Efficient Resampling for Fraud Detection During Anonymised Credit Card Transactions with Unbalanced Datasets

Petr Mrozek
College of Science and Engineering
University of Derby
Derby, United Kingdom
p.mrozek1@unimail.derby.ac.uk

John Panneerselvam
School of Informatics
University of Leicester
Leicester, United Kingdom
j.panneerselvam@leicester.ac.uk

Ovidiu Bagdasar
College of Science and Engineering
University of Derby
Derby, United Kingdom
o.bagdasar@derby.ac.uk

***Abstract*— The rapid growth of e-commerce and online shopping have resulted in an unprecedented increase in the amount of money that is annually lost to credit card fraudsters. In an attempt to address credit card fraud, researchers are leveraging the application of various machine learning techniques for efficiently detecting and preventing fraudulent credit card transactions. One of the prevalent common issues around the analytics of credit card transactions is the highly unbalanced nature of the datasets, which is frequently associated with the binary classification problems. This paper intends to review, analyse and implement a selection of notable machine learning algorithms such as Logistic Regression, Random Forest, K-Nearest Neighbours and Stochastic Gradient Descent, with the motivation of empirically evaluating their efficiencies in handling unbalanced datasets whilst detecting credit card fraud transactions. A publicly available dataset comprising 284807 transactions of European cardholders is analysed and trained with the studied machine learning techniques to detect fraudulent transactions. Furthermore, this paper also evaluates the incorporation of two notable resampling methods, namely Random Under-sampling and Synthetic Majority Oversampling Techniques (SMOTE) in the aforementioned algorithms, in order to analyse their efficiency in handling unbalanced datasets. The proposed resampling methods significantly increased the detection ability, the most successful technique of combination of Random Forest with Random Under-sampling achieved the recall score of 100% in contrast to the recall score 77% of model without resampling technique. The key contribution of this paper is the postulation of efficient machine learning algorithms together with suitable resampling methods, suitable for credit card fraud detection with unbalanced dataset.**

Keywords*— *Resampling, Oversampling, Unbalanced, Under-sampling

I. INTRODUCTION

E-commerce and its applications using the Internet have seen an unprecedented growth in recent years. The global online retail reached a staggering value of 1.85 trillion pounds (GBP) in 2017, with a forecast of 4.9 trillion by 2021 [1]. The rise of e-commerce has led many other industries, particularly the financial sector, to use the Internet as their primary medium of business transactions. Herein, data-driven approaches are now increasingly being used to facilitate various services such as online transactions, authenticating users, verifying credit card information, and identifying and preventing fraudulent transactions [2]. Genuine online credit card transactions, which allow transferring of money from the customer's bank accounts to the retailers [3], are pivotal in e-commerce. The increasing number of online transactions also introduces various levels of challenges whilst achieving genuine transactions, amid malicious users intending to capture user and card information. The total amount of fraudulent transactions in 2020 is expected to hit 25 billion pounds and further predicted to exceed 30 billion pounds per annum by 2027 [4].

Worth to note that the number of Card-Not-Present (CNP) frauds has increased by 2.1% just over four years, thus CPN transactions are considered to be the most serious threat for credit card industry [5]. Recently, strategies to analyse, detect and prevent fraud transactions by exploiting various machine learning techniques gains attention. Clustering techniques, classification algorithms, and neural networks are the ones considered to be effective and are becoming widely used for dealing with fraud transactions [6]. The use of advanced statistical learning algorithms, especially supervised and unsupervised machine learning algorithms, facilitates an efficient detection of patterns and anomalies and further provide useful insights extracted from large-scale datasets. Data-driven models for fraud detection offer higher precision of detection, operational efficiency, real-time detection and prevention of frauds, while being cost efficient [7].

Despite the existing efforts aimed at developing efficient data-driven approaches for fraud detection, such as model-based reasoning or descriptive data mining techniques, a reliable model with better practicality is still a requirement. Fraud detection methods and techniques are considered as intellectual assets of banks and other financial institutions [8]. Data-driven methods are bound to various requirements and preserving the user privacy and not disclosing the same is pivotal while analysing relevant information. To this end, a complete dataset that describes all the desired information for performing comprehensive analytics is often not available to data analysts. It is obvious that the quality of the dataset significantly influences to accuracy of data-driven models.

Training datasets are usually generated from hundreds of millions of transactions, often lack the desired sensitive and critical user information, which results in imbalanced datasets. This imbalance means that one or more variables with an instance occur more frequently than the others. Such as imbalance is also known as the skewed distribution and it is a very common issue in classification algorithms [7].

To overcome this issue, the imbalanced datasets can be subjected to various methods of under-sampling and oversampling during the pre-processing stage, so as to arrive at a complete balanced dataset for further advanced processing to produce accurate results. Furthermore, the selection of appropriate variables is another commonly encountered issue, which results from the interchangeable features between fraud and legitimate transactions. Finally, the durability and sustainability of the models adds complexities during the development stage, as the models are affected by trend variation caused by changes in the behaviours of both genuine users and fraudsters [4].

