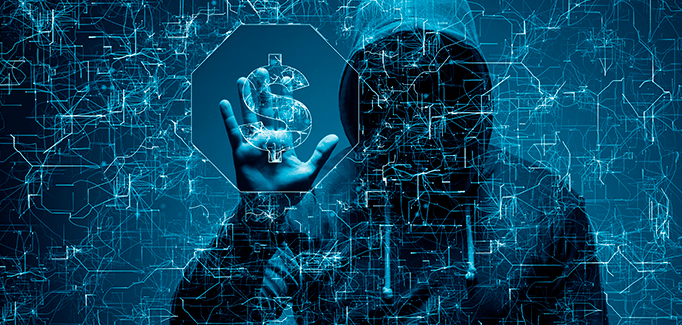
**Project of Capstone**



**St. Clair College, Windsor, ON**

Submitted To

**Profs. Muhammad Shahid**

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

**Date – 04/17/2020**

Final Project Report

DAB 402

CAPSTONE PROJECT

Content:

* **Topic (Problem Statement)**
* **Analytical Problem**
* **Literature Review**
* **Ethical Concerns**
* **Dataset**
* **Methodology**
* Data Overview
* How the dataset fit the problem
* Cleaning and Preprocessing
* Balancing
* Scaling
* Modeling
* **Conclusion**
* **Goals**
* **Challenges**
* **Results**
* **Contributions**
* **References**

Credit Card Fraud Detection

|  |
| --- |
|  |

**Problem Statement:**

**Introduction:** The most complex challenge engineering has ever faced is electronic communication and computing. As the banking industry has growing rapidly the last few years. Banking and electronic industries move part of their business, or the entire business, towards online services providing e-commerce, information and communication service for the purpose of allowing their customer better efficiency and accessibility. This evolution is a big step forward the efficiency, accessibility and profitability point of view but every good thing is come with some drawbacks. The problem behind that is making transaction through the internet lies in the fact that neither the card nor the cardholder needs to be present at the point-of-sale. Therefore, it is impossible for the merchant to check whether the customer is genuine or not. Credit card fraud detection is the most serious problem throughout the internet. Companies and industries lose huge amount annually due credit card frauds. 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, unauthorized financial fraud losses across payment cards and remote banking totaled £844.8 million in the United Kingdom. Whereas banks and card companies prevented £1.66 billion in unauthorized fraud in 2018. That is the equivalent to £2 in every £3 of attempted fraud being stopped

The main thing that we analyze through the article is that the main reason of fraud in banking is the online transactions and crypto jacking. Crypto jacking helps the hackers to freeze the system and take all the data of the banks through the networks.

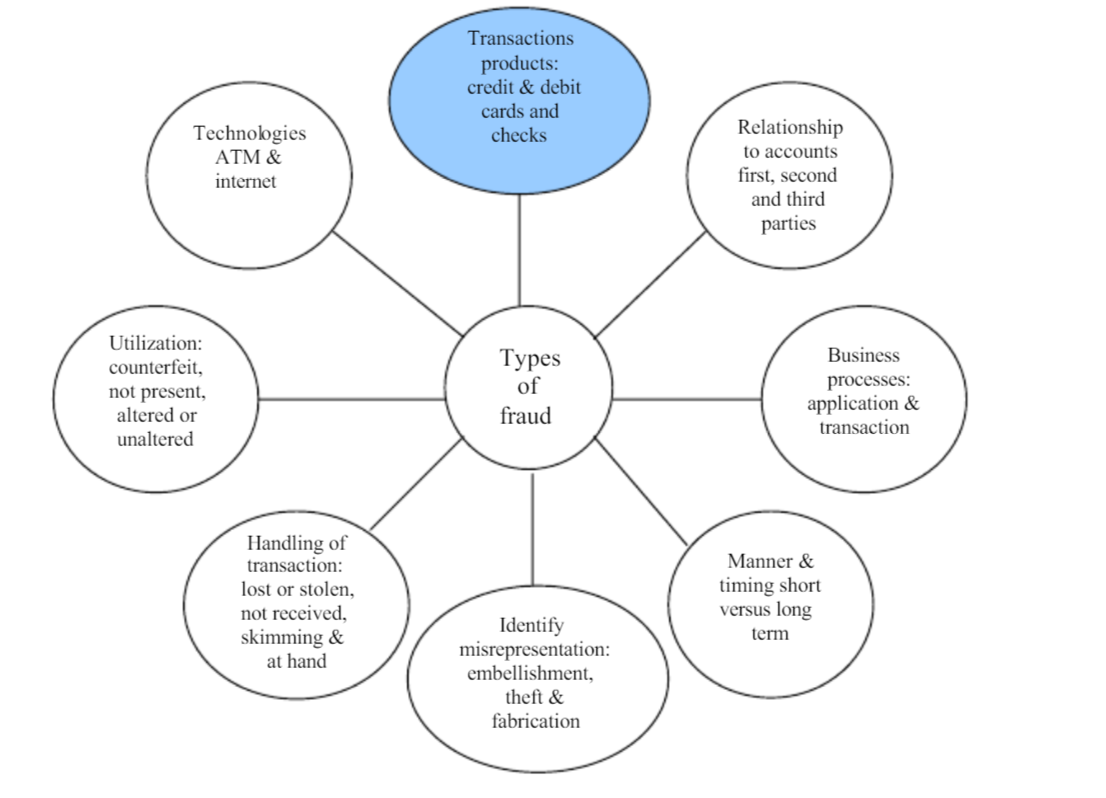
**Analytical Problem:**

* 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

**Literature Reviews:**

**Summary 1:** [**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 2: Critical Success Factor for Preventing e- Banking Fraud.**

Security issues are significant obstructions to online banking and online business activities among shoppers (Khasawneh, 2009) with misrepresentation featured as a significant hazard related with payment system (Roberds, 1998). To make sure about an e-banking framework, IBM set accentuation on characterizing clear targets. This is accomplished by understanding the business objectives, targets and basic achievement factors when arranging the security procedure, just as the effect on the business on the off chance that they are not accomplished (International Business Machines (IBM), 2001). There has been insignificant research identified with associations experience on fraud detection and the basic achievement factors for e-banking extortion anticipation measures. Henceforth the elements that have been distinguished require further examination to comprehend their criticality.

Past innovation, other viable approaches to control security dangers should be directed. This can be accomplished by having versatile strategies, methodology and controls (Titrade, 2000). The issue of correspondence was found to assume a significant job in ebanking security notwithstanding hierarchical adaptability, accessibility of assets, ebanking venture arrangement, support from top administration, data straightforwardness and security information and mindfulness (Koskosas, 2011). This commendation results from crafted by Akindele (2011) where it was discovered that absence of satisfactory preparing, insufficient correspondence, and powerless authority styles of administrators and directors as all reasons for extortion. In the UK, web-based banking has seen up to a 32% decrease in misrepresentation and this has been ascribed to expanded client mindfulness and extortion discovery programming in banks (UK Fraud Action, 2010). In this manner showing that past the innovative viewpoints, there is critical effect from client mindfulness and presentation to extortion prudent steps.

