Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-020-00430-y

ORIGINAL RESEARCH

Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection

Fayaz Itoo[1] · **Meenakshi**[2] · **Satwinder Singh**[2]

Received: 5 November 2019 / Accepted: 22 January 2020

Abstract Financial fraud is a threat which is increasing on a greater pace and has a very bad impact over the economy, collaborative institutions and administration. Credit card transactions are increasing faster because of the advancement in internet technology which leads to high dependence over internet. With the up-gradation of technology and increase in usage of credit cards, fraud rates become challenge for economy. With inclusion of new security features in credit card transactions the fraudsters are also developing new patterns or loopholes to chase the transactions. As a result of which behavior of frauds and normal transactions change constantly. Also the problem with the credit card data is that it is highly skewed which leads to inefficient prediction of fraudulent transactions. In order to achieve the better result, imbalanced or skewed data is pre-processed with the re-sampling (over-sampling or under sampling) technique for better results. The three different proportions of datasets were used in this study and random under-sampling technique was used for skewed dataset. This work uses the three machine learning algorithms namely: logistic regression, Naïve Bayes and K-nearest neighbour. The performance of these algorithms is recorded with their comparative analysis. The work is implemented in python and the performance of the algorithms is measured based on accuracy, sensitivity, specificity, precision, F-measure and area under curve. On the basis these measurements logistic regression based model for prediction of fraudulent was found to be a better in comparison to other prediction models developed from Naïve Bayes and K-nearest neighbour. Better results are also seen by applying under sampling techniques over the data before developing the prediction model.

Keywords Credit card fraud · Fraud detection · Random under-sampling · Logistic regression · Naïve Bayes · KNN

1 Introduction

Fraud are typically outlined as criminal duplicity with purpose of obtaining gain. With the fast growing dependence on internet technology, the rate at which credit card frauds happen has also increased at an alarming rate. Almost all modes of transactions be it online or offline are made via credit cards. External credit card fraud detection enjoys the inclination of majority of research work. There are two types of credit card frauds; one is inner card fraud and the other is external card fraud. Inner card fraud takes place when a false individuality is used to commit fraud because of mutual accord between cardholders and bank on the other hand fraud that is categorised as external consists of taking the credit card to induce cash through dubious means [1]. Credit card fraud might be considered as vital issue and amounts to a huge worth for banking organisations and card establishment companies. With this colossal disadvantage present in transaction system, banking organisations grade credit card fraud as a grave issue, to curb the menace they have fully-fledged security systems to keep a check on transactions and spot the frauds as quickly as possible upon conceived. Fraud detection is necessary so that we can impede the impact of dubious transactions on services of delivery, costs, and company name. Due to use of Machine learning, it has been helpful

✉ Satwinder Singh
satwindercse@gmail.com

[1] Central University of Punjab, Bathinda, India

[2] A.P. Department of Computer Science and Technology, Central University of Punjab, Bathinda, India

Published online: 15 February 2020

in finding variety of the mandatory business issues such as detecting email spam, targeted product recommendation, correct diagnosing etc. The promotion of machine learning has been attributed to the increasing process power, availableness of huge information and improvement in statistical modelling [2]. The most difficult thing for banks and commerce industry is the fraud management. The quantity of transactions has multiplied because of associate excessive number of payment channels—credit/debit cards, smartphones, and kiosks. At the same time, criminals became adept at finding loopholes. Hence it is not easy to authenticate transactions or it is very tough for businesses to authenticate transactions [3].

Data researchers have been quite successful in resolving this problem with machine learning and predictive analytics. The problem that comes with the credit card fraud detection is the skewed or unbalanced data and the algorithms treat the minority categories as a noise and only predicts the majority category accurately not the minority category [4, 5]. So in the case of a skewed data there are various resampling techniques which can be applied on a skewed or imbalanced data and can produce better results. The imbalanced problem of the dataset can also be solved by a technique called ensemble learning framework which guarantees the uprightness of test features and the method depends on training set split and congregate [6]. In order to overcome with the problem of the false alarm rates and increase the efficiency of credit card fraud detection rate various approaches like outlier detection methods have been used [7]. These approaches can optimise the best solution and their implementation on bank credit card fraud detection system (CCFDS) are very useful in detecting and preventing the fraudulent transaction [7].

