Improving the classification of credit card fraud detection using Dimensionality Reduction Methods

Ashish Kumar[1], Shivank Kumar Soni[2], Chetan Agrawal[3]
[1,2,3] *Department of Computer Science, RITS, RGPV, Bhopal, India*
([1]ashish.java100@gmail.com, [2]shivanksoni@gmail.com, [3]chetan.agrawal12@gmail.com)

***Abstract*—** In today's world machine learning techniques are widely used for prediction and classification purposes. Some of the applications based on machine learning classification are identification of diseases, intrusion detection in network, sentiment analysis and many more. In this paper the main focus will be on credit card fraud detection approach through machine learning. This work is used to analyze and predict frauds done in credit cards and whole classification will be done by machine learning algorithms with feature selection techniques. The proposed work will show how feature selection will improve the accuracy of classification algorithms. This paper explores the execution of enhanced Naïve Bayes, K-Nearest Neighbor, Random forest and logistic regression on exceptionally skewed credit card extortion data by applying dimensionality decrease strategies. In this work feature selection method is applied for dimensionality reduction. Dataset of credit card exchanges is sourced from European cardholders containing 284,807 exchanges. The execution of the methods is assessed dependent on accuracy, affectability, precision and specificity. The outcomes shows of optimal accuracy for Naïve Bayes, logistic regression classifier, random forest and K-nearest Neighbor are 97.85%, 99.96%, 99.95% and 99.5% individually.

***Keywords*—** *Credit card, Fraud, Naïve Bayes, Random Forest, Accuracy*

I. INTRODUCTION

Financial extortion is one of the emerging problems today which is regularly developing danger with critical results in the business account, corporate account, association and government. In this financial extortion, credit card fraud is another criminal activity that is creating big problems to banking industry [1]. Reliability on the internet has increased by expanded credit card exchanges. As credit card exchanges turn into the most predominant method of installment for both on the web and disconnected exchange, credit card extortion rate also quickens. Credit card misrepresentation can come in either internal card extortion or external card misrepresentation. Inward card extortion happens because of assent among cardholders and bank by utilizing false character to submit misrepresentation while the external card extortion includes the utilization of stolen credit card to get money through questionable methods. Lots of inquiries have been done for the evolution of external card misrepresentation which represents lion's share of credit card cheats. Identifying frauds through traditional strategies like manual prediction is very difficult and wasteful, even this is the big problem that's why big data problems are not able to solve by manual techniques. Financial establishments have centered thoughtfulness regarding later computational techniques to deal with credit card extortion issue. Data mining procedure is one eminent strategies utilized in taking care of credit extortion discovery issue [2].

There are two classes like real (veritable) and deceitful exchanges through which exchanges are classified by credit card classification strategies [3]. Various different types of systems are developed for credit card fraud recognition techniques like Support Vector Machines [4], Itemset mining [5], Genetic algorithm [6], Decision Tree [7], Artificial Neural Network [8], etc. There are some more analysis that is basically by logistic regression [9] and naive Bayes is done in [10]. Nowadays dimensionality reduction methods are used with traditional classification algorithm so that they can readily recognize credit card extortion [11] while neural network and logistic regression is connected on credit card misrepresentation identification issue. Various challenges are related with credit card discovery, specifically fake conduct profile are dynamic, that is deceitful exchanges will in general look like real ones; credit card exchange datasets are infrequently accessible and exceedingly imbalanced (or skewed); optimal feature (variables) choice for the models; reasonable measurement to evaluate execution of strategies on skewed credit card misrepresentation data. Credit card extortion identification execution is significantly influenced by sort of sampling approach utilized, determination of variables and location technique(s) utilized. This paper seeks to complete similar analysis of credit card misrepresentation discovery utilizing naive Bayes, k-nearest neighbor and logistic regression procedures on exceptionally skewed data dependent on accuracy, affectability, and specificity.