To this end, deploying the most suitable sampling techniques on imbalanced datasets is an important requirement as it significantly affects the overall accuracy and dependability of the detection models. Herein, this paper empirically evaluates the integration of notable sampling techniques such as Random under-sampling and Synthetic Minority Over-sampling Technique (SMOTE) with various machine learning methodologies, such as Logistic Regression, Random Forest Classifier, K-Nearest Neighbour and Stochastic Gradient Descent. The impact and efficiency of these machine learning methodologies for detecting fraud transactions are demonstrated through three implementations: Firstly, the machine learning methodologies are implemented solely without any sampling techniques; secondly, the random under-sampling technique is integrated; and thirdly, the SMOTE resampling is integrated as well with the machine learning methods, respectively. The outcomes of all the three forms of implementations are comprehensively discussed for each of the four studied machine learning methodologies, with the intention to showcase the efficiency of the machine learning models in their sole form, and to assess the impact which sampling techniques can provide when dealing with imbalanced datasets during online fraud detections.

The remainder of the paper is organised as follows: Section 2 reviews some existing works of online fraud detection based on data-driven approaches. Section 3 presents a background on the studied machine learning algorithms, while Section 4 introduces the sampling techniques for imbalanced datasets. Section 5 presents our evaluation strategy, introduces our dataset, and describes the results of our analytics. Section 6 concludes this paper and outlines some future research directions.

II. Background

This section presents the background information about the machine learning techniques evaluated in this paper.

A. Logistic Regression

Logistic Regression (LR) is a statistical linear supervised learning method used for binary classification [9] that works by arranging the calculation used to relegate perceptions, to form a discrete arrangement of classes [10], shown in (1).

It is usually used to predict the patterns in a dataset with unambiguous numeric attributes by performing regression on a group of variables. LR uses a non-linear function, called sigmoid for the prediction, defined as

$$P(X) = \frac{e^{a+bX}}{1+e^{a+bX}} \quad (1)$$

where, P represents the model probability, e is base of the natural logarithm, a and b are the parameters of the model. The value of this function is between 0 and 1. If the result is above 0.5, the classifier sets the predicted variable to 1; if the value is below 0.5, the result is classified as 0. The optimisation method giving the best fit parameters for the input function coefficients is used to train the classifier [11].

B. Random Forest Classifier

Decision tree (DT) models are known for their simplicity and are built based on human reasoning to deal with different attribute types. However, models with a single tree are often sensitive to the data subjectivity and suffer overfitting issues.

To solve this issue, ensemble methods combine multiple tree predictions, with each single tree predictor depending on its own random variable in independent datasets, and all the single trees are in a common forest of the same distribution [12-13]. Random Forest (RF) is a managed supervised technique which ensembles both classification and regression learning methods and it is suitable to solve the problems involving the dataset dividing to classes. A series of Decision Trees is used to predict the class, thus the majority of the voting techniques of all single trees are considered with the final predicted output class set from the most voted class.

Predictive class for new instances is obtained by aggregating every single tree output in the ensemble model. The advantage of RF is its computational efficiency, as each single tree is built independently of each other [14]. Furthermore, [15] emphasize the ability of RF to deal with unbalanced datasets, the common issue related to the dataset with fraudulent credit card transactions. Unbalanced datasets are dealt with by incorporating the highly differenced misclassification costs of credit card detection and implementation of RF bagging ensemble learning model. The model achieving the lowest positive (minority) class error has more emphasis and the tree with the lowest error is the most significant for defining the learning function. Random Forrest classifiers can be represented as in equation (2).

$$RFf_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features,\ k \in all\ trees} normfi_{jk}} \quad (2)$$

where, RFf_i represents the importance of feature I that is calculated from all trees in the model, $normfi_{ij}$ represents the normalised importance of features i and j, *and* $normfi_{jk}$ represents the normalised feature importance for i in tree j.

C. K-Nearest Neighbour

K-Nearest Neighbour classifier (KNN) is a widely used and regarded as a favourable algorithm for detection method and systems. KNN is a supervised statistical machine learning algorithm used for pattern recognition which is achieving consistently high performance while running without prior assumptions about the distribution of training dataset [16] and based on analogy learning [17]. The principle of this method works based on a calculation between two data points, when the Euclidean distance is used as a measurement technique for all instances in the dataset. Next, these distances are arranged in an increasing order [18]. The performance of the KNN algorithm is determined by the following three main factors: 1) the distance used for the location of the nearest neighbours; 2) the distance rule used to deliver a classification from the nearest neighbour; and 3) the k number of neighbours used for classification of the outcome variable [15]. When k = 1, the data point is assigned to the class of its nearest neighbour data point. Therefore, if the nearest neighbour is fraudulent, then the transaction is classified as fraud [10]. According to [16],

the *k* number should be small and odd to break the ties (usually 1, 3 or 5). However, the higher values of *k* are known to reduce the effect of the noisy dataset. KNN can be represented as in equation (3), by the formula

$$D_{(x,y)} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}\,, \tag{3}$$

where, *D* represents the distance between points *x* and *y*, and x_i, y_i represent the particular instances in the dataset.

D. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a popular classification machine learning algorithm which has not gained much emphasis for implementation in the detection of frauds in credit card transactions so far. SGD is an updated Gradient Descent algorithm, when the demand on computation time is decreased by calculation the cost error, corresponding gradient and updates on the weights for one data point are achieved at a time instead of the calculation exerted for the entire dataset simultaneously [19]. SGD can be represented as in equation (4)

$$V_t = \beta V_{t-1} + (1-\beta)\nabla_w L(W, X, y) \tag{4}$$

where, V_t represents the weighted average of the given sequence, β is the hyper-parameter, *L* is the loss function, *W* is the momentum, and ∇ is the weight-dependent gradient.