**Reference:** Santa Anita Park. 2020. Handicapping - Santa Anita Park. [online]

Available at: [<https://www.santaanita.com/horse-racing/handicapping/>](%3chttps:/www.santaanita.com/horse-racing/handicapping/%3e%20) [Accessed 31 March 2020].

**Summary 3: ELECTRONIC CRIME IN INDIAN BANKING**

We will probably ready to amount with however much exactness as could be expected the expense brought about by associations. at the point when they have digital assault. As far as we can tell, a conventional study approach would not catch the essential subtleties required to extrapolate cybercrime costs. Therefore, we decided to pursue field-based research that involved interviewing senior level personnel and collecting details about actual cyber crime incidents. In developing countries, like India, electronic crime is a serious problem because there is a lack of training on the subjects to investigate the electronic crime. The ATM fraud is not the sole problem of bank alone. It is a big threat and it requires a coordinated and cooperative action on the part of bank, customers and the law enforcement machinery.

The ATM frauds not only cause financial loss to banks, but they also undermine customers' confidence in the use of ATMs. The nature and extent of precautionary measures to be adopted will, however, depend upon the requirements of the respective banks. Credit card fraud can be devoted using a credit card or any similar payment mechanism as a fraudulent source of funds in a transaction. The reason may be to obtain goods without paying, or to obtain unauthorized funds from an account. The regulatory framework must also consider all the related issues like development of e-money, right to privacy of individual. International law and

international co-operation will go a long way in this regard. At last it can be concluded that to eliminate cybercrime from the cyber space is not a possible task, but it is possible to have a regular check on banking activities and transactions. The only promising step is to create awareness among people about their rights and duties and further making the application of Sai Om Journal of Commerce & Management

A Peer Reviewed National Journal VOL. 1, ISSUE 11 (November 2014) 14 Online ISSN 2347-7563 the laws progressively stringent to check wrongdoing. There is a need to acquire changes the Information Technology Act to make it increasingly successful to battle digital wrongdoing. As of late RBI has given rules recommending measures and revealing techniques for digital misrepresentation cases to be trailed by the banks.

**Reference:** CISO MAG | Cyber Security Magazine. 2020. Around 50,000 Cyber Frauds Reported in Indian Banks In 2018-19. [online]

Available at: [<https://www.cisomag.com/around-50000-cyber-frauds-reported-in-india-during-2018-19-rbi/>](%3chttps:/www.cisomag.com/around-50000-cyber-frauds-reported-in-india-during-2018-19-rbi/%3e%20) [Accessed 3 April 2020].

**Summary 4: 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 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,

[from https://www.trendmicro.co.uk/media/misc/automating-banking-fraud-via-ats-research-paper-en.pdf](from%20https:/www.trendmicro.co.uk/media/misc/automating-banking-fraud-via-ats-research-paper-en.pdf)

**Summary 5: 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, Random Forest and decision tree classifier for the fraud prediction.

The accuracy of the model is not accurate. According to the accuracy of the model we saw that the model is underfit. To fit the model, they used the confusion matrix.

The data is imbalanced as the non fraud cases were more than the fraud cases. So, it is hard to predict the fraud cases and it give us the error. It can’t predict what we want to know. So, they must scale the data with the help of balancing techniques.

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

Available at: <https://www.kaggle.com/mlg-ulb/creditcardfraud> [Accessed 23 Feb. 2020].7

# **Summary 6: 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 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)

**Summary 8: 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 9: 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 10: Credit Card Fraud Detection Using Meta-Learning: Issues 1 and Initial Results.**

Our analyses tried a few machine learning algorithms just as meta-learning systems on real world information. Not at all like many detailed investigations on data set collections, the set up and the assessment criteria of our tests right now to reflect then real-world setting and its resultant difficulties. The tests detailed here show: 50/50 appropriation of extortion/non misrepresentation preparing information will produce classifiers with the most noteworthy True Positive rate and low False Positive rate. Different scientists likewise announced comparative discoveries. Meta-learning with BAYES as a meta-student to consolidate base classifiers with the most elevated True Positive rates gained from 50/50 misrepresentation dispersion is the best strategy discovered up to this point.

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

Available at: [<https://www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-07-015.pdf>](%3chttps:/www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-07-015.pdf%3e%20) [Accessed 3 April 2020]

**Ethical Concerns:**  Any problem or a situation that need authorization from a person or organization that it is right or wrong.

**Consent:**

As this dataset is taken from 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. And it is clear we didn’t share any personal information of the person to the third party.

**Control:**

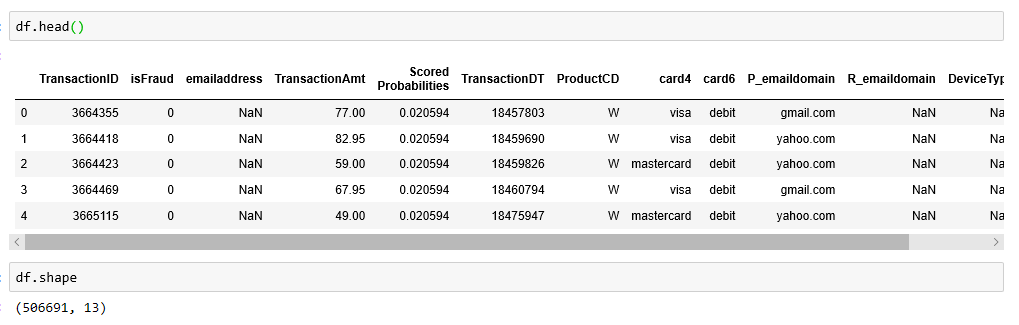
 Since our model helps only in detecting the intrusions but don’t block them. In real time all the data is monitored and controlled by data. World. And we didn’t change the original data. It remains the same.

**Consequences:**

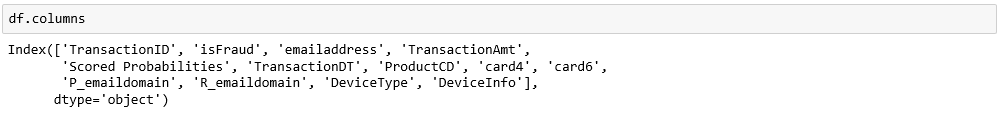
Fraud is the significant issue in our model. In some cases, fraud is less in the dataset. In this situation data is imbalanced must spend more time on deciding them otherwise we balanced the data.

**Dataset:**

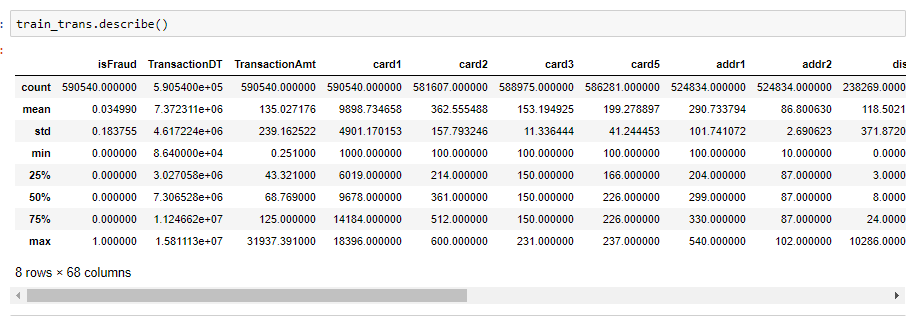
* We got the dataset name hiwott-cyber-security-dataset. This dataset contains 50,6691 rows and 13 columns. We have more data for test and train the model to predicting the accuracy.



