SIT384 Cyber security analytics

High Distinction Task 10.2HD: Case study

Task description:

Learning from an imbalanced dataset is a classic problem in machine learning tasks. In practical10, we balanced the training dataset by under-sampling the majority class. However, there are also other methods, such as oversampling the minority class or enlarging the feature space with new information based on the given data.

In addition to the supervised learning algorithms we have introduced, there are also performance-boosting techniques that often yield good results, such as like XGBoost, LightGBM or CatBoost.

In this task, you are given a dataset "creditcard.csv" used in practical10. Based on the code provided in "Case study Data-Challenge_Anomaly-Detection.ipynb", try to find the "best" classification model by conducting exploratory data analysis, engineering the features, and training models using boosting techniques (random forest, SVM, and XGBoost, LightGBM or CatBoost), as demonstrated in the case study.

You are given:

- Dataset: creditcard.csv (available from Task10.1D or practical10), Case study Data-Challenge_Anomaly-Detection.zip, and notebook: Case study Data-Challenge_Anomaly-Detection.ipynb.
- Split datasets: X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.3, random_state = 0)
- Other parameters of your setting

You are asked to:

- read and understand the given case study.
- follow the steps outlined in the case study to find the best model on the "creditcard.csv" dataset, using the AUC metric.
- explain every step of your work, similar to what the given case study does.
- analyse and compare the results and present your findings.

Note: Some steps outlined in the case study might not be suitable for the given "creditcard.csv" dataset due to differences between the case study dataset and the creditcard dataset. If this is the case, please clearly explain the alternative steps you took.

Submission:

Submit the following files to OnTrack:

1. Your jupyter notebook. You can edit your report in text cells as the given notebook does, or submit two separate files (notebook and report).

Check the following things before submitting:

1. Add proper comments to your code

Acknowledgement:

Special thanks to Delarue Simon for sharing his attempt at the Data Challenge - Supervised Anomaly Detection. We are grateful for his contribution and willingness to share his work.