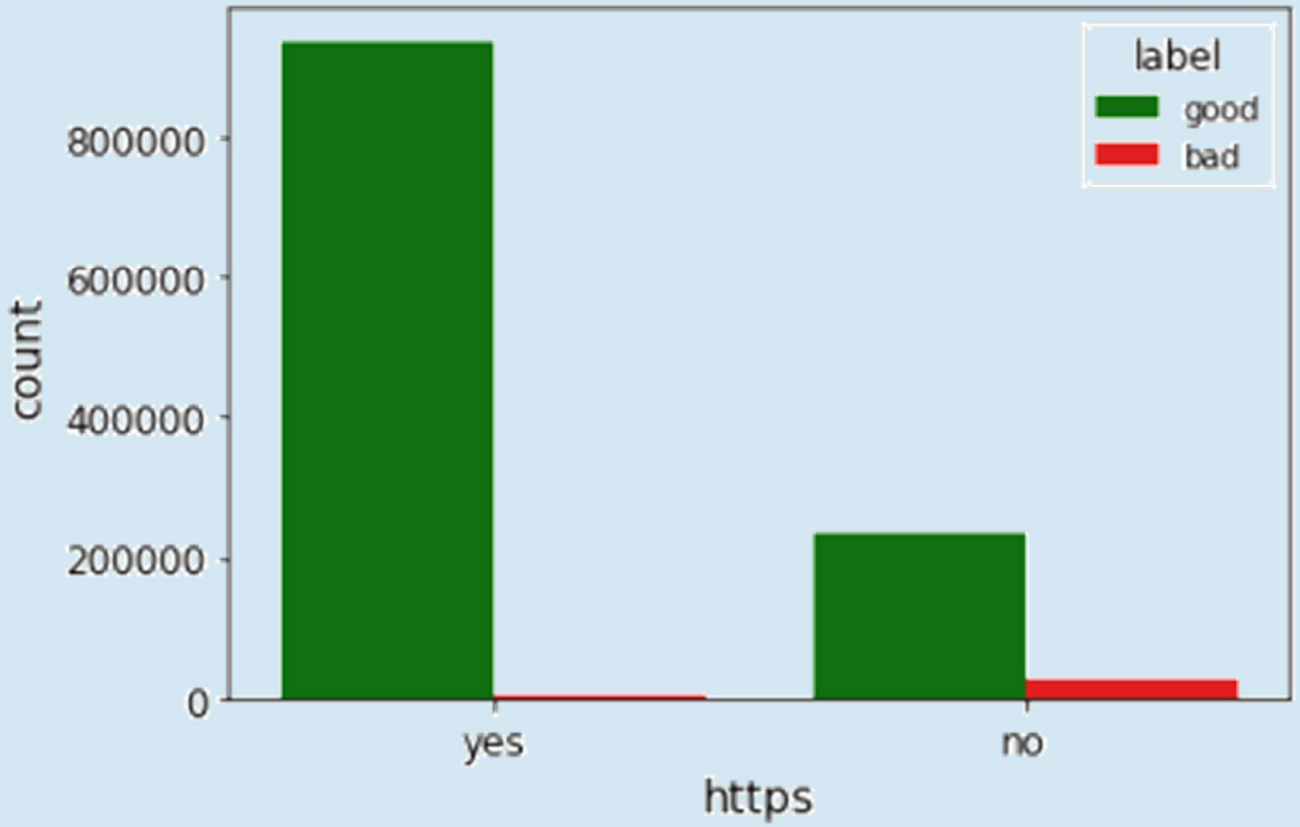
Analysis of WHO IS Registration Information: 'who_is' Attribute

Data Visualisation

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

Analysis of HTTP Status: 'https' Attribute



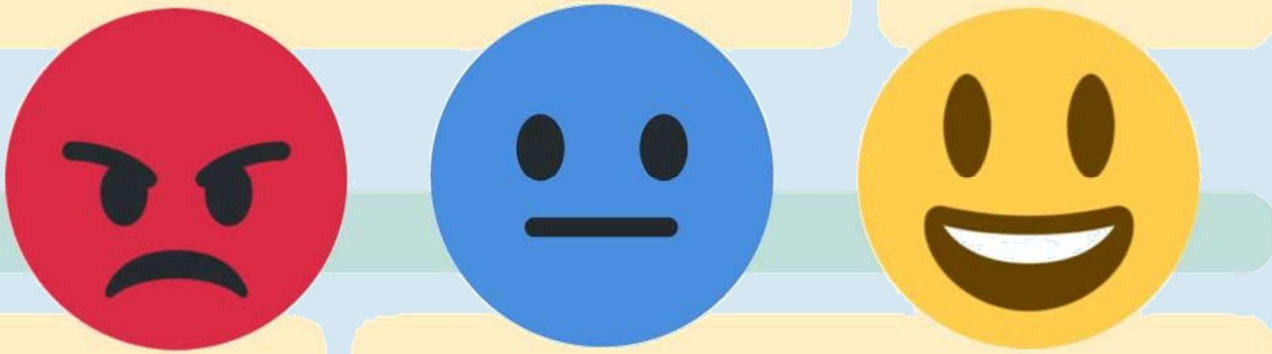
Data Visualisation

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

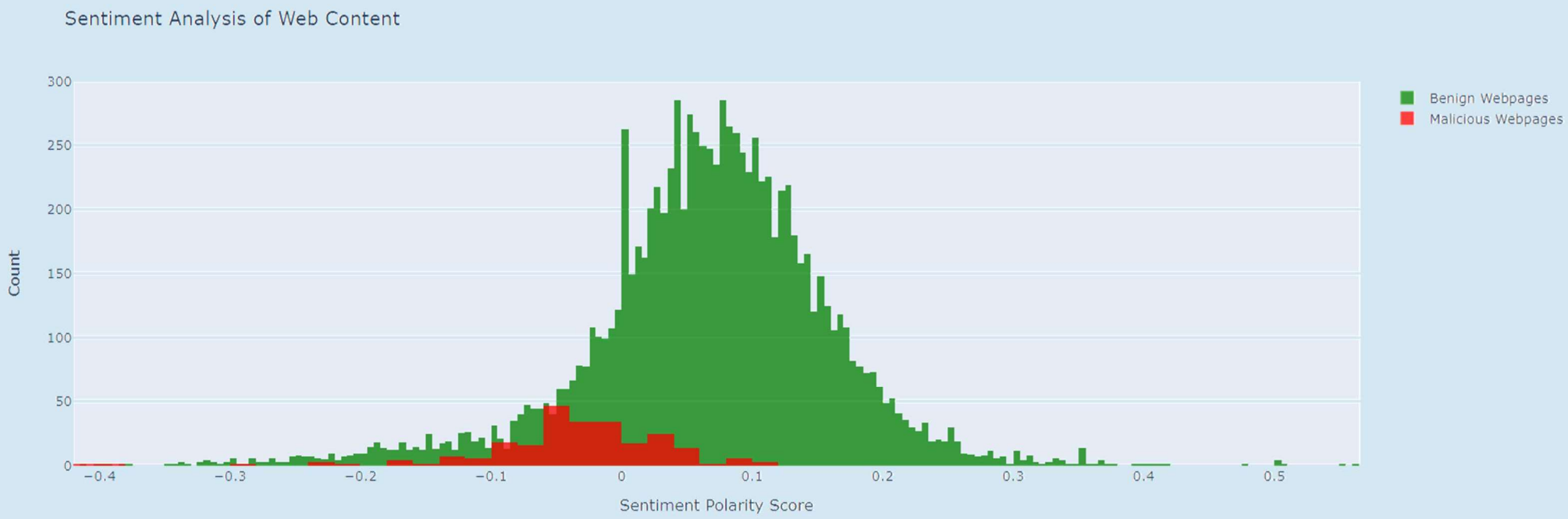
Analysis of Web Content (Raw Web content Including JavaScript)

