Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 173 (2020) 104–112

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on *Smart Sustainable Intelligent Computing and Applications* under ICITETM2020

Credit Card Fraud Detection using Pipeling and Ensemble Learning

Siddhant Bagga[a], Anish Goyal[a], Namita Gupta[b], Arvind Goyal[c]

[a] *Information Technology, Netaji Subhas University of Technology, Delhi, India*
[b] *Computer Science, Maharaja Agrasen Institute of Technology, Delhi, India*
[c] *Information Technology Services, Engineers India Limited, Delhi, India*

Abstract

Financial fraud is a problem that has proved to be a menace and has a huge impact on the financial industry. Data mining is one of the techniques which has played an important role in credit card fraud detection in transactions which are online. Credit card fraud detection has proved to be a challenge mainly due to the 2 problems that it poses - both the profiles of fraudulent and normal behaviours change and data sets used are highly skewed. The performance of fraud detection is affected by the variables used and the technique used to detect fraud. This paper compares the performance of logistic regression, K-nearest neighbors, random forest, naive bayes, multilayer perceptron, ada boost, quadrant discriminative analysis, pipelining and ensemble learning on the credit card fraud data.

Keywords: Credit card fraud detection; classification; en-semble learning;

1. Main Text

Financial fraud is a problem that has proved to be a menace and has far reaching consequences in the financial industry. Fraud is defined as any deception that is criminal and is carried out with the motive of acquiring financial gains. With the advent of technology, credit card transaction have become mainstream. This has also resulted in increased fraud rate. Credit card fraud can be due to use of card by unauthorized cardholder using false identity taking bank's official in confidence, or it may also be due to use of stolen credit cards.

Data mining technique is one of the most used techniques for solving the problem of credit card fraud detection. Credit card fraud detection is the process of detecting whether a transaction is genuine or fraudulent[2]. A card's

* Corresponding author. Tel.: +919718153376
E-mail address: siddhantbagga1@gmail.com

10.1016/j.procs.2020.06.014

spending behavior is analyzed to to detect fraud cases. A number of techniques have been used for this purpose, which include ANN[3], SVM[4], decision tree[5] etc. This paper evaluates and compares the performance of 9 techniques - logistic regression, K-nearest neighbors, random forest, naive bayes, multilayer perceptron, ada boost, quadrant discriminative analysis, pipelining and ensemble learning on the credit card fraud data in an attempt to find which is the optimal technique to use to solve the problem.

The major challenges associated with this problem is that fraudulent transactions often look like legitimate ones and the credit card databases are not easily available. If dataset is available , it tends to be highly imbalanced. The performance of fraud detection technique is greatly dependent on the variables used, sampling approach and detection technique used.

This paper uses accuracy, precision, recall, F1 score and confusion matrix to compare the performance of 9 different techniques in an attempt to find the most suitable one.

The rest of the paper is organised as follows. Section 2 gives a detailed review of the previous work down in this field. Section 3 describes the experimental setup including information about the dataset that has been used and also describes briefly the various techniques which are used in this paper. Section 4 discusses the work we propose in this paper. Section 5 and 6 then compars the results that are obtained from the 9 different techniques. Section 7 concludes the study.

2. Related Work

Credit card fraud detection is a binary classification problem. Here, a transaction is either fraudulent or legitimate.

Fraudulent credit card transactions are characterised by out of usual phenomenon. The problem lies in the fact that both fraudulent and legitimate transactions share the same kind of profile. Therefore, both profiles remain dynamic. This leads to a decrease in the number of true positives in case of fraudulent transactions.

Credit card fraud detections is based on the analysis of spending behaviour of a cardholder. Variables are chosen optimally such that they capture the distinct behavior of the credit card. The selection of variables directly affects the performance of the credit card fraud detection system due to the dynamic nature of the profile. Past and current transaction of a credit card are the selection basis for these variables. There are 5 type of transaction variables[6].

