Final Project – Phase 2

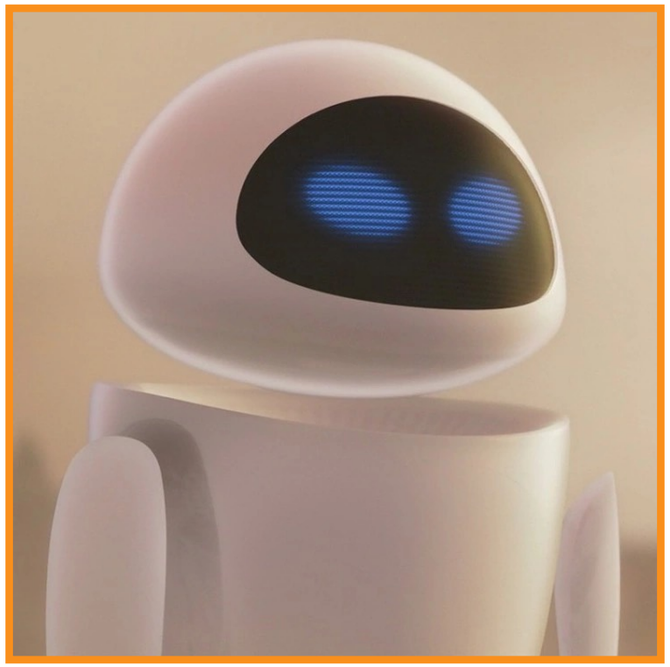
Modeling & Production

Supervisor: Sajjad Ramezani

Mohammad Hashemi – Parastoo Falak Aflaki

What we have done previously

Data Exploration

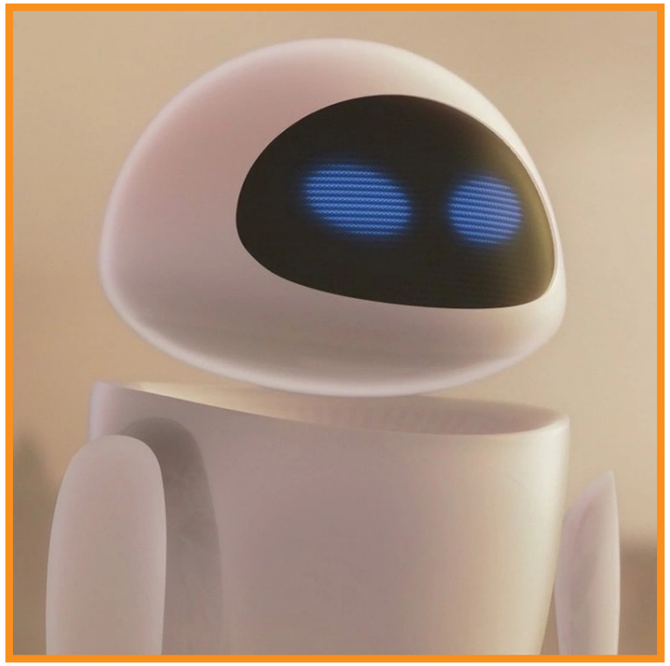


Feature Engineering

Baseline Models

What we have done previously

Data Exploration



Feature Engineering

Baseline Models

Feature Engineering

What We Have,

Vs.

What We Want to Do.

Feature Engineering

Session Identification:

Simply it can be performed by grouping all HTTP requests by their **IP**s and **user agents.**

Types of Features:

Per Request

Per Session

Features Per Session

Feature Engineering

1. Click rate – Higher click rate can only be achieved by an automated script.

```
user_df.sort_values("request_count", ascending=False)
```

ip	user_agent	requests_count
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	23912
207.213.207.102	**Googlebot-Image/1.0**	23627
207.213.207.116	**Googlebot-Image/1.0**	23380
207.213.207.130	**Googlebot-Image/1.0**	21494
207.213.207.144	**Googlebot-Image/1.0**	15363
...	**...**	...
35.202.69.199	**Mozilla/5.0 (compatible; heritrix/3.4.0-20200304 +https://zarebin.ir/)**	1
35.202.76.221	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	1
	okhttp/3.12.1	1
35.202.77.168	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	1
99.96.90.10	**Mozilla/5.0 (iPhone; CPU iPhone OS 14_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.1.1 Mobile/15E148 Safari/604.1**	1

51123 rows × 1 columns

Feature Engineering

2. STD of path’s depth – deeper requests usually indicates a human user.

```
user_df[user_df['requests_count'] > 5].sort_values(‘path_length_std', ascending=True)
```

ip	user_agent	requests_count	path_length_std
1.81.122.235	**Mozilla/5.0 (iPhone; CPU iPhone OS 14_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.1.1 Mobile/15E148 Safari/604.1**	7	0.000000
35.47.50.38	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	11	0.000000
35.47.49.41	**okhttp/3.12.1**	14	0.000000
35.47.49.38	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	25	0.000000
35.47.45.107	**okhttp/3.12.1**	6	0.000000
...	**...**	...	...
35.202.60.172	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A307FN) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/12.1 Chrome/79.0.3945.136 Mobile Safari/537.36**	6	1.366260
35.109.94.99	**Mozilla/5.0 (iPhone; CPU iPhone OS 10_2 like Mac OS X) AppleWebKit/602.1.50 (KHTML, like Gecko) GSA/68.0.234683655 Mobile/14C92 Safari/602.1**	6	1.366260
35.202.142.86	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A115F) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/13.2 Chrome/83.0.4103.106 Mobile Safari/537.36**	6	1.366260
127.227.58.62	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A107F) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/14.0 Chrome/87.0.4280.141 Mobile Safari/537.36**	7	1.463850
14.226.145.71	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.101 Safari/537.36**	15	2.153624

29956 rows × 2 columns

Feature Engineering

3. Percentage of 4xx status codes – Usually higher for crawlers as there is higher chances of hitting an outdated or deleted pages.

```
user_df[user_df['requests_count'] > 5].sort_values(['4xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)
35.124.193.182	**Dalvik/2.1.0 (Linux; U; Android 10; SM-A115F Build/QP1A.190711.020)**	104	0.00000	0.0	100.0
153.126.209.239	**Go-http-client/1.1**	35	0.00000	0.0	100.0
35.244.120.44	**Go-http-client/1.1**	18	0.00000	0.0	100.0
35.232.97.81	**okhttp/2.5.0**	9	0.00000	0.0	100.0
35.132.136.207	**MobileSafari/604.1 CFNetwork/1240.0.4 Darwin/20.5.0**	8	0.00000	0.0	100.0
...	**...**	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.00000	0.0	0.0
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.00000	0.0	0.0

29956 rows × 4 columns

Feature Engineering

4. Percentage of 3xx status codes.

