Report Assignment 2 (part 2)

ELG7186 - AI for Cyber Security Applications

Name: Abdallah Medhat Mohamed Rashed

Student number: 300273110

Algorithms:

I apply more algorithms on the data after read data, data visualization and preprocessing data to make binary classification such as CatBoostClassifier, random forest and logistic regression.

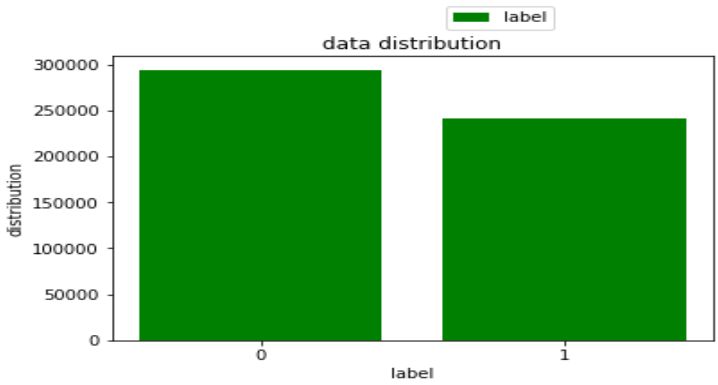
CatBoost: is supervised learning machine learning algorithm. It is used in classification and regression problems. It is used in solve a wide range of problems. It used ensemble methods which is Boosting [1]. It is used in python from catboost library.

Random forest: is supervised learning machine learning algorithm. It is used in classification and regression problems. It is consist of more decision tree. It used ensemble methods which is Bagging [2].it is used in python from scikit-learn library.

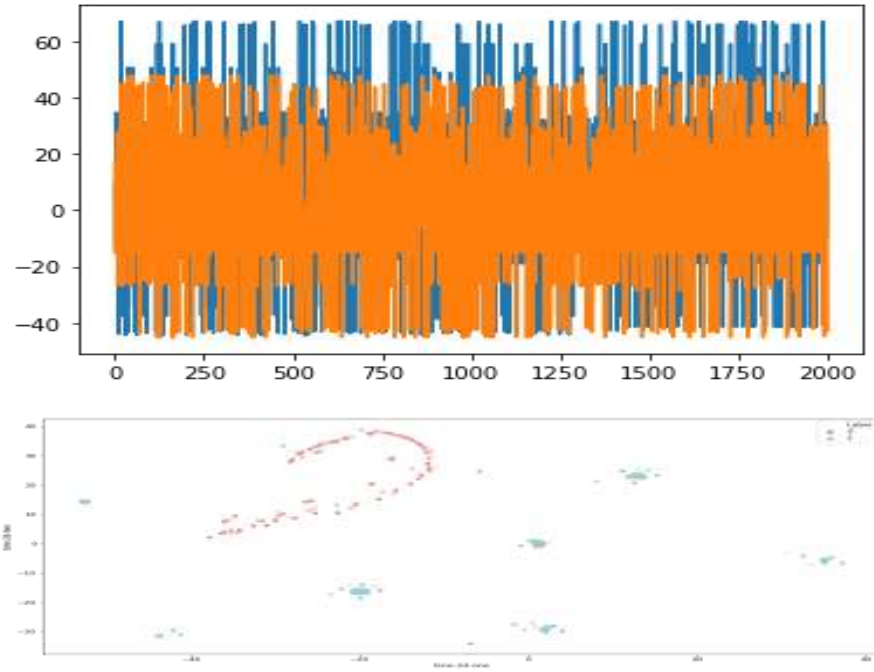
Logistic regression: is supervised learning machine learning algorithm. It is used in classification problems especially binary classification so I used it in my problems. It is used to find the relationship between features to classify classes [3]. It is used in python from scikit-learn library.

Experiments:

In the first, I do some data exploring to know imbalance data or balance data so I found the data is balanced which the number of classes in data is considered equal. The next figure show balance data.