2 Related work

Credit card fraud is on the rise as the number of online transactions are increasing. In order to prevent fraudulent transactions and to detect the credit card fraud there should be the most effective methods which are able to detect and prevent fraudulent transactions before they make a huge loss to the banks and credit card holders. There are various methods of fraud detection which are proposed by researchers and are somehow effective in credit card fraud detection but the real problem lies in availability of datasets due to security issues and datasets are very imbalanced. Some of the methods proposed for credit card fraud detection are Neural Networks, Fusion of Dempster Shafer, Bayesian Learning, Hidden Markow Model, Fuzzy Darwinian System, Outlier detection methods, Support Vector Machines, Genetic Algorithm, Covering Algorithm, Meta-Classifiers, Data Mining, ensemble Learning, machine Learning Algorithms (Random forest, Decision trees, Support Vector Machines, Bayesian Networks, MLP, Naïve Bayes and many more). Review of some of the research papers related to credit card fraud detection and its prevention is as follows:

Padvekar et al. [8] demonstrated that credit card misrepresentation are frequently distinguished utilizing hidden markov model all through transactions. hidden markov model gets a high misrepresentation inclusion joined with a low false alert rate. They utilized the scopes of exchange amount as the perception images, while the classifications of items are contemplated to be conditions of the HMM. They arranged a strategy for finding the expense profile of cardholders, correspondingly as use of this data in deciding the value of observation symbols and estimate of the model parameters. It's is also explained that how HMM is able to detect approaching transaction as fraudulent or not. Relative investigations uncover that the Accuracy of the framework is on the precarious edge of 80% over an extensive variety inside the data. The framework is furthermore ascendable for taking care of huge volumes of transactions.

Khare and Sait [9] examined and checked the presentation of Decision Tree, Random Forest, SVM and Logistic Regression classifier algorithms. The methods were used on the raw and pre-handled information. From the investigations the outcome that has been finished up is that Logistic regression has exactness of 97.7% while SVM indicates exactness of 97.5% and Decision tree demonstrates exactness of 95.5% yet the best outcomes are acquired by Random forest with an exact precision of 98.6%. The outcomes acquired therefore reason that Random forest demonstrates the most precise and high accuracy of 98.6% in issue of credit card fraud detection with dataset given by ULB.

Banerjee et al. [10] examined the various machine learning classifiers trained on a public dataset to analyse correlation of certain factors with fraudulence. The better metrics are used to determine false negative rates and the performance of random sampling was measured to deal with the class imbalance of the dataset. The support vector machine performed better for detecting credit card fraud under realistic conditions. The comparison between the deep learning and regression algorithmic models is done to determine which algorithm and combination of factors provides the most accurate method of classifying a credit-card transaction as fraudulent or non-fraudulent. The best algorithm for analysis of datasets with a close to 1:1 ratio of fraudulent to non-fraudulent transactions is the Random Forest Classifier, assuming the fraud-to-not fraud distribution of the testing and training set is the same.

Mishra and Ghorpade [4] analysed various classification techniques using various metrics for evaluating various

classifiers. The models were trained based on various classification and ensembling techniques. The models used were Logistic regression, Decision tree, Random Forest, Support Vector machines and various ensembling models. These models were trained and the results were obtained. Results obtained from the actual dataset were also good and with the recall of about 96% the Random Forest classifier performed better as compared to other classifiers.

Xuan et al. [11] utilized two sorts of random forest algorithms to prepare the features of typical and strange transactions. The two arbitrary random forest algorithms utilized are thought about which are distinctive in their base classifiers and their presentation is examined on credit card fraud detection. The two algorithms utilized are random-tree-based random forests and CART-based random forest whose preparation set originates from bootstrapped tests. The three experiments were performed for the two algorithms on different datasets with different proportions of datasets. The performance of these algorithms were measured for all the three experiments and the metrics which were added are intervention rate of transaction and average rate of model. Cart based random forest performed better in all the experiments performed.

3 Methodology

In this research work the machine learning classifiers namely: logistic regression (LR), K-nearest neighbour (KNN) and Naïve Bayes (NB)are put to application with Python serving as the language of implementation. The experiments are carried out and the evaluation of these experiments is done using the confusion matrix and performance comparison of the algorithms is analysed with the help of measures namely: accuracy, sensitivity, specificity, precision, F-measure and area under curve (AUC).

There are various stages which are involved in creating and processing of classifiers which include; gathering of data, pre-processing of data, training of algorithms, testing of algorithms and analysis of classifiers.