Sentiment Analysis

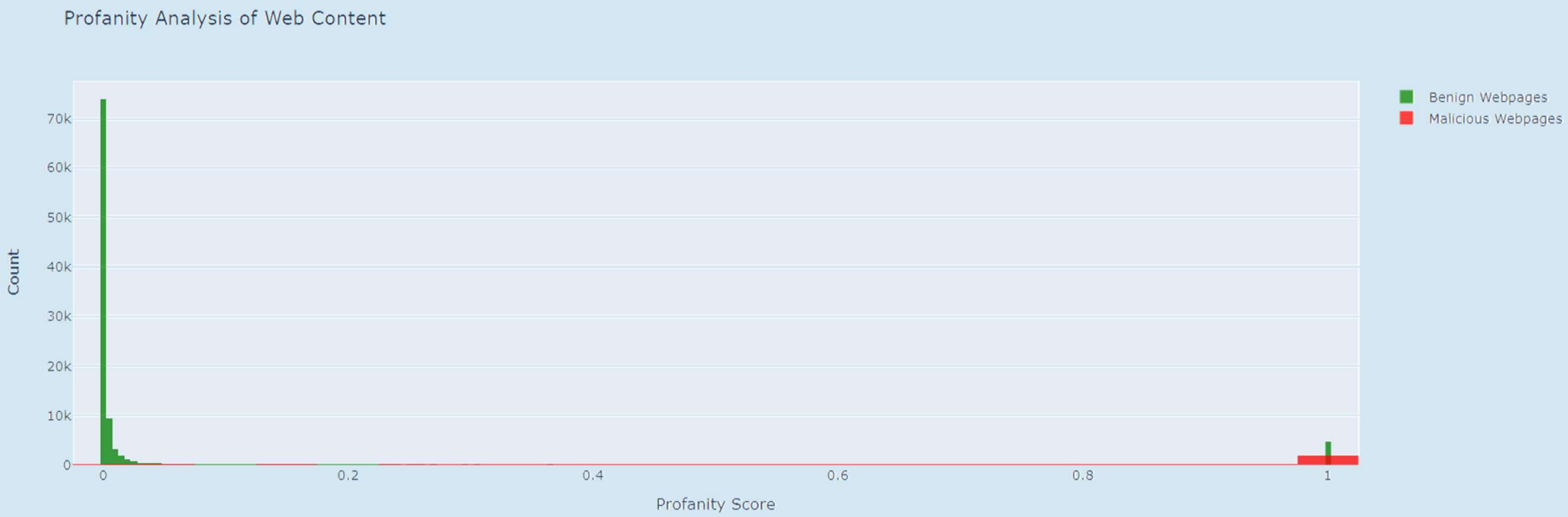


Negative Neutral Positive

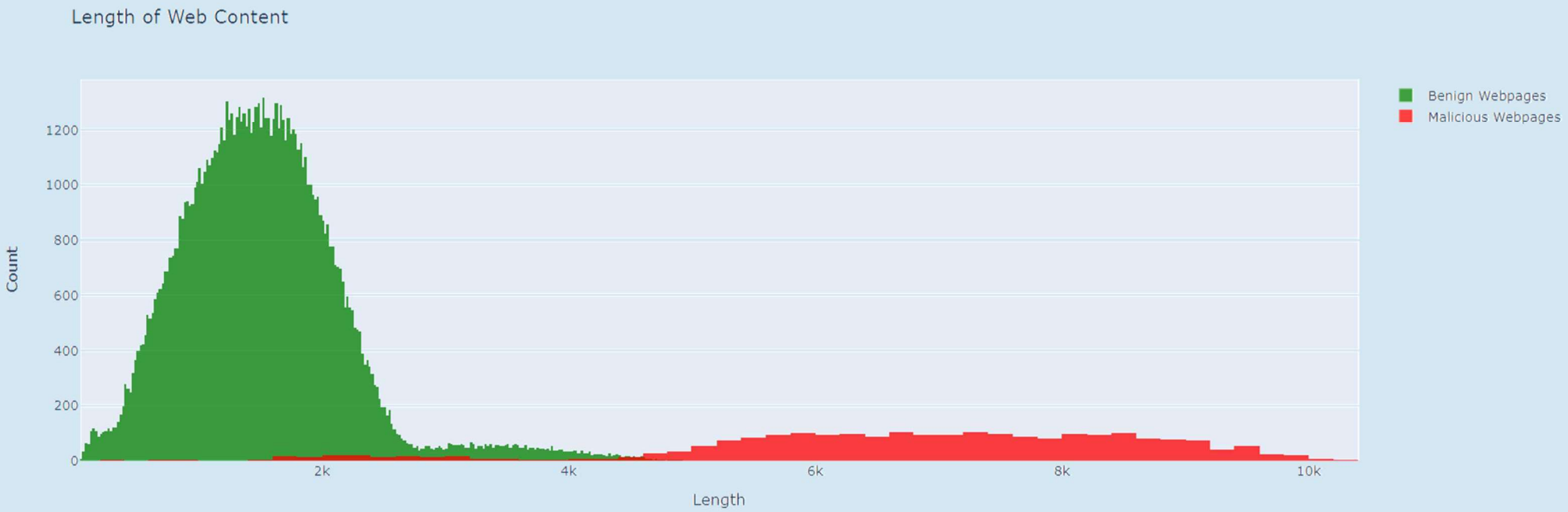
Analysis of Web Content (Raw Web content Including JavaScript)



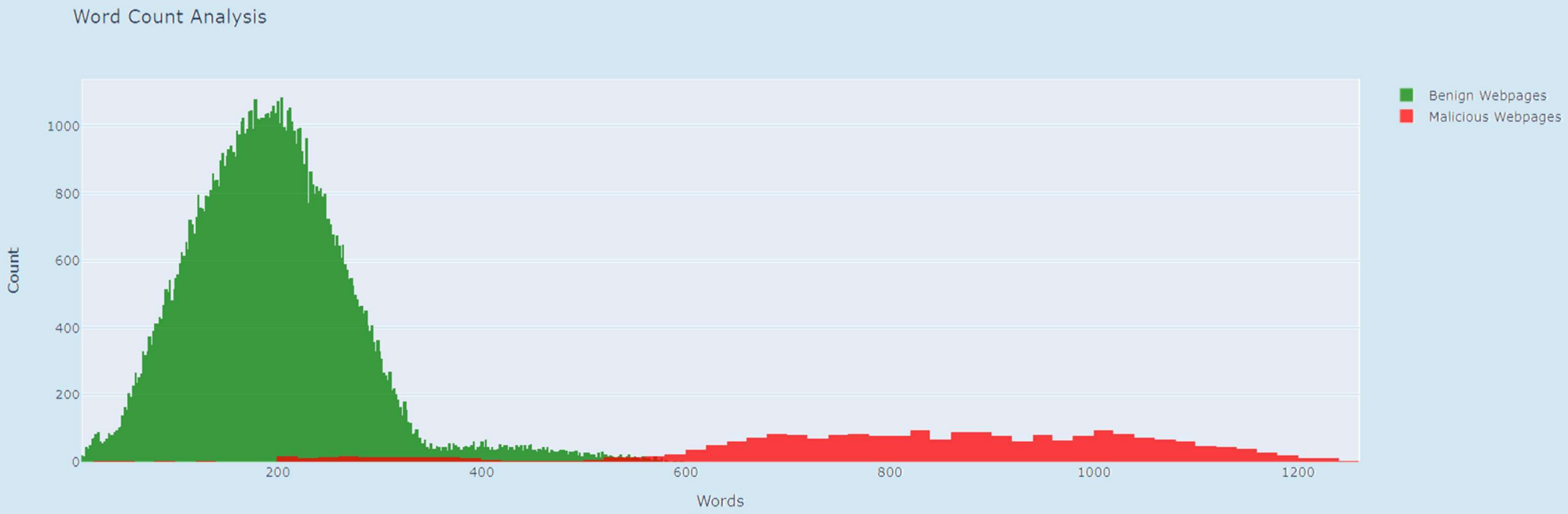
Analysis of Web Content (Raw Web content Including JavaScript)



Analysis of Web Content (Raw Web content Including JavaScript)



Analysis of Web Content (Raw Web content Including JavaScript)



Data Visualisation

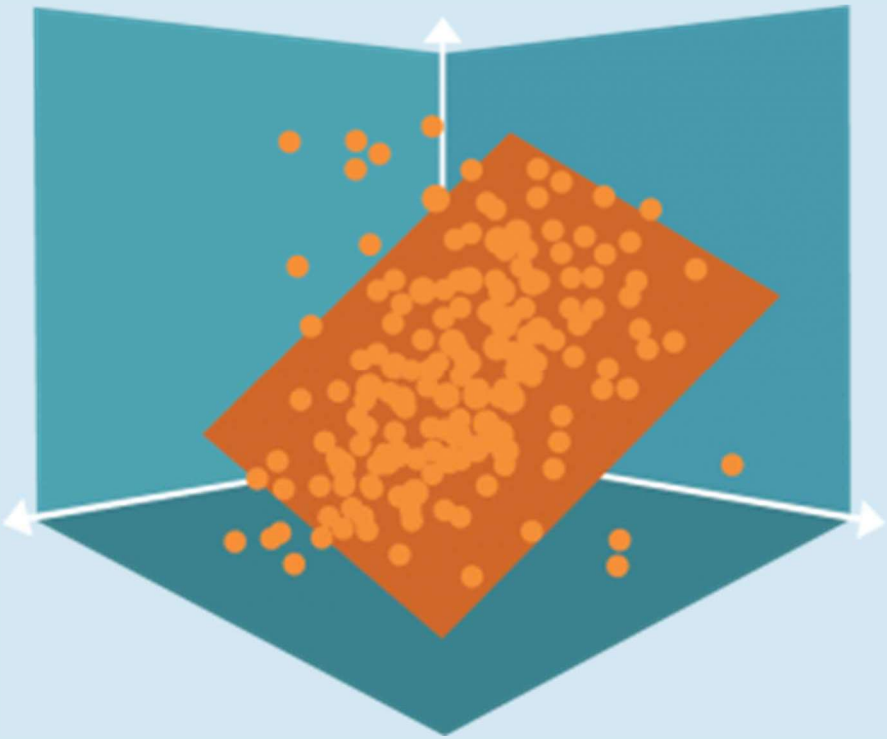
	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

1200000 rows × 12 columns

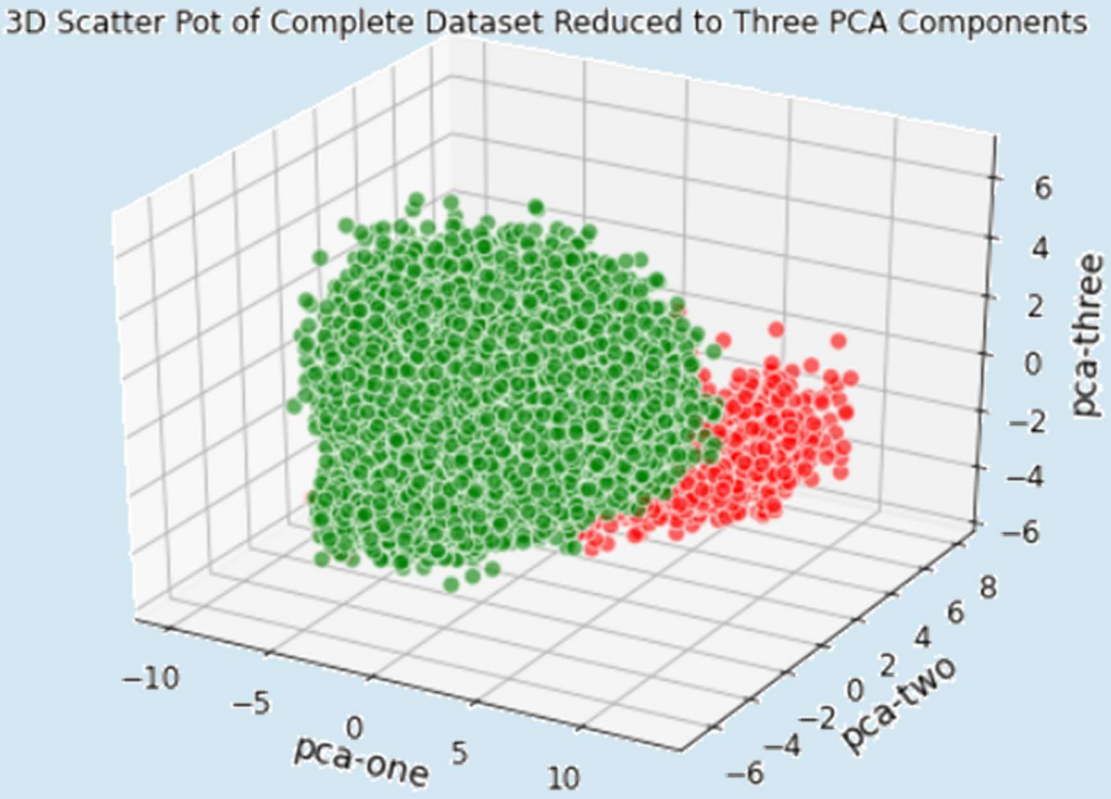
Analysis of Complete Dataset: Reducing all Attributes to 3 Dimensions Using PCA

