Detecting Phishing URL’s

Data Mining B-565

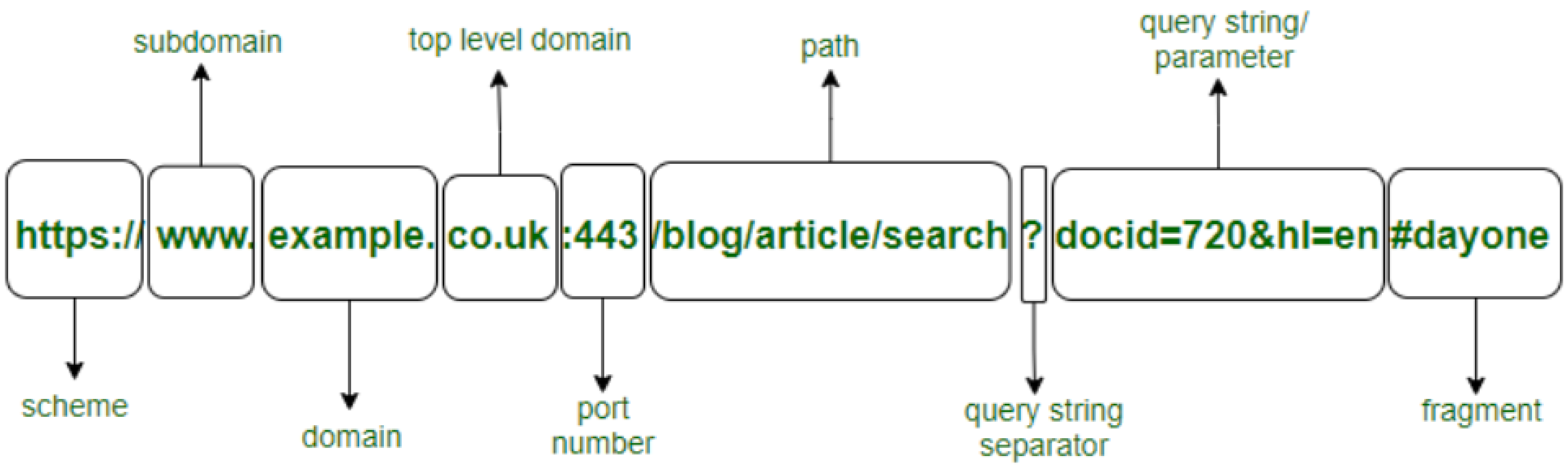
Pavuluru Rohith and Prajodh Pragath Sunder

Content

1. Intro to phishing URL’s
2. Dataset overview
3. Sampling
4. Modeling

URL

URL : https://www.example.co.uk:443/blog/article/search?docid=720&hl=en#dayone



GeeksforGeeks. (n.d.). Components of a URL. GeeksforGeeks. https://www.geeksforgeeks.org/components-of-a-url/

Types of URL’s

Benign

URLs that are harmless, safe, and not associated with any malicious activities.

Defacement

Web addresses that lead to websites or web pages that have been altered, vandalized, or defaced by unauthorized individuals or groups.

Malware

URLs that are associated with the distribution or hosting of malicious software.

Phishing

Phishing is a cyber attack technique where attackers try to trick individuals into revealing sensitive information.

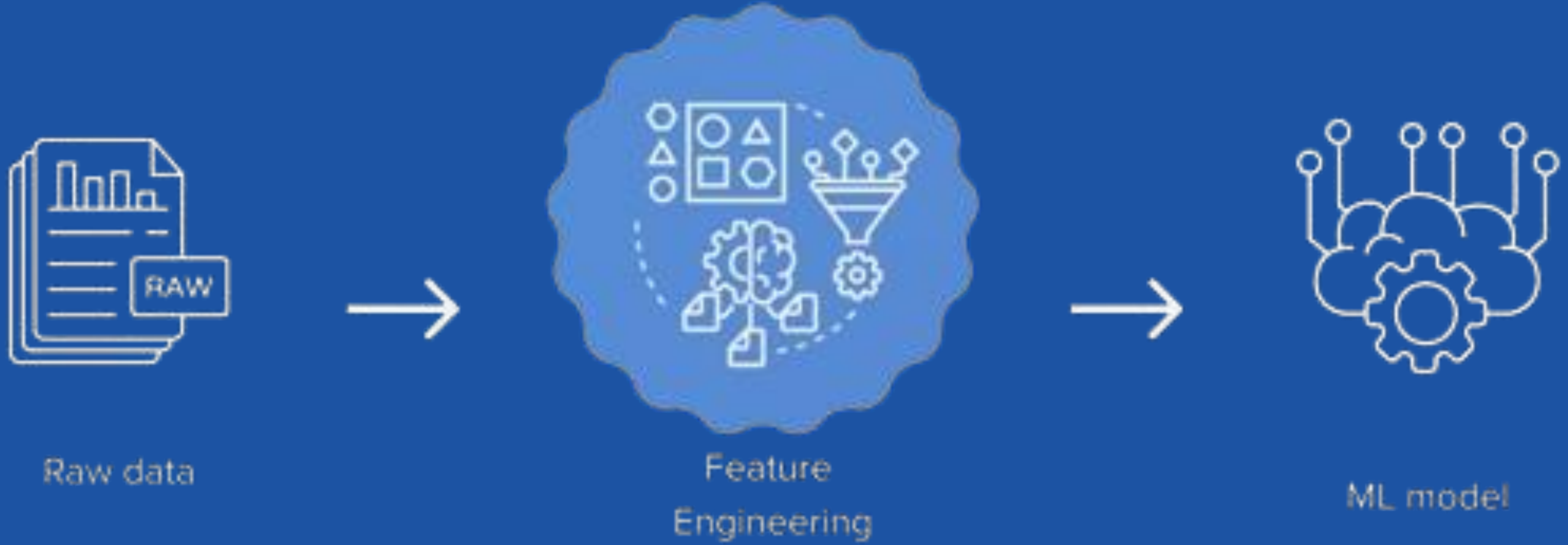
Dataset Overview

Sourced from kaggle’s “Malicious URLs” dataset which was extracted from URL dataset (ISCX-URL2016) of University of New Brunswick.

The dataset had 2 columns, URLs and the type of URLs.

The dataset has 651,191 URLs, out of which 428,103 benign or safe URLs, 96,457 defacement URLs, 94,111 phishing URLs, and 32,520 malware URLs.

Feature Engineering



Sampling

Random

A statistical technique where each member of the population has an equal chance of being selected to be part of the sample.

Random with replacement

the population has an equal chance of being selected for the sample on each draw, and after being selected, they are placed back into the population before the next draw.

Weighted

A statistical sampling technique in which different elements in a population are assigned different probabilities of being selected into the sample.

Stratified

The population can be divided into subgroups or strata, this method involves randomly selecting individuals from each stratum.

Models

Best estimator

- Logistic Regression
- Support vector machines
- Gradient boosting

Best Hyperparameter (Gradient boosting)

- Number of trees
- Learning rate
- Max depth of trees

Deep learning

Tensorflow neural network

Results