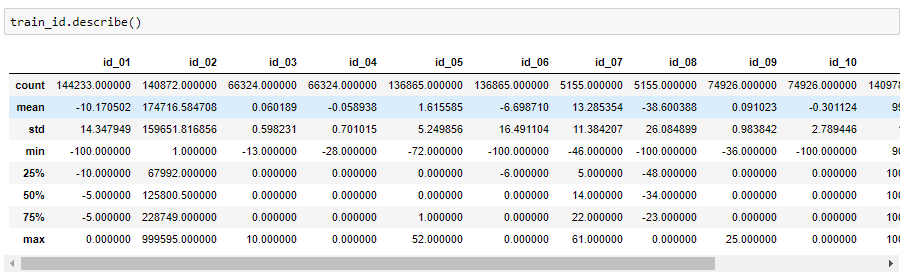
* The dataset contains the following columns



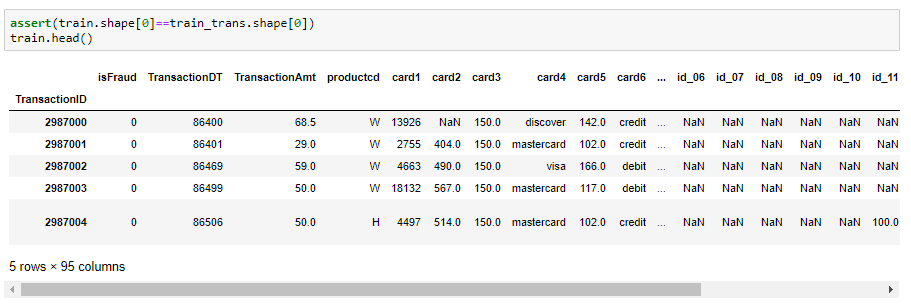
* We have datasets 4 datasets. Two train and test datasets for identity data and two train and test dataset for transaction.
* **Train data for transaction**



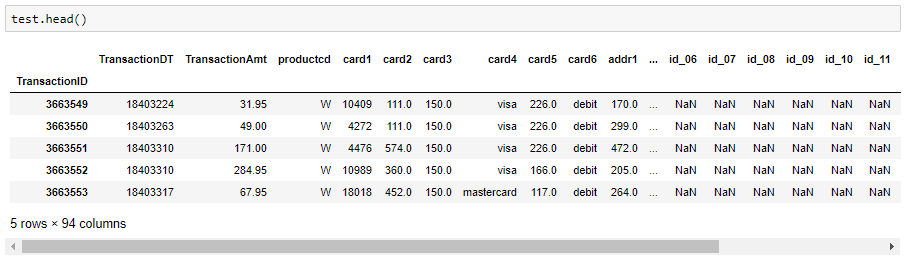
* **Train data for Identity**



* **After merging both train datasets of identity and transaction data.**



* **After merging both train datasets of identity and transaction data.**



**Methodology:**

* **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. Visa card had the maximum fraud and the amount was 59% while discover type had 2% of fraud cases only.

The above chart reveals that the percentage of fraud in our dataset is very less and the fraud number is near to 20663.

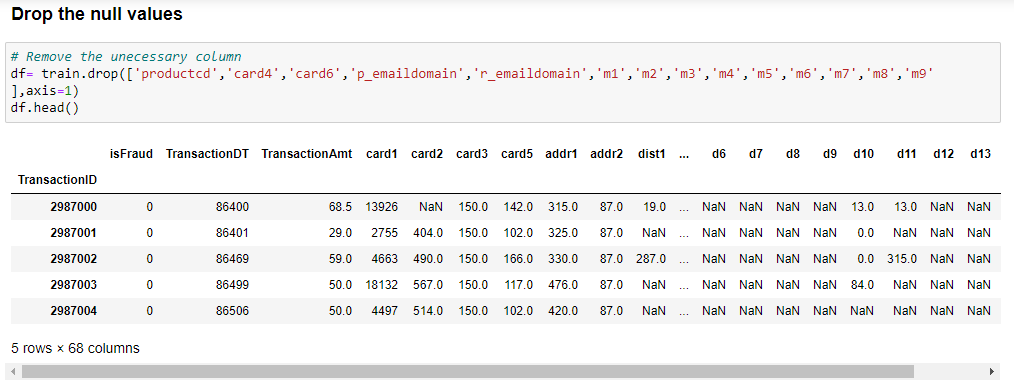
* **How the dataset fit the problem:**

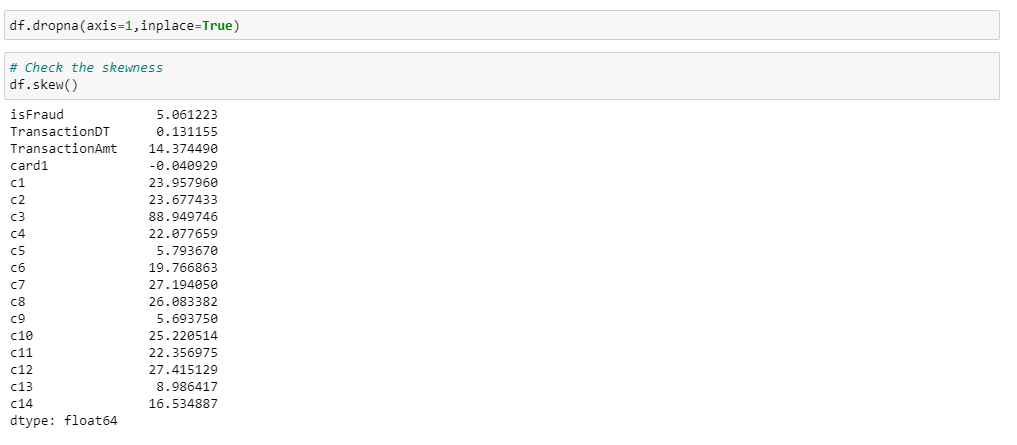
This dataset contains unnecessary columns and null values. So, we must do some cleaning and use pre-processing techniques for further analytics. And pre-procession is used to resolve the analytic problem. So, our topic relates to the analytical problem.

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 is some fraud happened due to the use of unsecure WIFI and networks.