Variables falling in all transaction types represent the normal card usage profile of a particular card. Spending habits of the card in accordance with geographical regions are depicted by regional statistics. The usage of a card in different categories of merchants is shown by merchant statistics. Time-based statistics type variables identify the usage profile of the cards with respect to time ranges. A person can have 2 or more credit cards each with different spending profile, so the study is not focused on personal profile, it is rather focused on card profile.

Due to the increase in the use of credit cards both as online and offline use of payments, fraud rates tend to increase. Detecting these fraudulent transactions has become a necessary task. Detecting these using traditional and manual methods is not only time consuming but is also inaccurate making it impractical to use. The advent of big data has now led the financial institutions to adopt intelligent techniques to solve the problem. There are two main categories of fraud detection: supervised and unsupervised. While the supervised technique uses fraudulent and legitimate samples to estimate the nature of a new transaction, unsupervised learning detects transactions as potential instances of fraudulent transactions. Many studies using different techniques have been carried out to solve this problem. Neural networks, Intelligent Decision Engines, Meta-learning agents, Bayesian network, Support Vector Machines, Adaptive Learning are some of these techiques.

3. Experimental Setup and Methods

3.1. Dataset

The source of the dataset is the ULB Machine Learning Group. It's description is found in [7] and consists of credit card transactions made by european cardholders occurring in 2 days in the month of September, 2013. There are a total of 284,807 anonymised transactions. The dataset contains a total of 0.172% of positive classes(fraud). This is an unbalanced dataset contains only numerical input variables which are the result of a PCA (Principal Component Analysis) transformation. The original features and and information are retracted due to the fact that it contained sensitive information. Features V1...V28 are the result of principal component analysis, 'Time' and 'Amount' being the only features on which PCA has not been applied. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Data is pre-processed for making the data normally distributed with zero mean and unit variance. Data is partitioned into training and test sets(33%).

For an example if there are 10 fraudulent transactions per 1,00,000 legitimate transactions, then even if the model predicts negative for all data, it will still be 99.999% accurate. Thus, the model will learn to predict all transactions as legitimate even if they are not. To combat this issue, the data must be balanced. In our model, data sampling is done using "ADASYN" [8] method to make the data balanced.

3.2. Methods used

3.2.1. Logistic Regression

Logistic regression [1] uses an approach which is functional to find the binary response probability based on a number of features. It uses the sigmoid function which is a nonlinear function to find the parameters which fit the best. The sigmoid function (sigma) along with the input (x) to the sigmoid function are given below:

$$\begin{aligned} \sigma(x) &= \frac{1}{(1 + l^{-x})} \\ x &= w_0 z_0 + w_1 z_1 + \cdots + w_n z_n \end{aligned} \tag{1}$$

The best coefficients w and input data which is a vector z are multiplied together multiply each element. These add up to get a number which finally determines a classification score of the target class. If the value of sigmoid comes out to be lesser than 0.5, then it is considered to be 0; otherwise its a 1.

3.2.2. Naive Bayes

Based on Bayesian theory, Naive Bayes [1] is a statistical approach which uses the highest probability to make decisions. Bayesian probability uses known values to estimates probabilities which are unknown. In this, logic and prior knowledge are applied to statements that are uncertain. This technique uses the conditional independence assumption among the features in the data. The conditional probabilities (3) and (4) of fraud and non fraud classes are used in the naive Bayes classifier.

$$\begin{aligned} P(c_i|f_k) &= \frac{P(f_k|c_i) * P(c_i)}{P(f_k)} \\ P(f_k|c_i) &= \textstyle\prod_{i=1}^{n} P(f_k|c_i),\ k = 1, 2, \ldots, n \end{aligned} \tag{2}$$

Where n is the maximum number of features, $P(f_k|c_i)$ is the probability of generating feature value f_k given class c_i, $P(c_i|f_k)$ is probability of feature value f_k being in class c_i, $P(c_i)$ and $P(f_k)$ are probability of occurrence of class c_i and probability of feature value f_k occurring respectively. The following classification rules are used by the classifier to perform binary classification.

