uOttawa

Professional Master's in Artificial Intelligence
AI for Cybersecurity Applications (ELG7186)

Subject: Assignment 2 – Part 2 (Report)

By
Mohamed Sayed Abdelwahab Hussein
Mhuss073@uottawa.ca
300273145

Under Supervision
Prof. Miguel Garzon

1- Overview

We developed a binary classifier model that can predict data exfiltration via DNS, and we used a data stream (local Kafka Server) to test our model as a real-world scenario.

2- Algorithms

We begin our methodology by loading and pre-processing the dataset; as shown in Figure 1, the target class distribution is almost balanced.

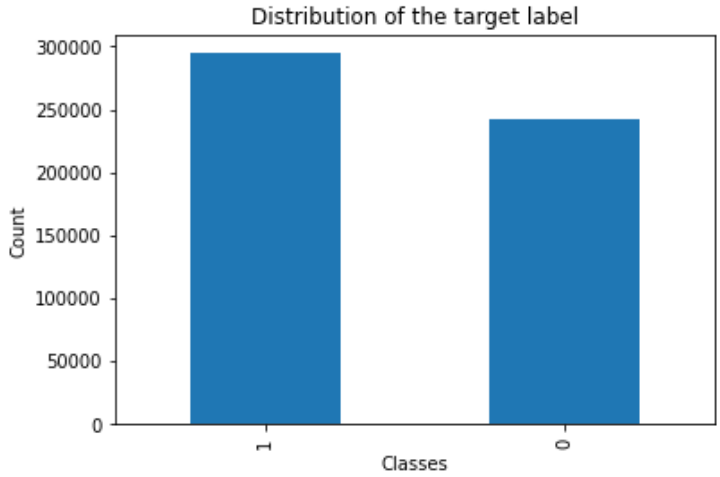


Figure 1. class distribution

To evaluate our model's performance, we divided the dataset into 80 % training and 20 % testing and compared three models to choose the best one.

2.1 CatBoost

CatBoost machine learning algorithm was recently released as open-source library. It can operate with a wide range of data formats to help businesses solve a variety of problems [1].

2.2 Random Forest

Random Forest is a classification technique that uses a large number of decision trees to classify data. When constructing each individual tree, it employs bagging and feature randomization in order to generate an uncorrelated forest of trees whose committee forecast is more accurate than any one tree's [2].

2.3 Logistic Regression

In its most basic form, logistic regression is a statistical model that utilises a logistic function to represent a binary dependent variable [3].

3- Experiments

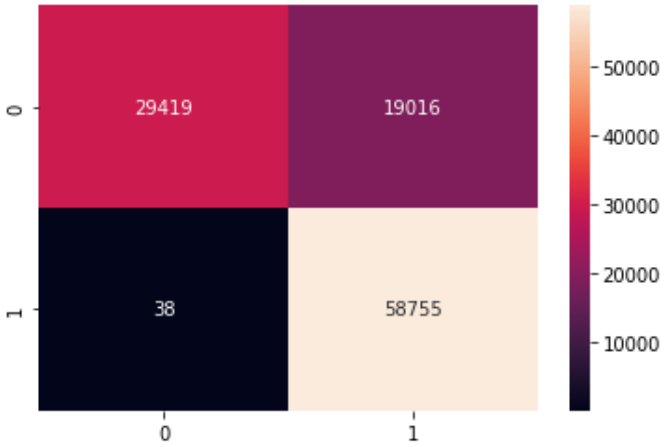
3.1 CatBoost

Hyper-parameters {'iterations': 1000,
'loss_function': 'Logloss',
'depth': 4,
'learning_rate': 0.02,
'train_dir': 'crossentropy',
'allow_writing_files': False,
'random_seed': 4}

We tested the algorithm using accuracy, precision, recall and F1 score.