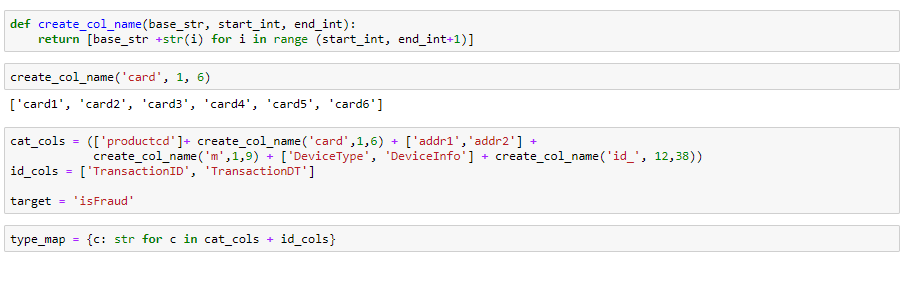
* **Cleaning and Pre-Processing:**
* **Remove the null values and unnecessary columns**





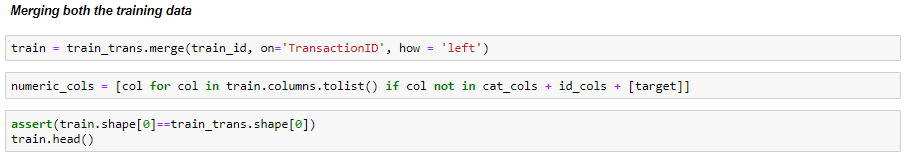
Above code is used to remove the unnecessary columns. As, these columns having the string data and we don’t want to use these. The 2nd code helps to remove the null values from the data and the 3rd code is used to check the skewness of the data.

* **Create the column names:**

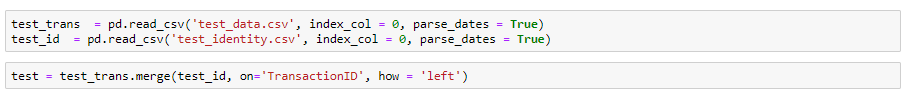


This code helps to create the column for different types of data. We create a column for different card type. In our dataset there are six different type of card, so we make the “Create\_col\_name” and put 1 to 6 card type. In “**cat\_cols**” we put all the categorical columns. And in “**id-cols**” we put the ‘transtionID’ and ‘transaction data’ and time column values. ‘is Fraud’ column is our target variable, so we create this one a “**target**” variable.

* **Merge both train and test data for identity and transaction data:**

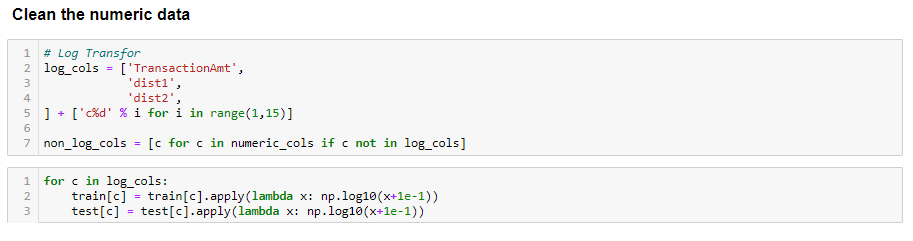


In the above we merge both train data of identity and transaction by merge code and put it into the train variable. We merge both the column with left join so no data will be left behind, and the common column was “TranactionID”. Then check shape for new data and the old data. If the shape for both one is equal, then the merging is right.



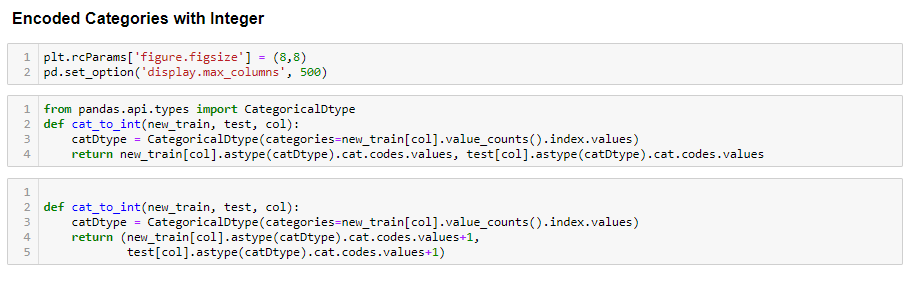
Also, merge both test dataset for identity and transaction data. And the code is given above.

* **Clean the Numeric Data:**



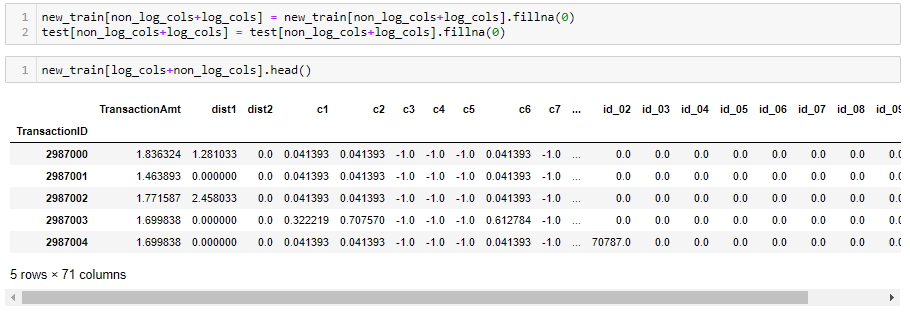
We have numeric values in our dataset. So, we put the numeric values in the “log\_cols”. And then create the array for this. While running the model python always shows an error. Therefore, we need to clean the numeric values.

* **Encoded the Categories into integer.**



Every time when we use the model to run in the python it always shows us an error can’t convert string into float. So, we encoded the categories into the integer.

* **Clean the null values**



Our data contains huge number of null values. So, we fill the null values with the mean of the values. Mostly in other cases we drop the null values but for the better accuracy we fill these with the mean values.

**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.

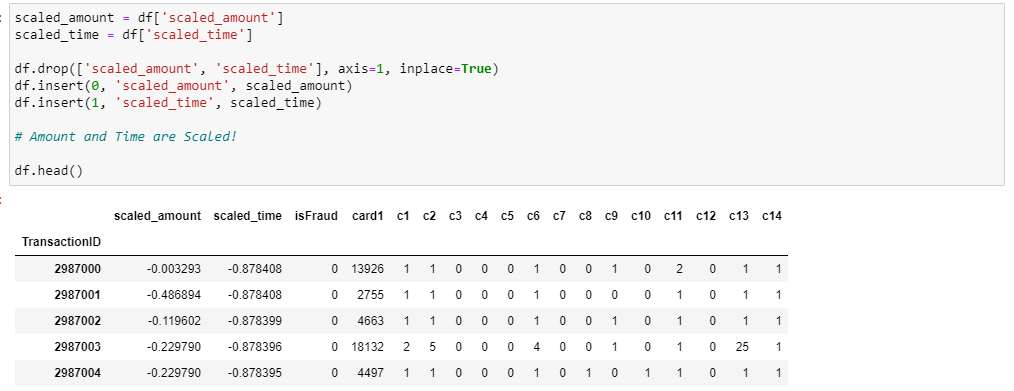
* **Check whether the data is balanced or not.**



