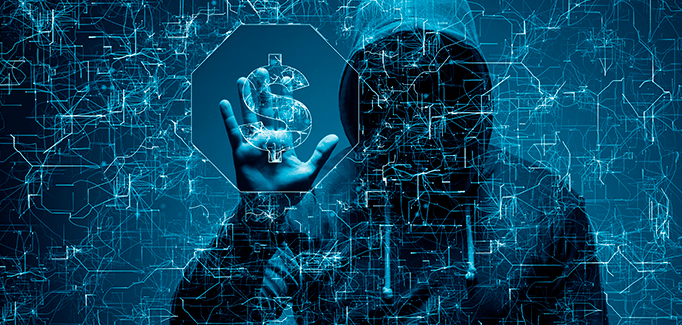
DAB 402 – CAPSTONE PROJECT

Final Project Report

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

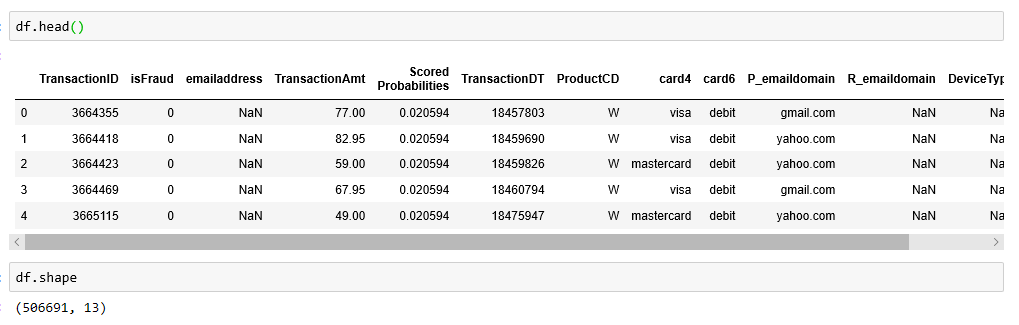
**Problem Statement:**

**Introduction:** Electronic figuring and correspondence represent probably the most unpredictable difficulties building has ever confronted. They run from securing the secrecy and honesty of transmitted data and dissuading fraud to forestalling the situation as of late performed in the Bruce Willis film "Live Free or Die Hard," in which programmers bring down the transportation framework, at that point correspondences, lastly the force matrix. The most intricate test building has ever confronted is electronic correspondence and figuring. Digital Crime is the most difficult issue in the current time. As indicated by U.S.A. government in excess of 600 billion measure of misrepresentation happened yearly and it is expanding by 1.2 billion. What's more, in excess of 200 billion measures of misrepresentation occurred in the financial segment. South Africa has as of late been beset by misrepresentation in credit and banking data from internet banking endorsers. Mastercard extortion is the point at which somebody utilizes your Visa or credit record to make a buy you didn't approve.  Fraudsters can likewise take your Mastercard account number, PIN and security code to make unapproved exchanges, without requiring your physical charge card. Fraud is one of the major ethical issues in the credit card industry. The main aims are, firstly, to identify the different types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. The sub-aim is to present, compare and analyze recently published findings in credit card fraud detection.

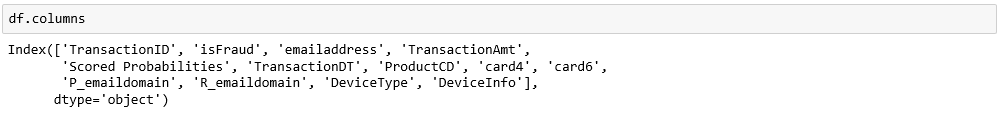
 In 2018, unauthorised financial fraud losses across payment cards and remote banking totalled £844.8 million in the United Kingdom. Whereas banks and card companies prevented £1.66 billion in unauthorised fraud in 2018. That is the equivalent to £2 in every £3 of attempted fraud being stopped