Test Accuracy = 0.8223038758533219

	precision	recall	f1-score	support
0	1.00	0.61	0.76	48435
1	0.76	1.00	0.86	58793
accuracy			0.82	107228
macro avg	0.88	0.80	0.81	107228
weighted avg	0.87	0.82	0.81	107228



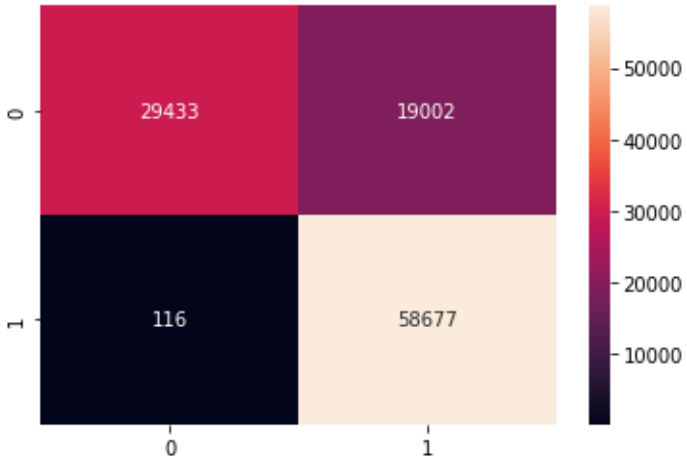
3.2 Random Forest

Hyper-parameters {n_estimators=1000, max_depth=4}

We tested the algorithm using accuracy, precision, recall and F1 score.

Test Accuracy = 0.8217070168239639

	precision	recall	f1-score	support
0	1.00	0.61	0.75	48435
1	0.76	1.00	0.86	58793
accuracy			0.82	107228
macro avg	0.88	0.80	0.81	107228
weighted avg	0.86	0.82	0.81	107228



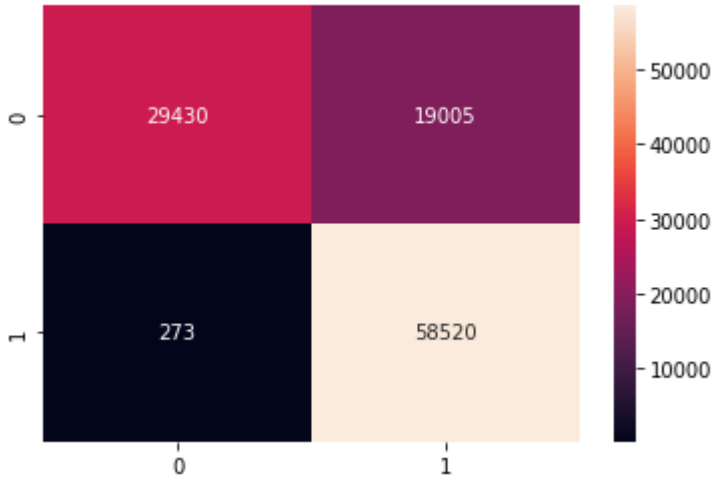
3.3 Logistic Regression

Hyper-Parameters {penalty= 'l2', C= 1.0}

We tested the algorithm using accuracy, precision, recall and F1 score.

Test Accuracy = 0.8202148692505689

	precision	recall	f1-score	support
0	0.99	0.61	0.75	48435
1	0.75	1.00	0.86	58793
accuracy			0.82	107228
macro avg	0.87	0.80	0.81	107228
weighted avg	0.86	0.82	0.81	107228



4- Discussion

As we can see from the above findings that the accuracy of all three models is almost the same. So, since our problem is sensitive to detecting DNS attacks, we need to look at different metrics to choose the optimal one. We should concentrate on False Negative and choose the minimum value. From confusion matrixes, we'll choose CatBoost as our champion model, which has a FP of 38, whereas Random Forest has a FP of 116, and Logistic Regression has a FP of 273.

5- References

[1] https://catboost.ai/

[2] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[3] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc