BOTNET DETECTION

LAINE B.

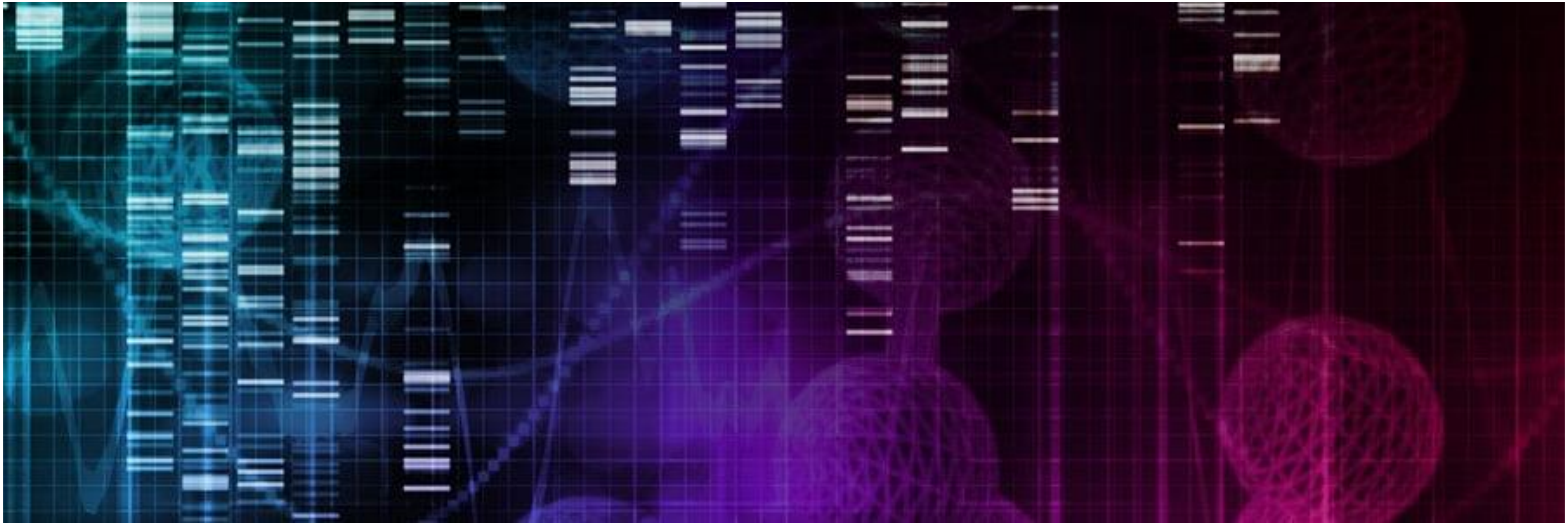
PROBLEM STATEMENT

- Malicious botnets emerged as early as 1999 and have been used for fraud, identity theft, and distributed denial-of-service attacks
- Increased connectivity and accessibility of processing power has expanded reach and frequency
  - Up to 50 million devices
  - Currently estimated to comprise 40 percent of all internet traffic
- Simple source and destination blacklists are not able to counter Domain Generation Algorithms and fast-flux techniques

MODELS

- Logistic Regression: feature analysis and covariance
- Decision Tree
- Gradient Boosting Machine

