Credit card fraud detection using Machine Learning Techniques:

A Comparative Analysis

John O. Awoyemi
Department of Computer Science
Federal University of Technology
Akure
Akure, Nigeria
johntobaonline@yahoo.com

Adebayo O. Adetunmbi
Department of Computer Science
Federal University of Technology
Akure
Akure, Nigeria
aoadetunmbi@futa.edu.ng

Samuel A. Oluwadare
Department of Computer Science
Federal University of Technology
Akure
Akure, Nigeria
saoluwadare@futa.edu.ng

***Abstract*—Financial fraud is an ever growing menace with far consequences in the financial industry. Data mining had played an imperative role in the detection of credit card fraud in online transactions. Credit card fraud detection, which is a data mining problem, becomes challenging due to two major reasons – first, the profiles of normal and fraudulent behaviours change constantly and secondly, credit card fraud data sets are highly skewed. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection technique(s) used. This paper investigates the performance of naïve bayes, k-nearest neighbor and logistic regression on highly skewed credit card fraud data. Dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions. A hybrid technique of under-sampling and oversampling is carried out on the skewed data. The three techniques are applied on the raw and preprocessed data. The work is implemented in Python. The performance of the techniques is evaluated based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and balanced classification rate. The results shows of optimal accuracy for naïve bayes, k-nearest neighbor and logistic regression classifiers are 97.92%, 97.69% and 54.86% respectively. The comparative results show that k-nearest neighbour performs better than naïve bayes and logistic regression techniques.**

Keywords—credit card fraud; data mining; naïve bayes; decision tree; logistic regression, comparative analysis

I. INTRODUCTION

Financial fraud is an ever growing menace with far reaching consequences in the finance industry, corporate organizations, and government. Fraud can be defined as criminal deception with intent of acquiring financial gain. High dependence on internet technology has enjoyed increased credit card transactions. As credit card transactions become the most prevailing mode of payment for both online and offline transaction, credit card fraud rate also accelerates. Credit card fraud can come in either inner card fraud or external card fraud. Inner card fraud occurs as a result of consent between cardholders and bank by using false identity to commit fraud while the external card fraud involves the use of stolen credit card to get cash through dubious means. A lot of researches have been devoted to detection of external card fraud which accounts for majority of credit card frauds. Detecting fraudulent transactions using traditional methods of manual detection is time consuming and inefficient, thus the advent of big data has made manual methods more impractical. However, financial institutions have focused attention to recent computational methodologies to handle credit card fraud problem.

Data mining technique is one notable methods used in solving credit fraud detection problem. Credit card fraud detection is the process of identifying those transactions that are fraudulent into two classes of legitimate (genuine) and fraudulent transactions [1]. Credit card fraud detection is based on analysis of a card's spending behaviour. Many techniques have been applied to credit card fraud detection, artificial neural network [2], genetic algorithm [3, 4], support vector machine [5], frequent itemset mining [6], decision tree [7], migrating birds optimization algorithm [8], naïve bayes [9]. A comparative analysis of logistic regression and naive bayes is carried out in [10]. The performance of bayesian and neural network [11] is evaluated on credit card fraud data. Decision tree, neural networks and logistic regression are tested for their applicability in fraud detections [12]. This paper [13] evaluates two advanced data mining approaches, support vector machines and random forests, together with logistic regression, as part of an attempt to better detect credit card fraud while neural network and logistic regression is applied on credit card fraud detection problem [14]. A number of challenges are associated with credit card detection, namely fraudulent behaviour profile are dynamic, that is fraudulent transactions tend to look like legitimate ones; credit card transaction datasets are rarely available and highly imbalanced (or skewed); optimal feature (variables) selection for the models; suitable metric to evaluate performance of techniques on skewed credit card fraud data. Credit card fraud detection performance is greatly affected by type of sampling approach used, selection of variables and detection technique(s) used. This study investigates the effect of hybrid sampling on performance of fraud detection of naïve bayes, k-nearest

978-1-5090-4642-3/17/$31.00 ©2017 IEEE

neighbour and logistic regression classifiers on highly skewed credit card fraud data.

This paper seeks to carry out comparative analysis of credit card fraud detection using naive bayes, k-nearest neighbor and logistic regression techniques on highly skewed data based on accuracy, sensitivity, specificity and Matthews's correlation coefficient (MCC) metrics. This paper extends the handling of highly imbalanced credit card fraud data in [33]. The imbalanced dataset used in this study which contains about 0.172% of fraud transactions is sampled in a hybrid approach. The positive class (fraud) is oversampled while the negative class (legitimate) is under-sampled by the same number of times to achieve two distributions of 34:66 and 10:90. The three techniques are applied to the data. The performance comparison of the three techniques is analyzed based on accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC) and balanced classification rate.

