**ABSTRACT**

It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. The project is mainly focussed on credit card fraud detection in real world. A phenomenal growth in the number of credit card transactions, has recently led to a considerable rise in fraudulent activities. The purpose is to obtain goods without paying, or to obtain unauthorized funds from an account. Implementation of efficient fraud detection systems has become imperative for all credit card issuing banks to minimize their losses. One of the most crucial challenges in making the business is that neither the card nor the cardholder needs to be present when the purchase is being made. This makes it impossible for the merchant to verify whether the customer making a purchase is the authentic cardholder or not. In this process, we have focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Random forest Classifier , Support Vector Classifier(SVC), K-Nearest Neighbor(KNN) and Decision Tree Algorithm on the PCA transformed Credit Card Transaction data..Finally optimize the accuracy of the result data. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision. Then processing of some of the attributes provided identifies the fraud detection and provides the graphical model visualization.

1. INTRODUCTION

A credit card is a payment card issued to the user as a system of payment. It permits the cardholder to acquire products and services based on the card holder’s purchase. A credit card is quite useful for day to day life. Fraud is as old as the human race. Fraud can be defined as a criminal deception with the intent of acquiring financial gain.'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future.In other words, Credit Card Fraud can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

1. LITERATURE REVIEW

Fraud is defined as an illegal deception intended to gain financial or personal gain. It's a planned conduct that goes against the law or a policy with the goal of gaining unjust financial gain. Data mining applications and adversarial detection are among the strategies used in this domain, according to a comprehensive survey undertaken by Clifton Phua and his colleagues. On a European dataset, classic methods such as Decision Tree, Support Vector, KNN, Random forest, and a mixture of particular classifiers were utilised, resulting in a recall of over 91 percent. Only after balancing the dataset by oversampling the data was high precision and recall achieved.

1. METHODOLOGY

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers.First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as V1-V28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one. We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it.

After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing any values in the dataset. This is done to ensure that we don’t require any missing value imputation and the machine learning algorithms can process the dataset smoothly.The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection. This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. We’ve used Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files.

The system architecture has following steps:

• Import of Necessary Packages

• Read the Dataset

• Exploratory Data Analysis i.e. finding null values, duplicate values etc.