DATA PREPROCESSING & EXPLORATORY ANALYSIS

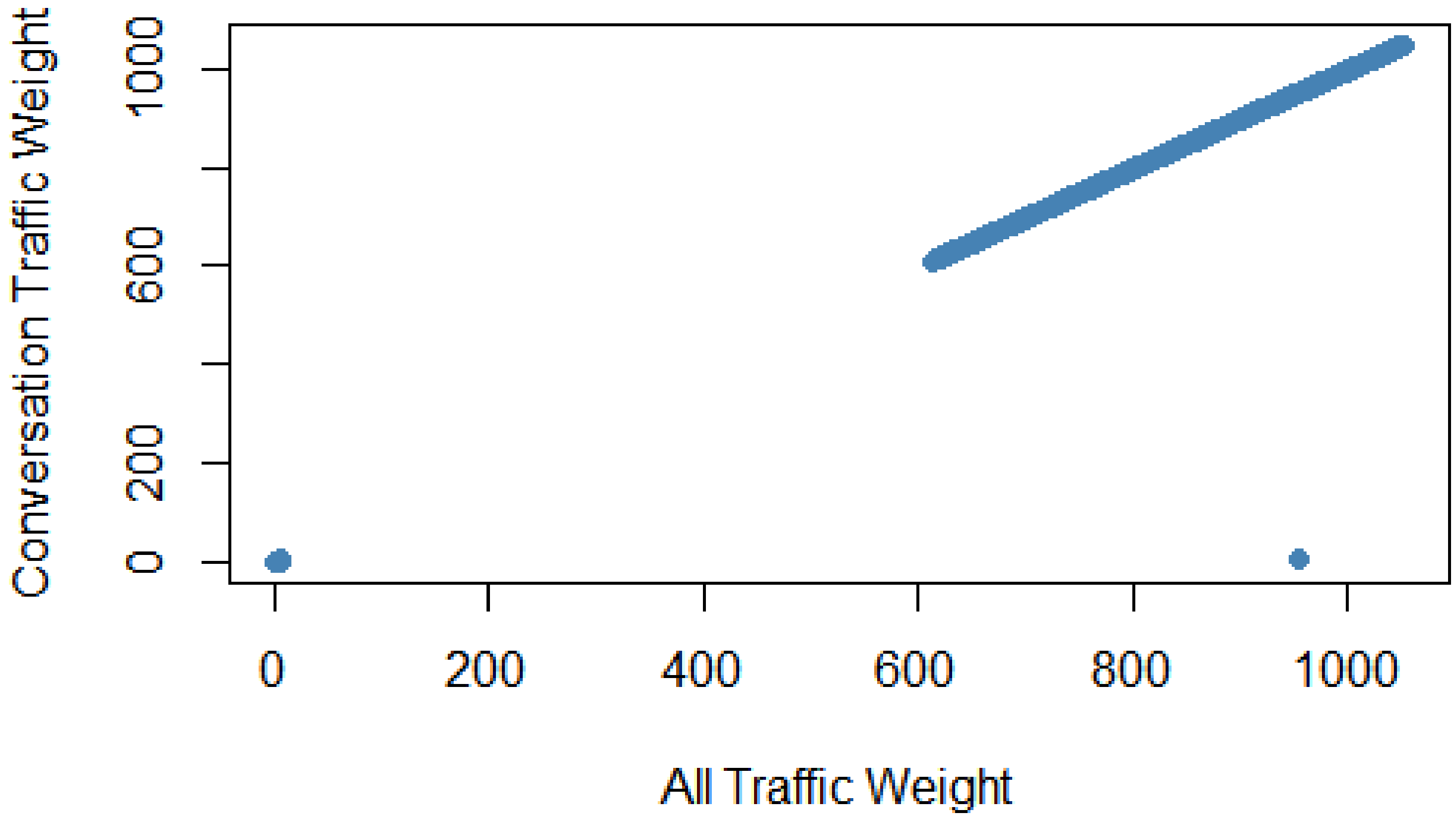


IoT Dataset for Intrusion Detection Systems (IDS)

- Integration of 9 IoT devices from the detection of IoT botnet attacks NBaIoT data set, including only 10-second time windows with a decay factor of 0.1 (L0.1)
- Response variable: *label* (0 = botnet traffic; 1 = benign)
- 23 calculated statistics across 4 stream aggregations:
  - H: recent traffic from the packet's host IP (all traffic from this source)
  - HH: recent traffic from the packet's host IP to the packet's destination IP (conversation traffic)
  - HpHp: recent traffic from the packet's host IP and port to the packet's destination IP and port
  - HH_jit: the jitter of traffic from the packet's host IP to the packet's destination IP (conversation jitter)
- Calculated statistics for stream aggregations:
  - weight: the number of items
  - mean
  - standard deviation
  - radius: root squared sum of the streams' variances
  - magnitude: root squared sum of the streams' means
  - pcc: an approximated covariance between two streams
- 2,426,574 observations

