ASSIGNMENT 2:

NAME: Rohan Bhowmick

SUID: 658096139

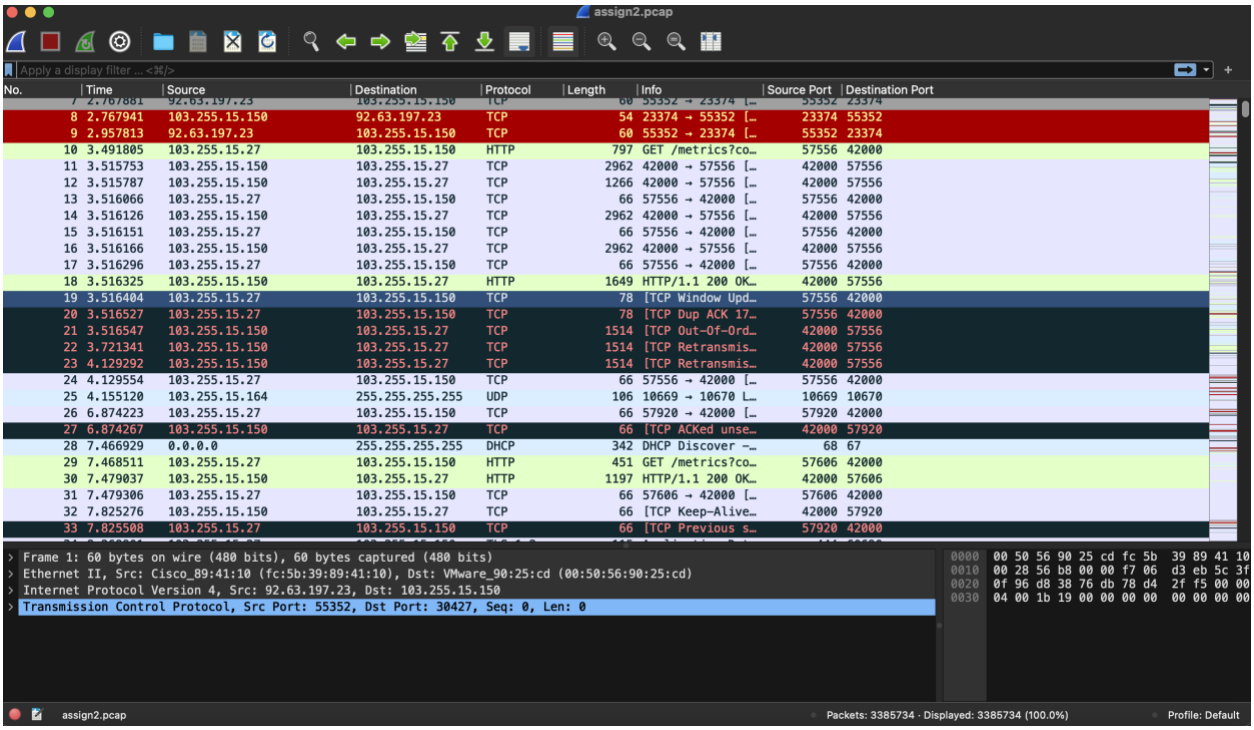
Email: rbhowmic@syr.edu

Solution of the assignment question and the code for it given below:

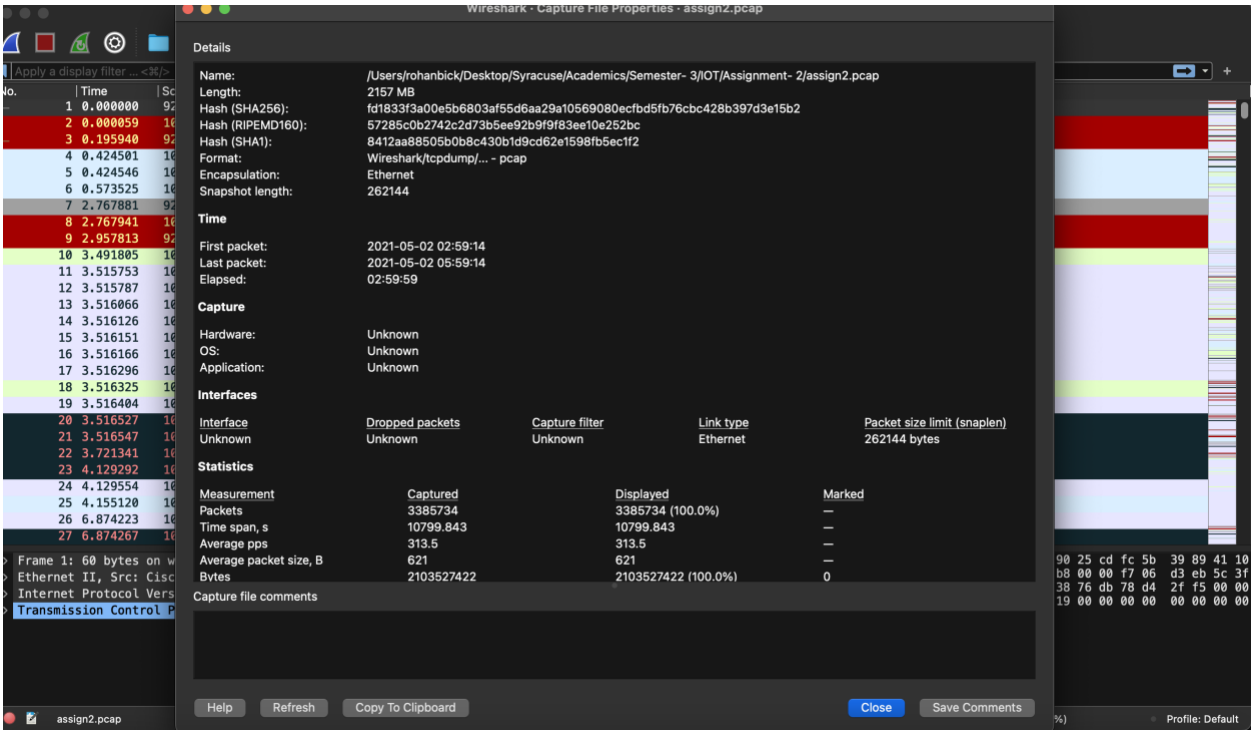
1) number of packets in total 3385734.

