Smridhi Patwari
95-767 Cybersecurity for AI and ML

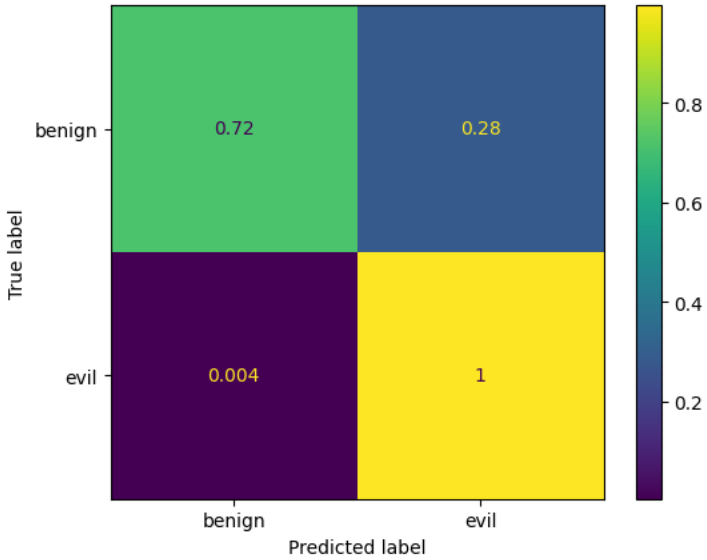
Assignment 1 Answers

1) **Interpretability - Describe what the visual shown in the last step (# Compute and plot performance metrics as a "confusion matrix") is representing. What meaning should a viewer draw from this and how is it determined?**

   The visual in the last step is called a confusion matrix. Its goal is to visually show the true positive, true negative, false positive, and false negative classification rates of the anomy detection model. The matrix is determined by running the trained model on a labelled evaluation set whereby each prediction is compared to the true label. The outcomes are put together in the 4 categories and normalised for easier comparison across classes.

   From the confusion matrix, 72 percent of the benign events are correctly identified as benign whereas 99.6 percent of the evil events are correctly predicted as evil. 0.4 percent of the benign events are wrongly identified as evil and 28 percent of the evil events are wrongly identified as benign.

   From the results, it seems like the model is pretty accurate in determining the evil events with a low false negative rates. It is generally good for security use cases since most of the malicious activity is being flagged. However, it does have a significantly high false positive rate, which means that security teams will have to spend additional resources on in investigating these false positive events. However, this could be intentional in some security use cases which might prioritize a more aggressive model with very high true positive rates for the evil class over the high false positive rates of the benign events being flagged as evil.



2) **Explainability - Describe for a novice end-user how this model classifies evil and benign traffic in step "Fit an anomaly detecting isolation forest model to the engineered features". Consider what are the model inputs/features, how does the model process them, what is the output displaying?**

The model isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This process is repeated to see how quickly each record can be separated from the rest of the data. If a traffic gets separated in only a few steps, it behaves in an unusual way and stands apart from the other records. However, this behaviour is rare compared to compared to normal system activity and as such the model is able to identify it as suspicious and labels it as evil. In contrast, normal traffic behaves similarly to many other records and is harder to separate, requiring more steps before it can be isolated. These records can be labelled as benign. After the model has performed these random splits and isolation, it outputs an anomaly score for each traffic record. The outputs are subsequently displayed, with a higher score (closer to 1) showing that the traffic record was isolated in a very few splits with extreme or rare feature values vs the lower score (close to 0) can be interpreted as the traffic record requiring many splits to isolate with feature values similar to many other records.

3) **If an attacker knew you were doing anomaly detection such as in this Assignment for a network they planned to attack, how might they disguise their attack? Make your answer specific to the data available in the BETH dataset.**

Since the isolation forest model classifies unusual traffic quickly, the attacker could use a slow attack strategy to avoid being detected. They could spread malicious actions over time by keeping packet rates and sizes consistent with benign traffic. The features with the *packet counts*, *byte totals*, *flow duration* can stay within normal limits such that each individual flow appears harmless. The attacker can also avoid any unique combination of features that could be easily classified by the isolation forest. These measures will increase the path length, eventually decreasing the anomaly score and fewer detections.

4) **As a countermeasure, what step or steps might you add to your analysis to prevent the attacker from performing such disguise?**

I would add feature-relationship checks that evaluate ratios between existing features. For example, ratios such as bytes sent to bytes received, bytes per packet, and packets sent per second can be introduced as new engineered features and used alongside the original features. This improves detection because attackers often focus on keeping individual feature values within normal ranges, especially when performing slow or stealthy attacks. However, they may overlook abnormal relationships between features, which can produce more extreme minimum or maximum values for these ratio-based features. As a result, these traffic records are isolated more quickly by the Isolation Forest and receive higher anomaly scores, making disguised attacks harder to evade.

5) **If you wanted to make this isolation forest model available to SOC analysts, how would you address the high rate of misclassifying benign records as**

malicious (i.e., false positives)? Your solution can include advice/guidance on how to use the model and/or technical workarounds.

Instead of alerting per record, one workaround would be to alert per source based on the source ID and destination IP, since many malicious attacks generate multiple flows with similar traffic patterns. Aggregating alerts in this way reduces noise from isolated benign outliers and helps SOC analysts focus on sources that repeatedly exhibit anomalous behavior. In addition, explainable AI techniques can be adopted to show why the model generated an alert, such as the anomaly score, the most influential features, and comparisons to normal traffic ranges. These explanations help analysts quickly understand the reason for the alert and decide whether targeted investigation and response are necessary, thereby reducing the impact of false positives.