```
user_df[user_df['requests_count'] > 5].sort_values(['3xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)
207.213.193.213	**Googlebot-Image/1.0**	71	0.257679	100.0	0.0
217.98.85.154	**Mozilla/5.0 (Linux; Android 10; SM-A105F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.88 Mobile Safari/537.36**	54	0.292582	100.0	0.0
14.240.9.74	**Mozilla/5.0 (Linux; Android 8.1.0; DUB-LX1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.181 Mobile Safari/537.36**	50	0.274048	100.0	0.0
35.108.36.67	**Mozilla/5.0 (Linux; Android 8.1.0; SM-G610F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.51 Mobile Safari/537.36**	39	0.269953	100.0	0.0
4.115.196.73	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	39	0.269953	100.0	0.0
...	**...**	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.000000	0.0	0.0
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.000000	0.0	0.0

29956 rows × 4 columns

Feature Engineering

4. Percentage of HTTP HEAD requests – most crawlers, in order to reduce the amount of data requested from a site, employ the HEAD method when requesting a page.

```
user_df[user_df['requests_count'] > 5].sort_values(['HEAD_count(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)
20.163.161.41	**Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X; en-us) AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile/11A465 Safari/9537.53**	37	0.538321	0.000000	0.000000	72.972973
36.67.23.210	**Go-http-client/2.0**	7582	0.475879	9.773147	1.609074	45.805856
60.148.0.167	**Go-http-client/2.0**	7351	0.479779	10.053054	1.714053	43.980411
20.92.247.170	**Go-http-client/2.0**	7273	0.482874	10.325863	0.000000	42.334662
76.212.164.3	**Go-http-client/2.0**	6549	0.487955	9.406016	1.878149	42.113300
...	**...**	...	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.000000	0.000000	0.000000	0.000000
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.000000	0.000000	0.000000	0.000000

29956 rows × 5 columns

Feature Engineering

5. Percentage of image requests – web crawlers usually ignore images.

```
user_df[user_df['requests_count'] > 20].sort_values(['image_count(%)', 'requests_count'
                                                    ,ascending=True).head(10)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)
102.29.29.19	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.000000	0.0	0.0	0.0	0.0
113.11.38.30	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.749603	0.0	0.0	0.0	0.0
113.111.195.219	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.000000	0.0	0.0	0.0	0.0
113.118.175.102	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.118.45.106	**okhttp/3.12.1**	21	0.436436	0.0	0.0	0.0	0.0
113.118.90.28	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.118.96.94	**okhttp/3.12.1**	21	0.358569	0.0	0.0	0.0	0.0
113.74.79.241	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.97.91.168	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.643650	0.0	0.0	0.0	0.0
127.166.17.33	**okhttp/3.12.1**	21	0.218218	0.0	0.0	0.0	0.0

Feature Engineering

6. Average & sum of response_length & response_time –

Human users retrieve info from the web via browser, so it forces the user’s session to request additional resource automatically.

```
user_df[(user_df['requests_count'] > 5) & (user_df['3xx_percentage(%)'] < 20)] \
                            .sort_values(['mean_response_time', 'mean_response_length'],
                            ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)	total_resp
29.240.244.96	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_1) AppleWebKit/601.2.4 (KHTML, like Gecko) Version/9.0.1 Safari/601.2.4 facebookexternalhit/1.1 Facebot Twitterbot/1.0	8	0.462910	0.00	75.000000	0.0	25.000000	
93.113.11.166	Mozilla/5.0 (iPhone; CPU iPhone OS 14_4 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0.3 Mobile/15E148 Safari/604.1	7	0.377964	0.00	0.000000	0.0	100.000000	
155.114.254.242	Mozilla/5.0 (iPhone; CPU iPhone OS 13_5_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/13.1.1 Mobile/15E148 Safari/604.1	16	0.250000	6.25	0.000000	0.0	43.750000	

Feature Engineering

7. Set the user agent attributes

browser	os	is_bot	is_pc
Mobile Safari	iOS	False	False
Mobile Safari	iOS	False	False
Samsung Internet	Android	False	False

Feature Engineering

8. Average of time between requests.

```
user_df.sort_values(['avg_time_diff'], ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)	total_response_
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	23912	0.015839	0.000000	0.012546	0.0	0.000000	219622
207.213.207.102	**Googlebot-Image/1.0**	23627	0.215708	99.297414	0.004232	0.0	5.091632	66
207.213.207.116	**Googlebot-Image/1.0**	23380	0.215150	99.328486	0.000000	0.0	5.038494	63
207.213.207.130	**Googlebot-Image/1.0**	21494	0.214493	99.390528	0.004652	0.0	5.038615	66
207.213.193.143	**Googlebot-Image/1.0**	13320	0.337257	98.791291	0.000000	0.0	13.250751	60
...	**...**	...	...	...	...	...	...	
148.197.248.86	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.000000	0.000000	0.000000	0.0	0.000000	
123.4.15.246	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	
217.49.61.35	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	
35.96.149.43	**Mozilla/5.0 (X11; Linux x86_64) app_version: 580**	5	0.894427	0.000000	0.000000	0.0	0.000000	
35.96.121.32	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	

31541 rows × 16 columns

```
git commit –m ‘feat: new features added’
```

Feature Engineering

9. Total number of “Robots.txt” requests per session. – Web administrators, through the Robots Exclusion Protocol, use a special-format file called robots.txt to indicate to visiting robots which parts of their sites should not be visited by the robot.

```
user_df.sort_values(['robots_txt_reqs'], ascending=False)
```

ip	user_agent	requests_count	path_length_std	4xx_percentage(%)	3xx_percentage(%)	HEAD_count(%)	image_count(%
67.149.194.62	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/92.0.4515.51 Safari/537.36**	61	0.682330	11.47541	22.950820	0.0	36.06557
79.130.18.121	**Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)**	244	0.562914	0.00000	2.868852	0.0	0.00000
226.152.248.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	85	0.723360	0.00000	20.000000	0.0	36.47058
118.151.92.167	**Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)**	11	0.000000	0.00000	45.454545	0.0	0.00000
207.213.207.3	**Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)**	4	0.000000	0.00000	100.000000	0.0	0.00000

Feature Engineering

10. Percentage of consecutive repeated requests per session. – A numerical attribute calculated as the number of repeated requests sent in sequence belonging to the same web directory sent by a user during a session.

Consecutive repeated	non-Consecutive repeated
cdn/profiles/12891213214	images/sanjagh_logo_purpule5.png
cdn/profiles/54151515151	fonts/sanjagh_icon_font_5.woff
cdn/profiles/6847463522	images/graystar_min.png

Feature Engineering

```
user_df.sort_values(['consq_rep_path_count'], ascending=False)
```

