ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20180427091048
Original scientific paper

Comparative Analysis of Different Distributions Dataset by Using Data Mining Techniques on Credit Card Fraud Detection

Oğuz ATA, Layth HAZIM

Abstract: Banks suffer multimillion-dollars losses each year for several reasons, the most important of which is due to credit card fraud. The issue is how to cope with the challenges we face with this kind of fraud. Skewed "class imbalance" is a very important challenge that faces this kind of fraud. Therefore, in this study, we explore four data mining techniques, namely naïve Bayesian (NB),Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Random Forest (RF), on actual credit card transactions from European cardholders. This paper offers four major contributions. First, we used under-sampling to balance the dataset because of the high imbalance class, implying skewed distribution. Second, we applied NB, SVM, KNN, and RF to under-sampled class to classify the transactions into fraudulent and genuine followed by testing the performance measures using a confusion matrix and comparing them. Third, we adopted cross-validation (CV) with 10 folds to test the accuracy of the four models with a standard deviation followed by comparing the results for all our models. Next, we examined these models against the entire dataset (skewed) using the confusion matrix and AUC (Area Under the ROC Curve) ranking measure to conclude the final results to determine which would be the best model for us to use with a particular type of fraud. The results showing the best accuracy for the NB, SVM, KNN and RF classifiers are 97,80%; 97,46%; 98,16% and 98,23%, respectively. The comparative results have been done by using four-division datasets (75:25), (90:10), (66:34) and (80:20) displayed that the RF performs better than NB, SVM, and KNN, and the results when utilizing our proposed models on the entire dataset (skewed), achieved preferable outcomes to the under-sampled dataset.

Keywords: credit card fraud detection; data mining; K-Nearest Neighbour; Naïve Bayesian; Random Forest; Support Vector Machine

1 INTRODUCTION

In recent years, there has been an increasing amount of literature on credit card fraud detection. The issue of credit card fraud has been studied in [1-5]. During the growth in credit card transactions, such as the electronic payment system, there has been an increase in credit card fraudcases, and 70 percent of US customers are most concerned about identity fraud [6, 7]. The Federal Trade Commission's online database of customer complaints has received 13 million complaints from 2012 to 2016, with 3 million in 2016 alone. Some of these, almost 42% were related to fraud and 13% were complaints concerning identity theft [8]. Thus, banks are attempting to decrease their losses from credit card fraud. Consideration should be given to the development of fraud detection methods such as data mining techniques as fraudsters also develop their fraud practices to avoid detection [9]. Hence, credit card fraud detection techniques require continuous innovation. This paper evaluates four techniques, including the naïve Bayesian, support vector machines, *k*-nearest neighbor algorithm, and random forests in an attempt to detect credit card fraud. Our study is based on real-life data transactions from European cardholders.

Credit card fraud may be divided into two types: 1) "Offline fraud", which is committed with a stolen physical card anywhere else such as call center, 2) "Online fraud", which is committed over the phone, shopping, on the Internet, or in the absence of a cardholder. Fraud detection may be supervised or unsupervised [9]. In supervised fraud detection methods, a database of known 'fraudulent / genuine' transactions is used to classify new transactions as being either 'fraudulent' or 'genuine'. In unsupervised fraud detection methods, there are no prior sets of genuine or fraudulent observations, which means that unusual or outlier transactions are identified as potential cases of fraudulent transactions [6]. Two fraud detection approaches perform a prediction of the possibility of fraud in any of the new transactions [10].

Credit card fraud detection depends on the analysis of cardholder's spending behavior. Most data mining techniques are applied to credit card fraud detection and support vector machines [6], [11-13]. Many researchers have used artificial neural networks and genetic algorithms [14-19], there have been credit card fraud detection comparative analyses using logistic regression, *k*-nearest neighbors and naïve Bayesian [20], credit card fraud detection using the *k*-nearest neighbor algorithm in [21, 22], hybrid approaches for detecting credit card fraud using random forest in [5], Bayesian network, decision tree, naïve Bayesian, K* models and support vector machine. The application of credit card fraud detection [23] is based on the bagging ensemble classifier, as well as using the hidden Markov model (HMM) in reference [24, 25], the migrating birds optimization algorithm in reference [26], and real-time credit card fraud detection using the computational intelligence self-organizing map (SOM) in reference [27, 28]. This paper [6] evaluates four techniques, including naïve Bayesian, support vector machines, the *k*-nearest neighbor algorithm, and random forests in an attempt to detect credit card fraud. It examines the performance of these techniques. However, we encountered many challenges in this study, including 'fraudulent' behavior appearing 'genuine' wherein real dataset transactions are not made available and results are not typically declared to the public even if found to be highly imbalanced (skewed) datas [1]. Therefore, feature selection is a problem in this study because of the large disparity in measurements and high dimensions of fraud dataset and the presence of numbers of 'features', 'attributes', 'inputs' which make application of "data mining" techniques and detection very difficult and complicated. We have chosen existing performance measures for the aggregating techniques. Four most commonly used measures are accuracy, sensitivity, specificity, and precision, all of which depend on true positives, false positives, true negatives and false negatives [4]. All these performance measures are affected by the

type of sampling used for the dataset. In this study, we have investigated the effect of aggregating sampling on the performance of fraud detection techniques, including the naïve Bayesian, support vector machines, *k*-nearest neighbor and random forest classifiers on highly imbalanced ('skewed') credit card fraud transactions, as well as their impact on used under-sampling fraud transactions.

