CMSC451 - CS22-316 Real Time Anomaly Detection Project Proposal

Problem Statement

Anyone with sensitive information requires a way to protect their information and assess threats the moment they occur. With that the goal of this project is to develop an application that reads event logs and alerts the user of anomalies when they happen. Anomaly detection of entity behavior from event logs is possible by using data that has been synthetically produced to benchmark machine learning and create the basis for the anomaly decision process. Using this benchmark, any new event logs will be compared against the machine learning model and produce the expected result. Though the goal is to do this in real time, near real time will be sufficiently produced using streaming services provided by Amazon Web Services.

Summary of Innovative Nature of Project/Furthers Research on a Topic

This project explores user and entity behavior analytics by comparison against the normal conduct of users using machine learning technologies to accurately alert the occurrence of anomalies in activity logs. The model generated by the machine learning also detects any anomalous behavior or deviations from typical patterns in real-time with the use of a data streaming service. As technology advances and entity behaviors become more complex, a real-time analytic approach to cyber security must be employed in order to maintain security standards.

How the Project Positively Impacts the World in Which We Live

In an age of rapidly advancing technology, we must be vigilant of entity behavior that may pose a threat to data and network security. Entity behavior can include network anomalies, application performance anomalies, web application security anomalies, etc. This project aims to detect these behaviors and ultimately prevent malicious actions. This proactive approach can have a positive impact on the world by helping to prevent financial disruption as well as maintaining reliability in the technology that we use in our daily lives. We frequently see news of cyber criminals attempting to gain access to personal information and sometimes even succeeding in their nefarious tasks. In such a perilous world, maintaining trust and reliability between a company and its users is proving to be a vital necessity.

Project Plan:

Expected Results
The goal of this project is to accurately predict anomalies in activity logs as the events occur (or near real time). Train the model on a large dataset in batch mode and predictions on test data should be on streaming data.
● Milestone 1: Identify a dataset for use in the project, design model architecture, perform baseline testing for validation of technique's effectiveness.
● Milestone 2: Perform real time anomaly detection with batch model training - This should include events from multiple data sources. Ex: if the user tries to perform event A followed by event B and C, then it is a more risky activity than just performing a single activity. These should be considered as anomalies based on their previous activity (baseline the user activity from a larger dataset during training the model, look for anomalies during prediction/test).

● The anomaly detection can be performed with predefined features and algorithms but would like to explore options on unsupervised learning as well (open to any thoughts)
● Milestone 3: Perform real time anomaly detection without any batch training, fully real time solution

Available Resources:
Any publicly available datasets can be used to detect anomalous events.
AWS Kinesis/Spark streams, spark streaming, sql or other similar streaming technologies can be used

Stakeholders:

1. Capital One:
● Geeta Shyamala (sponsor/mentor)
● Jake Jeffers (sponsor/mentor)
● Ben Polk (mentor)
● Andrew Kepley (mentor)
2. Incident Management Staff at Capital One
3. Cybersecurity Analysts at Capital One
4. Clients of Capital On