At the time of pre-processing of data, the data is transformed into viable format and is then sampled using the sampling techniques. A technique called random under-sampling is carried out on a dataset because of the highly imbalanced dataset which is more biased towards the negative cases (non-fraud cases) and due to the under-sampling of dataset, three sets of data distribution is achieved. The features selected are the principal components and these components are actually the product of principal component analysis dimensionality reduction resulting in 28 principal components which are represented as V1, V2 …, and V28. During the training stage the algorithms are given the input as the processed data. Testing datasets are assessed using trained model of the classifiers. Step by step methodology has been explained as follows. Figure 1 explains the flow diagram of research work.

3.1 Collection of dataset and pre-processing

The dataset is acquired from the Kaggle which hosts the dataset from credit card fraud detections [12]. The dataset is crafted from the MasterCard transactions of European cardholders on Sept 2013. The transactions that occurred for 2 days were recorded that amounts to 284,807 entries. The positive category (fraud cases) conjure 0.172% of the transactions information. The features are transformed and are reduced to 28 principal components as PCA is applied on them and are transformed into numerical input values. These principal components are named as V1, V2, V3 … and V28. The features include credit limit, gender, marital status, previous months bills, previous months payments, status of existing account, salary assignments, credit history, other credits existing, purpose, credit amount, present employment, savings account, personal status, other debtors, property, age in months, Housing, number of existing credits, Job, Telephone, foreign worker, ID, Credit card number, PIN, Time, Amount and Class. From the statistics of total entries and fraud cases it can be inferred that the dataset is very unbalanced and is inclined towards the

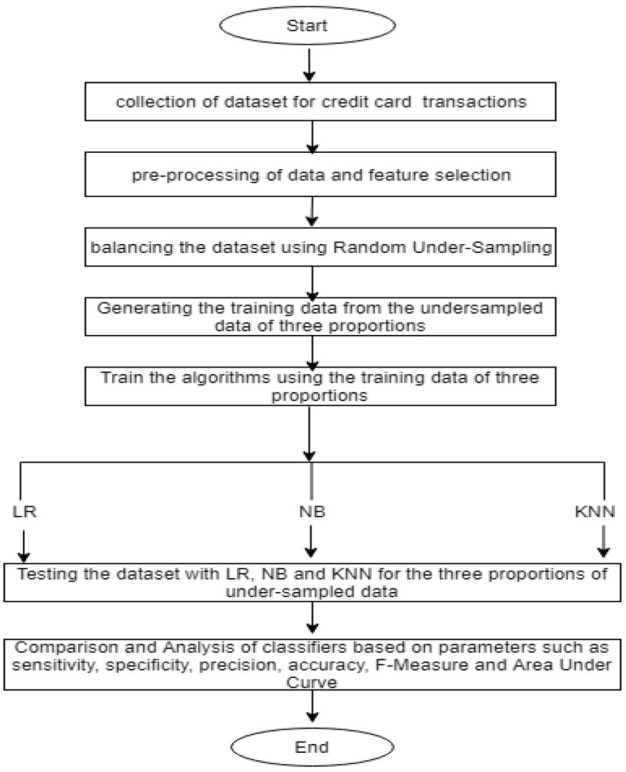

Fig. 1 Flow diagram of research work

negative class. The background details of the features are hidden and cannot be shown due to privacy issues. The time contains the seconds passed between every exchange and the primary exchange in the dataset. The 'Amount' feature is the exchange amount. The 'class' feature is used to represent whether the transaction is fraud or non-fraud and for the class value of 1 it represents the fraud transaction in the dataset and for the class value of 0 it represents the non-fraud transactions.

3.2 Under-sampling of dataset

To cope with unbalanced datasets, modifying classification algorithms in order to gain improvements or equalisation of classes within the training information conjointly known as data pre-processing is needed. Pre-processed information is provided as input to the machine learning rule because of its wide use. The chief target of knowledge pre-processing is increasing the number of the minority category or decreasing the frequency of the bulk class. This can be done with the aim of achieving same variety of instances for each the categories.

Random under-sampling (RUC) is a widely used resampling method and as such is used in our study. The choice of RUC is made on basis of its simplicity and effectiveness. The aim of RUC is to adjust class dispersion by means of arbitrarily dispensing with dominant part class precedents. The procedure is done until the greater part and minority occurrences are adjusted. RUC improves run time and capacity issues due to decreasing the quantity of preparing information tests in huge datasets. The lone limitation RUC suffers from is loss of some important information. In this study the dataset is distributed in three proportions taken as (fraud: non-fraud) ratio and the three proportions are: 50:50, 34:66 and 25:75. The results are taken for random under-sampling method for all the data distributions taken (Fig. 2).