In this study, we endeavor to make an analytical comparison of credit card fraud detection using NB, SVM, KNN and RF techniques on highly imbalanced data based on accuracy, sensitivity, specificity, and precision. The region under the ROC curve (AUC) is utilized as a standard measurement of classification performance [13], where we used AUC. Finally, to examine all techniques with skewed credit card transactions to obtain the best technique even we can advise to use with that type of fraud. This study enhances the handling of highly imbalanced credit card fraud data in [29]. This study also uses highly imbalanced dataset transactions which contain approximately 0,172% of fraud transactions being sampled in aggregating approaches. The fraud transactions indicate a positive class, while the negative class indicated the genuine, by using the under-sampling approach. The skewed or imbalanced transactions have been overcome as a part of preprocessing the dataset because of the small fraudulent credit card transaction percentages of the total number of transactions. A balancing handling mechanism is desired to make this data balanced with a '1:1' distribution between 'genuine' and 'fraudulent' classes to reshape any class imbalance [1, 11], where the distribution is in a '50:50' format. Four techniques were applied using the confusion matrix to calculate the accuracy, sensitivity, specificity, and precision to compare the performance of the four techniques afterward to additionally verify the performance measures. We applied cross-validation with 10 fold and Grid Search of the aggregating techniques and compared the performance. Finally, we applied our aggregating techniques to the imbalanced dataset and calculated the accuracy, sensitivity, specificity, precision, and AUC to compare each technique to obtain the most accurate one in this field of fraud.

The remainder of our paper includes the following: Section 2 presents a historical review of several techniques that have been used in credit card fraud detection, sampling approaches, and performance comparisons. Section 3 presents and describes the methodology, including data pre-processing, the under-sampling approach and the four classifier techniques of credit card fraud detection. Section 4 presents the results for our experimental setup, including illustrations, figures and a discussion about the comparison of the analyses. Section 5 presents a conclusion of the comparisons in our study and proposals for future work and research.

2 LITERATURE REVIEW

Credit card fraud detection is a binary classification problem in which a credit card transaction is labeled as either fraudulent or genuine. Data mining approaches are useful in this type of fraud detection because of their ability to identify small anomalies in huge data sets. In this section, we reviewed some previous researches that are relevant to this study [1].

2.1 Under-Sampling Approach

Under-sampling imbalanced classes mean deleting part of the data in the majority class or the negative class (genuine) [11]. Many researchers have used the under-sampling approach to balance the training data for fraud detection systems [1]. This approach has been used in reference [30]. They have used two sampling approaches, which are over-sampling and under-sampling. These sampling approaches are commonly used in machine learning algorithms to imbalanced classes and costs for misclassification. They studied cost curves to explore the interaction of undersampling and oversampling with the learner C4.5 of a decision tree. They concluded that under-sampling results in a reasonable sensitivity to variations in the costs of misclassification and class distributions, and Over-sampling has shown little sensitivity. In reference [31] they have employed three algorithms, which are logistic regression, C4.5 and random forest for cost-sensitive credit card fraud detection. They applied those algorithms on the full dataset and the under-sampled dataset; applying the under-sampling has given the best results. A comparative study [6] included testing different levels of under-sampling class distributions by data mining techniques. Comparison results showed that under-sampling generally performs better, compared to the hybrid under sampling and oversampling for credit card transactions using machine learning techniques [20] they also gave the same assessment of this approach and achieved two sets of distribution (10:90 and 34:64) for analysis. The paper in reference [32] discussed the effectiveness of undersampling on unbalanced classification. It has proposed an integrated analysis for two objects having the biggest effect on the efficiency of an under-sampling approach. This analysis increases the variance because of reducing the number of samples and counterfeiting (warping) of the posterior distribution. They concluded two main influences. It raises the classifier's variance and results in counterfeited (warped) posterior possibilities. Usually, the first influence is addressed using averaging methods for reducing the variability and the second needs the calibration of the possibility to the new priors of testing.

2.2 Credit Card Fraud Detection

Fraudsters are also increasing their attempts to get money because of growing use of online payment by credit cards. Through the significant contribution researchers in recent years are finding the best ways to reduce fraud by relying on data mining techniques or artificial intelligence techniques.

Datamining for credit card fraud [6] utilized three methods, support vector machines (SVM), random forest (RF) and logistic regression (LR) to evaluate the best one depending on the performance measures. They used under sampling of the imbalanced classes for their real transactions dataset from international companies with various proportions and they are divided into two subsets. They have applied three proposed techniques with cross-

validation performance and the results were: SVM (93,8 accuracy; 52,4 sensitivity; 98,4 specificity), RF (96,2 accuracy; 72,7 sensitivity; 98,7 specificity) and LR (94,7 accuracy; 65,4 sensitivity; 97,9 specificity). The authors in [5] proposed a hybrid approach of six well-known data mining techniques, namely, DT, RF, BN,NB, SVM, and their proposed model K* employed these models to detect credit card fraud. They combined an ensemble of the artificial intelligence (AI) models which have been applied into real life transactions from a leading bank in Turkey. The results in terms of performance measures were: DT results are 95,19 accuracy, 52,53 sensitivity, 97,35 specificit, RF results are 95,81 accuracy, 50,84 sensitivity, 98,09 specificity, BN results are 96,92 accuracy, 50,00 sensitivity, 99,30 specificity, NB results are 94,10 accuracy, 92,57 sensitivity, 94,18 specificity, SVM 94,17 accuracy, 66,89 sensitivity, 95,55 specificity and K* results are 91,37 accuracy, 73,14 sensitivity, 92,67 specificity.

In reference [33] they have investigated the efficiency of the personalized models in comparison with the aggregated structures in identifying fraud for various people. The authors used two techniques for comparison, which are random forest and Naive Bayesian. The dataset were collected from actual transactions and some other information via an on-line questionnaire. The performance results of their proposal has shown that RF is more efficient performance than the NB for the aggregated model whereas NB is more efficient performance in the personalized models, as follows: RF results are 91,09 accuracy, 91,1 sensitivity, 91,9 precision, NB results are 96,04 accuracy, 96,00 sensitivity, 95,9 precision and RF results are 96,18 accuracy, 96,00 sensitivity, 96,00 precision, NB results are 95,08 accuracy, 95,00 sensitivity, 95,00 precision. Researchers in [23] have proposed three techniques for credit card fraud detection, which are naive Bayesian, support vector machine and *k*-nearest neighbors. They used these models with a collaboration of an ensemble of the learning methods, the evaluation of performance is done on a real dataset transaction from UCSD-FICO competition, and the authors showed the bagging classifier based on the decision tree, as the best one for the fraud model.