Principal Component Analysis

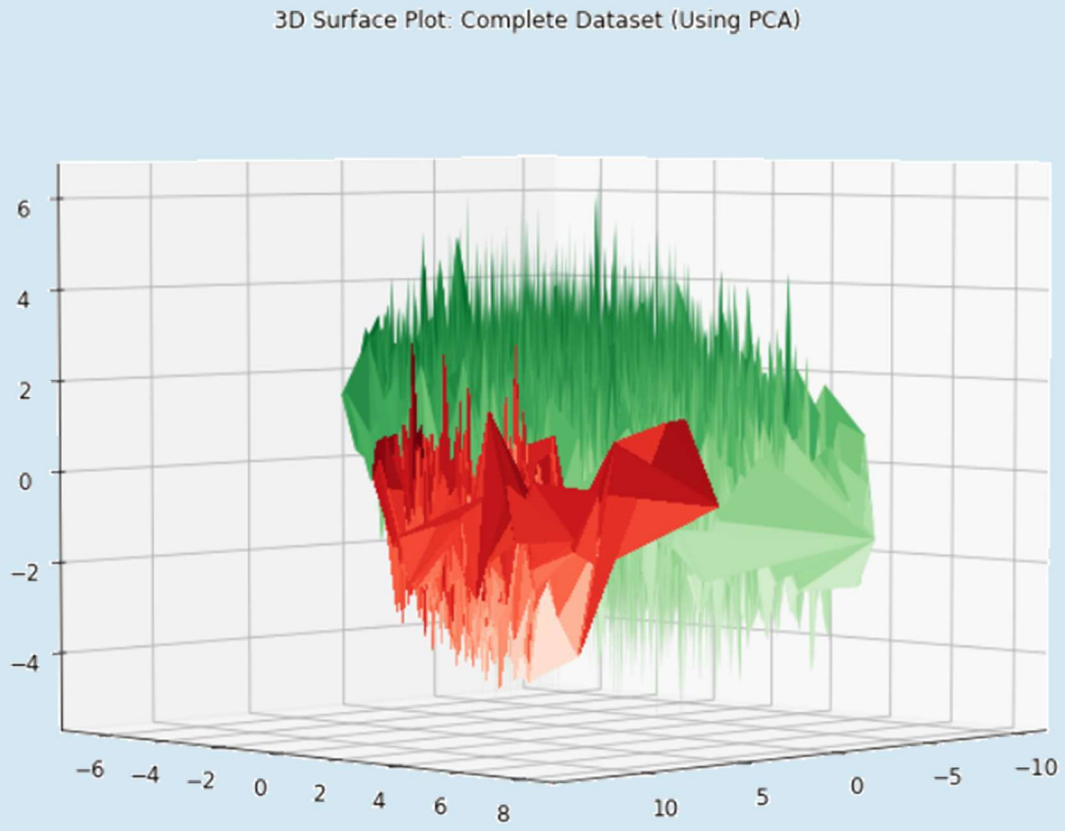
To reduce the number of dimensions (features) in a dataset



Analysis of Complete Dataset: Reducing all Attributes to 3 Dimensions Using PCA



Analysis of Complete Dataset: Reducing all Attributes to 3 Dimensions Using PCA



Feature Engineering for Dataset #1

1. url -> url profanity score.
   Eg http://www.badWord-vulgarity.com -> 0.98
   - profanity_check library (internal model is a linear SVM) (Zhou, 2019)[1]
2. ip_add -> ip_add_splits[4]. Eg 192.168.1.254 -> [192, 168, 1, 254]
   - Split IP addresses perform better than binary or one-hot encoding (Shao, 2019)[2]
3. Binary encoding for who_is and https. Eg “yes” -> 1, “no” -> 0
4. Ordinal encoding for geo_loc and tld.
   Eg if the column has categorical values [“a”, “b”, “c”], a row with a value “b” is mapped to 2.

1 – profanity-check 1.0.3 library on pypi. 2 – Encoding IP Adresss as a Feature for Network Intrusion Detection, Purdue University

Feature Engineering for Dataset #2

More challenging as only have 1 feature: raw URL string addresses

1. Scrap for WHOIS data: whois_exists (binary), days_since_creation, days_since_last_update, days_until_expiration
2. Lengths of: URL, path, hostname, first directory, subdomains. Eg "https://www.aprilsims.wordpress.com/2010/02/11/from-ghetto-to-greatness-by-kevin-brown/" has URL length: 87, path length: 52, hostname length: 27, first directory length: 4, subdomains: 5
3. Count of characters and strings: - @ ? % . + https www
4. Counts of digits and letters
5. Count of parameters (&) and fragments (#)
6. Use regex to determine if IP address is inside the URL. Eg: https://www.bgp.he.net/net/67.214.112.0/20
7. Use regex to determine if shortening services were used. Eg: http://bit.ly/e5w45f444444fff

Feature Engineering for Dataset #2

More challenging as only have 1 feature: raw URL string addresses

8. Calculate the entropy for each URL. Eg:

Low entropy:

```
http://pu-pa.eu/counter/?id=&amp;rnd=01
```

High entropy:

```
https://a.net-bz86.stream/claim/ytvt/index-2o-en-azwal-c1-2m.html?region=New%20York&amp;td=www
.bluecruisebooking.com&amp;brand=Samsung&amp;model=Galaxy%20S8&amp;cep=TToRcA-H4tzmqUHx1PS1ILy
AGNbsnrhwu3h5rknGclsYLc5qvY1EWJfRJ3y3qQbEEcT9vpRj0nhv7Bv01tWDMdCV2R8BPVLY9f-VcKOjB6HKDADURebde
fRsoM4kOCsblnWLvjyTQxjZH0xihj6tHZ5ySn4Jm9KAMlufXBQXKUPiGhjuRLQ39OjrmeYjySmGm9nUlvgAY8od67J69tf
z-EiYnNcBcsQ0yaP4MsguxKcOaWo6X74cS4FTX3FfnNsgC5ybBfYIbQ3MZH5FPt9cz64Qvqoffze9DXs2kOqEIwPeziz1x
3XAGFLNfVfTiFgJUPsSK69E9InRT_jir6W6S6t9Pv2pA2RK-CAbkhVFMDuHYtauXL8S2ZYbRSAG8Uc8lBdTZkJD533NfGL
qLuNHheu2QtiRGj4-VoqUccyU5Y8#
```

8. URL profanity score
9. TLD ordinal encodings

Methodology: Machine Learning

1. Split dataset into 2: Training set (80%) and Test set (20%)
2. Perform 5-fold cross-validation (when feasible) on training set to get optimal model parameters (using grid search)
3. With optimal parameters, train on full training set and get performance metrics on test set

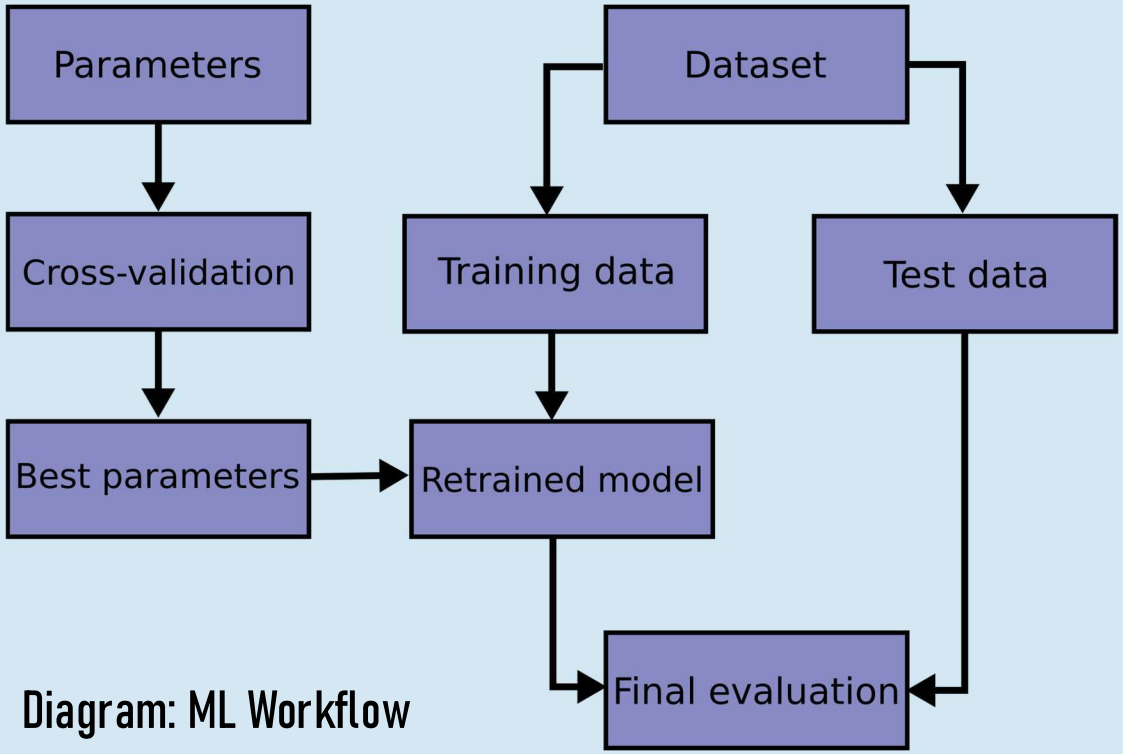


Diagram: ML Workflow

Methodology: Machine Learning

1. Split dataset into 2: Training set (80%) and Test set (20%)
2. Perform 5-fold cross-validation (when feasible) on training set to get optimal model parameters (using grid search)
3. With optimal parameters, train on full training set and get performance metrics on test set