ip	user_agent	requests_count	path_length_std	4xx_percentage(%)	3xx_percentage(%)	HEAD_count(%)	image_count(%)
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	23912	0.015839	0.012546	0.000000	0.000000	0.000000
36.67.23.210	**Go-http-client/2.0**	7582	0.475879	1.609074	9.773147	45.805856	0.000000
60.148.0.167	**Go-http-client/2.0**	7351	0.479779	1.714053	10.053054	43.980411	0.000000
20.92.247.170	**Go-http-client/2.0**	7273	0.482874	0.000000	10.325863	42.334662	0.000000
20.184.209.90	**Go-http-client/2.0**	2412	0.000000	0.000000	0.000000	0.000000	0.000000
76.212.164.3	**Go-http-client/2.0**	6549	0.487955	1.878149	9.406016	42.113300	0.000000
20.116.215.189	**Go-http-client/2.0**	6401	0.488124	0.000000	9.576629	41.696610	0.000000
20.117.146.75	**Go-http-client/2.0**	6067	0.489572	0.000000	10.004945	40.168123	0.000000
93.113.99.115	**Go-http-client/2.0**	5976	0.489754	0.000000	10.090361	39.909639	0.000000
207.213.207.102	**Googlebot-Image/1.0**	23627	0.215708	0.004232	99.297414	0.000000	5.091632
207.213.207.116	**Googlebot-Image/1.0**	23380	0.215150	0.000000	99.328486	0.000000	5.038494
207.213.207.130	**Googlebot-Image/1.0**	21494	0.214493	0.004652	99.390528	0.000000	5.038615
67.204.145.45	**Go-http-client/2.0**	1206	0.000000	0.000000	0.000000	0.000000	0.000000
209.231.83.221	**Go-http-client/2.0**	1206	0.000000	0.000000	0.000000	0.000000	0.000000
207.213.207.144	**Googlebot-Image/1.0**	15363	0.208714	0.000000	99.401159	0.000000	4.745167
207.213.193.143	**Googlebot-Image/1.0**	13320	0.337257	0.000000	98.791291	0.000000	13.250751
245.191.228.92	**FreshpingBot/1.0 (+https://freshping.io/)**	599	0.000000	0.000000	0.000000	0.000000	0.000000
109.165.215.198	**FreshpingBot/1.0 (+https://freshping.io/)**	598	0.000000	0.000000	0.000000	0.000000	0.000000
238.192.152.0	**FreshpingBot/1.0 (+https://freshping.io/)**	598	0.000000	0.000000	0.000000	0.000000	0.000000

ip	user_agent	requests_count	path_length_std	4xx_percentage(%)	3xx_percentage(%)	HEAD_count(%)	image_count(%)
20.92.247.146	sentry/21.4.1 (https://sentry.io)	23912	0.015839	0.012546	0.000000	0.000000	0.000000
36.67.23.210	Go-http-client/2.0	7582	0.475879	1.609074	9.773147	45.805856	0.000000
60.148.0.167	Go-http-client/2.0	7351	0.479779	1.714053	10.053054	43.980411	0.000000
20.92.247.170	Go-http-client/2.0	7273	0.482874	0.000000	10.325863	42.334662	0.000000
20.184.209.90	Go-http-client/2.0	2412	0.000000	0.000000	0.000000	0.000000	0.000000
76.212.164.3	Go-http-client/2.0	6549	0.487955	1.878149	9.406016	42.113300	0.000000
20.116.215.189	Go-http-client/2.0	6401	0.488124	0.000000	9.576629	41.696610	0.000000
20.117.146.75	Go-http-client/2.0	6067	0.489572	0.000000	10.004945	40.168123	0.000000
93.113.99.115	Go-http-client/2.0	5976	0.489754	0.000000	10.090361	39.909639	0.000000
207.213.207.102	Googlebot-Image/1.0	23627	0.215708	0.004232	99.297414	0.000000	5.091632
207.213.207.116	Googlebot-Image/1.0	23380	0.215150	0.000000	99.328486	0.000000	5.038494
207.213.207.130	Googlebot-Image/1.0	21494	0.214493	0.004652	99.390528	0.000000	5.038615
67.204.145.45	Go-http-client/2.0	1206	0.000000	0.000000	0.000000	0.000000	0.000000
209.231.83.221	Go-http-client/2.0	1206	0.000000	0.000000	0.000000	0.000000	0.000000
207.213.207.144	Googlebot-Image/1.0	15363	0.208714	0.000000	99.401159	0.000000	4.745167
207.213.193.143	Googlebot-Image/1.0	13320	0.337257	0.000000	98.791291	0.000000	13.250751
245.191.228.92	FreshpingBot/1.0 (+https://freshping.io/)	599	0.000000	0.000000	0.000000	0.000000	0.000000
109.165.215.198	FreshpingBot/1.0 (+https://freshping.io/)	598	0.000000	0.000000	0.000000	0.000000	0.000000
238.192.152.0	FreshpingBot/1.0 (+https://freshping.io/)	598	0.000000	0.000000	0.000000	0.000000	0.000000

And now

Modeling

Evaluation

Demo

Outline

Modeling

Evaluation

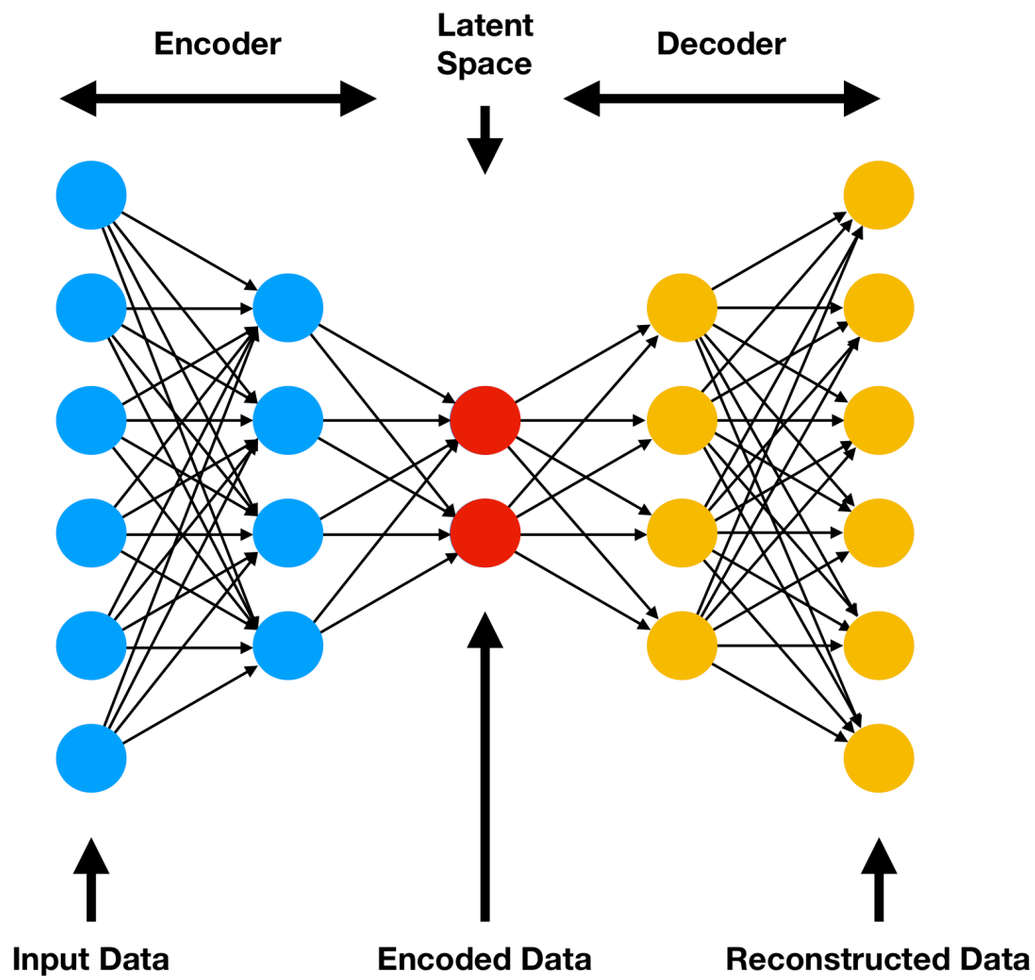
Demo

Modeling

Modeling

1. Auto encoder

It is a neural network architecture capable of discovering structure within data in order to develop a compressed representation of the input. Applications: **anomaly detection**, data denoising and etc.



