Classification of Malicious Activities

IDS Group 081 – Kuppuraj M, Sreenivasa R Pidugu, Penchal Poli,

Introduction to Data Science

M.Tech Data Science and Engineering – Cluster Batch 4

05-Mar-2021

Overview

- Objective: Perform an efficient classification of malicious activities using machine learning techniques.
- Methodology: Build a ML model using Logistic Regression and Decision Tree That can classify the activities.

Dataset

- How many features: 43
- Size of the dataset: 148515 records
- Multiple files: no
- What kind of data: numerical and character
- Balanced or imbalanced: imbalanced
- Distribution of Training set, validation set, testing set
- Missing data and preprocessing challenges: No missing data, had to perform feature engineering technique to convert string data to numerical data.

Methodology

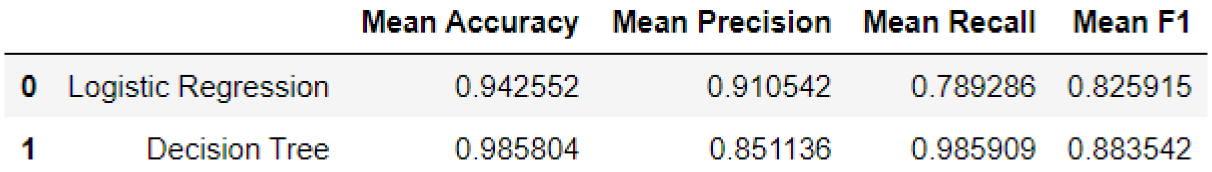
- The 2 classifiers used: Logistic Regression, Decision Tree
- Ensemble pipeline: No
- Other models considered: No
- Hyper-parameter tuning: No

Feature Engineering Techniques

- Features removed: “level”
- Feature creation: Yes, “attack”, “service”, “protocol_type”
- Feature ranking: Yes, using Pearson correlation
- Class imbalance treatment: No
- Any other: Plotted attack vs. protocol graph to find which protocol is highly related with the attacks.

Results

- Table for the evaluation metric for each ML technique used



		Mean Accuracy	Mean Precision	Mean Recall	Mean F1
0	Logistic Regression	0.942552	0.910542	0.789286	0.825915
1	Decision Tree	0.985804	0.851136	0.985909	0.883542

- Plot of the curves : No
- Conclusion: Based on the above results, we see the accuracy of the classification using Decision Tree is high compared to Logistic Regression, which can be used to predict the traffic that comes in.