**Dataset:**

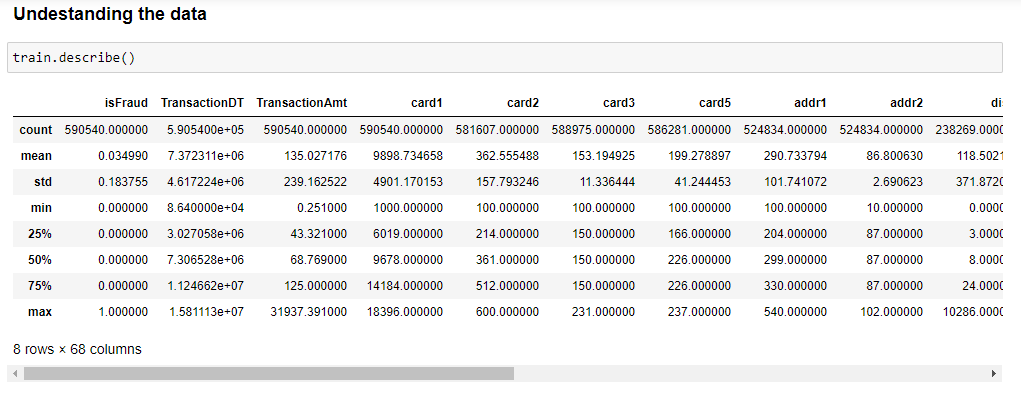
* We got the dataset name hiwott-cyber-security-dataset. This dataset contains 50,6691 rows and 13 columns. We have more data for test and train the model to predicting the accuracy.



* The dataset contains the following columns



* Train Datasets:



**Analytical Problem:**

* To choose this topic we must learn more about a problem before solving it. we will have to first collect data or information before analyzing it. Therefore, an important analytical skill is being able to collect data and research on problem. To select a problem, we must do data collection, investigation and prioritization on the topic.
* After selecting a problem, we must use critical thinking and communication skills to work on this topic within group. We need these tools to enhancing group productivity, reporting, surveying, teamwork, presentation, written communication, collaboration, problem-solving.
* According to above outputs the dataset is data driven problem and data-driven is a part of the analytical problem. To resolve this problem, we must use the analytical thinking. Analytical thinking helps to observe and identify this problem to develop more ideas about it. Gathering Information and developing solutions about the problem is also a part of the analytical problem. Analytical thinking is the ability to quickly identify cause and effect of this problem. Identify and examine the cause and effects while solving this problem. To gather the knowledge, we must do more research and information analysis.
* 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. This pre-processing and the cleaning of data will be done with the help of Excel tools. Before predicting the accuracy or using the accurate model we must understand the data.
* This dataset contains the huge number of rows. So, it is hard to analyses this data. And to create the co-relation between the feature we must use some analytics tools. To understand the data, we can use the Tableau and Excel. As tableau helps to visualize the things in proper way and good manner. By plotting the graph, we can create and find the correlation between the features.
* Now we step into the python tools. After the visualization and pre-processing we must do the feature selection. There are huge number of columns in each dataset so its little difficult to select the feature. We can use feature selection based on predicting our model.It is the process of reducing the number of input variables when developing a predictive model. And it helps us to improve performance of the model. In our project the main features are isFraud, Transaction amount and Transaction Time. And use different modeling techniques and consider the better one for further work.
* After feature selection we partition the data into test and train form for building different models. And select the best model which predict the best accuracy. We may use the regression and linear model. We used neural network based on the feature. And create a column which tells whether the fraud is happened or not while the transaction

**Data Overview:**

Our dataset contains 590540 rows in total and 82 columns i.e. good because it helps us to predict the good model. Most of the columns in 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 python 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 to calculate the number of frauds happened in different card types. Following graph shows the percentage of fraud vs card type.

Visa card having the maximum number of frauds while discover card type having minimum card type. Master card has the 38.07% of fraud among others.