The rest of this paper is organized as follows: Section II gives detailed review on credit card fraud, feature selection detection techniques and performance comparison. Section III describes the experimental setup approach including the data pre-processing and the three classifier methods on credit card fraud detection. Section IV reports the experimental results and discussion about the comparative analysis. Section V concludes the comparative study and suggests future areas of research.

II. Related Works

Classification of credit card transactions is mostly a binary classification problem. Here, credit card transaction is either as a legitimate transaction (negative class) or a fraudulent transaction (positive class). Fraud detection is generally viewed as a data mining classification problem, where the objective is to correctly classify the credit card transactions as legitimate or fraudulent [6].

A. Credit Card Fraud

Credit card frauds have been partitioned into two types: inner card fraud and external fraud [12, 15] while a broader classification have been done in three categories, that is, traditional card related frauds (application, stolen, account takeover, fake and counterfeit), merchant related frauds (merchant collusion and triangulation) and Internet frauds (site cloning, credit card generators and false merchant sites) [16]. It is reported in [17] that the total amount of fraud losses of banks and businesses around the world reached more than USD 16 billion in 2014 with an increase of nearly USD 2.5 billion in the previous year recorded losses, meaning that, each USD 100 is having 5.6 cents that was fraudulent, the report concluded.

Credit card transactions data are mainly characterized by an unusual phenomenon. Both legitimate transactions and fraudulent ones tend to share the same profile. Fraudsters learn new ways to mimic the spending behaviour of legitimate card (or cardholder). Thus, the profiles of normal and fraudulent behaviours are constantly dynamic. This inherent characteristic leads to a decrease in the number of true fraudulent cases identified in a pool of credit card transactions data leading to a highly skewed distribution towards the negative class (legitimate transactions). The credit card data investigated in [18] contains 20% of the positive cases, 0.025% positive cases [19] and below 0.005% positive cases [8]. The data used in this study has positive class (frauds) accounting for 0.172% of all transactions. A number of sampling approaches have been applied to the highly skewed credit card transactions data. A random sampling approach is used in [18, 20] and reports experimental results indicating that 50:50 artificially distribution of fraud/non-fraud training data generate classifiers with the highest true positive rate and low false positive rate. The paper [8] uses stratified sampling to under sample the legitimate records to a meaningful number. It experiment on 50:50, 10:90 and 1:99 distributions of fraud to legitimate cases reports that 10:90 distribution has the best performance (regarding the performance comparisons on the 1:99 set) as it is closest to the real distribution of frauds and legitimates. Stratified sampling is also applied in [21]. In this study, a hybrid of under-sampling the negative cases and oversampling the positive cases is carried in order to preserve valuable patterns from the data.

B. Feature (Variables) selection

The basis of credit card fraud detection lies in the analysis of cardholder's spending behaviour. This spending profile is analysed using optimal selection of variables that capture the unique behaviour of a credit card. The profile of both a legitimate and fraudulent transaction tends to be constantly changing. Thus, optimal selection of variables that greatly differentiates both profiles is needed to achieve efficient classification of credit card transaction. The variables that form the card usage profile and techniques used affect the performance of credit card fraud detection systems. These variables are derived from a combination of transaction and past transaction history of a credit card. These variables fall under five main variable types, namely all transactions statistics, regional statistics, merchant type statistics, time-based amount statistics and time-based number of transactions statistics [19].

The variables that fall under all transactions statistics type depict the general card usage profile of the card. The variables under regional statistics type show the spending habits of the card with taken into account the geographical regions. The variables under merchant statistics type show the usage of the card in different merchant categories. The variables of time-based statistics types identify the usage profile of the cards with respect to usage amounts versus time ranges or frequencies of usage versus time ranges. Most literature focused on cardholder profile rather than card profile. It is evident that a person can operate two or more credit cards for different purposes. Therefore, one can exhibit different spending profile on such cards. In this study, focus is beamed on card rather than cardholder because one credit card can only exhibit a unique spending profile while a cardholder can exhibit multiple behaviours on different cards. A total of 30 variables are used in [18], 27 variables in [19] and 20 variables are reduced to 16 relevant ones [6].