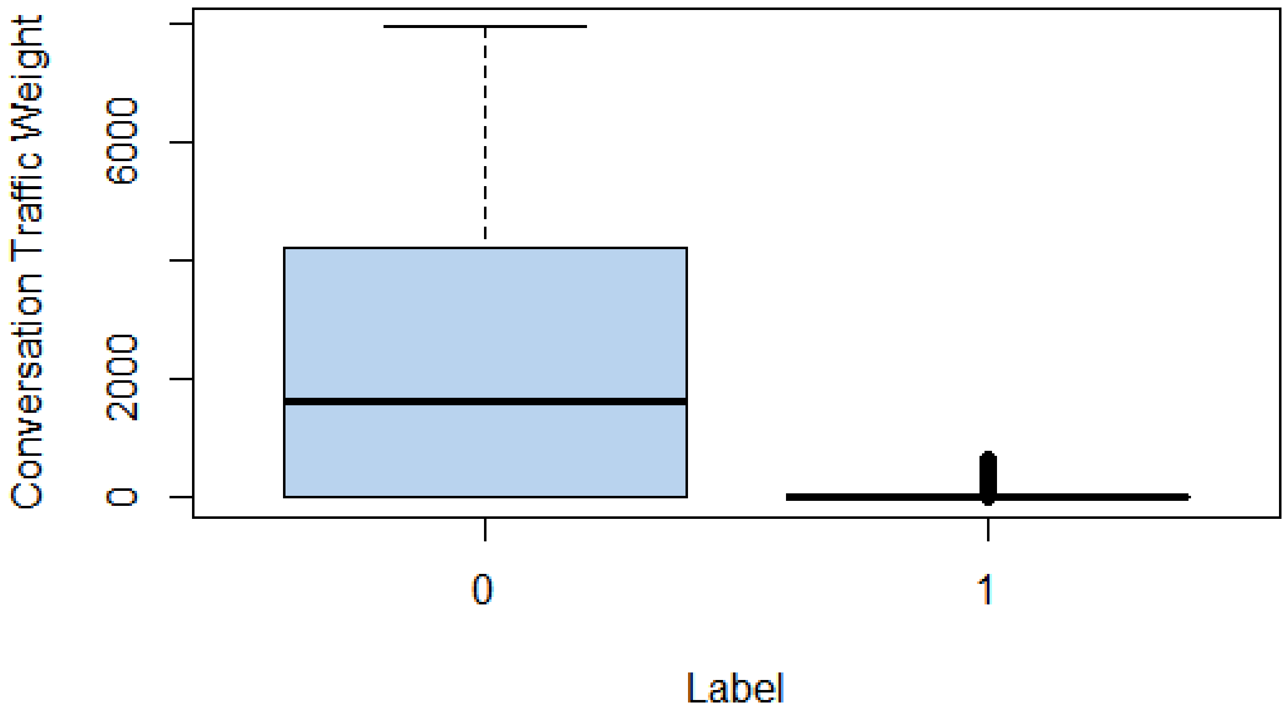
DATA PREPROCESSING

- Recoded "label" as a factor
- Examined correlation matrix and removed H (all recent traffic) due to nearly perfect correlation with HH (conversation traffic)
- Used absolute value transformation of "pcc" variables



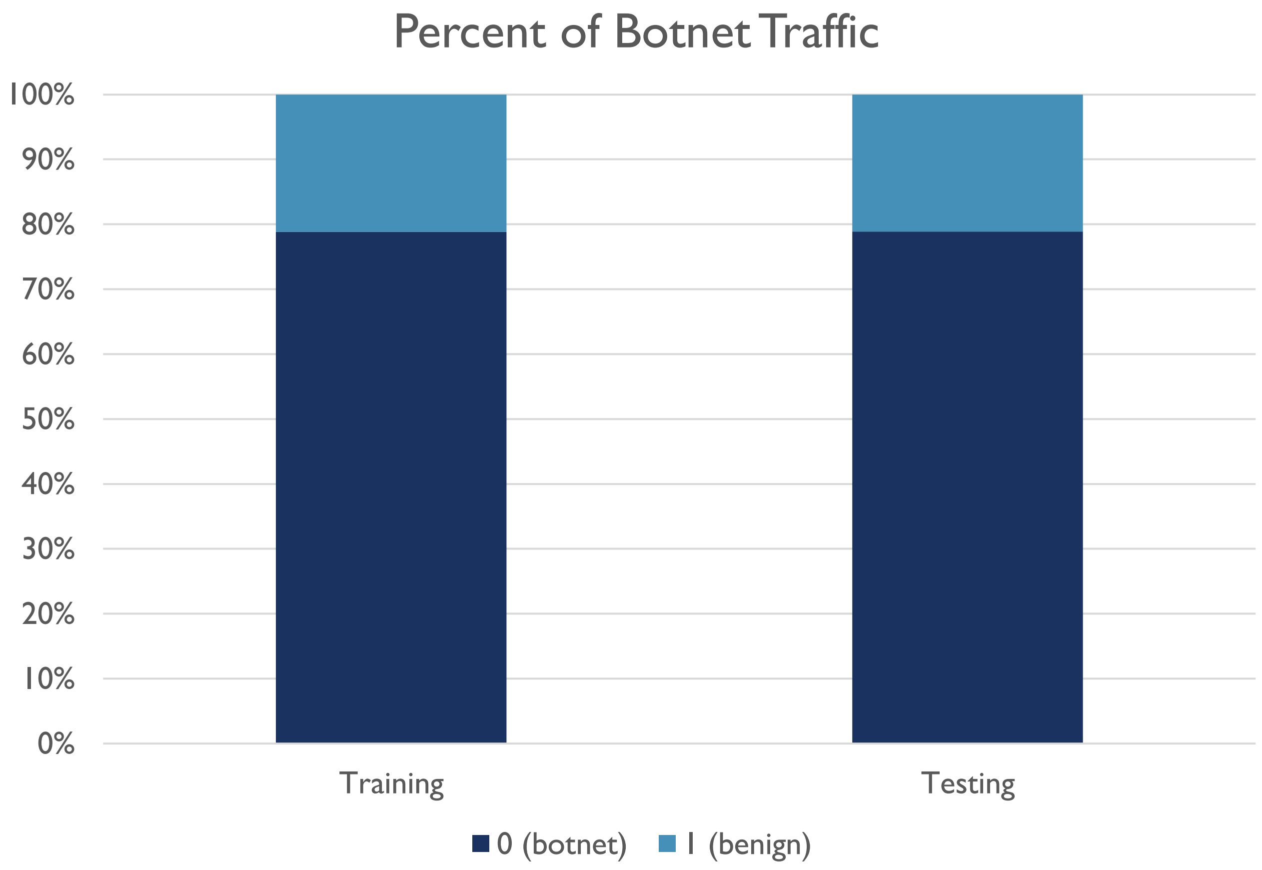
DATA LIMITATIONS

- The botnet traffic statistics differ significantly from regular traffic
  - Logistic regression resulted in probabilities of 0 or 1
  - Rule-based thresholds could misclassify a significant portion of infected networks, resulting in undetected intrusions

