Dataset Nutrition Label

Trends in Cybersecurity Data

About:

This dataset was created using data from two separate datasets (hereafter referred to as “dataset 1” and “dataset 2”) containing simulated data for cybersecurity attacks. This data was created based on real experiences, but the simulation provides us with a realistic estimation of cybersecurity attacks and their significance. As the world becomes more technologically advanced, more faith is being put into machines and more data is being shared. This data is important and although it helps optimize our daily lives, comes with a considerable amount of risk. There are many individuals that aim to take advantage of certain vulnerabilities in the system that we have created, so cybersecurity has become an increasingly important topic considering the information we are dealing with and the ease of access.

Data collection:

Dataset 1 compiles synthetic data using cGan
Unknown for dataset 2

Created By:

1 - Incribo
2 - Zunxhi Samniea

Data Creation Range:

Aug 2023 – present (updated monthly)

Sources:

Dataset 1:
https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks
Dataset 2:
https://www.kaggle.com/datasets/zunxhisamniea/cyber-threat-data-for-new-malware-attacks

Content:

The dataset brought together two dataframes that were both comprised of synthetically generated data based on real experiences. These frames were created to allow for an unobstructive method of exploring common cybersecurity vulnerabilities and trends. This dataset contains information to help identify the systems that were targeted, the methods people used to gain access, and allows for the analysis of specific attacks. This can be used to help recognize common attacks, vulnerabilities and other trends in this space to help develop better security solutions and identify common trends.

Biases:

- Source Bias
  - The data was created using synthetically generated programs that require being fed information, but we do not have the information the programs were fed.
- Sampling Bias
  - This data is taken from samples of larger datasets, which can skew the representation of the dataset as we are only looking at a small subset of the larger data.

Assumptions:

- Assumes that data properly reflects current state of cybersecurity.
- Assumes no recent major developments that have changed trends in cybersecurity.
- Assumes the data has been given proper diversity for better representation.

Limitations:

- Dataset 2 is not licensed
  - Usage may not be properly regulated.
- Synthetically generated
  - Limited how representative of the population the data.

Concerns:

- Possible misrepresentation
  - Especially since the data was not given proper sources, it is possible that this synthetic data may not properly represent greater populations.
- Provided IP addresses
  - Although synthetic, it is unclear if these IP addresses correlate to real individuals or if they were also synthetically generated, which could lead to privacy concerns.

Possible Applications:

- Identify current trends in cybersecurity
- Determine common vulnerabilities
- Intended to test and establish solutions to certain threats and model machine behavior

Proper Utilization:

Considering that this data was made synthetically and contains sensitive private information such as IP addresses and how cybersecurity attacks were performed, it is vital that this data is handled carefully. Ensure that in using this dataset you are following proper licensure agreements, upholding privacy standards and understand the implications of using synthetically generated data.