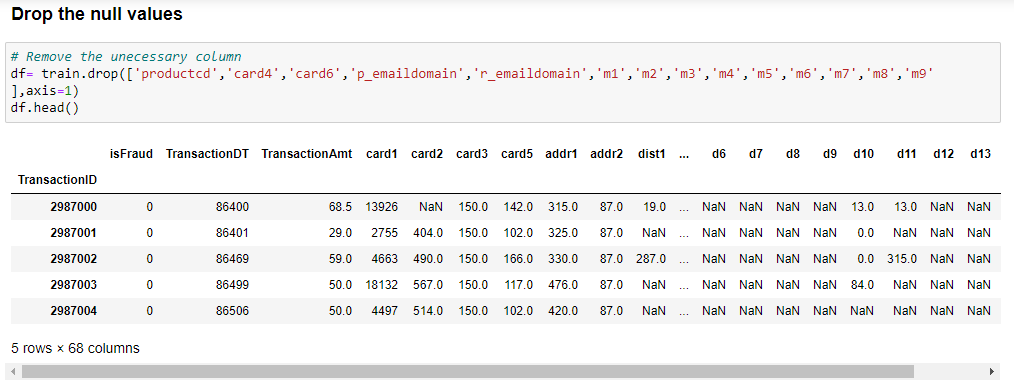
**How the dataset fit the problem:**

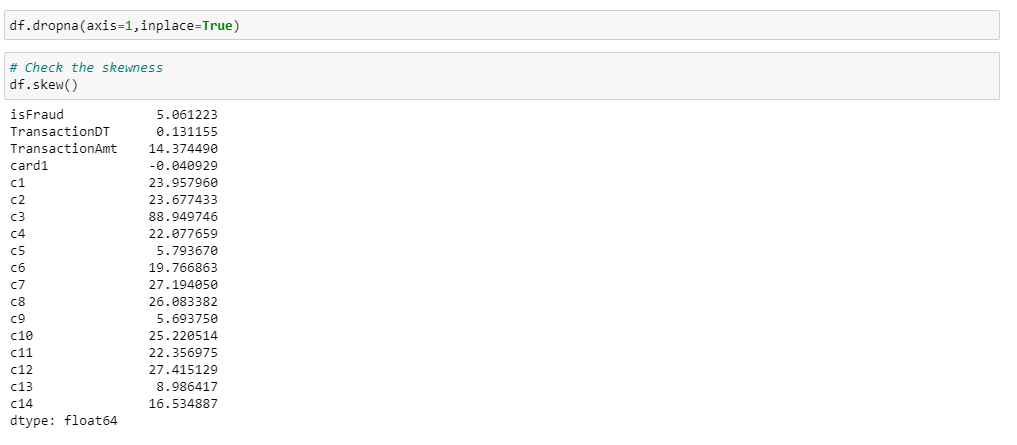
This dataset contains unnecessary columns and null values. So, we must do some cleaning and use pre-processing techniques for further analytics. And pre-procession is used to resolve the analytic problem. So, our topic relates to the analytical problem.

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.

**Cleaning and Pre-Processing:**

* **Remove the null values and unnecessary columns**





Above code is used to remove the unnecessary columns. As, these columns having the string data and we don’t want to use these. The 2nd code helps to remove the null values from the data and the 3rd code is used to check the skewness of the data.

* **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.

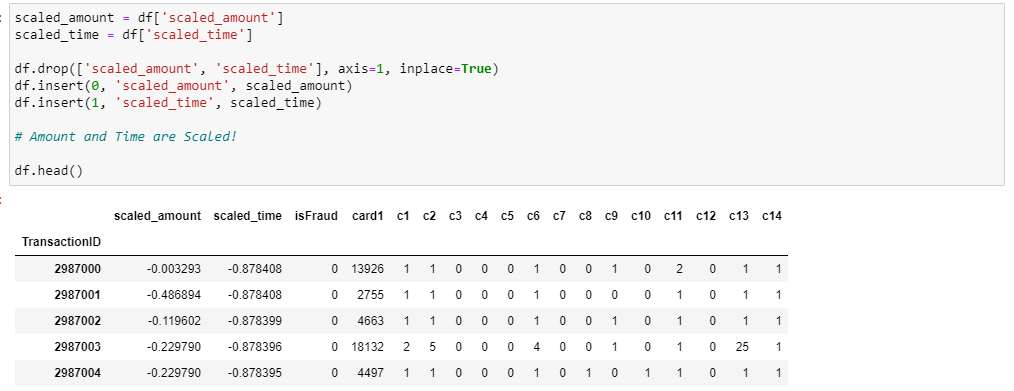
* **Check whether the data is balanced or not.**