The classification is C_1 if $P(c_1|f_k) > P(c_2|f_k)$

The classification is C_2 if $P(c_1|f_k) < P(c_2|f_k)$

3.2.3. Random Forest

Random forest is one of the supervised learning algorithms. It is used both for classification and regression. Random forest algorithm select random samples from the given dataset. It then constructs a decision tree for each and every sample and then gets a prediction from each of the decision tree. A vote is performed for each predicted result and the decision with the most votes is selected as the final prediction.

3.2.4. K-Nearest Neighbors

K-nearest neighbors classifier [1] uses similarity measures like Euclidean,Minkowski, or Manhattan distance and is a learning method which is instance based. Minkowski distance is suited well for categorical variables whereas Euclidean and Manhattan distance work well with variables which are continuous. In this paper we use Euclidean distance in the k nearest neighbors classifier. The Euclidean distance (D_{ij}) between two input vectors (X_i, X_j) is given as:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2} \text{ k=1,2,...,n} \qquad (3)$$

The Euclidean distance between current input and an input data point is calculated for each data point in the dataset. The calculated euclidean distances are arranged in increasing order and k items are selected which have the lowest distance to the input data. The classifier returns the majority class among these k data points as the classification for the input point.

3.2.5. Multi Layer Perceptron

Multi Layer Perceptron contains more than single neuron's linear layer. The simplest example can be of a 3 layer system which has the first layer as the input layer, last layer as the output layer and middle layer as the hidden layer. Input data is fed to the input layer and output data is collected from the output layer. Hidden layers can be increased as much as we want. Feed Forward network is most typical neural network. The ultimate goal is to approximate f() which is some function. For a classifier y = f(x) that gives an output value y for every input x, the multilayer perceptron finds a approximation to the classifier by defining a mapping and learning the best parameters for it. Many functions can be chained together in a MLP. Each layer in hidden layer performs a transformation of a linear sum of inputs which can be represented as y = f(WxT + b), where W represents the weights in the layer, x is a input vector which can also be the output of the previous layer, b is the bias vector and f is some activation function. Activation functions are functions which describe he relation between the input and the output in a non linear manner. The class score for each input is given by the output of the network. The performance is measured using a loss function. If the predicted and actual class does not correspond then the loss increases. To tackle the problem of overfitting and underfitting, an optimizer is used.

3.2.6. Ada Boost

Proposed by Yoav Freund and Robert Schapir, Adaptive Boosting is a type of ensemble boosting classifier. Multiple classifiers are combined in this method to get maximum accuracy. Ada Boost is an iterative method and combines multiple poorly performing classifiers to build up a strong classifier which gives high accuracy. The concept behind Ada Boost is to set the weights of each classifier and training the sample data in each iteration so as to ensure accurate predictions of unusual behavior. Any machine learning algorithm can be used as base classifier

3.2.7. Quadrant Discriminant Analysis

QDA assumes that the observations from each class of Y are from a gaussian distribution and that the covariance matrix is different for each class. The variance is not common for the predictor variables across the k levels in Y. It assumes that an observation fron the kth class will be of the form X = $N(\mu_k, \sum_k)$ where $\sum_k$ is the covariance matrix for class k. The classifier assigns each class an observation for which

$$\hat{\delta}_k(x) = -\frac{1}{2}x^T \textstyle\sum_k^{-1} x + x^T \sum_k^{-1} \hat{\mu_k} - \frac{1}{2}\hat{\mu_k}^T \sum_k^{-1} \hat{\mu_k} - \frac{1}{2} log|\sum_k| + log|\hat{\pi_k}|$$

is largest.

sklearn has been used for employing these classifiers. .fit and .predict functions were used. Default values as specifies in the sklearn library have been used for all parameters.