• Selecting Features (X) and the Target (y) columns

• Train Test Split will split the whole dataset into train and test data

• Build the model i.e. Training the model

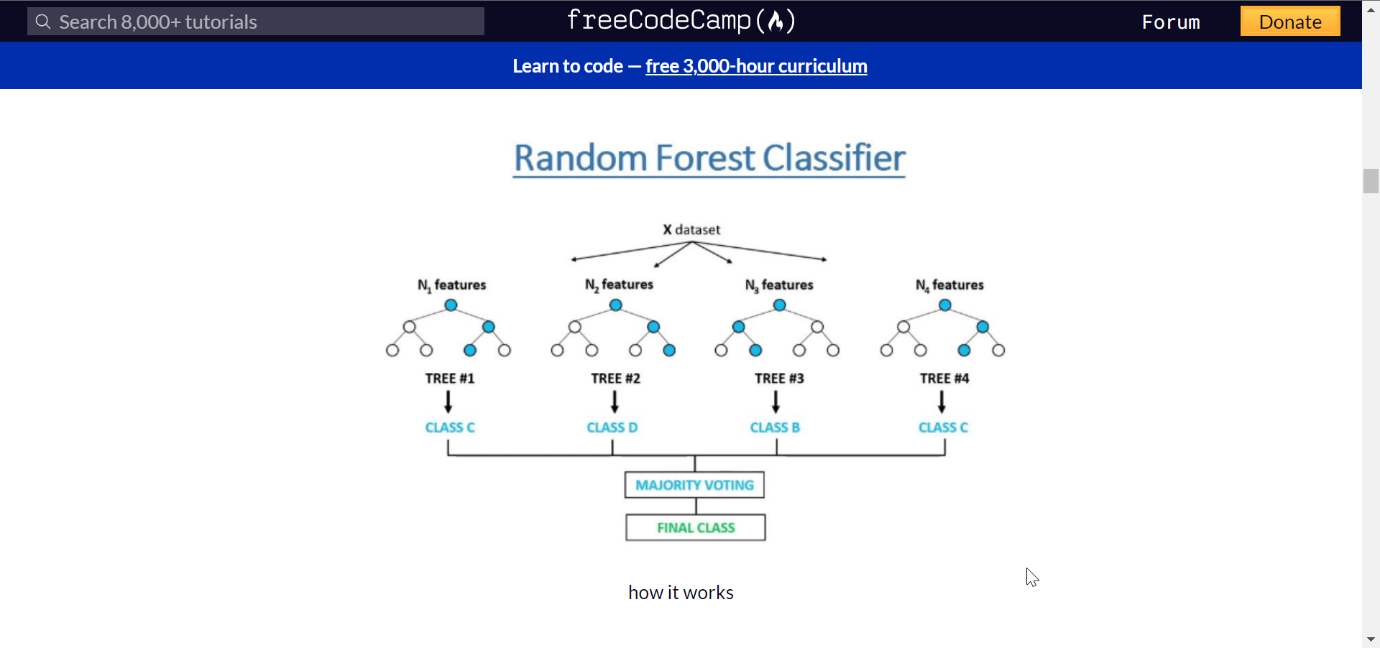
• Test the model i.e. Model prediction

• Evaluation of the system i.e. Accuracy score, F1- score etc.

Detailed explanations about the modules with for their algorithms and output graphs are given as follows:

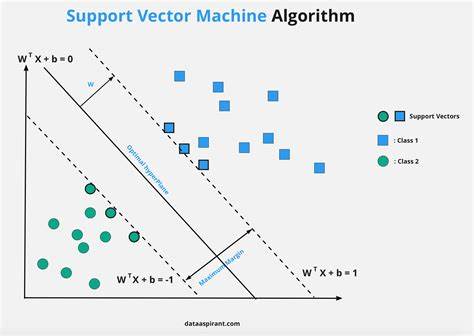
**Random Forest**

A random forest is a meta estimator that fits a number of 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 controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

****

**Support Vector Classifier**

Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.



**KNN Algorithm**

Various anomaly detection algorithms have exploited the concept of nearest neighbour analysis. Three primary elements influence the performance of the KNN algorithm:

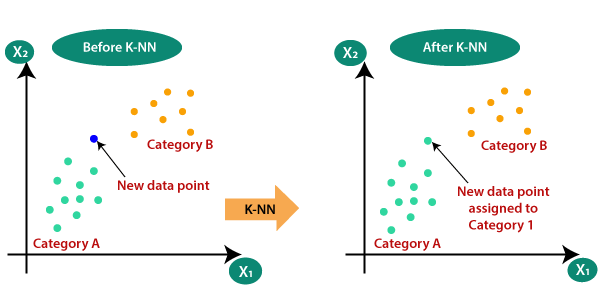
• The distance metric used to locate the nearest

neighbors.

• The distance rule that is used to classify k nearest

neighbours.

• The fresh sample was classified based on the number of neighbours it had.



**Decision Tree**

The training set is divided into nodes, each of which can contain all or most of one data category. Decision Tree is built by using recursive partitioning to classify the data. Firstly, an attribute is selected and its being the best attribute to split the data. It is split by minimizing the impurity at each step.

Impurity of a node is calculated by the entropy of data in the node. Entropy is a measure of uncertainty, in simple words, Entropy of the node is how much random data is in that node.

The lower the entropy the purer the node

**Root Node**: It depicts the maximum population of the dataset and this is then split into two or more

homogeneous groups.

**Splitting**: It is the splitting or distribution of a node

into two or more sub-nodes..