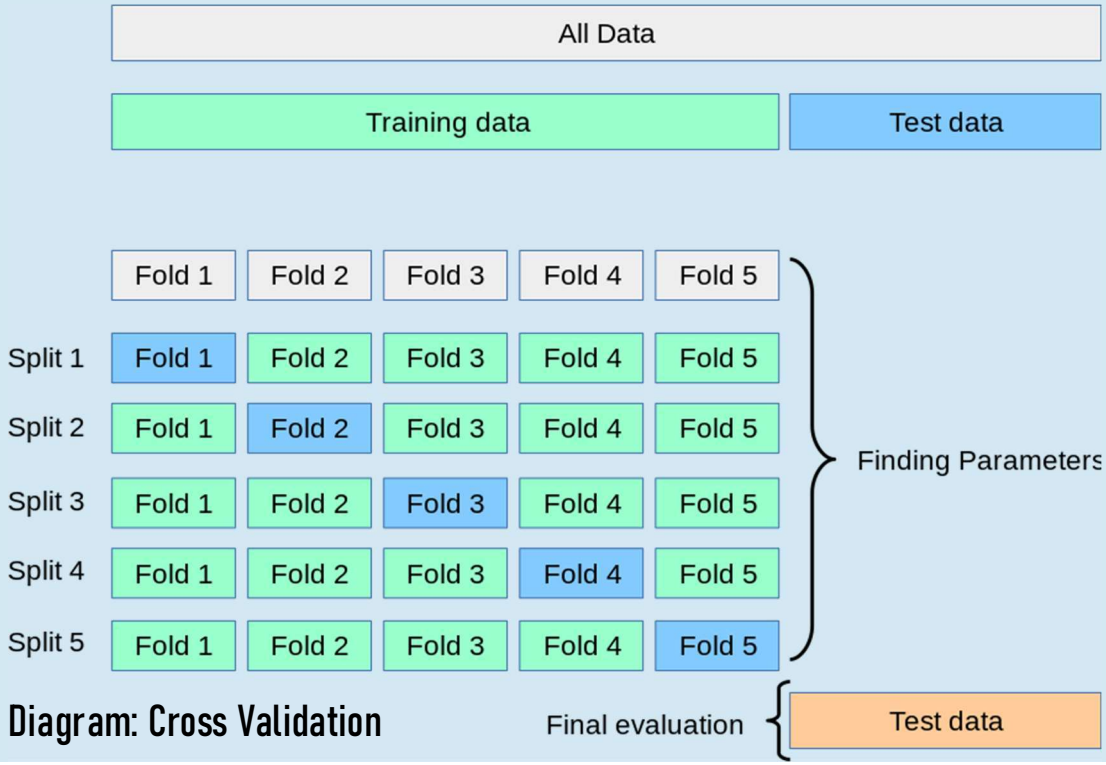
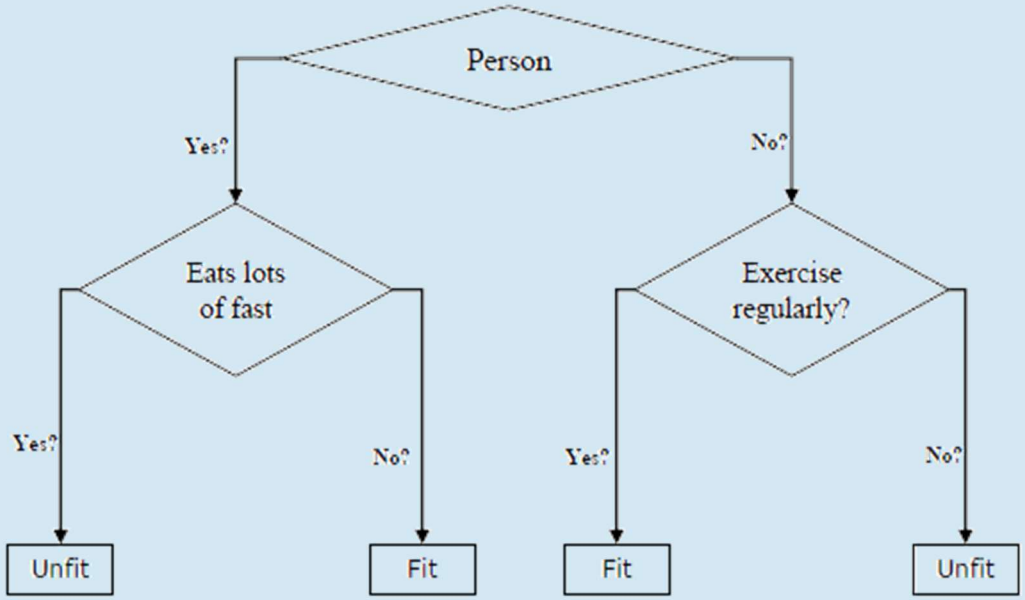


Diagram: Cross Validation

Machine Learning Models Primer: Decision Trees

- Essentially if-else statements used for prediction
- Pros: Easy to visualize and understand
- Cons: Easily overfits (generate too-complex trees). Need to know when to prune. Dependent on data quality.

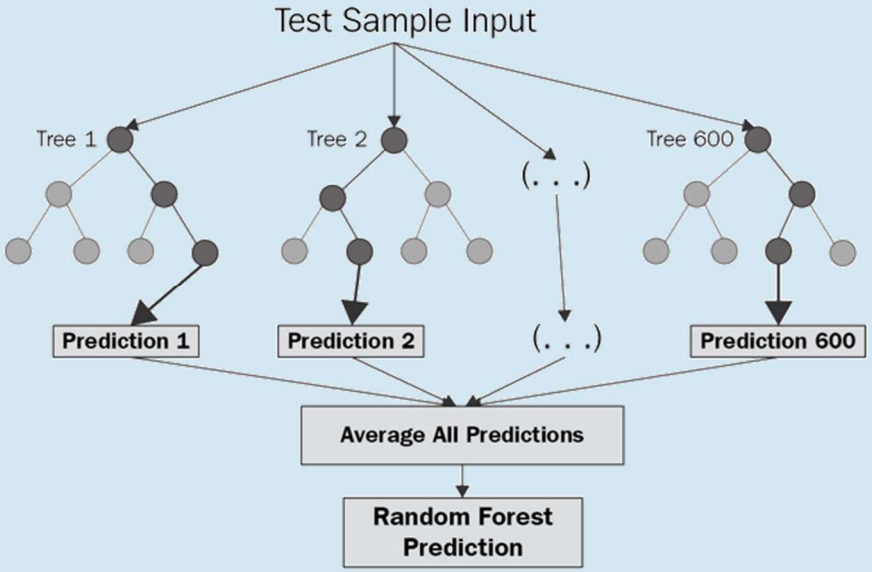
Decision Tree used for Classification



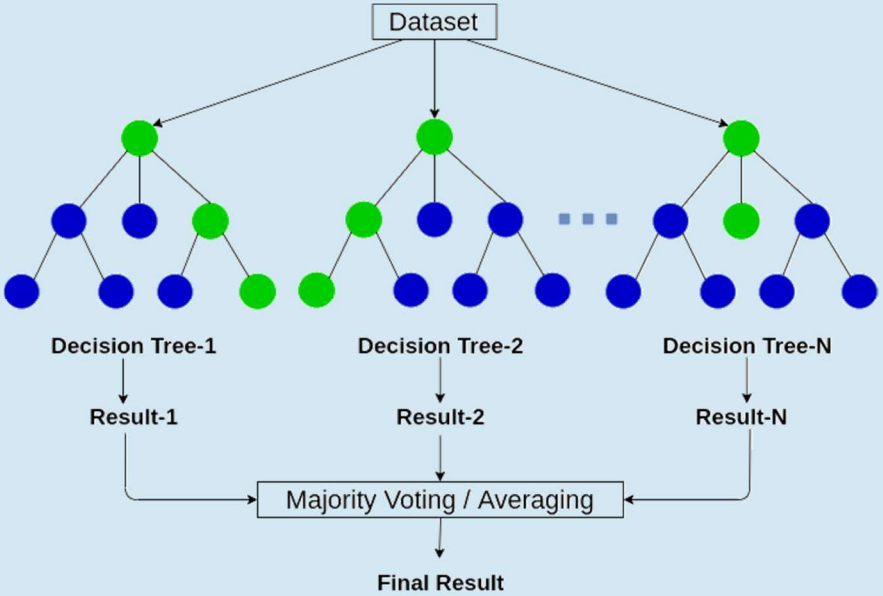
Machine Learning Models Primer: Random Forests

- Consists of a large number of uncorrelated decision trees
- Each individual tree makes a decision
- Overall prediction is based on the majority
- Uses ensemble learning i.e. using multiple models to get diversity