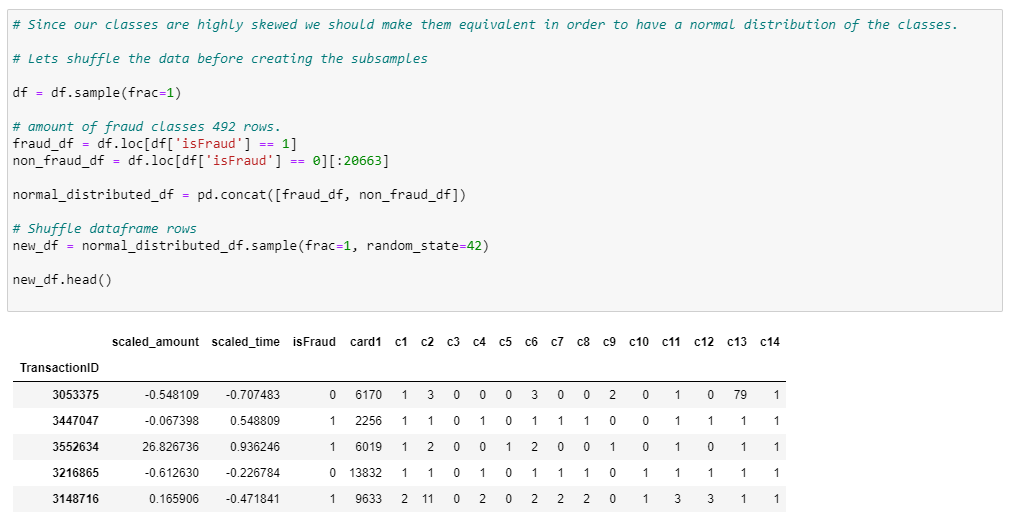
The data is not balanced. So, we used the standard scalar techniques to balance the data. As we saw in the previous graph it shows that data is highly imbalanced. 96.5% of the data is non fraud and only 3.5% of the data is fraud. The original data is having the more no fraud cases than the fraud cases. If we use this data, its shows errors and it overfit the algorithm. Also, it predicts the cases are no fraud, but we don’t want our model to predict that. Therefore, we must scale the data for fraud prediction.

**Scaling:**

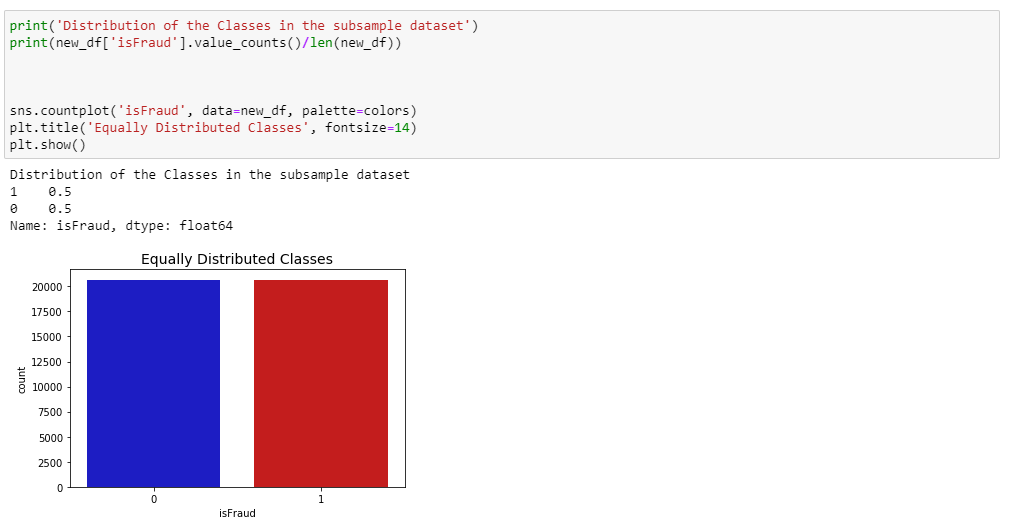
We can scale the data with the help of transaction time and transaction amount columns. We also create a sub sample of the data frame in order to equally distribute the no fraud and fraud cases.