The study in [34] used some classification methods, which are the artificial neural networks (ANN) and the logistic regression (LR) for creating the best model to detect credit card fraud, where they have concluded that the genetic algorithm is the best in their literature and they proposed to apply it on bank systems to predicted fraud soon after credit card transactions. The paper in [35] employed three supervised methods to predict credit card frauds, which are the logistic regression (LR), gradient boosted trees (GBT), and deep learning (DL),. Authors research also explores the benefits according to features by using the domain expertise and feature engineering to compare with the three techniques that have been mentioned above. They concluded that using domain expertise for feature engineering is the best and their results after applying the cross-validation with 5 fold were: LR (83,8), GBT (87,4) and DL (86,2). In [36] the authors have presented a survey of two techniques, which are the Hidden Markov Model (HMM) and k-means clustering, which have been adopted for the analysis of the spending behavior of cardholders. HMM categorized the cardholder's profile into low, medium and high, and then made clustering by using k-means clustering for the categorized cardholder behavior. HMM can detect the new arriving transaction as fraudulent or genuine.

Previously, we have historically reviewed comparative studies for credit card fraud detection, now we will review some historical studies for machine learning and features engineering. The study in [37] showed it is the way of extracting the proper traits from the transactions for constructing credit card fraud detection approach, by aggregating the transactions, and they expanded the transaction aggregation strategy, as proposed by creating a new group of properties according to the analysis of the time of the transaction by employing the "von Mises" distribution. Topological pattern in [38] discovered the 'topological patterns' of 'fraudulent financial reporting' FFR via dual 'GHSOM' ('Growing Hierarchical Self-Organizing Map') approach, as well as presenting an expert competitive feature extraction mechanism, which has been accurate in detecting the fraudulent and genuine by using the topological patterns for FFR and feature extraction. On the other hand, the authors in [39] proposed a linear discriminate as the fisher discriminant function to detect credit card fraud for the first time, their experiment which has been produced from the fisher discriminant function was more efficient for the fraudulent / genuine detection classifier. The study in [40] proposed a combination of the derived intrinsic features and network-based features for cardholders' behavior merchants, their results for the combination of the two types which are strongly tangled, and leads to the best performance models where the 'AUC' reaches higher than 0,98.

A new cost-sensitive decision tree in [41] compared the traditional popular classification method with the performance like precision and true positive rate to minimize the sum of misclassification costs, the outputs showed that the cost-sensitive decision tree may be ready and implemented in real transactions to avoid fraud for credit card transactions. The study in [42] applied the k-nearest neighbors (KNN) method and outlier detecting approach to put the optimal solution for credit card fraud issues, where those two methods minimized the false alarm rates and minimized the fraud detecting rate to prevent the fraudulent transactions. In [43] the researchers have implemented the self-organizing map (SOM) for credit card fraud detection because of the effectiveness of this approach, it is a part of the neural network and the unsupervised learning, focused on real-time credit card fraud detection. They concluded that the SOM is more accurate for the detection of the fraud due to using the clustering with that model.

3 METHODOLOGY

The objectives of this section are to describe the performance measures and examine the four data mining techniques (NB), (SVM), (KNN) and (RF) for credit card fraud detection so as to give the right advice for banks about which best technique to use to build their system. This section includes different stages: The first stage includes performance measures used for evaluating our comparison. The second stage is the dataset and sampling

dataset, which includes: how to sample the data set, how to collect this dataset and how to use the training dataset and testing dataset. The third stage explains the naïve Bayesian classifier. The fourth stage explains the support vector machine classifier. The fifth stage explains the *k*-nearest neighbor classifier. The sixth stage explains the random forest classifier and finally, the seventh stage describes the confusion matrix, cross-validation and AUC measures.

3.1 Performance Measures

In this study, we will use four well-known measures to evaluate the methodology, which are: accuracy, sensitivity, specificity and precision. These measures depend entirely on the four basic metrics (alarm rates), respectively 'true positive' (TP) the number of fraudulent transactions which were detected as a true alarm, 'False Positive' (FP) the number of genuine transactions which have been detected as a false alarm, 'True Negative' (TN) the number of genuine transactions that have been detected as true alarm and 'False Negative' (FN) are the number of missed fraudulent transactions [5], positives (P) mean the number of the "fraudulent transactions" and negatives (N) represents the number of the "genuine transactions" the total of P and N means all transactions. Evaluating the classification performance requires to know the meaning of each measure, where accuracy means the proportion of true alarm rates among all alarm rates, sensitivity (recall) the proportion of the true positives, which indicates the number of the fraudulent cases that are detected correctly, specificity represents the proportion of the true negatives, which indicates the number of genuine transactions which have been detected correctly too, and precision measures the proportion of true positive among all positives alarms. Below are the equations for each measure:

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

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

$$Sensitivity\ = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

Credit card fraud detection is a binary classification which means that the transactions for a credit card are divided into two sets, either (0) - genuine transactions or (1) - fraudulent transactions. Using the confusion matrix for our study the calculation of alarm rates will be different [6], where the values for each alarm rate are taken from its position in the confusion matrix such as the *TP*, where the value was (1, 1), *TN* (0, 0), *FP* (0, 1) and *FN* (1, 0).

We also used AUC (area under ROC curve) measures to examine the overall performance measures, the AUC is better than the accuracy measures for evaluating learning algorithms [44]. AUC has been tested on the positive class (fraudulent) *FP* and *TP*.