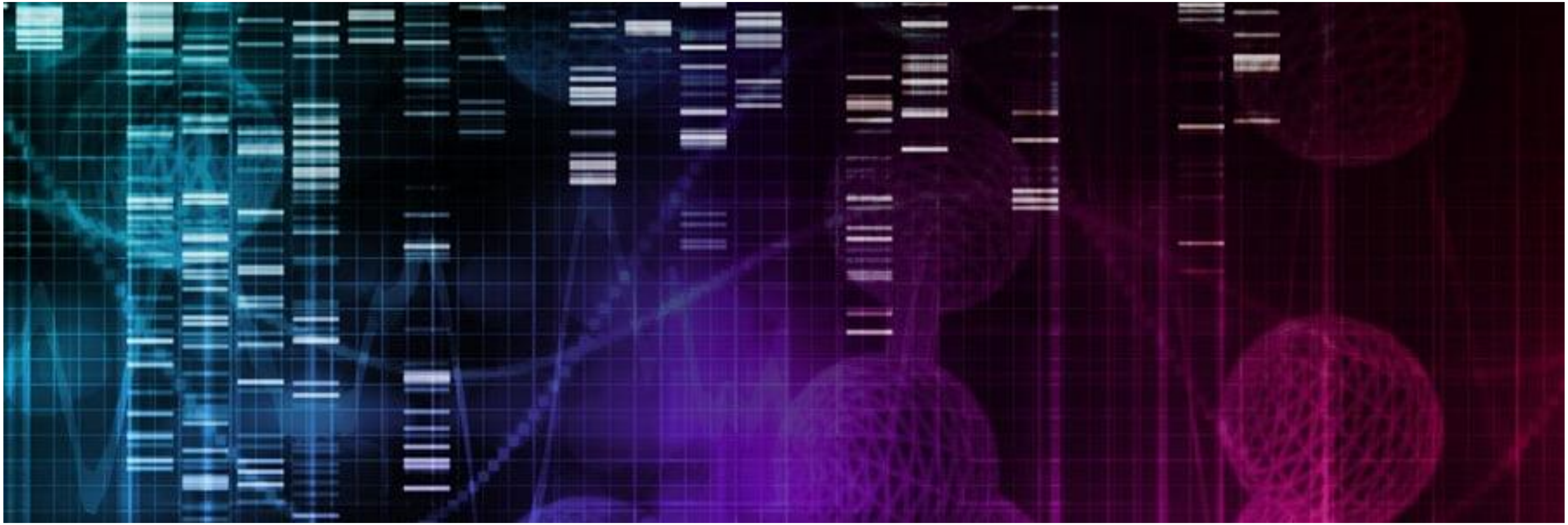


TRAINING AND TESTING SPLIT

- 70/30 training and testing split
  - Improved computational efficiency over 80% training data due to large number of observations
  - 1,698,601 training observations
  - 727,973 testing observations