III. Resampling of Umbalanced Datasets

To deal with imbalanced datasets, resampling techniques are applied to balance the class variable that can be later exploited to determine whether a given transaction is fraudulent or not. The random under-sampling and the Synthetic Minority Oversampling Technique (SMOTE) are the commonly used resampling techniques on imbalanced datasets. This section introduces the aforementioned resampling techniques in detail.

A. Random Under-sampling

Random Under-sampling is a method that works by selecting random examples from the majority class, so as to reduce the number of majority class in such a way to arrive at a balanced training set of equally represented majority and minority classes. It is one of the most popular method based on preserving the minority class and randomly selecting the instances from majority class for building a training set.

The limitation of the method is that some important or even critical information can be deleted from the majority class, as only a limited number of instances are randomly selected from the dataset and useful sampled data near the cross-edge of feature space are potentially overlapped and lost. To overcome this problem, [20] introduce the Gaussian Mixture Under-sampling (GMUS) model implementing the Gaussian Mixture Model (GMM) which select the instances near the cross-edge area of data distribution and thus ensure

Fig. 1. Random Under-Sampling

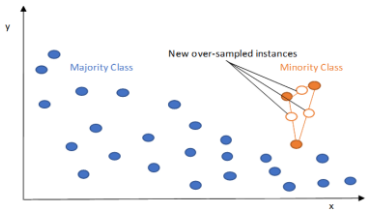


Fig. 2. SMOTE Resampling

that useful data points remain and will be sampled. GMM is an approach used for enabling the measure of uncertainty or probability that a data point belonging to a specific class using a soft classification, as an alternative method to hard classification methods, such as K-means where each data point is associated with one and only class [21].

This method has proven to be efficient in solving the over-fitting issue and increased the recognition rate of minority samples [20]. Generally, the Random Under-sampling technique is appropriate for the cases where there is a sufficient number of instances in the minority class [22].

B. SMOTE Resampling

Synthetic Minority Over-sampling Technique (SMOTE) works on the contrary to random under-sampling, as it oversamples the minority class instead of deleting the majority class [23]. The SMOTE resampling method uniquely selects the instances that require resampling, in order to eliminate overfitting issues witnessed in the traditional oversampling methods. Consider the scenario presented in Fig. 2, where the majority class (represented in blue) has a leading control on both axes. The SMOTE method draws synthetic lines between the data points of minority class and places the new instances on these lines. When the dataset contains more minority instances than shown in the graph, synthetic lines can origin diagonally to the opposite data points, resulting in a larger number of synthetic lines from one data point. The number of lines and new instances are determined by the model parameters in a similar way to the K-Nearest Neighbours algorithm, where the K parameter define the number of neighbours that should be accounted in the model [24]. The limitations come from datasets with high dimensional space, and the appropriate feature engineering reducing the number of variables is beneficial. The SMOTE technique performs with higher efficiency in a lower dimensional space, as the

converters will be closer to each other and resampling would be more accurate [25].

IV. Experiments

A. Experimental Strategy

Measuring the performance of the Machine Learning algorithms is pivotal for any analytics driven application. The unbalanced class dataset issue requires a specific approach for measuring the performance of the trained algorithms. Suppose an unbalanced dataset with 1000 instances encompasses 10 instances of minority class, i.e., 1% of the total instances. If the model classifies all the instances as majority class, the model can score with excellent predicting accuracy. Therefore, the following performance evaluation parameters are used to appropriately capture, visualise, and measure the results of predictive models.

- *Confusion Matrix* is an essential tool for visualising the prediction outcome. This is a matrix demonstrating the number of predicted outcomes for both classes compared to the actual number of both classes in the dataset.
- *Area Under the Receiver Operating Characteristics (AUROC)* is a performance measurement technique frequently used for describing the performance of classification tasks. The chart demonstrates both Area Under Curve (AUC) and Receiver Operating Characteristics (ROC). For better understanding, it is important to define the terms used in the chart.
- *F1 Score* determines the precision of the classifier by evaluating the proportion of instances predicted correctly against the proportion of significant instances those missed (Mishra, 2018).

B. Dataset

The dataset used for the analysis contains credit card transactions collected in September 2013 from European cardholders [26]. The dataset has a total of 284.807 transactions, of which 492 are identified as fraudulent. Therefore, the dataset is highly skewed, with the fraud transactions proportion of 0.172% of all the transactions. The dataset is subjected to a split of 70:30, for preparing the training and testing datasets respectively. Additionally, a K-Fold Cross-Validation is performed where appropriate, whilst testing the efficiencies of the machine learning models in order to achieve an unbiased comparison.

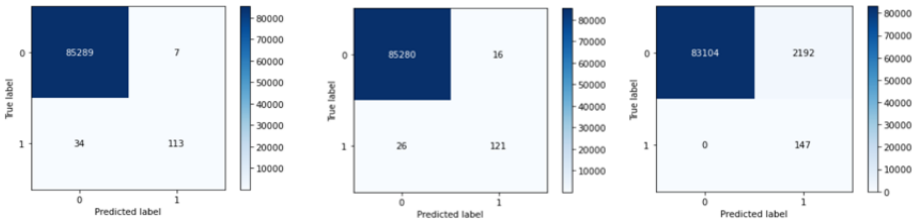
Fig. 4. Transactions Class Distribution

C. Evaluation of Logistic Regression

Fig. 3 and Fig. 4 illustrate the confusion matrix and the ROC curve comparison of Logistic Regression without resampling, with SMOTE resampling and with random under-sampling respectively.