Modeling

1. Auto encoder – Training configuration

Optimizer	Adam
Loss	MSE
Activation	ReLu
Epochs	20
Batch size	64
Train/test split	80/20

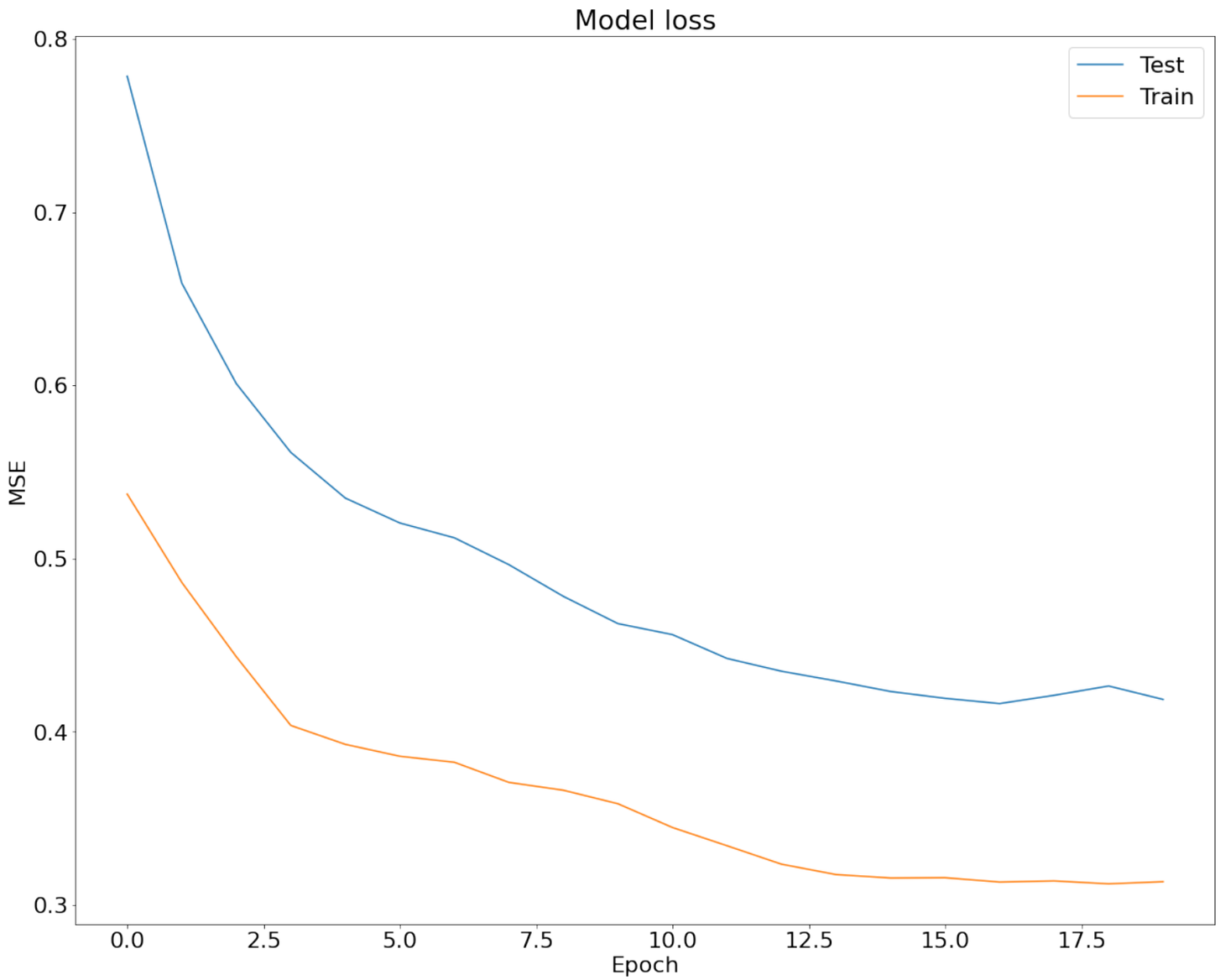
Modeling

1. Auto encoder – Architectures

# Neurons	Train loss	Test loss
[15, 7, 15]	0.42	0.48
[15, 3, 15]	**0.28**	**0.39**
[15, 7, 3, 7, 15]	0.29	0.43
[15, 7, 7, 7, 15]	0.31	0.42