C. Credit card Fraud Detection

As credit card becomes the most general mode of payment (both online and regular purchase), fraud rate tends to accelerate. Detecting fraudulent transactions using traditional methods of manual detection are time consuming and inaccurate, thus the advent of big data had made these manual methods more impractical. However, financial institutions have turned to intelligent techniques. These intelligent fraud techniques comprise of computational intelligence (CI)-based techniques. Statistical fraud detection methods have been divided into two broad categories: supervised and unsupervised [22]. In supervised fraud detection methods [13], models are estimated based on the samples of fraudulent and legitimate transactions to classify new transactions as fraudulent or legitimate while in unsupervised fraud detection, outliers' transactions are detected as potential instances of fraudulent transactions. A detailed discussion of supervised and unsupervised techniques is found in [23]. Quite a number of studies on a range of techniques have been carried out in solving credit card fraud detection problem. These techniques include but not limited to; neural network models (NN), Bayesian network (BN), intelligent decision engines (IDE), expert systems, meta-learning agents, machine learning, pattern recognition, rule-based systems, logic regression (LR), support vector machine (SVM), decision tree, k-nearest neighbor (kNN), meta learning strategy, adaptive learning etc. Some related works on comparative study of credit card fraud detection techniques are presented.

D. Comparative study

A study of the issues and results associated with credit card fraud detection using meta-learning is presented [18]. This study is geared towards investigating distribution of frauds and non-frauds that will lead to better performance, best learning algorithms between meta-learning strategy. The results show that given a skewed distribution in the original data, artificially more balanced training data leads to better classifiers. It demonstrate how meta-learning can be used to combine different classifiers and maintain, and in some cases, improve the performance of the best classifier. Multiple algorithms for fraud detection are investigated in [24] and results indicate that an adaptive solution can provide fraud filtering and case ordering functions for reducing the number of final-line fraud investigations necessary. A comparison of logistic regression and naive bayes is presented in [10]. The results of the analysis shows that even though the discriminative logistic regression algorithm has a lower asymptotic error, the generative naive Bayes classifier may also converge more quickly to its (higher) asymptotic error. There are a few cases reported in which logistic regression's performance underperformed that of naive Bayes, but this is observed primarily in particularly small datasets. Another comparative study on credit card fraud detection using Bayesian and neural networks is done [11]. The results report that Bayesian network performs better than neural network in detecting credit card fraud.

Back-propagation (BP), together with naive Bayesian (NB) and C4.5 algorithms are applied to skewed data partitions derived from minority oversampling with replacement [25]. The study shows that innovative use of naive Bayesian (NB), C4.5, and back-propagation (BP) classifiers to process the same partitioned numerical data has the potential of getting better cost savings. An adaptive and robust model learning method that is highly adaptive to concept changes and is robust to noise is presented [26]. The classifiers' weights are computed by logistic regression technique, which ensures good adaptability. Three different classification methods, decision tree, neural networks and logistic regression are tested for their applicability in fraud detections [12]. The results show that the proposed classifier of neural networks and logistic regression approaches outperform decision tree in solving the problem under investigation. A fusion approach using Dempster–Shafer theory and Bayesian learning for detecting credit card fraud is proposed [27].The results also show that use of Bayesian learning however, brings down the false positive rates to values close to 5%.

Detection of credit card fraud using decision trees and support vector machines is investigated [28] and the results show that the proposed classifiers of decision tree approaches outperform SVM approaches in solving the problem under investigation. As the training data scales, SVM based model detection accuracy equal that of the decision tree based models, but fall short in the number of frauds detected. This paper [13] evaluates the performance of logistic regression alongside two advanced data mining approaches, support vector machines and random forests in credit card fraud detection. The study shows that logistic regression maintained similar performance with different levels of under-sampling, while SVM performance tend to increase with lower proportion of fraud in the training data. Logistic regression shows appreciable performance, often surpassing that of the SVM models with different kernels. In another study, classification models based on Artificial Neural Networks (ANN) and Logistic Regression (LR) are developed and applied on credit card fraud detection problem [14] using a highly skewed data. The results show that the proposed ANN classifiers outperform LR classifiers in solving the problem under investigation. The logistic regression classifiers tend to over fit the training data as it increases. This is due to lack of adequate sampling in the work. A comparative assessment of supervised data mining techniques for fraud prevention is presented in [29]. The techniques evaluated are decision tree, neural network and naive bayes classifiers. It is reported that neural network classifiers are suitable for larger databases only and take long time to train the model. Bayesian classifiers are more accurate and much faster to train and suitable for different sizes of data but are slower when applied to new instances.