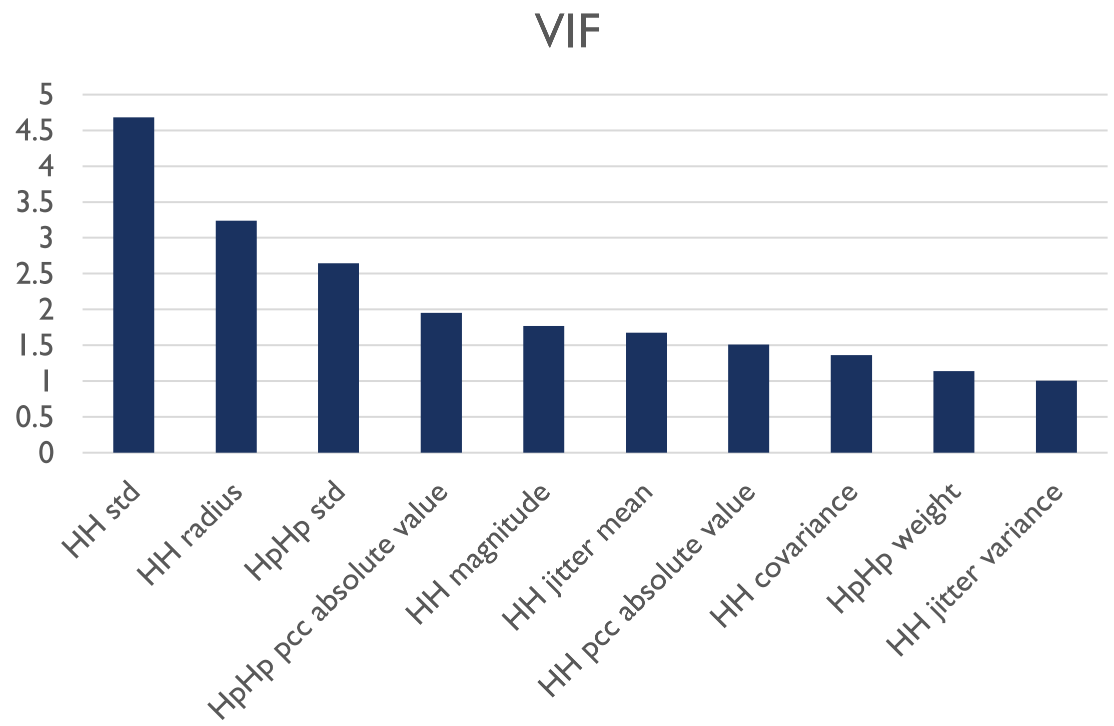


MODELING DETAILS & KEY INSIGHTS



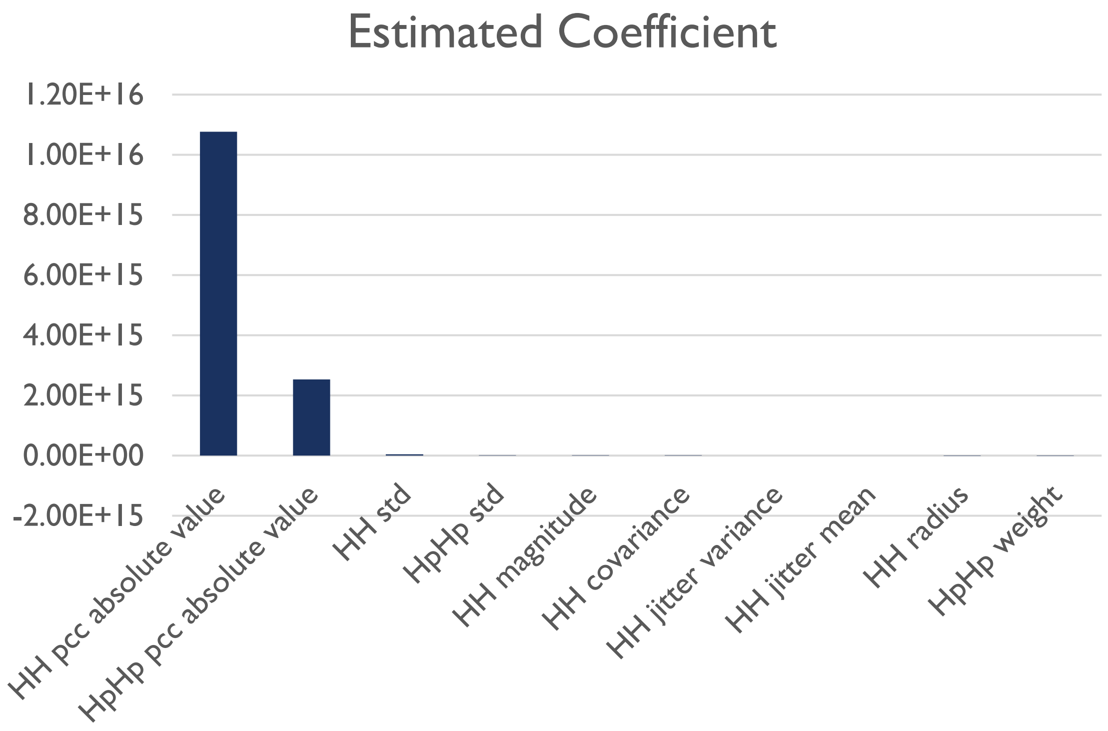
LOGISTIC REGRESSION: VARIABLE SELECTION

- Sequentially removed variables with high Variance Inflation Factors (VIF)
- 10 remaining variables with VIF less than 5.0
  - Subset used in subsequent models to avoid variable redundancy



