ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES
BERNWARD ASPRION, 29.11.2024

Informatik & Security

fh st.pölten

MCR DAS WS24
Homework2: Classification- Botnet Detection

Deadline: 11.12.2023 08:50

ML algorithms used for network anomaly detection, can also identify botnets.
The selected features in our example consist of the port number, values of network latency and network throughput.
In our threat model, anomalous values associated with these features are representative of the presence of a botnet and encoded with a “1”.

The decisive factor for the assessment is not the execution of program code. The decisive factor is the verbal description of evaluations, graphics and reasons why you did what you did.
Evaluations or plots without comments do not generate any points!

The use of **Classification-Notebook-V8.ipynb** (available on eCampus) is strongly recommended.

a) Load the data from “botnet-network-logs-v2.csv” in a DataFrame.
b) What scale of measurements do the variables have?
c) Analyze and manipulate the data, to get an understanding of the data and to prepare the data for modeling.
d) Split the data in a training and in a test set (75%-25%-Split).
   Use the stratify option in the method „train_test_split“ from sklearn (add: stratify = y).
   Read the documentation “train_test_split” from scikit learn (sklearn).
   What is the default value of the stratify-option?
   What does „stratify“ mean?
   Bonus points: check if the stratify-option worked like expected.
e) Use Logistic Regression, print the confusion matrices and the accuracies on test and training.
f) Chose a threshold with a minimum True Positive Rate (TPR) of 60% with logistic regression on train data. What TPR do you get on test with this threshold and logistic regression?
g) Use Decision Tree Classifier, print confusion matrices and accuracies on test and training.
h) Use Random Forest Classifier, print confusion matrices and accuracies on test and training.
i) From the results from (e) to (h) – which classifier would you prefer?
j) Use cross validation with 4 splits and 5 repeats, with the score “accuracy”, and the classifiers Logistic Regression, Decision Tree and Random Forest.
   From these results – which classifier would you prefer and why?
k) Create a fit of the preferred model (final model), that is to be deployed.

RULES
The same rules apply as for the first homework (see MCR-DAS-WS24-Homework1-v02.pdf).