Fraudulent Transactions Detection using Machine Learning

Asher George Jacob

Abstract: ***Information Technology has revolutionized and influenced each and every sector of the nation with banking sector not being an exception. India has moved from a manual, scale restricted space to an environment which has opened vistas of systems which are automated and computerized. Various customer-oriented products like ATM services, mobile banking, digital wallet online payment has made the life of the customers convenient. Lot of innovations have been made in the recent years across the banking sector through enablers like artificial intelligence, payment provisions, biometrics etc. Phenomenal improvements have been witnessed in the banking space with an increase in competition. However, making use of technology and sophisticated products pose a lot of discommodes. There are numerous challenges that need to be addressed. Credit frauds in the banking sector are seen as a common practice as numerous scams and scandals have been uncovered in the past few years. This research aims at identifying how frauds and scams in banks occur when customer makes payment through online mode. An attempt has been made in this research to detect such kinds of frauds using a machine learning algorithm. The random forest algorithm has been used for classification and detection of fraudulent transactions.***

Keywords: Artificial Intelligence, Scams, Scandals, Fraudulent Transaction, Machine Learning, Random Forest Algorithm

1. Introduction

India is one of the largest democratic countries as well as a potential global super- power of the world. There have been various key strategic developments attained by the country which has enabled it to achieve such significant position in the global world order. Establishment of advanced Banking system and incorporation of state-of-the-art technology has undoubtedly played an integral part in attaining this global recognition in the world. Banks act as an engine and propeller of growth in the economy. It acts as a trustee as it safeguards the savings of the people and further channelizes and processes it by converting them into investments. The role that banks therefore play cannot be undermined. Banking sector has been one of the biggest contributors amongst the service sectors in India. Today banks give a lot of emphasis on retaining customers than on mere customer acquisition. Banks therefore try their best to offer services which not only would provide ease and comfort to their customers but provide them with services which are easily accessible and available to all Integrating Information technology into banking practices has revolutionized the entire banking sector of the economy. Today banks offer to the customers various customer- oriented products like internet banking, ATM services, tele-banking, electronic payments etc. Online payments made through credit and debit cards has enabled the customers to pay their bills, file their income tax returns as well as trading of shares extremely easy and convenient. It has undoubtedly made our life easier and convenient. However, the stark reality is that with convenience comes challenges. In the recent times, banks have become a soft target for fraudsters, and they have misused and found ways to cheat and trick the innocent populace. Incorporating tight and robust security measures for banks by imposing stringent regulatory norms can play a key role in maintaining the health and stability of an economy. Impostors have found devious ways in carrying out duplicitous practices in the banking sector. Phishing e -mails, hacking of cards, making fake calls to steal PIN numbers etc. are few of the most popular and common fraudulent practices observed in this industry. These are generally caused by the paucity in the operational modus of the existing security management of the banking businesses. Thus, constant, and stringent observation, monitoring, examining, and scrutinizing the security system of the banking operations alone can aid in countering and offsetting such misconducts that is seen prevailing in the banking activities. This study thus aims at analyzing ways to identify online fraudulent practices and thereby ensue in attaining stability and effectual functioning of the domestic banking system in the country.

2. Need of the Study

A study by Statista has revealed the following glaring reality in context to the online banking fraudulent practices which makes this study essentially important and imperative as it tries finding solution to this grave problem that this sector has been going through.

- It has been observed that the use of credit and debit cards have increased phenomenally since the wake of Liberalisation, Privatisation and Globalization in 1991 and since then the number of cards being issued have risen phenomenally. It has been estimated that the number of credit cards would reach 44 million by 2025 and those of debit cards around 970 million.
- Between July 2019 and June 2020, it was observed that there was a total of 290 thousand banking related complaints received of which nearly one third of these complaints were concerned with the public sector banks and another 100 thousand complaints were from private sector banks.
- In financial year 2021, the Reserve Bank of India (RBI) reported a total of around 7,400 bank fraud cases amounting to around 1.38 trillion Indian rupees across India.
- In 2019 the state of Maharashtra alone recorded almost 552 cases out of the total 2000 cases of internet banking fraud cases which was reported from across all the states in India.

All these statistics certainly divulges the need to study in detail the ways to identify and unravel such malpractices being practised in this sector. Unless manpower, capital and finances are poured into examining and scrutinising financial fraud, the problem is bound to advance further and become more intricate.

