MY APPROACH

Initially, the complete dataset was studied. The number of features and the tuples/incidents were analyzed.

The training dataset consisted to target attribute consisting of Class 0 and Class 1.

Class 0 being not an hack incident and Class 1 was a hack incident.

From the dataset values, Class 0 had 1068 events and Class 1 had 22788 events.

1.DATASET STUDY

The observation was that, the dataset was highly imbalanced.The both Train.csv and Test.csv dataset consisted of NAN values which are not accepted by machine learning models.

So nan\_to\_num is used in which Nan is replaced by zero and infinte values are replaced by largest finite value.

2.FEATURE SELECTION

Now for creation of Machine learning model, attributes plays a major role for training the classifier. Visualization of dataset attributes were performed for Feature Selection.

Visualizations were, Histograms ,Box Plots (the screenshots are attached in the source folder).

Recursive Feature Elimination (RFE) was also used for the feature selection.

Finally , 10 Features were selected those are - X\_ 2, X\_3 , X\_4 , X\_5 , X\_6 , X\_7 , X\_9 , X\_11 , X\_13 , X\_14 .

(The Date and Incident\_ID features are not needed as those are not relevant for incident being hack or not.)

3.DATA PRE-PROCESSING

Standardization and Normalization were studied for the model creation.

But for this dataset, the values were not exponentially high.And the models were not having any improvement by Standardization and Normalization.

4.TRAINING DATASET

The dataset was then splitted tor training and testing purpose.

This was done just to understand Variouns Machine learning Models performance.

The Ration was 80:20 and again 70:30 was also checked for the performance of models.

But finally used 80:20 to avoid any Underfitting of Model.

(Note : Later at final stage the Complete dataset was given for training purpose and Test was performed on the Test.csv provided).

5.MACHINE LEARNING ALGORITHMS

Various Machine learning algorithms were studied for the dataset.

SVM, KNN, RandomForest, Logistic Regression, Linear Regression, MLP and Guassian Naive Bayes.

Confusion Matrics, Accuracy,Precision,Recall,Classification reports,F1 score and AUC curves were studied too.

Random Forest and KNN showed good results as compared to other ones.

Random Forest had a Recall score of 0.57 and Accuracy of 98 % for the Training Dataset whereas KNN should Accuracy of 96% but recall score was not that impressive(0.31).

Various Standardization and Normalization was carried out on data in order to increase the potential performance of Machine learning algorithms.But that was just not enough.

So decided to used Ensemble Learning.

6.ENSEMBLE LEARNING (Random Forest + KNN )

Ensemble learning was carried out , in which earlier Guassian Naive Bayes was included too but the results were affected giving less recall score. So the final Ensemble Model consisted of Random Forest and KNN machine learning algorithms.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*ENSEMBLED MODEL SHOWED RECALL SCORE OF 0.87\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

7.The Model was prepared and saved in local directory using Pickle.

Then the model was used to predict the target attribute (MULTIPLE\_OFFENSE) of Test.csv file provided.

And the predictions were saved in the predictions.csv file.