CS 4485 Project Meeting Notes
February 1st, 2024

OVERVIEW

LCCDE Paper

- ML Algorithms to look into:
    - XGBoost
    - LightGBM
    - CatBoost
- Framework
    - Figure 2 in the paper

Front End

- React - in github
- Log of potential attacks
- Attack or not based on what the ML model analyzes
- Will have multiple classifications

Backend

- Database: SQL is set up
- ML model

To set up:

- System Architecture
- Requirements

What do we take into account when considering attacks?

- https://www.kaggle.com/code/dheerendragupta/ids-using-lightgbm-and-xgb
- Sample Data
    - On github
    - 70/30 split
    - Kaggle Data
- Preprocessing
    - Data cleaning - handling missing values by mean imputation, etc.
    - Data Scaling and Normalisation - Scaling or Normalisation is a common preprocessing technique used in machine learning where the data is usually normalized to a scale of 0 to 1.
    - Data Encoding - Most of the models cannot process strings/objects. So the data needs to be transformed to numerical data. This process is known as data encoding(also data transformation).
    - Feature Selection - Removing redundant features or selecting the most "useful" features. We used recursive feature elimination for feature selection.
- Data labels
    - 'Flow Duration', 'Total Fwd Packets', 'Total Backward Packets', 'Total Length of Fwd Packets', 'Total Length of Bwd Packets', 'Fwd Packet Length Max', 'Fwd Packet Length Min', 'Fwd Packet Length Mean', 'Fwd Packet Length Std', 'Bwd Packet Length Max', 'Bwd Packet Length Min', 'Bwd Packet Length Mean', 'Bwd Packet Length Std', 'Flow Bytes/s', 'Flow Packets/s', 'Flow IAT Mean', 'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Total', 'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min', 'Bwd IAT Total', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max', 'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',

'Bwd URG Flags', 'Fwd Header Length', 'Bwd Header Length', 'Fwd Packets/s', 'Bwd Packets/s', 'Min Packet Length', 'Max Packet Length', 'Packet Length Mean', 'Packet Length Std', 'Packet Length Variance', 'FIN Flag Count', 'SYN Flag Count', 'RST Flag Count', 'PSH Flag Count', 'ACK Flag Count', 'URG Flag Count', 'CWE Flag Count', 'ECE Flag Count', 'Down/Up Ratio', 'Average Packet Size', 'Avg Fwd Segment Size', 'Avg Bwd Segment Size', 'Fwd Header Length.1', 'Fwd Avg Bytes/Bulk', 'Fwd Avg Packets/Bulk', 'Fwd Avg Bulk Rate', 'Bwd Avg Bytes/Bulk', 'Bwd Avg Packets/Bulk', 'Bwd Avg Bulk Rate', 'Subflow Fwd Packets', 'Subflow Fwd Bytes', 'Subflow Bwd Packets', 'Subflow Bwd Bytes', 'Init_Win_bytes_forward', 'Init_Win_bytes_backward', 'act_data_pkt_fwd', 'min_seg_size_forward', 'Active Mean', 'Active Std', 'Active Max', 'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min', 'Label'

- Data classification
    - 'BENIGN', 'DoS', 'PortScan', 'Bot', 'Infiltration', 'WebAttack', 'BruteForce'

QUESTIONS

- Live attacks? Or past attacks?
- Do we need to store data? - classification model
- What should the front end look like?
    - Functions of the front end?
- Should this be account based?
- Who is the user?
- Are there limitations on the training models?
- What is expected of the user?
    - Monitoring

    - Write reports
    - Team to handle next steps?
- Is there anything we have to do with the output? Any steps after detecting the attack?
    - Type of intrusion?

ACTION ITEMS: Due by next Thursday!

- Look into the Kaggle Code
    - https://www.kaggle.com/code/dheerendragupta/ids-using-lightgbm-and-xgb
- Look into data cleaning - Mr. Gupta
- LCCDE Github : https://github.com/SoftwareImpacts/SIMPAC-2022-260/tree/main