Predicting a real value



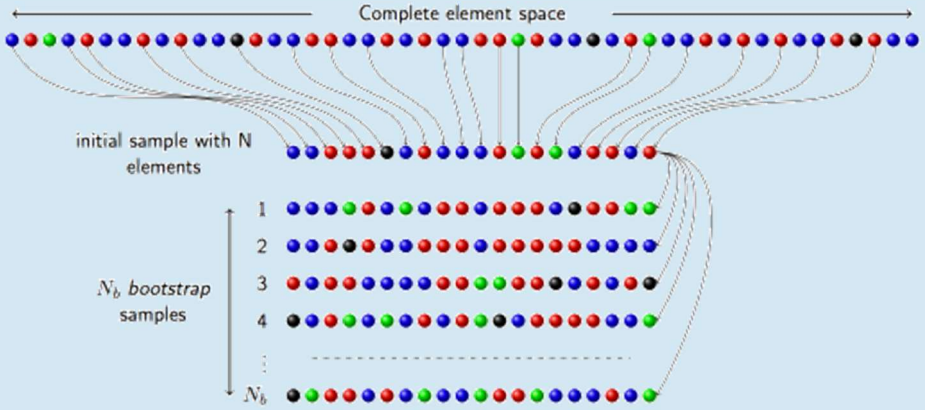
Predicting a class



Machine Learning Models Primer: Random Forests

- How is each tree able to maintain diversity?
- Via bagging (bootstrapping of samples + aggregation of votes) + Feature Randomization
  - Bootstrapping: Resampling technique to produce a dataset with replacement
  - Aggregation: Get a majority vote from each tree
  - Feature randomization: Some trees are restricted to a random subset of features

Bootstrapping Diagram



Pseudocode for Random Forests

For t= 1, 2, … T (total number of trees):

Select a random subset of features

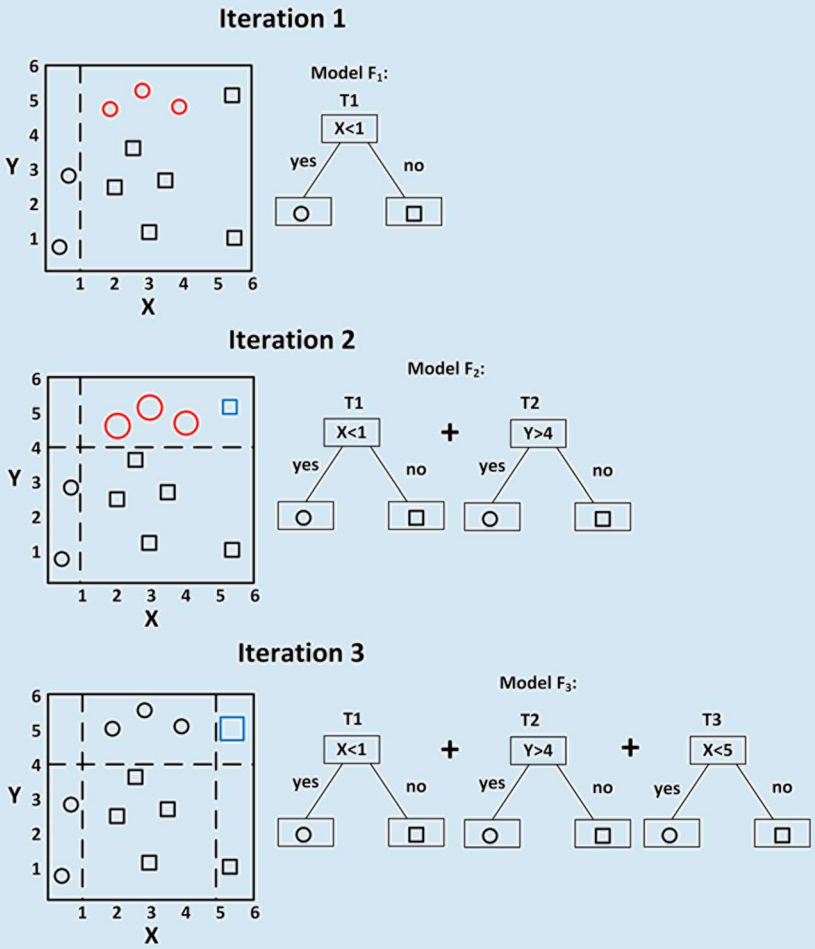
Sample a bootstrap data sample (restricted features)

Obtain a tree

Return Prediction(x) = Aggregated vote of all trees

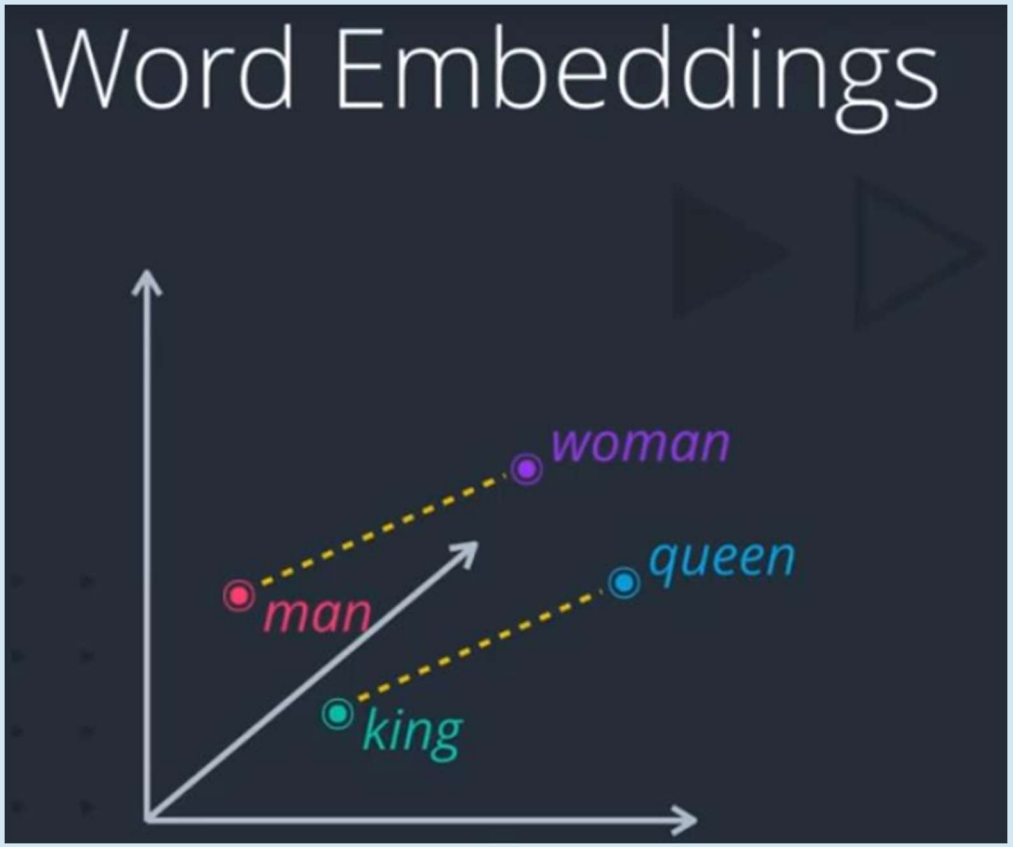
Machine Learning Models Primer: Gradient Boosting Decision Trees

- Also uses multiple tree classifiers
- Focus on the samples that are wrongly classified
- Give these samples more weight



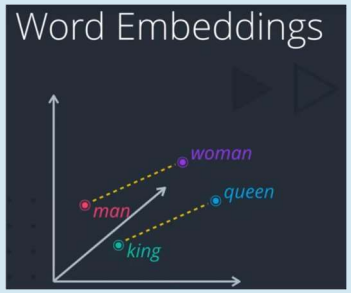
Machine Learning Models Primer: Word Embeddings

In this context (gender), man and king are grouped similarly



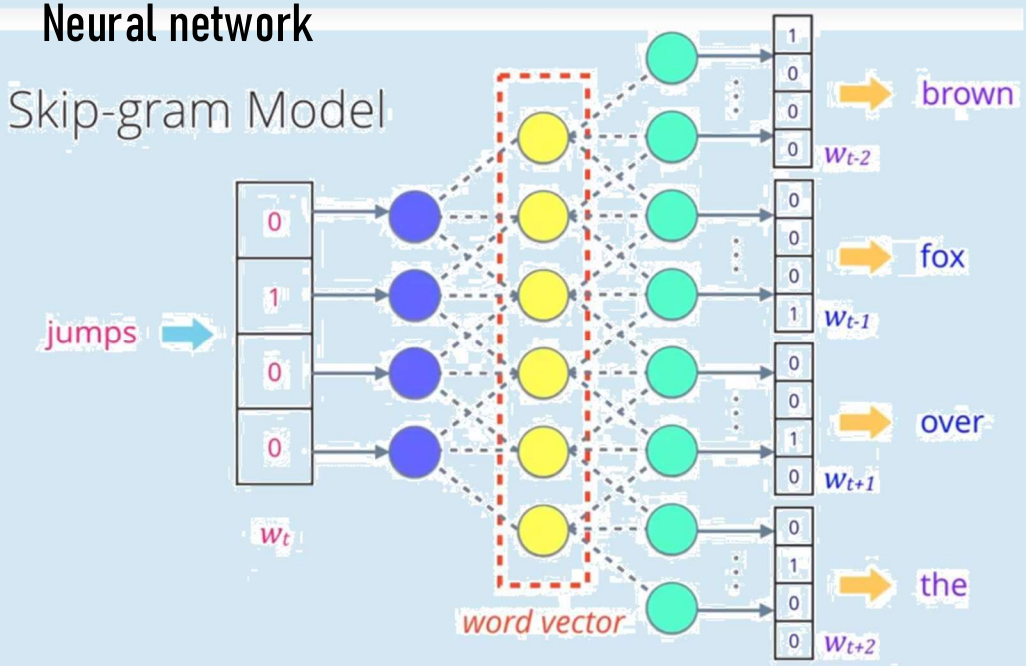
- Used in Natural Language Processing
- Allows words that have similar meanings to have similar representations