3.2 Dataset and Sampling Dataset

Highly imbalanced dataset for our study has been collected from the European cardholders in September 2013. The real credit card transactions and their description are included in this dataset [2], so it has been published in 2016. It was originally collected by a research collaboration of (ULB) Wordline and University Libre de Bruxelles to analyze big data and fraudulent transactions. Tab. 1 describes all thirty-one variables in the data, where the feature variables besides Time and Amount are displayed with an unknown description due to the protection of the sensitive information. These are not the original variables obtained during the collection of data. They have all been transformed with principal component analysis (PCA) to protect the true information from the analyst examining the data (or other third parties that may contribute to negative consequences). In other words, V1 - V28 are principal components holding the real data in some fashion. All twenty-eight (V_s) variables and Amount are categorized as numerical, while Class and Time are both integers. The dataset presents the transactions for two days 284,807, feature (Class) is used for binary classification among these features, it takes value 1 referring to the fraudulent transactions and 0 referring to the genuine transactions. These transactions consist of (492) fraud transactions, which represents almost (0,172%) of the total transactions.

This little number of class positives (fraud) are high imbalance class. In this case, we have to sample the skewed class among the existing approaches. We proposed in our study using the under-sampling imbalanced class. Under-sampling is a commonly used technique for imbalanced datasets to decrease the skew in the class distributions [29]. Under-sampling was used to remove the observation values from the majority class (genuine) randomly until the dataset reaches the balance because the minority class (fraudulent) is very small in comparison with the majority class, where there is equal proportion amongst the fraudulent / genuine (1:1), under-sampling is beneficial for handling the imbalanced datasets [41].

Table 1 Description of dataset and attributes

Attributes	Type	Description
Time	int	The time between each transaction
V1	num	Feature variable with unknown information
.	.	———
.	.	———
V28	num	Feature variable with unknown information
Amount	num	Total money spent
Class	int	Response attribute (0 = genuine and 1 = Fraudulent)

3.3 Naïve Bayesian Classifier (NB)

This classifier is a highly efficient probabilistic approach for supervised classification as well as a statistical method which uses class data from the training examples for the prediction of the future fraud class. The classification has been performed via implementing the Bayesian rule for the calculation of the possibility of the correct class given the specific features of the credit card transactions [45]. We used Gaussian Naive Bayesin in our study, this model extended the real-valued attributes. Gaussian distribution is the easiest and merely requires the

estimation of the mean and the standard deviation from the training data, following the equation of Gaussian Naive Bayes [46].

$$P(c_i|f) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{(c_i-\mu_f)^2}{2\sigma_f^2}\right)$$

Where i indicates either 0 for genuine transactions or 1 for the fraudulent transactions from the training data. These two values mean that the classification problem is binary as we mentioned, is a probability of feature value f being in class c_i, the μ_f and σ^2 are a mean and standard deviation calculating values of each input variable (c_i) for each class value.

If $P(c_1|f) > P(c_i|f)$ then the classification is c_1

If $P(c_1|f) < P(c_i|f)$ then the classification is c_2

The class c_i is a target or predicted class for classification where c_1 is the negative class (genuine) and c_2 is a positive class (fraudulent).

3.4 Support Vector Machine Classifier (SVM)

This is a supervised and statistical learning approach which has been used for a variety of classification problems successfully, suitable for binary classification problems as a credit card fraud detection. Support vector machines are linear classifiers which operate in a high dimensional property space which is a nonlinear mapping of the input space of the present problem [6], SVM is used for solving the non-linear classification problems. The advantages for the support vector machines are a result of two significant features. They possess kernel representation and margin optimization, where the kernel function is the trick which is used to convert the nonlinear problems to the linear problems. We can even extract optimal solutions for our problem, then, we can deal with the problem to find the (hyper-plan) with maximum separation margin between both classes to avoid any risk for overfitting the training instances. There are three functions to transform the 'nonlinear' to 'linear' classification, namely, 'polynomial function', 'radial basis function (Gaussians)' and 'sigmoid (neural net activation function)'. We used in our study radial basis function (RBF) because our dataset is nonlinear and it works with a wide variety of problems like credit card problems, following the equation for Gaussian's RBF:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma \geq 0$$

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$$

where, $\varphi : x \to H$ is a trick to transform the input space x into a higher dimensional spaceb H, $K(x_i, x_j)$ is the used kernel function and $\varphi(x_i)$ is the transformation function [46].

3.5 K-Nearest Neighbor (KNN)

This algorithm is a strong and largely used method in the detection systems. The KNN classifier is used successfully in the credit card fraud detection problem, that is always used as a benchmark for more complex classifiers such as Artificial Neural Networks and Support Vector Machines [42]. KNN is a supervised learning method, in this technique the new instance query will be classified depending on the well-known KNN distance measures such as Manhattan distance, Euclidean distance, and Minkowski distance. In our study we used the Euclidean distance between two instances (transactions), where each incoming transaction will be computed of its nearest point to the new incoming transaction to detect fraud, following its formulation, which is given by [20]:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}, k = 1, 2, ..., n$$

The KNN algorithm basic boils are going to form a majority vote between the K most similar instances to a given "predicted" observation. The Euclidean distance between two data points is new input data point with the current data point computed; the distances that have been computed are sorted and arranged incrementally and the lowest distances are selected with k-items to the input data point. The binary classification for our study means the negative class among these items is found and the KNN classifier returns the positive class such as the classification for the input data point. Parameters are chosen for k neighbors, and start from k = (1, 2, 3, 4, 5, 6, 7, …, 12) the classifier returns the k = 5, which is the best parameter for the accuracy performance, so, this is the parameter which is used in the classifier.

