Heather (00:12):

Welcome to The Hurricane Labs Podcast. I'm Heather, and today I'm chatting with two of our Splunk experts about machine learning and how it can be used in infosec. Tim, Josh, thanks for joining me today.

Tim (00:23):

Hey, Heather.

Josh (00:24):

Thanks for having us.

Heather (00:27):

Yeah, absolutely. So I guess the first question of the day is what exactly is machine learning and how does it connect or relate to InfoSec?

Tim (00:37):

Yeah, so machine learning it's basically using algorithms that can improve and self-learn over time. In InfoSec, we're typically talking about a specific subset of machine learning called supervised learning. And the goal in supervised machine learning is to be able to take a data sample with labels and puts it outputs and be able to predict the output of a future event. Typical use case would involve using an input of some sort of log data like authentications or system performance metrics. And then we want to predict whether or not that event should be investigated further because it's anomalous or atypical in some way.

Heather (01:17):

So then what are some benefits and drawbacks to using machine learning with insecurity?

Tim (01:23):

Ideally you can have an alert or a detection that can self-learn and can let you know about things without as much to analyst input. You know, you have an alert and it sort of self regulates itself and you can get good information from it without having to have a lot of input. That being said, the downside is getting alerts in that state takes quite a bit of work, you have to kind of treat each one individually and, you know, put a lot of thought and effort into tuning it to get it to the point where you can have an alert in that state.

Heather (02:04):

So basically it could eventually reduce the analyst contact time, but initially it takes a lot of that analyst time to like train the machine and get it to where it can handle that load.

Tim (02:16):

Yeah, that's correct. What you're really looking for is a problem that is solvable easily with machine learning and, you know, some in some cases that applies in other cases, it won't, and it's, you know, you kind of have to really look at your problem space to determine whether or not your problem is going to

be easily solvable with it. You know, there's a few concepts in machine learning about, you know, how you're labeling the data and how your functions are going to deal with new examples.

Heather (02:49):

Okay. So how then is machine learning used in Splunk?

Tim (02:55):

So Splunk has a couple of add-ons that allow you to use it. One is the machine learning toolkit, and that relies on another add-on which is called the Python for scientific computing add-on. The Python add-on basically that's a lot of libraries that are used in machine learning and allows you to do some things that aren't included in the Splunk native Python and the machine learning toolkit builds out on top of that and allows custom commands that lets you use some of these machine learning algorithms and also you know, give some nice charting and graphing to let you visualize those in a way that would normally require, you know, Python libraries to do so.

Heather (03:43):

So with all the training that machine learning has to have in order to get to where it can sort of function independently? I mean, that's like you have to filter out like a lot of noise, right?

Tim (03:54):

Yeah. And that's one of the challenges that comes along with machine learning. You know, you can only be it's only as good as the data that you're feeding these algorithms. So if your dataset has almost all negatives and only a few positives, it's harder for the algorithm to learn when the positives are occurring. So it might take a lot of time for that dataset to build up to the point where you can get accurate results from it.

Heather (04:23):

So then when you're setting it up within Splunk in order to, you know, sort of, you know, get it to its potential and sort of start filtering out that noise, I guess, what are the needs you know, the data needs and the setup needs for MLTK with Splunk.

Josh (04:43):

I think I can take that one. So the first thing that you need to do with a machine learning alert is really take a look at the data that you have. So for example, with authentication data, you want to see, like the failure reason. You may only want to look at failures from either like outside with public IP addresses or only looking at inside failures, maybe only certain types of failure reasons, like a bad password, or you may want to look at like the destinations that the failures are happening against. It's all about understanding, normalizing your data to remove any noise that's going on inside, because I mean, with windows authentication, you may even run into problems with different types of login failures being logged in various places, or log on failures, being logged twice on multiple DCs or on a DC and a workstation if you're logging that.

Tim (05:41):

And to kind of add onto what Josh was saying, you also have to be kind of realistic on what Splunk can do. It's built great as a data processing platform, but machine learning algorithms get into spaces where

you're doing a lot of floating port operations, which CPUs don't handle as well as some other processing units out there like GPUs, so that can slow things down. And, you know, you're also, if you've tried to search over, you know, a billion plus events and [inaudible] before, you know, that can also slow things down a bit.