Machine Learning Models Primer: Word Embeddings

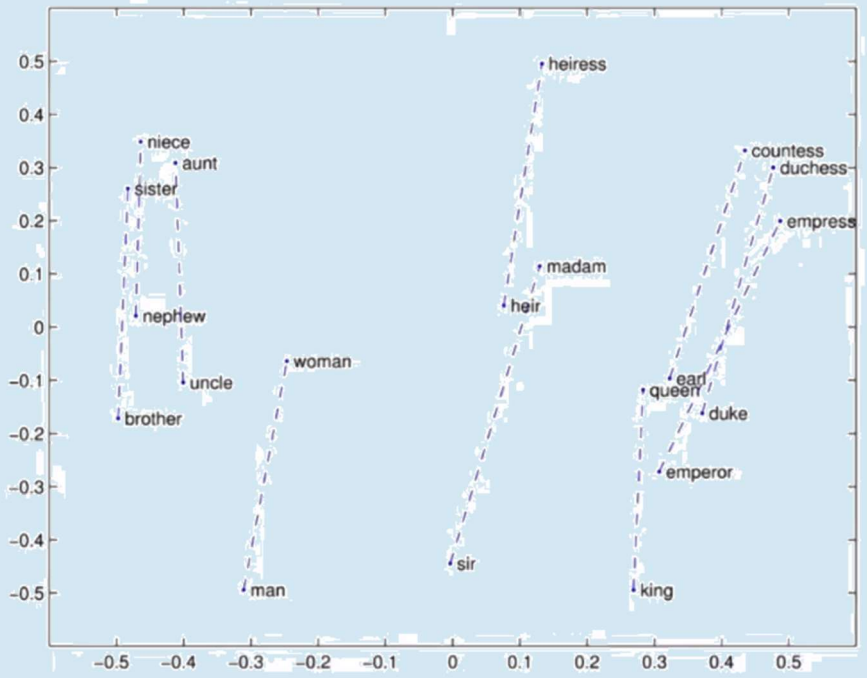


- Allows words that have similar meanings to have similar representations
- Can be learnt via neural networks or other methods (t-SNE)

Embeddings are learnt in the hidden layer Neural network

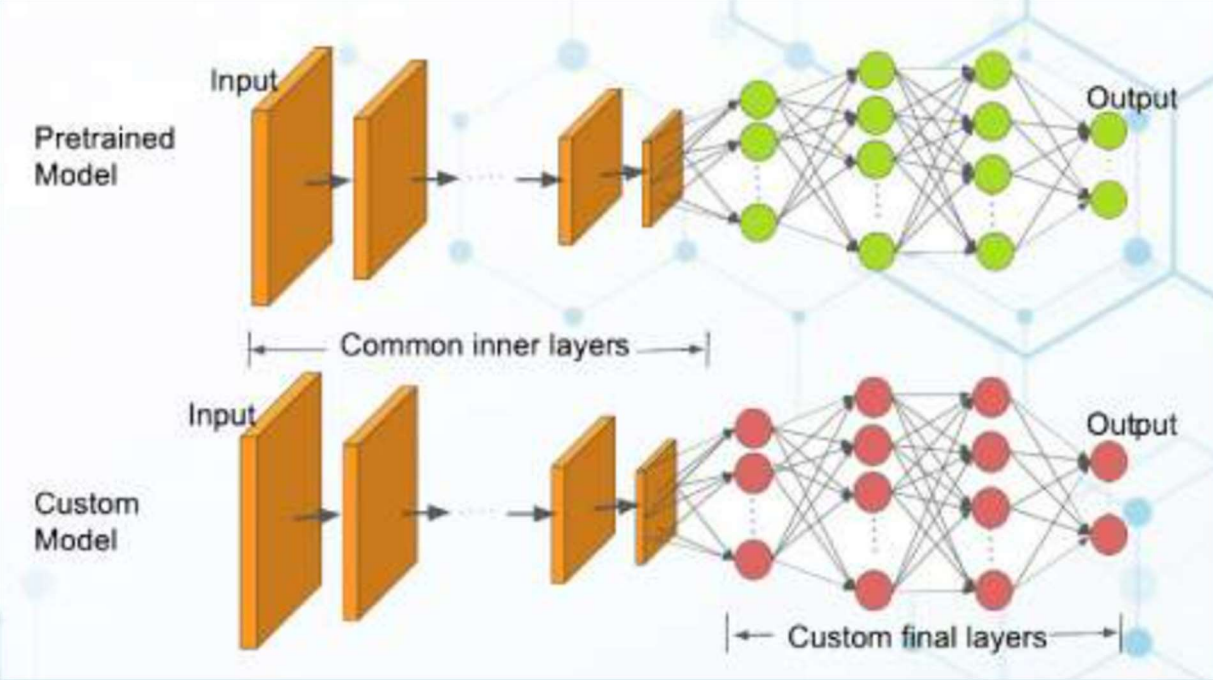


T-Distributed Stochastic Neighbor Embedding



Machine Learning Models Primer: Transfer Learning

- Use pre-trained neural network for our needs
- We only need to train the output layers
- Can leverage on pre-trained word embeddings



Methodology - Machine Learning

Splitting Dataset #1

Dataset #1

10 variables: ip_add, geo_loc, url_len, js_len, Js_obf_len, tld, who_is, https, content

Model #1

1 variable: content strings

Eg: "`decay suggest in 1315.. current constitution...on..}; sin(x) ontimeupdate evaluate operator`"

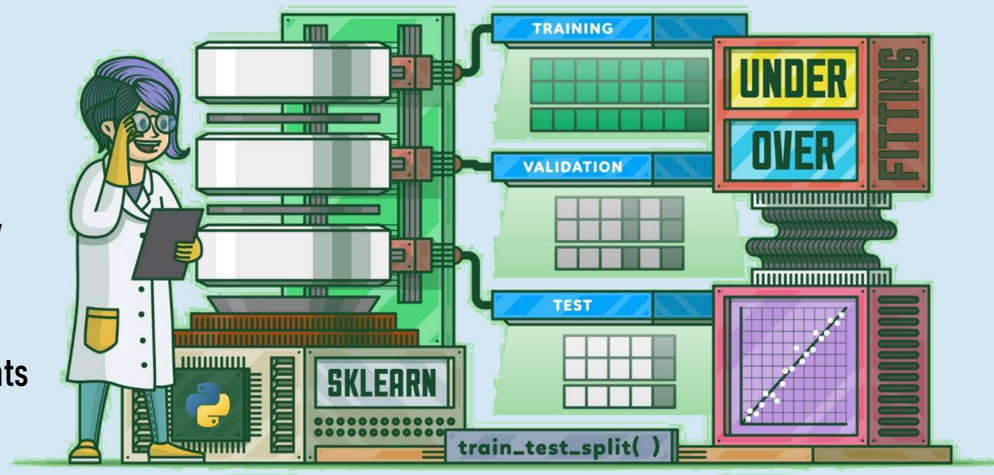
Focuses on Natural Language Processing, Word Embeddings

Model #2

9 variables:
ip_add, geo_loc, url_len,
js_len, Js_obf_len, tld, who_is, https

New features: ip_add_splits, profanity score

Variables can be converted to int/floats
Focus on traditional ML models



Methodology – Machine Learning (Model 1: NLP)

- Uses a pre-trained word embedding (trained on 130GB Google News corpus)

In [5]:
```
# Test pre-trained embeddings
import tensorflow_hub as hub

embed = hub.load("https://tfhub.dev/google/tf2-preview/gnews-swivel-20dim/1")
embeddings = embed(["cat is on the mat", "dog is in the fog"])
embeddings
```

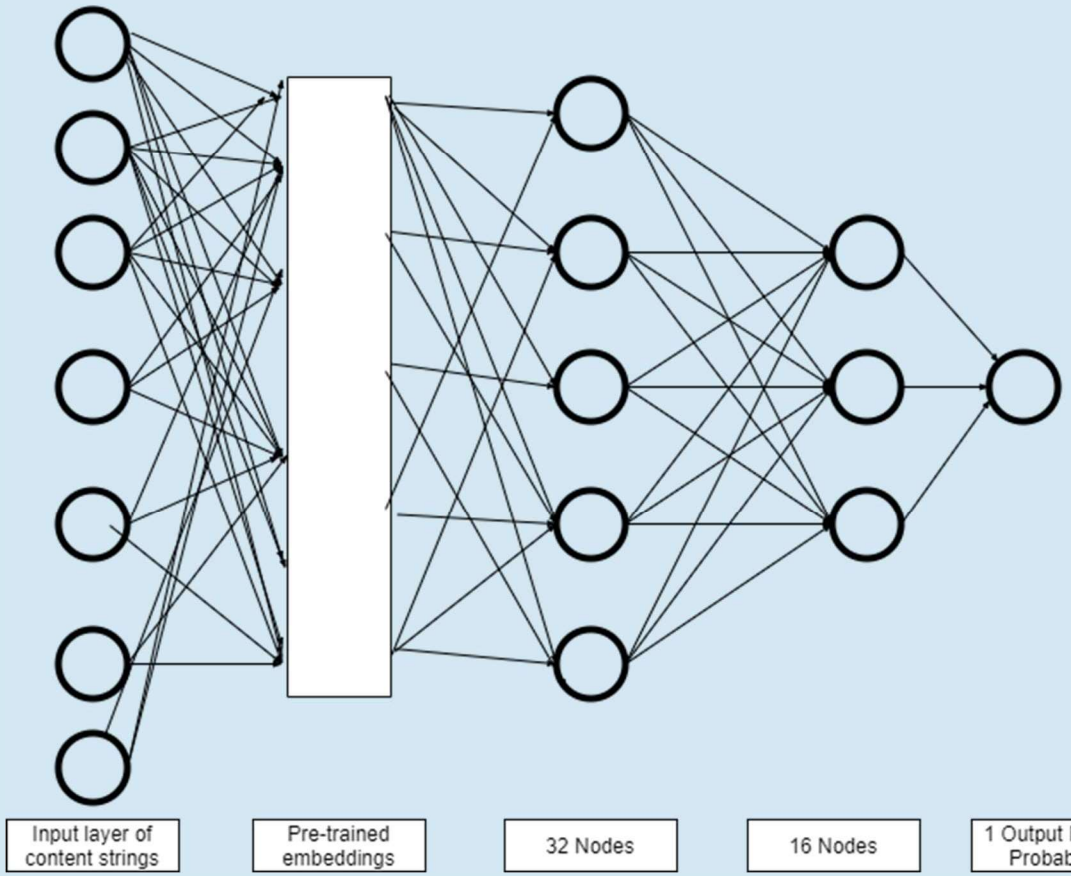
Out[5]:
```
<tf.Tensor: shape=(2, 20), dtype=float32, numpy=
array([[ 0.8666395 ,  0.35917717,  0.00579667,  0.681002  , -0.54226625,
         0.22343189, -0.38796625,  0.62195706,  0.22117122, -0.48538068,
        -1.2674141 ,  0.886369  , -0.32849073, -0.13924702, -0.53327686,
         0.5739708 , -0.05905761,  0.13629246, -1.1718255 , -0.31494334],
       [ 0.9602181 ,  0.62520486,  0.06261905,  0.37425604,  0.24782333,
        -0.39351934, -0.7418429 ,  0.56599647, -0.26197797, -0.69016844,
        -0.76565284,  0.71412426, -0.4537978 , -0.50701594, -0.8499377 ,
         0.8917156 , -0.30278975,  0.2149126 , -1.1098894 , -0.46719775]],
      dtype=float32)>
```