3.6 Random Forest (RF)

Random forest (RF) is an aggregate of the decision tree models or a combination of tree predictors [47]. It is using the average to improve predictive accuracy and control the overfitting, because the (RF) method is supervised. It trains each subsample (tree) of the original training set input on different bootstraps and the size of the sub-sample is the same as the original input data. After that, a random subsample of all the available features is used. This returns a forest of the decision trees that are very different from each other [48], every one of the trees in aggregate is produced from an arbitrary sub-sample of features. Because many studies recommended using this technique among different data mining techniques its performance has achieved the best accuracy. In the present study, every one of the trees in the set is constructed from a sample drawn with replacement (i.e. bootstrap sample) from the training group. The essential parameters used in our study estimator parameter are the number of the trees in the forest. Where the used random estimators start from E = (1, 10, 100, 1000) were the best accuracy which has been obtained resulted from E = 100. Criterion parameter is the function for measuring the quality of a split, selected from that parameter 'Gini'. Impurity is the best for the improvement of the performance, and max features

parameter is the number of properties to take under consideration when looking for the optimal split. Selected from that parameter 'auto' is the best for performance. These parameters are the most important for our study. The RF is the best technique in terms of performance for detecting fraud among the four techniques that we have used in our study.

4 RESULTS

In this section, we present our results of performance measures from four data mining methods: Naïve Bayesian, Support Vector Machines, K-Nearest Neighborand Random Forests for different divisions of training and testing dataset, evaluated from the training data which carry different levels of fraud cases using the confusion matrix (CM).

4.1 Experiment Results

First, we present in Tab. 2 our results for the performance measures, i.e. accuracy, sensitivity, specificity, precision, and AUC, respectively, for the four techniques and different divisions such as (90:10), (80:20), (75:25) and (66:34) after we followed the under-sampling approach to balance the dataset by removing several genuine classes to reach the number of the minority class. This means that the majority class becomes 492 transactions, equalling the number of the minority class, where we used 90% of the under sampled dataset to train 885 transactions and 10% of the under sampled dataset to test 99 transactions, 80% of the under sampled dataset to train 787 transactions and 20% of the under sampled dataset to test 197 transactions, 75% of the under sampled dataset to train 738 transactions and 25% of the under sampled dataset to test 246 transactions and 66% of the under sampled dataset to train 649 transactions and 34% of the under sampled dataset to test 335 transactions.

Table 2 Performance of under-sampling data set for four techniques

Models	Performance Measures of (90:10) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	91,91%	85,10%	98,07%	97,56%	92%
SVM	94,94%	95,74%	94,23%	93,75%	95%
KNN	96,96%	93,61%	100%	100%	97%
RF	97,97%	95,74%	100%	100%	98%
Models	Performance Measures of (80:20) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	90,86%	85,71%	95,28%	93,97%	90%
SVM	92,38%	94,50%	90,56%	89,58%	93%
KNN	95,43%	91,20%	99,05%	98,80%	95%
RF	97,46%	94,50%	100%	100%	97%
Models	Performance Measures of (75:25) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	89,43%	81,51%	96,85%	96,03%	89%
SVM	92,68%	94,11%	91,33%	91,05%	93%
KNN	95,12%	89,91%	100%	100%	95%
RF	96,74%	94,11%	99,21%	99,11%	97%
Models	Performance Measures of (66:34) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	89,55%	81,76%	97,57%	97,20%	90%
SVM	92,53%	91,17%	93,93%	93,93%	93%
KNN	94,02%	88,82%	99,39%	99,34%	94%
RF	94,62%	90,0%	99,39%	99,35%	95%

Note that in the results in Tab. 2, the (90:10) division produced better results than other divisions, where the RF technique performed better when we compared with other techniques in the performance evaluation on all divisions that have been used, so, the KNN also produced highest results, reaching 100 for specificity and precision compared with the SVMs and the NB, while the NB produced less efficient results. Moreover, the SVM performed better than the NB.

Table 3 Apply cross-validation performance measures of four techniques and standard deviation

Models	Performance Measures of (90:10) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	*Std.*
NB	91,41%	86,08%	95,58%	96,53%	2,58%
SVM	93,22%	91,03%	98,21%	95,34%	1,58%
KNN	92,66%	87,89%	97,22%	97,30%	2,45%
RF	92,21%	87,88%	97,09%	96,39%	3,29%
Models	Performance Measures of (80:20) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	*Std.*
NB	90,97%	85,53%	95,16%	96,55%	2,45%
SVM	92,51%	91,01%	97,92%	94,18%	2,90%
KNN	92,89%	88,78%	97,21%	97,00%	1,96%
RF	93,02%	89,03%	97,60%	97,87%	2,32%
Models	Performance Measures of (75:25) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	*Std.*
NB	90,66%	83,94%	95,35%	97,25%	3,28%
SVM	93,23%	90,91%	97,55%	95,50%	3,06%
KNN	93,90%	89,01%	96,86%	98,88%	1,52%
RF	92,83%	88,76%	97,49%	97,73%	2,78%
Models	Performance Measures of (66:34) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	*Std.*
NB	89,84%	82,93%	94,36%	96,25%	2,81%
SVM	93,05%	91,62%	98,06%	94,42%	2,52%
KNN	94,14%	89,14%	96,79%	99,02%	2,34%
RF	92,60%	89,14%	96,61%	96,68%	2,98%

Second, in Tab. 3, we presented the results for our models. However, here we applied cross-validation to obtain the traditional classification performance for the same proposed divisions used for training and testing the datasets as well as for the under-sampling. In this table, we added another measure, namely a standard deviation (*Std*), which measures the spread of the dataset. The dataset with the smaller (*Std*) has a narrower spread of measurements around the mean and it usually has comparatively low values. As we observe in this table, there is an increase in the accuracy and sensitivity of the NB technique as well as an increase in the accuracy for the SVM on all divisions approximately. However, in the remaining measures for the NB and the SVM, there is a notable decrease. While the performance measures of KNN and RF have decreased when applying the cross-validation. On the other hand, the standard deviation recorded different percentages in each division.

