## Project: Bitcoin Heist Ransomware Analysis

Name: Minal Dhansukhlal Student ID: 300318615

## 1. Introduction and discovery

Bitcoin is the best-known cryptocurrency introduced by Satoshi Nakamoto. Bitcoins is held on a public platform, and it is not owned by any government or company. On a blockchain, two unacquainted parties can create an immutable transaction that is permanently recorded on the ledger to be seen by the public.

Ransomware is a type of malware that infects a victim’s data and resources and demands ransom to release them. In two main types, ransomware can lock access to resources or encrypt their content. In this analysis there are 28 Ransomware families as the labels and White indicating that the transaction is not under the category of any ransomware family. The dataset has 2916697 transactions.

**Framing the problem**

The purpose of this analysis is to identify when the transaction is for a ransomware or it is a genuine transaction, so that we can prevent money heisted using ransomwares. We can understand how the ransomware families perform the heists by locking access or other methods thus enhancing better security for Bitcoin transactions. We can also identify other families of ransomware using the analysis to find similarities and suspicious transactions.

**Hypothesis 1**

**H0:** All features are used in predicting the labels of the transactions

**H1:** Not all features are used in predicting the labels of the transactions

**Hypothesis 2**

**H0:** Bitcoin Address is important feature in predicting the label

**H1:** Bitcoin Address is not an important feature in predicting the label

**Hypothesis 3**

**H0:** All White labeled transactions are not ransomware transactions

**H1:** Not all White labeled transactions are not ransomware transactions

## Data Preparation

**Data inventory**

The data was obtained from UCI website (link: [http://archive.ics.uci.edu/ml/datasets/BitcoinHeistRansomwareAddressDataset#](http://archive.ics.uci.edu/ml/datasets/BitcoinHeistRansomwareAddressDataset)).

I have grouped the records into four categories namely: Normal, Padua, Montreal, and Princeton. Further used label encoder to get integer labels to represent each family. Under each family there are categories.

I did not eliminate categories based on the value\_counts as I had shown in the proposal because I was eliminating many categories in the process. Graphical user interface, table

Description automatically generated

Label and Date were dropped after this stage.

## Model Implementation

In the proposal, my initial plan was to perform clustering after that perform classification on the clusters but since I had big dataset to cluster it was taking very long to analyze. Therefore, I decided to group the records into Ransomware families such as Montreal, Padua, Princeton and Normal (No Ransomware, white).

I performed two-stage classification by classifying the record twice:

**First Stage**: I performed classification using the 4 Ransomware family groups I had labeled. On the first stage the target is Ransomware\_families which will predict which record falls under which group. On this stage I used Pipeline to implement many classification algorithms with Scaling. I performed that with Address and without Address. The results showed that algorithm models performed better when Address was included. This shows that there is some significance of Address in the classification. Table

Description automatically generatedTable

Description automatically generated

Best performing model was XGBClassifier with 91% accuracy. Thereafter I used GridSearchCV to obtain best parameters for XGBClassifier. Text

Description automatically generated

**Second Stage:** Using if\_else condition, I obtained records of Montreal, Padua, and Princeton as separate data frames. The data frames were used to create models for each group. I used XGBClassifier with same parameters obtained in the earlier stage to model this groups with MinMax Scaler.

I performed classification without grouping the data, therefore 29 Categories as classes. The accuracy obtained was high, but it might be due to training with large dataset. As I counter checked the results of grouped categories and without grouped categories, I found that without grouped categories has higher chances of misclassifying.

## Results Interpretation and Implications

The accuracy obtained is good and giving good predictions. Hypotheses 1 and 2 are proved that all features are important, and address is making significant impact. Third hypothesis can be said to Null hypothesis is not rejected since during testing of White Labeled records show some white labeled records are predicted to be ransomware. The data is very big but is not well balanced.

Ransomware Families

Chart

Description automatically generatedTable

Description automatically generated

Grouped Categories based on Ransomware Families

Scatter chart

Description automatically generated Table

Description automatically generated

Chart

Description automatically generated Table

Description automatically generated

Chart, treemap chart

Description automatically generated Table

Description automatically generated

Without Grouped Categories

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

## Out of Sample Prediction

I created excel file with 5 records, I gave random values by mixing the values among itself. First, I used Ransomware family’s model which will label the record whether it falls into Normal, Montreal, Padua, or Princeton. Based on this classification I used if else and for loop to classify into categories of each group.

Table

Description automatically generated

Text

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

## Concluding Remarks

From the analysis, I have observed that chances of misclassification of the records are reduced if they are grouped, I was not able to implement clustering due to having large volume of records which was bringing difficulty to analyze.

Instead of grouping using the clusters, I grouped the records using the general Ransomware Family therefore having 4 groups. Because of having huge dataset, the accuracy with all categories is higher because its training from 50000 records while the groups might not show high accuracy, but it is training from lesser records due to grouping