Objectives of the Study

- To analyse the overall online banking frauds taking place in the eco-system.
- To examine the popular ways that online fraudulent practices are carried on.
- To develop an algorithm which will help in identifying such online bank frauds and thereby minimise such miscreant practices in this sector.
- To suggest policy initiatives for the smooth working of the online banking mechanism.

3. Literature Survey

Review of the literature reveals the kind of research that is already being carried in the area mentioned as well as highlights any gaps that may exist in the research. Following are few of the observations made by researchers on online fraud detection using machine learning models.

Table 1: Literature Survey

Sr. No	Researchers	Techniques	Observations
1	O. Shmatko Olexander, V. Fedorchenko, D. Prochukhan	Random Forest Classifier	The model works well for large training data, but the speed suffers during testing and application development.
2	K.S. Varun Kumar, V.G. Vijaya Kumar, A. Vijay Shankar, K. Pratibha	Logistic regression, Naïve Bayes, Decision trees, ANN	ANN model gave the best accuracy, recall and precision amongst all the techniques used. Various techniques such as average method, moving average or window method, naive method and sessional naive methods can be used to reduce the parameters in time series analysis.
3	Aditya Oza	Logistic regression, SVM	The model was able to detect the frauds with a very high accuracy rate and low false positives. To leverage the categorical features associated with the dataset, decision tree algorithms has been used.
4	Adi Saputra, Suharjito	Neural Network, Naïve Bayes, Random Forest, Decision Trees	SMOTE (Synthetic Minority Oversampling Technique) was used to work with the imbalanced dataset. Neural Network gave the best accuracy for detection of fraud in the dataset. The neural network model worked well because of the genetic algorithm which helps in improving the performance of ANN (Artificial Neural Network).

4. Research Methodology

4.1 Dataset Information:

The Credit card fraud detection dataset has been taken from the Kaggle Website. In September 2013 over a span of two days in Europe certain fraudulent transactions had been detected and based on the data obtained a dataset has been created. Through this dataset an attempt has been made to identify the fraudulent transactions using the various machine learning algorithm techniques. To conceal the individuals' personal details some of the information has not been stated in the dataset. It has therefore been classified into the 28 components v1 to v28.The dataset comprises of 2,84,807 rows and 31 columns in all. The columns have been categorized into various components ranging from time which is measured in seconds. The column also has various other categories ranging from v1, v2……. v28 of which the last two columns of the dataset comprises of amount and class. The column labelled as class comprises of values 0 and 1 where 0 represents normal transaction and 1 indicates fraudulent transaction.

4.2 Techniques

The Jupyter notebook IDE has been used for programming in Python language as it is easy to use and implement. Jupyter notebook has been divided into several individual blocks of code which can be run to give the desired output. The errors can thus be easily determined by running the individual block of code. Various packages can be easily imported in Jupyter notebook to carry out various data science techniques. Some of the packages imported in this project are numpy, pandas, Matplotlib and scikit-learn. Each of these data packages have unique features. The numpy package is helpful in working with numerical data while the panda package helps in the data science process which includes data cleaning, data transformation, data integration etc. The matplotlib package helps in the analysis of data by plotting various histograms, figures, charts, plots etc. The Scikit-learn (sklearn) provides efficient tools for various machine learning classification and regression models.

4.3 Procedure

To start with, the above-mentioned packages were imported into our project. By using pandas, we have loaded the dataset into our python project. Various data pre-processing techniques were applied such as identifying and handling the missing values, features scaling, splitting the dataset etc. As the dataset contains 0 null values there are no missing values in the entire dataset. By analyzing the column labelled as class it can be observed that there are a total of 2,84,315 non-fraudulent transactions and a total of 492 fraudulent transactions.