For the third way, we present in Tab. 4 the results of the four models when applied to the entire dataset, which is a skewed dataset. In terms of dataset divisions previously, we also divided the dataset into four sampling datasets of 90% for training and 10% for testing, which means 256,326 transactions for training and 28,481 for testing, 80% for training and 20% for testing, which means 227,845 transactions for training and 56,962 for testing, 75% for training and 25% for testing, which means 213,605 transactions for training and 71,202 for testing, 66% for training and 34% for testing, which means 187,972 transactions for training and 96,835 for testing. We observed that all performance measures for the four models have noted increases in all divisions. It is important to note in Tab. 4 that the precision measure has different

values than in the above tables. This difference is due to the precision measure depending on the true positive class (fraudulent) among all positives and the number of fraudulent transactions in the entire dataset being a very small approximate (0,172%) of all the transactions in comparison with the genuine transactions, as mentioned previously.

Table 4 Performance results for imbalanced dataset (skewed) distributions

Models	Performance Measures of (90:10) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	97,56%	89,09%	97,58%	6,65%	93%
SVM	97,19%	96,36%	97,19%	6,22%	97%
KNN	98,56%	100%	98,56%	11,85%	99%
RF	98,57%	100%	98,57%	11,93%	99%
Models	Performance Measures of (80:20) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	97,80%	88,11%	97,82	6,71%	93%
SVM	97,46%	93,06%	97,47%	6,14%	95%
KNN	98,16%	90,09%	98,17%	8,06%	94%
RF	98,23%	97,02%	98,23%	8,89%	98%
Models	Performance Measures of (75:25) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	97,46%	84,16%	97,48%	5,34%	91%
SVM	95,04%	95,0%	95,04%	3,13%	95%
KNN	97,55%	95,83%	97,55%	6,20%	97%
RF	97,70%	98,33%	97,70%	6,73%	98%
Models	Performance Measures of (66:34) Data Distribution				
	Accuracy	Sensitivity	Specificity	Precision	AUC
NB	97,70%	82,63%	97,73%	5,91%	90%
SVM	97,39%	90,41%	97,40%	5,67%	94%
KNN	97,97%	88,02%	97,98%	7,03%	93%
RF	98,25%	94,61%	98,26%	8,61%	96%

Finally, we present the area under the ROC curve (AUC) measure for our experimental techniques. To compare them, we propose to use the AUC measure for two types of datasets. The first use is with the under sampled data and the second use with the imbalanced data when used with under sampling data. The results for four models are illustrated as we showed in above Tab. 2, Tab. 3 and Tab. 4, where, as usual, the highest results were RF, KNN, SVM, and NB, respectively.

4.2 Comparative Results of Distributions

The performance evaluation of the four classifiers for the four divisions in terms of Accuracy as we have shown in Fig. 1, Fig. 2 and Fig. 3, the (90:10) distribution showed better performance of all three tested steps approximately, as the (80:20) distribution also produced high accuracy, especially the NB and SVM, while (66:34) produced the low accuracy of all four techniques used in our study.

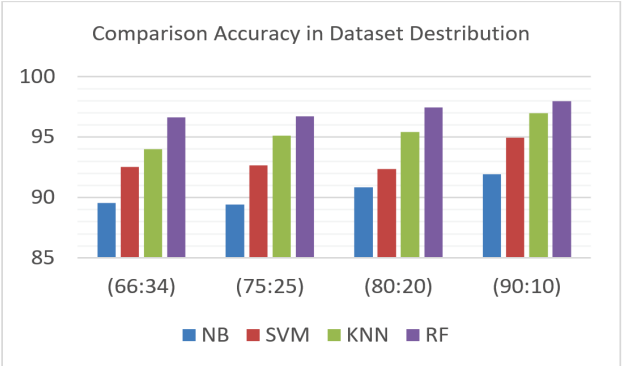

Figure 1 Performance of under-sampling data set for four techniques and four distributions

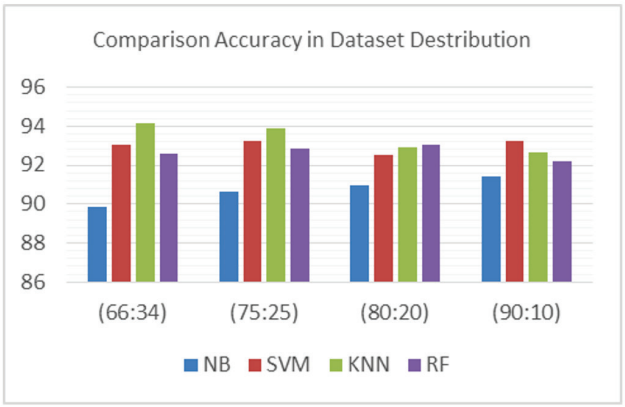

Figure 2 Apply cross-validation performance measures of four techniques and four distributions

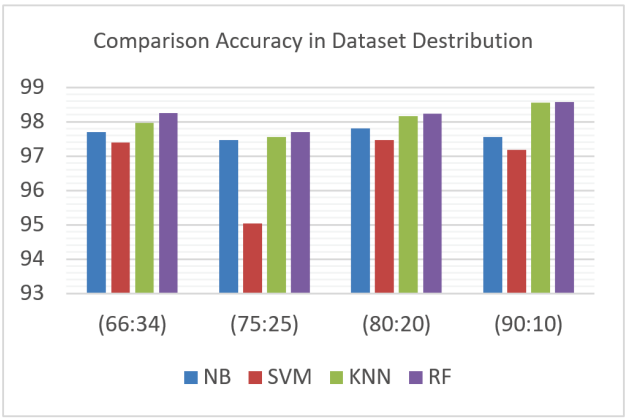

Figure 3 Performance results for imbalanced dataset (skewed) distributions and four distributions

5 CONCLUSION

In this study,we have investigated four classification methods, which are Naïve Bayesian (NB), Support Vector Machines (SVMs), K-Nearest Neighbor (KNN) and Random Forests (RF), and examined credit card fraud problems with binary classification, as this problem has become very common in banks. Our paper has contributed to three major trends. First, we have tested the four proposed techniques following the under-sampling dataset approach. Second, we have applied cross-validation with a 10 fold iteration and compared the performance of the four methods. Third, an examination has been made of the four classification methods while being applied to the entire dataset or skewed dataset with comparing their performances.