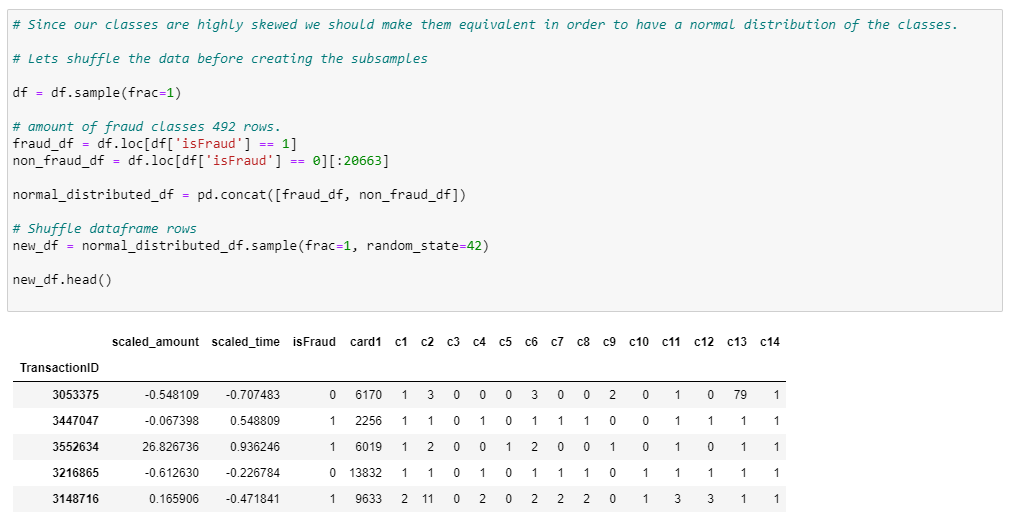
The data is not balanced. So, we used the standard scalar techniques to balance the data. As we saw in the previous graph it shows that data is highly imbalanced. 96.5% of the data is non fraud and only 3.5% of the data is fraud. The original data is having the more no fraud cases than the fraud cases. 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:**

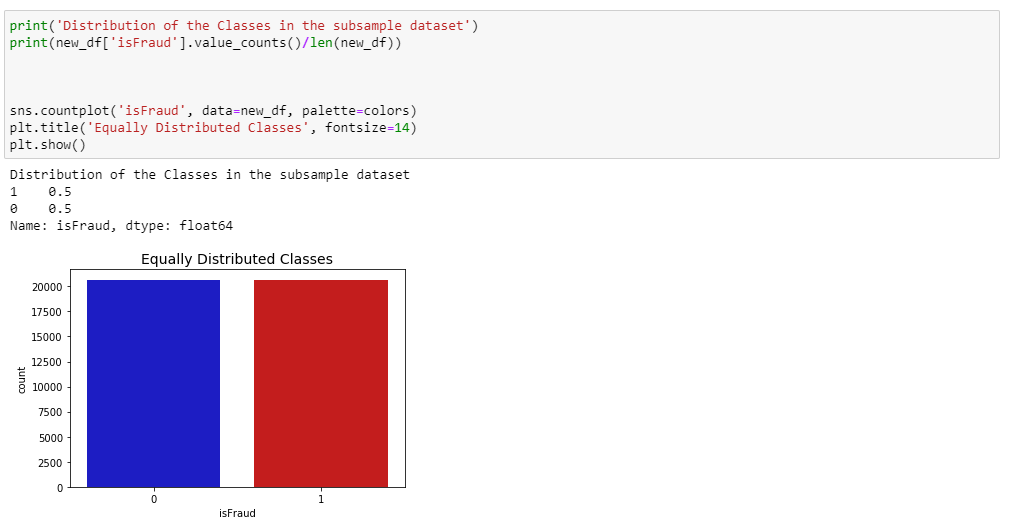
We can scale the data with the help of transaction time and transaction amount columns. We also create a sub sample of the data frame in order to equally distribute the no fraud and fraud cases.



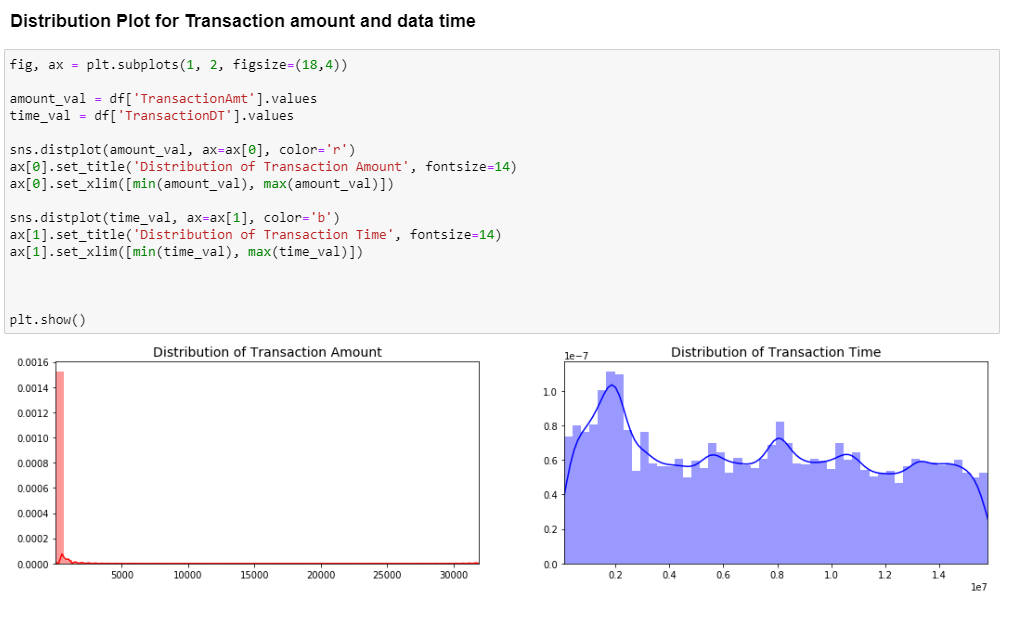
Our dataset contains 20663 fraud cases so we can get randomly 20663 of no fraud cases to create a new sub- sample. We concat the 20663 fraud cases to scale the data and the code is given below:



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:



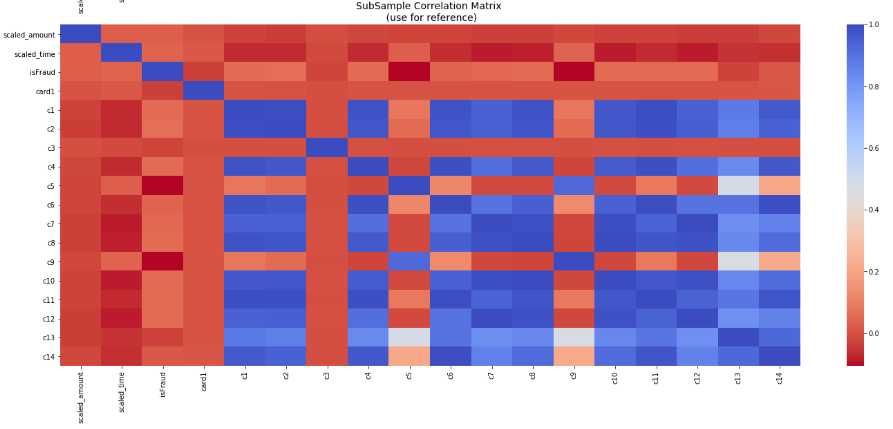
* **Plotting:**



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.