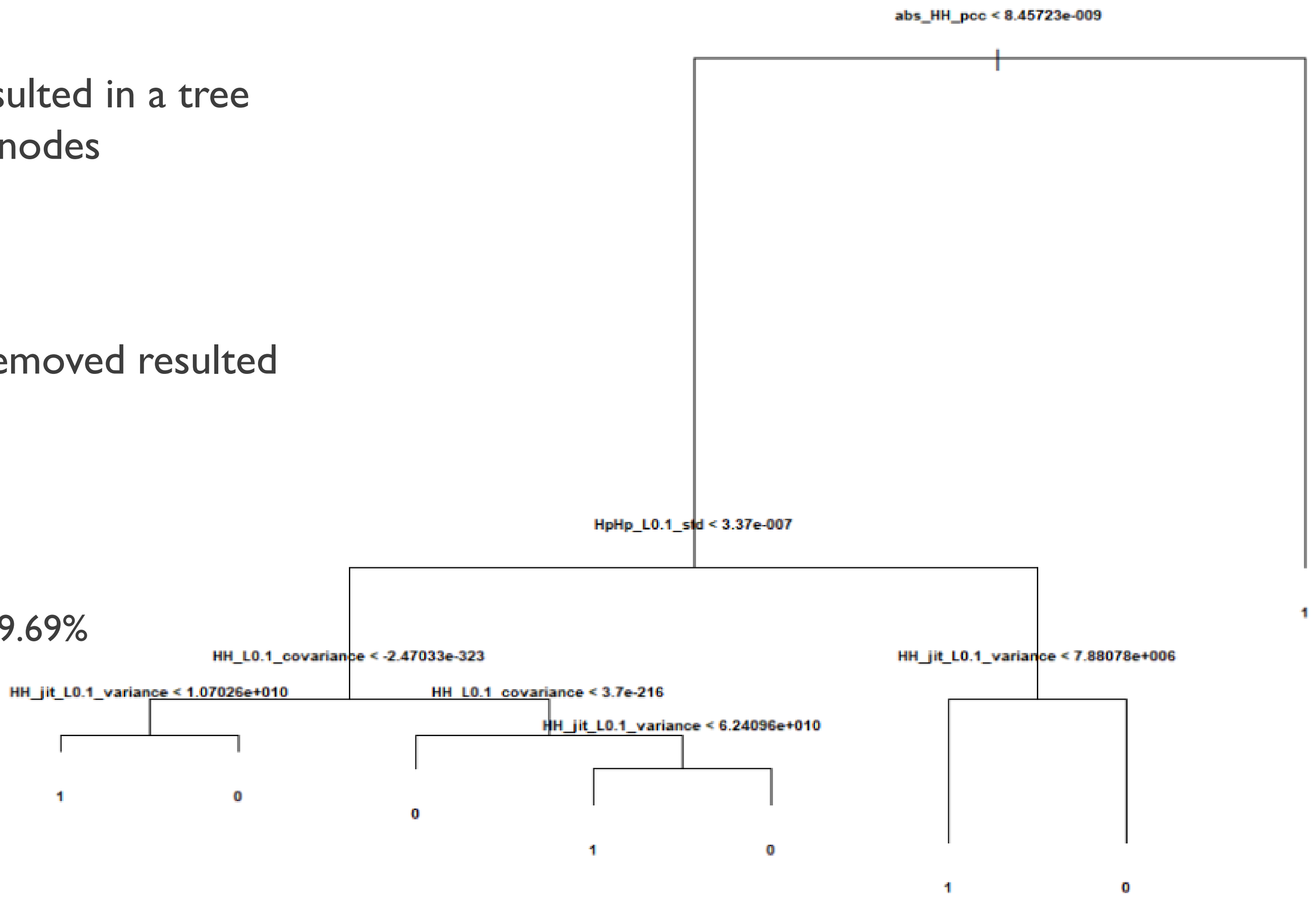
LOGISTIC REGRESSION: RESULTS

- All 10 remaining variables were highly significant
- Greatest estimated effect from the absolute value of the estimated covariance for conversation traffic by IP address
- Prediction Performance:
  - Overall Accuracy: 87.46%
  - “Negative Predictive Value”: 85.96%
    - Intrusions coded as “0”
    - Emphasis on detecting high percentage of intrusions due to low impact of false positives



