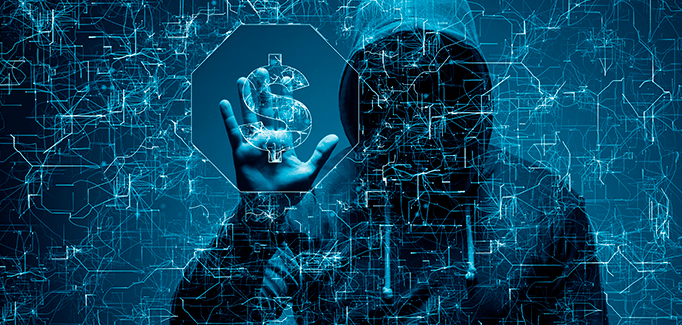
DAB 402 – CAPSTONE PROJECT

ASSESSEMENT 2 – Detailed data Assessment

Cyber Space Security in Banking (CSS)

Credit Card Fraud Detection



GROUP 13

|  |  |
| --- | --- |
| **NAMES** | **STUDENT ID:** |
| Jai Surya | 0731608 |
| Rajit Kumar | 0730468 |
| Amita Mehta | 0730478 |

***Date – 02/14/2020***

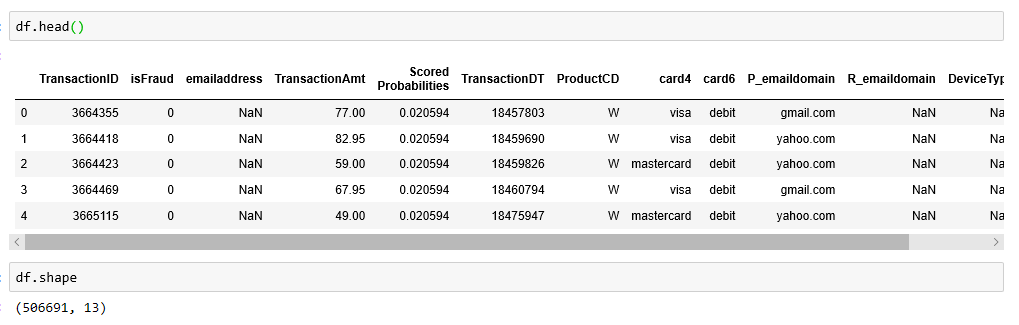
**Topic:** Credit Card Fraud Detection

**Problem Statement:**

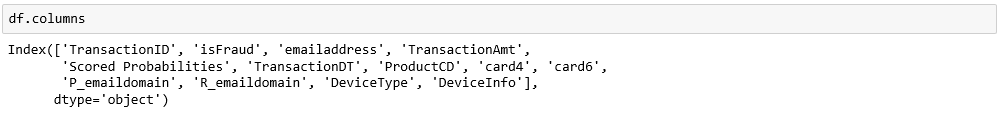
Electronic computing and communication pose some of the most complex challenges engineering has ever faced. They range from protecting the confidentiality and integrity of transmitted information and deterring identity theft to preventing the scenario recently dramatized in the Bruce Willis movie *"Live Free or Die Hard,"* in which hackers take down the transportation system, then communications, and finally the power grid. The most complex challenge engineering has ever faced is electronic communication and computing. **Cyber Crime** is the most serious problem in the present era. According to U.S.A. government more than 600 billion amount of fraud happened annually and it is increasing by 1.2 billion. And more than 200 billion amounts of fraud happened in the banking sector. South Africa has recently been afflicted by fraud in credit and banking information from online banking subscribers.

**Dataset:**

* We got the dataset name hiwott-cyber-security-dataset. This dataset contains 50,6691 rows and 13 columns. We have more data for test and train the model to predicting the accuracy.



* The dataset contains the following columns



**Data Quality:**

Our dataset contains 506692 rows in total and 13 columns i.e. good because it helps us to predict the good model. As all column or data is related to solve our problem. This dataset contains null values and the percentage of these value is high. So, we remove these. Apart of this we must do some cleaning with excel tools.

The motivation behind this project is preventing the people from credit card and online fraud. This dataset provides the clear face of the fraud to the banking industry. This dataset helps Electronic Engineering and cybersecurity researches to aware from this kind of fraud.