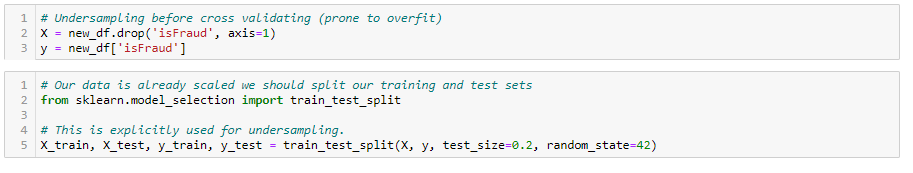


* **Modeling:**

For modeling we ran the different method of classification as this is the classification problem. We run the different model for neural network we run LSTM and CNN models. For classification problem we ran Logistic Regression, Decision Tree Classifier and Random forest Classifier. The data set is too big and containing huge number of columns, so we made the # different python files. In one file I rum the LSTM model and in the second one we did CNN model and in 3rd we did the classification models.

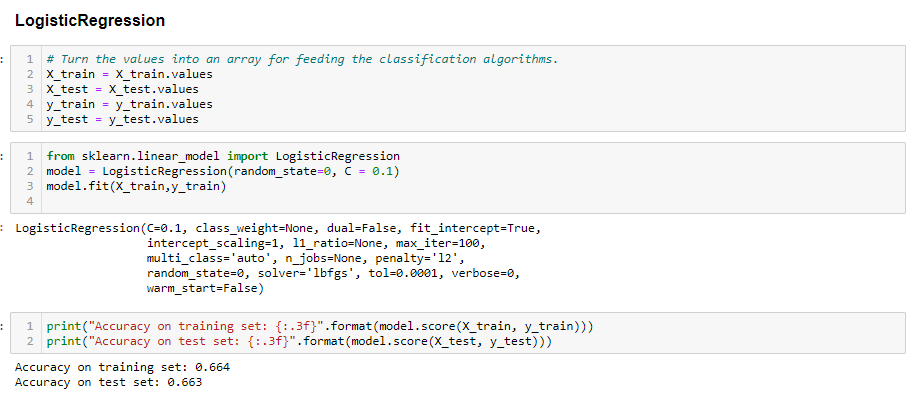
* **Split the data for model**

We split the data into train and test data for modeling. Train data predict or forecasting on the past data and test data help to evaluate the accuracy of the model.



This code helps us to split the data into train and test column. We drop the “isFraud” column from the train data. And made this column target variable and put into test.

* **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.



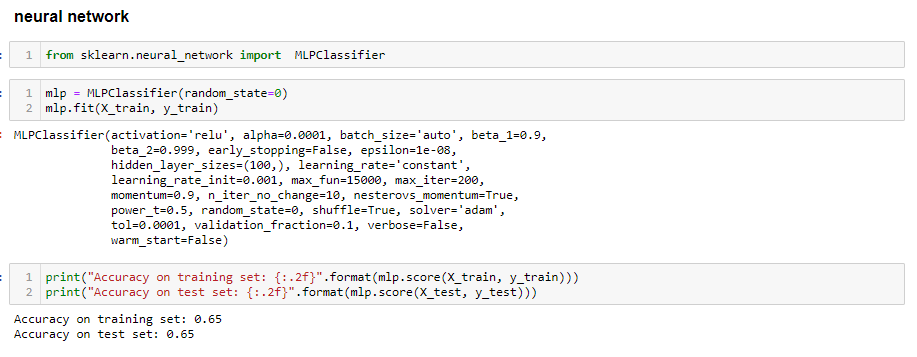
The accuracy of train and test set is very less. So, we reject this model.

* **Decision Tree Classifier:** A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.



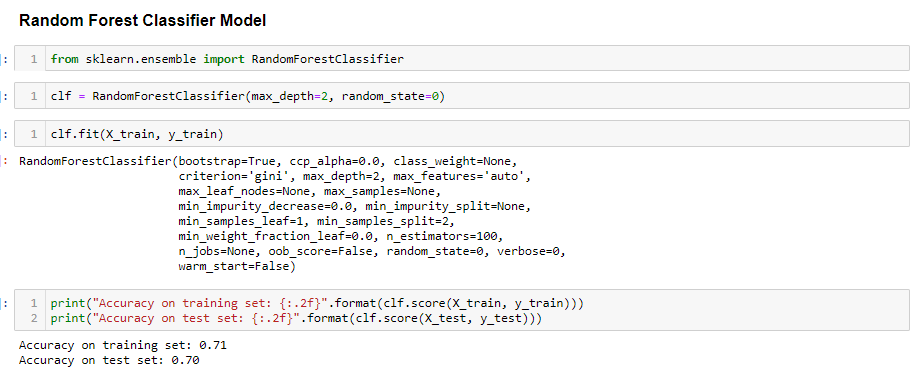
We choose this model as the best model among the other classification models as it predicts the better accuracy. Train and test set accuracy are better for prediction.

* **MLP Regressor:** MLP trains on two arrays: array X of size (n\_samples, n\_features), which holds the training samples represented as floating point feature vectors; and array y of size (n\_samples,), which holds the target values (class labels) for the training samples:



We reject this model as it didn’t predict the good accuracy for the model the accuracy for both train and test data are very low.

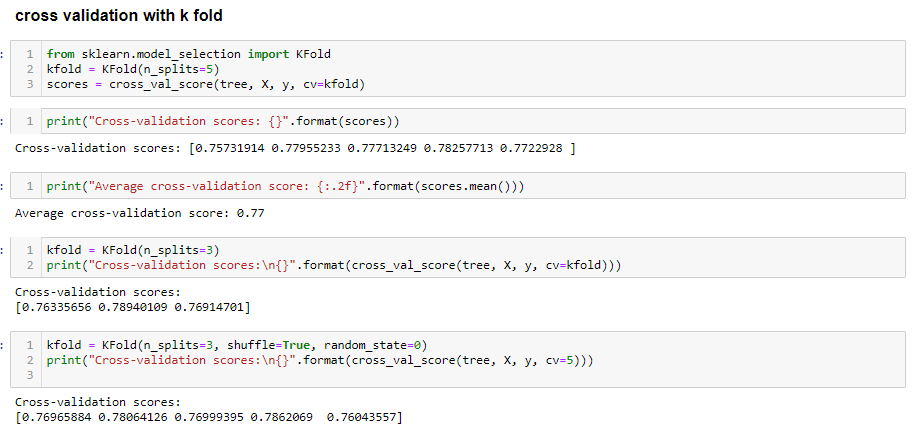
* **Random Forest Classifier:** A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement.



This model predicts the good accuracy for both train and test set. But the accuracy is lower than the Decision tree classifier. So, we reject this model too.

* **Cross Validation:** This method is used to increase theaccuracy for themodel. We can set k-fold parameters to increase the accuracy. As we know that decision tree classifier predicts the better accuracy. So, we used decision tree classifier for cross validation to increase the accuracy.