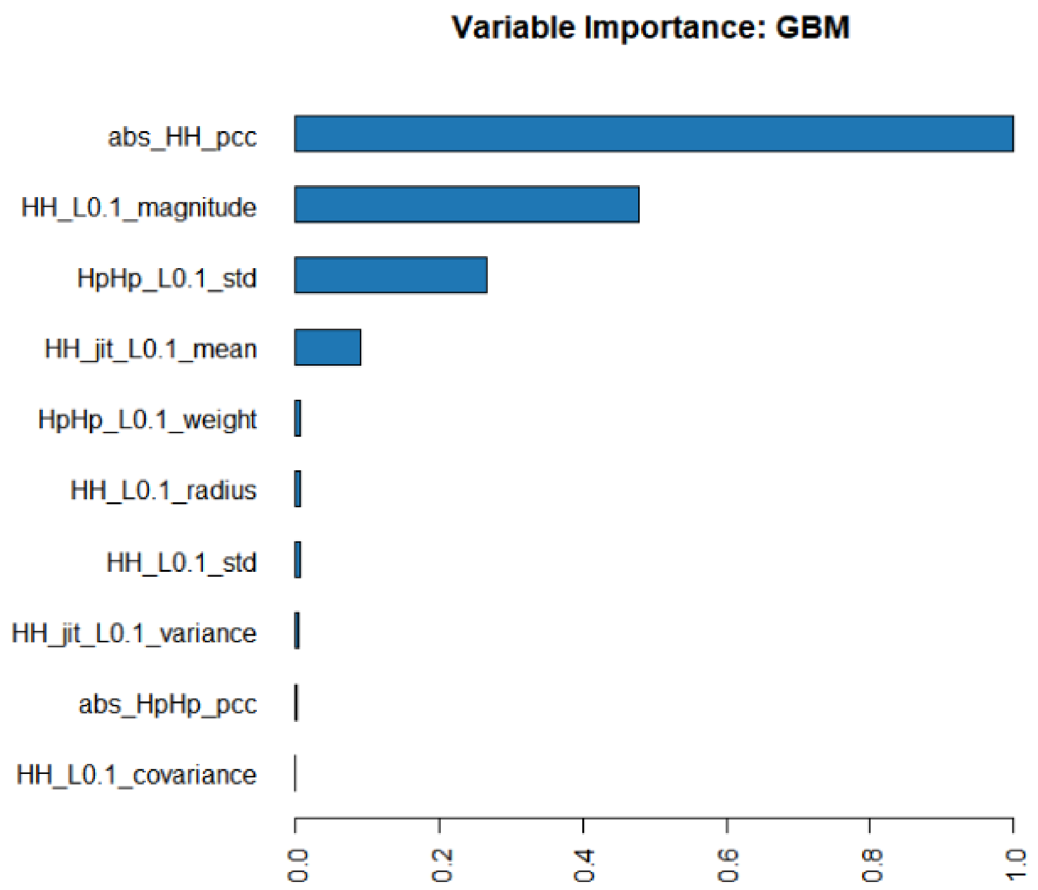
DECISION TREE

- Cross-validation with all variables resulted in a tree with 1 decision node and 2 terminal nodes
  - *HpHp weight* < 1
  - All traffic predicted as botnet traffic
- Cross-validation with *HpHp weight* removed resulted in a tree with 8 terminal nodes
- Prediction Performance:
  - Overall Accuracy: 99.29%
  - Percent of Botnet Traffic Detected: 99.69%



GRADIENT BOOSTING MACHINES (GBM)

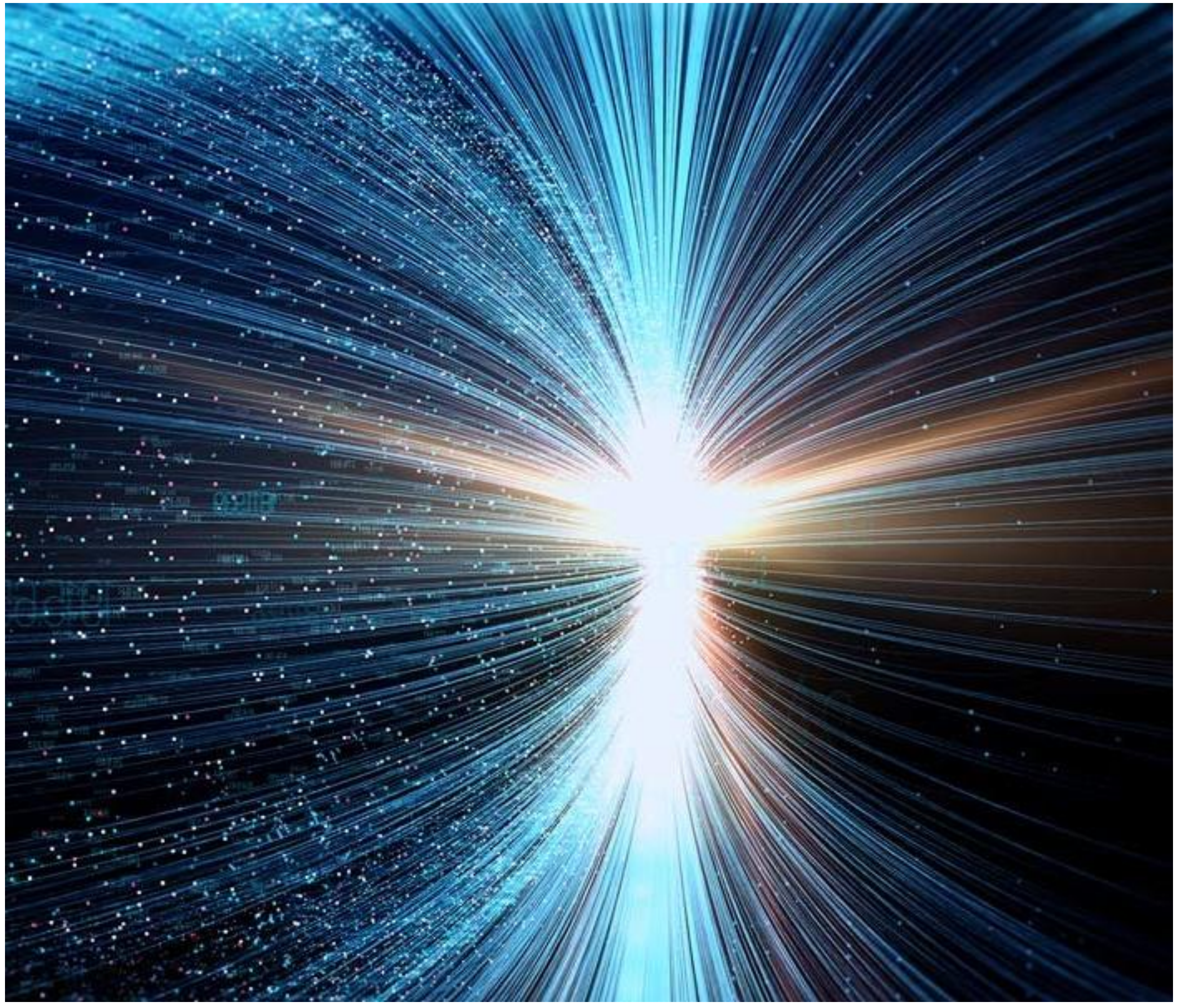
- *h2o* package in R
- 5-fold cross-validation
- 5 stopping rounds to prevent overfitting
  - 184 trees in final model
  - 25.6 mean leaves
- Prediction performance on testing data:
  - Overall Accuracy: 99.98%
  - Percent of Botnet Traffic Detected: 99.98%