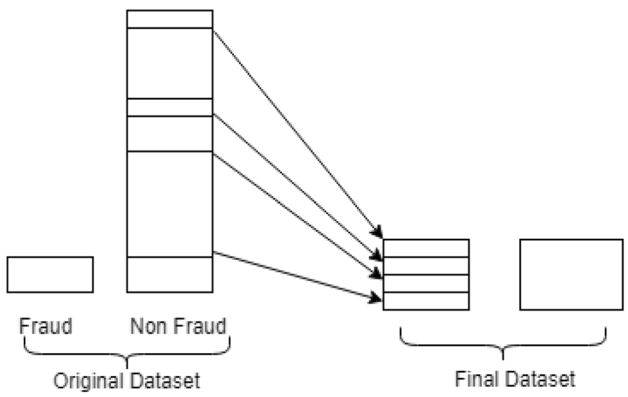

Fig. 2 Random under-sampling working of the dataset

3.3 Classification techniques

In the dataset of credit cards there are two values for classification of transactions which means that it is a binary classification problem where transactions are classified either as fraud (1) or non-fraud (0). After resampling of the data by under-sampling, the classifiers are trained using the training data to evaluate the methods. In this study classification techniques named as: logistic regression (LR), Naïve Bayes (NB) and K-nearest neighbour (KNN) are used.

3.4 Dataset division

The credit card dataset is split into two halves one training set and other testing set. In this study we chose the ratio 50:50, 34:66 and 25:75 (fraud: non fraud). Figure 3 below shows the overview of dataset division, which is split as training and testing and resampling (random under-sampling) is done. Also Tables 1, 2 and 3 shows more details about the dataset division.

In Table 1 division of the dataset is done by the ratio 50:50; it means that same number of fraud and non-fraud instances have been taken to train the three classification techniques.

In Table 2 division of the dataset (for fraud and non-fraud instances) is done in the ratio of 34:66; these numbers of instances have been taken to train the three classification techniques.

In Table 3 division of the dataset (for fraud and non-fraud instances) is done in the ratio of 25:75; these numbers of instances have been taken to train the three classification techniques.

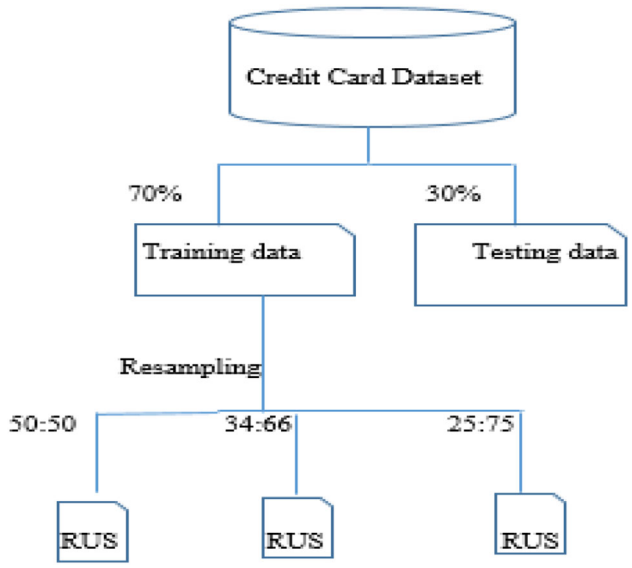

Fig. 3 Division of dataset

Table 1 DIvision of dataset by ratio 50:50

Data division	Training data	Resampling method RUS
Fraud	492	344
Non-fraud	284,315	344
Total	284,807	688

Table 2 Division of dataset by ratio 34:66

Data division	Training data	Resampling method RUS
Fraud	492	341
Non-fraud	284,315	692
Total	284,807	1033

Table 3 Division of dataset by ratio 25:75

Data division	Training data	Resampling method RUS
Fraud	492	353
Non-fraud	284,315	1024
Total	284,807	1377

Since 30% of dataset is used for testing the models and after resampling (random under-sampling) the number of fraud and non-fraud cases in the testing data for the three different proportions is given in Table 4.

Now onwards we will use A = 50:50, B = 34:66, C = 25:75.

After selection of the training and testing datasets, three different classification techniques namely LR, NB and KNN have been trained using the training dataset and we get the corresponding three models. Then testing dataset have been tested using these three models and then performance evaluation has been done.