A meta-classification strategy is applied in improving credit card fraud detection [30]. The approach consists of 3 base classifiers constructed using the decision tree, naïve Bayesian, and k-nearest neighbour algorithms. Using the naïve Bayesian algorithm as the meta-level algorithm to combine the

base classifier predictions, the result shows 28% improvement in performance. This paper [31] put a light on performance evaluation based on the correct and incorrect instances of data classification using Naïve Bayes and decision tree. The results show that the efficiency and accuracy of J48 is better than that of Naïve Bayes [31]. In this paper [19], new comparison measure that realistically represents the monetary gains and losses due to fraud detection shows that including the real cost by creating a cost sensitive system using a Bayes minimum risk classifier, gives rise to much better fraud detection results in the sense of higher savings.

III. Experimental Set up and Methods

This section describes the dataset used in the experiments and the three classifiers under study, namely; Naïve Bayes, k-Nearest Neighbour and Logistic Regression techniques. The different stages involved in generating the classifiers include; collection of data, preprocessing of data, analysis of data, training of the classifier algorithm and testing (evaluation). During the preprocessing stage, the data is converted into useable format fit and sampled. A hybrid of under-sampling (the negative cases) and over-sampling (the positive cases) is carried out to achieve two sets of data distributions. For the analysis stage, the feature selection and reduction is already carried out on the dataset using PCA. The training stage is where the classifier algorithms are developed and fed with the processed data. The experiments are evaluated using True positive, True Negative, False Positive and False Negative rates metric. The performance comparison of the classifiers is analyzed based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and balanced classification rate.

A. Dataset

The dataset is sourced from ULB Machine Learning Group and description is found in [32]. The dataset contains credit card transactions made by European cardholders in September 2013. This dataset presents transactions that occurred in two days, consisting of 284,807 transactions. The positive class (fraud cases) make up 0.172% of the transactions data. The dataset is highly unbalanced and skewed towards the positive class. It contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. Thus a total of 30 input features are utilized in this study. The details and background information of the features cannot be presented due to confidentiality issues. The time feature contains the seconds elapsed between each transaction and the first transaction in the dataset. The 'amount' feature is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (non fraud).

B. Hybrid Sampling of dataset

Data pre-processing is carried out on the data. A hybrid of under-sampling and over-sampling is carried out on the highly unbalanced dataset to achieve two sets of distribution (10:90 and 34:64) for analysis. This is done by stepwise addition and subtraction of a data point interpolated between existing data points till over-fitting threshold is reached.

$$PC_{new} = \sum_{i=1}^{n} PC + i \tag{1}$$

$$NC_{new} = \sum_{i=1}^{n} NC - i \tag{2}$$

$$n = \mathrm{mod}\left(\left(NC/PC\right)/2\right) \tag{3}$$

where PC_{new} is the new number of positive data point instances, NC_{new} is the new number of negative data points, n is the modulus of the ratio (NC/PC) of number of negative class to positive class, PC and NC is the number of positive and negative class data points in imbalanced dataset respectively.

C. Naïve Bayes Classifier

Naïve Bayes a statistical approach based on Bayesian theory, which chooses the decision based on highest probability. Bayesian probability estimates unknown probabilities from known values. It also allows prior knowledge and logic to be applied to uncertain statements. This technique has an assumption of conditional independence among features in the data. The Naïve Bayes classifier is based on the conditional probabilities (4) and (5) of the binary classes (fraud and non fraud).

$$P\left(c_i \mid f_k\right) = \frac{P\left(f_k \mid c_i\right) * P\left(c_i\right)}{P\left(f_k\right)} \tag{4}$$

$$P\left(f_k \mid \mathrm{c_i}\right) = \prod_{i=1}^{n} P\left(f_k \middle| c_i\right) \; k = 1,\ldots,n; i = 1,2 \tag{5}$$

where n represents maximum number of features (30), $P(c_i|f_k)$ is probability of feature value f_k being in class c_i, $P(f_k|c_i)$ is probability of generating feature value f_k given 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 classifier performs the binary classification based on Bayesian classification rule.

If $P(c_1|f_k) > P(c_2|f_k)$ *then the classification is* C_1

If $P(c_1|f_k) < P(c_2|f_k)$ *then the classification is* C_2

C_i is the target class for classification where C_1 is the negative class (non fraud cases) and C_2 is the positive class (fraud cases).

D. K-Nearest Neighbour Classifier

The k-nearest neighbour is an instance based learning which carries out its classification based on a similarity measure, like Euclidean, Mahanttan or Minkowski distance functions. The first two distance measures work well with continuous variables while the third suits categorical variables. The Euclidean distance measure is used in this study for the kNN classifier. The Euclidean distance (D_{ij}) between two input vectors (Xi, Xj) is given by:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk} \right)^2} \qquad \text{k=1,2,…,n} \tag{6}$$

For every data point in the dataset, the Euclidean distance between an input data point and current point is calculated. These distances are sorted in increasing order and k items with lowest distances to the input data point are selected. The majority class among these items is found and the classifier returns the majority class as the classification for the input point. Parameter tuning for k is carried out for k = 1, 3, 5, 7, 9, 11 and k = 3 showed optimal performance. Thus, value of k = 3 is used in the classifier.