1) Without Resampling

Without the resampling method, the logistic regression model demonstrates a typical trend in binary classification tasks with an unbalanced dataset. On the one hand, the accuracy score is very high as most of the cases are classified correctly. On the other hand, the model's ability to detect fraud transactions is arguable, as more than one-third of fraud transactions labelled as legitimate. The model's performance metrics are observed at an accuracy of 99.89%, precision of 73.46%, recall of 64.62% and F1 score of 68.84%. The AUC score 0.82 determines that there is an 82% chance that this model correctly distinguishes between the labels. With Cross Fold validation with 5 folds, the average accuracy score is observed at 99.57%, the average precision is 71.27%, average recall score 62.08% and average F1 score is 54.10%.



(a) (b) (c)

Fig. 3. Confusion Matrix for RF Classifier (a) without resampling (b) SMOTE resampling (c) Random Under-sampling

2) SMOTE Resampling

The SMOTE resampling method significantly increased the logistic regression model's ability to detect fraudulent transactions, as only one-seventh of all fraudulent cases remained uncovered. However, besides the fraudulent transactions, model incorrectly labelled more than 1500 legitimate transactions as a fraud. Thus, with an increased fraud detection ability, the number of incorrectly labelled legitimate transactions is increasing too. The model's performance metrics are observed at an accuracy of 98.14%, precision of 7.55%, recall of 87.07% and F1 score of 13.90%. An AUC score of 0.92 demonstrates that there is a 92% chance that the model correctly distinguishes between the labels. Surprisingly, the model's evaluation metrics have increased significantly after the Cross Fold validation. With Cross Fold validation with 5 fold, the average accuracy score is 97.17%, the average precision is 98.02%, average recall score is 96.29% and average F1 score is 97.15%.

3) Random Under-sampling Method

The Logistic Regression model with Random Under-sampling resampling method improved its ability to detect fraudulent transactions compared to both without resampling and with SMOTE resampling integration. Only 12 fraudulent transactions remanined undetected in the dataset, however, the number of incorrectly labelled legitimate transactions as fraud has at least doubled, compared to the case without resampling and the SMOTE resampling. The model's performance metrics are observed at an accuracy of 96.43%, precision of 4.22%, recall of 91.15% and F1 score of 8.07%. The AUC score 0.95 determines that there is a 95% chance that the model correctly distinguishes between the labels. This model also scores significantly higher with Cross Fold validation. With Cross Fold validation with 5 folds, the average accuracy score is 92.47%, the average precision is 94.93%, average recall score 89.84% and average F1 score is 92.23%.

Performance evaluation has shown that the logistic regression model without resampling method achieves balanced scores for all the measured performance metrics. The Random Under-sampling method had the highest ability to detect fraud transactions despite a large number of legitimate transactions incorrectly labelled as fraud.

D. Evaluation of Random Forest Classifier

On a coarse grain, the RF classifier outperformed the Logistic Regression by demonstrating a higher fraud detection ability. Subsequently, the RF classifier performed efficiently, with a shorter computation time, even with the integration of the SMOTE resampling methodology. Fig. 5 and Fig. 6 show the confusion matrix and the ROC curve comparison of Logistic Regression without resampling, with SMOTE resampling and with random under-sampling respectively.

1) Without Resampling

The RF model showed a higher fraud detection ability despite the imbalance in the dataset, without any resampling. Just a fourth of fraud transactions remained undetected and the model also characterised accuracy in terms of labelling legitimate transactions correctly. In general, the RF model is adequately balanced and accurate. The model's performance metrics are observed at an accuracy of 99.95%, precision of 94.16%, recall of 76.87% and F1 score of 84.64%. The AUC score of 0.88 demonstrates that there is a 91% chance that model correctly distinguishes between the labels.

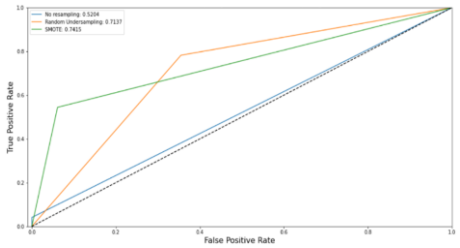
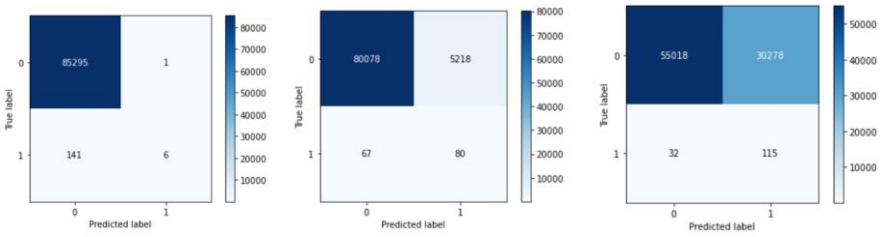
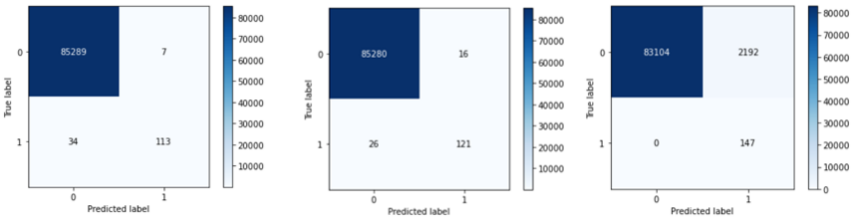