Table 4 Preparation of testing dataset

Data proportion	Fraud	Non-fraud	Total
50:50	35	261	296
34:66	137	306	443
25:75	141	450	591

3.5 Performance evaluation

Performance evaluations were done for the three different classification techniques namely LR, NB and KNN for the resampling technique (RUS) used. The four elementary matrices through which performance evaluations are predicted are as: True Positive (TP), True negative (TN), False positive (FP) and false Negative (FN). True positives are the cases which are predicted as positive and in reality they are positive as well. True negatives are cases anticipated appropriately as negative. False positive are cases anticipated as positive yet are negative cases. False negative are cases delegated negative yet are actually positive. The correlations between these metrics is given in an exceedingly confusion metrics. Additionally the achievement of three algorithms are compared in terms of sensitivity, specificity, accuracy, F-measure and area under curve (AUC). The metrics used are calculated using the confusion metrics as shown in the Table 5 below.

Accuracy Accuracy is defined as the ratio of total number of predicted transactions that are correct [13]

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Sensitivity The proportion of positive observed values correctly predicted as positive. It is also called as True Positive Rate (TPR) [13]

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \tag{2}$$

Specificity Specificity is defined as, with how much accuracy the negative (legitimate) cases are classified and in our caseit gives the accuracy on prediction of legitimate transactions classification. It is also called as True Negative Rate (TNR) [13]

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{3}$$

Precision The proportion of positive (fraud) predictions that are actually correct [13].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

F-measure F-measure gives the accuracy of the test which means that it gives the accuracy of experiments performed. It uses the both precision and recall to compute

Table 5 Confusion matrix of credit card dataset

	Predicted fraud	Predicted non-fraud
Actual fraud	TP	FN
Actual non-fraud	FP	TN

its value. The best value for f1 score is considered at value 1 [14].

$$\text{F-measure} = 2.\frac{Precision.Recall}{Precision + Recall} \quad (5)$$

Area under curve (AUC) AUC represents degree or measure of separability that is how much model is capable of differentiating between the classes [14].

$$\text{AUC} = \frac{1}{2}.(Sensitivity + Specificity) \quad (6)$$

4 Result and discussion

This part deals with the results gathered during experiments. In the Tables 6, 7 and 8 given below, the comparison results of all the three classifiers in the resampling technique used for the ratios; 50:50, 34:66 and 25:75 respectively are shown. Parameters chosen for comparison of results are sensitivity, specificity, accuracy, precision, F-measure and AUC. From the results obtained it is clear that in all the three proportions logistic regression (LR) dominates with higher accuracy for the random under-sampling method used.

(i) Comparison of classification techniques by ratio A
(ii) Comparison of classification techniques by ratio B

The value of the parameters for the three classifiers for the ratio B is given in Table 7.

(iii) Comparison of classification techniques by ratio C

The sensitivity comparison of each proportion is shown in Fig. 4. Which represent the better result for logistic regression in comparison to other techniques Naïve Bayes and KNN.

Specificity comparison of each proportion is represented graphically in Fig. 5 which shows parallel results for logistic regression and Naïve Bayes for each proportion of under sampling ratio A, B and C

As shown in Figs. 6, 7 and 8 in all the three proportions, logistic regression algorithm shows the better result. It gives the highest accuracy in all the three proportions as compared to Naïve Bayes and KNN (Fig. 6). However KNN showed the lowest accuracy in all the three proportions as compared to other two algorithms in Fig. 6. In 25:75 ratio represented by 'C' all the algorithms performed better in term of accuracy measurement. So this split or ratio is considered to be the best for further training and testing purpose. Same results with better performance of Logistic Regression can be seen for other performance measurement parameters (precision and F-measure) in Figs. 7 and 8 respectively.

4.1 Analysis of algorithms for the three proportions

The values of parameters for the three proportions are depicted in Tables 6, 7 and 8. The Logistic Regression shows higher values for all the parameters because it maximizes the conditional data likelihood function. The provisional data likelihood is the probability of the noticed Y values in the training data, constrained on their respective X values. The second reason is that the feature values in Logistic Regression are dependent and there is a much more correlation between these features which contribute to the prediction of new data point. Logistic Regression also shows the higher accuracy (91.2%, 92.3% and 95.9% in case of A, B and C ratios respectively) for a medium size dataset and it is able to estimate the patterns for the fraud data in the balanced dataset. Also, in this study data balancing was done by under sampling methodology for the fraud detection. Above might be the reasons for better performance of each parameter (sensitivity, specificity, precision, F-measure and AUC) in logistic regression The decision boundary is set by the maximum conditional data likelihood function.

