Brett Luskin

DATA 606 - Deliverable 1

Malware Detection with Machine Learning

The purpose of this project is to use machine learning algorithms to identify malware attacks based on the UNSW-NB15 dataset. This dataset is a combination of real network traffic and synthetic malware attacks that was created by the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS). There are 49 unique features to the dataset including the label. They consist largely of network data such as IP addresses, connection times, and number of connections.

The biggest challenge I will face in this project is domain knowledge. I have no experience dealing with cybersecurity issues, so this is an opportunity to expand on my skill set and knowledge base. While I have a basic understanding of the way networks work, some of the features in this dataset mean nothing to me at the start of this project. And while the dataset is daunting at first glance, the lack of domain knowledge is mitigated by a few factors. First, that this subject is well researched already. There will be an opportunity to read the works of those that have come before me. This will give me insight into what has been successful in the past, and possibly allow me to build on those methods. Second, that this boils down to a classification problem; a problem that I am trained to solve as a data scientist.

The primary goal of this project is to build multiple machine learning algorithms for this classification problem, optimize them, and compare their performance. Additionally, I will explain the differences in their performance. If time allows, a secondary goal would be to build a neural network and include this in my comparison of results. There is research on neural networks applied to this dataset available, so while I expect that I will be able to get a neural network working during the course of this project, I am not sure what to expect in my ability to optimize it. The last goal that I am not sure I will achieve, but that I want to aim for, is that by the end I will not only have replicated results based on existing research, but that I will be able to do some sort of iteration that advances what is possible in this space.

The timeline of this project is outlined by the course syllabus. EDA and research will be completed by March 1st. Model construction will last until the end of March. Execution and interpretation of the model results will be done by May 1st.