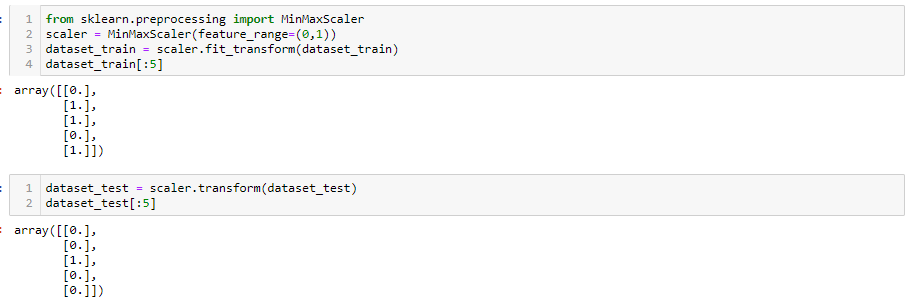
* **LSTM Model:** LSTM model is the part of the RNN (recurrent neural network). LSTM model is used to predict the best accuracy for the model. LSTM are the cells in the neural network. Its working is depending on the sequential data.

But the working for this model is slightly different from the other model that we used in our project. So, we must again split the data into train and test.

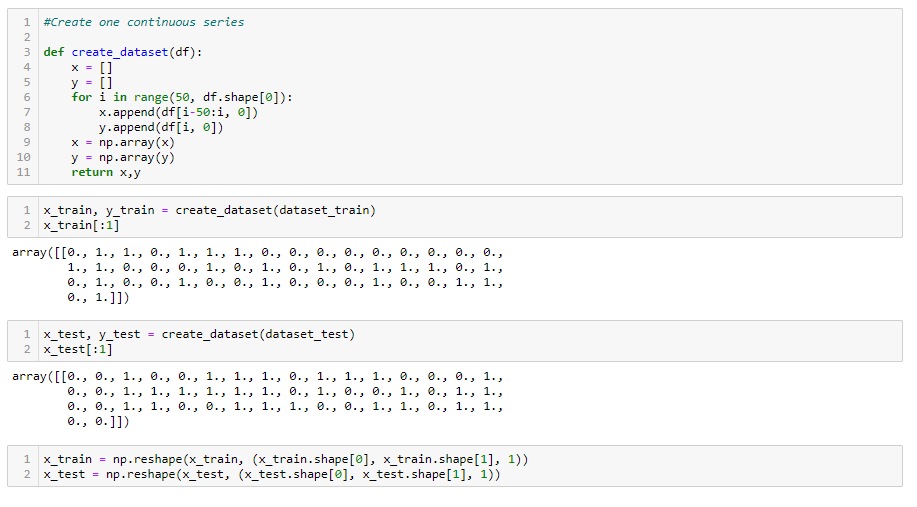
First, we set the X and y and then split the data into train and test set. Then check the shape of the data. And it is shown by the below code.



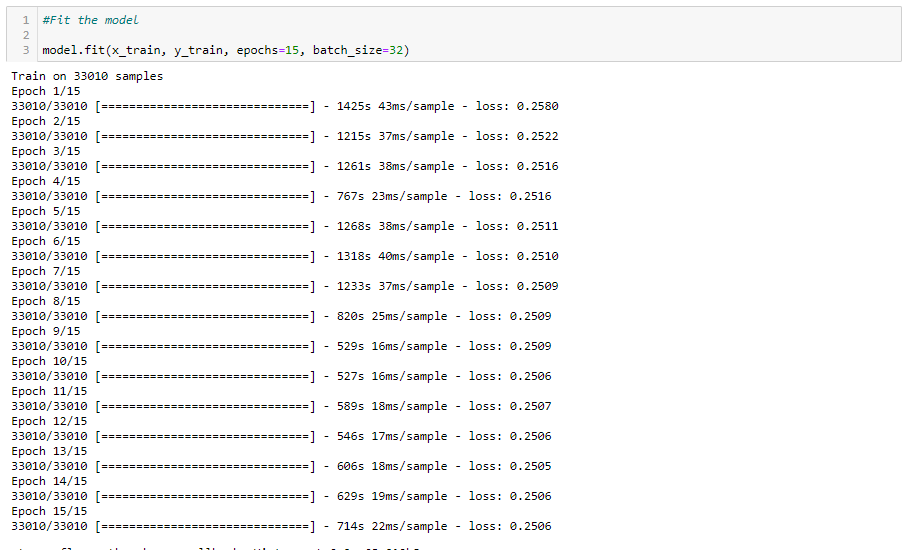
After that we scale the train and test set and transform it into array. Also reshape the values of the array.

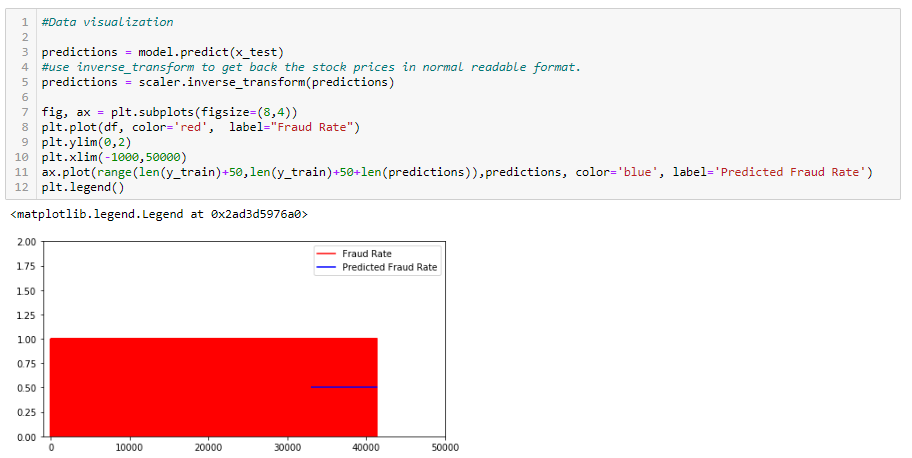


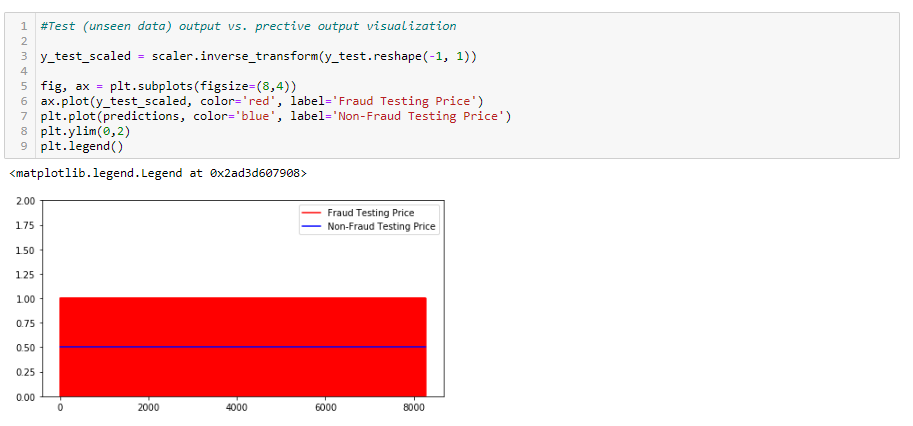
Now we reshape the values of the sets with using the bottom code.