As noticed in Tables 6, 7 and 8 the Naïve Bayes algorithm shows lower accuracy than the logistic regression and this might be due to the reason that the features are independent of each other and each feature for the Naïve Bayes classifier contributes individually for the prediction of new data point. Secondly, the features are not correlated and this supposition dramatically decreases the number of parameters that must be estimated to learn the classifier and this might be the reason algorithm shows sometime lower performance values for sensitivity, specificity, accuracy, precision, F-Measure and AUC compared to the logistic regression. For example sensitivity measure for Naïve Bayes is 0.757, 0.718 and 0.664 as compared to 0.878, 0.777 and 0.839 Sensitivity measure of logistic regression

Table 6 Comparison of classification techniques by ratio A

Resampling method: random under-sampling						
Techniques	Sensitivity	Specificity	Accuracy	Precision	F-measure	AUC
Logistic regression	0.878	0.949	0.912	0.951	0.913	0.914
Naïve Bayes	0.757	0.964	0.854	0.959	0.846	0.860
K-nearest neighbour	0.687	0.669	0.679	0.701	0.694	0.678

Table 7 Comparison of classification techniques by ratio B

Resampling method: random Under-Sampling						
Techniques	Sensitivity	Specificity	Accuracy	Precision	F-Measure	AUC
Logistic regression	0.777	**1.0**	0.923	**1.0**	0.875	0.888
Naïve Bayes	0.718	**1.0**	0.902	**1.0**	0.836	0.859
K-nearest neighbour	0.477	0.789	0.681	0.544	0.508	0.633

Bold values highlight the parameters

Table 8 Comparison of classification techniques by ratio C

Resampling method: random under-sampling						
Techniques	Sensitivity	Specificity	Accuracy	Precision	F-Measure	AUC
Logistic regression	0.839	0.997	**0.959**	0.991	0.909	**0.918**
Naïve Bayes	0.664	0.995	0.915	0.979	0.789	0.829
K-nearest neighbour	0.405	0.861	0.751	0.483	0.441	0.633