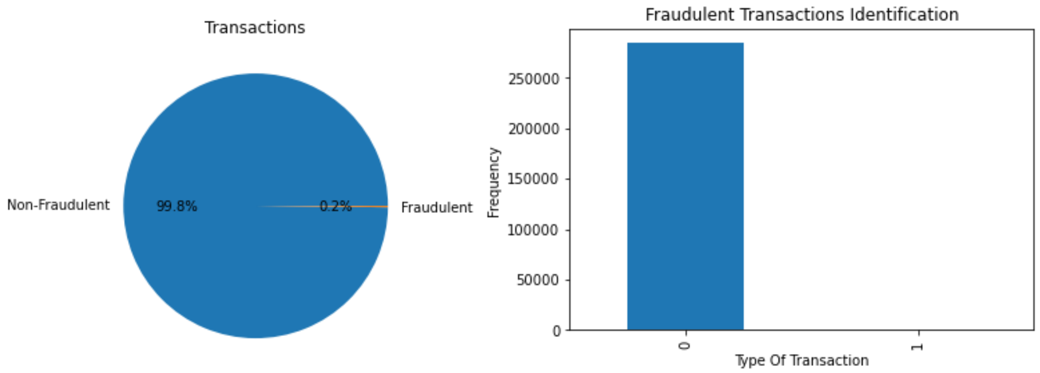


Figure 1: Transaction Identification

As the number of fraudulent transactions are extremely less as compared to the non-fraudulent transactions, the dataset is an example of imbalanced data. The problem to deal with imbalanced data is that the classifier tends to favor the majority class, in this case it is the class of non-fraudulent transactions. It also leads to large classification errors over the fraudulent cases. Therefore, to deal with the imbalanced data various techniques can be used such as resampling the training set, using K-fold Cross-validation, using right evaluation matrix, trying out different algorithms etc. In our project to deal with the imbalanced data we have used the under-sampling technique. We create a new sample dataset which consists of similar distribution of fraudulent and non-fraudulent transactions and apply the various machine learning algorithms on this new dataset.

5. Machine Learning Algorithms

The following algorithms have been used to detect the fraudulent transactions from the dataset.

1) Random Forest Algorithm
2) Logistic Regression
3) Decision Trees

1) Random Forest Algorithm:

Random Forest Algorithm as the name suggests selects randomly the samples from the dataset. It is a supervised learning algorithm which is used for classification as well as for regression. A decision tree is then formulated for the sample analyzed and the results are predicted. Random Forest is collection of decision trees trained through the bagging method or the pasting method. It is most ideal and convenient to use the random forest classifier for classification purposes. It includes all the hyperparameters of a decision tree classifier and bagging classifier as well. This algorithm helps in obtaining the most significant features from the dataset. As the algorithm takes the average of the predictions, there isn’t any possibility of over fitting.

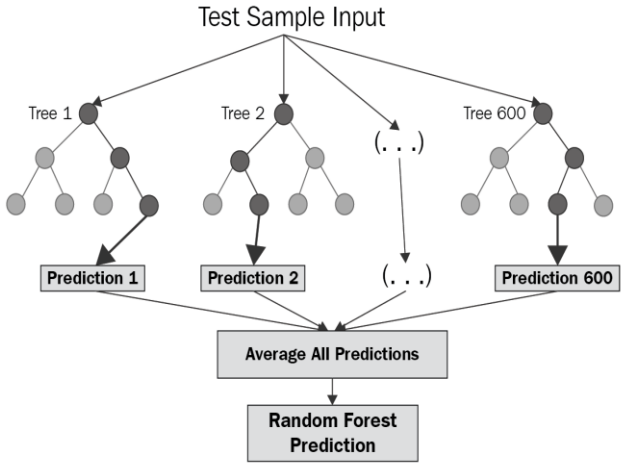


Figure 2: Random Forest

Source:https://medium.com/swlh/random-forest-and-its-implementation

2) Logistic Regression:

It is a supervised learning classification algorithm. It is a binary classifier which checks whether an object belongs to a specific class or not. If the estimated value is less than 50 % the prediction is that it is a negative class and it is labelled as 0, else it is positive class and labelled as 1. The logistic function is also known as the sigmoid function. It is mainly used for the categorical data prediction. Logistic Regression is classified as binary, multinomial and ordinal regression wherein a binary logistic regression there are only two scenarios, in multinominal three or more scenarios are possible without ordering and in ordinal three or more scenarios are possible with ordering. Logistic Regression is easy to implement and has a very good accuracy to predict the results.