5.1 Discussion

The results of this paper conclude that the performance measures of the proposed models gave better result when applied to the entire dataset than when they were used on the under sampled dataset, due to the under sampling approach that suffers from weakness when used with a huge dataset, where removing the number of majority class even equal to minority class has a great effect on the results. On the other hand, when we used cross validation, some of the techniques increased their efficiency and others decreased it. It has also been concluded from our comparative analysis, that the Random Forest (RF) technique is the best classification technique for credit card

fraud problems and has better results for all evaluated performances on the three examination results. Therefore, we advise the use of this technique with huge datasets with 100 estimators.

5.2 Comparison with the Previous Study

As we pointed out previously, to make sure that we have chosen the best method for the classification of credit card fraud detection as genuine and fraudulent, we compared our results with the previous works, which are shown in Tab. 5 below.

In the study [20] the same dataset has been used as the one in this study. They have obtained these results after testing two data distributions, (90:10) and (66:34), where they concluded that the best result was on the second data distribution. Compared to our study where four data distributions have been used (90:10), (80:20), (75:25) and (66:34), where the results were enhanced even when we used the same distributions as the abovementioned study, as shown in Tab. 5:

Table 5 Comparison with the previous study

Reference	Training & Testing	NB	SVM	KNN	RF
[6] (2011)	-	-	93,8	-	96,2
[33] (2012)	-	96,04	-	-	91,09
[5] (2016)	-	94,10	94,17	-	95,81
[20] (2017)	(66 : 34)	97,69	-	97,92	-
[20] (2017)	(90 : 10)	97,52	-	97,15	-
Our Study (2018)	(90 : 10)	97,56	97,19	98,56	98,57
Our Study (2018)	(66 : 34)	97,70	97,39	97,97	98,25
Our Study (2018)	(75 : 25)	97,46	95,04	97,55	97,7
Our Study (2018)	(80 : 20)	97,80	97,46	98,16	98,23

After making this comparison with the previous works, and comparing the detailed performance measures with the RF algorithm, we also concluded that the RF algorithm performs better than the other researches. We can advise as we mentioned, using this technique and applying it to the huge dataset directly without using the sampling approaches.

Acknowledgments

We thank and wish for an increase in knowledge to Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and GianlucaBontempi in [29] for supporting by the real dataset transactions from ULB Machine Learning Group and their description of dataset.

6 REFERENCES

[1] Maarof, M. A. & Abdallah, A. Z. A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 90-113. https://doi.org/10.1016/j.jnca.2016.04.007

[2] Boracchi, G., Caelen, O., Alippi, C., & Dal Pozzolo, A. (2017). Credit Card Fraud Detection: A Realistic Modeling a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems, 2162-237X. http://www.ieee.org/publications_standards/publications/rights/index.html

[3] Caelen, O., Le Borgne, Y.-A.Waterschoot, S., Bontempi, G., & DalPozzolo, A. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 4915-4928. https://doi.org/10.1016/j.eswa.2014.02.026

[4] West, J. & Bhattacharya, M. (2015). Intelligent financial fraud detection: A comprehensive review. *ScienceDirect*, 47-66. https://doi.org/10.1016/j.cose.2015.09.005

[5] Yiğit, K. & Mehmet, U. Ç. (2016). Hybrid approaches for detecting credit card fraud.wiley - Expert Systems, 1-13.

[6] Sanjeev, J., Kurian, T., & Christopher, W. S. B. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 602-613. https://doi.org/10.1016/j.dss.2010.08.008

[7] McAlearney, S. (2008). TJX Data Breach: Ignore Cost Lessons and Weep.

[8] Federal Trade Commission, Consumer Sentinel Network Data Book, January-December 2016, March 2017.

[9] Richard, J. B. & David, J. H. (2001). Unsupervised Profiling Methods for Fraud Detection.london.

[10] Richard, J. B. & David, J. H. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, 235-255. https://doi.org/10.1214/ss/1042727940

[11] Chen, T. S., Lin, C. C., & Chen, R. C. (2006). A New Binary Support Vector System for Increasing Detection Rate of Credit Card Fraud. *International Journal of Pattern Recognition and Artificial Intelligence*, 227-239. https://doi.org/10.1142/S0218001406004624

[12] Sahin, Y. & Duman, E. (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. *International Multiconference of Engineers and Computer Scientists (IMECS), 442-447.* https://doi.org/10.1109/INISTA.2011.5946108

[13] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 30-55. https://doi.org/10.1007/s10618-008-0116-z

[14] Devi, D. U. & Kalyani, K. R. (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm. *International Journal of Scientific & Engineering Research*, 1-6.

[15] Hamdi, O. E. D. M. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications*, 13057-13063. https://doi.org/10.1016/j.eswa.2011.04.110

[16] Xin, L. & Haiying, M. (2009). Application of Data Mining in Preventing Credit Card Fraud. *International conference on management and service science (MASS*), 1-6.

[17] Freisleben, B. & Aleskerov, B. R. E. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. *Computational Intelligence for Financial Engineering (CIFEr)*, 220-226.

[18] Duman, E. & Sahin, Y. (2011). Detecting Credit Card Fraud by ANN and Logistic Regression.international symposium on innovations in intelligent systems and applications (INISTA), 15-319.

[19] Carsten, A. W. P. (2008). *Credit Card Fraud Detection Using Artificial Neural Networks Tuned by Genetic Algorithms* (PhD thesis). Hong Kong University of Science and Technology.