Fig. 8. ROC Curve for KNN Models



(a) (b) (c)

Fig. 7. Confusion Matrix for KNN (a) without resampling (b) SMOTE resampling (c) Random Under-sampling



(a) (b) (c)

Fig. 5. Confusion Matrix for RF Classifier (a) without resampling (b) SMOTE resampling (c) Random Under-sampling

2) SMOTE Resampling Method

The confusion matrix of RF classifier without resampling method and SMOTE resampling method are observed to be very similar. The RF classifier with the SMOTE resampling technique achieved a slightly better fraud detection ability then without any resampling, on the other hand, it characterises more False Positive cases. One-fifth of the fraudulent transactions remained uncovered. The model's performance metrics are observed at an accuracy of 99.95%, precision of 89.70%, recall of 82.99% and F1 score of 86.21%. The AUC score of 0.91 determines that there is a 91% chance that model correctly distinguishes between the labels.

3) Random Under-sampling Method

The Random Forest model with the under-sampling methodology achieved a near perfect score in terms of detecting fraudulent transactions. All of the 147 fraud transactions are identified, despite more than 2000 incorrectly

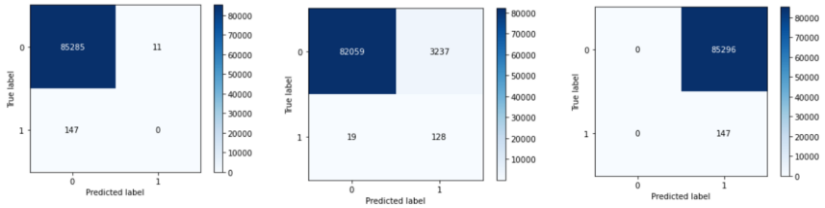
labelled legitimate transactions, this can be regarded as a convincing result. In the real-world scenario in the banking industry, it is probably more important to detect the fraudulent activities and proactively block such transactions. In the case of an incorrectly blocked legitimate transaction, the transaction can be approved again upon further verification. The model's performance metrics are observed at an accuracy of 97.43%, precision of 6.28%, recall of 100% and F1 score of 11.82%. The AUC score of 0.98 shows that there is a 98% chance for the model to correctly distinguish between labels. The ROC curve of Random Forest classifier with Random Under-sampling resampling method is demonstrated below.

The Random Forest classifier with the random under-sampling methodology comes as the first choice among the studied versions of the RF classifiers, as it discovered all the fraudulent transactions within the dataset. Nevertheless, both the other methods also achieved considerable results, with the RF without resampling method eventually outperformed the SMOTE methodology. To conclude, it is obvious that in the case of Random Forest classifiers, the under-sampling approach overcomes the issues of oversampling in binary classification task whilst dealing with an imbalanced dataset.

E. Evaluation of the KNN Model

Fig. 7 and Fig. 8 illustrate the confusion matrix and the ROC curve comparison of the KNN model without resampling, with SMOTE resampling and with random under-sampling respectively.

1) Without Resampling



(a) (b) (c)

Fig. 9. Confusion Matrix for SGD (a) without resampling (b) SMOTE resampling (c) Random Under-sampling

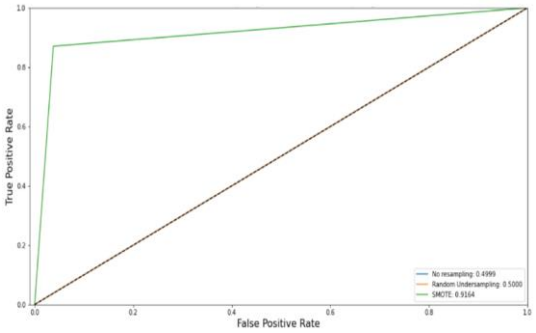


Fig. 10. ROC Curve for SGD Models

The KNN model without resampling has demonstrated the typical trend of classification algorithms when dealing with an unbalanced dataset, where the model labels most of the minority instances as majority class. Such a model usually characterises a very low ability to train as a fraud detector and it's practically is highly arguable in the banking industry. The model's performance metrics are observed at an accuracy of 99.83%, precision of 85.71%, recall of 4.08% and F1 score of 7.79%. The AUC score of 0.52 demonstrated that there is a 52% chance that the model correctly distinguishes between the labels. The performance metrics in accuracy and precision can be misleading, as this model loses the ability of initial requirement of classification between the data variables.