In [11]:
```
print('Difference:',np.sum(embeddings[1] - embeddings[0]))
```

```
Difference: -1.2328589
```

Methodology – Machine Learning (Model 1: NLP)

- Neural network architecture




```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
keras_layer (KerasLayer)     (None, 20)                400020
_________________________________________________________________
dense (Dense)                (None, 32)                672
_________________________________________________________________
dense_1 (Dense)              (None, 16)                528
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 17
=================================================================
Total params: 401,237
Trainable params: 401,237
Non-trainable params: 0
```

Best optimizer: RMSprop

```
Best: 0.998339 using {'optimizer': 'RMSprop'}
0.994321 (0.000263) with: {'optimizer': 'SGD'}
0.998339 (0.000164) with: {'optimizer': 'RMSprop'}
0.987725 (0.007276) with: {'optimizer': 'Adagrad'}
0.997896 (0.000230) with: {'optimizer': 'Adam'}
```

Methodology – Machine Learning (Model 2: Random Forest)

12 features used:

Feature	Variable	Feature	Variable
url_len	Int	profanity_score	Float
geo_loc	Float (encoded)	ip_split_1	Int
tld	Float (encoded)	ip_split_2	Int
who_is	Int (binary)	ip_split_3	Int
https	Int (binary)	ip_split_4	Int
js_len	Float	js_obf_len	Float

Methodology – Machine Learning (Model 2: Random Forest)

Random Forest parameters:

In [67]:
```
# Random Forest
param_grid=[{"n_estimators":[x for x in range(10, 120, 10)],
             "criterion":["gini", "entropy"]}]
forest_grid = GridSearchCV(estimator=RandomForestClassifier(random_state=42),param_grid=param_grid,cv=5)
forest_grid.fit(X_train,y_train)
```

Out[67]:
```
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=42),
             param_grid=[{'criterion': ['gini', 'entropy'],
                          'n_estimators': [10, 20, 30, 40, 50, 60, 70, 80, 90,
                                           100, 110]}])
```

In [68]:
```
print('Random forest best params: ', forest_grid.best_params_)
```

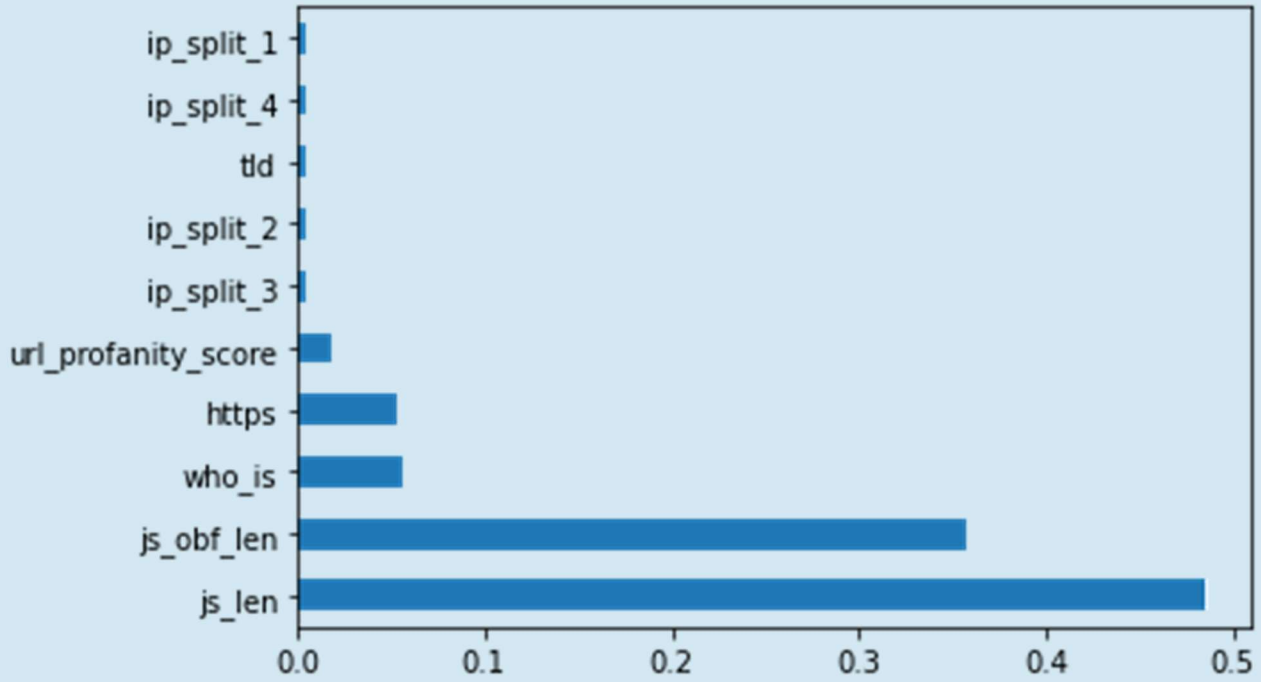
```
Random forest best params:  {'criterion': 'entropy', 'n_estimators': 110}
```

In [69]:
```
print('Random forest accuracy: ', forest_grid.score(X_test,y_test))
```

```
Random forest accuracy:  0.9990467875358491
```

Methodology – Machine Learning (Model 3: Random Forest)

Random Forest feature importance:



Methodology – Machine Learning (Model 3: Gradient Boosting Classifier)

28 features used:

Feature	Variable	Feature	Variable	Feature	Variable	Feature	Variable
whois exist	Int (binary)	path length	Int	url length	Int	hostname length	Int
whois days since creation	Int	whois days since last update	Int	whois days until expiration	Int	first directory length	Int
top level domain	Float (encoding)	top level domain length	Int	count of -	Int	count of @	Int
count of ?	Int	count of %	Int	count of .	Int	count of =	Int
https exist	Int (binary)	count of digits	Int	count of letters	Int	count of directories	Int
entropy	Float	count of parameters	Int	count of fragments	Int	ip address exist	Int (binary)
url shortening used	Int (binary)	url profanity score	Float	count www	Int	count of subdomains	Int

Methodology – Machine Learning (Model 3: Gradient Boosting Classifier)

Gradient Boosting Classifier parameters:

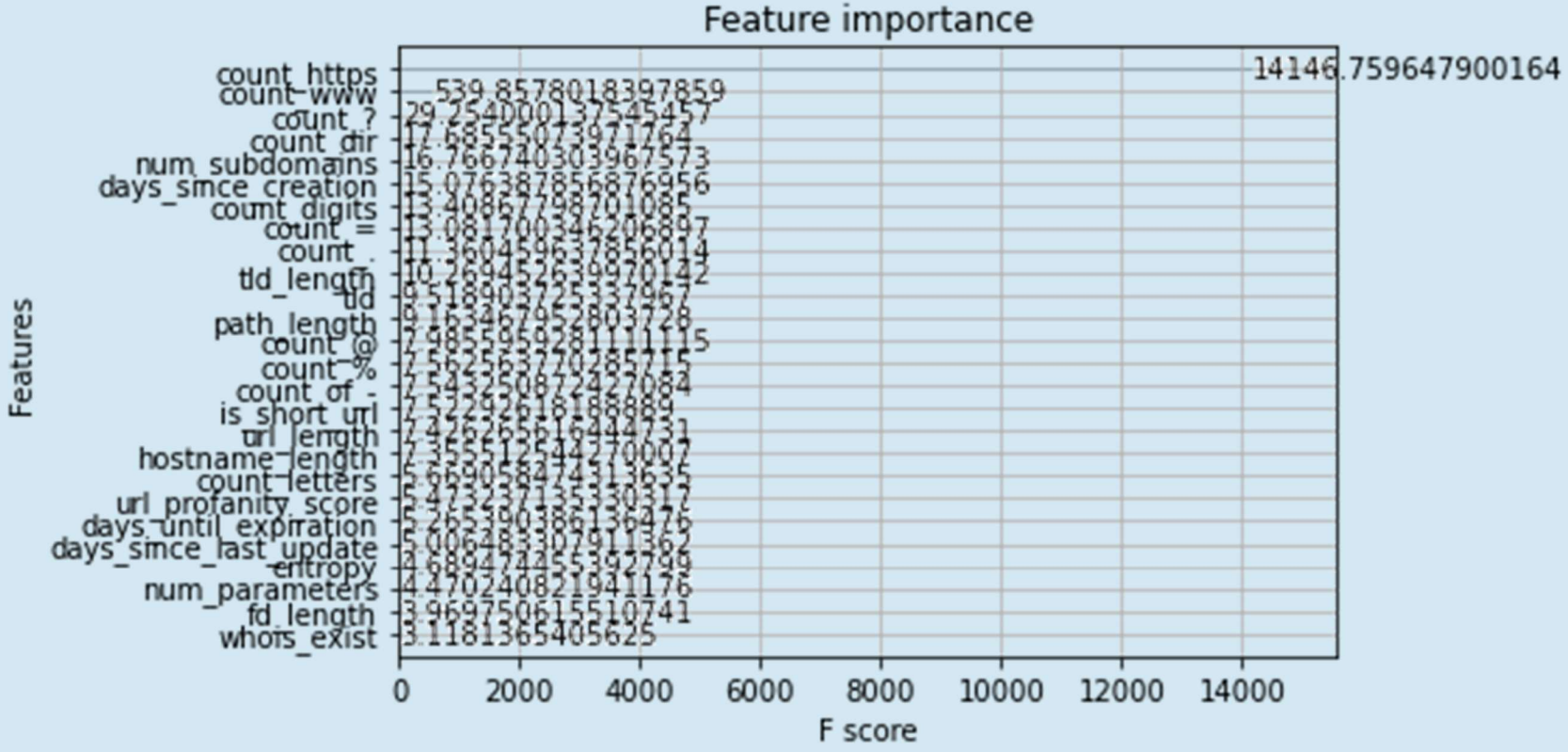
In [68]:
```
xgb_model = XGBClassifier()
params = {
        'min_child_weight': [1, 10],
        'gamma': [0.5, 5],
        'subsample': [0.6, 1.0],
        'max_depth': [3, 5]
        }
xgb_grid = GridSearchCV(estimator=xgb_model, param_grid=params, cv=5)
xgb_grid.fit(x_train,y_train)
```