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. LITERATURE SURVEY

Some of the well known methods used for credit card fraud detection are Logistic models, Bayesian belief network, neural

networks, and decision trees and all of them give original answers for the issue of recognition and classification of the fake data. Generally speaking, approaches connected for recognizing credit card fraud incorporate neural network, data mining, meta-learning, and support vector machine. Iyer et al describe the "Credit card fraud detection method by using Hidden Markov Model (HMM)" [12]. In this model they had used Hidden Markov Model (HMM) for modeling the sequence of operations in credit card transaction and it shows how this model can be used for detection of fraud transaction. This is trained with normal behavior of cardholder. S. Ghosh and Douglas L. Reilly et al describe the "Credit card fraud detection With Neural Network (NN)" [13]. This method is based on neural network in which credit card fraud detection was trained on large sample of labeled credit card account transactions and testing is performed on the holdout data that consist of all the accounts activity which is performed in the subsequent two months of time. Neural network was applied on the trained data which consist of record related to lost cards, stolen cards, application cards, counterfeit fraud and mail order fraud. It has significantly detected more fraud accounts with less false positives almost reduced by 20% as compared to rule based fraud detection procedures. In [14] proposed methods to detect fraud are presented. In this approach clustering is applied first to classify the legal and fraudulent transaction using regions clusterization of parameter value. After this Gaussian mixture model is applied to model the probability density of credit card users past behavior so that we can calculate the probability of current user behavior by detecting the abnormalities in the past behavior.

Lastly, Bayesian networks are used to describe the statistics of a particular user and the statistics of different fraud scenarios. Hilas and Mastorocostas [15] have proposed a methodology based on the client demonstrate recognizable proof. So as to test the capacity of each profile to segregate between genuine use and fraud, feed-forward neural network (FF-NN) is utilized as classifier. Panigrahi et al. (2009) [16] have proposed another technique for recognizing credit card fraud, which joins confirmations of the at various times conduct.

Kunal Goswami, Younghee Park and Chungsik Song [17] has developed feature set which can be compare with the state-of-the-art feature sets in detecting fraud. They consider feature set as the user's social interaction on the Yelp platform to observe whether the user is committing fraud. He concluded his work by computing F1 score obtained using neural networks is on par with all the well known methods for detecting fraud, a value of 0.95. The effectiveness of the feature set is in rivaling the other approaches to fraud detection. Masoumeh Zareapoor and Pourya Shamsolmoali [18] discuss about how various classification algorithm works during credit card fraud detection on the basis of confusion matrix parameters.

III. MATERIAL AND METHODS

A. *Feature selection methods*

In case of Big data dimensions are basically the features or attributes and there is very large number of dimensions [19]. It is very difficult to process high dimensional data, so to process high dimensional data feature selection and feature extraction methods [20]. These methods reduce the dimensions without the loss of information. After applying this method processing becomes easy and even many times performance also gets increased by machine learning algorithms. Feature is a one of a kind and quantifiable normal for a procedure that is noticeable. Whenever a credit card is utilized, the exchange data including various features, (for example, credit card ID, measure of the exchange, and so forth.) are spared in the database of the administration provider. Exact features unequivocally impact the execution of a fraud location system. Feature determination is the way toward choosing a subset of features out of a bigger set, and prompts a fruitful classification. The entire pursuit space contains all conceivable subsets of features, implying that its size is 2N, in which N is the quantity of features. Therefore feature choice is a NP-hard problem [21]. Repetitive and insignificant features are not valuable for classification, and they may even lessen the proficiency of the classifier with respect to the substantial inquiry space, which is the alleged revile of dimensionality.

B. *Dataset*

The dataset source commencing from ULB Machine Learning Gathering and depiction is found in [22]. The dataset contains credit card exchanges made by European cardholders in September 2013. This dataset presents exchanges that happened in two days, comprising of 284,807 exchanges. The positive class (fraud cases) makes up 0.172% of the exchanges data. The dataset is exceptionally unbalanced and skewed towards the positive class. It contains just numerical (ceaseless) input variables which are because of a Principal Component Analysis (PCA) feature choice change coming about to 28 principal components. Subsequently a total of 30 input features are used in this investigation. The subtleties and background data of the features can't be introduced because of confidentiality issues. The time feature contains the seconds passed between every exchange and the principal exchange in the dataset. The 'sum' feature is the exchange sum. Feature 'class' is the objective class for the paired classification and it takes value 1 for positive case (fraud) and 0 for negative case (non fraud).