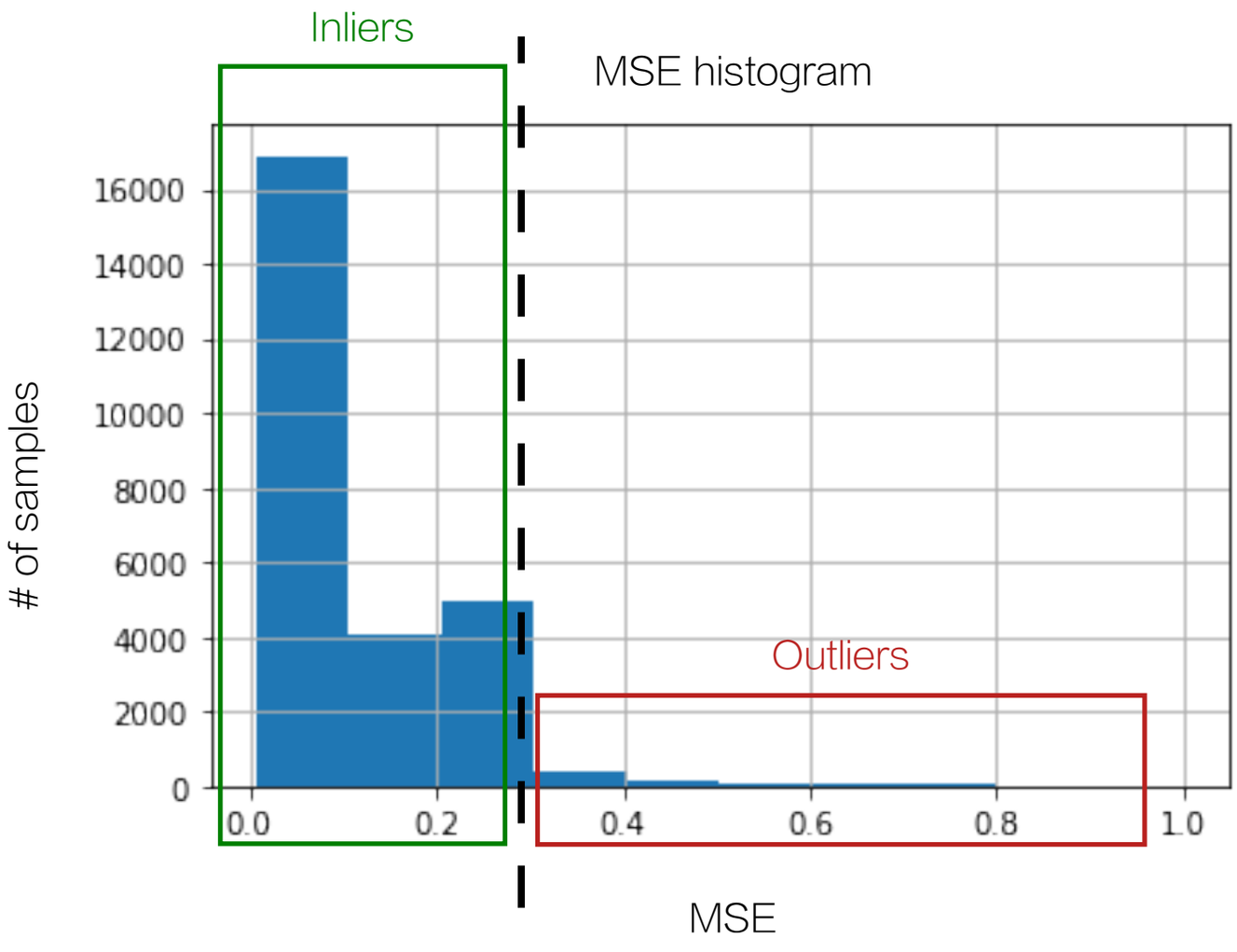
Modeling

1. Auto encoder – Training History



Modeling

1. Auto encoder – Anomaly score



One fact:

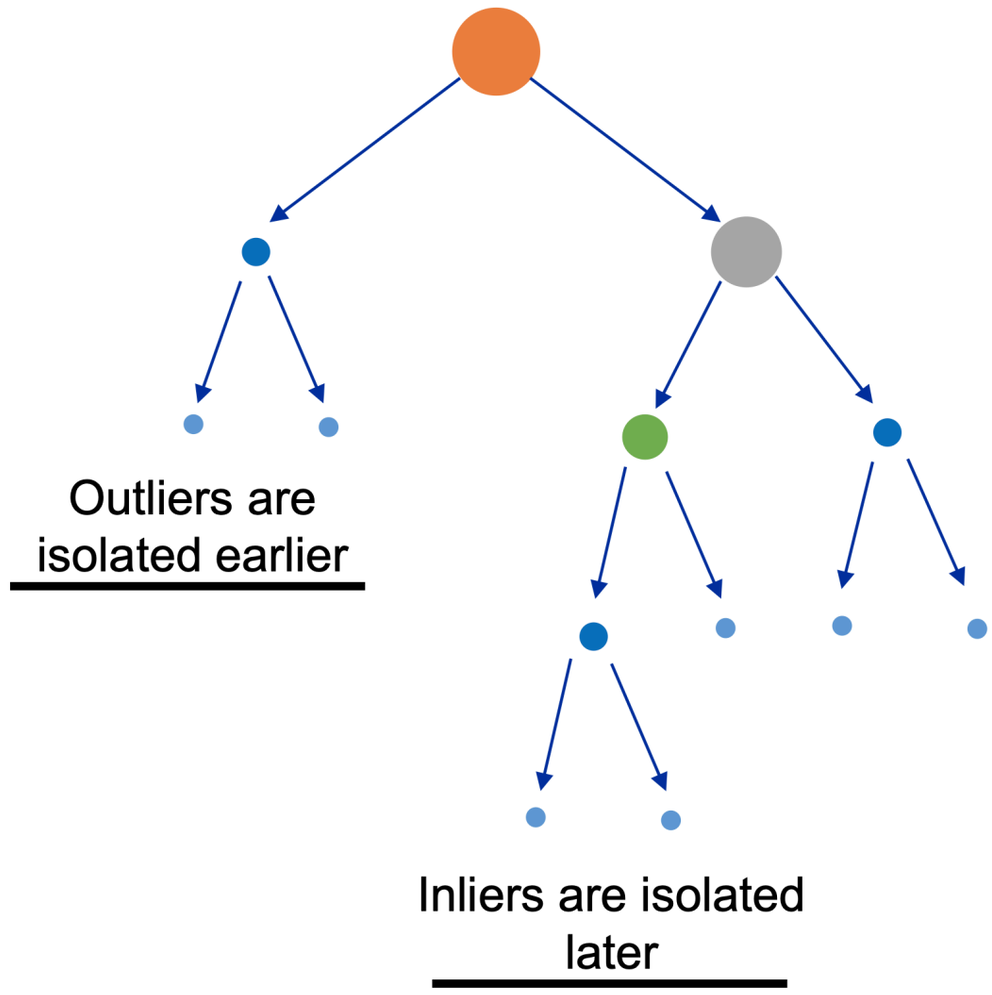
Approximately, **5%** of the dataset contains common bots, crawlers.

MSE threshold = 0.30

2. Isolation Forest

Since outliers have features X that differ significantly from most of the samples, they are isolated earlier in the hierarchy of a decision tree.

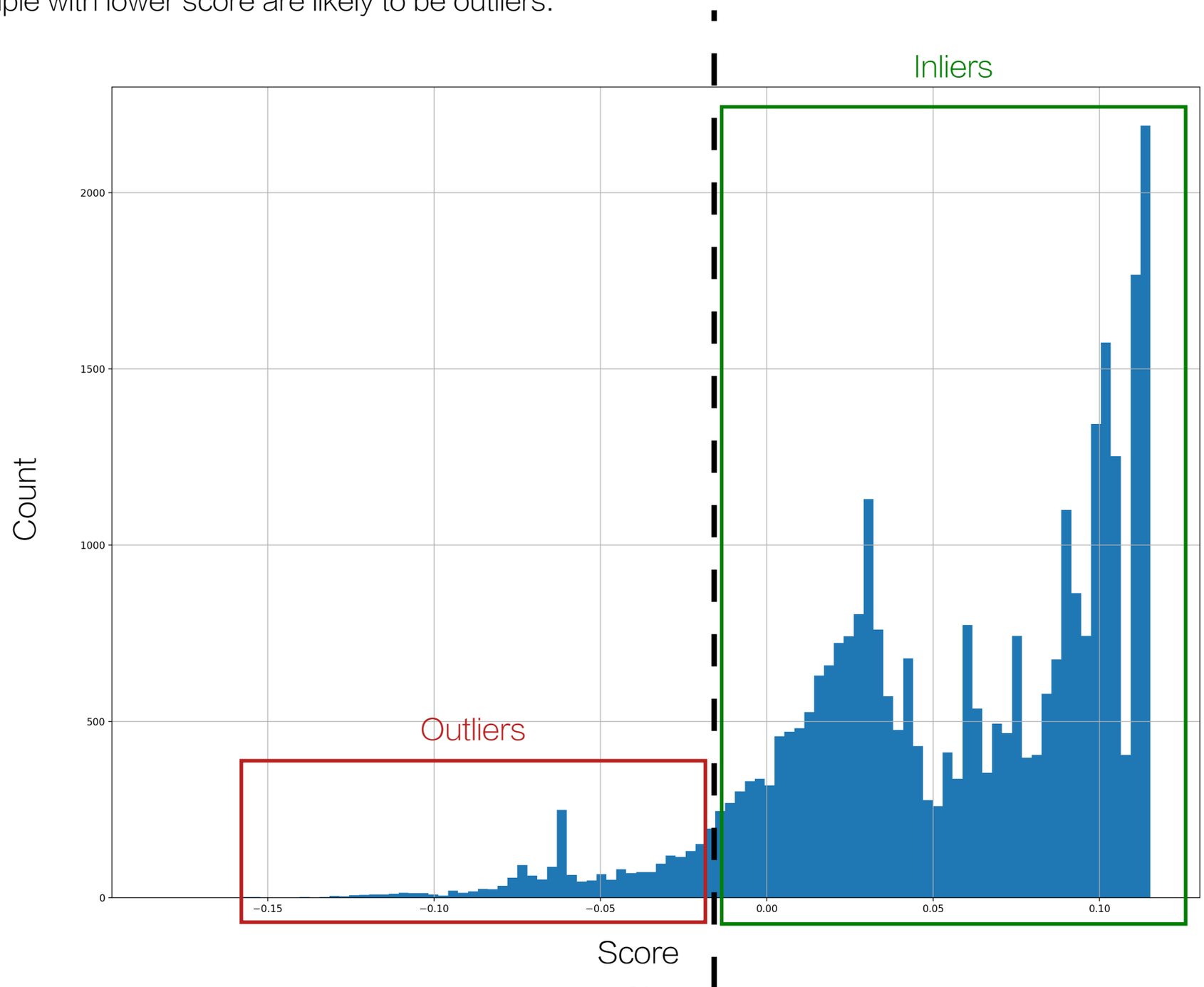
The Scikit-learn implementation provides a score for each sample that increases from -1 to 1 with the number of splits.



