The Problem:

URL’s have many different identifiers that can help distinguish them between being safe, or malicious. But not every URL can be easily identified as such. Each URL has a protocol, host name, its primary domain, top-level domain, path, and length. Each one of those can help us know if its malicious, but using machine learning algorithms, we can make the job easier for the average person and help identify malicious URL’s before they click on them.

According to statistics presented in a research paper I looked at, in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Hackers use malicious and misleading URLs to lure people to click the link and expose sensitive data.

Our Approach:

Features:

- Length of different part of the URL (hostname, path etc.)
- Protocol
- Get requests parameters
- Html from the web site (optional)
- Punctuation count
- Ip or domain name
- Shortened URL or full length

Our plan isn’t to come up with a new architecture, given the fact that this problem has been researched quite frequently, but to preprocess the data in a unique way and create new and more data specific features that can better help us identify the URLs as benign or malicious, using non deep learning methods such as Boosting, KNN, SVM, NB, unsupervised algorithms for outlier (anomaly) detection.

One of the major challenges in this project is the imbalance in the data and the False Negative mistakes which can be costly.
That’s the reason why we will evaluate the problem via Precision-Recall/ F-score.