**Decision Node**: The decision node is defined as a sub-node that splits into other sub-nodes.

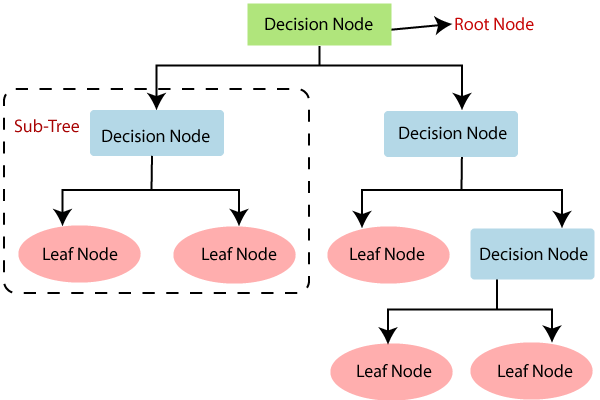
**Leaf/Terminal Node**: Leaf and Terminal nodes are

nodes that do not split.

**Pruning**: The process of eliminating sub-nodes from a decision node is referred to as pruning. Splitting is the polar opposite of pruning.

**Branch / Sub-Tree**: The term "branch" or "sub-tree" refers to a portion of the entire tree.

**Parent and Child Node**: A parent node is referred to as the parent node of sub-nodes, whilst sub-nodes are referred to as the child of a parent node.



1. IMPLEMENTATION  
     
   This idea is difficult to implement in real life because it requires the cooperation from banks, which aren’t willing to share information due to their market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results. As stated in one of these reference papers: “This technique was applied to a full application data set supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results obtained is presented below. After applying this technique, the level 1 list encompasses a few cases but with a high probability of being fraudsters. All individuals mentioned in this list had their cards closed to avoid any risk due to their high-risk profile.
2. EXPERIMENTAL RESULTS

To evaluate the results of the classification algorithms there are various parameter such as Accuracy score, classification report, F1-score, confusion matrix etc.

**Some important definitions**

**True positive(TP**)- It is an outcome in which the model accurately predicts the positive class.

**False positive(FP)-** It occurs when the positive class is predicted wrongly by the model.

**True negative(TN)-** It is an outcome in which the model accurately predicts the negative class.

**False negative(FN)-** It is an outcome in which the model predicts the negative class inaccurately.

•**Confusion Matrix -** It is a table that shows how well a classification model (or "classifier") performs on a set of test data for which the true values are known.

•**Accuracy-** The number of correct predictions divided by the total number of input samples is known as accuracy.

**Accuracy=TP+TN/TP+FN+FN+TN**

•**Precision (Specificity**)- It's the number of correct

positive outcomes divided by the classifier's projected number of positive finding.

**Precision=TP/TP+FP**

•**Recall (Sensitivity**) - It's calculated by dividing the number of correct positive results by the total number of relevant samples (all samples that should have been identified as positive).

**Recall=TP/TP+FN**

•**F1- score** - F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1].

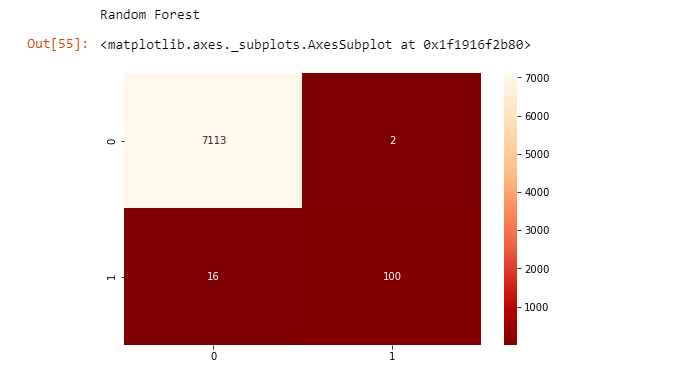
**F1 Score= 2\*(Recall\*Precision)/(Recall+**