4. Proposed Work

For the purpose of increasing the accuracy of fraud detection in credit card transactions, the following two methods have been applied.

4.1. Pipelining

Pipelining refers to application of series of transformations followed by a final classifier. Pipeline is used for assembling of several multiple processes for the purpose of cross validating them together simultaneously setting up of different parameters. In this work, 'selectKBest' (from sklearn) has been used which allows for the selection of features based on the k top scores. f-regression (sklearn) has been used for the purpose of carrying out feature selection. It is based on linear tests (regression) which are uni-variate. It is done for the purpose of finding the influence of each regressor. It's actually a scoring function. Finally, Random Forest Classifier has been used for the purpose of classification and prediction.

4.2. Ensemble Learning - Bagging Classifier

Ensemble methods refer to the combination of various estimator (base) which are developed with a specific learning method for the purpose of improving on the one individual single estimator. In this work, bagging classifier has been used in which fitting of the base classifiers is carried out on every randomized subset of the actual dataset. Then aggregation of each of the individual predictions is carried out. Aggregation can be carried out by two ways viz. voting or by taking the average and then a final prediction can be made. In order to decrease the variance of some kind of blackbox classifier (like decision tree), bagging classifier can be used because of the introduction of randomization in the development process and finally developing the ensemble out of it. 'Bagging' refers to the fact that the random subsets are taken out with replacement. In this work, Random Forest Classifier has been used as the base classifier. 15 base estimators have been taken.

5. Performance Evaluation

The experiments are evaluated using 4 basic metrics - True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP). The performance of the 7 methods is compared based on their accuracy, precision, f1 score and recall.

True positives are the cases which are actually positive and are also classified as positive. Similarly True negatives are the cases which are actually negative and are classified as negative. False positives are the cases which are actually negative but are classified as positives. Similarly False negatives are the cases which are actually positive but are classified as negative.

$$\begin{aligned}
&Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \\
&Precision = \frac{TP}{TP + FP} \\
&Recall = \frac{TP}{TP + FN} \\
&F1Score = \frac{2 * Precision * Recall}{Precision + Recall}
\end{aligned} \tag{4}$$

Accuracy is the measure of number of correct predictions divided by the total number of predicitons. Precision is the ratio of positive predictions to the total number of positive classes predicted. Recall is the ratio of positive predictions to the number of positive class values in the test data. F1 score depicts the balance between precision and recall.

MCC (Matthews Correlation Coefficient) is a balanced measure which uses TP, FP, TN, FN to measure the performance of a binary classifier if the classes have sizes very different from each other. MCC has values between -1 and 1. -1 value indicates a classifier which is completely wrong while 1 indicated a perfectly correct classifier.

The MCC formula is:

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (FN + TN) * (FP + TN) * (TP + FN)}} \tag{5}$$

BCR (Balanced classification rate) is another metric used for imbalanced datasets. It combines the specificity and sensitivity metrics as follows:

$$BCR = \frac{1}{2}(TPR + TNR)$$
$$BCR = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{6}$$

6. Results

In this study 9 classifier models are used on the dataset. 70% of the data is used to train the model while 30% is used for testing. Accuracy, precision, recall and F1 score is used to compare the classifier models. Class 0 represents a Non-Fraudulent transaction while Class 1 represents a fraudulent transaction.

Table 1. Accuracy result

Classifier	**Accuracy**
Logistic Regression	98.2%
Naive Bayes	99.6%
K-Nearest Neighbors	94.4%
MultiLayer Perceptron	98.4%
Ada Boost	98.5%
Quadrant Discriminant Analysis	97.3%
Random Forest	99.7%
Ensemble Learning	99.99%
Pipelining	99.9999%