2) SMOTE Resampling Method

The KNN model with SMOTE resampling demonstrated average results under all the measured aspects. It has detected more than half of the fraudulent transactions and there is no significant number of incorrectly labelled False Negative cases with regards to the overall number of transactions. However, there has already been presented with more efficient and competent models with enhanced detection ability and lower misclassification error. The model's performance metrics are observed at an accuracy of 93.81%, precision of 1.51%, recall of 54.42% and F1 score of 2.93%. The AUC score of 0.74 demonstrates that there is a 74% chance that model correctly distinguishes between the labels.

3) Random Under-sampling Method

The KNN model with Random Under-sampling improves its ability to detect the fraud transactions, as only one-fifth of the fraudulent instances remained uncovered. However, the number of incorrectly labelled legitimate transactions as fraud is very high, with more than 30 thousand instances with errors. Undoubtedly, this model would not be useful in practice, although a higher proportion of detected fraudulent transactions. The number of incorrectly detected and potentially blocked transactions would negatively impact the customer's experience. The model's performance metrics scored are observed at an accuracy of 64.52%, precision of 0.37%, recall of 78.23% and F1 score of 0.75%. The AUC score of 0.71 demonstrates that there is a 71% chance that model correctly distinguishes between the labels.

The K-Nearest Neighbour classification model has not scored well compared to the models studied so far. Random under-sampling and SMOTE resampling improved the performance of the model significantly, but these models are not suited for credit card detection, as the ability to efficiently detect the fraudulent transaction is low, with a higher misclassification error for legitimate transactions.

F. Evaluation of Stochastic Gradient Descent

Fig. 9 and Fig. 10 illustrate the confusion matrix and the ROC curve comparison of the SGD model without resampling, with SMOTE resampling and with random under-sampling respectively.

1) Without Resampling Method

The SGD model without resampling method scored poorly, as the model labels most of the instances as legitimate transactions and all the fraudulent transactions were labelled incorrectly. Therefore, the model has very poor ability to

detect the credit card frauds, despite of measured higher statistical accuracy. The model's performance metrics scored are observed at an accuracy of 99.81%, precision of 0%, recall of 0% and F1 score of 0%. The AUC score of 0.49 demonstrates that there is a 49% chance that model correctly distinguishes between the labels.

2) SMOTE Resampling Method

The SGD model with SMOTE resampling has achieved significantly improved results. Most of the fraudulent instances are detected, with only one-sixth of the fraud transactions remaining undetected. The misclassification error for False Negative instances has increased too, but overall performance is acceptable in comparison with previously presented models. The model's performance metrics are observed at an accuracy of 96.18%, precision of 3.80%, recall of 87.07% and F1 score of 7.28%. The AUC score of 0.91 demonstrates that there is a 91% chance that model correctly distinguishes between the labels.

3) Random Under-sampling Method

The SGD model with random under-sampling resampling methodology has proved to be useless, as it demonstrates a very poor ability to distinguish between the classes. It labels all the instances as fraudulent transactions. The model's performance metrics are observed at an accuracy of 0.17%, precision of 0.17%, recall of 100% and F1 score of 0.34%. The AUC score of 0.5 demonstrates that there is a 50% chance that the model correctly distinguishes between the labels. This model can hardly focus on a single performance indicator, and an overall consideration of measurement aspects is necessary for model evaluation.

TABLE I. PERFORMANCE COMPARISON OF THE MODELS

Model	Accuracy	Precision	Recall	F1 Score	AUC
LogReg without resampling	99.89%	73.46%	64.62%	68.84%	0.82
LogReg SMOTE resampling	98.14%	7.55%	87.07%	13.90%	0.92
LogReg Under-sampling	96.43%	4.22%	91.15%	0.08%	0.95
RFC without resampling	99.95%	94.16%	76.87%	84.64%	0.88
RFC SMOTE resampling	99.95%	89.70%	82.99%	86.21%	0.91
RFC Under-sampling	**97.43%**	**0.06%**	**100.00%**	**11.82%**	**0.98**
KNN without resampling	99.83%	85.71%	4.08%	7.79%	0.52
KNN SMOTE resampling	93.81%	1.51%	54.42%	2.93%	0.74
KNN Under-sampling	64.52%	0.37%	78.23%	0.75%	0.71
SGD without resampling	99.81%	0%	0%	0%	0.49
SGD with SMOTE resampling	96.18%	3.80%	87.07%	7.28%	0.91
SGD with Under-sampling	0.17%	0.17%	100%	0.34%	0.5

In contrast to previous evaluations, the random under-sampling seems to degrade the performance of the SGD model. Surprisingly, the oversampling approach works well.

G. Summary

Table 1 illustrates the performance metrics of all the studied models. Performance evaluations shown that the best score is achieved the Random Forest classifier with the integration of Random Under-sampling resampling strategy, with outstanding AUC score of 0.98. In other words, the model has a probability of 98% to correctly classify between the labels corresponding for fraudulent and legitimate transactions. This model discovered all the fraudulent transactions resulting in a 100% recall score. Logistic Regression integrated with Random Under-sampling and SMOTE resampling achieved the second and third best AUC scores of 0.95 and 0.92 respectively, demonstrating that the Logistic Regression is a simple but powerful tool for binary classification tasks. It is worthwhile to note that the resampling strategies should depend on the chosen model's characteristics and there is no optimal approach that can optimise the performance of all the models.