E. Logistic Regression Classifier

Logistic Regression which uses a functional approach to estimate the probability of a binary response based on one or more variables (features). It finds the best-fit parameters to a nonlinear function called the sigmoid. The sigmoid function (σ) and the input (x) to the sigmoid function are shown in (7) and (8).

$$\sigma(x) = \frac{1}{\left(1 + \ell^{-x}\right)} \tag{7}$$

$$x = w_0 z_0 + w_1 z_1 + \ldots + w_n z_n \tag{8}$$

The vector z is input data and the best coefficients w, is multiplied together multiply each element and adds up to get one number which determines the classifier classification of the target class. If the value of the sigmoid is more than 0.5, it's considered a 1; otherwise, it's a 0. An optimization method is used to train the classifier and find the best-fit parameters. The gradient ascent (9) and modified stochastic gradient ascent optimization methods were experimented on to evaluate their performance on the classifier.

$$w := w + \alpha \nabla_w f(w) \tag{9}$$

where the parameter ∇ is the magnitude of movement of the gradient ascent. The steps are continued until a stopping criterion is met. The optimization methods are investigated (for iterations 50 to 1000) to know if the parameters are converging. That is, are the parameters reaching a steady value, or are they constantly changing. At 100 iterations, steady values of parameters are achieved.

Stochastic gradient ascent incrementally updates the classifier as new data comes in rather than all at once. It starts with all weights set to 1. Then for every feature value in the dataset, the gradient ascent is calculated. The weights vector is updated by the product of alpha and gradient. Then weight vector is returned. The stochastic gradient ascent is used in this study because given the large size of data it updates the weights using only one instance at a time, thus reducing computational complexity.

IV. Performance Evaluation And Results

Four basic metrics are used in evaluating the experiments, namely True positive (TPR), True Negative (TNR), False Positive (FPR) and False Negative (FNR) rates metric respectively.

$$TPR = \frac{TP}{P} \tag{10}$$

$$TNR = \frac{TN}{N} \tag{11}$$

$$FPR = \frac{FP}{N} \tag{12}$$

$$FNR = \frac{FN}{P} \tag{13}$$

where TP, TN, FP and FN are the number of true positive, true negative, false positive and false negative test cases classified while P and N are the total number of positive and negative class cases under test. True positives are cases classified as positive which are actually positive. True negative are cases classified rightly as negative. False positive are cases classified as positive but are negative cases. False negative are cases classified as negative but are truly positive.

The performance of naïve bayes, k-nearest neighbour and logistic regression classifiers are evaluated based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient (MCC) and balanced classification rate. These evaluation metrics are implored based on their relevance in evaluating imbalanced binary classification problem.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{15}$$

$$Specificity = \frac{TN}{FP + TN} \tag{16}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{17}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{18}$$

$$BCR = \tfrac{1}{2} * \left(\frac{TP}{P} + \frac{TN}{N} \right) \tag{19}$$

Sensitivity (Recall) gives the accuracy on positive (fraud) cases classification. Specificity gives the accuracy on negative (legitimate) cases classification. Precision gives the accuracy in cases classified as fraud (positive). Matthews Correlation Coefficient (MCC) is an evaluation metric for binary classification problems. MCC is used mainly with unbalanced data sets because its evaluation consists of TP, FP, TN and FN. The MCC value is usually between -1 and +1; a +1 value represents excellent classification while a -1 value represents total distinction between classification and observation. Balanced classification rate represents the average of sensitivity and specificity which is the portion of negatives which are classified as negatives [33].

A. Results

In this study, three classifier models based on naive bayes, k-nearest neighbour and logistic regression are developed. To evaluate these models, 70% of the dataset is used for training while 30% is set aside for validating and testing. Accuracy, sensitivity, specificity, precision, Matthews correlation coefficient (MCC) and balanced classification rate are used to evaluate the performance of the three classifiers. The accuracy of the classifiers for the original 0.172:99.828 dataset distribution, the sampled 10:90 and 34:66 distributions are presented in Tables 1, 2 and 3 respectively.