Josh (06:22):

Yeah. And a lot of times with these machine learning algorithms, you want to split the data by something like time or maybe asset data use it or category asset category, things like that. And I know by default, there is a limit on some of the algorithms about how many groups that you can have. So you really need to be cognizant of like, when, when you split the data, you want to make sure that your data looks in some kind of normalized distribution. So there's no skews anywhere. No, no multiple peaks. You want, you want to get the data to a point where there's a single peak in the data so that the algorithm can look at it and correctly identify outliers.

Heather (07:02):

What then, knowing just the limitations of machine learning and what Splunk does best, what would be like the best use case scenarios for someone wanting to use Splunk's machine learning?

Josh (07:16):

Yeah, I would say once you have like a mature Splunk, so you have all your data normalization, you know, your data and what you want to look for with the machine learning algorithm, you have good asset data, you've looked at it, you understand where the peaks are, you've split your data out in a way where there's not too many groups and you don't have so much data going in. You can get some pretty good results out of it. The problem is getting there and the limits of the hardware with how much data is going in. I know that there is limits that you can set. I think the default max number of events that you can input into some of these algorithms is a hundred thousand. I'm not sure if you are able to increase that, how that would affect performance. You might need to beef up your search heads that are running those algorithms.

Tim (08:08):

Yeah. And I think one thing to keep in mind and emphasize that Josh thought is, you know, you just need to be realistic about what you're trying to do with this, if you're having an alert where you're looking at a trend of data over time, or you're looking at trying to predict, you know, if a hard drive is going to go down based on some input, output data or something like that with these ML algorithms, you know, your feature space in that is going to be low. So, you know, you're within the event limit, you're within the group limit that will work well. But if you're trying to, you know, do something more wide-ranging, like predict if a event coming in or amount of data transferred to a certain IP is anomalous, then you're probably going to get into the space where you're dealing with too many events, too many groups, and it's not going to work well in Splunk.

Heather (09:08):

It sounds like the key is not using machine learning for the sake of using machine learning, but rather like knowing what your goals and purposes are and that this is that this tool is the best one for the job.

Tim (09:20):

Yeah. There's definitely a Goldilocks zone was the phrase I thought, for some reason that you kind of have to have your problem space end for it to work in Splunk natively with these two add-ons.

Josh (09:36):

Yeah. I'd say if you're struggling to like reduce them, like specifically for anomaly detection, like if you have a lot of outliers that are coming out of this algorithm and you've done pretty much everything you can to like, try to split that up, you may need to take a step back and question whether machine learning is really the right answer to the solution. Like there are other options within Splunk to find anomalies that may work better, that are able to split by more groups. You may be able to calculate behavior per user even with some of the other options in Splunk.

Heather (10:12):

So it's just about choosing the right tool for the right job.

Tim (10:15):

Exactly. Like a lot of the options with MLTK might not be intuitive for the first time user and that you kind of need to dig into the documentation to get the most out of them.

Josh (10:26):

Yeah. Like initially looking at the app, like it's going to take a lot of time to dig into the documentation to understand exactly what you said as the cases are going to work, looking at blog posts about how it's used. There's a lot of extra steps that you need to take that I've mentioned before getting something useful out of it. And there are a lot of use cases outside of security, where I could think of like the machine learning toolkit being amazing for.

Tim (10:55):

Yeah. That's one of the difficulties is in the security space in particular, you'll have events that by necessity, will have a lot of groups, a lot of features, and it's difficult to do that cardinality reduction to get it to where you want. And, yeah, if you really want to put in the effort of doing getting it to work in the MLTK space, you have to be prepared to read not only the documentation on Splunk's website, but there's a lot of stuff in the scikit-learn library that is in Python that the MLTK or yeah. MLTK on provides a lot of wrappings around that to make that usable in Splunk.

Heather (11:44):

Alright. Well thank you both very much for stopping to chat with me about it today. I appreciate it.

Tim (11:51):

Yeah, of course. Thanks having us, Heather.

Josh (11:54):

Thank you.

Heather (11:55):

And that's all for today. Be sure to check out our links below Josh and Tim put together a collection of resources for you to help point you in the right direction in your machine learning adventure. In our next podcast, I'll be chatting with several of our team members about diversity in the future of infosec. So stay tuned and don't miss out. Until next time, stay safe.