Novartis – Hacked or Not

Introduction

With the evolution of technology, we are under a digital transformation. Digital payments have become a much more convenient way of transacting money. Though the technological revolution has been beneficial to the mankind, it also has some negative impacts. The risk of being hacked is increasing and the security of our information is a major concern. To provide us with a secure environment, we need the help of cyber security. Prevention is better than cure, Data Science helps us to predict an attack before its occurrence by efficiently handling the large amount of data available about the online payment portals and users.

Data

The data has been collected from Kaggle. It contains information about the hacking history of the company Novartis. The attributes are INCIDENT_ID, DATE and other variables X_1, X_2, X_3,......X_15. These variables are anonymized logging parameters such as details about the protocol, service provider, bytes of data shares and so on. The value of all of the attributes are numeric. The target variable is MULTIPLE OFFESNSE which is binary and takes the two values 0 (Not hacked) and 1 (Hacked).

We are given two datasets, the train data and the test data. Train data is the historical hacking record and test data is the data on which we need to apply our prediction algorithm.

Strategy

Firstly I have cleaned the data by removing the unnecessary columns such as INCIDENT_ID and DATE which are not responsible in predicting the result. I have also handled the missing data for the two datasets. I have then done Exploratory Data Analysis to analyse the data. I have tried to visualise the data to understand the significant parts. I have also applied different Machine Learning algorithms such as KNN, Naive Bayes, Logistic Regression and Random Forest to predict the output. The best fit model has been chosen on the basis of the highest accuracy score.

Once I had obtained the predicted values, I have encrypted the result using RSA Encryption scheme to increase its security.

Results

Random Forest is the model with the highest accuracy score as well as cross val score. This model was chosen to predict the results though other models were also executed on the data. The results have an accuracy of approximately 99% which shows that my prediction is almost accurate.