MODEL PERFORMANCE COMPARISON

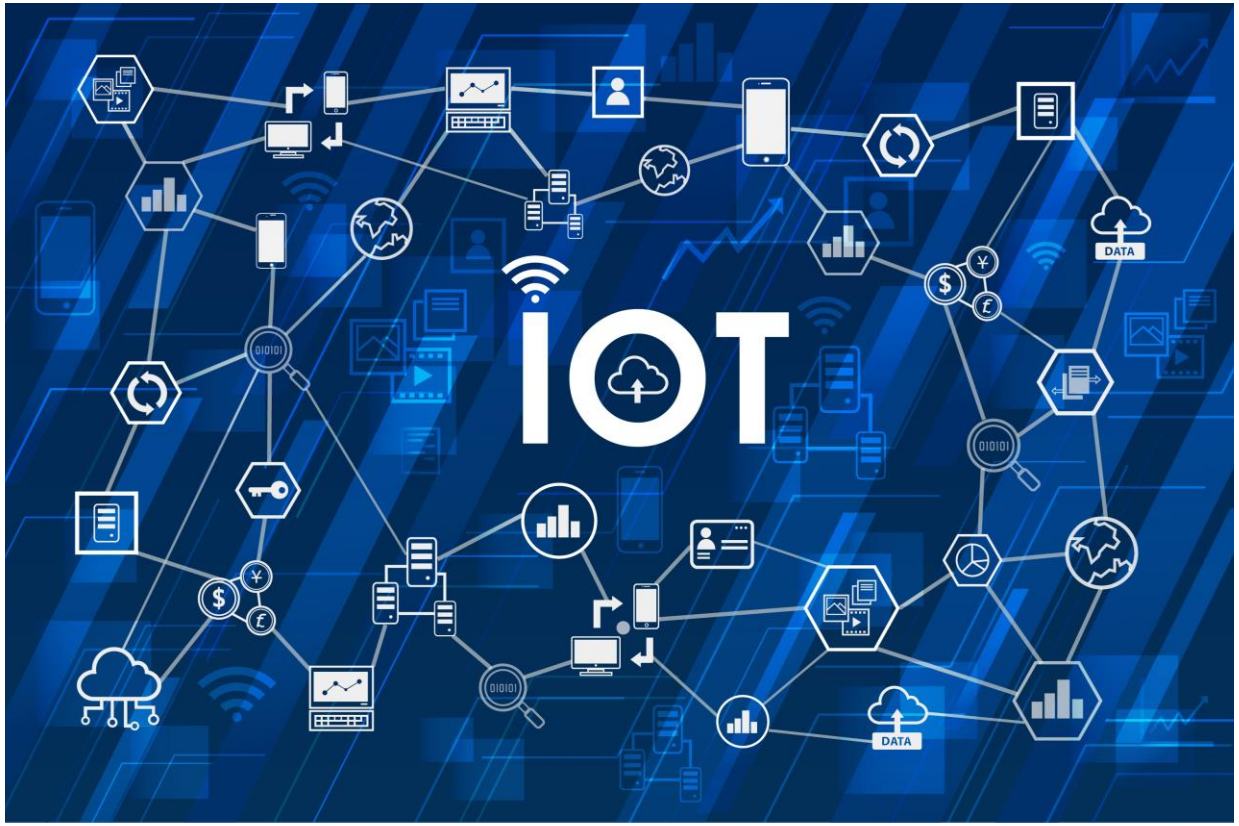
Model	Overall Accuracy	Botnet Detection Rate
Logistic Regression	87.46%	85.96%
Decision Tree	99.29%	99.69%
GBM	99.98%	99.98%

- Gradient boosting increases the botnet detection rate by 0.29%
- Small increase can translate to a large number of devices and detection within fewer time windows
  - Amounts to 145,000 devices in largest recorded botnet



CONCLUSIONS

- Models can be tailored to business needs
  - Less processing power in IoT devices
    - Test smaller subsets of features to implement real-time monitoring with lower computational demand
    - Hybrid rule-based and predictive modeling detection systems
- Data limited to IoT devices
  - Additional data collection necessary for networks with more complex devices and internet traffic



SOURCES

- A. Alhowaide, I. Alsmadi, J. Tang. "IoT dataset for Intrusion Detection Systems (IDS)", Kaggle.com, BotNeTIoT-L01_label_NoDuplicates.csv, https://www.kaggle.com/datasets/azalhowaide/iot-dataset-for-intrusion-detection-systems-ids
- A. Alhowaide, I. Alsmadi, J. Tang. "Towards the design of real-time autonomous IoT NIDS", Cluster Computing (2021), pages 1-14, Jan 2021.
- A. Alhowaide, I. Alsmadi, J. Tang, "Features Quality Impact on Cyber Physical Security Systems", 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Oct. 2019.
- Knecht, Tobias. "A Brief History of Bots and How They've Shaped the Internet Today." https://abusix.com/resources/botnets/a-brief-history-of-bots-and-how-theyve-shaped-t he-internet-today/