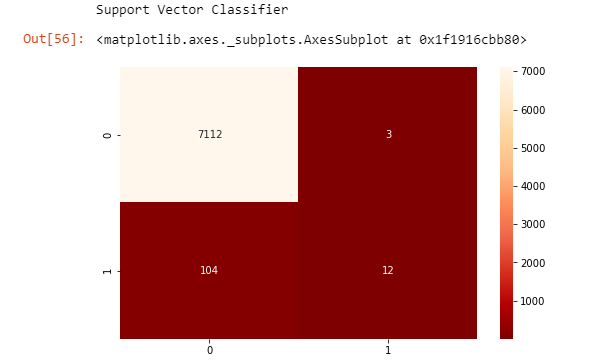
**Precision)**

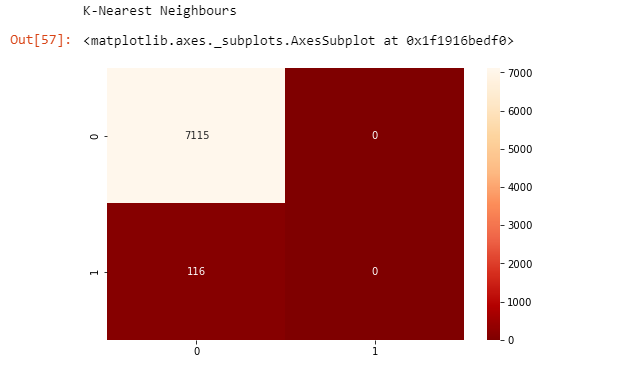
• **ROC-AUC Curve-** It is a performance metric for

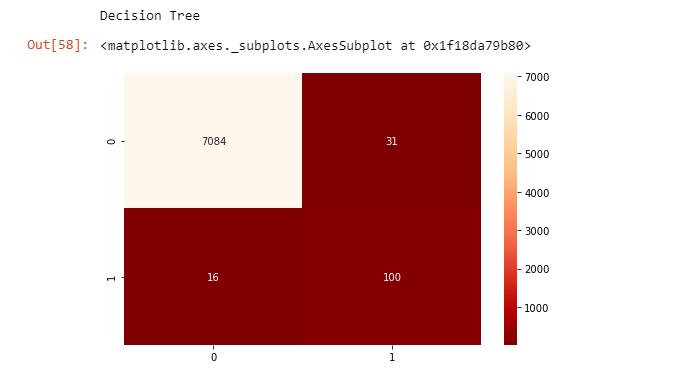
classifying issues at various thresholds. It's a

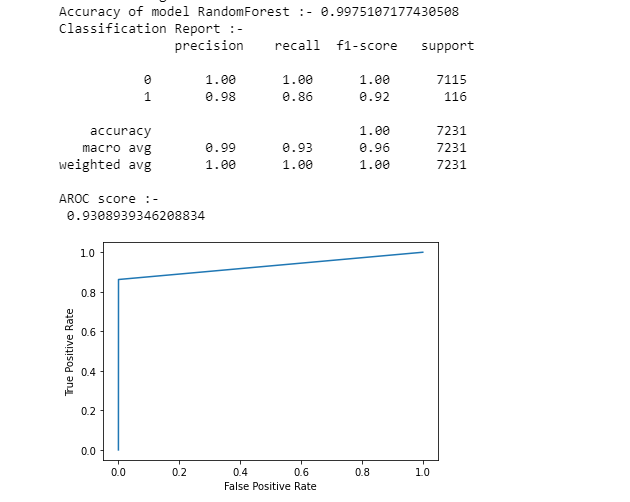
probability curve, and the AUC stands for the degree of separation. It expresses the model's capacity to distinguish across classes. The AUC indicates how well the model predicts 0s as 0s and 1s as 1s. TPR is plotted against FPR, with TPR on the y-axis and FPR on the x-axis.