An observation of the metric tables shows that there is significant improvement from the sampled dataset distribution of 10:90 to 34:66 for accuracy, sensitivity, specificity, Matthews correlation coefficient and balanced classification rate of the classifiers. This shows that a hybrid sampling (under-sampling and over-sampling) on a highly imbalanced dataset greatly improves the performance of binary classification. The true positive, true negative, false positive and false negative rates of the classifiers in each set of un-sampled and sampled data distribution is shown in Tables 4, 5 and 6. Logistic regression is the only technique that did not show better improvement in false negative rates from the 10:90 to 34:66 data distribution. However, it showed overall best performance in the un-sampled distribution.

TABLE 1. Accuracy result for un-sampled data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
Accuracy	0.9737	0.9691	0.9824
Sensitivity	0.8072	0.8835	0.9767
Specificity	0.9741	0.9711	0.9824
Precision	0.0505	0.4104	0.0873
Matthews Correlation Coefficient	+0.1979	+0.5903	+0.2893
Balanced Classification Rate	0.8907	0.9273	0.9796

TABLE 2. Accuracy result for 10:90 data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
Accuracy	0.9752	0.9715	0.3639
Sensitivity	0.8210	0.8285	0.7155
Specificity	0.9754	1.0000	0.2939
Precision	0.0546	1.0000	0.1678
Matthews Correlation Coefficient	+0.2080	+0.8950	+0.0077
Balanced Classificati on Rate	0.8975	0.9143	0.5047

TABLE 3. Accuracy result for 34:66 data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
Accuracy	0.9769	0.9792	0.5486
Sensitivity	0.9514	0.9375	0.5833
Specificity	0.9896	1.0000	0.5313
Precision	0.9786	1.0000	0.3836
Matthews Correlation Coefficient	+0.9478	+0.9535	+0.1080
Balanced Classification Rate	0.9705	0.9688	0.5573

TABLE 4. Basic metric rates for un-sampled data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
True Positive Rate	0.8072	0.8835	0.9767
False Positive Rate	0.0259	0.0288	0.0176
True Negative Rate	0.9741	0.9711	0.9824
False Negative Rate	0.1928	0.1165	0.0233

TABLE 5. Basic metric rates for 10:90 data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
True Positive Rate	0.8200	0.8285	0.7155
False Positive Rate	0.0250	0.000	0.7061
True Negative Rate	0.9750	1.0000	0.2939
False Negative Rate	0.1280	0.1715	0.2845

TABLE 6. Basic metric rates for 34:66 data distribution

Metrics	**Classifiers**		
	Naïve Bayes	***k-Nearest Neighbour***	***Logistic Regression***
True Positive Rate	0.9514	0.9375	0.5833
False Positive Rate	0.0104	0.000	0.4688
True Negative Rate	0.9896	1.0000	0.5313
False Negative Rate	0.0486	0.0625	0.4167

B. *Comparative Performance*

The performance evaluation of the three classifiers for the 34:66 data distribution is shown in figure 1. This data distribution showed better performance. The k-nearest neighbour technique showed superior performance across the evaluation metrics used. It reached the highest value for specificity and precision (that is 1.0) for the two data distributions. This is because the kNN classifier recorded no false positive in the classification. Naïve Bayes classifier only outperformed the kNN in accuracy for the 10:90 data distribution. The Logistic regression classifier showed the least performance among the three classifiers evaluated. However, there was significant improvement in performance between the two sets of sampled data distribution. Since not all related works carried out evaluation based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and balanced classification rate, thus other related works are compared with this study based on the basic true positive and false positive rates. Figures 2 and 3 show the TPR and FPR evaluation of proposed Naïve Bayes, kNN and LR classifiers against other related works. The related works are referenced using their reference number delimited within square brackets "[ ]".

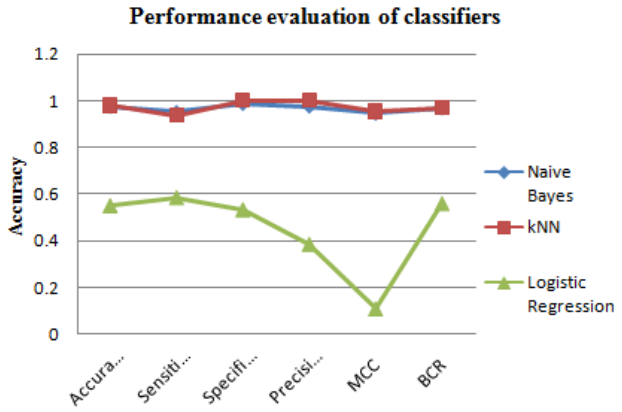

Figure 1. Performance evaluation chart for Naïve bayes, kNN and Logistic Regression