[20] Adebayo, O. A. & Samuel, A. (2017). Oluwadare John O. Awoyemi, "Credit card fraud detection using Machine Learning: A Comparative Analysis. Computing Networking and Informatics (ICCNI), 1-9.

[21] Naga, S., Mannem, P., & Ganji, V. R. (2012). Credit card fraud detection using anti-k nearest neighboralgorithm. *International Journal on Computer Science and Engineering (IJCSE)*, 1035-1039.

[22] Yu, W. F. & Wang, N. (2009). Research on Credit Card Fraud Detection Model Based on Distance Sum. *International Joint Conference on Artificial Intelligence*, 353-356.

[23] Zareapoora, M. & Shamsolmoali, P. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. *Procedia Computer Science*, 679-685. https://doi.org/10.1016/j.procs.2015.04.201

[24] Patil, V. & Bhusari, S. (2011). Application of Hidden Markov Model in credit card fraud detection. *International Journal of Distributed and Parallel Systems (IJDPS)*, 203-211. https://doi.org/10.5121/ijdps.2011.2618

[25] Thool, R. C. & Ingole, A.(2013). Credit Card Fraud Detection Using Hidden Markov Model and Its Performance. *International Journal of Advanced Research in Computer Science and Software Engineering*, 626-632.

[26] Ayse, B. & Ilker, E. E. D. (2013). A Novel and Successful Credit Card Fraud Detection. *IEEE 13th International Conference on Data Mining Workshops*, 1-10.

[27] Jon, T. S. Q, & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications*, 1721-1732. https://doi.org/10.1016/j.eswa.2007.08.093

[28] Dominik O. (2014). Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems*, 324-334. https://doi.org/10.1016/j.knosys.2014.07.008

[29] Andrea, D. P., Olivier, C., Reid, A. J., & Gianluca, B. (2015). Calibrating Probability with Undersampling for Unbalanced Classification. *IEEE Symposium Series on Computational Intelligence*, 159-166. https://doi.org/10.1109/SSCI.2015.33

[30] Holte, C. & Drummond, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why Under-Sampling beats Over-Sampling. *Workshop on Learning from Imbalanced Datasets II, (ICML),* Washington DC, 1-8.

[31] Stojanovic, A., Aouada, D., Ottersten, B., & Bahnsen, A. C. (2013). Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk. *International Conference on Machine Learning and Applications*, 333-338.

[32] Olivier, C. & Gianluca, B. A. D. P. (2015). When is Undersampling Effective in Unbalanced Classification Tasks?. *Machine Learning and Knowledge Discovery in Databases*, 200-215. https://doi.org/10.1007/978-3-319-23528-8_13

[33] Mohammed, I. A. & Soon, L. K. (2012). Credit Card Fraud Detection: Personalized or Aggregated Model. *International Conference on Mobile, Ubiquitous, and Intelligent Computin*g, 114-119.

[34] Ganesh, K. N. & Sena, P. V. (2013). Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit. *International Journal of Computer Science and Network Security (IJCSNS)*, 58-65.

[35] Cody, S., Muyang, S., Stephen, A., & Peter, B. G. R. (2017). Horse Race Analysis in Credit Card Fraud-Deep Learning, Logistic Regression, and Gradient Boosted Tree. *Systems and Information Engineering Design Symposium (SIEDS)*, pp. 117-121.

[36] Jabir, D. P. & Ali, H. E. A. M. A. Z. K. (2014). Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering. *International Journal of Advanced Research in Computer and Communication Engineering*, 5458-5461.

[37] Djamila, A., Aleksandar, S., & Björn, O. A. C. B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems With Applications*, 134-142. https://doi.org/10.1016/j.eswa.2015.12.030

[38] Tsaih, R. H., Yu, F., & Huang, S. Y. (2014). Topological pattern discovery and feature extraction for fraudulent. *Expert Systems with Applications*, 4360-4372. https://doi.org/10.1016/j.eswa.2014.01.012

[39] Nader, M. & Ekrem, D. (2015). Detecting credit card fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications*, 2510-2516. https://doi.org/10.1016/j.eswa.2014.10.037

[40] Cristián, B., Olivier, C., Tina, E. R., Leman, A., Monique, S., & Bart, B. V. V. V. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 38-48. https://doi.org/10.1016/j.dss.2015.04.013

[41] Serol, B. & Ekrem, D. Y. S. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 5916-5923. https://doi.org/10.1016/j.eswa.2013.05.021

[42] Malini, N. & Pushpa, M. (2017). Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection. *International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics*, 255-258. https://doi.org/10.1109/AEEICB.2017.7972424

[43] Mitali, B. & Suman. (2014). Credit Card Fraud Detection Using Self Organised Map. *International Journal of Information & Computation Technology*, 1343-1348.

[44] Huang J., Zhang, H., & Ling, C. X. (2003). AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. *International Joint Conferences on Artificial Intelligence*, 519-526.

[45] Joseph, K. F. P. (2011). Improving Credit Card Fraud Detection using a Meta-Learning Strategy, thesis ed., Chemical Engineering and Applied Chemistry, Ed.: University of Toronto.

[46] Naive Bayes. (2018, November, 01), Retrieved from http://scikitlearn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes

[47] Leo, B. (2001). Random Forests. Manufactured in The Netherlands, 5-32.

[48] Andrea, D. P. (2015). Adaptive Machine Learning for Credit Card Fraud Detection, Machine Learning Group, Ed.: Université Libre de Bruxelles.

Contact information:

Oğuz ATA, Asst. Prof. Dr.,
(Corresponding author)
Altinbas University, Institute of Science,
Dept. of Information Technologies,
Istanbul, Turkey
E-mail: oguzata@gmail.com, oguz.ata@altinbas.edu.tr

Layth HAZIM, M.Sc of IT,
Tikrit University, Cisco Networking Academy,
Dept. of Computer of Science,
Salah Al-Din, Iraq
E-mail: layth1985it@gmail.com