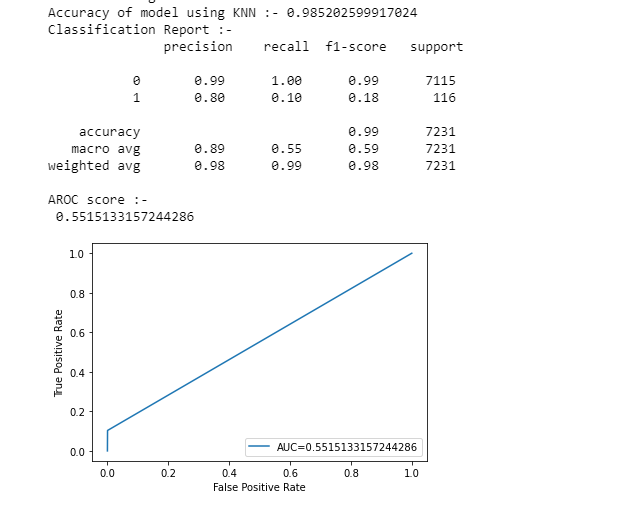


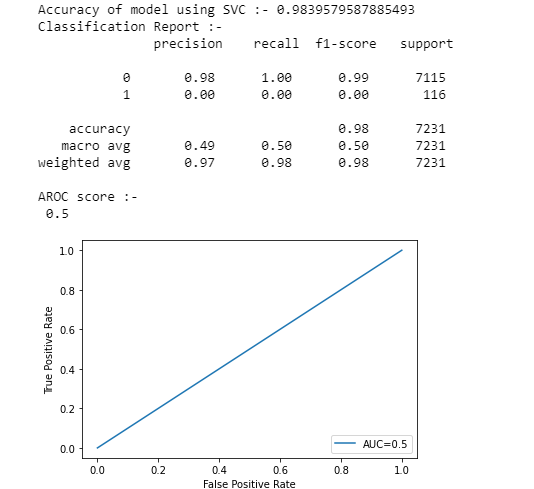


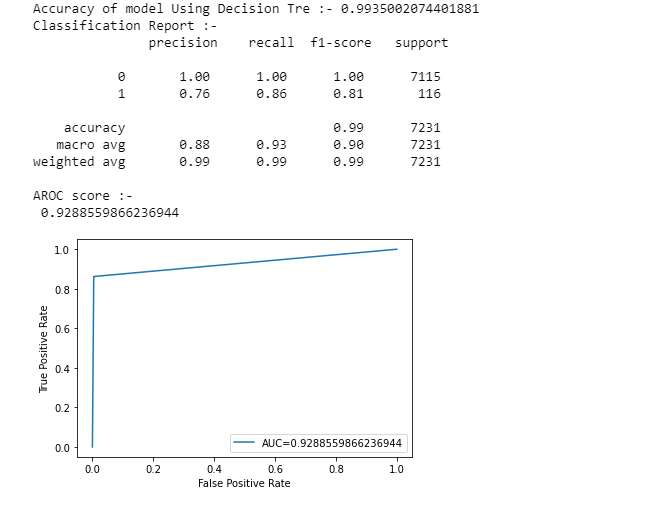




****

****

****

****

1. RESULT

* The code prints out the number of false positives it detected and compares it with the actual values.
* This is used to calculate the accuracy score and precision of the algorithms.
* The fraction of data we used for faster testing is 10% of the entire dataset.
* The complete dataset is also used at the end and both the results are printed.
* These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives.

1. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, explanation its implementation and experimentation results. While the Random Forest algorithm does reach over 99.75% accuracy, its precision remains at 98% when a tenth of the data set is taken into consideration. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

1. FUTURE WORK

While we couldn’t reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves

1. **REFERENCE**

[1] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na

Wang” published by 2009 International Joint Conference on Artificial Intelligence

[5] “Credit Card Fraud Detection through Parenclitic, Network Analysis By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008