Modeling

2. Isolation Forest

The sample with lower score are likely to be outliers.



Modeling

2. Isolation Forest

```
print(X['anomaly'].values_counts())
```

```
 1     27756
-1      3785
Name: anomaly, dtype: int64
```

Modeling

2. Isolation Forest – Results

Crawler

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norr
296503	233.46.142.110	2021-05-12 09:32:04+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
297902	233.46.142.110	2021-05-12 09:32:50+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
302321	233.46.142.110	2021-05-12 09:35:31+04:30	Get	101	api/v2/connect/1396318207	99	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	160053.0	0.
302369	233.46.142.110	2021-05-12 09:35:33+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	12.0	2.
303761	233.46.142.110	2021-05-12 09:36:14+04:30	Get	101	api/v2/connect/1944714213	465235	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	40604.0	0.
304029	233.46.142.110	2021-05-12 09:36:18+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
328160	233.46.142.110	2021-05-12 09:48:56+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	12.0	2.

Modeling

2. Isolation Forest – Results

Crawler

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norma
302565	148.208.107.7	2021-05-12 09:35:40+04:30	Get	404	favicon.ico	29827	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	4.0	0.03
302566	148.208.107.7	2021-05-12 09:35:40+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	20.0	2.87
309599	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/1330189377	5495	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309597	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/483825864	3813	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309596	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/718839810	5258	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309595	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/23998397	4944	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	28.0	0.00
309598	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/1803873472	6512	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	12.0	0.00

Modeling

2. Isolation Forest – Results

Normal

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norma
631560	4.138.32.12	2021-05-12 11:58:09+04:30	Get	200	pages/630180842	50797	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	16.0	0.0
631958	4.138.32.12	2021-05-12 11:58:19+04:30	Get	200	css/font_awesome.min.css	30891	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.7
631959	4.138.32.12	2021-05-12 11:58:19+04:30	Get	200	css/page.2f0fc69390da8cdff683.css	50880	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.4
634810	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	js/page.07cb314dc14eef820638.js	332023	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	28.0	1.4
634808	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/gadgets/join_pros3.jpg	34053	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.4
634807	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/sanjagh_logo_purpule5.png	4680	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	4.0	2.2
634809	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/default.jpg	20993	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.1

Modeling

2. Isolation Forest – Results

Normal

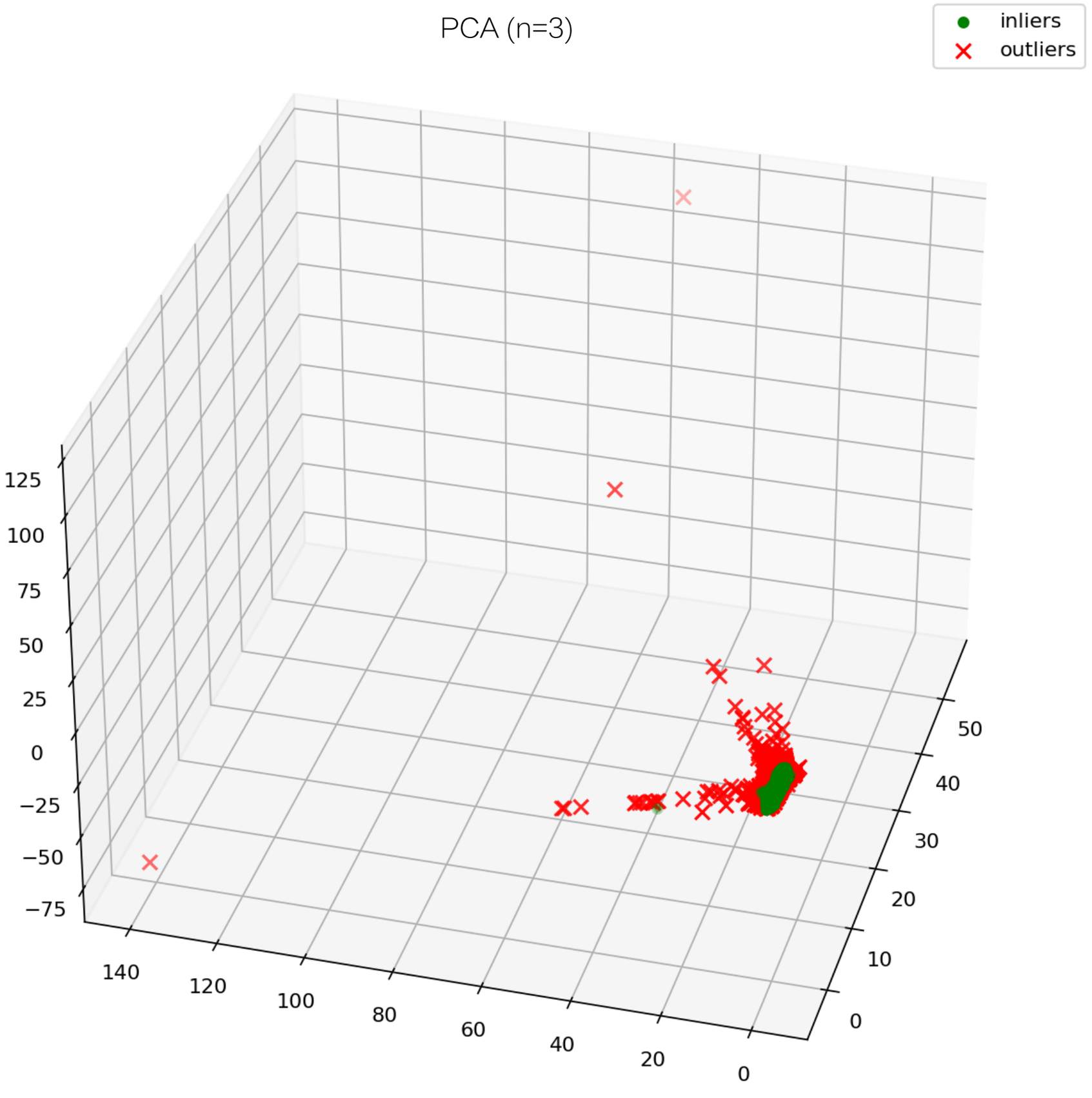
	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_nc
929570	37.199.253.251	2021-05-12 13:15:02+04:30	Get	200	pages/2098538394	52698	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	12.0	
929597	37.199.253.251	2021-05-12 13:15:03+04:30	Get	304	css/font_awesome.min.css	0	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	
929598	37.199.253.251	2021-05-12 13:15:03+04:30	Get	304	css/page.2f0fc69390da8cdff683.css	0	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	
929605	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/sanjaghmaglogo1.png	25889	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929606	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/gadgets/join_pros3.jpg	34053	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929607	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/default.jpg	20993	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929608	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/sanjagh_logo_purpule5.png	4680	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	