*MCC = Matthews Correlation Coefficient
*BCR = Balanced Classification Rate

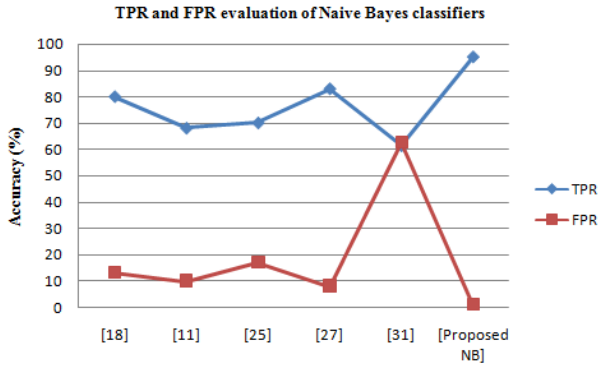

Figure 2. TPR and FPR evaluation of Naive Bayes classifiers

*TPR = True Positive Rate
*FPR = False Positive Rate
*Proposed NB = Proposed Naïve Bayes classifier

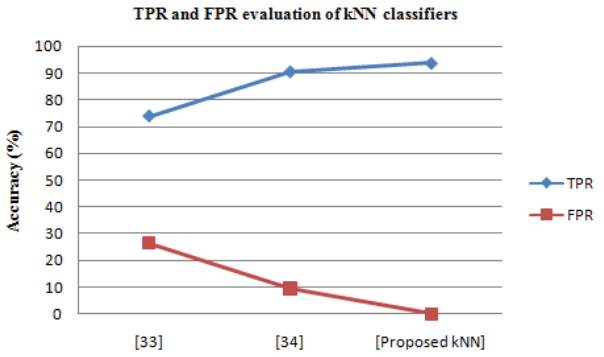

Figure 3.TPR and FPR evaluation of k-nearest neighbour classifiers

*TPR = True Positive Rate
*FPR = False Positive Rate
*Proposed kNN = Proposed k-nearest neighbor classifier

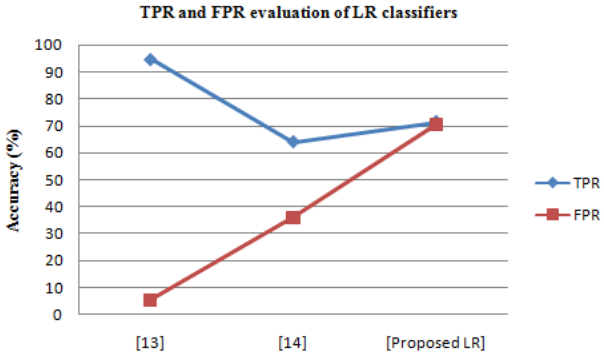

Figure 4.TPR and FPR evaluation of Logistic Regression classifiers

*TPR = True Positive Rate
*FPR = False Positive Rate
*Proposed LR = Proposed Logistic Rgression classifier

It could be observed that our proposed kNN classifier recorded zero false positive for both sets of data distributions (that is 10:90 and 34:66 datasets). Thus, the classifier shows better performance than reviewed works. The true positive and false positive rates evaluation on logistic regression with other works is shown in figure 4. There is an overlap between true positive and false positive rate for the 10:90 data distribution unlike in figures 2 and 3. This shows that the logistic regression classifier performs better on the un-sampled dataset than the two sampled sets.

V. CONCLUSION

This paper investigates the comparative performance of Naïve Bayes, K-nearest neighbor and Logistic regression models in binary classification of imbalanced credit card fraud data. The rationale for investigating these three techniques is due to less comparison they have attracted in past literature. However, a subsequent study to compare other single and ensemble techniques using our approach is underway. The contribution of the paper is summarized in the following:

1. Three classifiers based on different machine learning techniques (Naïve Bayes, K-nearest neighbours and

Logistic Regression) are trained on real life of credit card transactions data and their performances on credit card fraud detection evaluated and compared based on several relevant metrics.

2. The highly imbalanced dataset is sampled in a hybrid approach where the positive class is oversampled and the negative class under-sampled, achieving two sets of data distributions.
3. The performances of the three classifiers are examined on the two sets of data distributions using accuracy, sensitivity, specificity, precision, balanced classification rate and Matthews Correlation coefficient metrics.

Performance of classifiers varies across different evaluation metrics. Results from the experiment shows that the kNN shows significant performance for all metrics evaluated except for accuracy in the 10:90 data distribution. This study shows the effect of hybrid sampling on the performance of binary classification of imbalanced data. Expected future areas of research could be in examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data. Also effects of other sampling approaches can be investigated.

Acknowledgment

We wish to acknowledge Nwaiwu John C for his effort in the experimentation carried out and Pozzolo et al [32] for the source and description of the credit card fraud data.