   Dataset size: 2.2 Gb.

   Protocol used: HTTP.



Capture file properties of the .pacp file:



About the code:

- Using the extract_packet_features(packet) function, relevant features of a packet will be extracted. It returns to a dictionary containing the features that were retrieved from the packet after doing so.
- Packet length statistics can be computed and shown using the DataFrame's calcStats(df) method.
- The traffic labels 'Malicious' or 'Normal' are applied using the label_traffic(row) method and depend on the specific conditions. The label for the packet that was successful will be returned once it receives a row from the DataFrame that represents a packet.
- Preprocess a PCAP file using the method preprocess_pcap(pcap_file, output_csv) to extract features and store the results to a CSV file.
- To start the machine learning code, we divide the features and labels, then we perform a one-hot category column encoding. We train the model using training data and make predictions using testing data after dividing the dataset into training and testing portions.
- Following the prediction, we would compute the F1 score, accuracy, precision, and recall.

About the performance of the model:

- Accuracy: A high accuracy model suggested that the model was correctly predicting the outcome. Since most of the time our dataset is not well balanced, we also need to apply other methods like precision and recall in addition to accuracy.
- Precision: Precision is defined as the ratio of the number of true positives the model detects to the total number of true positives and false positives in the system. A higher precision indicates that our model is more adept at identifying fraudulent traffic.

- Recall: It is the system's capacity to record every positive class instance. The ratio of true positive to the total of false negative and true positive is what it is.
- F1-Score: It is the accurate and harmonic method of recollection. When attempting to strike a balance between recall and accuracy, the F1-score could be helpful.

Factors of strengths:

- **Random Forest's adaptability**
- **Managing Classification Features**
- **Measures of Evaluation**

Factors of weakness:

- **An Unbalanced Set of Data**: A dataset that is significantly uneven may have a biased model that favors predicting the majority class if the bulk of the samples fall into a single class (like "Normal"). Consequently, it is possible to detect poor communications with a high degree of accuracy but little effectiveness.
- **Engineering features**: The features that are included have a big impact on how effective the model is. To increase performance, make sure the pertinent features are being used and think about experimenting with additional features.
- **Using IP addresses correctly**: One-hot encoding of IP addresses can produce a variety of features, particularly if the dataset contains many unique IP addresses. This could lead to the dimensionality curse and affect model performance.

Some suggestions for Improvement:

- Enhance Feature Engineering
- Deal with the Unbalanced Dataset
- Evaluation continuity

In summary, the model's output should be evaluated using a variety of metrics, and the advantages and disadvantages of the strategy should inform upcoming changes and enhancements. Modifications may be made in response to new information, domain expertise, or variations in the properties of the network traffic.

Main Results and Conclusions:

- Model Operation: Metrics such as recall, accuracy, precision, and F1-score show how well the Random Forest model performs. It was emphasized how important recall and accuracy are, especially when it comes to spotting fraudulent or large traffic.
- Strengths: One-hot encoding handled categorical features well, and the model showed potential for adaptability. Numerous evaluation metrics facilitated an in-depth examination of the model's functionality.
- Limitations: Unbalanced datasets may cause bias towards the dominant class, which could impair the model's capacity to identify fraudulent traffic. Further optimization could be beneficial for feature engineering, particularly in IP address management. In certain situations, the Random Forest model's interpretability could be difficult.