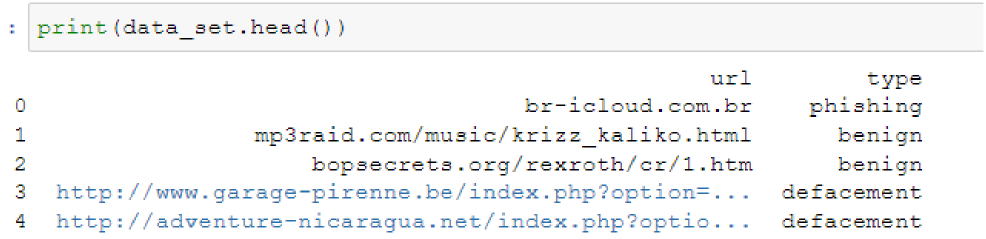
Irina Nazarova Sep 23, 2023

Machine Learning Model

Note: data set for the model training was taken from Kaggle.com

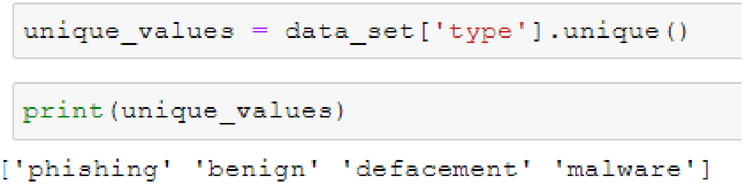


```
: print(data_set.head())

                                                  url        type
0                                    br-icloud.com.br    phishing
1                 mp3raid.com/music/krizz_kaliko.html      benign
2                     bopsecrets.org/rexroth/cr/1.htm      benign
3  http://www.garage-pirenne.be/index.php?option=...  defacement
4  http://adventure-nicaragua.net/index.php?optio...  defacement
```

Data set included two fields: URL and type.

There were four distinct website types:



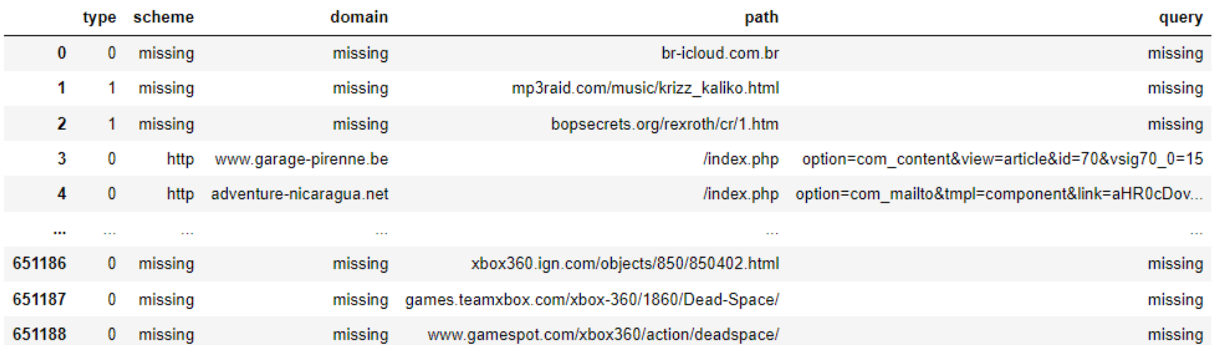
```
unique_values = data_set['type'].unique()
```

```
print(unique_values)
```

```
['phishing' 'benign' 'defacement' 'malware']
```

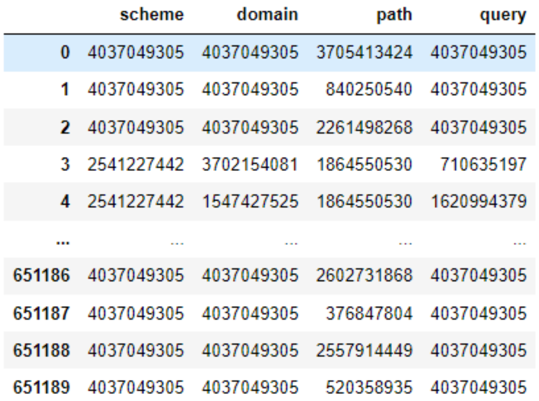
Phishing, defacement and malware types are put into one ‘malicious’ category