References

[1] 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*.

[2] 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

[3] RamaKalyani, K. and UmaDevi, D., (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm, *International Journal of Scientific & Engineering Research*, Vol. 3, Issue 7, pp. 1 – 6, ISSN 2229-5518

[4] Meshram, P. L., and Bhanarkar, P., (2012). Credit and ATM Card Fraud Detection Using Genetic Approach, *International Journal of Engineering Research & Technology (IJERT),* Vol. 1 Issue 10, pp. 1 – 5, ISSN: 2278-0181

[5] 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

[6] Seeja, K. R., and Zareapoor, M., (2014). FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, *The Scientific World Journal*, Hindawi Publishing Corporation, Volume 2014, Article ID 252797, pp. 1 – 10, http://dx.doi.org/10.1155/2014/252797

[7] 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

[8] Duman, E., Buyukkaya, A., & Elikucuk, I. (2013). A novel and successful credit card fraud detection system implemented in a turkish bank. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on* (pp. 162-171). IEEE.

[9] Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 677-685). Society for Industrial and Applied Mathematics.

[10] Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, *2*, 841-848.

[11] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies* (pp. 261-270).

[12] Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In *Service Systems and Service Management, 2007 International Conference on* (pp. 1-4). IEEE.

[13] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602-613.

[14] Sahin, Y. and Duman, E., (2011). Detecting credit card fraud by ANN and logistic regression. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on* (pp. 315-319). IEEE.

[15] Chaudhary, K. and Mallick, B., (2012). Credit Card Fraud: The study of its impact and detection techniques, *International Journal of Computer Science and Network (IJCSN)*, Volume 1, Issue 4, pp. 31 – 35, ISSN: 2277-5420

[16] Bhatla, T.P.; Prabhu, V.; and Dua, A. (2003). *Understanding credit card frauds*. Crads Business Review# 2003-1, Tata Consultancy Services

[17] The Nilson Report. (2015). U.S. Credit & Debit Cards 2015. David Robertson.

[18] Stolfo, S., Fan, D. W., Lee, W., Prodromidis, A., & Chan, P. (1997). Credit card fraud detection using meta-learning: Issues and initial results. In *AAAI-97 Workshop on Fraud Detection and Risk Management*.

[19] 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 12th International Conference on* (Vol. 1, pp. 333-338). IEEE.

[20] Pun, J. K. F. (2011). *Improving Credit Card Fraud Detection using a Meta-Learning Strategy* (Doctoral dissertation, University of Toronto).

[21] Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916-5923.

[22] Bolton, R. J. and Hand, D. J., (2001). Unsupervised profiling methods for fraud detection, Conference on Credit Scoring and Credit Control, Edinburgh.

[23] Kou, Y., Lu, C-T., Sinvongwattana, S. and Huang, Y-P., (2004). Survey of Fraud Detection Techniques, *In Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control,* Taipei, Taiwan, March 21-23.

[24] Wheeler, R., and Aitken, S. (2000). Multiple algorithms for fraud detection. *Knowledge-Based Systems*, *13*(2), 93-99. Elsevier

[25] Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, *6*(1), 50-59.

[26] Chu, F., Wang, Y., & Zaniolo, C. (2004). An adaptive learning approach for noisy data streams. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on* (pp. 351-354). IEEE

[27] Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion*, *10*(4), 354-363.

[28] Sahin, Y. and Duman, E., (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, *Proceedings of International Multi-Conference of Engineers and Computer Scientists (IMECS 2011)*, Mar. 16-18, Hong Kong, Vol. 1, pp. 1 - 6, ISBN: 978-988-18210-3-4, ISSN: 2078-0966 (Online)

[29] Sherly, K. K. (2012). A comparative assessment of supervised data mining techniques for fraud prevention. *TIST. Int. J. Sci. Tech. Res*, *1*(16).

[30] Pun, J., and Lawryshyn, Y. (2012). Improving credit card fraud detection using a meta-classification strategy. *International Journal of Computer Applications*, *56*(10).

[31] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

[32] 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*.

[33] Fahmi, M., Hamdy, A. and Nagati, K., (2016). Data Mining Techniques for Credit Card Fraud Detection: Empirical Study, *In Sustainable Vital Technologies in Engineering and Informatics BUE ACE1*, pp. 1 – 9, Elsevier Ltd.

[34] Islam, M. J., Wu, Q. M. J., Ahmadi, M. and Sid-Ahmed, M. A., (2007). Investigating the Performance of Naive- Bayes Classifiers and KNearestNeighbor Classifiers. IEEE, International Conference on Convergence Information Technology, pp. 1541-1546.