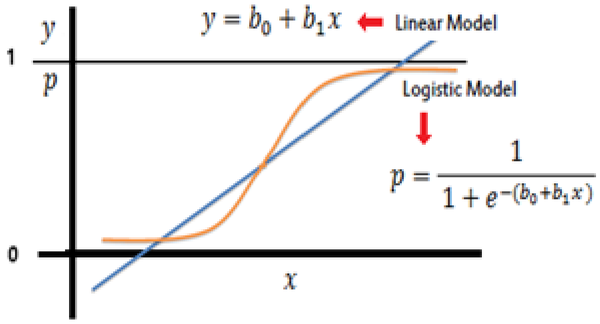


Figure 3: Logistic Regression

Source: https://www.saedsayad.com/logistic_regression.htm

3) Decision Trees:

Decision trees are supervised machine learning algorithms that can perform both classification and regression tasks. They have the capability to work with very complex datasets. They are the fundamental components used in the random forest algorithms. As the name suggests it builds the model in the form of a tree. It can be worked with numerical and categorical data. The main feature of decision tree is that it splits the data tree into individual subsets, classified by the noteworthy attribute in the entire dataset. Decision trees are simple, can be easily understood and does not require feature scaling.

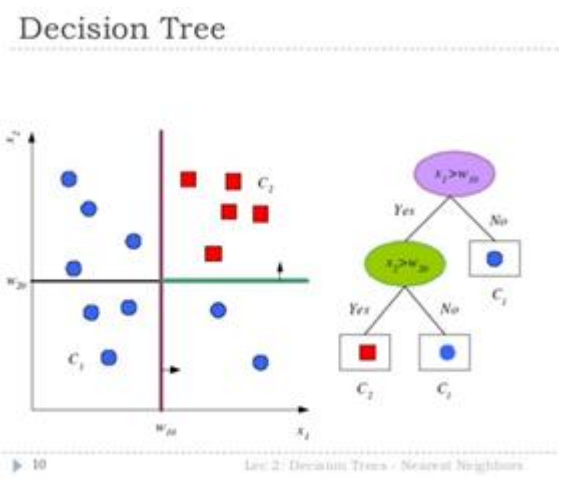


Figure 4: Decision Tree
https://www.lewisgavin.co.uk/Machine-Learning-Decision-Tree/

6. Results & Analysis

The dataset comprised of 492 fraudulent transactions and 2,84,315non-fraudulent transactions. These two categories of transactions were stored in new individual data frames.As the dataset loaded was highly imbalanced, the random under sampling technique was used to balance the data. Out of the total non- fraudulent transactions 492 transactions were randomly selected and stored in a new data frame. This data frame was then concatenated with the 492 fraudulent transactions to create a new dataset on which the machine learning algorithms were applied.

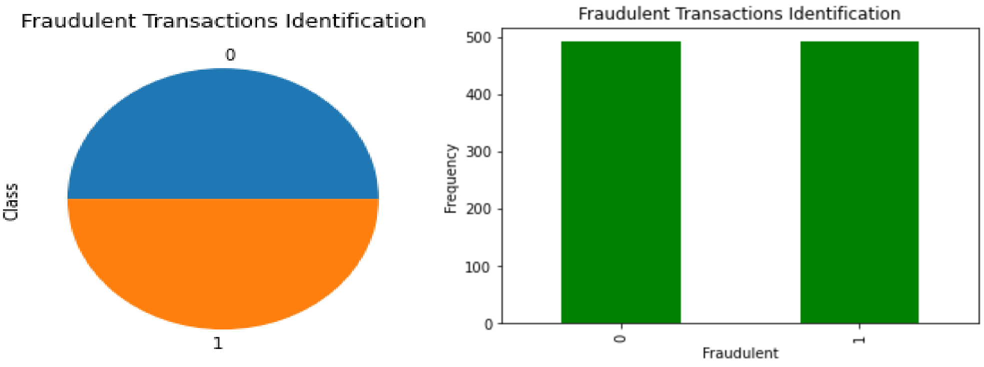
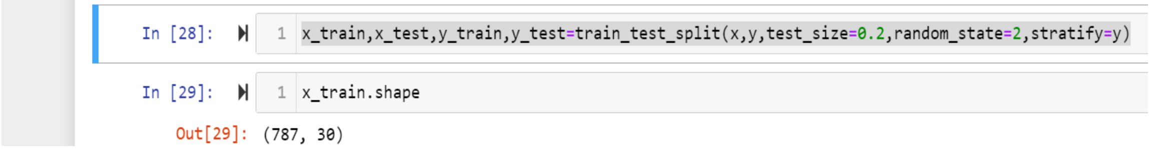


Figure 5: Balanced Dataset

Using the Scikit-learn library this new data has been split into train and test dataset.