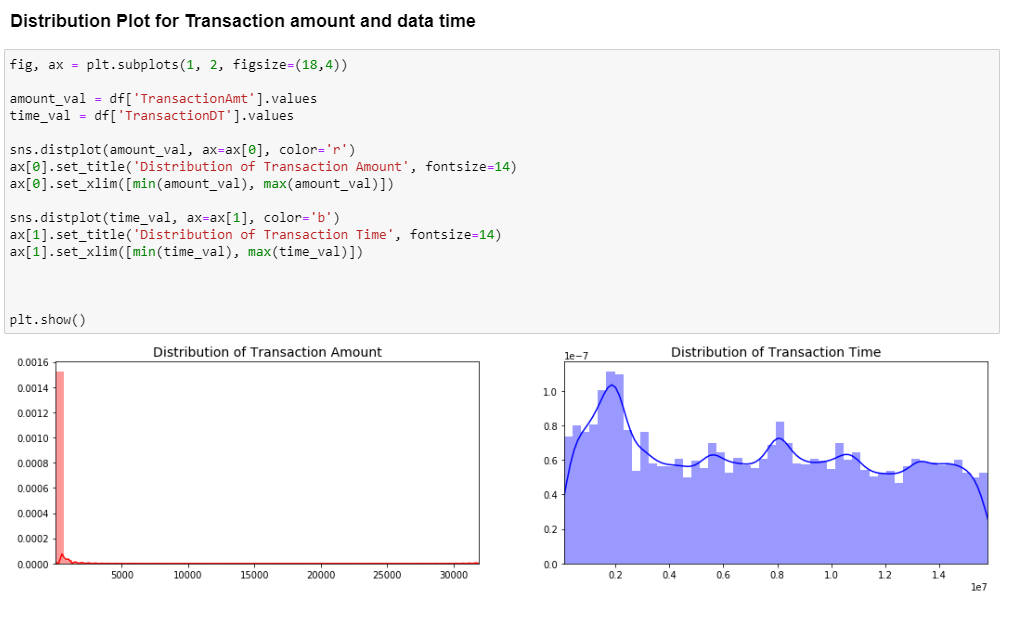
Our dataset contains 20663 fraud cases so we can get randomly 20663 of no fraud cases to create a new sub- sample. We concat the 20663 fraud cases to scale the data and the code is given below:



we will shuffle the data to check whether our models can maintain a certain accuracy when we run it. After Scaling the distribution of fraud cases is in 50-50 ratio and it shown by the following picture:



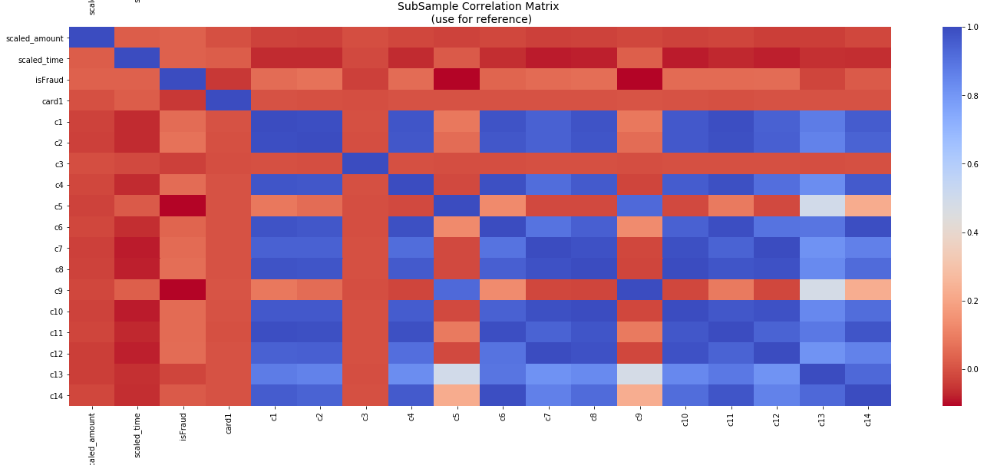
**Plotting:**



The above distribution shows the skewness of features in the data. There are techniques that can help us to make the distribution less skew.

**Correlation Matrix:** We use the correlation matrix to understand the data. It helps us to check which feature used to predict whether a transaction is fraud or not. If negative correlation has the lower values, then the more likely the result will be a fraud case. Otherwise, if the positive correlation has the higher value, then the more likely the result will be a fraud case.





**Literature Reviews:**

**Summary 1: The Effects of Cyber Threats on Customer’s Behavior in e-Banking Services**

The current project has clearly described the causes of cyber threats and the measures that must be taken to prevent cybercrime. In this project we have seen that the cybercrime is the major problem in the financial institutions in 21st century. The cybercrimes occur mainly for Identity theft, phishing, vishing, malware, hacking and cracking, social engineering, automating online banking fraud etc. The common security measure that must be taken for preventing from cybercrimes are: Securing the device using for online banking, protecting personal data, use strong password, upgrade system and software.

According to a survey it is proved that 70% customers are not aware or got limited awareness about cyber threats. Therefore, the online banking users need to keep extra care over their usage towards banking services, E-banking customers should be taken care and should educate them more about the cyber threats and secure process for online banking environment.

**Reference:** Ijeeee.org. 2020. [online]

Available at: [<http://www.ijeeee.org/vol7/414-IM023.pdf](%3chttp:/www.ijeeee.org/vol7/414-IM023.pdf)> [Accessed 31 March 2020].

**Summary 2: Cybercrime in Banking Industry and Its Impacts on Banking Industry**

This project described the outline of cybercrime in banking industry and its effects on banking industry globally. The major cybercrimes occur at credit card frauds and vishing. Banks should take adequate measures to educate customers through their websites about this banking fraud and secure banking options, the banks should work cooperatively with other banks to avoid cybercrimes. Banks should take strong measures for eradicating cyber fraud completely by hiring strong IT background employees with more knowledge in technology and cybercrime prevention tactics.

Following are the factors of cybercrime in banking:

* Fraudulent with use of Automated Teller Machine (ATM) cards and accounts
* Credit card frauds
* Frauds involving electronic funds transfers
* On Call Frauds
* Frauds relating to data Interchange.
* Email spam is the serious issue.
* Non- payment and non-delivery spam
* Investment scams
* Multi factors of password

**Reference:** [[file:///C:/Users/AMITA%20MEHTA/Downloads/cybercrime-in-banking-industry-and-its-impacts-on-banking-industry.pdf](file:///C:\Users\AMITA%20MEHTA\Downloads\cybercrime-in-banking-industry-and-its-impacts-on-banking-industry.pdf)](file:///C:\Users\AMITA%20MEHTA\Downloads\cybercrime-in-banking-industry-and-its-impacts-on-banking-industry.pdf)

**Summary 3: Anonymized credit card transactions labeled as fraudulent**

This project contains the data of credit card transaction made in September 2013 by European cardholders. The data contains 20 columns and 284,807rows. First column having data of number of seconds elapsed between this transaction and the first transaction in the dataset and rest of the columns made by the PCA use. The data is imbalance as the dataset contain the data of transaction happened in two days, where we found 492 frauds out of 284,807 transactions. It contains numerical data which are the result of pca transformation except amount and time. As they don’t want to disclose the personal information.

