Detecting Cryptojacking Malware

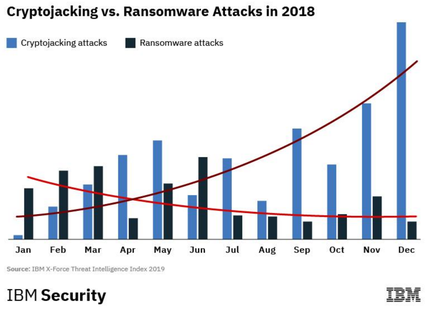
WOOSTER
Mathematical & Computational Sciences

Chase Fickes

Advised by Drew Pasteur and Max Taylor

Cryptocurrency

Cryptocurrency is an electronic currency that uses a decentralized system to self-govern the transactions of its users. To be decentralized means to not be controlled by a 3rd party. Cryptomining is the process of verifying transactions of cryptocurrency. This gives whoever verified the transactions a reward in cryptocurrency. Cryptomining uses a great amount of processing power. This leads to cryptojacking, which is the unauthorized cryptomining on another person's computer. In this case, the money all goes to the hacker, an, since cryptomining uses so much power, the victim's computer slows down and overheats which decreases the computer's lifespan.



Machine Learning

We can use machines to analyze data and classify it based on past trends. This is mostly done through machine learning, which uses data to "learn" from by analyzing patterns in an attempt to reapply them on new data in the future. We can use this for classification problems like ours which attempts to determine whether something is cryptojacking malware or not. There are many algorithms that are used for machine learning, but I mainly focused on using two to see which would perform better for my particular problem.

Decision Trees

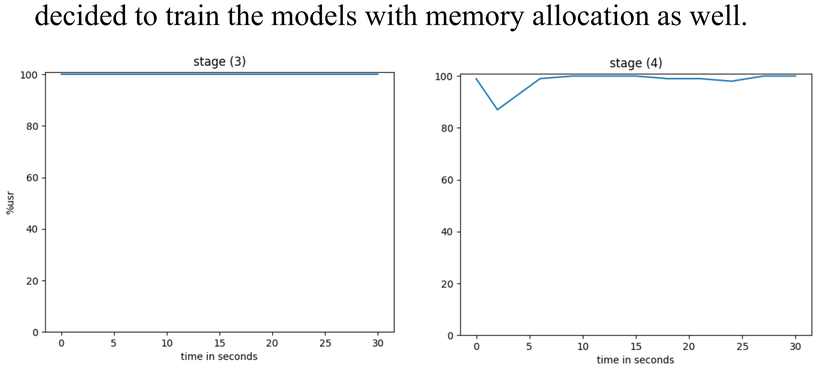
Decision trees are a type of machine learning algorithm that are commonly used in classification problems. They use the attributes of the data in the training set to come up with the optimal questions to ask when determining what class that an unknown data point is. As the name implies, decision trees are in the shape of trees, usually upside-down. They contain a root which has the first question, and each question leads to either another node containing a different question or the predicted class.

Support Vector Machines

These are another machine learning algorithm that are based on lines of regression. It is created by plotting data points from the training set on a graph based on their attributes, then finding the line that separates the classes of data. Each new point that you add to the graph will be predicted based on its relation to the line, whether it is above of below.

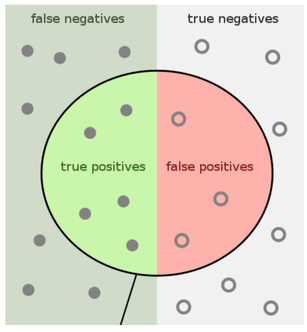
Project Design

For designing my project, I started by getting some statistics from my computer in different states. Since I was investigating CPU miners, I looked at the CPU allocation on these states. The states included one that was idle, a benchmark state used to emulate typical browser usage, a state with a cryptominer, and a state with an infinite while loop. The last two states were indistinguishable to some degree, so I decided to train the models with memory allocation as well.



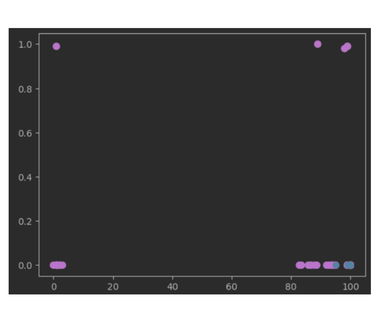
Evaluation

I used three main statistics to evaluate the models were accuracy, precision, and recall. Accuracy is the rate of successful predictions, precision is the rate of correct positive predictions, and recall is the rate of overall positives correctly identified.



Results

I trained 3 decision trees and 3 support vector machines, each with different training sets: a set that includes a single metric, only CPU allocation, and CPU allocation and memory allocation.



The decision trees did much better than the support vector machines, especially using the test data.

Single Metric			Tree	w/o Free		Tree	w/ Free	
Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
0.81	0.59	0.85	0.81	0.62	0.82	0.99	0.97	0.98

Single Metric			Tree	w/o Free		Tree	w/ Free	
Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
0.70	0.30	0.67	0.70	0.30	0.67	0.68	0.32	0.89

Single Metric			SVM	w/o Free		SVM	w/ Free	
Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
0.70	0.25	0.44	0.70	0.25	0.44	0.49	0.00	0.00

Future Work

If I were to continue on this project, I would like to further investigate the support vector machine models. Thought I can guess why they performed so poorly, I would like to have evidence to back up my claim and maybe fix it. I would also think it would be a good idea to use different variables in the training set than just the CPU and memory allocation. Based on the results, it is likely fine to remove most of the CPU allocation not based on the user.

References

- Tom Michael Mitchell. Machine learning. McGraw-Hill, 1997
- Lorne Lantz and Daniel Cawrey. Mastering Blockchain Electronic Resource: Unlocking the power of Cryptocurrencies, smart contracts, and Decentralized Applications. O'Reilly
- Carl Kingsford and Steven L Salzberg. "What are decision trees?" In: Nature Biotechnology 26.9 (2008), pp. 1011–1013
- Mustafa Qizilbash. Data entropy. url: https://www.linkedin.com/pulse/data-entropy-mustafa-qizilbash