IV. PROPOSED STRATEGY

The proposed strategy comprises of two primary parts, specifically feature determination which is a dimensionality decrease approach and classification. The initial segment of the proposed strategy incorporates division of the datasets and an all-encompassing wrapper technique that prompts choosing the best and the most effective features. The second part is the classification algorithm connected on the preprocessed dataset

which predicts the class whether that is fraudulent exchange or not. Let us see the flow chart for the proposed work.

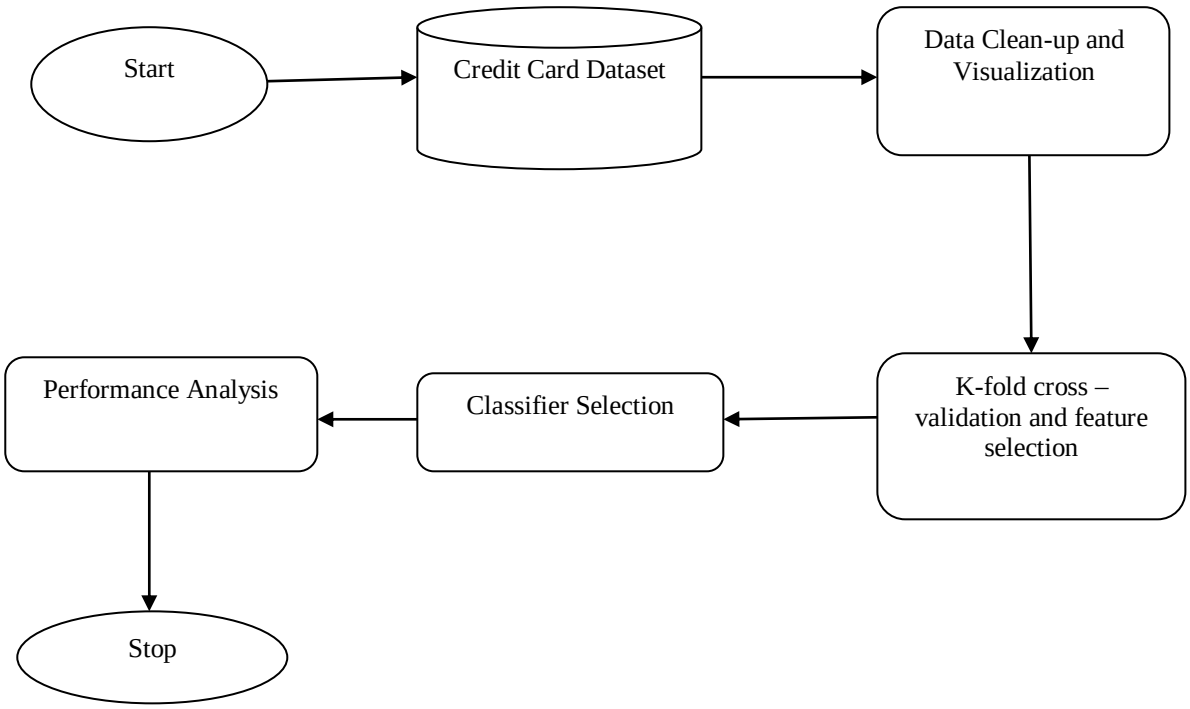

Figure 1: Flow chart of the proposed work

In this stage, to give steadiness on the best features for the final investigations, distinctive subsets of preparing dataset are made.

A. Algorithm

1. Import all libraries
2. Fetch the dataset i.e. credit card.csv
3. Split the dataset into train set and test set as x_train, y_train, x_test and y_test
4. Now feature selection approach for selecting the best feature using random forest classifier

```
clf_rf_5 = RandomForestClassifier()
clr_rf_5 = clf_rf_5.fit(x_train,y_train)
importances = clr_rf_5.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf_rf.estimators_],
axis=0)
indices = np.argsort(importances)[::-1]
```

5. Next classification algorithms will be applied on the reduced dataset.
6. Print accuracy, precision, recall and F-score.

V. RESULT ANALYSIS AND ITS PARAMETERS METRICS USED

Here, we utilize the Python version 3.6 for examination just as its parameter which is utilization of this examination. The arrangement of steps and all of the calculations with it will be appeared in this portion, in both parallel and sequential evaluation. The best four models for evaluations are likewise introduced here.