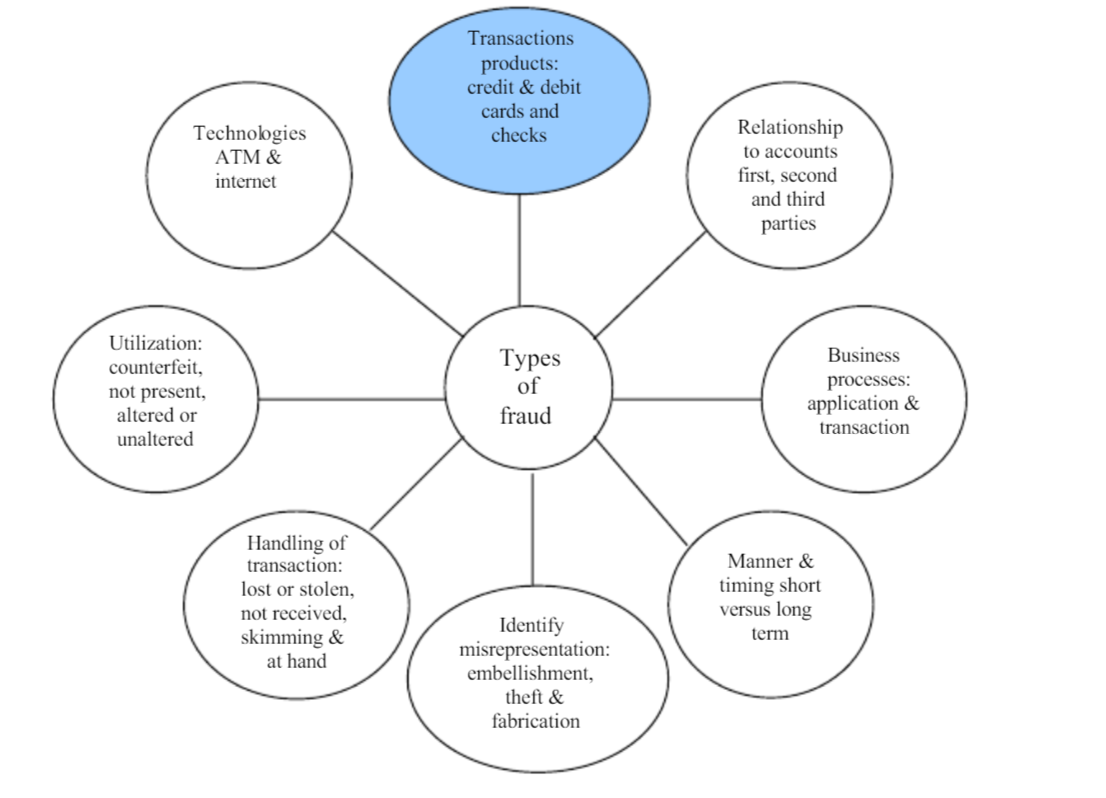
First, they balance the dataset and use the column created by the pca transformation for the fraud prediction. For class imbalance ratio, they use the accuracy using the Area Under the Precision-Recall Curve. They use the logistic regression, KNearest, supply vector and decision tree classifier for the fraud prediction.

**Reference:** Kaggle.com. (2020). *Credit Card Fraud Detection*. [online]

Available at: <https://www.kaggle.com/mlg-ulb/creditcardfraud> [Accessed 23 Feb. 2020].

**Summary 4:** [**Abstract data set for Credit card fraud detection**](https://www.kaggle.com/shubhamjoshi2130of/abstract-data-set-for-credit-card-fraud-detection)

In this project dataset contains 12 columns and 3076 rows. They iterate through all the columns and modify the data type to reduce the memory usage. They split the data into test and train for the model selection. They use counter vector classifier, logistic regression, gradient boosting classifier and MultinominalNB for the fraud prediction. But the accuracy for all the model are very less to find the best model they use the confusion matrix for the classification report. The main aim of this project, firstly to identify the different types of fraud credit card fraud and secondly, to review alternative techniques that have been used to identify the fraud. Following chart shows the different types fraud detection in credit card:



**Reference:** Kaggle.com. (2020). *Credit Card Fraud Detection*. [online]

Available at: <https://www.kaggle.com/aherparesh/credit-card-fraud-detection> [Accessed 23 Feb. 2020].