Modeling

2. Isolation Forest – Results

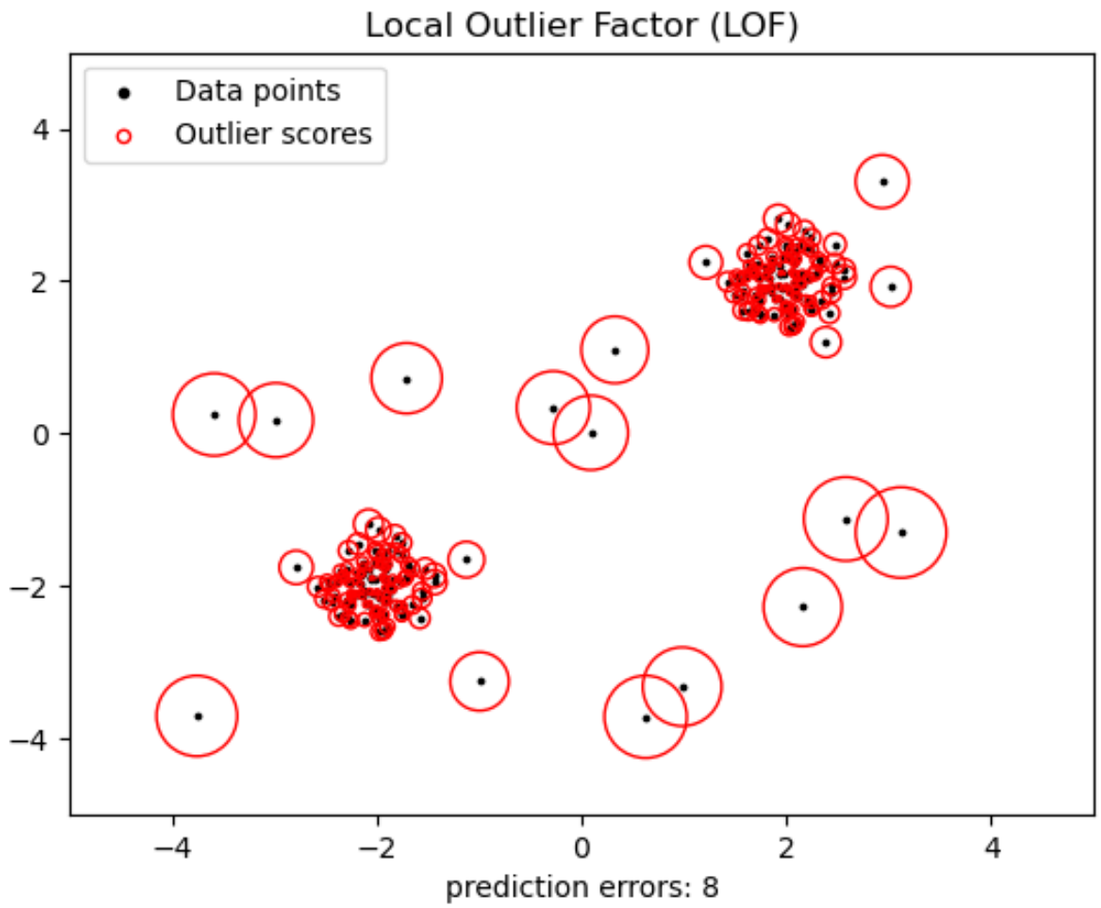
```
print(len(anomalies[anomalies['is_bot']])))
print(len(non_anomalies[non_anomalies['is_bot']])))
```

```
603
12
```


PCA (n=3)
inliers
outliers
125
100
75
50
25
0
−25
−50
−75
140
120
100
80
60
40
20
0
50
40
30
20
10
0

3. Local Outlier Factor

It measures the local deviation of density of a given sample with respect to its neighbor. It is local in that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood.