We will use the card type column, which help to calculate the number of fraud happened in different card types.

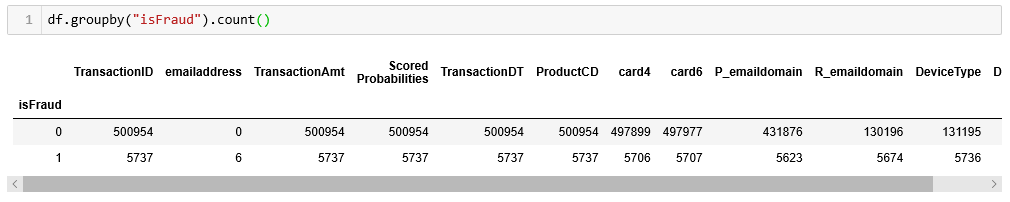
**How the dataset fit the problem:**

This dataset contains unnecessary columns and null values. So, we must do some cleaning and use pre-processing techniques for further analytics. And pre-procession is used to resolve the analytic problem. So, our topic relates to the analytical problem.

Our project is regarding prevention of fraud, so the basic information that we must optimize the fraud. Therefore, we need a column that can identify how many frauds is happened in our dataset. We use the account no. and the transaction column that will help us to find the fraud in the dataset. From some cases we also realize that there are some fraud happened due to the use of unsecure WIFI and networks.

**Reliability:**

The dataset become more reliable after balancing the data. Our data is imbalance because the fraud cases are more.

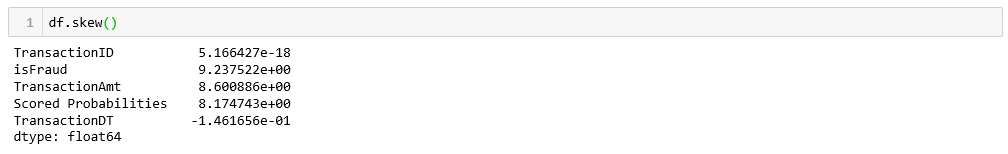




**Null Values:**



**Skewness:**



**Observations:**

The features are positively skewed i.e. distribution has tail on the right side for the positive ones. Some features are highly skewed to right. This can be addressed by transforming them using log transformation.

**Data Processing:**

Our data contains 13 features that are input for the model and being analyzed. We will remove the less important features from dataset. We can’t directly input the data to the model as features are of different datatypes. The dataset contains the numeric and non-numeric values.

**Feature Selection:**

It is necessary to select the important feature from dataset that can contribute to high performance detection model. We use the PCA and Random forest helps to reduce the dimensionality.

**Five C’s:**

**Consent:**

As this dataset is opensource and available in data. World. So, there is no need to have consent to collect the dataset.

**Consistency:**

The number of records in the dataset is reasonable and affordable to perform the experiments on this dataset without selecting random samples. As a result, the predictions or results of different models will be comparable and consistent.

**Clarity:**

It is clear on how we use the dataset. The data is used to identify Credit card fraud in cyber space security in banking.

**Control:**

 Since our model helps only in detecting the intrusions but don’t block them. In real time all the data is monitored and controlled by data.world.

**Consequences:**

Fraud is the significant issue in our model. In some cases, fraud is less in the dataset. In this situation data is imbalanced must spend more time on deciding them otherwise we balanced the data.

**Reference:**

* Pdfs.semanticscholar.org

<https://pdfs.semanticscholar.org/f920/211b1e8a3d2c3e3ac89caea3ba3caba98fd9.pdf>

* Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier

Credit Card Fraud Detection

[https://www.kaggle.com/mlg-ulb/creditcardfraud](javascript:openWebLink('https://www.kaggle.com/mlg-ulb/creditcardfraud'))

* Detecting credit card fraud by genetic algorithm and scatter search

#### Detecting credit card fraud by decision trees and support vector machines

* [https://openaccess.dogus.edu.tr/xmlui/handle/11376/2366#sthash.sDqj3beI.dpbs](javascript:openWebLink('https://openaccess.dogus.edu.tr/xmlui/handle/11376/2366#sthash.sDqj3beI.dpbs'))
* Dataset got from data.world
* <https://data.world/hiwott/cyber-security-dataset>