V. RELATED WORKS

Online credit card fraud detection with machine learning techniques is gaining its attention due to the loss of revenue caused by credit card fraud. The most common data-driven strategies in the state-of-art to detect credit card fraud use a combination of machine learning algorithms such as Random Forest classifier, Support Vector Machine (SVM) and clustering methods including K-Nearest Neighbour. Furthermore, deep learning techniques are also increasingly implemented either individually or in cooperation with other methods. For instance, a series of nonlinear processing units are integrated into layers for variable extraction and transformation [27]. The KNN algorithm is deemed to be powerful and accurate when implemented for a range of classification problems [14][17][27]. [29] demonstrated the performance consistency of the KNN algorithm whilst delivering a higher classification accuracy.

The RF classifier is also regarded as a valuable and suitable technique especially for binary unbalanced dataset [30]. The RF classifier was shown to perform better under skewed distributions [31]. Surprisingly, the logistic regression technique is often underrated, owing to its average performance in the terms of its accuracy and precision, and often significantly outperformed by the other algorithms [32][11]. A classification algorithm usually focuses more on the majority class as the intention of the model is to minimise the overall classification error [33].

Resampling techniques including both the oversampling and under-sampling have been used in imbalanced datasets to overcome such issues around the classification. [23] postulated SMOTE as an efficient strategy to resampling imbalanced datasets, and demonstrated that SMOTE can outperform the random oversampling (ROS). While the SMOTE model is acknowledged for its ability to randomly replicate the minority class to balance the dataset, it is also addressed to result in overfitting the classifier [34]. The under-sampling strategy has proven to perform better than oversampling, as the former is claimed to work better on unbalanced datasets [20]. The random under-sampling method (RUS) is one of the most popular method that work by preserving the minority class and randomly selecting the instances from majority class for building a training set. But, the randomness in the RUS model has been a common issue,

and to overcome this problem, [20] introduced the Gaussian Mixture Under-sampling (GMUS) model by implementing the Gaussian Mixture Model (GMM), which select the instances near the cross-edge area of data distribution to ensure that useful data points remain and are sampled. GMM approach enable the measure of uncertainty or probability belonging to a specific class using a soft classification as an alternative method to hard classification methods, such as K-means where each data point is associated with one and only class [21]. This method has proven to be efficient in solving the over-fitting issue and increased the recognition rate of minority samples [20].

Despite the existing efforts of detection methods using machine learning algorithms, an efficient model with the ability to balance the unbalanced datasets is still a lacking. In the same way, many of the researched sources use the dataset which is not available to the public [15], [35]-[36]. Thus this paper emphasises the need for more open research in the context of online fraud detection, particularly when with unbalanced dataset that characterises ambiguity.

VI. Conclusion

This paper empirically evaluated four different machine learning techniques solely and in integration with notable resampling techniques to evaluate their efficiencies whilst classifying online credit card fraud transactions with unbalanced dataset. The dataset is composed of a total 284.807 transactions, where fraudulent transactions accounted only for 0.172% or 492 transactions, providing highly unbalanced class distribution. While all the studied machine learning algorithms exhibited their analytics efficiency to different extent, this paper identified the most optimal result is achieved with the Random Forest classifier model when applied in integration with the Random Under-sampling Resampling methodology, as this model achieved a 100% recall accuracy and an AUC score of 0.98. A Logistic Regression model along with the Random Under-sampling and SMOTE Resampling methods respectively also exhibited considerable performances after the RF classifier model. As a future work, we plan to build a novel machine learning model with suitable resampling techniques for a generalised implementation in credit card fraud transaction analytics with diverse range of heterogeneous datasets.

References

[1] Y. Vakulenko, P. Shams, D. Hellström and K. Hjort, "Online retail experience and customer satisfaction: the mediating role of last mile delivery", The International Review of Retail, Distribution and Consumer Research, vol. 29, no. 3, pp. 306-320, 2019. Available: 10.1080/09593969.2019.1598466.

[2] Q. Farooq, P. Fu, Y. Hao, T. Jonathan and Y. Zhang, "A Review of Management and Importance of E-Commerce Implementation in Service Delivery of Private Express Enterprises of China", SAGE Open, vol. 9, no. 1, p. 215824401882419, 2019. Available: 10.1177/2158244018824194.

[3] U. Porwal and S. Makund, "Credit Card Fraud Detection in E-commerce", in International Conference On Trust, Security And Privacy In Computing And Communications, 2019.

[4] Nilson Report – Card Fraud Losses Reach $27.85 Billion", Nilsonreport.com, 2020. [Online]. Available: https://nilsonreport.com/mention/407/1link.

[5] A. Yeşilkanat, B. Bayram, B. Köroğlu and S. Arslan, "An Adaptive Approach on Credit Card Fraud Detection Using Transaction Aggregation and Word Embeddings", in IFIP International Conference on Artificial Intelligence Applications and Innovations. Istanbul: Applied AI and R&D Department, 2020.

[6] S. Georgieva, M. Markova and V. Pavlov, "Using neural network for credit card fraud detection", American Institut of Physics, 2019.

[7] S. Makki, "An Efficient Classification Model For Analyzing Skewed Data To Detect Frauds In The Financial Sector", PhD, Universite de Lyon, 2019.

[8] P. Shimpi, "Survey on Credit Card Fraud Detection Techniques", International Journal Of Engineering And Computer Science, 2016. Available: 10.18535/ijecs/v4i11.25.