# **Summary 5: A Review of Data Mining-Based Financial Fraud Detection Research**

This project is regarding the rapid development of smart phones and the plethora of smart phone applications becoming developed, the lines between the cellular network and the internet are being blurred, rendering the phone a not-so-trusted device. In this article they show how existing smart phone banking applications can be tampered to capture user information and password. In this article they use the column which contain the data of the smart phone i.e. with which network the device was connected. And fraud happened. They use clustering for pre-processing and then use the logistic regression, decision tree, knn models for the prediction.

**Reference:** Ieeexplore.ieee.org. (2020). *Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk - IEEE Conference Publication*. [online]

Available at: [https://ieeexplore.ieee.org/abstract/document/6784638](https://ieeexplore.ieee.org/abstract/document/6784638%20) [Accessed 24 Feb. 2020].

**Summary 6: Automating Online Banking Fraud**

This project has described about the automates transfer system (ATS’s), which is highly used by cybercriminals in conjunction with Spy Eye Zues malware variants as a part of web inject files. We have also seen why some countries are targeting as compare to others. This project predicted that ATS could be a better source of the income and cybercriminals Will continue to improve ATS. ATS performs fraudulent transactions in the background and so it is hard to determine ATS infection. The cybercriminal underground is the place to find people coding Web Inject files and ATSs.

A Web Inject file is basically a text file with a lot of JavaScript and HTML code. This file allows cybercriminals to target specific organizations (e.g., banks) and inject specific code into victims’ browsers so they can modify the web pages the users access in real time. Web Inject file users can easily make fake pop-ups that ask victims for specific credentials (e.g., social security numbers and mothers’ maiden names) appear. Web Inject files have all of the code required to fool victims into thinking the pop-ups they see are real.

In this project, it clearly described that It is suggestible to check the bank account via checking balances over the phones or monitoring bank statements sent through emails instead of checking them in online.

At the end financially institution will get benefited from analyzing ATS attack method to identify whether they must modify or supplement the current security control.

**Reference:** (2020). Retrieved 24 February 2020,

Available at: [from https://www.trendmicro.co.uk/media/misc/automating-banking-fraud-via-ats-research-paper-en.pdf](file:///C:\Users\jaisu\Downloads\Desktop\Project\from%20https:\www.trendmicro.co.uk\media\misc\automating-banking-fraud-via-ats-research-paper-en.pdf)

**Summary 7:** **BANKRUPTCY PREDICTION SYSTEM FOR CREDIT CARD USING MACHINE LEARNING TECHNIQUES: A SURVEY**

In this article we studied that they are classifying the transaction, they are predicting that a transaction is high risk transaction or low risk transaction. They have dataset with 100000 rows and 20 columns, and they are using naive Bayes classifier.

They first collected the data then did some pre-processing after that they selected the features of their interest using feature selection. Then they did classification using naive Bayes classifier after that all they used confusion matrix to check the performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **No. of Features** | **Total Instances** | **No. of Instances (Yes)** | **No. of Instances (No)** |
| USCD-FICO | 20 | 10,0000 | 97346 (97.35%) | 2654 (2.65%) |
|  |  |  |  |  |

**Reference:** Irjeas.org. (2020). [online]

Available at[: http://www.irjeas.org/wp-content/uploads/admin/volume7/V7I1/IRJEAS04V7I101190319000002.pdf](https://stclairconnect-my.sharepoint.com/personal/w0730468_myscc_ca/Documents/:%20http:/www.irjeas.org/wp-content/uploads/admin/volume7/V7I1/IRJEAS04V7I101190319000002.pdf)