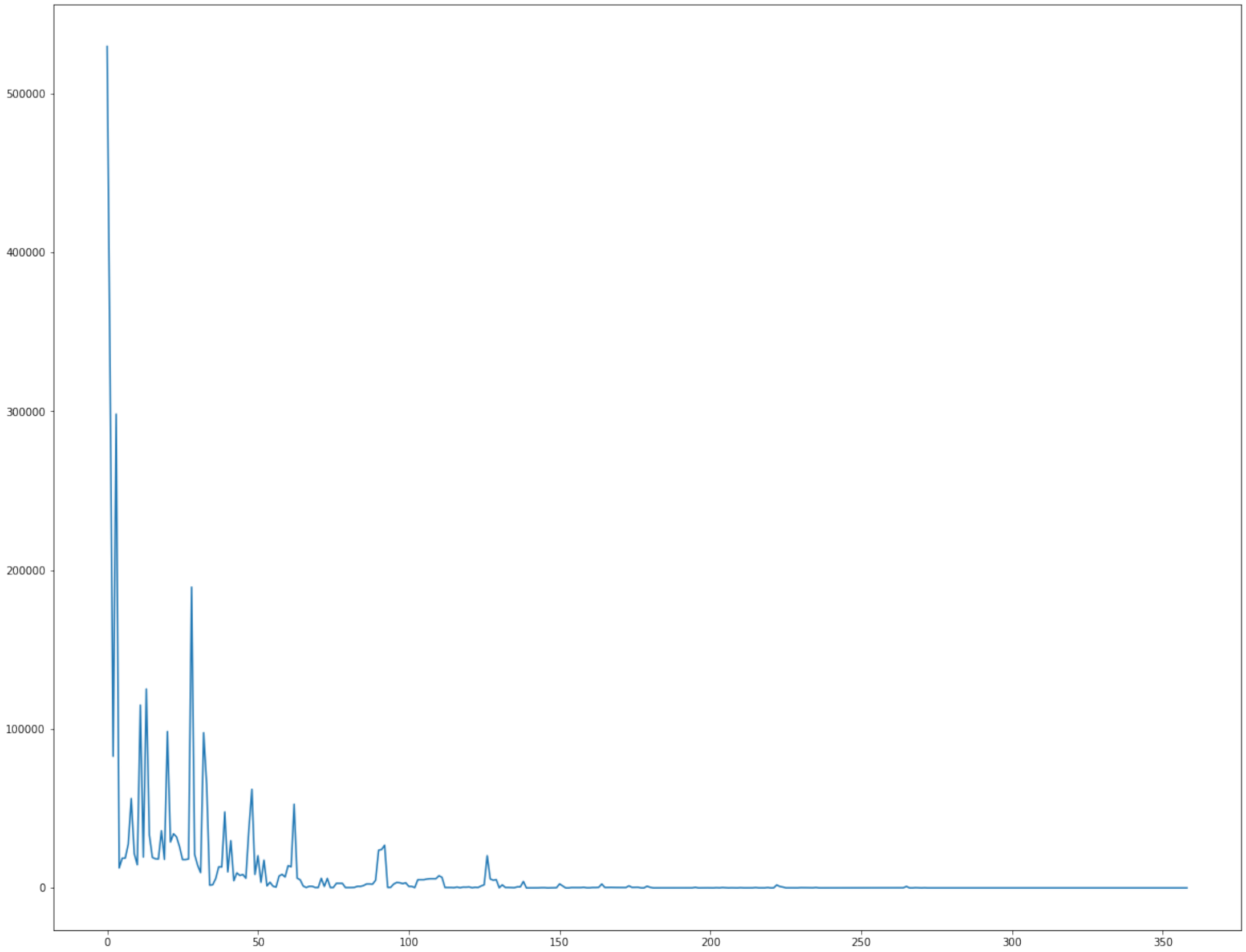


Modeling

4. Take advantage of PATHS!

ip	user_agent	time	method	status_code	path
153.126.83.145	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	200	pages/13405616
	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	304	js/page.07cb314dc14eef820638.js
	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	304	css/page.2f0fc69390da8cdff683.css
	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	304	css/font_awesome.min.css
	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	304	images/sanjaghmaglogo1.png
	Mozilla/5.0 (Linux; Android 10; SM-N960F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Mobile Safari/537.36	2021-05-12 05:43:17+04:30	Get	304	images/gadgets/join_pros3.jpg

		time	method	status_code	path
ip	**user_agent**				
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	2021-05-12 05:08:46+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:08:46+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:08:46+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:09:16+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:09:16+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:09:16+04:30	Get	200	js/profession.c67de06df71c34fc126d.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:52+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:52+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:53+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:53+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:53+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:53+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:10:53+04:30	Get	200	js/profile.bd977d36866688b90b03.js
	sentry/21.4.1 (https://sentry.io)	2021-05-12 05:11:50+04:30	Get	200	js/profession.c67de06df71c34fc126d.js


500000
400000
300000
200000
100000
0
0
50
100
150
200
250
300
350

Modeling

4. Take advantage of PATHS! – Results

Crawler

ip	user_agent	time	method	status_code	path	response_length	response_time
207.213.193.220	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.90 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	2021-05-12 06:30:39+04:30	Get	200	amp/blog/article/1782595231	92731	144.0
	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.90 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	2021-05-12 06:30:39+04:30	Get	200	amp/blog/article/735452053	107418	124.0
	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.90 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	2021-05-12 06:30:43+04:30	Get	200	amp/blog/article/1279149218	99336	124.0
	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.90 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	2021-05-12 06:30:43+04:30	Get	200	amp/blog/article/1211446147	99454	128.0

Modeling

4. Take advantage of PATHS! – Results

Crawler

ip	user_agent	time	method	status_code	path	response_length	response_time
123.107.128.103	**Go-http-client/2.0**	2021-05-12 13:38:06+04:30	Get	200	robots.txt	4825	4.0
	Go-http-client/2.0	2021-05-12 13:38:07+04:30	Get	200	sitemap.xml	790	8.0
	Go-http-client/2.0	2021-05-12 13:38:07+04:30	Get	200	sitemap.xml	790	4.0
	Go-http-client/2.0	2021-05-12 13:38:07+04:30	Get	200	sitemaps/basicSitemap.xml	100388	116.0
	Go-http-client/2.0	2021-05-12 13:38:12+04:30	Get	200	sitemaps/forumSitemap.xml	224081	108.0

4. Take advantage of PATHS! – Results

Crawler

ip	user_agent	time	method	status_code	path	response_length	response_time	path_count_normalized	path_length
21.101.108.189	**python-requests/2.18.1**	2021-05-12 06:13:05+04:30	Get	307		0	4.0	0.481101	1
	python-requests/2.18.1	2021-05-12 06:13:05+04:30	Get	200	877499224	63379	28.0	2.670010	1
	python-requests/2.18.1	2021-05-12 06:13:06+04:30	Get	200	877499224	63379	28.0	2.670010	1
	python-requests/2.18.1	2021-05-12 06:13:06+04:30	Get	200	877499224	63379	36.0	2.670010	1

Outline

Modeling

Evaluation

Demo

Evaluation

Life is not easy for any for us.

Marie Curie.

Evaluation

1. Known crawlers.

```
test_outliers[test_outliers['user_agent'].str.contains('python', case=False)]
```

	ip	user_agent	requests_count	path_length_std	4xx_percentage(%)	3xx_percentage(%)	HEAD_count(%)	image_count(%)	total_response_l
258	21.101.108.189	python-requests/2.18.1	4	0.0	0.0	25.0	0.0	0.0	1

```
train_outliers[train_outliers['user_agent'].str.contains('blackberry', case=False)]
```

	ip	user_agent	requests_count	path_length_std	4xx_percentage(%)	3xx_percentage(%)	HEAD_count(%)	image_count(%)	total_response_leng
284	14.9.86.233	Mozilla/5.0 (BlackBerry; U; BlackBerry 9900; e...	117	0.516883	23.076923	48.717949	0.0	49.57265	272615

Evaluation

2. What we expected.

Average of	# of requests	Path length STD	Percentage of 4xx	Percentage of 3xx	Percentage of HEAD reques	consecutive repeated requests	robots.txt requests	Percentage of image requests
Outliers	231	0.43	3.21%	9.33%	0.34%	0.81	0.08	9.74%
Inliers	25	0.39	0.68%	26%	0.00003	0.62	0.0	28.16%

Evaluation

3. Classification report!

We selected the **150** top most confident predictions(High scores) and tagged them manually.

Accuracy	90%
Precision	85.71%
Recall	100%
F1-score	92.30%

Evaluation

4. Bots!

Bots	GoogleBot	Go-http-client	Freshping Bot	Twitter Bot
Outliers	100	15	6	15
Inliers	29	5	0	14

Outline

Modeling

Evaluation

Demo

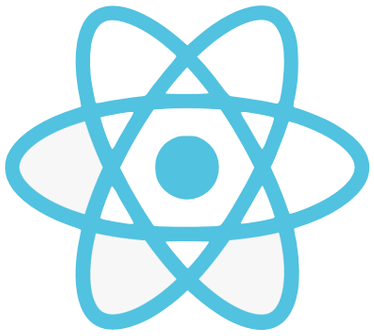
Demo

Demo

Flask

We also developed a Webpage to play around with. **[Link]**

Check the source code if you are interested. **[Link]**



React

Thanks.

Q/A

Mohammad Hashemi – Parastoo FalakAflaki