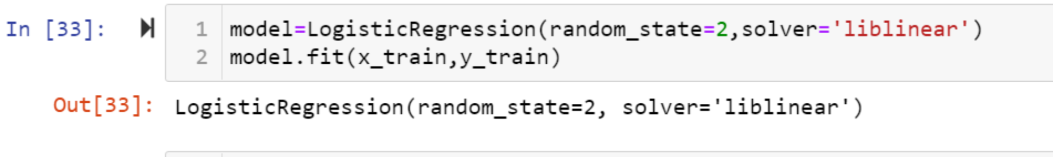


```
In [28]:  x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=2,stratify=y)

In [29]:  x_train.shape

Out[29]: (787, 30)
```

The following are the results acquired from the analysis carried out by using various machine learning algorithms.

- **Logistic Regression:**

The LogisticRegression model was imported from the Scikit-learn library and applied to the training dataset.



```
In [33]:  model=LogisticRegression(random_state=2,solver='liblinear')
          model.fit(x_train,y_train)

Out[33]: LogisticRegression(random_state=2, solver='liblinear')
```

The following evaluation metrics was obtained when the model was applied to the test dataset.

Table 2: Evaluation Metrics-Logistic Regression
Test Data

Evaluation Metrics	Result
Accuracy	0.9137055837563451
Precision	0.8571428571428571
Recall	0.9655172413793104
F1	0.908108108108108

The confusion matrix was plotted on the test data.

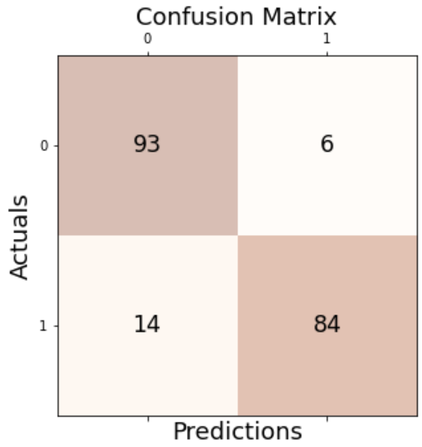
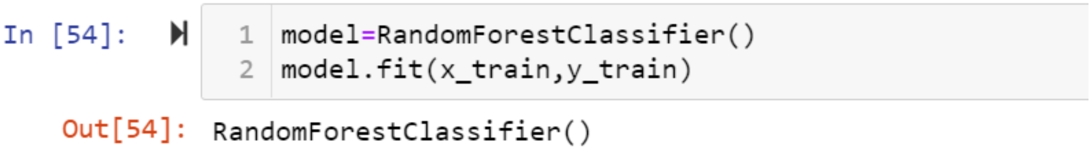


Figure 6: Confusion Matrix- Logistic Regression

- **Random Forest**

The RandomForestClassifier model was imported from the Scikit-learn library and applied to the training dataset.



```
In [54]:    model=RandomForestClassifier()
            model.fit(x_train,y_train)

Out[54]:  RandomForestClassifier()
```

The following evaluation metrics was obtained when the model was applied to the test dataset.

Table 3: Evaluation Metrics-Random Forest
Test Data

Evaluation Metrics	Result
Accuracy	0.9289340101522843
Precision	0.8673469387755102
Recall	0.9883720930232558
F1	0.9239130434782609

The confusion matrix was plotted on the test data.

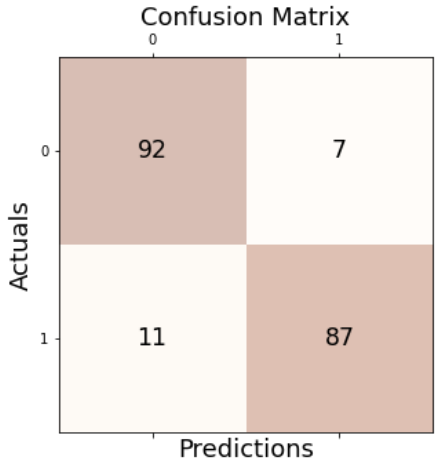
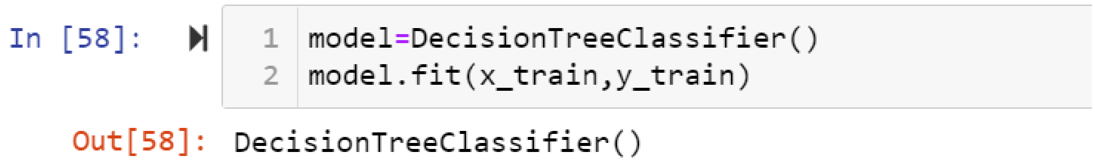


Figure 7: Confusion Matrix-Random Forest

- **Decision Tree**

The Decision Tree Classifier model was imported from the Scikit-learn library and applied to the training dataset.



```
In [58]:   model=DecisionTreeClassifier()
           model.fit(x_train,y_train)

Out[58]: DecisionTreeClassifier()
```

The following evaluation metrics was obtained when the model was applied to the test dataset.

