Week 1 Report

Haoyang Han

In this week we clarified the outline of this project and distinguished what specific kind of model should be used for this project. Since we looked through the database given and couldn't find the correct 'label'(as described, binary label of whether there exists an 'attack' or not), it's hard to say it's a supervised learning ML problem(binary classification) or unsupervised ML problem(clustering). So in this week, we explored the dataset, give description to important schema, and give a templates of how to do binary classification.(see appendix)

For supervised ML problem, our initial approach is to try logistic regression. To train the model, we could use cross validation on training data. If the probability of 'attacked' predicted is larger than 0.5, then we would say this place should be anomaly. We could use mis-classification rate for evaluation metric, but certainly F-1 score and AUC could also be good alternatives. Then we could gradually improve model to decision tree, random forest and other advanced models.

For unsupervised ML problem, we should explore more about the dataset to make sure what we classified. For example, in schema torgi, table gps_observation_points, a lot of rows do not have moving data, and if we add those columns into dataset, then those columns(or datas) would greatly disrupt the final output. So, if we distinguish that this problem is unsupervised in next week, we really should focus on ETL process during future several weeks. Then we should try k-means and GMM.

Next week, we should focus on:

1. Ask about whether we have a 'label' to distinguish whether this is a classification problem or clustering problem.
2. Explore the dataset properly and select the feature needed for model fitting.
3. Provide the pipeline for this prediction problem(baseline), and provide template of how to optimize the model performance step by step.

Below are the links I provide as a template for next week's modeling:

Outline and database exploration:
https://docs.google.com/document/d/1TKU468N8eggEg4Ba6lITCUl7qIlVz_B_9lYhpGm2OSU/edit?usp=sharing

Template:

https://github.com/HaoyangHan/sofwerx_intern/blob/master/Week%201/predictive%20pipeline.ipynb