In [72]:
```
print('Gradient boosting classifier best params: ', xgb_grid.best_params_)
print('Gradient boosting accuracy: ', xgb_grid.score(x_test, y_test))
```

```
Gradient boosting classifier best params:  {'gamma': 0.5, 'max_depth': 5, 'min_child_weight': 1, 'subsample': 1.0}
Gradient boosting accuracy:  0.997834199653472
```

Methodology – Machine Learning (Model 3: Gradient Boosting Classifier)

Gradient Boosting Classifier feature importance:



Methodology – Final Model Deployment

Model	Most important variable	Implication
1: NLP Neural Network	content string of website	Website must be active
2: Random Forest Classifier	JS code variables are most important	Website must be active
3: Gradient Boosting Classifier	'https' string most important	Website need not be active

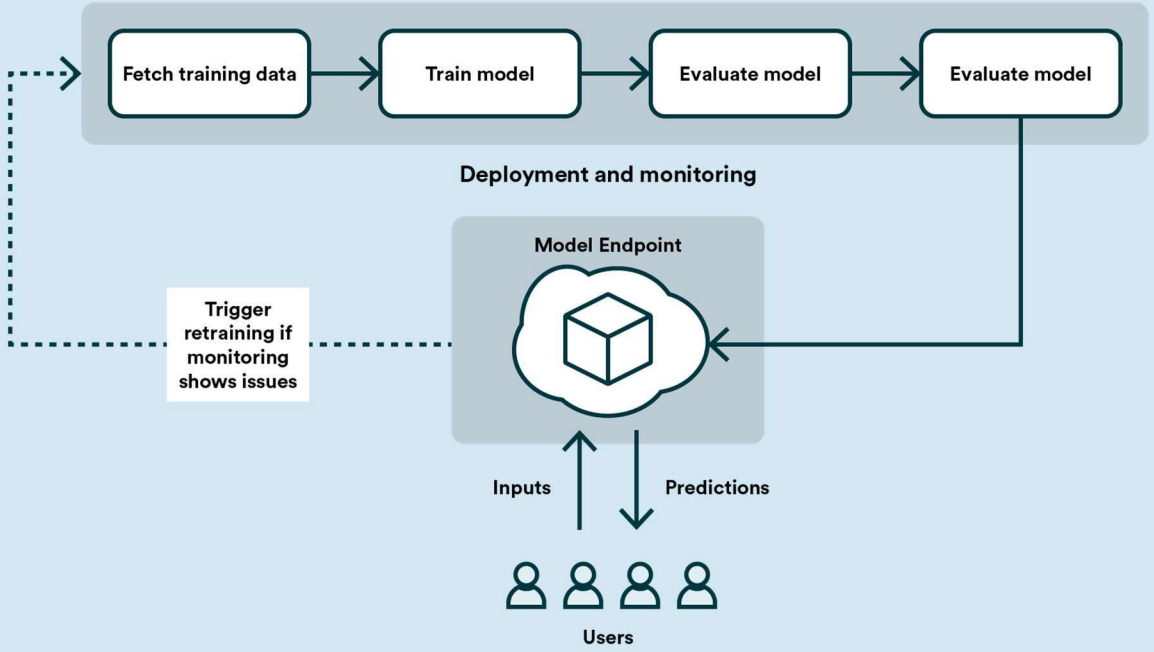
Methodology – Final Model Deployment

If website is active:

Take a majority vote using all 3 models

Elif website is inactive:

Only use model 3 (which has more offline features)



Methodology – Final Model Deployment (Example)

URL	URL Details	Classification	Explanation
google.com	Google's website	Benign	Website did not contain malicious features
http://172.253.118.101/	Google's website but in IP address format	Benign	Website did not contain malicious features
http://102.253.118.101/	HTTP website that does not exist	Malicious	Website was inaccessible. Model 3 classified as malicious due to IP address
https://102.253.118.101/	HTTPS website that does not exist	Malicious	Website was inaccessible. Model 3 classified as malicious due to IP address despite having 'https' in URL

Model Deployment (CLI App)

Input

Script loads ML models

Required Column values extracted from URL input

Values piped into models

Output

```
python prediction_cli.py -w http://www.google.com
```


```
Prediction for http://google.com : benign
```

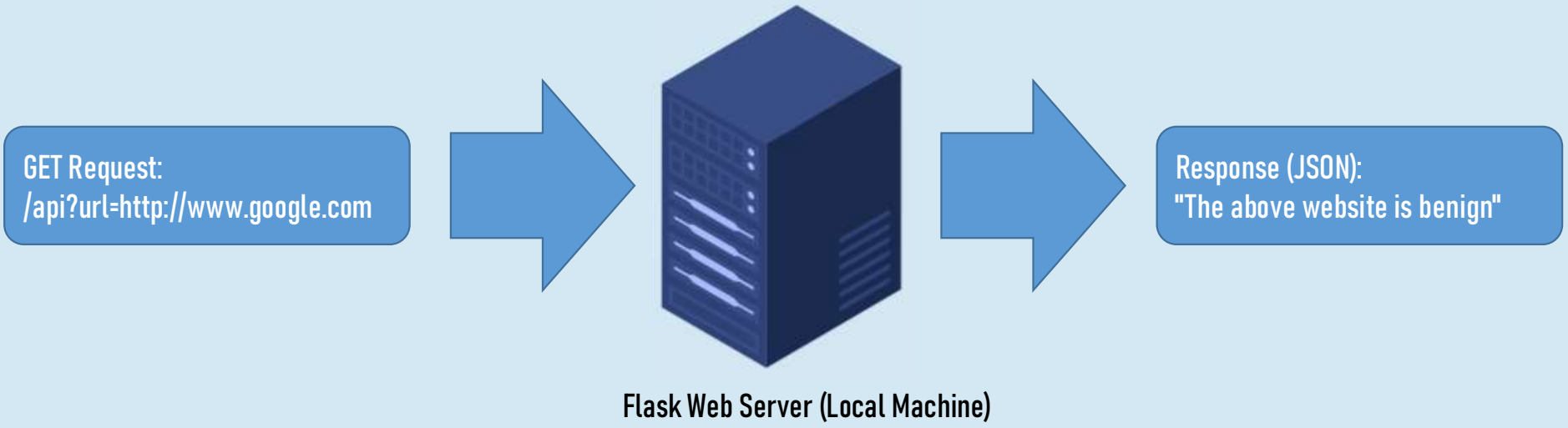
Model Deployment (Flask Web Server)

- Flask
  - Python web framework
    - Set of resources and tools for developers to build and manage web apps
  - Easy to implement, high scalability
  - Supports standard HTML methods (GET, POST)

- Flask Web Server
  - Transition from CLI App to API (Application Programming Interface)
    - Allows 2 applications to talk to each other

Model Deployment (Flask Web Server)

- Adapt CLI App into API using Flask Web Server



Model Deployment (Chrome Extension)

- Browser of choice: Google Chrome
  - Dominating market share (70%)
- Robust framework for extension implementation
  - Support industry standard languages (HTML, Javascript)
  - Easy for developers to upload code to test
- Integrates well with a Flask as back end

chrome

Model Deployment (Chrome Extension)

- Develop Chrome Extension that calls API
- Easy to use
  - Push of a button, no need to type command
- Simple UI that displays result
  - 'Malicious' / 'Benign'

Model Deployment (Chrome Extension)

Chrome Extension UI (popup.html)

Javascript Backend (popup.js): Calls API, display response in UI

Chrome Extension UI with result(popup.html)

Model Deployment (Chrome Extension)

Live Demo

Thank you!

Model Limitation

Class Imbalance Problem

Dataset	Benign	Malicious
Dataset #1	1,526,619 (97.74%)	35,315 (2.26%)
Dataset #2	345,738 (76.80%)	104,438 (23.20%)

Possible Solution: Synthetic Minority Oversampling Technique (SMOTE)

- Oversample (duplicate) the minority class to correct class imbalance
- Given a minority example, generate more samples by looking at neighbors