Bold values highlight the parameters

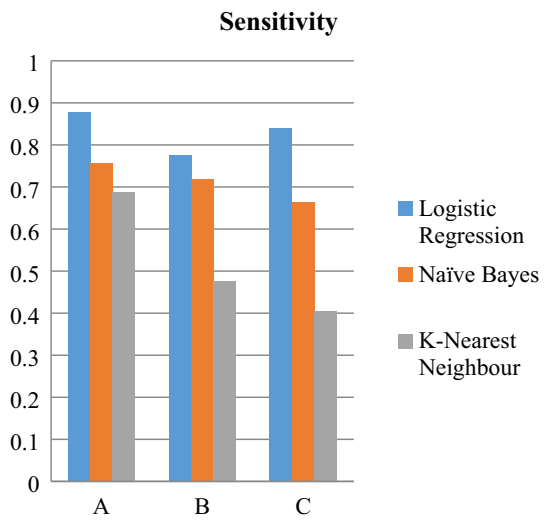

Fig. 4 Sensitivity

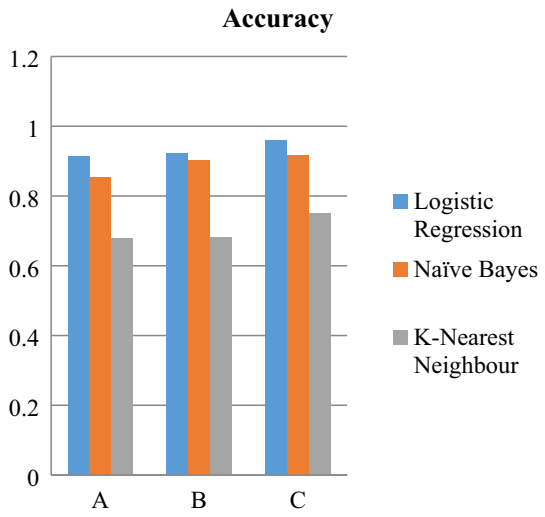

Fig. 6 Accuracy

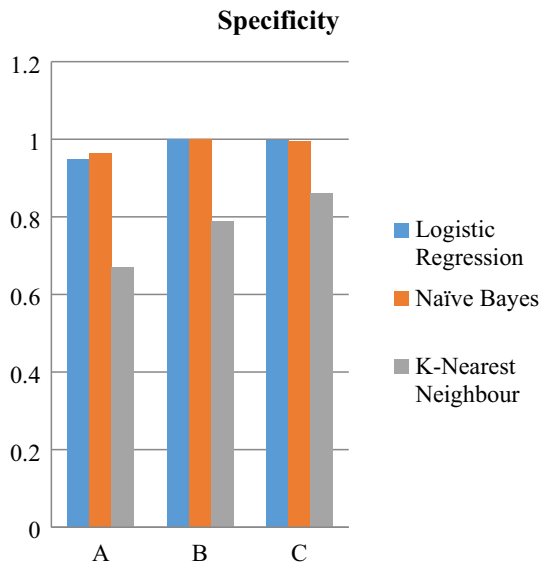

Fig. 5 Specificity

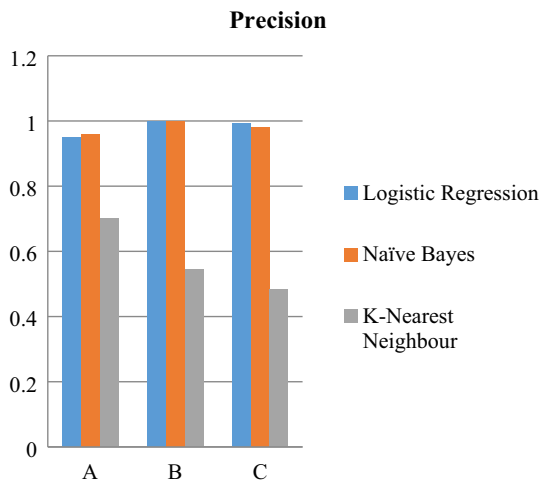

Fig. 7 Precision

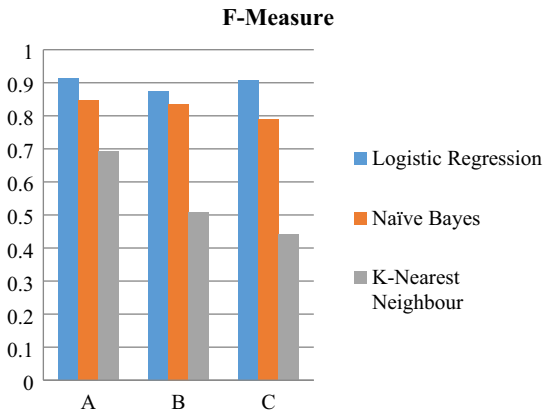

Fig. 8 F-measure

for respectively three different ratios of 50:50, 34:66 and 25:75. Also Naïve Bayes algorithm has a greater bias but lower variance than logistic regression which might be support the under sampling methodology of data balancing.

K-nearest neighbour requires a distance or measure the separation characterized between two information. In procedure of KNN, it characterizes any approaching transaction by ascertaining separation of closest point to new approaching transaction. At that point if the closest neighbour be deceitful, then the transaction demonstrates as a fraud. The estimation of K is utilized as, a little and odd to break the ties (normally 1, 3 or 5). Bigger K values can lessen the impact of boisterous dataset. In this algorithm, distance between two information instances is determined utilizing Euclidean distance. For multivariate information, distance is typically determined for each instance and after that consolidated. The algorithm shows the poor accuracy for the proportion C (50:50) in visualising the results from bar graphs in Figs. 4, 5, 6, 7 and 8 and this is due to the small sample of training data as there is much more similarity between the fraud and non-fraud cases and the algorithm does not efficiently differentiates the patterns in fraud and non-fraud cases.

As it is clear from the results of the experiments that the accuracy of classifiers increases as the training data is increased. Table 6 summarises the information for the ratio 34:66 for the Random Under-sampling (RUS). In this proportion logistic regression and Naïve Bayes have the same specificity rate which is 1.0 and it means that both the classifiers classified the negative cases (non-fraud) with 100% accuracy. This might be due to reason, as the training data sample increases the accuracy of both the classifiers increases. Since, both the algorithms estimate the priori probability which increases with number of samples in the training data and thus helps in classifying the data samples more accurately. It is depicted by the AUC values which is 0.89 and 0.85 for logistic regression and Naïve Bayes respectively. Table 6 clearly shows better performance for all the metrics (else than Accuracy and Precision) as compared to other proportions shown in Tables 7 and 8. The KNN showed the accuracy of 75% for 50:50 proportion which is better than the other proportions as it differentiates the classes more accurately when the training data increases. This is clear that when the training data is increased the algorithms are performing better and this shows that the data proportion of 25:75 is better for training the classifiers. For all the proportions taken LR performs very well with at most 95% correctness. The accuracy of all the classifiers for all the three data proportions is shown in the Fig. 6.

