Faculty of Engineering
School of Electrical Engineering and Computer Science

Assignment 3 Description Continuous Evaluation with Kafka

Course	ELG7186 – AI for Cybersecurity Applications
Academic year	2022-2023
Semester	Fall
Instructor	Paula Branco
Deadline	November 5th 2022 11:59 PM EDT

In this assignment, you will implement a binary classifier aiming at predicting data exfiltration via DNS. In this exercise there will be two tasks graded individually. You are expected to implement two predictive modeling solutions: the static model, and another solution that adapts through time.

For both problems you will have 2 independent datasets, that should be treated accordingly:

- A "static_dataset.csv" file which you can use to train a static model.
- A "kafka_dataset.csv" file that you should treat as a data stream (local Kafka Server) which will be used to evaluate the dynamic model.

Part I (Static Model):

Welcome, you are now part of the Canadian Centre for Cyber Security, yesterday the Prime Minister approved the financial support to create our group and our first task will be to create a static Machine Learning Model based on batch data. You will be provided with top secret files obtained from our allies Ring Canada (RC) and the Cyber Threat Intelligence (CTI). The dataset provided to you has DNS traffic generated by exfiltrating various file types ranging from small to large sizes. Your job here will be to follow the standard ML cycle and train the best model possible. Perform the following analysis and steps to your data and your model:

1. Data Analysis: Using the file called "static_dataset.csv", check using plots and statistical tools the distribution of each feature and the target variable, validate if your dataset is imbalanced and if you have any type of data skewed pattern. Justify all your findings with plots or graphs.
2. Feature engineering and data cleaning: Analyze the data inside the .csv file (static_dataset.csv) and transform the variables that contain string values, so that all of them can be used in the model. Check for missing values and categorical values. Tip: you might find useful an embedding technique or hashing
3. Feature Filtering/Selection: Use at least two different statistical techniques to evaluate which features are the best to train your model, give a final list of those you will use and justify your answer based on the results obtained from your analysis.
4. Model Training: Split the data using a method you find suitable and justify it. Normalize your data and train the selected model. In this point it is crucial that

you analyze which metrics and performance evaluation will be the best for your model. Select two learning algorithms and provide a comparison of their metrics. Select the one that shows the best performance to be reused on the second part.

5. Model evaluation: It will not be enough to split your data into train and test to justify the results of your model. Plot the performance metrics and perform a critical analysis of them. Save your model for using it in the second part.

Part II (Dynamic Model):

The Chief of Cyber Security has read your report and the managing directors are threatening with closing down our branch because even though your results were outstanding, we are not using enough technological tools as they would like. In order to fulfill the board's requests our analysis team has proposed to make an alternative analysis using Online Learning. The purpose of this second task will be to simulate a real-life scenario of a constant data stream. Make predictions on the new data that arrives in real-time and register the performance of two different models.

- Static model – will be the same that you saved in the previous task, and it will not change at all.
- Dynamic model – At the beginning, this model will be the same as the static, loaded from our previous task, you will then use windows of 1,000 observations and make an analysis of whether or not, it is a good idea to retrain the model. If needed, retrain and plot your results

For this task you will need to follow the next instructions:

1. Follow the instructions on the pdf called "setup_instructions.pdf", and make sure to install dockers, create the images, install Kafka and the dependencies.
2. Use the .csv file called "Kafka_dataset.csv" and use it according to the instructions to deploy de producer's task.
3. Load your .h5 (or other extension) model to the notebook
4. Make sure to create two models, one for the dynamic analysis and other for the static analysis.
5. Run the consumer's code and validate you are receiving the data stream.
6. Create the necessary code to append 1,000 observations of data streaming information and use them as a window.
7. Evaluate the dynamic model's performance and create a decision boundary to retrain the model or keep it as is.
8. Evaluate the performance of each model on each window.
9. Use the necessary arrays to store the performance metrics you consider best to evaluate the performance of both models and create a time-based plot comparing both performances from both models.
10. Decide when it is better to retrain the model and justify why.
11. Describe your results and make a conclusion of whether the dynamic implementation is better or not, compared to the traditional static model.

Report and deliverables:

Your performance will be graded with two notebooks and one report.

- Work with a .ipynb file. In both cases your code should be clean, it should have comments every few lines clearly describing the section's purpose. Please try to be as neat and efficient as possible. In this file we should be able to appreciate how you developed the exercise and also see the results of your code (prints and plots). You will only need to work on the consumer's file as the producer's file will be separated.
- Hand in a five pages report where you explain all points mentioned before and make sure all points in the grading document are covered. Be efficient with the number of graphs and plots you put in this section, you can use the appendix section described before, for providing more graphical information, but this section should not contain any analysis or description.

Please make sure you hand in a compressed folder with the following files:

- Static_model.ipynb
- Dynamic_model.ipynb
- PDF report max 5 pages

Doubts please contact TA Rodrigo Ledesma, rlede046@uottawa.ca