# Summer Of Code Internship Report –Trust Lab(2025)

Title: ML Integration with SOC

We were given 6 problem statements that needs to be resolved before the end of the internship. The 6 problem statements are namely:

* Phishing URL Detection
* Spam Email Detection
* Anomaly Detection
* Log Parsing
* Data Leak Prevention
* Firewall Optimization

We have finished Building the models for the first two problem statements

1)Phishing URL detection:

The model extracts unique features from the URL which totals to 46,and then when used together with Logistic regression, helps to classify URLs between phishing and non phishing URLs.

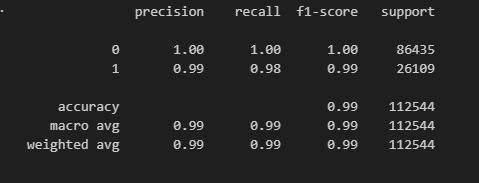


Fig.1: Classification Report for Phishing URL Detection model

Model Summary:  
 The model achieved a recall of 98% at a precision of 99%, and uses approximately about <400MB of RAM.

2)Spam Email Detection:

We use TF-IDF vectorizer to extract valuable keywords from the email. Logistic Regression is then used to classify the emails based on the extracted features.

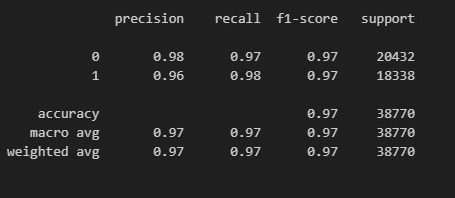


Fig.2: Classification Report for Spam Email Detection model

Model Summary:  
 The model has a recall of 98% with precision of 97% and utilizes minimal RAM usage.

Conclusion:

We used Logistic regression along with various other feature extraction methods to effectively minimize the RAM utilization while maximizing recall and precision of the model.

Next Steps:

We are currently working on the preliminary stage of Anomaly detection. The log needs to be thoroughly scoured so as to not lose any critical information while extracting features from the logs. Once this is completed, we will proceed sequentially with Log Parsing, Data Leak Prevention, and Firewall Optimization models

Team Primum: Sinan & Paul