Finding the Needle in the Haystack

Anomaly Detection in the Cybersecurity Industry

AT A GLANCE

Capstone Company: Rapid7 **Location:** Boston, USA

Problem Statement: Develop and implement an algorithm that reduces the amount of time necessary for analysts to detect intrusions of bad actors in client networks and facilitates the detection of new intrusions not discovered before.

Timeline & Milestones

February	March	April	May	June	July	August
Feb 6: Initial kick off meeting	**Feb 28**: Learn data structure	**Apr 3**: Exploratory presentation due	**May 8**: Phase 1 presentation, Interim report, summer work plan due		**Aug 8-15**: Final report, Showcase presentation, Project poster due	
Feb 20: Project charter and workplan due	**Mar 4**: Operational on data usage	**Apr 25**: Understanding analytics @ Rapid7 presentation due			**Jul 14**: Mid-summer project status due	**Aug 23**: Final Capstone Presentation

The Team:

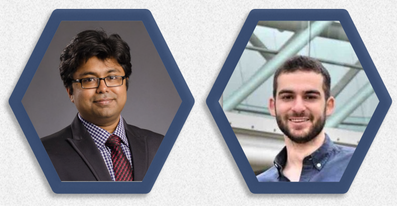
MBAn

Jocelyn Beauchesne | Johnny Oh

Rapid7

Roy Hodgman | Vasudha Shivamoggi

Faculty Advisors



Rahul Mazumder | Hussein Hazimeh

IMPACT

1. Direct Labor Savings

$1M+

Impact to Security Analysts:

Created machine learning models to classify 90% of the data as "normal", significantly reducing manual review of client network data at a projected $1M+

2. Avoidance Savings

$35M+

Impact to Clients:

By laying the foundation for modern machine learning in cybersecurity and driving initial findings, Rapid7 and their clients can avoid costs of at least $35M+ annually

3. Novel ML Tools

patentable research

Impact to Rapid7:

Packaged flexible, scalable, and auto-tuned machine learning pipeline to be used on any data sets for future anomaly detection use cases

DATASET

Rapid7 deploys "hunts" on client computer networks, which are deep downloads of computer behavior in a two week period; we used this data to conduct our anomaly detection analysis:

Hunt Data Statistics

- 4TB+ of data from 300+ clients
- Unstructured raw data
- 100% unlabeled without prior examples of intrusions

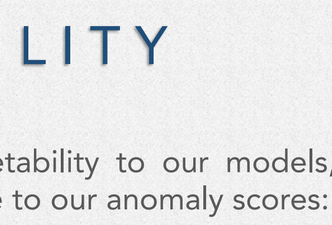
TWO-PART METHODOLOGY

1 Tree Approach to Contextualize Data

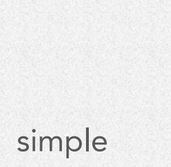
To contextualize data, we used paths to create a tree grouping similar processes together into leaves

Path	Command Line	Signing Details	MD5	Modules	Network
chrome.exe	option = private navigation	valid	2jljljiwldk24830	1. Network Driver 2. GIF Loader 3. Ad Blocker	IP[country = Russia]]= 12.35.534.34

1 **Tree Structure Using Paths**

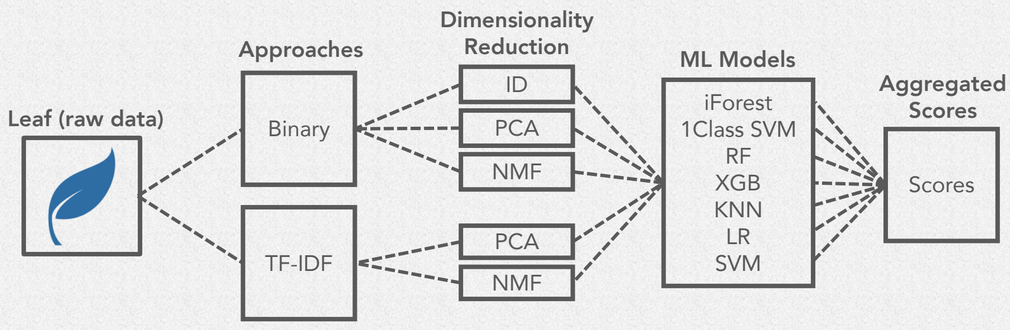


2 **Confidence Interval With Outlier**



2 Machine Learning Pipeline

Within each leaf, we processed the data to **reduce dimensionality** and executed machine learning models **automatically tuned with Bayesian optimization** to **identify anomalies** and **designate anomaly scores**

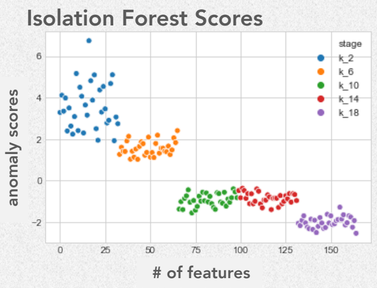


KEY RESULTS

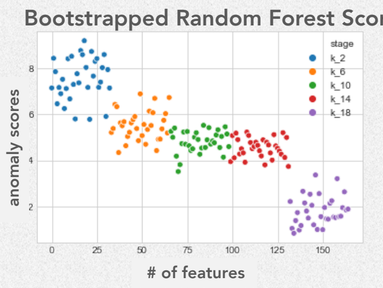
Synthetic Data

To test model efficacy, we ran multiple analyses on synthetic data; **85% accuracy on synthetic anomalies**

Isolation Forest Scores



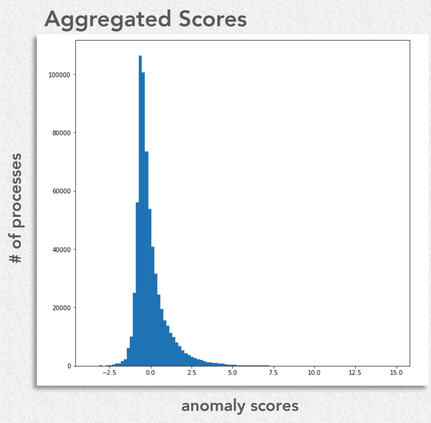
Bootstrapped Random Forest Scores



Client Data

The anomaly distribution score is narrow: **95% of client data lies within ±2.25σ** of the mean

Aggregated Scores



Examples of Anomalous Activities

1. Data transfer software **connected to Russia and Ukraine**
2. **Unregistered remote control software** installed on user files
3. **Malware** installed on user desktops

INTERPRETABILITY

In order to add a layer of interpretability to our models, we devised two simple approaches to tie feature importance to our anomaly scores:

- **Interpretable regression model** against anomaly scores on the original features
- **Individual feature anomaly scores** for each process

FUTURE DIRECTION

- Confirm results with additional analyses using synthetic data
- Develop cross feature interpretability
- Apply machine learning pipeline on other datasets for future research
- Create database of labeled intrusions and hacks for improved machine learning
- Continue developing key relationships with security analysts for feedback