A. Confusion Matrix:

A confusion matrix is a unique of prediction result on a classification trouble. The numeral of right along with wrong predictions are aggregate up with consider values well as broken down through each class. This is the key in to the confusion matrix. The confusion matrix exhibits the technique in which your classification model is confounded while it makes predictions. It gives us drawing nearer not just into the mistakes individual made through a classifier however further unmistakably the kind of blunders that are being made.

	Set 1(Yes) **Predicted**	**Set 2(No)** **Predicted**
Set 1(Yes) **Actual**	TP	FN
Set 2(No) **Actual**	FP	TN

Here,
Set 1: Positive (Yes)
Set 2: Negative (No)

Explanation of the Terms:
Positive (P): Test is positive (for case : is an apple).
Negative (N): Test is not positive (for case: is not an apple).
True Positive (TP): Test is positive, along with is predicted to be positive.
False Negative (FN): Test is positive, other than is predicted negative.
True Negative (TN): Test is negative, along with is predicted to be negative.
False Positive (FP): Test is negative, other than is predicted positive.

B. Classification Accuracy:

Classification Accuracy is recognized during the relation: Although, there are harms throughout the accuracy. It is assumed that corresponding costs for commonly kind of errors. 99% accuracy could be good, excellent, poor, middling, or else dreadful depending primary the difficulty.

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

At this point, we have utilize classification algorithms accessible inside meticulous library. Originally, we will estimation the confusion matrix consequent that we will compute the accuracy throughout via function or else confusion metrics. accuracy_score;

According to Logistic Regression

	Yes	No
Yes	**77979**	2
No	**30**	73

Accuracy Score: 99.95901849290507%
Precision score: 0.9733333333333334
Recall score: 0.7087378640776699
F1 Score: 0.8202247191011237

According to Naive Bayes demonstrate the confusion matrix

	Yes	No
Yes	**76323**	**1658**
No	**20**	83

Accuracy Score: 97.85103222170996%
Precision score: 0.04767375071797817
Recall score: 0.8058252427184466
F1 Score: 0.09002169197396963

According to Random Forest, we will demonstrate the productivity of Dataset which is given below:

	Yes	No
Yes	**77981**	**0**
No	**24**	79

Accuracy Score: 99.9692638696788%
Precision score: 1.0
Recall score: 0.7669902912621359
F1 Score: 0.868131868131868
As indicated by K- nearest neighbour

	Yes	No
Yes	**77976**	**5**
No	**31**	72

Accuracy Score: 99.95389580451821%
Precision score: 0.935064935064935
Recall score: 0.6990291262135923
F1 Score: 0.8

Logistic Regression, Naïve Bayes, Random Forest and K-nearest neighbor classifier were implemented and compared to each other in terms of accuracy score. The comparison of classifiers results are shown in the following table.

Method	Base Methods	Proposed Methods
Logistic Regression	0.9824	0.9995
Naive Bayes	0.9737	0.9785
K- nearest neighbor	0.9691	0.9995
Random Forest	-----	0.9996

Table 1: Comparisons of previous as well as present result on given data set

The above table shows how our proposed methods perform better than the based method. In the given work we had also implemented the work on the basis of accuracy, precision, recall and F1 score.. The proposed work performs better than the base method and Random forest classifier outperforms other classifier like Naïve Bayes, Logistic regression and K-Nearest Neighbor. Graph comparison is shown in the given below figures for accuracy, precision, recall and F1 score.