Table 4: Evaluation Metrics-Decision Tree

Test Data

Evaluation Metrics	Result
Accuracy	0.8984771573604061
Precision	0.8877551020408163
Recall	0.90625
F1	0.8969072164948454

The confusion matrix was plotted on the test data.

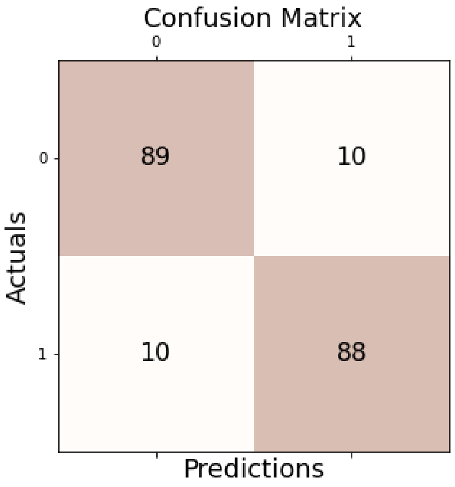


Figure 8 : Confusion Matrix- Decision Trees

7. Comparative Analysis of various models used

Table 5: Comparative Analysis

Evaluation Metrics	Logistic Regression	Random Forest	Decision Tree
Accuracy	0.9137055837563451	0.9289340101522843	0.8984771573604061
Precision	0.8571428571428571	0.8673469387755102	0.8877551020408163
Recall	0.9655172413793104	0.9883720930232558	0.90625
F1	0.908108108108108	0.9239130434782609	0.8969072164948454

8. Conclusion

The creditcard dataset was a highly imbalanced dataset. This research made use of the random undersampling technique to balance the data. The following machine learning algorithms have been implemented in this study. They are Logistic Regression, Random Forest and Decision Trees. Through this study it was observed that the Random Forest algorithm gave the best accuracy with a value of 92%, followed by logistic regression and decision tree with a value of 91% and 89% respectively. With regard toprecision, it was observed that decision tree gave the best results with a value of 88% followed by Random Forest and Logistic Regression with a value of 86% and 85% respectively. The recall scores were highest for Random Forest with a value of 98% followed by Logistic Regression and decision tree with a value of 96% and 90% respectively. A similar result was observed for F1 as well. Random Forest gave the best results with a value of 92% followed by Logistic Regression and Decision Tree with a value of 90% and 89% respectively. This proves that the random undersampling technique method is effective in increasing the performance of unbalanced data classification. To obtain improved results the concept of deep learning and neural networks can be used. With increasing facilities provided by the banking sector there is an equally increasing need to make the system more robust. Making payments by using the credit card is one of the most popular methods used by customers nowadays which however is not bereft of theft and ambiguities. Therefore, it certainly demands the need to foster a system which can unravel such fraudulent banking transactions that are carried out using credit cards.

References

[1] https://www.statista.com/topics/8143/credit-and-debit-card-market-in-india

[2] Scientific Collection «InterConf», (71): with the Proceedings of the 3 rd. International Scientific and Practical Conference «Current Issues and Prospects for the Development of Scientific Research» (August 19-20, 2021). Orléans,France:Epi, 2021. 412 p. ISBN 978-2-7045-4521-6 DOI 10.51582/interconf.19-20.08.2021.

[3] Varun Kumar K S, Vijaya Kumar V G, Vijay Shankar A, Pratibha K ,"Credit Card Fraud Detection using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9 Issue 07, July-2020.

[4] FraudDetectionusing Machine Learning- Aditya Oza - aditya19@stanford.edu

[5] Adi Saputra , Suharjito , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10,No. 9,2019

Author Profile

Mr. Asher George Jacob, Status: Undergraduate student of Computer Science Engineering