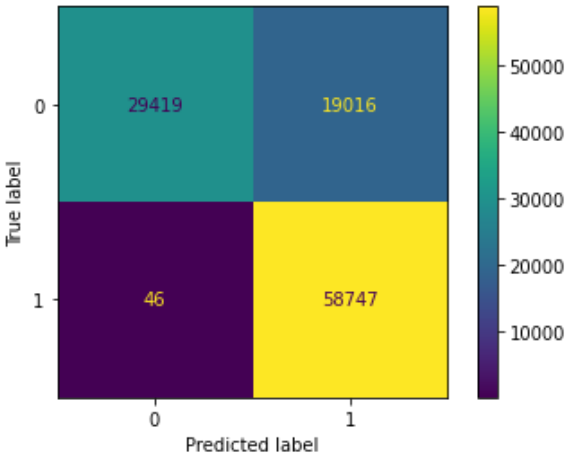
Then I apply T-SNE to check data sequence or not sequence and check linearity of data or not linearity. So I found the data is sequence and the data not linearity which is not splitting data by line. The first figure is showed the sequence data and the second figure show the not linearity data.



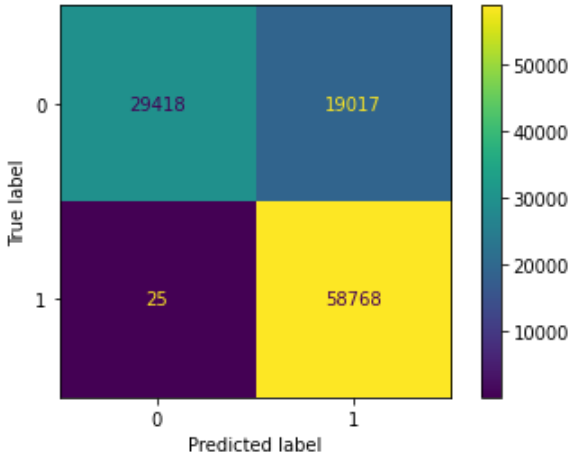
After founding data is not linearity, the data is need strong machine learning algorithms to obtain to high accuracy. I apply catBoost with hyper parameter (iterations=1000,verbose=0,learning_rate=0.01and random_state=0), random forest with hyper parameter (n_estimators=450, n_jobs=-1 and random_state=0) and logistic regression with default parameter expect random forest = 0 on the data after applying data preprocessing such as remove some columns which are timestamp, longest_word and sld so I found the accuracy and f1-score of catBoost and random forest are same (accuracy = 0.822 and f1-score =0.860) but the accuracy and f1 score of logistic regression is less

(accuracy = 0.820 and f1-score =0.858) because of the not linearity data. So I save the catBoost to predict streaming data from Kafka server.

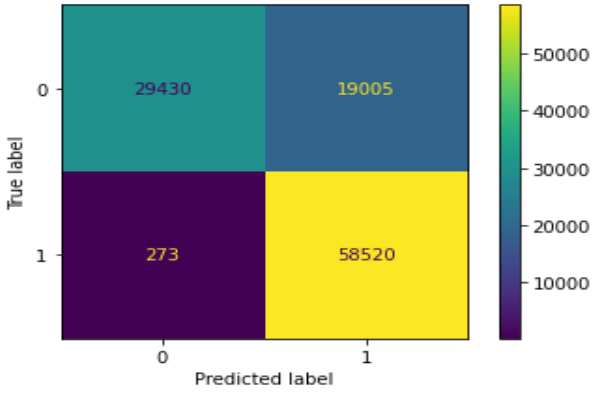
CatBoost



Random forest



Logistic regression



Reference:

[1] *CatBoost | CatBoost Categorical Features*. Analytics Vidhya. (2022). Retrieved 14 March 2022, from https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/.

[2] Education, I. (2022). *What is Random Forest?*. Ibm.com. Retrieved 14 March 2022, from https://www.ibm.com/cloud/learn/random-forest.

[3] *What is Logistic Regression? - Statistics Solutions*. Statistics Solutions. (2022). Retrieved 14 March 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.