Capstone Project 2: Proposal

Machine Learning for Cyber Security: Intrusion Detection

Client: online company looking for a data science consultant to assist them with their web security problems.

Goal: The goal is the help built a system for intrusion detection, to detect when someone has intruded in the network.

Potential data set:
https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

Overview:

Cyber attacks are increasing as more and more data is being processed throughout industries. Many companies are being targeted by detrimental attacks especially in critical business sectors such as banking. It is critical to find solutions to prevent these attacks before any loss can occur. With new improvements in machine learning and data science, data security is becoming a key issue to resolve. Data scientists can apply their knowledge to the cybersecurity field to help protect attacks and identify suspicious behavior. The goal is to identify threats, stop intrusions and attacks, properly identify malware and spam, and prevent fraud. A wide range of problems that fit a data science problem.

For this project I will be analyzing and create an intrusion detection system using a plethora of machine learning techniques. I will dive into a supervised learning approach in order to make predictions on classification of anomaly data, a unsupervised learning approach to determine a way to detect anomalies instead of predicting them, and I will create a deep neural network of anomaly data.

An intrusion detection system (IDS) monitors the network traffic looking for suspicious activity, which could represent an attack of unauthorized access. Traditional systems were designed to detect known attacks but cannot identify unknown threats. The problem stems from sophisticated attackers who can bypass common IDS techniques. Most techniques used in today's IDS are not able to deal with the dynamic and complex nature of cyber attacks on computer networks. Hences, a well defined more intelligent solution is needed. A good alternative solution to this problem is applying machine learning to this area of cybersecurity, this result in higher detection rates, lower false alarm rates and reasonable computation cost.

Steps to solve problem:

- Data wrangling/exploration
- Supervised learning
    - SVM
    - KNN
- Unsupervised learning
    - Clustering
- Deep neural net

Deliverables

- Recommendations / visualizations / model

Dataset

- For this project we are going to utilize the UNSW-NB 15 dataset created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviors. Found here: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
- This data has nine types of attacks (Fuzzers, Analysis, Backdoor, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms)
- There are a total of 49 features.
- A partition from this dataset is configured as a training set and testing set.
    - The number of records in the training set is 175,341
    - The number of records in the testing set is 82,332