Dataset was transformed by separating components of URL



	type	scheme	domain	path	query
0	0	missing	missing	br-icloud.com.br	missing
1	1	missing	missing	mp3raid.com/music/krizz_kaliko.html	missing
2	1	missing	missing	bopsecrets.org/rexroth/cr/1.htm	missing
3	0	http	www.garage-pirenne.be	/index.php	option=com_content&view=article&id=70&vsig70_0=15
4	0	http	adventure-nicaragua.net	/index.php	option=com_mailto&tmpl=component&link=aHR0cDov...
...	...	...	...	...	...
651186	0	missing	missing	xbox360.ign.com/objects/850/850402.html	missing
651187	0	missing	missing	games.teamxbox.com/xbox-360/1860/Dead-Space/	missing
651188	0	missing	missing	www.gamespot.com/xbox360/action/deadspace/	missing

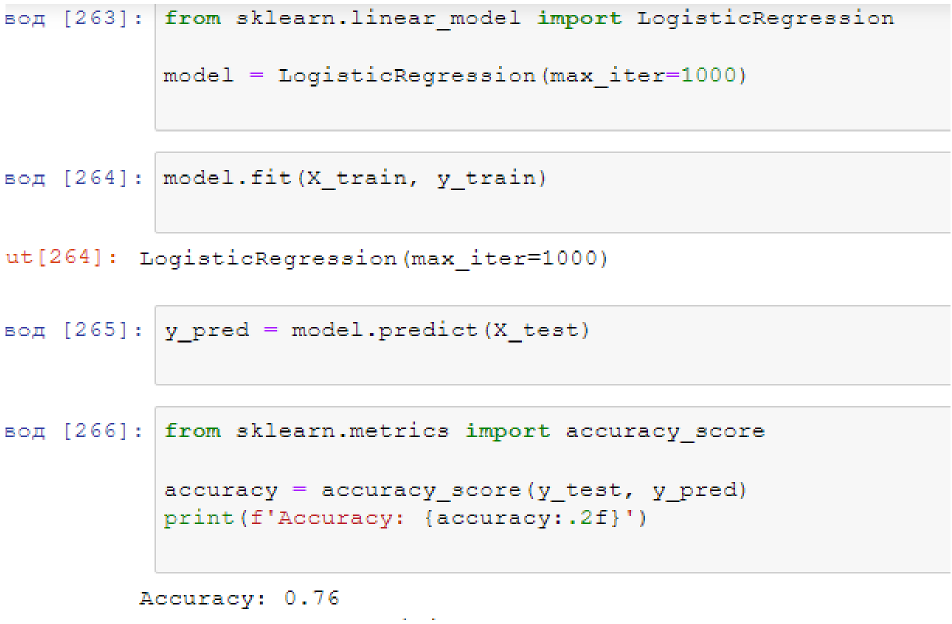
Features (scheme, domain path and query) were encoded to later feed it to machine learning model



	scheme	domain	path	query
0	4037049305	4037049305	3705413424	4037049305
1	4037049305	4037049305	840250540	4037049305
2	4037049305	4037049305	2261498268	4037049305
3	2541227442	3702154081	1864550530	710635197
4	2541227442	1547427525	1864550530	1620994379
...	...	...	...	...
651186	4037049305	4037049305	2602731868	4037049305
651187	4037049305	4037049305	376847804	4037049305
651188	4037049305	4037049305	2557914449	4037049305
651189	4037049305	4037049305	520358935	4037049305

Algorithm Chosen:

Logistic Regression. It is a perfect algorithm for supervised learning with a goal of classifying something (in this case – websites into categories 'malicious', 'benign')



```
вод [263]: from sklearn.linear_model import LogisticRegression

           model = LogisticRegression(max_iter=1000)

вод [264]: model.fit(X_train, y_train)

ut[264]: LogisticRegression(max_iter=1000)

вод [265]: y_pred = model.predict(X_test)

вод [266]: from sklearn.metrics import accuracy_score

           accuracy = accuracy_score(y_test, y_pred)
           print(f'Accuracy: {accuracy:.2f}')

         Accuracy: 0.76
```

Accuracy of the model was 76%

Analysis:

It is very hard to accurately determine the maliciousness of a website, knowing only URL. In a future, this model could be improved by including additional parameters connected with network activity, traffic and behavioral analysis of the website.