[9] R. Popat and J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning", in International Conference on Trends in Electronics and Informatics (ICOEI), 2018.

[10] S. Sasank, R. Sahith, K. Abhinav and M. Belwal, "Credit Card Fraud Detection Using Various Classification and Sampling Techniques: A Comparative Study", in International Conference on Communication and Electronics Systems (ICCES), 2019.

[11] A. Nadim, I. Sayem, A. Mutsuddy and M. Chowdhury, "Analysis of Machine Learning Techniques for Credit Card Fraud Detection", in International Conference on Machine Learning and Data Engineering (iCMLDE), 2019.

[12] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random Forest for Credit Card Fraud Detection", in IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018.

[13] S. Shubho, R. Razib, N. Rudro, A. Saha, S. Khan and S. Ahmed, "Performance Analysis of NB Tree, REP Tree and Random Tree Classifiers for Credit Card Fraud Data", in International Conference on Computer and Information Technology (ICCIT), 2019.

[14] D. Prusti and S. Rath, "Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning techniques", in International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019.

[15] D. Devi, S. Biswas and B. Purskayastha, "A Cost-sensitive weighted Random Forest Technique for Credit Card Fraud Detection", in International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019.

[16] S. Darwish, "A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking", Journal of Ambient Intelligence and Humanized Computing, 2020. Available: 10.1007/s12652-020-01759-9.

[17] S. Rajora et al., "Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance", in Symposium Series on Computational Intelligence (SSCI), 2018.

[18] A. Singh and A. Jain, "An Empirical Study of AML Approach for Credit Card Fraud Detection–Financial Transactions", International Journal of Computers Communications & Control, vol. 14, no. 6, pp. 670-690, 2019. Available: 10.15837/ijccc.2019.6.3498.

[19] S. Dutta, "Why Stochastic Gradient Descent Works?", Medium, 2019. [Online]. Available: https://towardsdatascience.com/https-towardsdatascience-com-why-stochastic-gradient-descent-works-9af5b9de09b8.

[20] F. Zhang, G. Liu, Z. Li, C. Yan and C. Jiang, "GMM-based Under-sampling and Its Application for Credit Card Fraud Detection", in International Joint Conference on Neural Networks (IJCNN), 2019.

[21] O. Carrasco, "Gaussian Mixture Models Explained", Medium, 2019. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95.

[22] [21]M. Bach, A. Werner and M. Palt, "The Proposal of Under-sampling Method for Learning from Imbalanced Datasets", Procedia Computer Science, vol. 159, pp. 125-134, 2019. Available: 10.1016/j.procs.2019.09.167.

[23] S. Taneja, B. Suri and C. Kothari, "Application of Balancing Techniques with Ensemble Approach for Credit Card Fraud Detection", in International Conference on Computing, Power and Communication Technologies (GUCON), 2019.

[24] R. Kunert, "SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line · Rich Data", Rikunert.com, 2017. Available: https://rikunert.com/SMOTE_explained.

[25] M. Jagelid and M. Movin, "A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning", Bachelor, KHT Royal Institute of Technology, 2017.

[26] A. Pozzolo, R. Johnson, O. Caelen and G. Bontempi, "Calibrating Probability with Under-sampling for Unbalanced Classification", in Symposium Series on Computational Intelligence (SSCI), 2015.

[27] N. Shirodkar, P. Mandrekar, R. Mandrekar, R. Sakhalkar, K. Kumar and S. Aswale, "Credit Card Fraud Detection Techniques – A Survey", in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.

[28] S. Padmanabhuni, A. Kandukuri, D. Prusti and S. Rath, "Detecting Default Payment Fraud in Credit Cards", in IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), 2019.

[29] I. Sadgali, N. Sael and F. Benabbou, "Adaptive Model for Credit Card Fraud Detection", International Journal of Interactive Mobile Technologies (iJIM), vol. 14, no. 03, 2020. Available: 10.3991/ijim.v14i03.11763.

[30] S. Salazar, G. Safont and L. Vergara, "A new method for fraud detection in credit cards based on transaction dynamics in subspaces", in International Conference on Computational Science and Computational Intelligence (CSCI), 2019.

[31] O. Ata and L. Hazim, "Comparative Analysis of Different Distributions Dataset by Using Data Mining Techniques on Credit Card Fraud Detection", Tehnicki vjesnik - Technical Gazette, vol. 27, no. 2, 2020. Available: 10.17559/tv-20180427091048.

[32] H. Najadat, O. Altiti, A. Aqouleh and M. Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning", in International Conference on Information and Communication Systems (ICICS), 2020.

[33] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep Learning Detecting Fraud in Credit Card Transactions", in Systems and Information Engineering Design Symposium (SIEDS), 2018.

[34] H. Zhou, L. Wei, G. Chen, P. Lin and Z. Lin, "Credit Card Fraud Identification Based on Principal Component Analysis and Improved Adaboost Algorithm", in International Conference on Intelligent Computing, Automation and Systems (ICICAS), 2019.

[35] I. Sadgali, N. Sael and F. Benabbou, "Fraud detection in credit card transaction using neural networks", in International Conference on Smart City Applications, 2019.

[36] L. Zheng et al., "A New Credit Card Fraud Detecting Method Based on Behaviour Certificate", in International Conference on Networking, Sensing and Control (ICNSC), 2020