5 Conclusion and future work

The research work was carried out with the purpose of comparing the ability of machine learning algorithms as to how accurately they differentiate and classify the fraud and non-fraud transactions of the credit card dataset with random under sampling method (RUS) and to check out if the performance is improved or not. Logistic Regression (LR) showed the optimal performance for all the data proportions as compared to Naïve Bayes (NB) and K-Nearest Neighbour (KNN). LR was successful in getting higher accuracy as compared to Naïve Bayes and KNN. The LR showed the maximum accuracy of 95%, NB showed 91% and KNN 75%. Also LR technique shows the better Sensitivity, Specificity, Precision and F-Measure as compare to NB and K-NN technique. It has also been observed that being a supervised techniques (LR and Naïve Bayes) shows a better results in each case as compared to un-supervised technique K-NN.

There can be other resampling methods as well which could be put to application for the skewed dataset for credit card fraud detection (CCFD). The resampling methods could be improved to get better results. Also using our statistics could be compared with the other techniques like Random-Forest, SVC, Decision-Tress, Neural Network and Genetic Algorithm. The main limitation of Random Undersampling is that some information could be lost and new resampling methods could be devised for achieving optimal results which can prove helpful in credit card fraud detection (CCFD) in future. Likewise our results might be useful and can offer further help to the association to assemble a vastly improved credit card fraud detection system (CCFDS) which can be better in dealing with the skewed information and utilize the better measurements to assess the outcomes.

References

1. Kundu A, Panigrahi S, Sural S, Majumdar AK (2009) BLAST-SSAHA hybridization for credit card fraud detection. IEEE Trans Dependable Secure Comput 6(4):309–315
2. Guo T, Li G-Y (2008) Neural data mining for credit card fraud detection. In: International conference on machine learning and cybernetics
3. Ghobadi F, Rohani M (2016) Cost sensitive modeling of credit card fraud using neural network strategy. In: International conference of signal processing and intelligent systems (ICSPIS)
4. Mishra A, Ghorpade C (2018) Credit card fraud detection on the skewed data using various classification and ensemble techniques. In: 2018 IEEE International students' conference on electrical, electronics and computer science, SCEECS 2018
5. Raj SE, Portia AA (2011) Analysis on credit card fraud detection methods. In: International conference on computer, communication and electrical technology
6. Wang H, Zhu P, Zou X, Qin S (2018) An ensemble learning framework for credit card fraud detection based on training set partitioning and clustering. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, pp 94–98. https://doi.org/10.1109/SmartWorld.2018.00051
7. Malini N, Pushpa M (2017) Analysis on credit card fraud identification techniques based on KNN and outlier detection. In: Proceedings of the 3rd IEEE international conference on advances in electrical and electronics, information, communication and bio-informatics, AEEICB 2017, pp 255–258
8. Padvekar SA, Kangane PM, Jadhav KV (2016) Credit card fraud detection system. Int J Eng Comput Sci 5(4):16183–16186
9. Khare N, Sait SY (2018) Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math 118(20):825–838
10. Banerjee R, Bourla G, Chen S, Kashyap M, Purohit S, Battipaglia J (2018) Comparative analysis of machine learning algorithms through credit card fraud detection. New Jersey's Governor's School of Engineering and Technology, Piscataway, pp 1–10
11. Xuan S, Liu G, Li Z, Zheng L, Wang S, Jiang C (2018) Random forest for credit card fraud detection. In: ICNSC 2018—15th IEEE International conference on networking, sensing and control, pp 1–6
12. Hordri NF, Yuhaniz SS, Firdaus N, Azmi M, Shamsuddin SM (2018) Handling class imbalance in credit card fraud using resampling methods. Int J Adv Comput Sci Appl 9(11):390–396
13. Awoyemi JO, Adetunmbi AO, Oluwadare SA (2017) Credit card fraud detection using machine learning techniques: a comparative analysis. In: Proceedings of the IEEE international conference on computing, networking and informatics, ICCNI 2017, vol 2017–Jan, pp 1–9
14. Hordri NF, Yuhaniz SS, Azmi NFM, Shamsuddin SM (2018) Handling class imbalance in credit card fraud using resampling methods. Int J Adv Comput Sci Appl 9(11):390–396