* **LSTM Coding**





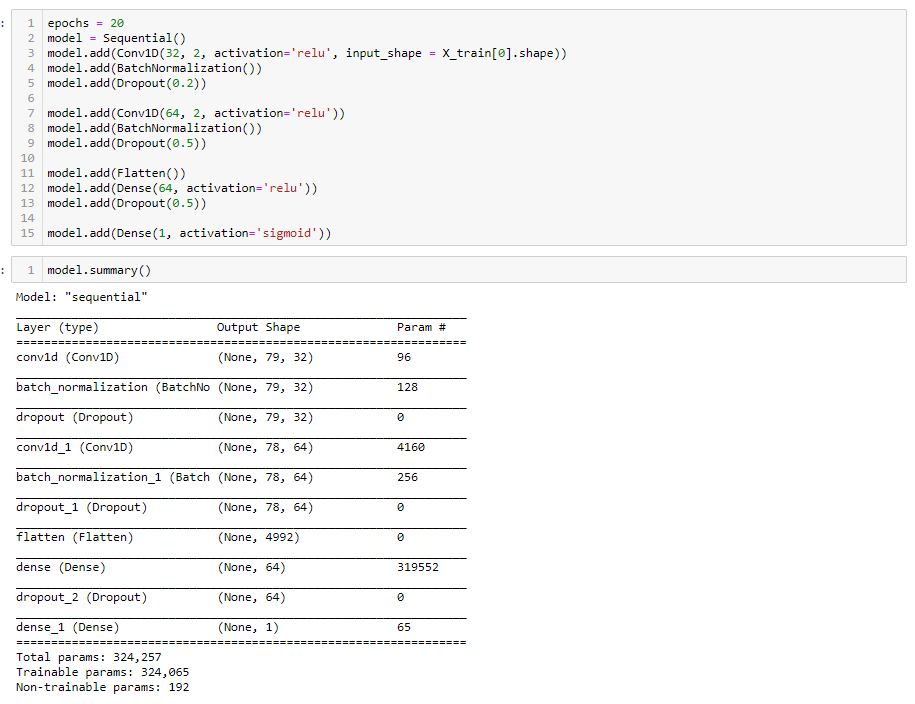


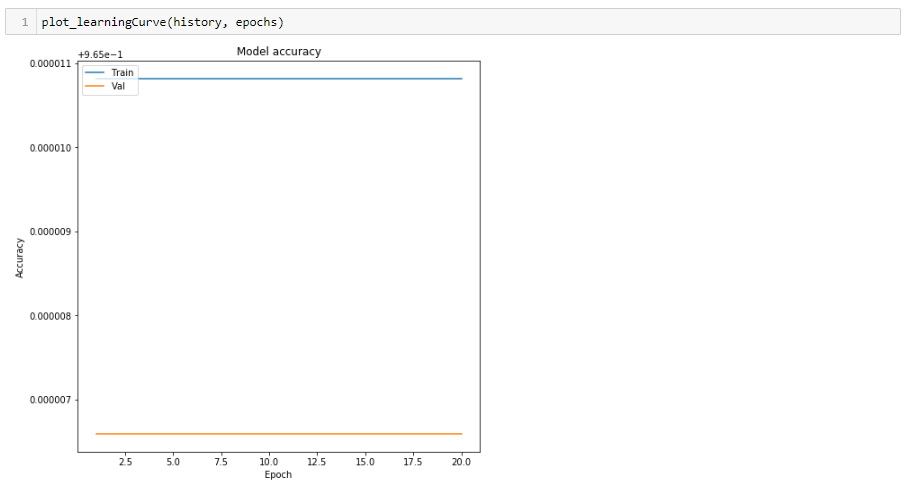
The LSTM model set on the epoch value equal to 15 as it predicts the best accuracy for model. And it creates the output in a sequence and creates the history.

With help of history of the accuracy and val\_accuracy we create the plot for the accuracy. The fraud is set on 0.5 accuracy. And this predict the best model for pour project.

* **CNN (Convolutional Neural Network):** It is another form of the neural network, but the working CNN is slightly different from the other models. As in neural network every Input layer node relates to hidden layers node. But in CNN it is not as the neural network. CNN is another form of the deep learning. So, we use this model to predict the best accuracy for our problem. Most anomaly detection approaches using MLP can detect novel real-time attacks, but still has high false alarm rates. Most attacks are composed of a series of anomaly events. These attacks are called time-delayed attacks, which current neural network IDSs cannot identify efficiently.

For preparing the data or split into the test and train are like the LSTM as shown above:





We put the epoch equal to 20 to predict the better accuracy. But it didn’t give us the better accuracy at all the time the history came with fraud 0.5 accuracy. So, we reject this model and accept the LSTM and Decision Tree Classifier Models.

**Conclusion:**

We achieved everything we wanted to in this project. We can now say that, prediction of fraud is not only based on one or two factors. It is based on many factors account type, Transaction amount, transaction date time and account’s previous records etc.

We used many models to get a good model and we came with the Decision Tree Classifier and LSTM, that have good accuracy on training and test data.

Because the Decision Tree Classifier predict the better accuracy for the data

**Contribution:**

* AMITA MEHTA
  + Literature Review, Coding, Prepare the presentation, Clean the data, ran the basic models and help Jai to evaluate the model to increase the accuracies.
* JAI SURYA
  + Work on the python file, work on the neural network and help Amita to run the models.
* RAJIT KUMAR
  + Cleaning and Pre-processing, Scaling, Create Correlations, modeling and help in preparation of final report.

**Challenges:**

* Finding the topic. And it is an analytical problem or not
* Finding the dataset.
* Find the relative data for comparison between the two field. Because this is very vast field to work.
* We got a **credit card fraud dataset** from **Kaggle**. because every project in Kaggle was done with the same dataset. So, we refuse that one.
* dataset from the open data source was good and relate to our problem but that havingS less column which doesn’t help us in this project.
* We search and read number of articles on the cybercrime topic and finally we choose this topic as this is the measure factor.
* *Ts difficult to find or do the unique or different work among the others. As a* lot of work is done on the internet regarding this project.

**Goals:**

* We analysis our data to detect the fraud in credit card transactions. Do some python coding if it requires. To solve this problem, we are planning to create a survey that help us to find the rate of frauds among the peoples. And ask for the solutions.
* Look at the places that are highlighted in frauds and crimes. Provide them the solution to solve this problem.
* Our aim is to detect the 100% of the frauds in online transaction while minimize the incorrect fraud classification.
* Response activities effectively coordinated with internal and external stakeholders.

**References:**

* Data.world. (2020). Cyber Security Dataset - dataset by hiwott.

**Available at:** <https://data.world/hiwott/cyber-security-dataset>

* Ieeexplore.ieee.org. (2020). Credit card fraud detection with a neural-network - IEEE Conference Publication.

**Available at:**

[Final Project Report.docxcessed](Final%20Project%20Report.docx)  24 Feb. 2020].

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

**Available at**[: http://www.irjeas.org/wp- content/uploads/admin/volume7/V7I1/IRJEAS04V7I101190319000002.pdf](:%20http://www.irjeas.org/wp-%20%20content/uploads/admin/volume7/V7I1/IRJEAS04V7I101190319000002.pdf)