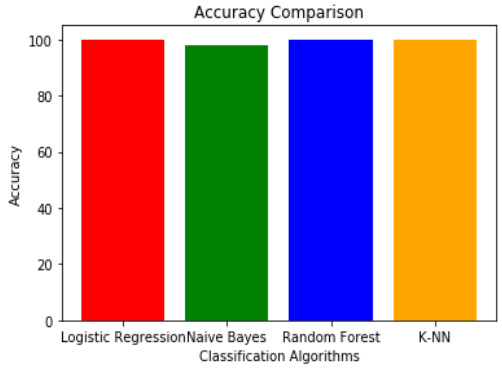

Figure 2- Accuracy comparison of classifiers

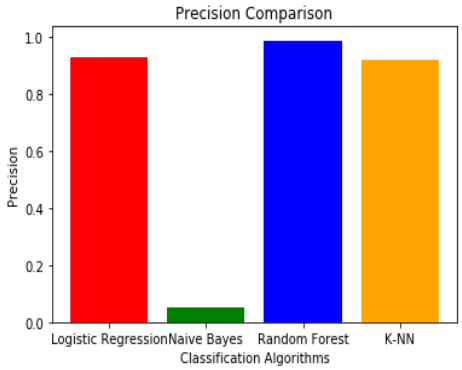

Figure 3- Precision comparison of classifiers

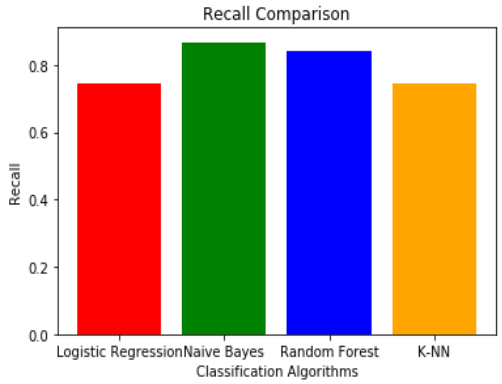

Figure 4- Recall comparison of classifiers

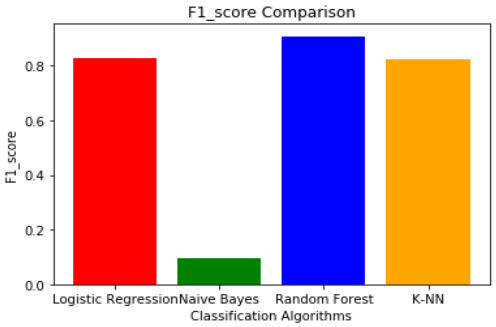

Figure 5- F1 score comparison of classifiers

VI. CONCLUSION AND FUTURE WORK

In this work we have shown how credit card classification using features selection method. Feature selection method reduced the dimensions due to which classifiers performance has been enhanced. 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. In the future we will apply deep learning methods which may helps us to perform classification without dimensionality reduction method.

References

[1] L. Delamaire, H. Abdou, and J. Pointon, "Credit card fraud and detection techniques: a review," Banks Bank Syst., 2009.

[2] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., 2011.

[3] Data Mining: Practical Machine Learning Tools and Techniques. 2016.

[4] D. V. and D. R., "BEHAVIOR BASED CREDIT CARD FRAUD DETECTION USING SUPPORT VECTOR MACHINES," ICTACT J. Soft Comput., 2016.

[5] K. R. Seeja and M. Zareapoor, "FraudMiner: A novel credit card fraud detection model based on frequent itemset mining," Sci. World J., 2014.

[6] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," Expert Syst. Appl., 2011.

[7] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," Expert Syst. Appl., 2013.

[8] L. S. Raghavendra Patidar, "Credit Card Fraud Detection Using Neural Network," India Int. J. Soft Comput. Eng., 2011.

[9] Y. Sahin and E. Duman, "Detecting credit card fraud by ANN and logistic regression," in INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications, 2011.

[10] M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the performance of Naive- Bayes

classifiers and K- nearest neighbor classifiers," in 2007 International Conference on Convergence Information Technology, ICCIT 2007, 2007.

[11] A. Ghodsi, "Dimensionality Reduction A Short Tutorial," Science (80-.)., 2006.

[12] D. Iyer, A. Mohanpurkar, S. Janardhan, D. Rathod, and A. Sardeshmukh, "Credit card fraud detection using hidden Markov model," in Proceedings of the 2011 World Congress on Information and Communication Technologies, WICT 2011, 2011.

[13] Ghosh and Reilly, "Credit card fraud detection with a neural-network," in Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94, 1994.

[14] V. Dheepa and R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods," ternational J. Recent Trends Eng., 2016.

[15] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," Knowledge-Based Syst., 2008.

[16] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection: A fusion approach using