Table 2. Precision result

Classifier	**Precision**	
	Class 0	**Class 1**
Logistic Regression	1.00	0.07
Naive Bayes	1.00	0.26
K-Nearest Neighbors	1.00	0.02
MultiLayer Perceptron	1.00	0.08
Ada Boost	1.00	0.09
Quadrant Discriminant Analysis	1.00	0.05
Random Forest	1.00	0.31
Ensemble Learning	1.00	0.76
Pipelining	1.00	0.84

As can be seen in tables 1, 2, 3, 4, the proposed method Ensemble Learning has an accuracy of 99.99%. Accuracy is the most important metric while comparing the performance of different classifier models. Ensemble learning also

Table 3. Recall result

Classifier	**Recall**	
	Class 0	**Class 1**
Logistic Regression	0.98	0.91
Naive Bayes	1.00	0.85
K-Nearest Neighbors	0.94	0.55
MultiLayer Perceptron	0.98	0.90
Ada Boost	0.99	0.91
Quadrant Discriminant Analysis	0.97	0.91
Random Forest	1.00	0.89
Ensemble Learning	1.00	0.87
Pipelining	1.00	0.86

Table 4. F1 Score

Classifier	**F1 Score**	
	Class 0	**Class 1**
Logistic Regression	0.99	0.14
Naive Bayes	1.00	0.40
K-Nearest Neighbors	0.97	0.03
MultiLayer Perceptron	0.99	0.15
Ada Boost	0.99	0.16
Quadrant Discriminant Analysis	0.99	0.10
Random Forest	1.00	0.46
Ensemble Learning	1.00	0.81
Pipelining	1.00	0.85

has precision and F1 score much greater than all the other classifiers. The precision for Class 1 is much better in the proposed models as compared to the others. But these metrics do not prove that our proposed model is better than others because of the fact that there is a huge difference in the sizes of input classes. To compare the performance of all the models we use 2 much better and balanced metric: MCC and BCR. The metric values of the models are show in the following graphs.

A MCR or BCR value close to 1 signifies an almost perfect model whereas values close to -1 signify models which are almost completely wrong. From the graphs we can see that the proposed models - Ensemble Learning and Pipelining performed significantly better than all the other models while K-Nearest neighbors model performed the worst.

7. Conclusion

This paper successfully investigates the performance of Logistic Regression, Naïve Bayes, K nearest neighbours, Multi Layer Perceptron, Ada Boost, Quadrant Discriminant Analysis, Random Forests, Pipelining and Ensemble Learning in determining fraudulent credit card transactions.

1. 9 different classifier models are trained on real life dataset and their performances are evaluated based on various parameters and metrics.
2. The dataset is highly imbalanced. ADASYN method is used to make this dataset balanced.
3. The performance of the classifiers are examined using precision, accuracy, recall, F1 score, Matthews correlation coefficient and Balanced Classification Rate.

Based on different metrics, the performance of Pipelining method was found out to be the best.

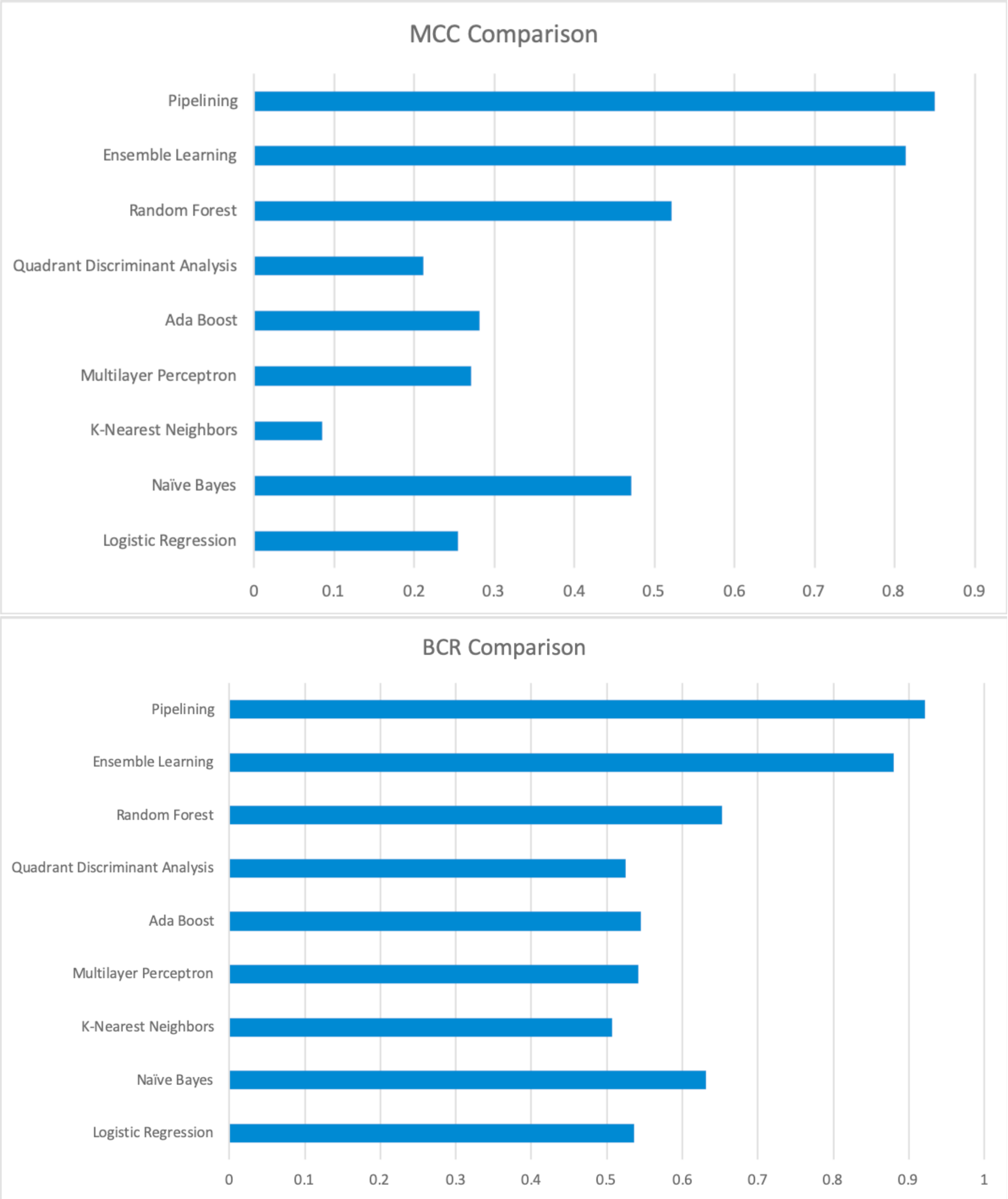

Fig. 1. MCC and BCR Comparison

References

[1] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, ”Credit card fraud detection using machine learning techniques: A comparative analysis,” 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-9. doi: 10.1109/ICCNI.2017.8123782

[2] Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., (2002). Credit card fraud detection using Bayesian and neural networks. Proceeding International NAISO Congress on Neuro Fuzzy Technologies.

[3] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311 – 322

[4] Singh, G., Gupta, R., Rastogi, A., Chandel, M. D. S., and Riyaz, A., (2012). A Machine Learning Approach for Detection of Fraud based on SVM, International Journal of Scientific Engineering and Technology, Volume No.1, Issue No.3, pp. 194-198, ISSN : 2277-1581

[5] Patil, S., Somavanshi, H., Gaikwad, J., Deshmane, A., and Badgujar, R., (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4, Issue 4, pp. 92-95, ISSN: 2320-088X

[6] Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. In Machine Learning and Applications (ICMLA), 2013. IEEE.

[7] Pozzolo, A. D., Caelen, O., Johnson, R. A., and Bontempi, G., (2015). Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE.

[8] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, ”ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328. doi: 10.1109/IJCNN.2008.4633969